Non-linear regression models for Approximate Bayesian Computation

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

Keywords Likelihood-free inference, curse of dimensionality, feed forward neural networks, heteroscedasticity, coalescent models.

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 | Conditional density estimation | Implicit statistical models | Importance sampling | non-linear regression | indirect inference

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 bypasses explicit likelihood functions by simulating from an *implicit statistical model* – that is, a model defined in term of a stochastic generating mechanism.

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). 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 standard ABC (Tavaré et al. 1997; Pritchard et al. 1999; Beaumont et al. 2002) 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; Bortot et al. 2007). The MCMC-ABC algorithm takes into account the distance between the simulated and the observed summary statistics into the Metropolis-Hastings acceptance rule. 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; Beaumont et al. 2008; Toni et al. 2009).

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 (see e.g. Härdle et al. 2004). 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.

In the 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. In the 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 ABC with rejection and to ABC with local linear adjustment.

2 Method

Rejection and weighting. In ABC, we assume that there is a multidimensional parameter of interest ϕ, and the observed value $\mathbf{s}$ of a set of summary statistics $\mathbf{S}$ is calculated for the data. To make statistical inference, a rejection-sampling method generates random draws $(\phi_i, \mathbf{s}_i)$ where ϕ_i is sampled from the prior distribution, and $\mathbf{s}_i$ is measured from synthetic data, simulated from a generative model with parameter ϕ_i. Setting the tolerance error to a value δ and denoting by $\|.\|$ the Euclidean norm, only parameters ϕ_i such that $\|\mathbf{s} - \mathbf{s}_i\| \leq \delta$ are retained. Because the summary statistics may span different scales, norms that use rescaled distances are often considered in place of the Euclidean distance. In our application of ABC, we rescale distances by the median absolute deviation of the simulated summary statistics $\mathbf{s}_i$, $i = 1, \ldots, M$. For this basic rejection algorithm, the accepted ϕ_i form a random sample from the approximate posterior distribution defined as

$$p_\delta(\phi|\mathbf{s}) \propto \Pr(\|\mathbf{s} - \mathbf{s}_i\| \leq \delta|\phi)\, p(\phi)$$

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

$$p(\mathbf{s}|\phi) \approx \Pr(\|\mathbf{s} - \mathbf{s}_i\| \leq \delta|\phi)\,.$$

If the summary statistics are sufficient with respect to the parameter ϕ, the approximate posterior distribution converges to the true posterior distribution as δ goes to 0. In addition, the approximate posterior distribution corresponds to the prior distribution when δ is large. Note that more generic interpretations of the ABC algorithm can be found in Wilkinson (2008).

Beaumont et al. (2002) introduced a first improvement of the standard rejection ABC algorithm in which the parameters ϕ_i were weighted by the values $K_\delta(\|\mathbf{s}_i - \mathbf{s}\|)$, where K_δ is the Epanechnikov kernel. Using this weighting scheme, an estimator of the posterior mean was then given by $\sum_i K_\delta(\|\mathbf{s}_i - \mathbf{s}\|)\phi_i / \sum_i K_\delta(\|\mathbf{s}_i - \mathbf{s}\|)$. 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). Note that the Nadaraya-Watson smoother is known to be subject to the curse of dimensionality since the rate of convergence of the estimator decreases dramatically as the dimension of the summary statistics increases (see e.g. Härdle et al. 2004).

A local linear model. 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

$$\phi_i = \alpha + (\mathbf{s}_i - \mathbf{s})^T \beta + \zeta_i, \quad i = 1, \ldots, M, \tag{1}$$

where α is an intercept, β is a vector of regression coefficients, and the ζ_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 parameters (α, β) are inferred by minimizing the weighted least-squares criterion

$$\sum_{i=1}^{n} \{\phi_i - (\alpha + (\mathbf{s}_i - \mathbf{s}_{obs})^t \beta)\}^2 K_\delta(\|\mathbf{s}_i - \mathbf{s}\|). \tag{2}$$

If equation (1) exactly describes the relationship between ϕ and $\mathbf{s}$, 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 ζ_i's, the parameters are adjusted as

$$\phi_i^* = \phi_i - (\mathbf{s}_i - \mathbf{s})^T \hat{\beta}. \tag{3}$$

Weighted by $K_\delta(\|\mathbf{s}_i - \mathbf{s}\|)$, the ϕ_i^*'s, $i = 1, \ldots, M$, provide an approximate sample from the posterior distribution. In this approach, the choice of δ involves a bias-variance trade-off: Increasing δ 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, ϕ_i, in equation (1) (see e.g. Fan and Yao 1998). The new regression model takes the form of a nonlinear conditional heteroscedastic model

$$\phi_i = m(\mathbf{s}_i) + \sigma(\mathbf{s}_i) \times \zeta_i, \quad i = 1, \ldots, M, \tag{4}$$

where $m(\mathbf{s}_i)$ denotes the conditional expectation, $\mathrm{E}[\phi|\mathbf{S} = \mathbf{s}_i]$, and $\sigma^2(\mathbf{s}_i)$ denotes the conditional variance, $\mathrm{Var}[\phi|\mathbf{S} = \mathbf{s}_i]$.

The conditional expectation can be estimated as $\hat{m}(\mathbf{s}_i)$ by fitting a flexible non-linear regression model. The variance term is then estimated using a second regression model for the log-residuals

$$\log(\phi_i - \hat{m}(\mathbf{s}_i))^2 = \log\sigma^2(\mathbf{s}_i) + \xi_i, \quad i = 1, \ldots, M, \tag{5}$$

where the ξ_i's are independent random variates with mean zero and common variance. In our forthcoming examples, we consider feed-forward neural network (FFNN) regression models (Ripley 1996; Bishop 2006). This choice is motivated by the fact that FFNNs include the possibility to reduce the dimensionality of the set of summary statistics via internal projections on lower dimensional subspaces (Bishop, 2006).

In FFNN regression models, linear combinations of the inputs — i.e. the summary statistics — are transformed to compute the values z_j, $j = 1, \ldots, H$ at the H *hidden units*

$$z_j = h(\sum_{k=1}^{D} w_{jk}^{(1)}\mathbf{s}^k + w_{j0}^{(1)}), \quad j = 1, \ldots, H, \tag{6}$$

where $\mathbf{s}^k$, $k = 1, \ldots, D$, is the k^{th} component of $\mathbf{s}$, H is the number of hidden units in the network, D is the dimension of the vector $\mathbf{s}$ of summary statistics, the $w_{jk}^{(1)}$'s, $k = 0, \ldots, D$, $j = 1, \ldots, H$ are the weights of the first layer of the neural network, and h denotes the logistic function. Because the number of hidden units H is typically smaller that the number of summary statistics D, equation (6) reduces the initial dimension of the vector of summary statistics. The values z_i, $i = 1, \ldots, H$, of the hidden units are then linearly combined to produce the output $g_{\mathbf{w}}(\mathbf{s})$ of the FFNN

$$g_{\mathbf{w}}(\mathbf{s}) = \sum_{j=1}^{H}(w_j^{(2)} z_z + w_0^{(2)}), \tag{7}$$

where the $w_j^{(2)}$'s, $j = 0, \ldots, H$, are the weights of the second layer of the neural network. Note that logistic regression can easily be performed within the FFNN framework by transforming the linear combination of equation (7) with a logistic function. The extension to general discrete parameters is obtained using a softmax transformation (see Bishop 2006).

We use FFNNs for fitting both $m(\mathbf{s}_i)$ and $\log \sigma^2(\mathbf{s}_i)$ (Nix and Weigend 1995). The weights $\mathbf{w}$ of a first FFNN are found by minimizing the regularized least-squares criterion

$$\sum_{i=1}^{n} \{\phi_i - g_{\mathbf{w}}(\mathbf{s})\}^2 K_\delta(\|\mathbf{s}_i - \mathbf{s}\|) + \lambda \|\mathbf{w}\|^2,$$

where λ represents the regularization parameter, called weight-decay parameter in the neural network literature. The weights of a second FFNN estimating the conditional variance, are found by minimizing

$$\sum_{i=1}^{n} \{\log((\phi_i - \hat{m}(\mathbf{s}_i))^2) - g_{\mathbf{w}'}(\mathbf{s})\}^2 K_\delta(\|\mathbf{s}_i - \mathbf{s}\|) + \lambda \|\mathbf{w}'\|^2.$$

Similarly to equation (3), parameter adjustment under the NCH model can be performed as follows

$$\phi_i^* = \hat{m}(\mathbf{s}) + (\phi_i - \hat{m}(\mathbf{s}_i)) \times \frac{\hat{\sigma}(\mathbf{s})}{\hat{\sigma}(\mathbf{s}_i)}, \quad i = 1, \ldots, M. \tag{8}$$

Assuming that $\phi_i = m(\mathbf{s}_i) + \sigma(\mathbf{s}_i)\zeta_i$ corresponds to the true relationship between ϕ_i and $\mathbf{s}_i$, then (ϕ_i^*) forms a random sample from the distribution $p(\phi|\mathbf{s})$ provided that $\hat{m}$ could be considered equal to m and $\hat{\sigma}$ equal to σ.

Similarly to the LocL ABC method, a tolerance error, δ, is allowed, and the adjusted parameters, ϕ_i^*, are weighted by $K_\delta(\|\mathbf{s}_i - \mathbf{s}\|)$. Furthermore, to warrant that the adjusted parameters, ϕ_i^*, obtained from equation (3) or (8), 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. These transformations have the further potential advantage of stabilizing the variance of the response when performing regression (Box and Cox 1964).

Iterated importance sampling. A second change to the LocL ABC algorithm converts the single-stage regression based ABC method into a multistage algorithm in which estimations are improved iteratively (Liu 2001; Sisson et al. 2007). In practice, we implemented a two-stage algorithm. The logic of using a two-stage algorithm is that the second run can control a first run with a high acceptance rate, 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 NCH algorithm can be described as follows: Starting from a sample (ϕ_i^1) obtained from a first NCH ABC run, the adaptive step of the algorithm consists of estimating the support Δ_1 of the sample. Then new parameters are proposed from the conditional prior distribution p_{Δ_1} given that they fall in Δ_1. This can be implemented using a simple rejection step. For ϕ having a moderate number of dimension, this is usually achieved at a computational cost which can be considered significantly lower than the cost of simulating from the generating distribution, $p(\mathbf{s}|\phi)$. Using this new set of parameters, a second sample, (ϕ_i^2), can be formed using the NCH method again. To compensate for the fact that we do not sample from the prior distribution, each value ϕ_i^2 should in principle be weighted by $p(\phi_i^2)/p_{\Delta_1}(\phi_i^2)$. Here, the importance weights need not to be computed because we have $p_{\Delta_1}(\phi_i^2) = p(\phi_i^2)/p(\Delta_1)$, which means that the ratio $p(\phi_i^2)/p_{\Delta_1}(\phi_i^2)$ does not depend on i. In this two-stages approach, we suppose that Δ_1 approximates the support of $p(\mathbf{s}|\phi)$ accurately. For multi-dimensional parameters ϕ, we estimate the support of the distribution using support vector machines (SVM, Schölkopf et al. 2001).

3 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 approximate 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 squares adjustment for neural networks was implemented using the R package `nnet` (R Core Team 2008). Model choice for neural networks was based on a Bayesian (or regularization) approach (Ripley 1996). We used $H = 4$ hidden units and the weight-decay regularization parameter was set equal to $\lambda = 0.001$. Increasing the weight-decay parameter will increase the bias of the estimator, but it will also decrease its variance. At this stage, cross-validation might be a useful alternative approach to the choice of a FFNN model, but the previous values for λ and H proved to work well in the examples considered here. In the ANCH algorithm, the support of the conditional density was estimated using a SVM ν-regression algorithm ($\nu = 0.005$) as implemented in the R package `e1071` based on the public library `libsvm` (Chang and Lin 2001).

Two 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 (see the Supplementary Material for further information).

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 θ 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, $\mathbf{s}$, is computed as the number of segregating sites. Note that $\mathbf{s}$ is not a sufficient statistic with respect to θ (Fu and Li 1993). The generating mechanism for $\mathbf{s}$ 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 $(j-1)/2$, $j = 2, \dots, n$.

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

The simulation of L_n can be derived from the formula $L_n = \sum_{j=2}^{n} jY_j$ that gives the total length of the tree as a function of the inter-coalescence times denoted by Y_j, $j = 2, \dots, n$. The inter-coalescence times correspond to the times during which the sample has j ancestors, $j = 2, \dots, n$ (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 $j(j-1)/2$, $j = 2, \dots, n$. A more detailed description of the coalescent process can be found in (Tavaré 2004).

We computed the posterior distribution of the effective mutation rate θ given the observation of $\mathbf{s} = 10$ segregating sites in a sample of $n = 100$ DNA sequences. The prior distribution for the parameter θ was taken to be the exponential distribution of mean 50, which was meant to represent a vague prior. An 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 distribution using a total of 2,000 simulations of the bivariate vector $(\theta, \mathbf{s})$. 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,...,5, with the corresponding empirical quantiles, Q_k^0, obtained from the exact sample. For values of the tolerance rate between 0 and 1 and for each quantile, Q_k, the accuracy of each algorithm was assessed by the relative median absolute error defined as the median of $(Q_k - Q_k^0)/Q_k^0$ computed over 150 runs. In addition, we measured the discrepancy between each approximate distribution and the empirical posterior distribution using the sum of the RMAE's 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. A total of 1,000 replicates were used in the first run and the tolerance rate, denotes as P_δ, was set to 75%.

Comparisons with standard rejection algorithms were first conducted. We found that the posterior distribution obtained from the rejection methods deviated from the empirical posterior distribution significantly for tolerance 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, curves with diamonds). 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, curve with triangles point-up). The adaptive NCH algorithm achieved even superior performances for values of the tolerance rate ranging between 0 and 90% (Figure 1B, curve with triangles point-down). 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. We note, however, that for the smallest tolerance rate, the RMAE is smaller in the LocL model than in the NCH models. Because of the additional level of complexity introduced in FFNNs, FFNNs require more data points to be trained than local linear regression, but this not an issue for higher tolerance rates.

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 α 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 with n tips 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, S_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 $S_1 - S_0$ 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 three ABC algorithms, and we compared these 5 values for the conditional distributions of each parameter t_0, N_A, and α. 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 ($P_\delta \geq 50\%$), we observed a strong agreement with the values estimated by the LocL model (Figure 2, green curves) used with small tolerance rates ($P_\delta = 5\%$), 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 α (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 $\mathbf{s} = 10$ segregating sites, we ran the ANCH algorithm using 200 simulations of the implicit model at each step (tolerance rate $P_\delta = 85\%$). 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 (Gourieroux et al. 1993) 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 ϕ 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 means of finding 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 θ_3. Let W_n be the inter-arrival time of the nth 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 integration.

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 θ_1, $\theta_2 - \theta_1$, and for θ_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 θ_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 posterior quantile estimates came from the use of local optimization during the learning phase of the feed-forward neural networks. For θ_1 and θ_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.

4 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 benefited 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. Increasing the number of summary statistics has a dramatic effect on the variability of the estimators of the conditional mean $\hat{m}(\mathbf{s}_i)$ and variance $\hat{\sigma}^2(\mathbf{s}_i)$ and consequently inflates the variances of the estimated posterior distributions. The variance can be reduced with the Bayesian predictive approach of Ripley (1996) that consists of training a large number of FFNNs for each conditional regression and averaging the results over the replicate networks. In addition Bayesian neural network theory includes general rules for choosing appropriate regularization parameters 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, Toni et al. 2009, Blum and Tran 2008) or for the evolution of protein networks (Ratmann et al. 2007). It has also recently been suggested in the context of compositional data (Butler and Glasbey 2008), and applied in extreme value theory (Bortot et al. 2007) and Gibbs random fields (Grelaud et al. 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. An R code for performing ABC with the NCH model is available at the author's website.

Acknowledgements The authors acknowledge the support of the ANR grant MAEV and M. B. acknowledges the support of the Rhône-Alpes Institute of Complex Systems (IXXI). We are also grateful to Oscar Gaggiotti and two anonymous reviewers for their helpful comments.

References

[1] Beaumont MA, Zhang W, Balding DJ (2002) Approximate Bayesian computation in population genetics. Genetics 162:2025–2035.

[2] Beaumont MA (2007) Joint determination of topology, divergence time, and immigration in population trees. In Matsumura S, Forster P, Renfrew C. *Simulation, Genetics and Human Prehistory.* McDonald Institute Monographs: Cambridge McDonald Institute for Archeological Research, UK. In press.

[3] Beaumont MA, Cornuet J-M, Marin J-M, Robert CP (2008) Adaptivity for ABC algorithms: the ABC-PMC scheme. arXiv:0805.2256

[4] Bellman RE (1961) *Adaptive Control Processes.* Princeton University Press, Princeton.

[5] Bishop CM (2006) *Pattern Recognition and Machine Learning.* Springer, New York.

[6] Blum MGB, Tran VC (2008) Approximate Bayesian Computation for epidemiological models: Application to the Cuban HIV-AIDS epidemic with contact-tracing and unobserved infectious population. arXiv:0810.0896

[7] Box GEP, Cox DR (1964) An analysis of transformations. J Roy Statist Soc B 26: 211–246.

[8] Bortot P, Coles SG and Sisson SA (2007) Inference for stereological extremes. J Amer Statistical Assoc 102: 84–92.

[9] Butler A, Glasbey CA (2008) A latent Gaussian model for compositional data with structural zeroes. Journal of the Royal Statistical Society Series C (Applied Statistics). To appear

[10] Chang CC, Lin CJ (2001) *LIBSVM: A Library For Support Vector Machines*. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[11] Diggle PJ, Gratton RJ (1984) Monte Carlo methods of inference for implicit statistical models. J Roy Statist Soc B 46:193–227.

[12] Fagundes NJR, Ray N, Beaumont M, Neuenschwander S, Salzano SM, Bonatto SL, Excoffier L (2007) Statistical evaluation of alternative models of human evolution. Proc Natl Acad Sci USA 104:17614-17619.

[13] Fan J, Yao Q (1998) Efficient estimation of conditional variance functions in Stochastic regression. Biometrika 85:645-660

[14] Friedman JH, Stuetze W (1981) Projection pursuit regression. J Am Stat Assoc 76:817–823.

[15] Fu Y-X, Li W-H (1993) Maximum Likelihood Estimation of Population Parameters. Genetics 134: 1261-1270.

[16] Fu Y-X, Li W-H (1997) Estimating the age of the common ancestor of a sample of DNA sequences. Mol Biol Evol 14:195–199.

[17] Gelman A, Carlin JB, Stern HS, and Rubin DB (2003) *Bayesian Data Analysis*, 2nd edition, Chapman & Hall, London.

[18] Grelaud A, Robert CP, Marin J-M, Rodolphe F, Taly J-F (2008) ABC methods for model choice in Gibbs random fields. arXiv:0807.2767

[19] Gourieroux C, Monfort A, Renault E (1993) Indirect Inference. J Appl Econom 8:85–118.

[20] Härdle W, M Müller, Sperlich S, Werwatz A (2004) Nonparametric and Semiparametric Models. Springer-Verlag.

[21] Heggland K and Frigessi A (2004) Estimating functions in indirect inference. J Roy Statist Soc B 66:447-462.

[22] Hey J, Nielsen R (2007) Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. Proc Natl Acad Sci USA 104:2785–2790.

[23] King JP, Kimmel M, Chakraborty R (2000) A power analysis of microsatellite-based statistics for inferring past population growth. Mol Biol Evol 17:1859-1868

[24] Kuhner MK (2006) LAMARC 2.0: maximum likelihood and Bayesian estimation of population parameters. Bioinformatics 22:768-770.

[25] Liu JS (2001) Monte Carlo Strategies in Scientific Computing, Springer, New-York.

[26] Marjoram P, Molitor J, Plagnol V, Tavaré S (2003) Markov chain Monte Carlo without likelihoods. Proc Natl Acad Sci USA 100: 15324–15328.

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

[28] Nadaraya, E A (1964) On Estimating Regression Theory of Probability and its Applications 9: 141-142.

[29] Nix DA, Weigend AS (1995) Learning local error bars for nonlinear regression. In Tesauro G, Touretzky D, Leen T eds *Advances in Neural Information Processing Systems* 7 (NIPS'94), MIT Press, Cambridge MA, pp. 489–496.

[30] Ohta T, Kimura M (1973) A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. Genet Res 22:201–204.

[31] Pritchard JK, Feldman MW (1996) Statistics for microsatellite variation based on coalescence. Theor Pop Biol 50: 325-344

[32] Pritchard JK, Seielstad MT, Perez-Lezaun A, Feldman MW (1999) Population growth of human Y chromosomes: a study of Y chromosome microsatellites. Mol Biol Evol 16:1791–1798.

[33] R Development Core Team (2008) R: A Language and Environment for Statistical. R Foundation for Statistical Computing, Vienna, Austria.

[34] Ratmann O, Jørgensen O, Hinkley T, Stumpf M, Richardson S, Wiuf C (2007) Using likelihood-free inference to compare evolutionary dynamics of the protein networks of *H. pylori* and *P. falciparum*. PLoS Comp Biol, 3:e230

[35] Reich DE, Goldstein DB (1998) Genetic evidence for a Paleolithic human population expansion in Africa. Proc Natl Acad Sci USA 95: 8119-8123

[36] Ripley BD (1996) Pattern Recognition and Neural Networks. Oxford University Press, Oxford, UK.

[37] Robert CP, Casella G (2004) Monte Carlo Statistical Methods 2nd Edition. Springer-Verlag, New York.

[38] Schölkopf B, Platt JC, Shawe-Taylor J, Smola AJ (2001) Estimating the support of a high-dimensional distribution. Neural Computation 13:1443–1471.

[39] Shriver MD, Jin L, Ferrell RE, Deka R (1997) Microsatellite data support an early population expansion in Africa. Genome Res 7:586-591

[40] Sisson SA, Fan Y, Tanaka MM (2007) Sequential Monte Carlo without likelihoods. Proc Natl Acad Sci USA 104:1760–1765.

[41] Stephens M, Donnelly P (2000) Inference in molecular population genetics. J Roy Statist Soc Ser B 62: 605-635.

[42] Tanaka M, Francis A, Luciani F, Sisson S (2006) Estimating tuberculosis transmission parameters from genotype data using approximate Bayesian computation. Genetics 173:1511–1520.

[43] Tavaré S (2004) Ancestral Inference in Population Genetics. J Picard ed in *Lectures on Probability Theory and Statistics*, Springer-Verlag, Berlin. pp. 1–188.

[44] Toni T, Welch D, Strelkowa N, Stumpf MPH (2009) Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. J R Soc Interface 6:187-202.

[45] Tavaré S, Balding DJ, Griffiths RC, Donnelly P (1997) Inferring coalescence times from DNA sequence data. Genetics 145: 505–518.

[46] Vapnik VN (1998) *Statistical Learning Theory.* Wiley: New-York.

[47] Watson GS (1964) Smooth regression analysis. Shankya Series A 26: 359–372

[48] Weiss G, von Haeseler A (1998) Inference of population history using a likelihood approach. Genetics 149:1539-1546.

[49] Wilkinson RD (2008). Approximate Bayesian computation (ABC) gives exact results under the assumption of model error. arXiv:0811.3355

[50] Wilson IJ, Weale M E, Balding DJ (1998) Inferences from DNA data: population histories, evolutionary processes and forensic match probabilities. J Roy Statist Soc Ser A 166:155-187.

[51] Zhivotovsky LA, Bennett L, Bowcock AM, Feldman MW (2000) Human population expansion and microsatellite variation. Mol Biol Evol 17: 757-767

[52] Zhivotovsky LA, Rosenberg NA, Feldman MW (2003) Features of evolution and expansion of modern humans, inferred from genome-wide microsatellite markers. Am J Hum Genet 72:1171–1186.

Figure legends

Figure 1. Relative median absolute error (RMAE) when estimating the quantiles of the posterior distribution of θ as a function of the tolerance rate. In panel A), the errors obtained with the regression algorithms are compared to the errors obtained with the local linear adjustment. In panel B), the different regression methods are compared. 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 θ_1, θ_2, and θ_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 θ_3 as a function of the tolerance rate. The different ABC methods were run 100 times each using 2,000 simulations at each run.

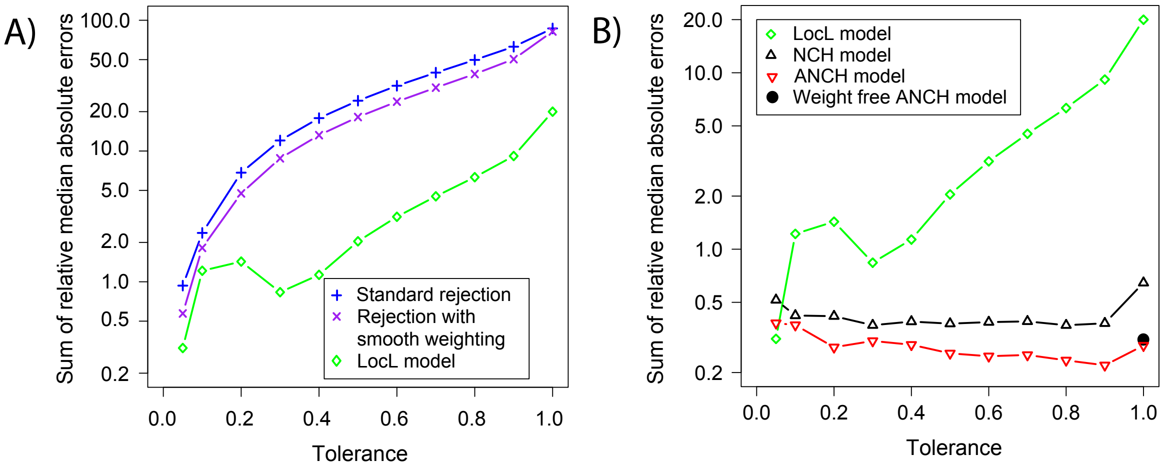
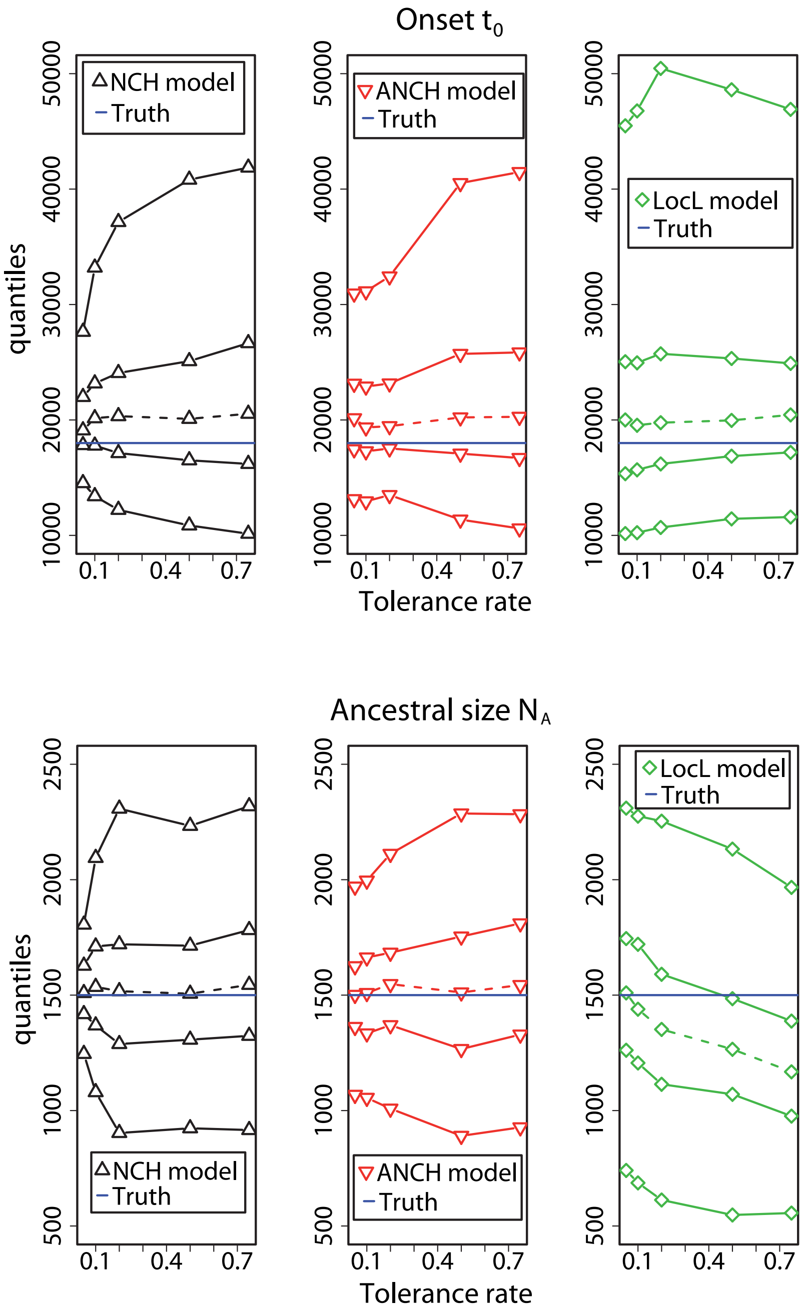

A)
Sum of relative median absolute errors
100.0
50.0
20.0
10.0
5.0
2.0
1.0
0.5
0.2
0.0
0.2
0.4
0.6
0.8
1.0
Tolerance
+ Standard rejection
× Rejection with smooth weighting
◇ LocL model
B)
Sum of relative median absolute errors
20.0
10.0
5.0
2.0
1.0
0.5
0.2
0.0
0.2
0.4
0.6
0.8
1.0
Tolerance
◇ LocL model
△ NCH model
▽ ANCH model
● Weight free ANCH model

Figure 1:



Onset t_0
quantiles
NCH model
Truth
ANCH model
Truth
LocL model
Truth
10000
20000
30000
40000
50000
0.1
0.4
0.7
Tolerance rate
Ancestral size N_A
quantiles
NCH model
Truth
ANCH model
Truth
LocL model
Truth
500
1000
1500
2000
2500
0.1
0.4
0.7
Tolerance rate

Figure 2:

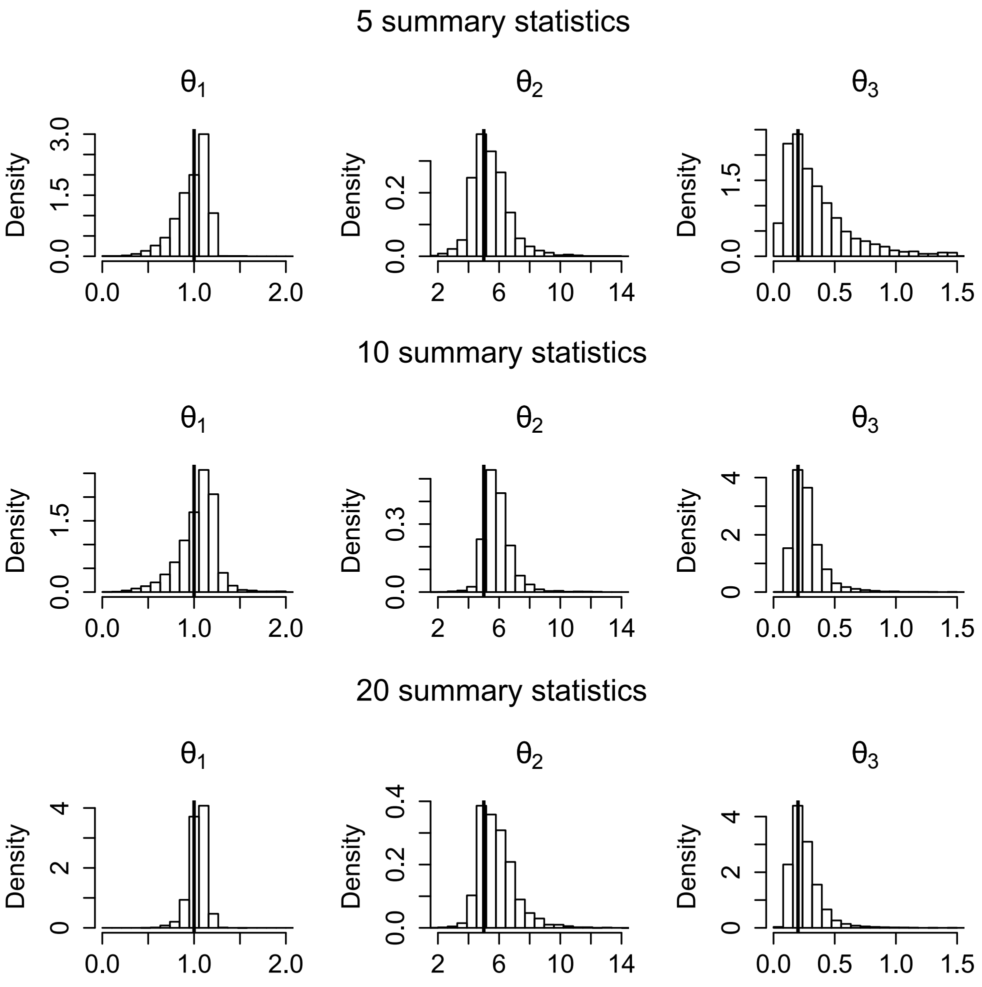

5 summary statistics
θ_1
θ_2
θ_3
Density
0.0
1.5
3.0
0.0
1.0
2.0
0.0
0.2
2
6
10
14
0.0
1.5
0.0
0.5
1.0
1.5
10 summary statistics
θ_1
θ_2
θ_3
0.0
1.5
0.0
1.0
2.0
0.0
0.3
2
6
10
14
0
2
4
0.0
0.5
1.0
1.5
20 summary statistics
θ_1
θ_2
θ_3
0
2
4
0.0
1.0
2.0
0.0
0.2
0.4
2
6
10
14
0
2
4
0.0
0.5
1.0
1.5

Figure 3:

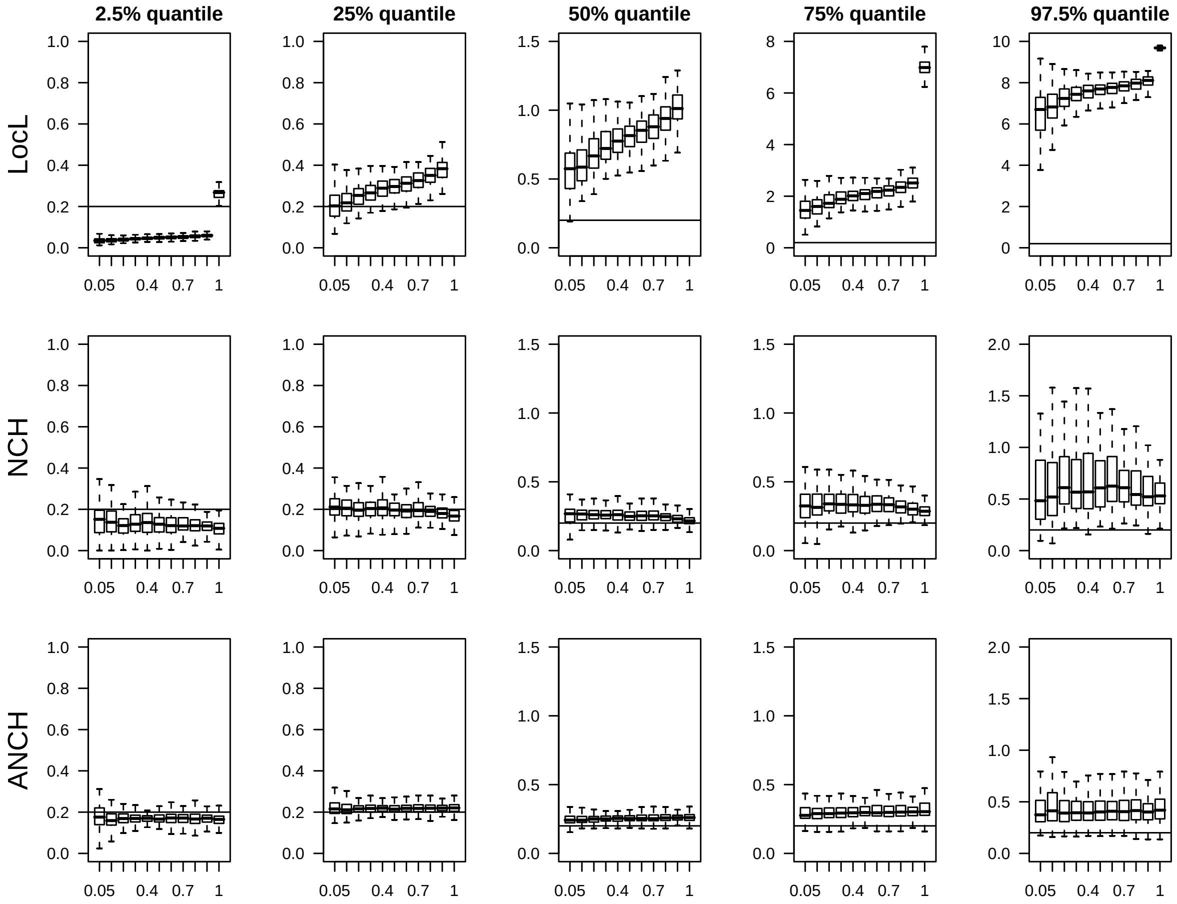

2.5% quantile
25% quantile
50% quantile
75% quantile
97.5% quantile
LocL
NCH
ANCH
0.0
0.2
0.4
0.6
0.8
1.0
1.5
2.0
0
2
4
6
8
10
0.05
0.4
0.7
1

Figure 4:

Table caption

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

Table 1: