Approximate Maximum Likelihood Estimation

Johanna Bertl*
Aarhus University, Denmark

Gregory Ewing
École polytechnique fédérale de Lausanne, Switzerland

Carolin Kosiol Andreas Futschik
Vetmeduni Vienna, Austria Johannes Kepler University Linz, Austria

July 17, 2015

Abstract

In recent years, methods of approximate parameter estimation have attracted considerable interest in complex problems where exact likelihoods are hard to obtain. In their most basic form, Bayesian methods such as Approximate Bayesian Computation (ABC) involve sampling from the parameter space and keeping those parameters that produce data that fit sufficiently well to the actually observed data. Exploring the whole parameter space, however, makes this approach inefficient in high dimensional problems. This led to the proposal of more sophisticated iterative methods of inference such as particle filters.

Here, we propose an alternative approach that is based on stochastic gradient methods and applicable both in a frequentist and a Bayesian setting. By moving along a simulated gradient, the algorithm produces a sequence of estimates that will eventually converge either to the maximum likelihood estimate or to the maximum of the posterior distribution, in each case under a set of observed summary statistics. To avoid reaching only a local maximum, we propose to run the algorithm from a set of random starting values.

As good tuning of the algorithm is important, we explored several tuning strategies, and propose a set of guidelines that worked best in our simulations. We investigate the performance of our approach in simulation studies, and also apply the algorithm to two models with intractable likelihood functions. First, we present an application to inference in the context of queuing systems. We also re-analyze population genetic data and estimate parameters describing the demographic history of Sumatran and Bornean orang-utan populations.

* johanna.bertl@clin.au.dk
1 Introduction

Both in the Bayesian as well as in the frequentist framework, statistical inference commonly uses the likelihood function. However, under a wide range of complex models, no explicit formula is available. This occurs for example when modelling population genetic and evolutionary processes (Marjoram and Tavaré 2006), in spatial statistics (Soubeyrand et al. 2009), and in queuing systems (Heggland and Fregessi 2004). Furthermore, such situations also occur with dynamical systems as used for example in systems biology (Toni et al. 2009), and epidemiology (McKinley et al. 2009). Here, we consider statistical models with an intractable distribution theory, but a known data generating process under which simulations can be obtained.

A recent approach to overcome this problem is Approximate Bayesian Computation (ABC) (Beaumont et al. 2002). Its basis is a rejection algorithm: Parameter values are randomly drawn from the prior distribution, data sets are then simulated under these values. To reduce complexity, informative but low dimensional summaries are derived from the data sets. All parameter values that gave rise to summary statistics similar to those computed for the observed data are then accepted as a sample from the posterior distribution. Hence, ABC does not involve any kind of direct approximation of the likelihood, but with a uniform prior (with support on a compact subset of the parameter space taken large enough to contain the maximum likelihood estimator) it 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 (Creel and Kristensen 2013; Rubio and Johansen 2013).

ABC has successfully been used in population genetic applications and also in other fields (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 needs to be relaxed, but this increases the bias in the approximation to the posterior distribution. To reduce the sampling space, combinations of ABC and iterative Monte Carlo methods such as particle filters and Markov Chain Monte Carlo have been proposed (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 whole parameter space, we propose a stochastic gradient algorithm that approximates the maximum likelihood estimate. Similar to ABC, it relies on lower-dimensional summary statistics of the data. In the simultaneous perturbations algorithm, in each iteration two noisy evaluations of the likelihood are sufficient to obtain an ascent direction (Spall 1992). To this end the likelihood is approximated by kernel density estimation on summary statistics of data simulated under the parameter value of interest.

Our algorithm 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 1-dimensional i.i.d. data. In principle, our approach can also be applied in the context of indirect inference, where summary statistics are derived from a tractable auxiliary model (Drovandi et al. [2011]).

2 Method

We will start by describing the so called simultaneous perturbation algorithm for optimizing the expected value of a random function using simulations. Next, we explain how this algorithm can be adapted to obtain maximum likelihood estimates. Details on the tuning can be found in the following section.

2.1 General simultaneous perturbation algorithm

Let \( Y \in \mathbb{R} \) be a random variable depending on \( \Theta \in \mathbb{R}^p \). The function \( L(\Theta) = E(Y \mid \Theta) \) shall be maximized in \( \Theta \), but \( L(\Theta) \) as well as the gradient \( \nabla L(\Theta) \) are unknown. If a realization \( y(\Theta) \) of \( Y \mid \Theta \) can be observed for any value of \( \Theta \), the maximizer \( \arg \max_{\Theta \in \mathbb{R}^p} L(\Theta) \) can be approximated by the simultaneous perturbation algorithm (Spall [1992]).

Similar to a deterministic gradient algorithm, it is based on the recursion

\[
\Theta_k = \Theta_{k-1} + a_k \nabla L(\Theta_{k-1}),
\]

where \( a_k \in \mathbb{R}^+ \) is a decreasing sequence. However, as \( \nabla L(\Theta_{k-1}) \) is unknown, it is substituted by a rescaled putative ascent direction that is randomly chosen among the vertices of the unit hypercube. Thus, in iteration \( k \), a random vector \( \delta_k \) with elements

\[
\delta_k^{(i)} = \begin{cases} 
-1 & \text{with probability } 1/2, \\
+1 & \text{with probability } 1/2,
\end{cases}
\]

for \( i = 1, \ldots, p \) is generated, and the ascent direction is approximated by

\[
\delta_k \frac{y(\Theta_k + c_k \delta_k) - y(\Theta_k - c_k \delta_k)}{2c_k},
\]

with \( c_k \in \mathbb{R}^+ \) being a decreasing sequence.

2.2 Approximate maximum likelihood algorithm

Suppose, data \( D_{obs} \) are observed under model \( \mathcal{M} \) with unknown parameter vector \( \Theta \in \mathbb{R}^p \). Let \( L(\Theta; D_{obs}) = p(D_{obs} \mid \Theta) \) denote the likelihood of \( \Theta \). 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 \( L(\Theta; S_{obs}) = p(S_{obs} \mid \Theta) \), an approximation to the
original likelihood that uses a $d$-dimensional vector of summary statistics $S_{\text{obs}}$ instead of the original data $D_{\text{obs}}$.

An estimate $\hat{L}(\Theta; S_{\text{obs}})$ of $L(\Theta; S_{\text{obs}})$ can be obtained from simulations under $\mathcal{M}(\Theta)$ using kernel density estimation. By setting $y(\Theta) = \hat{L}(\Theta; S_{\text{obs}})$, we adopt the simultaneous perturbation algorithm to approximate the maximum likelihood estimate $\hat{\Theta}_{\text{ML}}$ of $\Theta$. In more detail, the algorithm is as follows:

**Algorithm (AML).** Let $a_k, c_k \in \mathbb{R}^+$ be two decreasing sequences. Let $H_k$ be a sequence of symmetric positive definite $d \times d$ matrices and $\kappa$ a $d$-dimensional kernel function satisfying $\int_{\mathbb{R}^d} \kappa(x) dx = 1$. Choose a starting value $\Theta_0 \in \mathbb{R}^p$.

For $k = 1, 2, \ldots, K$:

1. **Choice of a likely ascent direction in $\Theta_{k-1}$**:
   
   (a) Generate a random vector $\delta_k$ as defined in equation (1).
   
   (b) Simulate datasets $D_1^-, \ldots, D_n^-$ from $\mathcal{M}(\Theta^-)$ and $D_1^+, \ldots, D_n^+$ from $\mathcal{M}(\Theta^+)$ with $\Theta^\pm = \Theta_{k-1} \pm c_k \delta_k$.
   
   (c) Compute summary statistics $S_j^-$ on dataset $D_j^-$ and $S_j^+$ on $D_j^+$ for $j = 1, \ldots, n$.
   
   (d) 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 the summary statistics $S_1^-, \ldots, S_n^-$ and $S_1^+, \ldots, S_n^+$, respectively, with multivariate kernel density estimation (Wand and Jones, 1995):

   $$
   \hat{L}(\Theta^-; S_{\text{obs}}) = \frac{1}{n \sqrt{\det(H_k)}} \sum_{j=1}^n \kappa \left( H_k^{-1/2} (S_{\text{obs}} - S_j^-) \right)
   $$

   and analogously for $\hat{L}(\Theta^+, S_{\text{obs}})$.
   
   (e) Estimate the ascent direction $\nabla l(\Theta_{k-1}; S_{\text{obs}})$ by

   $$
   \hat{\nabla}_{\delta_k} \hat{l}(\Theta_{k-1}; S_{\text{obs}}) = \delta_k \frac{\log \hat{L}(\Theta^+; S_{\text{obs}}) - \log \hat{L}(\Theta^-; S_{\text{obs}})}{2c_k}
   $$

2. **Updating $\Theta_k$**:

   $$
   \Theta_k = \Theta_{k-1} + a_k \hat{\nabla}_{\delta_k} \hat{l}(\Theta_{k-1}; S_{\text{obs}})
   $$

Then, the approximate maximum likelihood (AML) estimate is $\hat{\Theta}_{\text{AML}} := \Theta_K$.

If, in a Bayesian setting, a prior distribution $\pi(\Theta)$ has been specified, the algorithm can be modified to approximate the maximum of the posterior distribution, $\hat{\Theta}_{\text{MAP}}$, by multiplying $\hat{L}(\Theta^-, S_{\text{obs}})$ and $\hat{L}(\Theta^+, S_{\text{obs}})$ by $\pi(\Theta^-)$ and $\pi(\Theta^+)$, respectively.
2.3 Parametric bootstrap

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

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

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

where $q(\beta)(\hat{\Theta}_{AML}^*)$ denotes the $\beta$ quantile of $\hat{\Theta}_{AML,1}^*, \ldots, \hat{\Theta}_{AML,B}^*$ (Davison and Hinkley, 1997).

3 Tuning guidelines

The performance of the algorithm strongly depends on the choice of the sequences $a_k$ and $c_k$. However, their optimal values depend on the unknown likelihood function. Another challenge in the practical application is the stochasticity: large steps can by chance lead very far away from the maximum.

3.1 Choice of $a_k$ and $c_k$ and convergence diagnostics

Here, we consider sequences of the form

$$a_k = \frac{a}{(k + A)^\alpha}, \text{ and } c_k = \frac{c}{k^\gamma}$$

as proposed in Spall (2003). To reflect the varying slope of $L$ in different directions of the parameter space, we choose $a \in \mathbb{R}^p$ (this is equivalent to scaling the space accordingly). Choosing $\alpha = 1$ and $\gamma = 1/6$ ensures optimal convergence speed (Spall, 2003). The optimal choice of $a$ and $c$ depends on the unknown shape of $L$. Therefore, we propose the following heuristic based on suggestions in Spall (2003) 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 $K$ be the number of planned iterations and $b \in \mathbb{R}^p$ a vector that gives the desired stepsizes in early iterations in each dimension. Choose a starting value $\Theta_0$.

1. Set $c$ to a small percentage of the total parameter range (usually between 1% and 5%) to obtain $c_1$.
2. Set $A = \lfloor 0.1 * K \rfloor$. 

5
Choose $a$:

(a) Estimate $\nabla l(\Theta_0; S_{\text{obs}})$ by the median of $n_1$ finite differences approximations (step 1 of the AML algorithm), $\hat{\nabla}_c \hat{l}(\Theta_0, S_{\text{obs}})$.

(b) Set

$$a^{(i)} = \frac{b^{(i)}(A + 1)^\alpha}{(\hat{\nabla}_c \hat{l}(\Theta_0; S_{\text{obs}}))^{(i)}}$$

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

As $a$ is determined using information about the likelihood in $\Theta_0$ 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 $K_0$ iterations the following three tests are conducted on the preceding $K_0$ iterates:

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

  $$\Theta^{(i)}_k = \Theta^{(i)}_{k-1} + \beta + \epsilon_k,$$

  where $\beta$ 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 $\Delta_k = \Theta^{(i)}_k - \Theta^{(i)}_{k-1}$. If a trend is detected, $a^{(i)}$ is increased by a fixed factor $f \in \mathbb{R}^+$.  

- **Range test (too large $a$):** For each dimension $i = 1, \ldots, p$, $a^{(i)}$ is set to $a^{(i)}/f$ if the trajectory of $\Theta^{(i)}_k$ spans more than 70% of the parameter range.

- **Convergence test:** Simulate $n_2$ likelihood estimates at $\Theta_{k-K_0}$ and at $\Theta_k$. Growth of the likelihood is then tested by a one-sided Welch’s $t$-test. (A standard hypothesis 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. In the applications of the algorithm presented in the article, $f = 1.5$.

### 3.2 Kernel density estimation

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 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 1d:

If \( \hat{L}(\Theta^-; S_{\text{obs}}) \approx 0 \) or/and \( \hat{L}(\Theta^+; S_{\text{obs}}) \approx 0 \) (with “\( \approx \)” denoting “numerically equivalent to”), find

\[
S^-_{\min} := \arg \min_{S^-_j} \{ \| S^-_j - S_{\text{obs}} \| : j = 1, \ldots, n \}
\]
\[
S^+_{\min} := \arg \min_{S^+_j} \{ \| S^+_j - S_{\text{obs}} \| : j = 1, \ldots, n \}
\]

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

To be computationally efficient, we estimate a diagonal bandwidth matrix \( H_k \) using a multivariate extension of Silverman’s rule (Silverman, 1986; Hardle et al., 2004). Using new estimates in each iteration introduces an additional level of noise that can be reduced by using a moving average of bandwidth estimates.

### 3.3 Starting points

To avoid starting in very unfavourable regions of the likelihood, a set of random points should be drawn from the parameter space and their simulated likelihood values compared to find useful starting points. This strategy also helps to avoid that the algorithm reaches only a local maximum.

### 3.4 Constraints

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 \( \Theta^- \) as well as \( \Theta^+ \) are in the feasible set (Sadegh, 1997). Even if there are no imperative constraints, it is advisable to restrict the parameter 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 10% of the range of the feasible set in each dimension.
4 Examples

To study the performance of the AML algorithm, we test it on different applications. The first example is 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. Second, we apply the algorithm to a queuing process where the exact evaluation of the likelihood can require a prohibitive number of steps. The third example is from population genetics. Here, 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

When the AML algorithm is used to estimate the mean vector $\Theta = (\mu_1, \ldots, \mu_{10})$ of a 10-dimensional normal distribution with a diagonal VC matrix using the arithmetic means $\bar{X} = (\bar{X}_1, \ldots, \bar{X}_{10})$ as summary statistics, convergence is achieved quickly and the results are extremely accurate.

One dataset is simulated under a 10-dimensional normal distribution such that the maximum likelihood estimator for $\Theta$ is $\hat{\Theta}_{ML} = \bar{X} \sim \mathcal{N}(5 \cdot \mathbf{1}_{10}, I_{10})$ 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 $\hat{\Theta}_{AML}$, the AML algorithm is run 1000 times on this dataset with summary statistics $S = \bar{X}$. For each AML estimate, 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 algorithm is run for at least 10000 iterations and stopped as soon as convergence is reached (for $\approx 90\%$ of the sequences within 11000 iterations; convergence is tested every $K_0 = 1000$ iterations). Again, the likelihood is simulated on each result and the one with the highest likelihood is considered as a realization of $\hat{\Theta}_{AML}$. For each evaluation of the likelihood, $n = 100$ datasets are simulated. Based on these 1000 realizations of $\hat{\Theta}_{AML}$, the density, bias and standard error of $\hat{\Theta}_{AML}$ are estimated for each dimension (tab. 1, fig. 1).

The densities of the 10 components of $\hat{\Theta}_{AML}$ are symmetric around the maximum likelihood estimate with a standard error that is nearly 20 times smaller than the standard error of $\hat{\Theta}_{ML}$ and a negligible bias. 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%.

To investigate the impact of the choice of summary statistics on the results, we repeat the experiment with the following set of summary statistics:
\[ X_1 \ X_2 \ X_3 \ X_4 \ X_5 \ X_7 \ X_9 \ X_6 \ X_8 \ X_{10} \]

(2)

For \( \approx 90\% \) of the sequences, convergence was detected within 14000 iterations. Bootstrap confidence intervals are obtained for 100 datasets. Their coverage probability matches the nominal 95\% confidence level closely (tab. 1). The components behave very similarly to the previous simulations, except for the estimates of the density for components 9 and 10 (fig. 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 \( \hat{\Theta}_{ML} \). To investigate how fast the bias decreases with the number of iterations, we re-run the above described algorithm for 100000 iterations without earlier stopping (fig. 3).

Both bias and standard error decrease with the number of iterations.

Table 1: Properties of \( \hat{\Theta}_{AML} \) in the 10-dimensional normal distribution model using 2 different sets of summary statistics

| dim. | \( \hat{b} \) | \( \hat{se} \) | \( \hat{p} \) | \( \hat{b} \) | \( \hat{se} \) | \( \hat{p} \) |
|------|-------------|-------------|-------------|-------------|-------------|-------------|
| 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 (\( \hat{se} \)) of \( \hat{\Theta}_{AML} \), estimated from 1000 runs of the AML algorithm. Coverage probability of bootstrap 95\% confidence intervals (\( \hat{p} \)) with \( B = 100 \), estimated from 100 datasets.

4.2 Queueing process

The theory of queueing processes is a very prominent and well-studied field of applied probability with applications in many different areas like production management, communication and information technology, and health care (Kingman 2009; Gross et al. 2008). Here, we consider a first-come-first-serve queuing process with a single server, a Markovian distribution of interarrival times of customers and a general distribution of service times (M/G/1). It is modelled by the following sequences of random variables: The service time
Figure 1: Density of the components of $\hat{\Theta}_{AML}$ obtained with $S = \bar{X}$ in one dataset estimated from 1000 converged sequences with a minimum length of 10000 iterations by kernel density estimation. Vertical dashed line: $\hat{\Theta}_{ML}$.

Figure 2: Density of the components of $\hat{\Theta}_{AML}$ obtained with $S$ as in eq. (2) in one dataset estimated from 1000 converged sequences with a minimum length of 10000 iterations by kernel density estimation. Vertical dashed line: $\hat{\Theta}_{ML}$.
Figure 3: Absolute bias and standard error of the AML estimator of $\mu_9$ and $\mu_{10}$ using the summary statistics in eq. (2) estimated from 1000 runs of the algorithm on the same dataset.

$U_m$, the interarrival time $W_m$ and the interdeparture time $Y_m$ for customer $m = 1, \ldots, M$ [Heggland and Frigessi (2004)]. $U_m$ and $W_m$ are independent and their distributions are known except for the parameters. Here we assume $U_m \sim U(\theta_1, \theta_1 + \theta_2)$ and $W_m \sim \text{exp}(\theta_3)$. $Y_m$ is defined as

$$Y_m = \begin{cases} 
U_m & \text{if } \sum_{i=1}^{m} W_i \leq \sum_{i=1}^{m-1} Y_i, \\
U_m + \sum_{i=1}^{m} W_i - \sum_{i=1}^{m-1} Y_i & \text{otherwise}
\end{cases}$$

for $m = 1, \ldots, M$.

If only the interdeparture times $Y_m$ are observed, the evaluation of the likelihood of the parameter vector $\Theta = (\theta_1, \theta_2, \theta_3)$ requires an exponential number of steps in $M$ [Heggland and Frigessi (2004)]. Approximate inference has been conducted with indirect inference [Heggland and Frigessi (2004)] and ABC methods (Blum and François (2010)).

To estimate the distribution of $\hat{\Theta}_{\text{AML}}$ in this model, we simulate 100 datasets of size $M = 100$ under the same parameter value $\Theta = (1, 4, 0.2)$ and compute $\hat{\Theta}_{\text{AML}}$ for each of them. The summary statistics are the minimum, maximum and the quartiles of $Y_1, \ldots, Y_M$ (Blum and François (2010)). To estimate the standard error of $\hat{\Theta}_{\text{AML}}$ per dataset, we re-run the algorithm five times on each dataset.

Each time, we randomly draw 100 starting values from the search space $(0,10) \times (0,10) \times (0.05, 10)$ and run the AML algorithm on the best five starting values for 5000 iterations (runtime of an implementation in R, version 3.0.1 [R Core Team (2013): 1.9 hours on a single core). Finally, we estimate the
likelihood on the 5 results and the best one is used as a realization of \( \hat{\Theta}_{\text{AML}} \). Each likelihood estimation is based on \( n = 100 \) simulations. For 3 datasets, the AML algorithm is run 100 times to obtain kernel density estimates of the density of \( \hat{\Theta}_{\text{AML}} \).

**Table 2: Properties of \( \hat{\Theta}_{\text{AML}} \) in the M/G/1 queue**

| \( \theta_1 \) | true | 1 | space | (0, 10) | mean | 0.942 | median | 0.948 | bias | -0.058 | mean se | 0.062 | total se | 0.098 |
| \( \theta_2 \) | true | 5 | space | (0, 10) | mean | 5.031 | median | 4.891 | bias | 0.031 | mean se | 0.437 | total se | 1.046 |
| \( \theta_3 \) | true | 0.2 | space | (0.05, 10) | mean | 0.229 | median | 0.226 | bias | 0.029 | mean se | 0.004 | total se | 0.024 |

Figures: marginal densities of components of \( \hat{\Theta}_{\text{AML}} \), estimated by kernel density estimation. Black solid line: density of \( \hat{\Theta}_{\text{AML}} \) estimated over 100 simulated data sets. Coloured dotted and dashed lines: density of \( \hat{\Theta}_{\text{AML}} \) estimated separately in 3 example datasets. Vertical black line: true parameter value.

Summaries: \textit{true}: true parameter value; \textit{space}: search space; \textit{mean}: mean of \( \hat{\Theta}_{\text{AML}} \); \textit{median}: median of \( \hat{\Theta}_{\text{AML}} \); \textit{bias}: bias of \( \hat{\Theta}_{\text{AML}} \); \textit{mean se}: mean standard error of \( \hat{\Theta}_{\text{AML}} \) per dataset; \textit{total se}: standard error of \( \hat{\Theta}_{\text{AML}} \).

In all three dimensions, the standard error of \( \hat{\Theta}_{\text{AML}} \) is considerably larger across datasets than within dataset (tab. 2). This indicates that the approximation algorithm only adds a small additional error to the error of \( \hat{\Theta}_{\text{ML}} \). In all dimensions the density of \( \hat{\Theta}_{\text{AML}} \) is slightly asymmetric around the true value, but only the bias of \( \theta_3 \) is of the same order as the standard error. It is probably caused by a strongly skewed likelihood function \cite{Blum2010} that emphasises the error of the finite differences approximation to the slope. Additional simulations of the likelihood on a grid between \( \theta_3 \) and \( \theta_3 \) locate the maximum around 0.209. Re-running the algorithm for \( K = 10000 \) iterations...
reduces the bias from 0.029 to 0.024. The parameter estimate with the largest standard error is $\hat{\theta}_2$ which is also the most difficult parameter to estimate with ABC (Blum and François 2010).

4.3 Population genetics

The goal of population genetics is to understand how evolutionary forces like mutation, recombination, selection and genetic drift shape the variation within a population. A major challenge in the application of population genetic theory is to infer parameters of the evolutionary history of a population from DNA sequences of present-day samples only. The questions range from estimating single parameters like the mutation rate or the recombination rate to determining a complex demographic structure with multiple parameters.

Since the introduction of the coalescent by Kingman in the early eighties (Kingman 1982a,b,c), a very elegant and flexible stochastic process is at hand to model the ancestral history of a population. The coalescent can capture population structure, varying population size and different mating schemes. Extensions to incorporate recombination and selection exist as well (Nordborg 2007; Wakeley 2008).

However, statistical inference on data obtained from a coalescent process is difficult: Under the coalescent, computing the likelihood of a parameter vector is a computationally very expensive task even for a single locus, since all possible ancestral trees need to be taken into account (Stephens 2007). The number of binary tree topologies explodes with the number of sampled individuals, $n$, e. g. with $n = 10$, the number of possible topologies is $2,571,912,000$, with $n = 1000$, it is already $3.01 \cdot 10^{4831}$ (Wakeley 2008, p. 83, table 3.2).

Consequently, exact computation of the likelihood is feasible in reasonable time for a very restricted set of models and a small number of haplotypes only (Griffiths and Tavare 1994; Wu 2010). To make use of the large DNA data sets generated by modern sequencing technology, approximate methods have been developed in frequentist as well as in Bayesian frameworks. In fact, the need for inferential tools that allow for large genomic data sets was a major driver in the development of ABC. The focus of the early ABC papers was clearly on population genetic applications (Weiss and von Haeseler 1998; Pritchard et al. 1999; Beaumont et al. 2002) and since then it has found a large variety of applications in the field of population genetics (Csilléry et al. 2010; Beaumont 2010; Bertorelle et al. 2010).

4.3.1 Population genetic history of 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 that led to a cooler, drier and more seasonal climate might have contracted rain forest and isolated the population 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 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); the only analyses that have been performed genome-wide. Both studies use DaDi (Gutenkunst et al., 2009), a state-of-art software that has been widely used for demographic inference. It is based on the diffusion approximation to the coalescent. The orang-utan SNP data consisting of 4-fold degenerate (synonymous) sites are taken from the 10 sequenced individuals (5 individuals each per orang-utan population, haploid sample size 10, Locke et al., 2011).

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 $\tau$ 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 $\mu_{BS}$ (fraction of the Bornean population that is replaced by Sumatran migrants per generation) and $\mu_{SB}$ (vice versa; fig. 4a).

Figure 4: (a) Isolation-migration model for the ancestral history of orang-utans. $N_A$, effective size of the ancestral population; $\mu_{BS}$ ($\mu_{SB}$), fraction of the Bornean (Sumatran) population that is replaced by Sumatran (Bornean) migrants per generation (backwards migration rate); $\tau$, 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).

$N_A$ is set to the present effective population size that we obtain using the number of SNPs in our considered data set and assuming an average per gener-
ation mutation rate per nucleotide of $2 \cdot 10^{-8}$ and a generation time of 20 years (Locke et al., 2011), so $N_A = 17400$.

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 (fig. 4b). As the ancestral state is unknown, we use the folded binned JSFS that has 28 entries. 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 algorithm with the JSFS as summary statistics was implemented as an extra module in the msms package, a coalescent simulation program (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.

### 4.3.2 Simulations

Before applying the AML algorithm to the actual orang-utan DNA sequences, we tested it on simulated data. Under the described IM model with parameters $N_B = N_S = 17400$, $\mu_{BS} = \mu_{SB} = 1.44 \cdot 10^{-5}$ and $\tau = 695000$, we simulated 25 datasets with 25 haploid sequences per deme, each of them consisting of 75 loci with 130 SNPs each.

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 5 values with the highest likelihood estimates were used as starting points for the AML algorithm. The algorithm converged after 3000-25000 iterations (average: $\approx 8000$ iterations; average runtime of msms: 11.7 hours on a single core).

For the migration rates $\mu_{SB}$ and $\mu_{BS}$, the per dataset variation of the estimates is considerably smaller than the total variation (tab. 3). This suggests that the approximation error of the AML algorithm is small in comparison to the error of $\hat{\Theta}_{ML}$. For the split time $\tau$ 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 $\hat{\Theta}_{ML}$ 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. Only in 2 of the 25 datasets, the likelihood at the true value is higher, but not significantly so, whereas it is significantly lower in 12 of them (significance was tested with a one-sided Welch test using a Bonferroni-correction to account for multiple testing). This suggests that the AML algorithm usually produces estimates that
are closer to the maximum likelihood estimate than the true parameter value is.

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 \( \tau \) twice as large. Here, all parameter estimates, especially \( \tau \), \( N_B \) and \( N_S \) had larger standard errors and biases (tab. 4). Apparently, the estimation problem is more difficult for more distant split times. This can 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 \( \tau \): a longer divergence time allows for more migration events that might facilitate their analysis.

### 4.3.3 Real data

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

To study the distribution of \( \hat{\Theta}_{AML} \) on this dataset, the algorithm is run 20 times with the same scheme as in the simulations, so we can estimate the standard error of \( \hat{\Theta}_{AML} \) in this single dataset. The best one is used as the final parameter estimate \( \hat{\Theta}_{AML} \) and is used for bootstrapping.

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 (tab. 5).

In Locke et al. (2011) and Ma et al. (2013), parameters of two different IM models are estimated, denote the estimates \( \hat{\Theta}_1 \) and \( \hat{\Theta}_2 \). Scaled to the ancestral population size \( N_A = 17372 \), the estimates are shown in tab. 6.

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 \( \log \hat{L} (\hat{\Theta}_{AML}) = -162.732 \) \((se = 7.258)\), it seems that \( \hat{\Theta}_{AML} \) 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 and that the authors report convergence problems of DaDi in this model.

For model 2, the ancestral population splits in two subpopulations of size \( s \) and \( 1 - s \) relative to the ancestral population and the subpopulations experience exponential growth. Here a direct likelihood comparison with the DaDi estimates given in Locke et al. (2011) is impossible. However, a rough comparison shows that the AML estimates for \( \tau \), \( N_B \) and \( N_S \) lie between \( \hat{\Theta}_1 \) and \( \hat{\Theta}_2 \) and for \( \mu_{BS} \) and \( \mu_{SB} \) they are of similar size.

### 4.3.4 Orang-utan data set

This real data is based on data sets of 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
Table 3: Properties of $\hat{\Theta}_{\text{AML}}$ in the IM model with short divergence time ($\tau = 695000$ years)

|                | true          | space                        | mean  | median | bias   | mean se | total se |
|----------------|---------------|------------------------------|-------|--------|--------|---------|----------|
| $\mu_{BS}$     | 1.44e-05      | (1.44e-06, 0.000144)         | 1.3e-05| 1.21e-05| -1.43e-06| 2e-06   | 4.48e-06 |
| $\mu_{SB}$     | 1.44e-05      | (1.44e-06, 0.000144)         | 1.43e-05| 1.39e-05| -7.2e-08| 2.19e-06| 3.77e-06 |
| $\tau$         | 695000        | (139000, 6950000)            | 1020000| 946000 | 330000 | 305000  | 389000   |
| $N_S$          | 17400         | (1740, 174000)               | 21700 | 21000  | 4370   | 3070    | 4100     |
| $N_B$          | 17400         | (1740, 174000)               | 22400 | 21800  | 5070   | 3160    | 4550     |

Figures and summaries: As in Tab. 2.
Table 4: Properties of $\hat{\Theta}_{AML}$ in the IM model with long divergence time ($\tau = 1390000$ years)

|          | true     | space               | mean    | median   | bias     | mean se | total se |
|----------|----------|---------------------|---------|----------|----------|---------|----------|
| $\mu_{BS}$ | 1.44e-05 | (1.44e-06, 0.000144) | 1.24e-05 | 1.12e-05 | -1.97e-06 | 3.38e-06 | 4.99e-06 |
| $\mu_{SB}$ | 1.44e-05 | (1.44e-06, 0.000144) | 1.25e-05 | 1.17e-05 | -1.88e-06 | 3.28e-06 | 4.38e-06 |
| $\tau$   | 1390000  | (139000, 6950000)    | 2360000 | 2280000  | 967000   | 876000  | 1e+06    |
| $N_S$    | 17400    | (1740, 174000)      | 25700   | 24600    | 8350     | 6620    | 7560     |
| $N_B$    | 17400    | (1740, 174000)      | 26000   | 25200    | 8670     | 6680    | 7870     |

Figures and summaries: As in Tab. 2.
Table 5: Parameter estimates for the ancestral history of orang-utans

|            | $\mu_{BS}$ | $\mu_{SB}$ | $\tau$ | $N_S$ | $N_B$ |
|------------|------------|------------|--------|-------|-------|
| $\hat{\Theta}_{AML}$ | $5.023\times 10^{-6}$ | $4.600\times 10^{-6}$ | $1300681$ | $52998$ | $21971$ |
| $\hat{\Theta}^*_{AML}$ | $4.277\times 10^{-6}$ | $3.806\times 10^{-6}$ | $1402715$ | $52715$ | $22233$ |
| $\hat{se}^*$ | $1.244\times 10^{-6}$ | $9.366\times 10^{-7}$ | $208391$ | $7223$ | $2779$ |
| $\hat{se}$ | $1.992\times 10^{-7}$ | $1.066\times 10^{-7}$ | $194868$ | $6083$ | $2963$ |
| lower       | $0$        | $1.633\times 10^{-7}$ | $715590$ | $31476$ | $13290$ |
| upper       | $6.627\times 10^{-6}$ | $4.931\times 10^{-6}$ | $1820852$ | $67118$ | $27426$ |

1This confidence interval was shifted to the positive range. The original value was $(-9.09\times 10^{-7}, 5.72\times 10^{-6})$.

$\hat{\Theta}_{AML}$: approximate maximum likelihood estimate; $\hat{\Theta}^*_{AML}$: bootstrap bias corrected estimate; $\hat{se}^*$, bootstrap standard error of $\hat{\Theta}_{AML}$; $\hat{se}$, 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, so the coverage probabilities are 99% in each dimension (Davison and Hinkley, 1997).

Table 6: Results from (Locke et al., 2011, Tab. S21-1) (same results for Model 2 reported in Ma et al. (2013)), scaled with $N_e = 17400$.

|            | $\mu_{BS}$ | $\mu_{SB}$ | $\tau$ | $N_S$ | $N_B$ |
|------------|------------|------------|--------|-------|-------|
| Model 1    | $9.085\times 10^{-7}$ | $7.853\times 10^{-7}$ | $6948778$ | $129889$ | $50934$ |
| Model 2    | $1.518\times 10^{-5}$ | $2.269\times 10^{-5}$ | $630931$ | $35976$ | $10093$ |
| $\hat{\Theta}_{AML}$ | $5.023\times 10^{-6}$ | $4.600\times 10^{-6}$ | $1300681$ | $52998$ | $21971$ |

Model 1: IM model in figure S4; 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.
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 (9,695 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 5 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 4-fold degenerate synonymous sites that are polymorphic in the orang-utan populations were selected.

5 Discussion

In this article, we propose an algorithm to approximate the maximum likelihood estimator in models with an intractable likelihood. Simulations and kernel density estimates of the likelihood are used to obtain the ascent directions in a stochastic approximation algorithm, so it is flexibly applicable to a wide variety of models.

Alternative simulation based approximate maximum likelihood methods have been proposed that estimate the likelihood surface from samples from the whole parameter space that are obtained in an ABC like fashion (Creel and Kristensen 2013; Rubio and Johansen 2013) or using MCMC (de Valpine 2004). The maximum likelihood estimator is obtained subsequently by standard numerical optimization. Conversely, our method only uses simulations and estimates of the likelihood along the trajectory of the stochastic approximation algorithm, thereby approximating the maximum likelihood estimator more efficiently. This is along the lines of Diggle and Gratton (1984) where a stochastic version of the Nelder-Mead algorithm is used, but through the use of summary statistics we can drop the restrictive i.i.d. assumption. 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. A thorough comparison of the properties of different approximate maximum likelihood methods is beyond the scope of this paper, but the three presented examples show that the algorithm provides fast and reliable approximations to the corresponding maximum likelihood estimators.

The population genetics application where we estimate parameters of the evolutionary history of orang-utans demonstrates 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]), 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 the normal distribution example, the 10-dimensional maximum likelihood estimate is approximated very precisely even though the search space spans 200 times the standard error of $\hat{\Theta}_{\text{ML}}$ in each dimension.

However, we also observe a bias for a few of the estimated parameters. Partly, for example for one of the parameters of the queuing process, this can be attributed to a 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], 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 longer runs of the algorithm.

As sufficiently informative summary statistics are crucial for the reliability of both maximum likelihood and Bayesian estimates, 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 approximate maximum likelihood algorithm and also its robustness in highly stochastic settings. With the implementation in the coalescent simulation program msms (Ewing and Hermisson 2010), its potential for population genetic applications can easily be explored further.
Acknowledgments

Johanna Bertl 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). 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.

References

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

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

M. A. Beaumont, J.-M. Cornuet, J.-M. Marin, and C. P. Robert. Adaptive approximate Bayesian computation. Biometrika, 2009.

G. Bertorelle, A. Benazzo, and S. Mona. ABC as a flexible framework to estimate demography over space and time: some cons, many pros. Molecular Ecology, 19(13):2609–2625, 2010.

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

M. G. B. Blum, M. A. Nunes, D. Prangle, and S. A. Sisson. A comparative review of dimension reduction methods in approximate Bayesian computation. Statistical Science, 28(2):189–208, 2013.

M. Creel and D. Kristensen. Indirect Likelihood Inference (revised). UFAE and IAE Working Papers 931.13, Unitat de Fonaments de l’Anàlisi Econòmica (UAB) and Institut d’Anàlisi Econòmica (CSIC), June 2013. URL http://ideas.repec.org/p/aub/autbar/931.13.html

K. Csilléry, M. Blum, O. E. Gaggiotti, and O. François. Approximate Bayesian computation (ABC) in practice. Trends in Ecology and Evolution, 25:410–418, 2010.

A. C. Davison and D. V. Hinkley. Bootstrap Methods and Their Applications. Cambridge University Press, Cambridge, 1997.

N. De Maio, C. Schlötterer, and C. Kosiol. Linking great apes genome evolution across time scales using polymorphism-aware phylogenetic models. Molecular Biology and Evolution, 30(10):2249–2262, 2013.
P. de Valpine. Monte carlo state-space likelihoods by weighted posterior kernel density estimation. *Journal of the American Statistical Association*, 99(466):523–536, 2004.

P. J. Diggle and R. J. Gratton. Monte Carlo methods of inference for implicit statistical models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 46(2):193–227, 1984.

C. C. Drovandi, A. N. Pettitt, and M. J. Faddy. Approximate Bayesian computation using indirect inference. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 60(3):317–337, 2011.

E. Ehrlich, A. Jasra, and N. Kantas. Gradient free parameter estimation for hidden markov models with intractable likelihoods. *Methodology and Computing in Applied Probability*, pages 1–35, 2013.

G. Ewing and J. Hermisson. MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. *Bioinformatics*, 26(16):2064–2065, 2010.

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

J.-D. Fermanian and B. Salanié. A nonparametric simulated maximum likelihood estimation method. *Econometric Theory*, 20(4):701–734, 2004.

R. C. Griffiths and S. Tavaré. Ancestral inference in population genetics. *Statistical Science*, 9(3):307–319, 1994.

D. Gross, J. J. Shortle, J. M. Thompson, and C. M. Harris. *Fundamentals of queueing theory*. Wiley, 4th edition, 2008.

R. N. Gutenkunst, R. D. Hernandez, S. H. Williamson, and C. D. Bustamante. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genetics*, 5(10), 2009.

W. Härdle, M. Müller, S. Sperlich, and A. Werwatz. *Nonparametric and semiparametric models*. Springer Series in Statistics. Springer, 2004.

K. Heggland and A. Frigessi. Estimating functions in indirect inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(2):447–462, 2004.

J. F. C. Kingman. The coalescent. *Stochastic Processes and their Applications*, 13:235–248, 1982a.

J. F. C. Kingman. Exchangeability and the evolution of large populations. In G. Koch and F. Spizzichino, editors, *Exchangeability in Probability and Statistics*, pages 97–112. North-Holland, Amsterdam, 1982b.
J. F. C. Kingman. On the genealogy of large populations. *Journal of Applied Probability*, 19:27–43, 1982c.

J. F. C. Kingman. The first Erlang century – and the next. *Queueing Systems*, 63:3–12, 2009.

D. P. Locke, 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, et al. Comparative and demographic analysis of orang-utan genomes. *Nature*, 469:529–533, 2011.

X. Ma, 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. Population genomic analysis reveals a rich speciation and demographic history of orang-utans (*Pongo pygmaeus* and *Pongo abelii*). *PLOS ONE*, 8, 2013.

P. Marjoram and S. Tavaré. Modern computational approaches for analysing molecular genetic variation data. *Nature Reviews Genetics*, 7:759–770, 2006.

T. McKinley, A. R. Cook, and R. Deardon. Inference in epidemic models without likelihoods. *The International Journal of Biostatistics*, 5, 2009.

L. N. Naduvilezhath, L. E. Rose, and D. Metzler. Jaatha: a fast composite-likelihood approach to estimate demographic parameters. *Molecular Ecology*, 20:2709–2723, 2011.

M. Nordborg. Coalescent theory. In D. J. Balding, M. Bishop, and C. Cannings, editors, *Handbook of Statistical Genetics*, volume 2, pages 843–877. John Wiley & Sons, third edition, 2007.

J. K. Pritchard, M. T. Scielstad, A. Perez-Lezaun, and M. W. Feldman. Population growth of human Y chromosomes: A study of Y chromosome microsatellites. *Molecular Biology and Evolution*, 16(12):1791–1798, 1999.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. URL [http://www.R-project.org/](http://www.R-project.org/).

F. J. Rubio and A. M. Johansen. A simple approach to maximum intractable likelihood estimation. *Electronic Journal of Statistics*, 7:1632–1654, 2013.

P. Sadegh. Constrained optimization via stochastic approximation with a simultaneous perturbation gradient approximation. *Automatica*, 33(5):889–892, 1997.

B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, 1986.

S. Soubeyrand, F. Carpentier, N. Desassis, and J. Chadoèuf. Inference with a contrast-based posterior distribution and application in spatial statistics. *Statistical Methodology*, 6(5):466–477, 2009.
J. C. Spall. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Transactions on Automatic Control*, 37 (3):352–355, 1992.

J. C. Spall. *Introduction to Stochastic Search and Optimization: Estimation, Simulation and Control*. Wiley, 2003.

M. Stephens. Inference under the coalescent. In D. J. Balding, M. Bishop, and C. Cannings, editors, *Handbook of Statistical Genetics*, volume 2, pages 878–908. John Wiley & Sons, third edition, 2007.

A. Tellier, P. Pfaffelhuber, B. Haubold, L. Naduvilezhath, L. E. Rose, T. Städler, W. Stephan, and D. Metzler. Estimating parameters of speciation models based on refined summaries of the joint site-frequency spectrum. *PLOS ONE*, 6(5), 05 2011.

T. Toni, D. Welch, N. Strelkowa, A. Ipsen, and M. P. H. Stumpf. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface*, 6:187–202, 2009.

J. Wakeley. *Coalescent Theory*. Roberts & Company Publishers, 2008.

M. P. Wand and M. C. Jones. *Kernel Smoothing*. Chapman & Hall, 1995.

D. Wegmann, C. Leuenberger, and L. Excoffier. Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood. *Genetics*, 182(4):1207–1218, 2009.

G. Weiss and A. von Haeseler. Inference of population history using a likelihood approach. *Genetics*, 149(3):1539–1546, 1998.

S. Wellek. *Testing Statistical Hypotheses of Equivalence and Noninferiority*. CRC Press, Taylor & Francis, 2010.

Y. Wu. Exact computation of coalescent likelihood for panmictic and subdivided populations under the infinite sites model. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 7(4):611–618, 2010.