Johanna Bertl, Gregory Ewing, Carolin Kosiol and Andreas Futschik*

Approximate maximum likelihood estimation for population genetic inference

https://doi.org/10.1515/sagmb-2017-0016

Abstract: In many population genetic problems, parameter estimation is obstructed by an intractable likelihood function. Therefore, approximate estimation methods have been developed, and with growing computational power, sampling-based methods became popular. However, these methods such as Approximate Bayesian Computation (ABC) can be inefficient in high-dimensional problems. This led to the development of more sophisticated iterative estimation methods like particle filters. Here, we propose an alternative approach that is based on stochastic approximation. By moving along a simulated gradient or ascent direction, the algorithm produces a sequence of estimates that eventually converges to the maximum likelihood estimate, given a set of observed summary statistics. This strategy does not sample much from low-likelihood regions of the parameter space, and is fast, even when many summary statistics are involved. We put considerable efforts into providing tuning guidelines that improve the robustness and lead to good performance on problems with high-dimensional summary statistics and a low signal-to-noise ratio. We then investigate the performance of our resulting approach and study its properties in simulations. Finally, we re-estimate parameters describing the demographic history of Bornean and Sumatran orang-utans.

Keywords: approximate inference; isolation-migration model; maximum likelihood estimation; orang-utans; population genetics; stochastic approximation.

1 Introduction

Both in the Bayesian as well as in the frequentist framework, statistical inference commonly uses the likelihood function. Under a wide range of complex models however, no explicit formula is available. An important example is the coalescent process that is commonly used to model the evolutionary history of a sample of DNA sequences from a population (Marjoram and Tavaré, 2006). Even for relatively simple demographic models, the likelihood function consists of a computationally infeasible number of terms (Stephens, 2007). Similar situations occur with dynamical systems as used for example in systems biology (Toni et al., 2009) and epidemiology (McKinley et al., 2009), but also in other fields like spatial statistics (Soubeyrand et al., 2009) and queuing systems (Heggland and Frigessi, 2004). Here, we consider statistical models with an intractable distribution theory, but a known data generating process under which data can be simulated.

A recent approach to overcome this problem is Approximate Bayesian Computation (ABC) (Beaumont et al., 2002). In its most basic form, it can be described by the following rejection algorithm: parameter values are randomly drawn from the prior distribution, data sets are then simulated under these values. To reduce

***Corresponding author: Andreas Futschik,** Department of Applied Statistics, Johannes Kepler University Linz, Altenberger Straße 69, 4040 Linz, Austria; and UC Santa Barbara, Kavli Institute for Theoretical Physics, CA 93106, USA, e-mail: andreas.futschik@jku.at. http://orcid.org/0000-0002-7980-0304

Johanna Bertl: Department of Molecular Medicine (MOMA), Aarhus University Hospital, Palle Juul-Jensens Boulevard 99, 8200 Aarhus N, Denmark

Gregory Ewing: École polytechnique fédérale de Lausanne, 1015 Lausanne, Switzerland

Carolin Kosiol: Centre for Biological Diverstity, University of St Andrews, St Andrews, Fife KY16 9TH, UK; and Institute of Population Genetics, Vetmeduni Vienna, Veterinärplatz 1, 1210 Wien, Austria

complexity, informative but low dimensional summaries are derived from the data sets. All parameter values that gave rise to summary statistics sufficiently similar to those computed for the observed data are then accepted as a sample from the posterior distribution. Hence, with a uniform prior (with support on a compact subset of the parameter space taken large enough to contain the maximum likelihood estimator), ABC can be used to obtain a simulation-based approximation of the likelihood surface and, subsequently, the maximum likelihood estimate for the observed summary statistic values (Rubio and Johansen, 2013; see de Valpine, 2004 for a Markov Chain Monte Carlo (MCMC) approach).

The need for inferential tools in population genetics that allow for the analysis of large genomic datasets was a major driver in the development of ABC, and it has successfully been used in population genetic applications and also in other fields, see e.g. Beaumont (2010). However, the sampling scheme can be inefficient in high-dimensional problems: the higher the dimension of the parameter space and the summary statistics, the lower is the acceptance probability of a simulated data set. Consequently, more data sets need to be simulated to achieve a good estimate of the posterior distribution. Alternatively, the acceptance threshold can be relaxed, but this increases the bias in the approximation to the posterior distribution. To reduce the sampling burden, combinations of ABC and iterative Monte Carlo methods such as particle filters and MCMC have been proposed that focus the sampling on the relevant regions of the parameter space (Beaumont et al., 2009; Wegmann et al., 2009).

Here, we follow an alternative approach to obtain an approximate maximum likelihood estimate. Instead of using random samples from the entire parameter space, we adapt stochastic approximation methods and propose two algorithms to approximate the maximum likelihood estimate. Similar to ABC, they rely on lower-dimensional summary statistics of the data. With the classical stochastic gradient algorithm in k dimensions, the gradient is approximated in each iteration by $2k$ evaluations of the likelihood (Kiefer and Wolfowitz, 1952; Blum, 1954). For each parameter value of interest, they can be obtained by kernel density estimation on summary statistics of simulated datasets. Alternatively, we use a simultaneous perturbations algorithm, where independent of the parameter dimension only two noisy measurements are sufficient to obtain an ascent direction (Spall, 1992). Stochastic gradient methods have also received recent interest in the context of regression, as documented for instance by theoretical studies of Dieuleveut and Bach (2016); Bach (2014).

Our approach is related to a method suggested in Diggle and Gratton (1984) (see also Fermanian and Salanié, 2004). There, an approximate maximum likelihood estimate is obtained using a stochastic version of the Nelder-Mead algorithm. However, the authors explore only applications to one-dimensional *i.i.d.* data. In the context of indirect inference, Creel and Kristensen (2013) propose a simulated maximum indirect likelihood (SMIL) estimator that is also based on approximations of the likelihood by kernel density estimation on simulated summary statistics, but they do not use stochastic approximation for obtaining estimates. In Meeds et al. (2015), a (Bayesian) MCMC algorithm is proposed where stochastic gradients are obtained in a similar fashion as here. In the setting of hidden Markov models, Ehrlich et al. (2013) have proposed a recursive maximum likelihood algorithm that also combines ABC methodology with the simultaneous perturbations algorithm.

Here we focus on population genetic inference which is often quite challenging. In principle however, our approach can also be applied in other contexts. We briefly mention that the proposed algorithms could also be applied in connection with indirect inference, where summary statistics are derived from a tractable auxiliary model (Drovandi et al., 2011).

2 Method

We start by describing two approximate maximum likelihood (AML) algorithms for optimizing the expected value of a random function using noisy observations, the AML-finite difference (FD) and AML-stochastic perturbations (SP), respectively. Next, we explain how these algorithms can be adapted to obtain maximum likelihood estimates in complex models with intractable likelihood functions.

2.1 Maximization by stochastic approximation

Here, we summarize stochastic approximation in its classic form. In Subsection 2.2., we discuss how stochastic approximation can be adopted to obtain maximum likelihood estimates.

Let $Y \in \mathbb{R}$ be a random variable depending on $\Theta \in \mathbb{R}^p$. The function $L(\Theta) = E(Y|\Theta)$ is to be maximized in Θ, but $L(\Theta)$ as well as the gradient $\nabla L(\Theta)$ are unknown. If realizations $y(\Theta)$ of $Y|\Theta$ can be obtained for any value of Θ, $\arg\max_{\Theta\in\mathbb{R}^p} L(\Theta)$ can be approximated by stochastic approximation methods (see Spall, 2003, sections 6 and 7, for an overview).

Similar to gradient algorithms in a deterministic setting, the stochastic approximation algorithms are based on the recursion

$$\Theta_n = \Theta_{n-1} + a_n \nabla L(\Theta_{n-1}) \quad \text{for } n \in \mathbb{N}, \text{ and a decreasing sequence } a_n \in \mathbb{R}^+$$

starting with some $\Theta_0 \in \mathbb{R}^p$.

The unknown gradient can be substituted by an approximation based on observed finite differences. For $\Theta \in \mathbb{R}$ this algorithm was first described in Kiefer and Wolfowitz (1952), followed by a multivariate version in Blum (1954), where the gradient is approximated by finite differences in each dimension. In iteration n, each element $l = 1, \ldots, p$ of the gradient is approximated by

$$\left(\hat{\nabla}_{c_n} L(\Theta_n)\right)^{(l)} = \frac{y(\Theta_n + c_n e_l) - y(\Theta_n - c_n e_l)}{2c_n},$$

where e_l is the lth unit vector of length p, $c_n \in \mathbb{R}^+$ is a decreasing sequence and $y(\Theta_n + c_n e_l)$ and $y(\Theta_n - c_n e_l)$ are independent realizations of $Y|\Theta_n + c_n e_l$ and $Y|\Theta_n - c_n e_l$, respectively.

Thus, for each iteration of this algorithm, $2p$ observations of Y are needed. A computationally more efficient method was introduced by Spall (1992): The finite differences approximation of the slope is only obtained along one direction that is randomly chosen among the vertices of the unit hypercube, so only two observations per iteration are necessary. More specifically, in iteration n, a random vector with elements

$$\delta_n^{(l)} = \begin{cases} -1 & \text{with probability } 1/2, \\ +1 & \text{with probability } 1/2, \end{cases}$$

for $l = 1, \ldots, p$ is generated. Then, the gradient is approximated by

$$\hat{\nabla}_{c_n} L(\Theta_n) = \delta_n \frac{y(\Theta_n + c_n\delta_n) - y(\Theta_n - c_n\delta_n)}{2c_n}.$$

Spall showed that for a given number of simulations this algorithm reaches a smaller asymptotic mean squared error in a large class of problems than the original version. Phrased differently, fewer realizations of Y are usually necessary to reach the same level of accuracy. See Spall, 2003, Section 7, for a more detailed overview.

2.2 Approximate maximum likelihood (AML) algorithms

Suppose, data D_{obs} are observed under model $\mathcal{M}$ with unknown parameter vector $\Theta \in \mathbb{R}^p$. Let $L(\Theta; D_{\text{obs}}) = p(D_{\text{obs}}|\Theta)$ denote the likelihood of Θ. For complex models often there is no closed form expression for the likelihood, and for high dimensional data sets, the likelihood can be difficult to estimate. As in ABC, we therefore consider $L(\Theta; s_{\text{obs}}) = p(s_{\text{obs}}|\Theta)$, an approximation to the original likelihood that uses a d-dimensional vector of summary statistics s_{obs} instead of the original data D_{obs}. The more informative s_{obs} is about D_{obs}, the more accurate is this approximation. In the rare cases where s_{obs} can be chosen as a sufficient statistic for Θ, $L(\Theta; D_{\text{obs}}) \propto L(\Theta; s_{\text{obs}})$.

An estimate $\hat{L}(\Theta; s_{\text{obs}})$ of $L(\Theta; s_{\text{obs}})$ can be obtained from simulations under the model $\mathcal{M}(\Theta)$ using kernel density estimation. With $\hat{L}(\Theta; s_{\text{obs}})$ as our objective function, we approximate the maximum likelihood estimate $\hat{\Theta}_{\text{ML}}$ of Θ using the following algorithm:

Algorithm (AML-FD). *Let $a_n, c_n \in \mathbb{R}^+$ be two decreasing sequences and $k_n \in \mathbb{N}$ a non-decreasing sequence. Let H_n be a sequence of symmetric positive definite $d \times d$ matrices, the bandwidth matrices, and κ a d-variate kernel function satisfying $\int_{\mathbb{R}^d} \kappa(x)dx = 1$. Define $\kappa_{H_n}(x) := (\det H_n)^{-1/2}\kappa\left(H_n^{-1/2}x\right)$. Choose a starting value Θ_0.*

For $n = 1, 2, \ldots, N$:

1. *Simulation of the gradient in Θ_{n-1}:*

   For $l = 1, \ldots, p$:

   (a) *Simulate datasets $D_1^-, \ldots, D_{k_n}^-$ from $\mathcal{M}(\Theta^-)$ and $D_1^+, \ldots, D_{k_n}^+$ from $\mathcal{M}(\Theta^+)$ with $\Theta^\pm = \Theta_{n-1} \pm c_n e_l$.*

   (b) *Compute the summary statistics S_j^- on dataset D_j^- and S_j^+ on D_j^+ for $j = 1, \ldots, k_n$.*

   (c) *Estimate the likelihood $\hat{L}(\Theta^-; s_{\text{obs}}) = \hat{p}(s_{\text{obs}}|\Theta^-)$ and $\hat{L}(\Theta^+; s_{\text{obs}}) = \hat{p}(s_{\text{obs}}|\Theta^+)$ from $S_1^-, \ldots, S_{k_n}^-$ and $S_1^+, \ldots, S_{k_n}^+$, respectively, with kernel density estimation (e.g. Wand and Jones, 1995, equation 4.1):*

$$\hat{L}(\Theta^-; s_{\text{obs}}) = \frac{1}{k_n}\sum_{j=1}^{k_n} \kappa_{H_n}\left(s_{\text{obs}} - S_j^-\right)$$

   and analogously for $\hat{L}(\Theta^+; s_{\text{obs}})$.

   (d) *Compute the l'th element of the finite differences approximation of the gradient of the likelihood, $\nabla L(\Theta_{n-1}; s_{\text{obs}})$:*

$$\left(\hat{\nabla}_{c_n}\hat{L}(\Theta_{n-1}; s_{\text{obs}})\right)^{(l)} = \frac{\hat{L}(\Theta^+; s_{\text{obs}}) - \hat{L}(\Theta^-; s_{\text{obs}})}{2c_n}$$

2. *Updating Θ_n:*

$$\Theta_n = \Theta_{n-1} + a_n\hat{\nabla}_{c_n}\hat{L}(\Theta_{n-1}; s_{\text{obs}})$$

The approximate maximum likelihood estimate is obtained at the final step of this algorithm, i.e. $\hat{\Theta}_{\text{AML-FD,N}} := \Theta_N$. Notice however, that more sophisticated versions of the algorithm involve averaging over the last s steps.

Alternatively, an algorithm based on Spall's simultaneous perturbations method can be defined as follows:

Algorithm (AML-SP). *Let $a_n, c_n \in \mathbb{R}^+$ be two decreasing sequences and $k_n \in \mathbb{N}$ a non-decreasing sequence. Let H_n be a sequence of symmetric positive definite $d \times d$ matrices, the bandwidth matrices, and κ a d-variate kernel function satisfying $\int_{\mathbb{R}^d} \kappa(x)dx = 1$. Define $\kappa_{H_n}(x) := (\det H_n)^{-1/2}\kappa\left(H_n^{-1/2}x\right)$. Choose a starting value Θ_0.*

For $n = 1, 2, \ldots, N$:

1. *Simulation of the ascent direction in Θ_{n-1}:*

   (o) *Generate a p-dimensional random vector δ_n with elements*

$$\delta_n^{(l)} = \begin{cases} -1 & \text{with probability } 1/2, \\ +1 & \text{with probability } 1/2, \end{cases} \tag{1}$$

   for $l = 1, \ldots, p$.

   (a) *Simulate datasets $D_1^-, \ldots, D_{k_n}^-$ from $\mathcal{M}(\Theta^-)$ and $D_1^+, \ldots, D_{k_n}^+$ from $\mathcal{M}(\Theta^+)$ with $\Theta^\pm = \Theta_{n-1} \pm c_n\delta_n$.*

   (b) *Compute the summary statistics S_j^- on dataset D_j^- and S_j^+ on D_j^+ for $j = 1, \ldots, k_n$.*

(c) *Estimate the likelihood* $\hat{L}(\Theta^-; s_{\text{obs}}) = \hat{p}(s_{\text{obs}}|\Theta^-)$ *and* $\hat{L}(\Theta^+; s_{\text{obs}}) = \hat{p}(s_{\text{obs}}|\Theta^+)$ *from* $S_1^-, \ldots, S_{k_n}^-$ *and* $S_1^+, \ldots, S_{k_n}^+$*, respectively, with kernel density estimation:*

$$\hat{L}(\Theta^-; s_{\text{obs}}) = \frac{1}{k_n} \sum_{j=1}^{k_n} \kappa_{H_n}\left(s_{\text{obs}} - S_j^-\right) \tag{2}$$

and analogously for $\hat{L}(\Theta^+; s_{\text{obs}})$.

(d) *Compute the finite differences approximation of the slope of the likelihood* $L(\Theta_{n-1}; s_{\text{obs}})$ *along* δ_n*:*

$$\hat{\nabla}_{c_n} \hat{L}(\Theta_{n-1}; s_{\text{obs}}) = \delta_n \frac{\hat{L}(\Theta^+; s_{\text{obs}}) - \hat{L}(\Theta^-; s_{\text{obs}})}{2c_n}$$

2. *Updating* Θ_n*:*

$$\Theta_n = \Theta_{n-1} + a_n \hat{\nabla}_{c_n} \hat{L}(\Theta_{n-1}; s_{\text{obs}})$$

Here, the approximate maximum likelihood estimate is denoted $\hat{\Theta}_{\text{AML-SP,N}} := \Theta_N$. In more general statements, we denote both approximate maximum likelihood estimates by $\hat{\Theta}_{\text{AML}}$.

Instead of the likelihood, one might want to equivalently maximize the log-likelihood, which can be preferable from a numerical point of view. This can be done both with the AML-FD and the AML-SP algorithm by replacing $\hat{L}(\Theta^+; s_{\text{obs}})$ and $\hat{L}(\Theta^-; s_{\text{obs}})$ by their logarithms in step 1d.

In a Bayesian setting with a prior distribution $\pi(\Theta)$, the algorithms can be modified to approximate the maximum of the posterior distribution, $\hat{\Theta}_{\text{MAP}}$, by multiplying $\hat{L}(\Theta^-, s_{\text{obs}})$ and $\hat{L}(\Theta^+, s_{\text{obs}})$ with $\pi(\Theta^-)$ and $\pi(\Theta^+)$, respectively.

2.3 Parametric bootstrap

Confidence intervals and estimates of the bias and standard error of $\hat{\Theta}_{\text{AML}}$ can be obtained by parametric bootstrap: B bootstrap datasets are simulated from the model $\mathcal{M}(\hat{\Theta}_{\text{AML}})$ and the AML algorithm is run on each dataset to obtain the bootstrap estimates $\hat{\Theta}^\star_{\text{AML},1}, \ldots, \hat{\Theta}^\star_{\text{AML},B}$. These estimates reflect both the error of the maximum likelihood estimator as well as the approximation error of the algorithm.

The bias can be estimated by

$$\hat{b}^\star = \bar{\Theta}^\star - \hat{\Theta}_{\text{AML}},$$

where $\bar{\Theta}^\star = \left(\sum_{i=1}^B \hat{\Theta}_i^\star\right)/B$. The corrected estimator is

$$\hat{\Theta}^\star_{\text{AML}} = \hat{\Theta}_{\text{AML}} - \hat{b}^\star.$$

The standard error of $\hat{\Theta}_{\text{AML}}$ can be estimated by the standard deviation of $\hat{\Theta}_1^\star, \ldots, \hat{\Theta}_B^\star$,

$$\widehat{se}^\star = \sqrt{\frac{1}{B-1} \sum_{i=1}^B \left(\hat{\Theta}_i^\star - \bar{\Theta}^\star\right)^2}.$$

We compute separate bootstrap confidence intervals for each parameter that are based on the assumption that the distribution of $\hat{\Theta}_{\text{AML}} - \Theta$ can be approximated sufficiently well by the distribution of $\hat{\Theta}^\star_{\text{AML}} - \hat{\Theta}_{\text{AML}}$, where $\hat{\Theta}^\star_{\text{AML}}$ is the bootstrap estimator. Then, a two-sided $(1-\alpha)$-confidence interval is defined as

$$\left[2\hat{\Theta}_{\text{AML}} - q_{(1-\alpha/2)}(\hat{\Theta}^\star_{\text{AML}}),\ 2\hat{\Theta}_{\text{AML}} - q_{(\alpha/2)}(\hat{\Theta}^\star_{\text{AML}})\right],$$

where $q_{(\beta)}(\hat{\Theta}^{\star}_{\mathrm{AML}})$ denotes the β quantile of $\hat{\Theta}^{\star}_{\mathrm{AML},1}, \ldots, \hat{\Theta}^{\star}_{\mathrm{AML},B}$ (Davison and Hinkley, 1997, sections 2.2.1, 2.4).

3 Tuning guidelines

Confirming general observations made at the end of Section 2.1, the experience from our simulations suggests that the simultaneous perturbation approach AML-SP usually converges faster than AML-FD. Notice however, that we encountered a slightly higher proportion of runs that did not converge with AML-SP (see Section 4.1).

For both algorithms the performance strongly depends on proper tuning of the four main parameters a_n (step size), c_n (gradient approximation), k_n (size of the simulated samples), and H_n (bandwidth of the kernel density estimate). A proper choice of the kernel function can also be helpful. Another challenge we encountered in our population genetic application was stochasticity: occasional large steps can by chance lead very far away from the maximum.

To facilitate tuning, we first summarize general guidelines based on theory. As the resulting optimal choices depend on the unknown likelihood function, we supplement these recommendations by practical tuning guidelines. We found our proposed practical guidelines helpful to ensure both a high proportion of convergent runs, and (in case of convergence) rapid convergence to solutions that are sufficiently close to the true maximum of the likelihood. We first state these recommendations in general terms. As they require additional parameter choices, we then state actual values we found to work well in our examples.

The guidelines ensure some degree of robustness, as they are adaptive. In particular, they take into account problems occuring in recent steps of the algorithm, and include local information on the likelihood surface for instance for deciding when to stop, when adapting the step sizes, and when choosing the kernel bandwidth. We can imagine that even with our proposed adaptive tuning guidelines, convergence problems might still occur in some applications, in particular when the likelihood surface is very complex. However, we do not see this as a major drawback of our method for two reasons:

- It is always possible (and recommended) to estimate the likelihood at stopping, for different runs starting from randomly selected starting points. A certain proportion of runs converging to the same point which also provides the highest estimated likelihood value would give some confidence in this estimate. (In principle the final estimate could still be a local optimum however, as global optimization can sometimes be a very hard task.)
- In an initial step our algorithms can be applied to datasets simulated under the model of interest, where the true solution is known. In cases where frequent convergence problems show up with the recommended parameter settings, tuning parameters could be changed. If multiple local optima show up, further independent runs from additional starting points would be an option.

3.1 Theoretical guidelines for choosing tuning parameters

The stochastic approximation part of our approximate maximum likelihood algorithm requires two tuning parameter sequences, the stepsize a_n and the parameter c_n responsible for the bias-variance trade-off of the finite difference approximations to the derivative. According to Spall (2003, chap. 7), tuning constants satisfying $a_n, c_n \to 0$ for $n \to \infty$, as well as $\sum_{n=1}^{\infty} a_n < \infty$ and $\sum_{n=1}^{\infty} \frac{a_n^2}{c_n^2} < \infty$ ensure convergence of the algorithm. (Notice that there are further technical conditions for convergence, such as a unique optimum and a uniformly bounded Hessian matrix.) It can be shown that for

$$a_n = \frac{a}{(n+A)^{\alpha}} \text{ and } c_n = \frac{c}{n^{\gamma}}, \tag{3}$$

optimum rates of convergence can be obtained when $\alpha = 1$ and $\gamma = 1/6$. As discussed in Spall (2003, p. 187), however, smaller values for α and γ often lead to a better practical performance. The constant A is usually

taken as a small fraction of the maximum number of iterations N and guards against too large jumps in the initial steps.

Another important factor affecting the algorithm is the precision of the kernel density estimate used. Under smoothness conditions on the underlying likelihood, the kernel density estimate used for estimating the likelihood is consistent, if $k_n \to 0$, $H_n \to 0$ and $nH_n^d \to \infty$ in the case of d summary statistics. Notice however that for stochastic approximation to work, the variance of the estimates does not need to tend to zero. In particular large sample sizes k_n are not needed in the individual steps. Nevertheless good estimates of the gradient are obviously helpful. If we measure the quality of the pointwise density estimates by the the mean squared error (MSE), optimizing the MSE [see Rosenblatt (1991)] suggests to choose $H_n = cn^{-(4+d)}$, (for each diagonal enrtry of H_n) with a constant $c^\star$ depending on the unknown likelihood function.

As $c^\star$ is unknown in practice, we use a data driven bandwidth selection rule with our algorithm. Many different methods to estimate bandwidth matrices for multivariate kernel density estimation have been proposed (see Scott, 2015, for an overview), and in principle any method that fulfills the convergence criteria is valid to estimate H_n. For some types of densities, non-sparse bandwidth matrices can give a substantial gain in efficiency compared to e.g. diagonal matrices (Wand and Jones, 1995, section 4.6). However, as the bandwidth matrix is computed for each likelihood estimate, computational efficiency is important. In our examples, we estimate a diagonal bandwidth matrix H_n using a multivariate extension of Silverman's rule (Härdle et al., 2004, equation 3.69). Using new bandwidth matrices in each iteration introduces an additional level of noise that can be reduced by using a moving average of bandwidth estimates.

In density estimation, the choice of a specific kernel function is usually considered less important than the bandwidth choice. The kernel choice plays a more important role when used with our stochastic gradient algorithm however. Indeed, to enable the estimation of the gradient even far away from the maximum, a kernel function with infinite support is helpful. The Gaussian kernel is an obvious option, but when the log-likelihood is used, the high rate of decay can by chance cause very large steps leading away from the maximum. Therefore, we use the following modification of the Gaussian kernel:

$$\kappa(H^{1/2}x) \propto \begin{cases} \exp\left(-\frac{1}{2}x'H^{-1}x\right) & \text{if } x'H^{-1}x < 1 \\ \exp\left(-\frac{1}{2}\sqrt{x'H^{-1}x}\right) & \text{otherwise.} \end{cases}$$

In degenerate cases where the likelihood evaluates to zero numerically, we replace the classical kernel density estimate by a nearest neighbor estimate in step 1(c):

If $\hat{L}(\Theta^-; s_{\text{obs}}) \approx 0$ *or/and* $\hat{L}(\Theta^+; s_{\text{obs}}) \approx 0$ (*with "≈" denoting "numerically equivalent to"*), *find*

$$S^-_{\min} := \arg\min_{S_j^-}\left\{\left\|S_j^- - s_{\text{obs}}\right\| : j = 1, \ldots, k_n\right\}$$
$$S^+_{\min} := \arg\min_{S_j^+}\left\{\left\|S_j^+ - s_{\text{obs}}\right\| : j = 1, \ldots, k_n\right\}$$

and recompute the kernel density estimate $\hat{L}(\Theta^-; s)$ *and* $\hat{L}(\Theta^+; s)$ *in step 1c using* $S^-_{\min}$ *and* $S^+_{\min}$, *respectively, as the only observation.*

3.2 Heuristic tuning guidelines

Here, we summarize some practical recommendations that turned out to be useful in our simulations.

As the parameters we want to estimate may often lie in regions of different length, and to reflect the varying slope of L in different directions of the parameter space, a good choice of a and c can speed up convergence. We therefore replace $a \in \mathbb{R}^+$ in eq. (3) by a p-dimensional diagonal matrix $\mathbf{a} = \text{diag}\left(a^{(1)}, \ldots, a^{(p)}\right)$ with $a^{(i)} \in \mathbb{R}^+$ for $i = 1, \ldots, p$ (this is equivalent to scaling the space accordingly). The optimal choice of $\mathbf{a}$ and

c depends on the unknown shape of L. Therefore, we propose the following heuristic based on suggestions in Spall (2003, sections 6.6, 7.5.1, 7.5.2) and our own experience to determine these values as well as $A \in \mathbb{N}$, a shift parameter to avoid too fast decay of the step size in the first iterations.

Let N be the number of planned iterations and $b \in \mathbb{R}^p$ a vector that gives the desired stepsize in early iterations in each dimension. Choose a starting value Θ_0.

1. *Set c to a small percentage of the total parameter range (or the search space, see Section 3.4), to obtain c_1.*
2. *Set $A = \lfloor 0.1 * N \rfloor$.*
3. *Choose* **a**:
   (a) *Estimate $\nabla L(\Theta_0; s_{\text{obs}})$ by the median of K_1 finite differences approximations (step 1 of the AML algorithm), $\tilde{\hat{\nabla}}_{c_1} \hat{L}(\Theta_0; s_{\text{obs}})$.*
   (b) *Set*

$$a^{(i)} = \left| \frac{b^{(i)}(A+1)^{\alpha}}{\left(\tilde{\hat{\nabla}}_{c_1} \hat{L}(\Theta_0; s_{\text{obs}})\right)^{(i)}} \right| \text{ for } i = 1, \ldots, p.$$

As **a** is determined using information about the likelihood in Θ_0 only, it might not be adequate in other regions of the parameter space. To be able to distinguish convergence from a too small step size, we simultaneously monitor the growth of the likelihood function and the trend in the parameter estimates to adjust a if necessary. Every N_0 iterations the following three tests are conducted on the preceding N_0 iterates:

- ***Trend test (too small $a^{(i)}$):*** *For each dimension $i = 1, \ldots, p$ a trend in $\Theta_k^{(i)}$ is tested using the standard random walk model*

$$\Theta_k^{(i)} = \Theta_{k-1}^{(i)} + \beta + \epsilon_k,$$

  where β denotes the trend and $\epsilon_k \sim N(0, \sigma^2)$. The null hypothesis $\beta = 0$ can be tested by a t-test on the differences $\Delta_k = \Theta_k^{(i)} - \Theta_{k-1}^{(i)}$. If a trend is detected, $a^{(i)}$ is increased by a fixed factor $f \in \mathbb{R}^+$.
- ***Range test (too large $a^{(i)}$):*** *For each dimension $i = 1, \ldots, p$, $a^{(i)}$ is set to $a^{(i)}/f$ if the trajectory of $\Theta_k^{(i)}$ spans more than π_a% of the parameter range.*
- ***Convergence test:*** *Simulate K_2 likelihood estimates at Θ_{n-N_0} and at Θ_n. Growth of the likelihood is then tested by a one-sided Welch's t-test. (Testing for equivalence could be used instead, see Wellek, 2010.)*

We conclude that the algorithm has converged to a maximum only if the convergence test did not reject the null hypothesis three times in a row and at the same time no adjustments of **a** were necessary.

The values of the tuning parameters we used in our examples presented in Section 4 are summarized in Table 1.

Table 1: Setting of the tuning parameters in the examples.

Parameter	Normal distribution	Orang-utan data
N	Max. 100,000	Max. 50,000
Search space	$[-100, 100] \times \ldots \times [-100, 100]$	Scaled to $[0, 1] \times \ldots \times [0, 1]$
c	$c = 2$	$c = 0.05$
A	500	500
K_1	100	
K_2	25	
N_0	1000	
f	1.5	
π_a	70%	
$\pi_{\max}$	10%	

3.3 Starting points

To avoid starting in regions of very low likelihood, a set of random points should be drawn from the parameter space and their simulated likelihood values compared to find good starting points. This strategy also helps to avoid that the algorithm reaches only a local maximum. More details can be found in Section 4.

3.4 Constraints on the parameters

The parameter space will usually be subject to constraints (e.g. rates are positive quantities). They can be incorporated by projecting the iterate to the closest point such that both Θ^- as well as Θ^+ are in the feasible set (Sadegh, 1997). Even if there are no imperative constraints, it is advisable to restrict the search space to a range of plausible values to prevent the algorithm from trailing off at random in regions of very low likelihood.

To reduce the effect of large steps within the boundaries, we clamp the step size at $\pi_{\max}\%$ of the range of the search space in each dimension.

4 Examples

To study the performance of the AML algorithm, we first test it on the multivariate normal distribution. While there is no need for simulation based inference for normal models, it allows us to compare the properties of the AML estimator and the maximum likelihood estimator. We also compare the convergence speed of the two different algorithms here. Then, we apply it to an example from population genetics. For this purpose we use both simulated data, where the true parameter values are known, as well as DNA sequence data from a sample of Bornean and Sumatran orang-utans and estimate parameters of their ancestral history.

4.1 Multivariate normal distribution

One dataset, consisting of *i.i.d.* draws from a 10-dimensional normal distribution, is simulated such that the maximum likelihood estimator for Θ is $\hat{\Theta}_{\text{ML}} = \bar{X} \sim \mathcal{N}(5 \cdot \mathbf{1}_{10}, I_{10})$ where I_{10} denotes the 10-dimensional identity matrix and $\mathbf{1}_{10}$ the 10-dimensional vector with 1 in each component. To estimate the distribution of $\hat{\Theta}_{\text{AML-SP}}$, the AML-SP algorithm is run 1000 times on this dataset with summary statistics $S = \bar{X}$.

At start, 1000 points are drawn randomly on $(-100, 100) \times \ldots \times (-100, 100)$. For each of them, the likelihood is simulated and the 5 points with the highest likelihood estimate are used as starting points. On each of them, the AML-SP algorithm is run with $k_n = 100$ for at least 10,000 iterations and stopped as soon as convergence is reached (for ≈90% of the sequences within 11,000 iterations). Again, the likelihood is simulated on each of the five results and the one with the highest likelihood is considered as a realization of $\hat{\Theta}_{\text{AML-SP}}$. Based on these 1000 realizations of $\hat{\Theta}_{\text{AML-SP}}$, the density, bias and standard error of $\hat{\Theta}_{\text{AML-SP}}$ are estimated for each dimension (Table 2, Figure 1).

The results are extremely accurate: The densities of the 10 components of $\hat{\Theta}_{\text{AML-SP}}$ are symmetric around the maximum likelihood estimate with a negligible bias. Compared to the standard error of $\hat{\Theta}_{\text{ML}}$, which by construction equals 1, the standard error of $\hat{\Theta}_{\text{AML-SP}}$ is nearly 20 times smaller. Bootstrap confidence intervals that were obtained from $B = 100$ bootstrap samples for 100 datasets simulated under the same model as above show a close approximation to the intended coverage probability of 95% (Table 2).

To investigate the impact of the choice of summary statistics on the results, we repeat the experiment with the following set of summary statistics:

Table 2: Properties of $\hat{\Theta}_{\text{AML-SP}}$ in the 10-dimensional normal distribution model using two different sets of summary statistics.

dim.	$S = \bar{X}$			$S^\star$ (eq. 4)		
	$\hat{b}$	$\widehat{se}$	$\hat{p}$	$\hat{b}$	$\widehat{se}$	$\hat{p}$
1	−0.0016	0.0546	93	−0.0007	0.0660	93
2	0.0017	0.0544	94	−0.0002	0.0638	93
3	0.0004	0.0547	94	0.0007	0.0635	97
4	−0.0033	0.0567	90	−0.0032	0.0789	98
5	−0.0015	0.0584	95	−0.0000	0.0748	90
6	−0.0000	0.0565	94	0.0044	0.0757	90
7	0.0017	0.0557	96	0.0035	0.0686	95
8	−0.0007	0.0559	95	−0.0030	0.1140	96
9	−0.0013	0.0554	91	0.0809	0.0658	98
10	0.0001	0.0554	99	−0.0595	0.0922	92

Bias ($\hat{b}$) and standard error ($\widehat{se}$) of $\hat{\Theta}_{\text{AML-SP}}$, estimated from 1000 runs of the AML-SP algorithm. Coverage probability of bootstrap 95% confidence intervals ($\hat{p}$) with $B = 100$, estimated from 100 datasets.

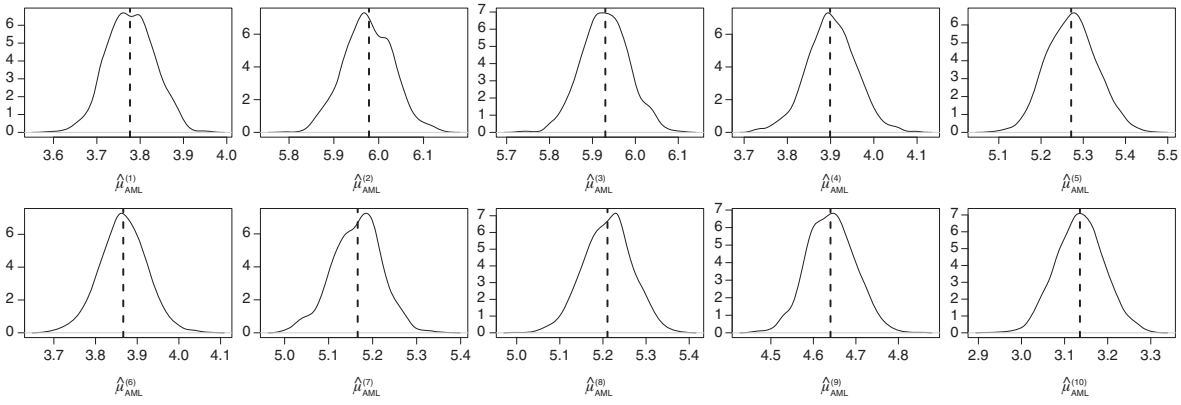


Figure 1: Density of the components of $\hat{\Theta}_{\text{AML-SP}}$ obtained with $S = \bar{X}$ in one dataset estimated from 1000 converged sequences with a miminum length of 10,000 iterations by kernel density estimation. Vertical dashed line: $\hat{\Theta}_{\text{ML}}$.

$$S^\star = \begin{pmatrix} \bar{X}_1 \\ \bar{X}_2 + \bar{X}_3 \\ \bar{X}_4 + \bar{X}_5 \\ \bar{X}_7 \\ \bar{X}_9 \\ \bar{X}_2 - \bar{X}_3 \\ \bar{X}_5 + \bar{X}_6 \\ \bar{X}_7 + \bar{X}_8 \\ \bar{X}_9 \cdot \bar{X}_{10} \\ \bar{X}_6 + \bar{X}_4 \end{pmatrix} \tag{4}$$

For ≈90% of the sequences, convergence was detected within 14,000 iterations. Bootstrap confidence intervals are obtained for 100 simulated example datasets. Their coverage probability matches the nominal 95% confidence level closely (Table 2). The components behave very similar to the previous simulations, except for the estimates of the density for components 9 and 10 (Figure 2). Compared to the simulations with $S = \bar{X}$, the bias of components 9 and 10 is considerably increased, but it is still much smaller than the standard error of $\hat{\Theta}_{\text{ML}}$. To investigate how fast the bias decreases with the number of iterations, we re-run the above described algorithm for 100,000 iterations without earlier stopping (Figure 3). Both bias and standard error decrease with the number of iterations.

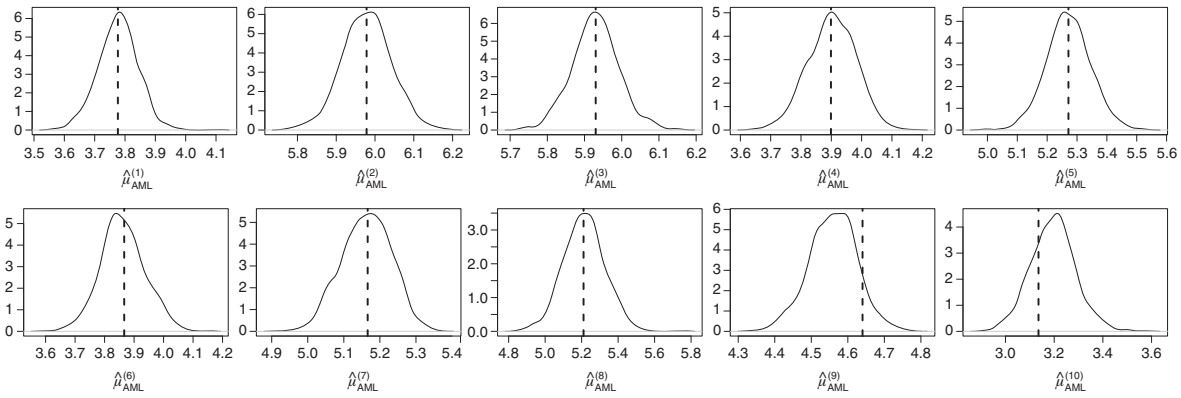


Figure 2: Density of the components of $\hat{\theta}_{\text{AML-SP}}$ obtained with $S^\star$ (eq. 4) in one dataset estimated from 1000 converged sequences with a miminum length of 10,000 iterations by kernel density estimation. Vertical dashed line: $\hat{\theta}_{\text{ML}}$.

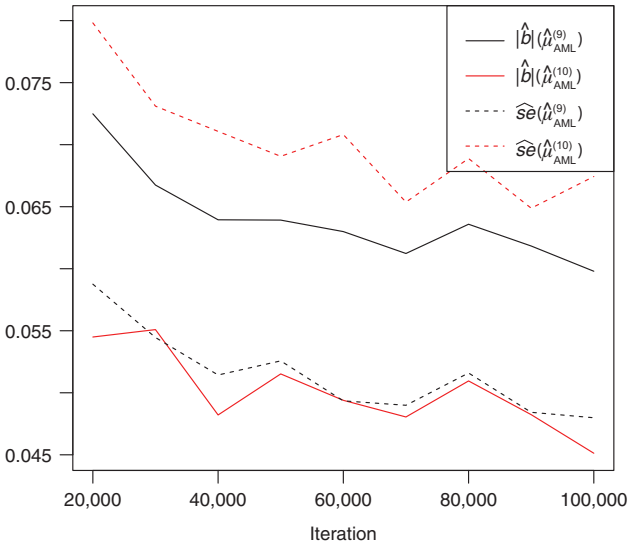


Figure 3: Absolute bias and standard error of the AML estimator of μ_9 and μ_{10} using the summary statistics in eq. (4) estimated from 1000 runs of the algorithm on the same dataset.

For the comparison of the AML-FD and AML-SP algorithm, 1000 datasets are simulated under the same normal distribution model as above. $S^\star$ (eq. 4) is used as our vector of summary statistics.

For each dataset, one random starting value is drawn on the search space $(0, 10) \times \ldots \times (0, 10)$ and the two approximate maximum likelihood algorithms AML-FD and SP are run on each of them with $k_n = 50$. Different values of c are tested in short preliminary runs and a small set of good values is used. Here, the convergence diagnostics and the methods to make the algorithms more robust are not used (except from shifting the iterates back into the search space if necessary).

For each algorithm, 5 million datasets are simulated. This results in different numbers of iterations (N): For the FD algorithm, $N = 5000$, for the SP algorithm, $N = 50{,}000$. The results of the different algorithms are compared after the same number of simulations. The corresponding runtimes are shown in Table 3. As expected, one iteration of the FD algorithm takes approximately 10 times longer than one iteration of the SP algorithm: the likelihood is estimated at 20 points in each iteration in the FD algorithm compared to 2 points in the SP algorithm.

In the worst case ($c = 1$, SP), more than 50% of the runs enter regions of approximately zero likelihood.

Table 3: Runtimes and errors.

Type	Parameters	K	Total runtime	Runtime per iteration	Errors
FD	$c = 1$	5000	1580.39	0.3161	29
	$c = 2$	5000	1650.04	0.3300	22
	$c = 0.5$	5000	1654.64	0.3309	116
SP	$c = 1$	50,000	1549.79	0.0310	558
	$c = 2$	50,000	1545.65	0.0309	364
	$c = 0.5$	50,000	1607.65	0.0322	228

Runtime is measured in seconds. Errors: number of runs (out of 1000) that enter regions where the likelihood estimate is zero.

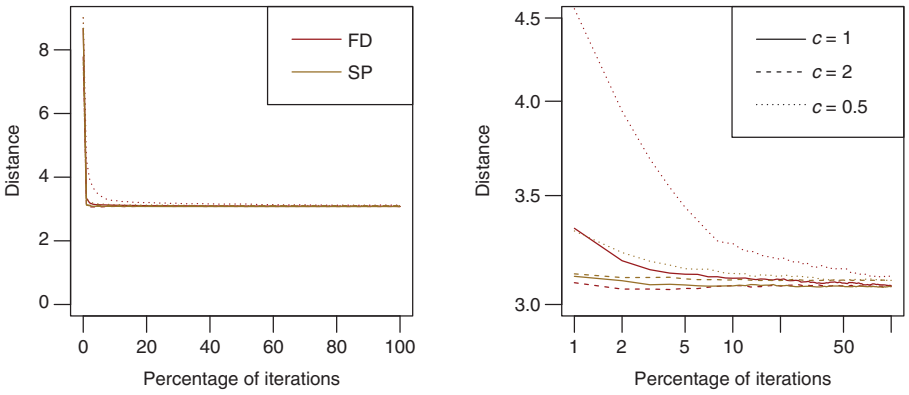


Figure 4: Euclidean distance of $\hat{\theta}_n$ to $\hat{\theta}_{\text{ML}}$. The distance is plotted only every percent of the iterations. Right panel: same as left panel, but with both axes on a log-scale.

For $c = 2$, both algorithms behave very similar (Figure 4). For $c = 1$, the AML-SP algorithm converges slightly faster and for $c = 0.5$ it converges considerably faster and the AML-FD algorithm needs many more iterations to reach the same accuracy. This reflects its theoretical properties derived in Spall (1992).

However, the AML-SP algorithm is less stable than AML-FD and can more easily trail off randomly to regions of very low likelihood. The SP algorithm even enters regions of zero likelihood (at standard double precision) considerably more frequently (Table 3), causing problems with the computation of the log-likelihood. In our simulations we could easily overcome these problems by the adjustments proposed in Sections 3.2–3.4 that make the algorithm more robust. But in regions of very low likelihood, small differences in the likelihood values will still lead to large gradients of the log likelihood, and thus to large steps. Then, it can help to reduce $\pi_{\max}$.

In the following examples, we use the AML-SP algorithm and to simplify the notation, define $\hat{\Theta}_{\text{AML}} := \hat{\Theta}_{\text{AML-SP}}$.

4.2 The evolutionary history of Bornean and Sumatran orang-utans

Pongo pygmaeus and *Pongo abelii*, Bornean and Sumatran orang-utans, respectively, are Asian great apes whose distributions are exclusive to the islands of Borneo and Sumatra. Recurring glacial periods led to a cooler, drier, and more seasonal climate. Consequently the rain forest might have contracted and led to isolated populations of orang-utans. At the same time, the sea level dropped and land bridges among islands created opportunities for migration among previously isolated populations. However, whether glacial periods have been an isolating or a connecting factor remains poorly understood. Therefore, there has been a considerable interest in using genetic data to understand the demographic history despite the computational difficulties involved in such a population genetic analysis. We will compare our results to the analysis of the orang-utan genome paper (Locke et al., 2011) and a more comprehensive study by Ma et al. (Ma et al., 2013);

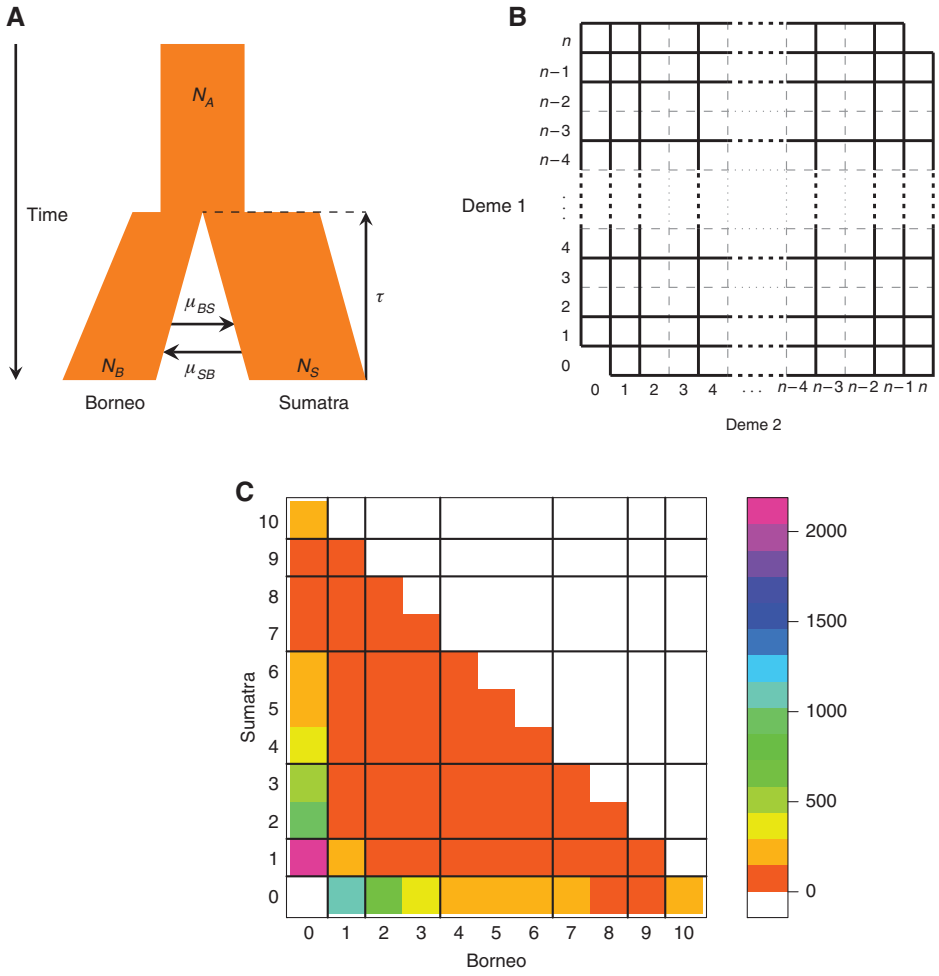


Figure 5: (A) Isolation-migration model for the ancestral history of orang-utans. Notation: N_A, effective size of the ancestral population; μ_{BS} (μ_{SB}), fraction of the Bornean (Sumatran) population that is replaced by Sumatran (Bornean) migrants per generation (backwards migration rate); τ, split time in years; N_B (N_S), effective population size in Borneo (Sumatra). (B) Binned joint site frequency spectrum (adapted from Naduvilezhath et al., 2011). (C) Folded joint site frequency spectrum of biallelic SNPs at four-fold degenerate sites in the Bornean and Sumatran orang-utan samples.

these are analyses that have been performed genome-wide. Both studies use ∂a∂i (Gutenkunst et al., 2009), a popular software that has been widely used for demographic inference. ∂a∂i is based on a numerical solution of the Wright-Fisher diffusion approximation. We use the orang-utan data set of Locke et al. (2011), consisting of SNP data of four-fold degenerate (synonymous) sites taken from 10 sequenced individuals (five individuals each per orang-utan population, haploid sample size 10). See the Appendix for more details on the dataset.

As in Locke et al. (2011) and Ma et al. (2013), we consider an Isolation-Migration (IM) model where a panmictic ancestral population of effective size N_A splits τ years ago into two distinct populations of constant effective size N_B (the Bornean population) and N_S (the Sumatran population) with backward migration rates μ_{BS} (fraction of the Bornean population that is replaced by Sumatran migrants per generation) and μ_{SB} (vice versa; Figure 5A).

N_A is set to the present effective population size that we obtain using the number of SNPs in our data set and assuming an average per generation mutation rate per nucleotide of $2 \cdot 10^{-8}$ and a generation time of 20 years (Locke et al., 2011), so $N_A = 17{,}400$.

There are no sufficient summary statistics at hand, but for the IM model the joint site frequency spectrum (JSFS) between the two populations was reported to be a particularly informative summary statistic (Tellier et al., 2011). However, for N samples in each of the two demes, the JSFS has $(N+1)^2 - 2$ entries, so

even for small datasets it is very high-dimensional. To reduce this to more manageable levels we follow Naduvilezhath et al. (2011) and bin categories of entries (Figure 5B). As the ancestral state is unknown, we use the folded binned JSFS that has 28 entries (see Figure 5C for the folded JSFS of the two population samples). To incorporate observations of multiple unlinked loci, mean and standard deviation across loci are computed for each bin, so the final summary statistics vector is of length 56.

The AML-SP algorithm with the JSFS as summary statistics was used together with the coalescent simulation program `msms` (Ewing and Hermisson, 2010). This allows for complex demographic models and permits fast summary statistic evaluation without using external programs and the associated performance penalties.

4.2.1 Simulations

Before applying the AML-SP algorithm to the actual orang-utan DNA sequences, we tested it on simulated data. Under the described IM model with parameters $N_B = N_S = 17{,}400$, $\mu_{BS} = \mu_{SB} = 1.44 \cdot 10^{-5}$ and $\tau = 695{,}000$, we simulated 25 datasets with 25 haploid sequences per deme, each of them consisting of 75 loci with 130 SNPs each. We define loci as unlinked stretches of DNA sequence, which are so short that recombination within a locus can be disregarded.

For each dataset, 25 AML estimates were obtained with the same scheme: 1000 random starting points were drawn from the parameter space; the likelihood was estimated with $n = 40$ simulations. Then, the five values with the highest likelihood estimates were used as starting points for the AML algorithm. The algorithm converged after 3000–25,000 iterations (average: ≈8000 iterations; average runtime of our algorithm: 11.7 h on a single core).

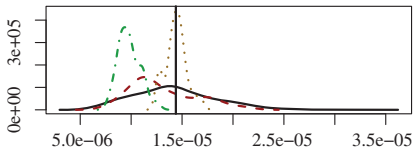
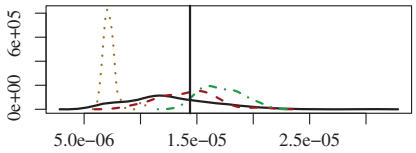
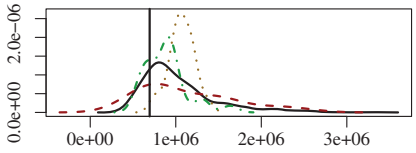
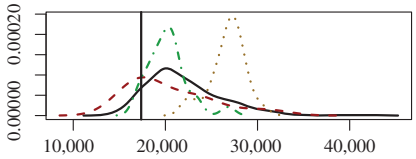
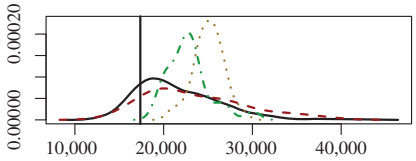
For the migration rates μ_{SB} and μ_{BS}, the per dataset variation of the estimates is considerably smaller than the total variation (Table 4). This suggests that the approximation error of the AML algorithm is small in comparison to the error of $\hat{\Theta}_{\mathrm{ML}}$. For the split time τ and the population sizes N_S and N_B, the difference is less pronounced, but still apparent. For all parameters, the average bias of the estimates is either smaller or of approximately the same size as the standard error. As the maximum likelihood estimate itself cannot be computed, it is impossible to disentangle the bias of $\hat{\Theta}_{\mathrm{ML}}$ and an additional bias introduced by the AML algorithm.

As an alternative measure of performance, we compare the likelihood estimated at the true parameter values and the AML estimate with the highest estimated likelihood on each dataset (Table 5). In none of the 25 simulated datasets the likelihood at the true value is higher, whereas it is significantly lower in 11 of them (significance was tested with a two-sided Welch test using a Bonferroni-correction to account for multiple testing, i.e. for the 100 tests in Tables 5 and 7, a test with p-value <0.00025 is considered significant at an overall significance level of 5%). This suggests that the AML algorithm usually produces estimates that are closer to the maximum likelihood estimate than the true parameter value is.

Similarly, we compare our results to estimates obtained with $\delta a \delta$i, $\hat{\Theta}_{\delta a \delta i}$ (Gutenkunst et al., 2009). As $\delta a \delta$i is based on the diffusion approximation, whereas we are using the coalescent, we do not expect the maximum likelihood estimators to be equal, and the coalescent likelihood should be higher at $\hat{\Theta}_{\mathrm{AML}}$, if the algorithm performs well. In all the 25 simulated datasets the coalescent likelihood at $\hat{\Theta}_{\delta a \delta i}$ is lower than at $\hat{\Theta}_{\mathrm{AML}}$ with the highest likelihood, and the difference is significant in 13 datasets (Table 5). To make sure that the results are not biased due to possible convergence problems of $\delta a \delta$i, we use the true Θ as starting point. In none of the runs, $\delta a \delta$i reported convergence problems.

To investigate the impact of the underlying parameter value on the quality of the estimates, simulation results were obtained also for 25 datasets simulated with the divergence time τ twice as large. Here, all parameter estimates, especially τ, N_B and N_S had larger standard errors and biases (Table 6). Apparently, the estimation problem is more difficult for more distant split times. This may be caused by a flatter likelihood surface and by stronger random noise in the simulations. Only the migration rates are hardly affected by the large τ: a longer divergence time allows for more migration events that might facilitate their analysis. The higher level of difficulty shows up also when comparing the likelihood at the true value and at $\hat{\Theta}_{\mathrm{AML}}$: in 13

Table 4: Properties of $\hat{\Theta}_{\mathrm{AML}}$ in the IM model with short divergence time ($\tau = 695{,}000$ years).

μ_{BS}		True	1.44e−05
		Space	(1.44e−06, 0.000144)
		Mean	1.42e−05
		Median	1.39e−05
		Bias	−1.74e−07
		Mean *se*	2.1e−06
		Total *se*	4.09e−06
μ_{SB}		True	1.44e−05
		Space	(1.44e−06, 0.000144)
		Mean	1.26e−05
		Median	1.22e−05
		Bias	−1.78e−06
		Mean *se*	2.02e−06
		Total *se*	3.81e−06
τ		True	695,000
		Space	(139,000, 6,950,000)
		Mean	1,030,000
		Median	920,000
		Bias	334,000
		Mean *se*	310,000
		Total *se*	437,000
N_S		True	17,400
		Space	(1740, 174,000)
		Mean	22,000
		Median	21,200
		Bias	4610
		Mean *se*	3060
		Total *se*	4130
N_B		True	17,400
		Space	(1740, 174,000)
		Mean	21,600
		Median	20,600
		Bias	4230
		Mean *se*	3090
		Total *se*	4790

Figures: marginal densities of components of $\hat{\Theta}_{\mathrm{AML}}$, estimated by kernel density estimation. Black solid line: density of $\hat{\Theta}_{\mathrm{AML}}$. Coloured dotted and dashed lines: density of $\hat{\Theta}_{\mathrm{AML}}$ in three example datasets. Vertical black line: true parameter value. Summaries: true, true parameter value; space, search space; mean, mean of $\hat{\Theta}_{\mathrm{AML}}$; median, median of $\hat{\Theta}_{\mathrm{AML}}$; bias, bias of $\hat{\Theta}_{\mathrm{AML}}$; mean *se*, mean standard error of $\hat{\Theta}_{\mathrm{AML}}$ per dataset; total *se*, standard error of $\hat{\Theta}_{\mathrm{AML}}$.

out of 25 datasets, the likelihood at Θ is significantly lower than at $\hat{\Theta}_{\mathrm{AML}}$, whereas it is significantly higher in three cases (Table 7). The likelihood at $\hat{\Theta}_{\delta a \delta i}$ is significantly lower in 15 datasets and never significantly higher.

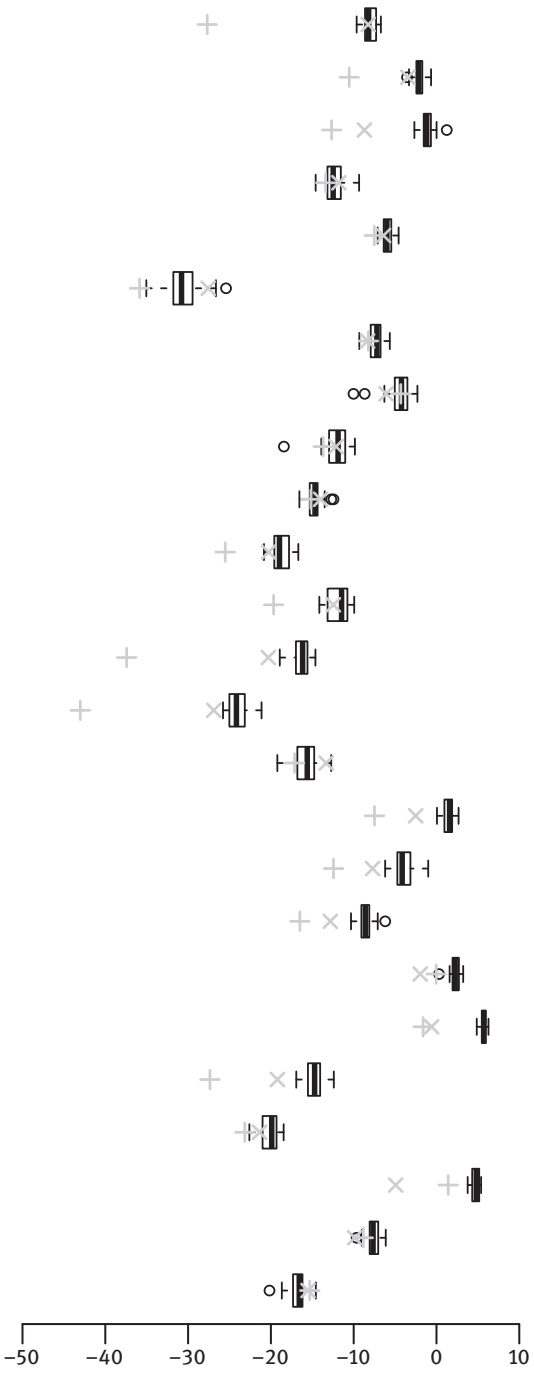
4.2.2 Real data

We model the ancestral history of orang-utans with the same model and use the same summary statistics as in the simulations. Also the datasets are simulated in the same manner.

To study the distribution of $\hat{\Theta}_{\mathrm{AML}}$ on this dataset, the algorithm has been run 20 times with the same scheme as in the simulations. The estimate with the highest likelihood is further used for bootstrapping.

Table 5: Comparison of likelihood estimates at Θ, $\hat{\Theta}_{\text{AML}}$ and $\hat{\Theta}_{\delta a \delta i}$ in the IM-model with short divergence time ($\tau = 695{,}000$ years).

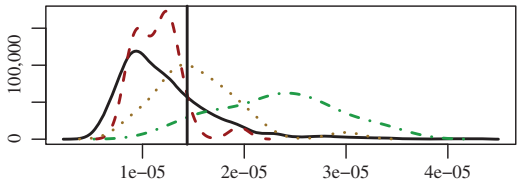
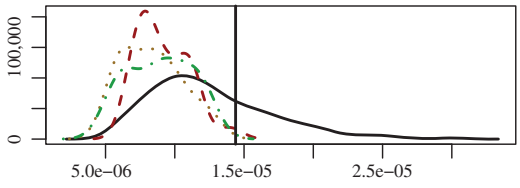
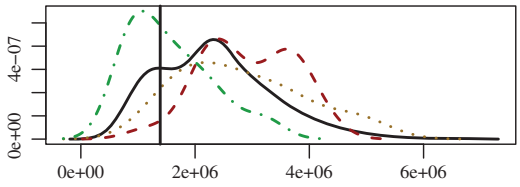
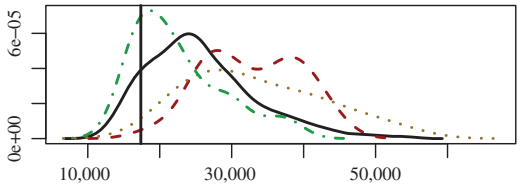
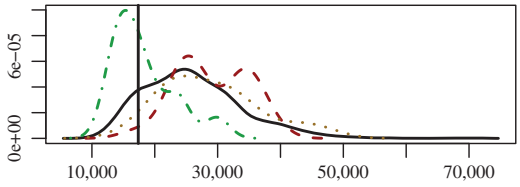
Dataset	Rank $L(\Theta)$	p-Value	Rank $L(\hat{\Theta}_{\delta a \delta i})$	p-Value	Boxplot
1	12	0.05182	1	0.00000	
2	2	0.00175	1	0.00000	
3	1	0.00000	1	0.00000	
4	19	0.01791	4	0.00686	
5	7	0.01618	1	0.00023	
6	24	0.14270	1	0.01523	
7	3	0.00069	4	0.00405	
8	4	0.00003	12	0.00413	
9	13	0.05292	3	0.02100	
10	21	0.12235	11	0.09453	
11	5	0.00037	1	0.00294	
12	9	0.00613	1	0.00000	
13	1	0.00009	1	0.00000	
14	1	0.00054	1	0.00000	
15	25	0.32213	7	0.01725	
16	1	0.00000	1	0.00000	
17	1	0.00000	1	0.00000	
18	1	0.00000	1	0.00000	
19	1	0.00000	1	0.00008	
20	1	0.00000	1	0.00000	
21	1	0.00000	1	0.00000	
22	6	0.03779	1	0.09898	
23	1	0.00000	1	0.00000	
24	1	0.00015	4	0.01592	
25	23	0.24288	24	0.38258	



Rank: rank of $\hat{L}(\Theta)$ or $\hat{L}(\hat{\Theta}_{\delta a \delta i})$, respectively, among $\hat{L}(\hat{\Theta}_{\text{AML},1}), \ldots, \hat{L}(\hat{\Theta}_{\text{AML},25})$, where $\hat{\Theta}_{\text{AML},j}$ denotes the j'th realization of $\hat{\Theta}_{\text{AML}}$ in the dataset at hand; p-value: p-value of Welch's t-test for $H_0 : L(\Theta) = L(\hat{\Theta}_{\text{AML,max}})$ vs. $H_1 : L(\Theta) \neq L(\hat{\Theta}_{\text{AML,max}})$ with $\hat{\Theta}_{\text{AML,max}} = \arg\max\{\hat{L}(\hat{\Theta}_{\text{AML},j}) : j = 1, \ldots, 25\}$; boxplot: boxplot of $\log \hat{L}(\hat{\Theta}_{\text{AML},1}), \ldots, \log \hat{L}(\hat{\Theta}_{\text{AML},25})$; grey $\times$, $\hat{L}(\Theta)$; grey $+$, $\hat{L}(\hat{\Theta}_{\delta a \delta i})$.

Confidence intervals are obtained by parametric bootstrap with $B = 1000$ bootstrap datasets. The bootstrap replicates are also used for bias correction and estimation of the standard error (Table 8).

Table 6: Properties of $\hat{\Theta}_{\mathrm{AML}}$ in the IM model with long divergence time ($\tau = 1{,}390{,}000$ years).

μ_{BS}		True	1.44e−05
		Space	(1.44e−06, 0.000144)
		Mean	1.24e−05
		Median	1.12e−05
		Bias	−1.97e−06
		Mean *se*	3.38e−06
		Total *se*	4.99e−06
μ_{SB}		True	1.44e−05
		Space	(1.44e−06, 0.000144)
		Mean	1.25e−05
		Median	1.17e−05
		Bias	−1.88e−06
		Mean *se*	3.28e−06
		Total *se*	4.38e−06
τ		True	1,390,000
		Space	(139,000, 6,950,000)
		Mean	2,360,000
		Median	2,280,000
		Bias	967,000
		Mean *se*	876,000
		Total *se*	1e+06
N_S		True	17,400
		Space	(1740, 174,000)
		Mean	25,700
		Median	24,600
		Bias	8350
		Mean *se*	6620
		Total *se*	7560
N_B		True	17,400
		Space	(1740, 174,000)
		Mean	26,000
		Median	25,200
		Bias	8670
		Mean *se*	6680
		Total *se*	7870

Figures and summaries as in Table 4.

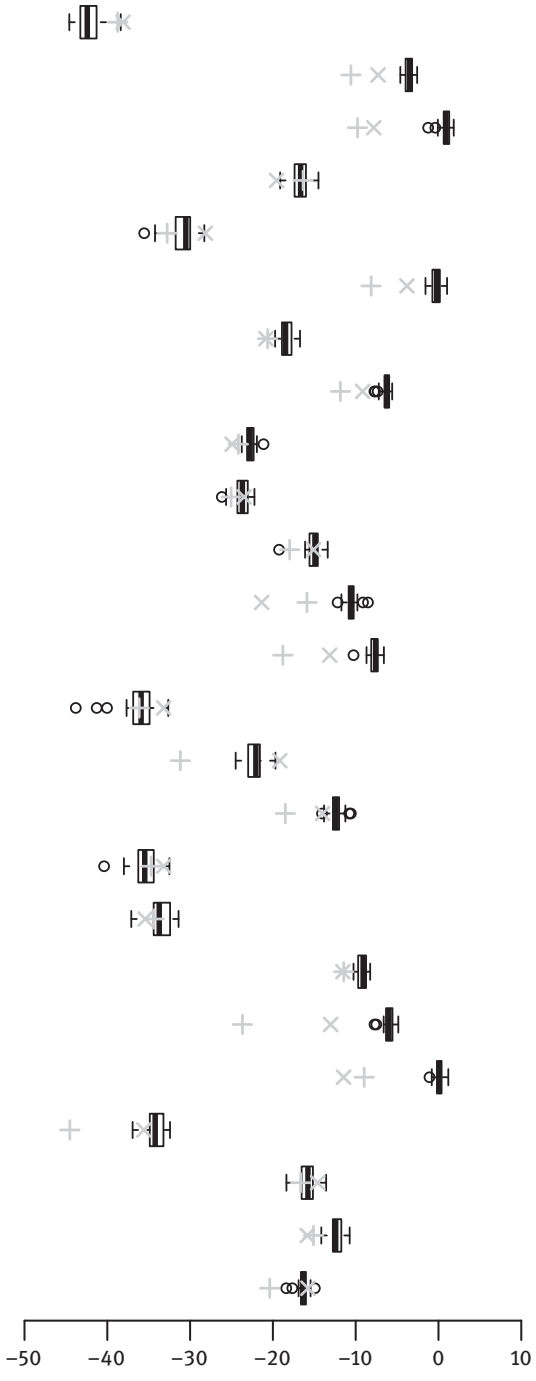
In Locke et al. (2011) and Ma et al. (2013), parameters of two different IM models are estimated; we denote them $\hat{\Theta}_1$ and $\hat{\Theta}_2$. The estimates are scaled to the ancestral population size $N_A = 17{,}372$, and shown in Table 9.

Their model 1 is identical to the model considered here, so we simulate the likelihood at $\hat{\Theta}_1$ within our framework for comparison. Since $\log \hat{L}(\hat{\Theta}_1) = -217.015$ ($se = 7.739$) is significantly lower than $\log \hat{L}(\hat{\Theta}_{\mathrm{AML}}) = -162.732$ ($se = 7.258$), it seems that $\hat{\Theta}_{\mathrm{AML}}$ is closer to the maximum likelihood estimate than the competing estimate. Note, however, that we are only using a subset of the data to avoid sites under selection (see the Appendix for details) and that the authors report convergence problems of ∂a∂i in this model.

For model 2, the ancestral population splits in two subpopulations of relative sizes s and $1 - s$, and the subpopulations experience exponential growth. A direct comparison of the likelihoods is impossible here, because our results were obtained under a different model. However, a rough comparison with the ∂a∂i estimates given in Locke et al. (2011) shows that the AML estimates for τ, N_B and N_S lie between $\hat{\Theta}_1$ and $\hat{\Theta}_2$ and for μ_{BS} and μ_{SB} they are of similar size.

Table 7: Comparison of likelihood estimates at Θ, $\hat{\Theta}_{\text{AML}}$ and $\hat{\Theta}_{\delta a \delta i}$ in the IM-model with $\tau = 1{,}390{,}000$ years.

Dataset	Rank $L(\Theta)$	p-Value	Rank $L(\hat{\Theta}_{\delta a \delta i})$	p-Value	Boxplot
1	26	0.56775	25	0.40934	
2	1	0.00000	1	0.00000	
3	1	0.00000	1	0.00000	
4	1	0.00002	16	0.06099	
5	26	0.52427	6	0.00986	
6	1	0.00000	1	0.00000	
7	1	0.00020	1	0.00036	
8	1	0.00001	1	0.00000	
9	1	0.00279	1	0.01271	
10	15	0.23090	3	0.05010	
11	13	0.04446	2	0.00001	
12	1	0.00000	1	0.00000	
13	1	0.00000	1	0.00000	
14	25	0.36420	12	0.01573	
15	26	0.66990	1	0.00000	
16	2	0.00015	1	0.00000	
17	23	0.31862	17	0.06723	
18	4	0.01892	8	0.06263	
19	1	0.00000	1	0.00001	
20	1	0.00000	1	0.00000	
21	1	0.00000	1	0.00000	
22	6	0.03567	1	0.00000	
23	23	0.24540	6	0.02087	
24	1	0.00000	1	0.00000	
25	23	0.17278	1	0.00000	



Figures and summaries as in Table 5.

5 Discussion

In this article, two related algorithms to approximate the maximum likelihood estimator in models with an intractable likelihood are proposed and carefully investigated. They rely on summary statistics computed from simulated data under the model at hand. Therefore, the methods are flexible and applicable to a wide

Table 8: Parameter estimates for the ancestral history of orang-utans.

	μ_{BS}	μ_{SB}	τ	N_S	N_B
$\hat{\theta}_{AML}$	5.023e−06	4.600e−06	1,300,681	52,998	21,971
$\hat{\theta}^*_{AML}$	4.277e−06	3.806e−06	1,402,715	52,715	22,233
$\widehat{se}^*$	1.244e−06	9.366e−07	208,391	7223	2779
$\widehat{se}$	1.992e−07	1.066e−07	194,868	6083	2963
Lower	0[a]	5.72e−06	715,590	31,476	13,290
Upper	6.627e−06	4.931e−06	1,820,852	67,118	27,426

[a]This confidence interval was cut off at zero. $\hat{\theta}_{AML}$, approximate maximum likelihood estimate; $\hat{\theta}^*_{AML}$, bootstrap bias corrected estimate; $\widehat{se}^*$, bootstrap standard error of $\hat{\theta}_{AML}$; $\widehat{se}$, standard error of $\hat{\theta}_{AML}$ in this dataset, estimated from 20 replicates of $\hat{\theta}_{AML}$; lower and upper limits of the 95% simultaneous bootstrap confidence intervals. All bootstrap results were obtained with $B = 1000$ bootstrap replicates. The simultaneous 95% confidence intervals are computed following a simple Bonferroni argument, using coverage probabilities of 99% in each dimension (Davison and Hinkley, 1997, p. 232).

Table 9: Comparison of $\hat{\theta}_{AML}$ with results from (Locke et al., 2011, Tab. S21-1) [same results for Model 2 reported in Ma et al. (2013)], scaled with $N_e = 17{,}400$.

	μ_{BS}	μ_{SB}	τ	N_S	N_B
Model 1	9.085e−07	7.853e−07	6,948,778	129,889	50,934
Model 2	1.518e−05	2.269e−05	630,931	35,976	10,093
$\hat{\theta}_{AML}$	5.023e−06	4.600e−06	1,300,681	52,998	21,971

Model 1: IM model as shown in Figure 5. Model 2: IM-model where the ancestral population splits in two subpopulations with a ratio of $s = 0.503$ (estimated) going to Borneo and $1 - s$ to Sumatra and exponential growth in both subpopulations (Locke et al., 2011, Fig. S21-3). Here, N_B and N_S are the present population sizes.

variety of problems. We provide examples that show that they reliably approximate the maximum likelihood estimate in challenging applications. Based on extensive simulations, we provide tuning guidelines that make the algorithms run efficiently and reliably.

Alternative simulation based approximate maximum likelihood methods have been proposed that estimate the likelihood surface in an ABC like fashion (Creel and Kristensen, 2013; Rubio and Johansen, 2013) or using MCMC (de Valpine, 2004) by sampling from the whole parameter space. The maximum likelihood estimator is obtained subsequently by standard numerical optimization. Leaving aside the practical challenges of actually computing the MLE, they study the asymptotic properties of their estimator for an increasing number of simulations (Rubio and Johansen, 2013) and observations (Creel and Kristensen, 2013).

By providing tunig guidelines that reduce the number of simulations in low-likelihood regions, our method presented here complements these results and emphasizes the practical applicability in high-dimensional problems.

More generally speaking, our method is related to the class of simulated minimum distance estimators, as described in Forneron and Ng (2015), among them the classical indirect inference estimators (Gouriéroux et al., 1993). The use of an approximation to the likelihood function as distance measure connects our estimator to the realm of maximum likelihood estimation, and also to ABC methods (Gutmann and Corander, 2016).

For our considered population genetic application, we estimate parameters of the evolutionary history of orang-utans and demonstrate that very high-dimensional summary statistics (here: 56 dimensions) can be used successfully without any dimension-reduction techniques. Usually, high-dimensional kernel density estimation is not recommended because of the curse of dimensionality (e.g. Wand and Jones, 1995, p. 90), but stochastic approximation algorithms are explicitly designed to cope with noisy measurements. To this end, we also introduce modifications of the algorithm that reduce the impact of single noisy likelihood estimates. In our experience, this is crucial in settings with a low signal-to-noise ratio.

Furthermore, the examples show that the AML algorithm performs well in problems with a high-dimensional and large parameter space: In our toy example involving the normal distribution, the 10-dimensional

maximum likelihood estimate is approximated very precisely even though the search space spans 200 times the standard error of $\hat{\Theta}_{\mathrm{ML}}$ in each dimension.

However, we also observe a bias for a few of the estimated parameters. Partly, this can be attributed to the bias of the maximum likelihood estimator itself. In addition, it is known that the finite differences approximation to the gradient in the Kiefer-Wolfowitz algorithm causes a bias that vanishes only asymptotically (Spall, 2003, section 6.4.1), and that is possibly increased by the finite-sample bias of the kernel density estimator. In most cases though, the bias is smaller than the standard error of the approximate maximum likelihood estimator and can be made still smaller by carrying out sufficiently long runs of the algorithm.

As informative summary statistics are crucial, the quality of the estimates obtained from our AML algorithm will also depend on an appropriate choice of summary statistics. This has been discussed extensively in the context of ABC (Fearnhead and Prangle, 2012; Blum et al., 2013). General results and algorithms to choose a small set of informative summary statistics should carry over to the AML algorithm.

In addition to the point estimate, we suggest to obtain confidence intervals by parametric bootstrap. The bootstrap replicates can also be used for bias correction. Resampling in models where the data have a complex internal structure catches both the noise of the maximum likelihood estimator as well as the approximation error. Alternatively, the AML algorithm may also complement the information obtained via ABC in a Bayesian framework: the location of the maximum a posteriori estimate can be obtained from the AML algorithm.

The presented work shows the broad applicability of the AML algorithm and also its robustness in settings with high-dimensional summary statistics and a low signal-to-noise ratio.

Acknowledgment: JB was supported by the Vienna Graduate School of Population Genetics [Austrian Science Fund (FWF): W1225-B20] and worked on this project while employed at the Department of Statistics and Operations Research, University of Vienna, Austria. The computational results presented have partly been achieved using the Vienna Scientific Cluster (VSC) and the GenomeDK HPC cluster at Aarhus University. The orang-utan SNP data was kindly provided by X. Ma. Parts of this article have been published in the PhD thesis of Johanna Bertl at the University of Vienna (Bertl, 2014). CK has partially been funded by the Austrian Science Fund (FWF-P24551) and by the Vienna Science and Technology Fund (WWTF) through project MA16-061. The research of AF was supported in part by the National Science Foundation under Grant No. NSF PHY-1125915. We thank Shoji Taniguchi who read the manuscript very thoroughly and pointed us to a few errors.

Appendix

Orang-Utan SNP data

This real data is based on two publications, De Maio et al. (2013) and Ma et al. (2013). For the first, CCDS alignments of *H. sapiens, P. troglodytes* and *P. abelii* (references hg18, panTro2 and ponAbe2) were downloaded from the UCSC genome browser (http://genome.ucsc.edu). Only CCDS alignments satisfying the following requirements were retained for the subsequent analyses: divergence from human reference below 10%, no gene duplication in any species, start and stop codons conserved, no frame-shifting gaps, no gap longer than 30 bases, no nonsense codon, no gene shorter than 21 bases, no gene with different number of exons in different species, or genes in different chromosomes in different species (chromosomes 2a and 2b in non-humans were identified with human chromosome 2). From the remaining CCDSs (9695 genes, 79,677 exons) we extracted synonymous sites. We only considered third codon positions where the first two nucleotides of the same codon were conserved in the alignment, as well as the first position of the next codon.

Furthermore, orang-utan SNP data for the two (Bornean and Sumatran) populations considered, each with five sequenced individuals Locke et al. (2011), were kindly provided by X. Ma and are available online (http://www.ncbi.nlm.nih.gov/projects/SNP/snp_viewTable.cgi?type=contact&handle=WUGSC_SNP&batch_id=1054968). The final total number of synonymous sites included was 1,950,006. Among them, a subset of 9750 four-fold degenerate synonymous sites that are polymorphic in the orang-utan populations were selected.

References

Bach, F. (2014): "Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression," J. Mach. Learn. Res., 15, 595–627.

Beaumont, M. A. (2010): "Approximate Bayesian computation in evolution and ecology," Annu. Rev. Ecol. Evol. S., 41, 379–406.

Beaumont, M. A., W. Zhang and D. J. Balding (2002): "Approximate Bayesian computation in population genetics," Genetics, 162, 2025–2035.

Beaumont, M. A., J.-M. Cornuet, J.-M. Marin and C. P. Robert (2009): "Adaptive approximate Bayesian computation," Biometrika, 96, 983–990.

Bertl, J. (2014): "An approximate maximum likelihood algorithm with case studies," PhD thesis, University of Vienna.

Blum, J .R. (1954): "Multidimensional stochastic approximation methods," Ann. Math. Stat., 25, 737–744.

Blum, M. G. B., M. A. Nunes, D. Prangle and S. A. Sisson (2013): "A comparative review of dimension reduction methods in approximate Bayesian computation," Stat. Sci., 28, 189–208.

Creel, M. and D. Kristensen (2013): "Indirect Likelihood Inference (revised)," UFAE and IAE Working Papers 931.13. URL http://ideas.repec.org/p/aub/autbar/931.13.html.

Davison, A. C. and D. V. Hinkley (1997): Bootstrap methods and their applications, Cambridge University Press, Cambridge.

De Maio, N., C. Schlötterer and C. Kosiol (2013): "Linking great apes genome evolution across time scales using polymorphism-aware phylogenetic models," Mol. Biol. Evol., 30, 2249–2262.

De Valpine, P. (2004): "Monte Carlo state-space likelihoods by weighted posterior kernel density estimation," J. Am. Stat. Assoc., 99, 523–536.

Dieuleveut, A. and F. Bach (2016): "Non-parametric stochastic approximation with large step sizes," Ann. Stat., 44, 1363–1399.

Diggle P. J. and R. J. Gratton (1984): "Monte Carlo methods of inference for implicit statistical models," J. R. Stat. Soc. B, 46, 193–227.

Drovandi, C. C., A. N. Pettitt and M. J. Faddy (2011): "Approximate Bayesian computation using indirect inference," J. R. Stat. Soc. C, 60, 317–337.

Ehrlich, E., A. Jasra and N. Kantas (2013): "Gradient free parameter estimation for hidden Markov models with intractable likelihoods," Methodol. Comput. Appl. Probab., 17, 1–35.

Ewing, G. and J. Hermisson (2010): "MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus," Bioinformatics, 26, 2064–2065.

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

Fermanian, J.-D. and B. Salanié (2004): "A nonparametric simulated maximum likelihood estimation method," Economet. Theor., 20, 701–734.

Forneron , J.-J. and S. Ng (2015): "The ABC of simulation estimation with auxiliary statistics," Technical report, arXiv.

Gouriéroux, C., A. Monfort and E. Renault (1993): "Indirect inference," J. Appl. Econometr., 8, 85–118.

Gutenkunst, R. N., R. D. Hernandez, S. H. Williamson and C. D. Bustamante (2009): "Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data," PLoS Genet., 5, 1–11.

Gutmann, M. U. and J. Corander (2016): "Bayesian optimization for likelihood-free inference of simulator-based statistical models," J. Mach. Learn. Res., 17, 1–47.

Härdle, W., M. Müller, S. Sperlich and A. Werwatz (2004): Nonparametric and semiparametric models, Springer Series in Statistics. Springer, New York.

Heggland , K. and A. Frigessi (2004): "Estimating functions in indirect inference," J. R. Stat. Soc. B, 66, 447–462.

Kiefer , J. and J. Wolfowitz (1952): "Stochastic estimation of the maximum of a regression function," Ann. Math. Stat., 23, 462–466.

Locke, D. P., L. W. Hillier, W. C. Warren, K. C. Worley, L. V. Nazareth, D. M. Muzny, S.-P. Yang, Z. Wang, A. T. Chinwalla, P. Minx, M. Mitreva, L. Cook, K. D. Delehaunty, C. Fronick, H. Schmidt, L. A. Fulton, R. S. Fulton, J. O. Nelson, V. Magrini, C. Pohl, T. A. Graves, C. Markovic, A. Cree, H. H. Dinh, J. Hume, C. L. Kovar, G. R. Fowler, G. Lunter, S. Meader, A. Heger, C. P. Ponting, T. Marques-Bonet, C. Alkan, L. Chen, Z. Cheng, J. M. Kidd, E. E. Eichler, S. White, S. Searle, A. J. Vilella, Y. Chen, P. Flicek, J. Ma, B. Raney, B. Suh, R. Burhans, J. Herrero, D. Haussler, R. Faria, O. Fernando, F. Darré, D. Farré, E. Gazave, M. Oliva, A. Navarro, R. Roberto, O. Capozzi, N. Archidiacono, G. Della Valle, S. Purgato, M. Rocchi, M. K. Konkel, J. A. Walker, B. Ullmer, M. A. Batzer, A. F. Smit, R. Hubley, C. Casola, D. R. Schrider, M. W. Hahn, V. Quesada, X. S. Puente, G. R. Ordoñez, C. López-Otín, T. Vinar, B. Brejova, A. Ratan, R. S. Harris, W. Miller, C. Kosiol, H. A. Lawson, V. Taliwal, A. L. Martins, A. Siepel, A. Roychoudhury, X. Ma, J. Degenhardt, C. D. Bustamante, R. N. Gutenkunst, T. Mailund, J. Y. Dutheil, A. Hobolth, M. H. Schierup, O. A. Ryder, Y. Yoshinaga, P. J. de Jong, G. M. Weinstock, J. Rogers, E. R. Mardis, R. A. Gibbs and R. K. Wilson (2011): "Comparative and demographic analysis of orang-utan genomes," Nature, 469, 529–533.

Ma, X., J. L. Kelly, K. Eilertson, S. Musharoff, J. D. Degenhardt, A. L. Martins, T. Vinar, C. Kosiol, A. Siepel, R. N. Gutenkunst and C. D. Bustamante (2013): "Population genomic analysis reveals a rich speciation and demographic history of orang-utans (*Pongo pygmaeus* and *Pongo abelii*)," PLoS One, 8, 1–11.

Marjoram, P. and S. Tavaré (2006): "Modern computational approaches for analysing molecular genetic variation data," Nat. Rev. Genet., 7, 759–770.

McKinley, T., A. R. Cook and R. Deardon (2009): "Inference in epidemic models without likelihoods," Int. J. Biostat., 5, 1–37.

Meeds, E., R. Leenders and M. Welling (2015): "Hamiltonian ABC," arXiv preprint, (arXiv:1503.01916).

Naduvilezhath, L. N., L. E. Rose and D. Metzler (2011): "Jaatha: a fast composite-likelihood approach to estimate demographic parameters," Mol. Ecol., 20, 2709–2723.

Rosenblatt, M. (1991): Stochastic curve estimation, IMS, Hayward, CA.

Rubio, F. J. and A. M. Johansen (2013): "A simple approach to maximum intractable likelihood estimation," Electron. J. Stat., 7, 1632–1654.

Sadegh, P. (1997): "Constrained optimization via stochastic approximation with a simultaneous perturbation gradient approximation," Automatica, 33, 889–892.

Scott, D. (2015): Multivariate density estimation: theory, practice, and visualization, Wiley Series in Probability and Statistics. Wiley, New York.

Soubeyrand, S., F. Carpentier, N. Desassis and J. Chadœuf (2009): "Inference with a contrast-based posterior distribution and application in spatial statistics," Stat. Methodol., 6, 466–477.

Spall, J. C. (1992): "Multivariate stochastic approximation using a simultaneous perturbation gradient approximation," IEEE T. Automat. Contr., 37, 352–355.

Spall, J. C. (2003): Introduction to stochastic search and optimization: estimation, simulation and control, Wiley, New York.

Stephens, M. (2007): "Inference under the coalescent," In: Balding, D. J., Bishop, M., and Cannings, C. (Eds.), Handbook of statistical genetics, volume 2, John Wiley & Sons, New York, third edition, pp. 878–908.

Tellier, A., P. Pfaffelhuber, B. Haubold, L. Naduvilezhath, L. E. Rose, T. Städler, W. Stephan and D. Metzler (2011): "Estimating parameters of speciation models based on refined summaries of the joint site-frequency spectrum," PLoS One, 6, 5.

Toni, T., D. Welch, N. Strelkowa, A. Ipsen and M. P. H. Stumpf (2009): "Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems," J. Roy. Soc. Interface, 6, 187–202.

Wand, M. P. and M. C. Jones (1995): Kernel smoothing, Chapman & Hall, Boca Raton.

Wegmann, D., C. Leuenberger and L. Excoffier (2009): "Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood," Genetics, 182, 1207–1218.

Wellek, S. (2010): Testing statistical hypotheses of equivalence and noninferiority, CRC Press, Taylor & Francis, Boca Raton.