Likelihood-free Bayesian inference without rejection

Michael GB Blum*,
Olivier François†

*Centre National de la Recherche Scientifique, †Institut National Polytechnique de Grenoble, TIMC-IMAG, Faculty of Medicine of Grenoble, 38706 La Tronche France.

Corresponding author: Michael Blum.
Address: TIMC-IMAG, Faculté de Médecine de Grenoble, 38706 La Tronche, France.
Tel: +33 456 520 065. Fax: +33 456 520 055.
E.mail: michael.blum@imag.fr
Abstract

Approximate Bayesian inference on the basis of summary statistics is well-suited to complex problems for which the likelihood is either mathematically or computationally intractable. However the methods that use rejection suffer from the curse of dimensionality when the number of summary statistics is increased. Here we propose a machine-learning approach to the estimation of the posterior density by introducing two innovations. The new method fits a nonlinear conditional heteroscedastic regression of the parameter on the summary statistics, and then adaptively improves estimation using importance sampling. The new algorithm is compared to the state-of-the-art approximate Bayesian methods, and achieves considerable reduction of the computational burden in two examples of inference in statistical genetics and in a queueing model.

Keywords : Approximate Bayesian computation | Ancestral inference | Conditional density estimation | Implicit statistical models | Importance sampling
1 Introduction

Making use of simulations to perform non-Bayesian inference in models for which the likelihood is neither analytically solvable nor computationally tractable has a well-established methodology that finds its roots at least in the seminal papers of Diggle and Gratton (1984) and Gourieroux et al. (1993). This approach bypass explicit likelihood functions by fitting an *implicit statistical model* – that is, a model defined in term of a stochastic generating mechanism – to the observed data.

In the Bayesian setting, there has been a growing interest in implicit statistical models for demographic inference in population genetics (Marjoram and Tavaré 2006). Statistical inference with population-genetic data usually requires prior knowledge on genealogical trees. As the tree is usually considered as a nuisance parameter, Bayesian Monte Carlo is a natural approach to average over high-dimensional tree space. Although many likelihood-based methods have been proposed in this framework, these methods are restricted to particular demographic and genetic processes (Stephens and Donnelly 2000; Wilson et al. 2003; Kuhner 2006; Hey and Nielsen 2007; Fearnhead 2008). Alternatively, likelihood-free methods, named approximate Bayesian computation (ABC) after Beaumont et al. (2002), have gained many advocates in the recent years. The principle of ABC relies on the simulation of large numbers of data sets using parameters drawn from the prior distribution. A set of summary statistics is then calculated for each simulated sample, and compared with the values for the observed sample. Parameters that have generated summary statistics close enough to the observed data
are retained to form an approximate sample from the posterior distribution.

Approximate Bayesian estimation algorithms – that were originally all based on rejection algorithms – can be classified into three broad categories, resembling the mainstream methods that are applied in standard computational Bayesian statistics (Gelman et al. 2003). The first class of methods relies on the direct rejection algorithm as described in the previous paragraph (Tavaré et al. 1997; Pritchard et al. 1999). The second class of algorithms mimics Markov chain Monte Carlo methods (MCMC, Robert and Casella 2004), embedding simulations from the implicit model in the updating step of the stochastic algorithm (Marjoram et al. 2003). The MCMC-ABC algorithm modifies the Metropolis-Hastings acceptance rule in order to incorporate the distance to the observed summary statistics. The third class of algorithms shares similarity with the recently introduced sequential Monte Carlo samplers (SMC, Liu 2001). The main SMC-ABC algorithm combines ideas underlying rejection methods and sequential importance sampling (Sisson et al. 2007).

Nevertheless, a severe limitation of rejection-based generative algorithms arises when the dimensionality of the set of summary statistics increases. Because the three classes of methods attempt to sample from a small multidimensional sphere around the observed summary statistics, all of them suffer from the curse of dimensionality. To overcome this problem, Beaumont et al. (2002) allowed larger acceptance rates in the rejection algorithm, ranging up to 20 percent of the simulated values, and then performed local linear adjustment in order to correct for the discrepancy between the simulated
and the observed statistics. Here, we address the curse of dimensionality issue by adopting a machine learning perspective constructing a functional relationship between the generated set of summary statistics and the model parameters. Assuming perfect construction, this relationship could ideally be utilized to produce samples from the posterior distribution by exploiting information not restricted to a subset of generated values, but to the extended set.

At a first stage, our approach infers the functional relationship linking the summary statistics to the model parameters by considering a flexible nonlinear conditional heteroscedastic (NCH) model. Flexible regression models, like neural networks, are exploited to reduce dimension and to better account for the correlation within the set of summary statistics. At a second stage, we introduce an adaptive version of the NCH (ANCH) algorithm using importance sampling. The rationale of the adaptive algorithm is to iteratively limit the discrepancy between the sampling distribution and the posterior distribution, which may be particularly useful when the prior distribution is vague. In two historical examples of demographic inference in population genetics and in an example of a queueing process, we provide evidence that the NCH and the ANCH algorithms reduce the computational burden when compared to the state-of-the-art ABC methods.

2 Method

Standard rejection. In ABC, we assume that there is a multidimensional parameter of interest \( \phi \), and the observed value \( s \) of a set of summary statis-
tics $S$ is calculated for the data. To make statistical inference, a naive rejection-sampling method generates random draws $(\phi_i, s_i)$ where $\phi_i$ is sampled from the prior distribution, and $s_i$ is measured from synthetic data, simulated from a generative model with parameter $\phi_i$. Fixing the tolerance error to the value $\delta$ and denoting by $\|\cdot\|$ the Euclidean norm, only parameters $\phi_i$ such that $\|s - s_i\| \leq \delta$ are retained. Because the summary statistics may span different scales, norms that use re-scaled distances are often considered in place of the Euclidean distance. In our application of the ABC, we re-scale distances by the mean absolute deviation of the simulated summary statistics. The accepted $\phi_i$ then form a random sample from the approximate posterior distribution defined as

$$p_\delta(\phi|s) \propto \Pr(\|s - s_i\| \leq \delta|\phi) p(\phi)$$

where $p(\phi)$ denotes the prior distribution. Compared to the exact expression of the posterior distribution, the likelihood is replaced by

$$p(s|\phi) \approx \Pr(\|s - s_i\| \leq \delta|\phi).$$

If the summary statistics are sufficient for the parameter $\phi$, the approximate posterior distribution converges to the posterior distribution as $\delta$ goes to 0. In addition, the approximate posterior distribution corresponds to the prior distribution when $\delta$ is large.

Beaumont et al. (2002) introduced a first improvement of the standard rejection ABC algorithm in which the parameters $\phi_i$ were weighted by the values $K_\delta(\|s_i - s\|)$, where $K_\delta$ is the Epanechnikov kernel. Using that weighting scheme, an estimator of the posterior mean was then given
by $\sum_i K_\delta(||s_i - s||)\phi_i / \sum_i K_\delta(||s_i - s||)$. Although this was not originally stated, it can be seen that it corresponds to the Nadaraya-Watson estimator, a classic approach to nonparametric regression (Nadaraya, 1964; Watson 1964).

To avoid the curse of dimensionality, Beaumont et al. (2002) also described the posterior density as a homoscedastic linear regression model (in fact, a local-linear model) of the form

$$\phi_i = \alpha + (s_i - s)^T \beta + \zeta_i, \quad i = 1, \ldots, M,$$

where $\alpha$ is an intercept, $\beta$ is a vector of regression coefficients, and the $\zeta_i$'s are independent random variates with mean zero and common variance. We further refer to this algorithm as the LocL ABC model. In the LocL model, the observations are weighted by $K_\delta(||s_i - s||)$, and $(\alpha, \beta)$ are inferred by the weighted least-square estimates $(\hat{\alpha}, \hat{\beta})$. If equation (1) exactly describes the relationship between $\phi$ and $s$, random draws of the posterior distribution can simply be obtained as $\alpha + \zeta_i$, for $i = 1, \ldots, M$. Using the empirical residuals in place of the $\zeta_i$'s, the parameters are adjusted as

$$\phi_i^* = \phi_i - (s_i - s)^T \hat{\beta}$$

to provide an approximate sample from the posterior distribution. In this approach, the choice of $\delta$ involves a bias-variance trade-off: Increasing $\delta$ reduces variance thanks to a larger sample size for fitting the regression, but also increases bias arising from departures from linearity and homoscedasticity.
A nonlinear conditional heteroscedastic model. In this study, we introduce an important modification on the previously described adjustment-based ABC method for conditional density estimation. In order to minimize departures from linearity and homoscedasticity, we propose to model both the location and the scale of the response parameter, \( \phi_i \), in equation (1). The new regression model takes the form of a nonlinear conditional heteroscedastic (NCH) model

\[
\phi_i = m(s_i) + \sigma(s_i) \times \zeta_i, \quad i = 1, \ldots, M, \tag{3}
\]

where \( m(s_i) \) denotes the conditional expectation, \( E[\phi|S = s_i] \), and \( \sigma^2(s_i) \) denotes the conditional variance, \( \text{Var}[\phi|S = s_i] \).

The conditional expectation can be estimated as \( \hat{m}(s_i) \) by adjusting a flexible non-linear regression model using a least-square method. The variance term is then estimated using a second regression model for the log-residuals

\[
\log(\phi_i - \hat{m}(s_i))^2 = \log \sigma^2(s_i) + \xi_i, \quad i = 1, \ldots, M, \tag{4}
\]

where the \( \xi_i \)'s are independent random variates with mean zero and common variance. In this study, we consider feed-forward neural network models for fitting both \( m(s_i) \) and \( \log \sigma^2(s_i) \) (Ripley 1996; Nix and Weigend 1995; Bishop 2006). This choice is motivated by the fact that feed-forward neural networks include the possibility to reduce the dimensionality of the set of summary statistics via internal projections on lower dimensional subspaces.

Similarly to equation (2), parameter adjustment under the NCH model can be performed as follows

\[
\phi_i^* = \hat{m} + (\phi_i - \hat{m}) \times \frac{\hat{\sigma}(s_i)}{\sigma(s_i)}, \quad i = 1, \ldots, M. \tag{5}
\]
Note that \((\phi_i^*)\) forms a random sample from the true posterior distribution if the distribution of \(\phi\) given \(s\) belongs to the same location-scale family for all \(s\), and if the sample is large enough so that \(\hat{m}\) could be considered equal to \(m\) and \(\hat{\sigma}\) equal to \(\sigma\).

To enable comparisons with the LocL ABC method, a tolerance error, \(\delta\), is allowed, and the simulated parameters, \(\phi_i\), are weighted by \(K_\delta(||s_i - s||)\). Furthermore, to warrant that the adjusted parameters, \(\phi^*_i\), obtained from equation (2) or (5) fall in the support of the prior distribution, we sometimes consider transformations of the original responses. Parameters that lie in an interval are transformed via the logit function, and nonnegative parameters are transformed using a logarithm. For local-linear regression, the logarithmic transformation has the further potential advantage of making the homoscedastic hypothesis more likely (Box and Cox 1964).

**Iterated importance sampling.** ABC methods usually concentrate their computational efforts on a single run. A second change to the ABC algorithm converts the single-stage regression based ABC method into a multi-stage algorithm resembling SMC algorithms, in which estimations are improved iteratively (Liu 2001; Sisson et al. 2007). In practice, we implemented a two-stage algorithm. The extension of the algorithm to more than one stage of adaptation is straightforward. The logic of using a two-stage algorithm is that the second run can control a first tolerant run, and adaptively builds a better approximation of the posterior distribution. If the two empirical distributions obtained after each step of the ANCH algorithm agree, then
the results can be pooled to form a larger approximate sample from the
posterior distribution. The adaptive step uses a sampling scheme in which
the parameter values $\phi_i$ are sampled from an importance distribution, $q(\phi)$,
rather than the prior distribution, $p(\phi)$. To compensate for the fact that we
do not sample from the prior distribution, each value $\phi_i$ is then weighted by
$p(\phi_i)/q(\phi_i)$ in sampling from the posterior distribution.

The adaptive NCH (ANCH) algorithm can be described as follows: Starting
from a sample $(\phi_1^i)$ obtained from a first NCH ABC run, the adaptive
step of the algorithm consists of estimating the support $\Delta_1$ of the sample.
For multi-dimensional parameters, the support of a multi-dimensional dis-
tribution can be estimated using support vector machines (SVM, Schölkopf
et al. 2001). Then new parameters are proposed from the conditional prior
distribution given that they fall in $\Delta_1$. Using this new set of parameters,
a second sample, $(\phi_2^i)$, can be formed using the NCH method again. A
noticeable point is that the importance weights need not to be effectively
computed because we have $q(\phi_2^i) = p(\phi_2^i)/p(\Delta_1)$, which means that the ratio
$p(\phi_2^i)/q(\phi_2^i)$ does not depend on $i$.

Examples of implicit statistical models

In this section, we present three examples of implicit statistical models, two
of which have received considerable attention in population genetics, and
the last one has served to illustrate indirect inference (Gourieroux et al.
1993; Heggland and Frigessi 2004). Using these examples, we performed an
empirical evaluation of the relative performance of three regression-based ap-
proximate Bayesian algorithms, the local linear regression model (LocL ABC model), the non-linear conditional heteroscedastic model (NCH model), and its adaptive implementation (ANCH model). We used the R programming language to implement the LocL, NCH and ANCH algorithms. Least-square adjustment for neural networks was implemented using the R package nnet (R Core Team 2007). Model choice for neural networks was based on a Bayesian (or regularization) approach (Ripley 1996). We used 4 hidden units and the weight-decay regularization parameter was set to equal to $\lambda = 0.001$.

At this stage, cross-validation might be a useful alternative approach to neural network model choice, but the previous values proved to work well in the examples considered here. In the ANCH algorithm, the support of the conditional density was estimated using a SVM $\nu$-regression algorithm ($\nu = 0.005$) as implemented in the R package e1071 based on the public library libsvm (Chang and Lin 2001).

**Examples in population genetics.** There has been tremendous interest in simulation-based inference methods in evolutionary biology during the last decade (Fu and Li 1997; Pritchard et al. 1999; Fagundes et al. 2007). In these applications, the inference of demographic and genetic parameters depends on the so-called coalescent approximation which describes, in a probabilistic fashion, the ancestry of genes represented in a sample. Coalescent models provide good examples of implicit statistical models for which a straightforward stochastic generating mechanism exists, but the likelihood is usually computationally intractable.
Example 1. Given a set of $n$ DNA sequences, the first problem concerns the estimation of the effective mutation rate, $\theta > 0$, under the infinitely-many-sites model. In this model, mutations occur at rate $\theta$ at DNA sites that have not been hit by mutation before. If a site is affected by mutation, it is said to be segregating in the sample. In this example, the summary statistic, $s$, is computed as the number of segregating sites. Note that $s$ is not a sufficient statistic (Fu and Li 1993). The generating mechanism for $s$ can be described as follows.

1. Simulate $L_n$, the length of the genealogical tree of the $n$ sequences, as the sum of independent exponential random variables of rate $(j - 1)/2$, $j = 2, \ldots, n$.

2. Generate $s$ according to a Poisson distribution of mean $\theta L_n/2$.

The simulation of $L_n$ can be derived from the formula for the total length in a tree, $L_n = \sum_{j=2}^{n} jY_j$, where the $Y_j$’s, the inter-coalescence times, correspond to the times during which the sample has $j$ ancestors, $j = 2, \ldots, n$ (Tavaré et al. 1997). In a wide range of models in population genetics, the inter-coalescence times form independent random variables distributed according to the exponential distributions of rate $j(j - 1)/2$, $j = 2, \ldots, n$. A more detailed description of the coalescent process can be found in (Tavaré 2004).

We computed the posterior distribution of the effective mutation rate $\theta$ given the observation of $s = 10$ segregating sites in a sample of $n = 100$ DNA sequences. The prior distribution for the parameter $\theta$ was taken to be the exponential distribution of mean 50, which was meant to represent a
vague prior. An exact sample from the posterior distribution was obtained
using a direct rejection algorithm accepting only parameters that produced 10
segregating sites exactly. Ten millions of replicates were generated resulting
in a sample of size 39,059 after rejection.

For the ABC algorithms, we performed inference of the posterior distribu-
tion using a total of 2,000 replicates of the summary statistic. We recorded
the three quartiles and the 0.025 and 0.975 quantiles of the approximate
posterior distributions computed by the three algorithms, and we compared
these 5 quantiles \( Q_k, k=1,\ldots,5 \), with the corresponding empirical quantiles,
\( Q^0_k \), obtained from the exact sample. For values of the tolerance rate be-
tween 0 and 1 and for each quantile, \( Q_k \), the accuracy of each algorithm was
assessed by the relative median absolute error

\[
\text{RMAE} = \text{median} \left| \frac{Q_k - Q^0_k}{Q^0_k} \right|
\]

computed over 150 runs. In addition, we measured the discrepancy between
each approximate distribution and the empirical posterior distribution using
the sum of RMAE over all quantiles. In the ANCH algorithm, the support
was estimated as the range of the empirical distribution – i.e., the \((0, \max)\)
interval – obtained after a first run using 1,000 replicates (tolerance rate
\( P_\delta = 75\% \)).

Comparisons with standard rejection algorithms were first conducted.
We found that the posterior distribution obtained from the rejection meth-
ods deviated from the empirical posterior distribution significantly for tol-
erance rates larger than 10\% (Figure 1A). The LocL model approximated
the posterior distribution accurately for small tolerance rates (\( \leq 20\% \)), but
the performances of the LocL method deteriorated as the tolerance rate increased (Figure 1, green curves). The performances of the NCH model were significantly less sensitive to the tolerance rate, staying at values close to the optimum achieved by the LocL model (Figure 1B, black curve). The adaptive NCH algorithm achieved even superior performances for values of the tolerance rate ranging between 0 and 90% (Figure 1B, red curve). The black dot in Figure 1B represents the performance of the ANCH algorithm without weighting and allowing total acceptance. Having eliminated the concept of rejection in the approximate Bayesian algorithm, i.e. setting $P_\delta = 1$, the accuracy of the algorithm remained close to the optimum achieved by all algorithms. This first example illustrates the benefit of the NCH model over the LocL model and the other rejection methods. The additional gain of the adaptive step stems from the use of a vague prior, which, in this case, gave low weight to the region of posterior values.

**Example 2.** Turning to a more complex problem in which the posterior distribution could not be estimated easily, we considered an exponentially growing population model with 3 parameters. Similarly to Weiss and von Haeseler (1998), a population of size $N_A$ started to grow exponentially $t_0$ years ago to reach a present size of $N$ individuals where $N = N_A/\alpha$ for a value $\alpha \in (0, 1)$. We performed inference on the two parameters $N_A$, $t_0$ and $\alpha$ was considered as a nuisance parameter. In this example the data consisted of a sample of $n$ individuals genotyped at a multilocus subset of independent microsatellite markers (see e.g. Zhivotovsky et al. 2003). Microsatellite
loci are characterized by a motif of two to four nucleotides that may repeat itself several times, and the data are recorded as number of repeats for each individual.

The generating mechanism for the implicit model can be described as follows.

1. Simulate candidate coalescent genealogies in a growing population for each marker,

2. Superimpose mutations on the tree branches according to a specific mutation model.

Step 1 requires simulating coalescence times in a coalescent model with varying population size (see Tavaré 2004). In step 2, we used the single-step mutation model, that can be viewed as a simple random walk for which the +1 and -1 steps are equally likely (Ohta and Kimura 1973).

To capture the pattern of genetic variation, we computed six summary statistics previously reported to be sensitive to the genetic diversity in the sample and to the intensity of the demographic expansion. The amount of genetic diversity was measured by the mean (over the loci) of the variance in the number of repeats and by the mean of their heterozygosities (Pritchard and Feldmann 1996). For the demographic pattern, we used two imbalance indices studied by King et al. (2000), the interlocus statistic introduced by Reich and Goldstein (1998), and the expansion index of Zhivotovsky et al. (2000). We also computed a seventh summary statistic based on an observation of Shriver et al. (1997) who studied the distribution, $P_K$, of
pairwise comparisons that differ by $K$ repeat units. This distribution has its peak at the value 0 for a recent expansion, and the peak shifts to the value 1 for older expansions. To compute the seventh statistic, we averaged the quantity $P_1 - P_0$ over all the loci.

We took a uniform prior distribution ranging from 0 to 100,000 years for the onset of the expansion, a uniform distribution over the interval (0, 10,000) for the ancestral population size, and a uniform distribution over the interval (1, 6) for $-\log_{10}(\alpha)$.

One hundred test data sets were generated using $t_0 = 18,000$ years for the date of onset of the expansion, $N_A = 1,500$ for the ancestral population size, and $\alpha = 0.0012$ ($\log_{10}(\alpha) = -2.92$) for the ratio of the ancestral size to the present population size. These values were very similar to those used in Pritchard et al. (1999) in a study of the Y chromosome in humans. For each algorithm, we computed the posterior distribution of the 3 parameters given the observation of the 7 summary statistics in a sample of $n = 100$ individuals surveyed at 50 microsatellite loci.

For the three algorithms, we generated samples from the posterior distribution using 2,000 replicates from the implicit model. We recorded the quartiles and the 0.025 and 0.975 quantiles of the output distributions for the 3 approximate Bayesian algorithms, and we compared these 5 values for the conditional distributions of each parameter $t_0$, $N_A$, and $\alpha$. The median value of each quantile was then computed over 100 runs.

For each of the three parameters, the median estimates of the quantiles of the marginal posterior distribution were very similar in the NCH model
and in the ANCH implementation (Figure 2, black and red curves). For large tolerance rates ($P_δ \geq 50\%$), we observed a strong agreement with the values estimated by the LocL model (Figure 2, green curves) used with small tolerance rates ($P_δ = 5\%$), indicating that the NCH model can efficiently exploit simulation results that fall far apart from the observed values of the summary statistics. The performances of the LocL model decreased as the tolerance rate increased above 20\%, and the estimation of the conditional distribution of the ancestral size provided evidence that the bias was substantial (Figure 2, bottom right panel). The three algorithms gave similar results regarding the estimation of $\alpha$ (results not shown).

To further compare the performances of the NCH model and its ANCH variant, we studied a particular simulated data set corresponding to the same ground truth as before. After running the three algorithms 100 times for each tolerance rate, the variance of each quantile in the posterior distribution was of order three times higher in the NCH model than in the ANCH algorithm (2,000 simulations for each algorithm, Table 1). Clear evidence of the stabilization phenomenon was also obtained under the infinitely-many-sites model. Given $s = 10$ segregating sites, we ran the ANCH algorithm using 200 simulations of the implicit model at each step (tolerance rate $P_δ = 85\%$). To compare estimations obtained after the initial step with those obtained after adapting the support, we replicated the estimation procedure 100 times. The reduction in variance ranged from a factor 2.7 to a factor 34.7. The highest reduction in variance was obtained for the estimation of the upper quantile.
Example 3. Our third example arose from a totally different context, and was formerly studied by Heggland and Frigessi (2004) using indirect inference. The connection of indirect inference to ABC is the following. Indirect inference is a non-Bayesian method that proceeds with 3 steps. 1) an auxiliary model is introduced, usually as a simplified version of the true model. 2) Estimates of the parameters in the auxiliary model are obtained and play the role of summary statistics. These estimates can be obtained, for example, by maximizing the likelihood in the auxiliary model. 3) An estimate of the parameter $\phi$ is built by minimizing a weighted Euclidean distance between simulated summary statistics and the observed summary statistics. Note that the introduction of an auxiliary model can also be a useful mean to find informative summary statistics for ABC methods.

The model considered in (Heggland and Frigessi 2004) was a queueing system with a first-come-first-serve single-server queue (G/G/1). The service times were uniformly distributed in the interval $[\theta_1, \theta_2]$, and the inter-arrival times had exponential distribution with rate $\theta_3$. Let $W_n$ be the inter-arrival time of the $n$th customer and $U_n$ be the corresponding service time. The process of inter-departure times $\{Y_n, n = 1, 2, \ldots\}$ can be described by the following generative algorithm

$$Y_n = \begin{cases} U_n, & \text{if } \sum_{i=1}^{n} W_i \leq \sum_{i=1}^{n-1} Y_i, \\ U_n + \sum_{i=1}^{n} W_i - \sum_{i=1}^{n-1} Y_i, & \text{otherwise.} \end{cases}$$

Bayesian inference on $(\theta_1, \theta_2, \theta_3)$ was done by assuming that only the inter-departure times were observed. Because the inter-arrival times were unobserved, likelihood-based inference would involve high-dimensional integra-
tion, or using missing data algorithms.

We generated a test data set with $n = 50$ successive inter-departure time observation using $\theta_1 = 1$, $\theta_2 = 5$, and $\theta_3 = 0.2$. We set a uniform prior over $[0, 10]$ for $\theta_1$, $\theta_2 - \theta_1$, and for $\theta_3$. To investigate the sensitivity of the NCH model to the number of summary statistics, we ran the ABC-NCH algorithm using 5, 10 and 20 summary statistics. Here, the set of summary statistics included the minimum and the maximum of the inter-departure times and the 3, 8 and 18 equidistant quantiles of the inter-departure times. We used 10,000 replicates and the tolerance rate was set to the value $P_\delta = 50\%$ resulting in a posterior sample of size 5,000.

Figure 3 shows that posterior distributions of the 3 parameters had their mode and median values close to the ground truth values whatever the number of summary statistics. This provided evidence that the NCH ABC algorithm was robust to an increase in the dimensionality of the set of summary statistics. In addition, we found that the posterior distributions were more concentrated around the true values when 20 summary statistics were used.

To investigate the variability of the ABC algorithms from one run to the other, we ran the LocL, NCH and ANCH algorithms 100 times on the same data set. We used 2,000 replicates (a rather small number) in order to observed an exaggerated variability, and we varied the tolerance rate from 0 to 1.

Figure 4 displays estimated posterior quantiles for $\theta_3$. The LocL ABC algorithm was the less variable algorithm, but the posterior credibility intervals produced by this method were wider than those produced by the NCH
and ANCH algorithms. The latter were less sensitive to the variation of the
tolerance rates, and the medians of the posterior distributions were closer to
the true value in the non-linear models than in the LocL model. Compared
to the NCH algorithm, the ANCH algorithm reduced the variance of the
quantile estimates. We suspect that the variance of the quantile estimates
came from the use of local optimization during the learning phase of the
feed-forward neural networks. Averaging the quantile estimates using multiple runs of the NCH or the ANCH algorithms is therefore likely to provide better results than relying on a single run on the replicates. For $\theta_1$ and $\theta_2$, the LocL ABC and the NCH ABC algorithms led to similar approximate posterior distributions when the tolerance rate was set at values close to zero in the LocL ABC algorithm.

3 Discussion

Approximate Bayesian computation encompasses a wide range of useful methods for making inference in implicit statistical models. In this context rejection algorithms have greatly benefit from ideas coming from regression-based conditional density estimation. So far conditional density estimation in ABC approaches has relied on linear adjustment exclusively (Beaumont et al. 2002). While the linear regression-based ABC method can approximate posterior distribution accurately, this is usually achieved at the expense of a heavy computational load. For example, using the LocL ABC method for estimating parameter in complex models of modern human expansion, Fagundes et al. (2007) required an amount of computational time equivalent
to 10 CPU-months. To increase the tolerance of the algorithm, we have proposed to use non-linear regression-based ABC. In three examples, non-linear neural networks proved to be able to reduce computational generation costs significantly.

A heuristic reason why neural networks worked well when the number of summary statistics was large is that their first layer allows a nonlinear projection onto a subspace of much lower dimensionality, and nonlinear regression can then be performed using the reduced number of projection variables. In addition Bayesian neural network theory includes general rules for choosing appropriate regularization parameters (Ripley 1996) which makes the method rather automatic. Compared to other regression models, neural networks share many properties of projection pursuit regression (Friedman and Stuelze 1981), which may then lead to equivalent performances. As well SVM have gained increased popularity in machine learning approaches during the recent years (Vapnik 1998), and the algorithms described here could be modified to include SVM regression without change in spirit.

A second justification for using feed-forward neural networks is their ability to implement probabilistic outputs, hence allowing a unified Bayesian treatment of model choice. Indeed model choice may be performed by considering the model itself as an additional parameter to infer. Beaumont (2007) proposed to estimate the posterior probability of each candidate model by an approach based on a weighted multinomial logistic regression procedure. This approach is an extension of logistic regression to more than two categories, and it is equivalent to the use of a multinomial log-linear model. As
they pertain to a more flexible class of models, neural networks may achieve
equal or better predictive values than multinomial logistic regression (Ripley
1996).

The ABC approach has been recently used in connected domains like
human population genetics (Pritchard et al. 1999, Fagundes et al. 2007),
epidemiology (Tanaka et al. 2006) or for the evolution of protein networks
(Ratman et al. 2007). It has also recently been applied to the analysis
of compositional data (Butler and Glasbey 2008). Although inference from
synthetic data that mimic observations has a long lasting record in frequentist
statistics (Diggle and Gratton 1984, Gourieroux et al. 1993), ABC is still
in its infancy. Because the ABC method combines the power of simulating
from stochastic individual-based models with sound methodological grounds
from Bayesian theory, it has the potential to open doors to inference in many
complex models in ecology, evolution, and epidemiology, or other domains
like the social science. Improved statistical ABC models, like those presented
in this study, will then be useful to deal with increased model complexity,
and with the need to raise the dimension of the vector of summary statistics.
References


767

Figure 1. Relative median absolute error (RMAE) when estimating the quantiles of the posterior distribution as a function of the tolerance rate. For the quartiles and for the 0.025 and 0.975 quantiles, relative errors between approximate quantiles computed by the ABC methods and empirical values from the posterior distribution were computed over the 150 replicates. The sum of the RMAE’s was obtained by summing, over the 5 quantiles, the values of the relative median absolute differences. The black dot corresponds to the rejection-free ANCH algorithm (tolerance rate $P_\delta = 100\%$) without including the Epanechnikov weights.

Figure 2. The posterior quantiles of the date of onset of expansion and the ancestral population size for the NCH and the ANCH methods (left) and LocL ABC method (right). The quantiles are plotted against the tolerance rate for the NCH and the ANCH methods (left) and LocL ABC method (right). For each algorithm, the curves represent the 0.025, 0.25, 0.5, 0.75, and 0.975 quantiles of the posterior distribution. These values correspond to the median over the 100 test data sets. The ground truth values are represented as the blue lines. The values of the quantiles estimated by the NCH and ANCH methods using a tolerance rate $P_\delta = 75\%$ match with those obtained from the LocL algorithm using $P_\delta = 5\%$.

Figure 3. The posterior quantiles of $\theta_1$, $\theta_2$, and $\theta_3$, using the NCH ABC
method, with 5, 10, and 20 summary statistics. The vertical lines correspond to the true values of the parameters that were used when simulating the data set. The tolerance rate was set at 50% and a total of 10,000 simulations were performed.

Figure 4. The boxplot of the posterior quantiles for $\theta_3$ as a function of the tolerance rate. The different ABC methods were run 100 times each using 2,000 simulations at each run.
Figure 1:
Figure 2:
Figure 3:
Figure 4:
Table 1.  The benefit of adapting the support. Infinitely-many-sites: Ratios of variances for the quantiles estimated after the first NCH step and the second NCH step of the ANCH algorithm (acceptance rate $P_\delta = 85\%$ in each step). Expansion model: Ratios of variances for the quantiles estimated by the NCH and the ANCH algorithm (2,000 simulations in each algorithm, $P_\delta = 75\%$). P-values were computed according to the F-test.
### Posterior distribution quantiles

<table>
<thead>
<tr>
<th>Model parameters</th>
<th>2.5%</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
<th>97.5%</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>INFINITELY-MANY-SITES</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mutation rate $\theta$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Var. ratio</td>
<td>2.75</td>
<td>3.16</td>
<td>3.37</td>
<td>5.46</td>
<td>34.76</td>
</tr>
<tr>
<td>$P$-values</td>
<td>1.27e-06</td>
<td>1.67e-08</td>
<td>1.76e-09</td>
<td>6.88e-15</td>
<td>0</td>
</tr>
<tr>
<td><strong>EXPANSION MODEL</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Onset $t_0$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Var. ratio</td>
<td>1.84</td>
<td>2.95</td>
<td>3.69</td>
<td>3.53</td>
<td>2.97</td>
</tr>
<tr>
<td>$P$-values</td>
<td>1.28e-03</td>
<td>8.12e-08</td>
<td>1.74e-10</td>
<td>6.16e-10</td>
<td>6.12e-08</td>
</tr>
<tr>
<td>Ancestral pop. size $N_A$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Var. ratio</td>
<td>1.65</td>
<td>3.23</td>
<td>6.83</td>
<td>4.58</td>
<td>3.20</td>
</tr>
<tr>
<td>$P$-values</td>
<td>6.20e-03</td>
<td>7.15e-09</td>
<td>0</td>
<td>2.3e-13</td>
<td>9.34e-09</td>
</tr>
<tr>
<td>Ratio of pop. sizes $\alpha$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Var. ratio</td>
<td>0.09</td>
<td>0.87</td>
<td>3.83</td>
<td>1.97</td>
<td>7.66</td>
</tr>
<tr>
<td>$P$-values</td>
<td>1</td>
<td>0.75</td>
<td>5.89e-11</td>
<td>4.24e-04</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 1: