Published in "Materials Characterization 61(6): 653-660 , 2010"
which should be cited to refer to this work.

Bayesian Computation and Model Selection Without Likelihoods

Christoph Leuenberger*,[1,2] and Daniel Wegmann†,[1]

*Département de Mathématiques, Université de Fribourg, 1200 Fribourg, Switzerland and †Computational and Molecular Population Genetics Laboratory, Institute of Ecology and Evolution, University of Bern, 3012 Bern, Switzerland

ABSTRACT

Until recently, the use of Bayesian inference was limited to a few cases because for many realistic probability models the likelihood function cannot be calculated analytically. The situation changed with the advent of likelihood-free inference algorithms, often subsumed under the term approximate Bayesian computation (ABC). A key innovation was the use of a postsampling regression adjustment, allowing larger tolerance values and as such shifting computation time to realistic orders of magnitude. Here we propose a reformulation of the regression adjustment in terms of a general linear model (GLM). This allows the integration into the sound theoretical framework of Bayesian statistics and the use of its methods, including model selection via Bayes factors. We then apply the proposed methodology to the question of population subdivision among western chimpanzees, *Pan troglodytes verus*.

WITH the advent of ever more powerful computers and the refinement of algorithms like MCMC or Gibbs sampling, Bayesian statistics have become an important tool for scientific inference during the past two decades. Consider a model $\mathcal{M}$ creating data $\mathcal{D}$ (DNA sequence data, for example) determined by parameters $\boldsymbol{\theta}$ from some (bounded) parameter space $\Pi \subset \mathbb{R}^m$ whose joint prior density we denote by $\pi(\boldsymbol{\theta})$. The quantity of interest is the posterior distribution of the parameters, which can be calculated by Bayes rule as

$$\pi(\boldsymbol{\theta} \mid \mathcal{D}) = c \cdot f_{\mathcal{M}}(\mathcal{D} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta}),$$

where $f_{\mathcal{M}}(\mathcal{D} \mid \boldsymbol{\theta})$ is the likelihood of the data and $c = \int_{\Pi} f_{\mathcal{M}}(\mathcal{D} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})\,d\boldsymbol{\theta}$ is a normalizing constant. Direct use of this formula, however, is often prevented by the fact that the likelihood function cannot be calculated analytically for many realistic probability models. In these cases one is obliged to use stochastic simulation. Tavaré *et al.* (1997) propose a rejection sampling method for simulating a posterior random sample where the full data $\mathcal{D}$ are replaced by a summary statistic s (like the number of segregating sites in their setting). Even if the statistic does not capture the full information contained in the data $\mathcal{D}$, rejection sampling allows for the simulation of approximate posterior distributions of the parameters in question (the scaled mutation rate in their model). This approach was extended to multiple-parameter models with multivariate summary statistics $\mathbf{s} = (s_1, \ldots, s_n)^T$ by Weiss and von Haeseler (1998). In their setting a candidate vector $\boldsymbol{\theta}$ of parameters is simulated from a prior distribution and is accepted if its corresponding vector of summary statistics is sufficiently close to the observed summary statistics $\mathbf{s}_{\text{obs}}$ with respect to some metric in the space of $\mathbf{s}$, *i.e.*, if $\text{dist}(\mathbf{s}, \mathbf{s}_{\text{obs}}) < \epsilon$ for a fixed tolerance ϵ. We suppose that the likelihood $f_{\mathcal{M}}(\mathbf{s} \mid \boldsymbol{\theta})$ of the full model is continuous and nonzero around $\mathbf{s}_{\text{obs}}$. In practice the summary statistics are often discrete but the range of values is large enough to be approximated by real numbers. The likelihood of the truncated model $\mathcal{M}_{\epsilon}(\mathbf{s}_{\text{obs}})$ obtained by this acceptance–rejection process is given by

$$f_{\epsilon}(\mathbf{s} \mid \boldsymbol{\theta}) = \text{Ind}(\mathbf{s} \in \mathcal{B}_{\epsilon}(\mathbf{s}_{\text{obs}})) \cdot f_{\mathcal{M}}(\mathbf{s} \mid \boldsymbol{\theta}) \cdot \left(\int_{\mathcal{B}_{\epsilon}} f_{\mathcal{M}}(\mathbf{s} \mid \boldsymbol{\theta})\,d\mathbf{s} \right)^{-1}, \quad (1)$$

where $\mathcal{B}_{\epsilon} = \mathcal{B}_{\epsilon}(\mathbf{s}_{\text{obs}}) = \{\mathbf{s} \in \mathbb{R}^n \mid \text{dist}(\mathbf{s}, \mathbf{s}_{\text{obs}}) < \epsilon\}$ is the ϵ-ball in the space of summary statistics and $\text{Ind}(\cdot)$ is the indicator function. Observe that $f_{\epsilon}(\mathbf{s} \mid \boldsymbol{\theta})$ degenerates to a (Dirac) point measure centered at $\mathbf{s}_{\text{obs}}$ as $\epsilon \to 0$. If the parameters are generated from a prior $\pi(\boldsymbol{\theta})$, then the distribution of the parameters retained after the rejection process outlined above is given by

$$\pi_{\epsilon}(\boldsymbol{\theta}) = \frac{\pi(\boldsymbol{\theta}) \int_{\mathcal{B}_{\epsilon}} f_{\mathcal{M}}(\mathbf{s} \mid \boldsymbol{\theta})\,d\mathbf{s}}{\int_{\Pi} \pi(\boldsymbol{\theta}) \int_{\mathcal{B}_{\epsilon}} f_{\mathcal{M}}(\mathbf{s} \mid \boldsymbol{\theta})\,d\mathbf{s}\,d\boldsymbol{\theta}}. \quad (2)$$

We call this density the *truncated prior*. Combining (1) and (2) we get

$$\begin{aligned} \pi(\boldsymbol{\theta} \mid \mathbf{s}_{\text{obs}}) &= \frac{f_{\mathcal{M}}(\mathbf{s}_{\text{obs}} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int_{\Pi} f_{\mathcal{M}}(\mathbf{s}_{\text{obs}} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})\,d\boldsymbol{\theta}} \\ &= \frac{f_{\epsilon}(\mathbf{s}_{\text{obs}} \mid \boldsymbol{\theta})\pi_{\epsilon}(\boldsymbol{\theta})}{\int_{\Pi} f_{\epsilon}(\mathbf{s}_{\text{obs}} \mid \boldsymbol{\theta})\pi_{\epsilon}(\boldsymbol{\theta})\,d\boldsymbol{\theta}}. \end{aligned} \quad (3)$$

Thus the posterior distribution of the parameters under the model $\mathcal{M}$ for $\mathbf{s} = \mathbf{s}_{\text{obs}}$ given the prior $\pi(\boldsymbol{\theta})$ is exactly

[1]These authors contributed equally to this work.

[2]*Corresponding author:* Département de Mathématiques, Université de Fribourg, Fribourg, Switzerland. E-mail: christoph.leuenberger@unifr.ch

equal to the posterior distribution under the truncated model $\mathcal{M}_\epsilon(\mathbf{s}_{\text{obs}})$ given the truncated prior $\pi_\epsilon(\boldsymbol{\theta})$. If we can estimate the truncated prior and make an educated guess for a parametric statistical model of $M_\epsilon(\mathbf{s}_{\text{obs}})$, we arrive at a reasonable approximation of the posterior $\pi(\boldsymbol{\theta} \mid \mathbf{s}_{\text{obs}})$ even if the likelihood of the full model $\mathcal{M}$ is unknown. It is to be expected that due to the localization process the truncated model will exhibit a simpler structure than the full model $\mathcal{M}$ and thus be easier to estimate.

Estimating $\pi_\epsilon(\boldsymbol{\theta})$ is straightforward, at least when the summary statistics can be sampled from $\mathcal{M}$ in a reasonable amount of time: Sample the parameters from the prior $\pi(\boldsymbol{\theta})$, create their respective statistics $\mathbf{s}$ from $\mathcal{M}$, and save those parameters whose statistics lie in $\mathcal{B}_\epsilon(\mathbf{s}_{\text{obs}})$ in a list $\mathcal{P} = \{\boldsymbol{\theta}^1, \ldots, \boldsymbol{\theta}^N\}$. The empirical distribution of these retained parameters yields an estimate of $\pi_\epsilon(\boldsymbol{\theta})$. If the tolerance ϵ is small, then one can assume that $f_{\mathcal{M}}(\mathbf{s} \mid \boldsymbol{\theta})$ is close to some (unknown) constant over the whole range of $\mathcal{B}_\epsilon(\mathbf{s}_{\text{obs}})$. Under that assumption, Equation 3 shows that $\pi(\boldsymbol{\theta} \mid \mathbf{s}_{\text{obs}}) \approx \pi_\epsilon(\boldsymbol{\theta})$. However, when the dimension n of summary statistics is high (and for more complex models dimensions like $n = 50$ are not unusual), the "curse of dimensionality" implies that the tolerance must be chosen rather large or else the acceptance rate becomes prohibitively low. This, however, distorts the precision of the approximation of the posterior distribution by the truncated prior (see Wegmann *et al.* 2009). This situation can be partially alleviated by speeding up the sampling process; such methods are subsumed under the term *approximate Bayesian computation* (ABC). Marjoram *et al.* (2003) develop a variant of the classical Metropolis–Hastings algorithm (termed ABC–MCMC in Sisson *et al.* 2007), which allows them to sample directly from the truncated prior $\pi_\epsilon(\boldsymbol{\theta})$. In Sisson *et al.* (2007) a sequential Monte Carlo sampler is proposed, requiring substantially less iterations than ABC–MCMC. But even when such methods are applied, the assumption that $f_{\mathcal{M}}(s \mid \boldsymbol{\theta})$ is constant over the ϵ-ball is a very rough one, indeed.

To take into account the variation of $f_{\mathcal{M}}(\mathbf{s} \mid \boldsymbol{\theta})$ within the ϵ-ball, a postsampling regression adjustment (termed ABC-REG in the following) of the sample $\mathcal{P}$ of retained parameters is introduced in the important article by Beaumont *et al.* (2002). Basically, they postulate a (locally) linear dependence between the parameters $\boldsymbol{\theta}$ and their associated summary statistics $\mathbf{s}$. More precisely, the (local) model they implicitly assume is of the form $\boldsymbol{\theta} = \mathbf{M}\mathbf{s} + \mathbf{m}_0 + \boldsymbol{\epsilon}$, where $\mathbf{M}$ is a matrix of regression coefficients, $\mathbf{m}_0$ a constant vector, and $\boldsymbol{\epsilon}$ a random vector of zero mean. Computer simulations suggest that for many population models ABC–REG yields posterior marginal densities that have narrower highest posterior density (HPD) regions and are more closely centered around the true parameter values than the empirical posterior densities directly produced by ABC samplers (Wegmann *et al.* 2009). An attractive feature of ABC–REG is that the posterior adjustment is performed directly on the simulated parameters, which makes estimation of the marginal posteriors of individual parameters particularly easy. The method can also be extended to more complex, nonlinear models as demonstrated, *e.g.*, in Blum and Francois (2009). In extreme situations, however, ABC–REG may yield posteriors that are nonzero in parameter regions where the priors actually vanish (see Figure 1B for an illustration of this phenomenon). Moreover, it is not clear how ABC–REG could yield an estimate of the marginal density of model $\mathcal{M}$ at $\mathbf{s}_{\text{obs}}$, information that is useful for model comparison.

In contrast to ABC–REG we treat the parameters $\boldsymbol{\theta}$ as exogenous and the summary statistics $\mathbf{s}$ as endogenous variables and we stipulate for $\mathcal{M}_\epsilon(\mathbf{s}_{\text{obs}})$ a general linear model (GLM in the literature—not to be confused with the generalized linear models that unfortunately share the same abbreviation). To be precise, we assume the summary statistics $\mathbf{s}$ created by the truncated model's likelihood $f_\epsilon(\mathbf{s} \mid \boldsymbol{\theta})$ to satisfy

$$\mathbf{s} \mid \boldsymbol{\theta} = \mathbf{C}\boldsymbol{\theta} + \mathbf{c}_0 + \boldsymbol{\epsilon}, \tag{4}$$

where $\mathbf{C}$ is a $n \times m$ matrix of constants, $\mathbf{c}_0$ an $n \times 1$ vector, and $\boldsymbol{\epsilon}$ a random vector with a multivariate normal distribution of zero mean and covariance matrix $\boldsymbol{\Sigma}_s$:

$$\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_s).$$

A GLM has the advantage of taking into account not only the (local) linearity, but also the strong correlation normally present between the components of the summary statistics. Of course, the model assumption (4) can never represent the full truth since its statistics are in principle unbounded whereas the likelihood $f_\epsilon(\mathbf{s} \mid \boldsymbol{\theta})$ is supported on the ϵ-ball around $\mathbf{s}_{\text{obs}}$. But since the multivariate Gaussians will fall off rapidly in practice and not reach far out off the boundary of $\mathcal{B}_\epsilon(\mathbf{s}_{\text{obs}})$, this is a disadvantage we can live with. In particular, the ordinary least squares (OLS) estimate outlined below implies that for $\epsilon \to 0$ the constant $\mathbf{c}_0$ tends to $\mathbf{s}_{\text{obs}}$ whereas the design matrix $\mathbf{C}$ and the covariance matrix $\boldsymbol{\Sigma}_s$ both vanish. This means that in the limit of zero tolerance $\epsilon = 0$ our model assumption yields the true posterior distribution of $\mathcal{M}$.

THEORY

In this section we describe the above methodology—referred to as ABC–GLM in the following—in more detail. The basic two-step procedure of ABC–GLM may be summarized as follows.

GLM1: Given a model $\mathcal{M}$ creating summary statistics $\mathbf{s}$ and given a value of observed summary statistics $\mathbf{s}_{\text{obs}}$, create a sample of retained parameters $\boldsymbol{\theta}^j$, $j = 1, \ldots, N$, with the aid of some ABC sampler (rejection sampling, ABC–MCMC, or ABC–PRC) based on a prior distribution $\pi(\boldsymbol{\theta})$ and some choice of the tolerance $\epsilon > 0$.

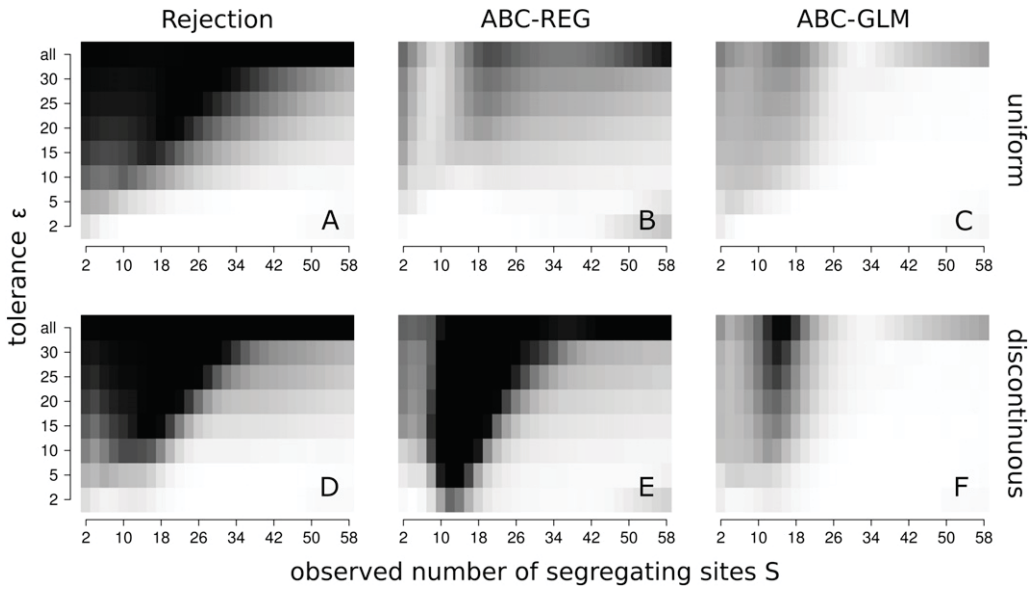

FIGURE 1.—Comparison of rejection (A and D), ABC–REG (B and E), and ABC–GLM (C and F) posteriors with those obtained from analytical likelihood calculations. We estimated the population–mutation parameter $\theta = 4N\mu$ of a panmictic population for different observed numbers of segregating sites (see text). Shades indicate the L_1 distance between the inferred and the analytically calculated posterior. White corresponds to an exact match (zero distance) and darker gray shades indicate larger distances. If the inferred posterior differs from the analytical more than the prior does, squares are marked in black. The top row (A–C) corresponds to cases with a uniform prior $\theta \sim \text{Unif}([0.005, 10])$ and the bottom row (D–F) to cases with a discontinuous prior $\boldsymbol{\theta} \sim \text{Unif}([0.005, 3] \cup [6, 10])$ with "gap." The tolerance ϵ is given as the absolute distance in number of segregating sites. Shown are averages over 25 independent estimations. To have a fair comparison, we adjusted the smoothing parameters (bandwidths) to get the best results for all approaches.

GLM2: Estimate the truncated model $\mathcal{M}_\epsilon(\mathbf{s}_{\text{obs}})$ as a general linear model and determine, on the basis of the sample $\boldsymbol{\theta}^j$, from the truncated prior $\pi_\epsilon(\boldsymbol{\theta})$ an approximation to the posterior $\pi(\boldsymbol{\theta} \mid \mathbf{s}_{\text{obs}})$ according to Equation 3.

Let us look more closely at these two steps.

GLM1: ABC sampling: We refer the reader to MARJORAM *et al.* (2003) and SISSON *et al.* (2007) for details concerning ABC algorithms and to MARJORAM and TAVARÉ (2006) for a comprehensive review of computational methods for genetic data analysis. In practice, the dimension of the summary statistics is often reduced by a principal components analysis (PCA). PCA also has a certain decorrelation effect. A more sophisticated method of reducing the dimension of summary statistics, based on partial least squares (PLS), is described in WEGMANN *et al.* (2009). In a recent preprint, VOGL *et al.* (C. VOGL, C. FUTSCHIK and C. SCHLOETTERER, unpublished data) propose a Box–Cox-type transformation of the summary statistics that makes the likelihood close to multivariate Gaussian. This transformation might be especially efficient in our context as we assume normality of the error terms in our model assumption.

To fix the notation, let $\mathcal{P} = \{\boldsymbol{\theta}^1, \ldots, \boldsymbol{\theta}^N\}$ be a sample of vector-valued parameters created by some ABC algorithm simulating from some prior $\pi(\boldsymbol{\theta})$ and $\mathcal{S} = \{\mathbf{s}^1, \ldots, \mathbf{s}^N\}$ be the sample of associated summary statistics produced by the model $\mathcal{M}$. Each parameter $\boldsymbol{\theta}^j$ is an m-dimensional column vector $\boldsymbol{\theta}^j = (\theta^j, \ldots, \theta^j_m)^t$ and each summary statistic is an n-dimensional column vector $\mathbf{s}^j = (s^j_1, \ldots, s^j_n)^t \in \mathcal{B}_\epsilon(\mathbf{s}_{\text{obs}})$. The samples $\mathcal{P}$ and $\mathcal{S}$ can thus be viewed as $m \times N$ and $n \times N$ matrices $\mathbf{P}$ and $\mathbf{S}$, respectively.

The empirical estimate of the truncated prior $\pi_\epsilon(\boldsymbol{\theta})$ is given by the discrete distribution that puts a point mass of $1/N$ on each value $\boldsymbol{\theta}^j \in \mathcal{P}$. We smooth out this empirical distribution by placing a sharp Gaussian peak over each parameter value $\boldsymbol{\theta}^j$. More precisely, we set

$$\pi_\epsilon(\boldsymbol{\theta}) = \frac{1}{N} \sum_{j=1}^{N} \phi(\boldsymbol{\theta} - \boldsymbol{\theta}^j, \boldsymbol{\Sigma}_\theta), \tag{5}$$

where

$$\phi(\boldsymbol{\theta} - \boldsymbol{\theta}^j, \boldsymbol{\Sigma}_\theta) = \frac{1}{|\, 2\pi\boldsymbol{\Sigma}_\theta|^{\,1/2}} e^{-(1/2)(\boldsymbol{\theta}-\boldsymbol{\theta}^j)^t \boldsymbol{\Sigma}_\theta^{-1} (\boldsymbol{\theta}-\boldsymbol{\theta}^j)}$$

and

$$\boldsymbol{\Sigma}_\theta = \text{diag}(\sigma_1, \ldots, \sigma_m)$$

is the covariance matrix of ϕ that determines the width of the Gaussian peaks. The larger the number N of sampled parameter values is, the sharper the peaks can be chosen to still get a rather smooth π_ϵ. If the parameter domain Π is normalized to $[0, 1]^m$, say, then a reasonable choice is $\sigma_k = 1/N$. Otherwise, σ_k should be adapted to the parameter range of the parameter component θ_k. Too small values of σ_k will result in wiggly posterior curves, and too large values might unduly smear out the curves. The best advice is to run the calculations with several choices for $\boldsymbol{\Sigma}_\theta$. If π_ϵ induces a correlation between parameters, a nondiagonal $\boldsymbol{\Sigma}_\theta$ might be beneficial. In practice, however, the posterior estimates are most sensitive to the diagonal values of $\boldsymbol{\Sigma}_\theta$.

GLM2: general linear model: As explained in the Introduction, we assume the truncated model $\mathcal{M}_\epsilon(\mathbf{s}_{\text{obs}})$ to be normal linear; *i.e.*, the random vectors $\mathbf{s}$ satisfy (4). The covariance matrix $\boldsymbol{\Sigma}_s$ encapsulates the

strong correlations normally present between the components of the summary statistics. $\mathbf{C}$, $\mathbf{c}_0$, and $\mathbf{\Sigma}_s$ can be estimated by standard multivariate regression analysis (OLS) from the sample $\mathcal{P}$, $\mathcal{S}$ created in step GLM1. [Strictly speaking, one must redo an ABC sample from uniform priors over Π to get an unbiased estimate of the GLM if the prior $\pi(\boldsymbol{\theta})$ is not uniform already. On the other hand, ordinary least-squares estimators are quite insensitive to the prior's influence. In practice, one can as well use the sample $\mathcal{P}$ to do the estimate. We applied both estimation methods to the toy models presented in the EXAMPLES FROM POPULATION GENETICS section and found no significant difference between the estimated posteriors. The same holds true for the so-called feasible generalized least-squares (FGLS) estimator; see GREENE (2003). In this two-stage algorithm the covariance matrix is first estimated as in our setting but in a second round the design matrix $\mathbf{C}$ is newly estimated. When we applied FGLS to our toy models, we found a difference in the estimated matrices only after the eighth significant decimal. FGLS is a more efficient estimator only when the sample sizes are relatively small as is often the case in economical data sets but not in ABC situations. In theory, both OLS and FGLS are consistent estimators but FGLS is more efficient.] To be specific, set $\mathbf{X} = (\mathbf{1}\dot{:}\mathbf{P}^t)$, where $\mathbf{1}$ is an $N \times 1$ vector of 1's. $\mathbf{C}$ and $\mathbf{c}_0$ are determined by the usual least-squares estimator

$$(\hat{\mathbf{c}}_0\dot{:}\hat{\mathbf{C}}) = \mathbf{S}\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1},$$

and for $\mathbf{\Sigma}_s$ we have the estimate

$$\hat{\mathbf{\Sigma}}_s = \frac{1}{N-m}\hat{\mathbf{R}}^t\hat{\mathbf{R}}, \tag{6}$$

where $\hat{\mathbf{R}} = \mathbf{S}^t - \mathbf{X}\cdot(\hat{\mathbf{c}}_0\dot{:}\hat{\mathbf{C}})^t$ are the residuals. The likelihood for this model—dropping the hats on the matrices to unburden the notation—is given by

$$f_\epsilon(\mathbf{s}\,|\,\boldsymbol{\theta}) = |2\pi\mathbf{\Sigma}_s|^{-1/2}\cdot e^{-(1/2)(\mathbf{s}-\mathbf{C}\boldsymbol{\theta}-\mathbf{c}_0)^t\mathbf{\Sigma}_s^{-1}(\mathbf{s}-\mathbf{C}\boldsymbol{\theta}-\mathbf{c}_0)}. \tag{7}$$

An exhaustive treatment of linear models in a Bayesian (econometric) context is given in Zellner's book (ZELLNER 1971).

Recall from (3) that for a prior $\pi(\boldsymbol{\theta})$ and an observed summary statistic $\mathbf{s}_{\text{obs}}$, the parameter's posterior distribution for our full model $\mathcal{M}$ is given by

$$\pi(\boldsymbol{\theta}\,|\,\mathbf{s}_{\text{obs}}) = c\cdot f_\epsilon(\mathbf{s}_{\text{obs}}\,|\,\boldsymbol{\theta})\pi_\epsilon(\boldsymbol{\theta}), \tag{8}$$

where $f_\epsilon(\mathbf{s}_{\text{obs}}\,|\,\boldsymbol{\theta})$ is the likelihood of the truncated model $\mathcal{M}_\epsilon(\mathbf{s}_{\text{obs}})$ given by (7) and $\pi_\epsilon(\boldsymbol{\theta})$ is the estimated (and smoothed) truncated prior given by (5).

Performing some matrix algebra (see APPENDIX A), one can show that the posterior (8) is—up to a multiplicative constant—of the form $\sum_{i=j}^{N}\exp(-\frac{1}{2}Q_j)$, where

$$\begin{aligned} Q_j = {} & (\boldsymbol{\theta}-\mathbf{t}^j)^t\mathbf{T}^{-1}(\boldsymbol{\theta}-\mathbf{t}^j) + \ \ldots \\ & \ldots + (\mathbf{s}_{\text{obs}}-\mathbf{c}_0)^t\mathbf{\Sigma}_s^{-1}(\mathbf{s}_{\text{obs}}-\mathbf{c}_0) + \ \ldots \\ & \ldots + (\boldsymbol{\theta}^j)^t\mathbf{\Sigma}_\theta^{-1}\boldsymbol{\theta}^j - (\mathbf{v}^j)^t\mathbf{T}\mathbf{v}^j. \end{aligned}$$

Here $\mathbf{T}$, $\mathbf{t}^j$, and $\mathbf{v}^j$ are given by

$$\mathbf{T} = (\mathbf{C}^t\mathbf{\Sigma}_s^{-1}\mathbf{C} + \mathbf{\Sigma}_\theta^{-1})^{-1} \tag{9}$$

and $\mathbf{t}^j = \mathbf{T}\mathbf{v}^j$, where

$$\mathbf{v}^j = \mathbf{C}^t\mathbf{\Sigma}_s^{-1}(\mathbf{s}_{\text{obs}}-\mathbf{c}_0) + \mathbf{\Sigma}_\theta^{-1}\boldsymbol{\theta}^j. \tag{10}$$

From this we get

$$\pi(\theta\,|\,\mathbf{s}_{\text{obs}}) \propto \sum_{j=1}^{N} c(\boldsymbol{\theta}^j)e^{-(1/2)(\boldsymbol{\theta}-\mathbf{t}^j)^t\mathbf{T}^{-1}(\boldsymbol{\theta}-\mathbf{t}^j)}, \tag{11}$$

where

$$c(\boldsymbol{\theta}^j) = \exp\left[-\frac{1}{2}((\boldsymbol{\theta}^j)^t\mathbf{\Sigma}_\theta^{-1}\boldsymbol{\theta}^j - (\mathbf{v}^j)^t\mathbf{T}\mathbf{v}^j)\right]. \tag{12}$$

When the number of parameters exceeds two, graphical visualization of the posterior distribution becomes impractical and marginal distributions must be calculated. The marginal posterior density of the parameter θ_k is defined by

$$\pi(\theta_k\,|\,\mathbf{s}) = \int_{\mathbb{R}^{m-1}}\pi(\boldsymbol{\theta}\,|\,\mathbf{s})d\boldsymbol{\theta}_{-k},$$

where integration is performed along all parameters except θ_k.

Recall that the marginal distribution of a multivariate normal $\mathcal{N}(\boldsymbol{\mu},\mathbf{\Sigma})$ with respect to the kth component is the univariate normal density $\mathcal{N}(\mu_k,\sigma_{k,k})$. Using this fact, it is not hard to show that the marginal posterior of parameter θ_k is given by

$$\pi(\theta_k\,|\,\mathbf{s}_{\text{obs}}) = a\cdot\sum_{j=1}^{N} c(\boldsymbol{\theta}^j)\exp\left(-\frac{(\theta_k-t_k^j)^2}{2\tau_{k,k}}\right), \tag{13}$$

where $\tau_{k,k}$ is the kth diagonal element of the matrix $\mathbf{T}$, t_k^j is the kth component of the vector $\mathbf{t}^j$, and $c(\boldsymbol{\theta}^j)$ is still determined according to (12). The normalizing constant a could, in principle, be determined analytically but is in practice more easily recovered by a numerical integration. Strictly speaking, the integration should be done only over the bounded parameter domain Π and not over the whole of $\mathbb{R}^m$. But this no longer allows for an analytic form of the marginal posterior distribution. For

large values of N the diagonal elements in the matrix Σ_θ can be chosen so small that the error is in any case negligible.

Model selection: The principal difficulty of model selection methods in nonparametric settings is that it is nearly impossible to estimate the likelihood of $\mathcal{M}$ at $\mathbf{s}_{\text{obs}}$ due to the high dimension of the summary statistics (curse of dimensionality); see Beaumont (2007) for an approach based on multinomial logit. Parametric models on the other hand lend themselves readily to model selection via Bayes factors. Given the model $\mathcal{M}$, one must determine the marginal density

$$f_{\mathcal{M}}(\mathbf{s}_{\text{obs}}) = \int_{\Pi} f(\mathbf{s}_{\text{obs}} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}.$$

It is easy to check from (1) and (2) that

$$f_{\mathcal{M}}(\mathbf{s}_{\text{obs}}) = A_\epsilon(\mathbf{s}_{\text{obs}}, \pi) \cdot \int_{\Pi} f_\epsilon(\mathbf{s}_{\text{obs}} \mid \boldsymbol{\theta})\pi_\epsilon(\boldsymbol{\theta})\, d\boldsymbol{\theta}.$$

Here

$$A_\epsilon(\mathbf{s}_{\text{obs}}, \pi) := \int_{\Pi} \pi(\boldsymbol{\theta}) \int_{\mathcal{B}_\epsilon} f_{\mathcal{M}}(\mathbf{s} \mid \boldsymbol{\theta})\, d\mathbf{s}\, d\boldsymbol{\theta} \qquad (14)$$

is the acceptance rate p of the rejection process. It can easily be estimated with aid of ABC–REJ: Sample parameters from the prior $\pi(\boldsymbol{\theta})$ create the corresponding statistics $\mathbf{s}$ from $\mathcal{M}$ and count what fraction of the statistics fall into the ϵ-ball $\mathcal{B}_\epsilon$ centered at $\mathbf{s}_{\text{obs}}$.

If we assume the underlying model of $\mathcal{M}_\epsilon(\mathbf{s}_{\text{obs}})$ to be our GLM, then the marginal density of $\mathcal{M}$ at $\mathbf{s}_{\text{obs}}$ can be estimated as

$$f_{\mathcal{M}}(\mathbf{s}_{\text{obs}}) = \frac{A_\epsilon(\mathbf{s}_{\text{obs}}, \pi)}{N\ |2\pi\mathbf{D}|^{1/2}} \sum_{j=1}^{N} e^{-(1/2)(\mathbf{s}_{\text{obs}} - \mathbf{m}^j)^t \mathbf{D}^{-1} (\mathbf{s}_{\text{obs}} - \mathbf{m}^j)}, \qquad (15)$$

where the sum runs over the parameter sample $\mathcal{P} = \{\boldsymbol{\theta}^1, \ldots, \boldsymbol{\theta}^N\}$,

$$\mathbf{D} = \boldsymbol{\Sigma}_s + \mathbf{C}\boldsymbol{\Sigma}_\theta \mathbf{C}^t$$

and

$$\mathbf{m}^j = \mathbf{c}_0 + \mathbf{C}\boldsymbol{\theta}^j.$$

For two models $\mathcal{M}_A$ and $\mathcal{M}_B$ with prior probabilities π_A and $\pi_B = 1 - \pi_A$, the Bayes factor B_{AB} in favor of model $\mathcal{M}_A$ over model $\mathcal{M}_B$ is

$$B_{AB} = \frac{f_{\mathcal{M}_A}(\mathbf{s}_{\text{obs}})}{f_{\mathcal{M}_B}(\mathbf{s}_{\text{obs}})}, \qquad (16)$$

where the marginal densities $f_{\mathcal{M}_A}$ and $f_{\mathcal{M}_B}$ are calculated according to (15). The posterior probability of model $\mathcal{M}_A$ is

$$f(\mathcal{M}_A \mid \mathbf{s}_{\text{obs}}) = \frac{B_{AB}\pi_A}{B_{AB}\pi_A + \pi_B}.$$

EXAMPLES FROM POPULATION GENETICS

Toy models: In Figure 1 we present the comparison of posteriors obtained with rejection sampling, ABC–REG and ABC–GLM, with those determined analytically ("true posteriors"). As a toy model we inferred the population–mutation parameter $\theta = 4N\mu$ of a panmictic population model from the number of segregating sites S of a sample of sequences with 10,000 bp for different observed values and tolerance levels. Estimations are always based on 5000 simulations with $\text{dist}(S, S_{\text{obs}}) < \epsilon$, and we report the average of 25 independent replications per data point. Estimation bias of the different approaches was assessed by computing the total variation distance between the inferred posterior and the true one obtained from analytical calculations using the likelihood function introduced by Watterson (1975). Recall that the L_1-distance of two densities $f(\theta)$ and $g(\theta)$ is given by

$$d_1(f, g) = \frac{1}{2} \int |f(\theta) - g(\theta)|\, d\theta.$$

It is equal to 1 when f and g have disjoint supports and it vanishes when the functions are identical.

When we used a uniform prior $\theta \sim \text{Unif}([0.005, 10])$ (Figure 1, A–C), both ABC–REG and ABC–GLM give comparable results and improve the posterior estimation compared to the simple rejection algorithm except for very low tolerance values ϵ where the rejection algorithm is expected to be very close to the true posterior. The average total variation distances over all observed data sets and tolerance values ϵ are 0.236, 0.130, and 0.091 for the rejection algorithm, ABC–REG, and ABC–GLM, respectively. Note that perfect matches between the approximate and the true posteriors are difficult to obtain because all approximate posteriors depend on a smoothing step that may not give accurate results close to borders of their supports. However, when we used a discontinuous prior $\theta \sim \text{Unif}([0.005, 3] \cup [6, 10])$ with an admittedly extremely artificial "gap" in the middle, we observed a quite distinct pattern (Figure 1, D and E). One clearly recognizes that posteriors inferred with ABC–REG are frequently misplaced and often even farther away from the true posterior (in total variation distance) than the prior, especially for cases where the likelihood of the observed data is maximal within the gap. The reason for this is that in the regression step of ABC–REG parameter values may easily be shifted outside the prior support. This behavior of ABC–REG has been observed earlier (Beaumont *et al.* 2002; Estoup *et al.* 2004; Tallmon *et al.* 2004) and as an *ad hoc* solution

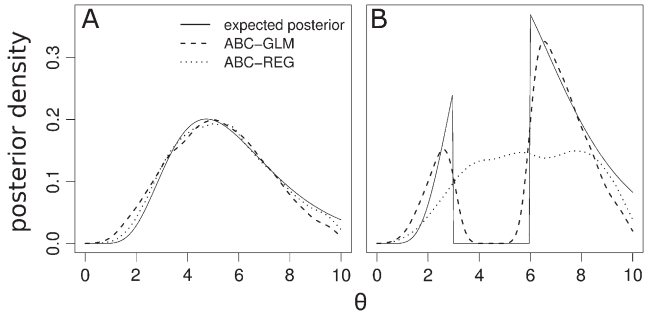

FIGURE 2.—Example posteriors for uniform (A) and discontinuous (B) priors. The model is the same as in Figure 1. Posterior estimates using ABC–GLM and ABC–REG for $S_{obs} = 16$ were based on 5000 simulations with dist(S, S_{obs}) $<$ 10. ABC–REG posteriors were smoothed with a bandwidth of 0.4, and the width of the Dirac peaks in the ABC–GLM approach was set to 10^{-5}.

HAMILTON *et al.* (2006) proposed to transform the parameter values prior to the regression step by a transformation of the form $y = -\ln(\tan(((x-a)/(b-a))(\pi/2))^{-1})$, where a and b are the lower and upper borders of the prior support interval. For more complex priors—like the discontinuous prior used here—this transformation may not work. ABC–GLM is much less affected by the gap prior than ABC–REG. The average total variation distances over all observed data sets and tolerance values ϵ are 0.221, 0.246, and 0.094 for the rejection algorithm, ABC–REG, and ABC–GLM, respectively. Example posteriors with $S_{obs} = 16$ based on 5000 simulations with dist(S, S_{obs}) < 10 are shown in Figure 2.

The success of ABC–GLM depends on how well a general linear model fits the truncated model $\mathcal{M}_\epsilon(\mathbf{s}_{obs})$. Under the null hypothesis that the fit is perfect the estimated residuals $\mathbf{r}_j$ (see Equation 6) are independently multivariate normally distributed random vectors. Hence the Mahalanobis distances

$$d_j = \mathbf{r}_j^t \boldsymbol{\Sigma}_s^{-1} \mathbf{r}_j \sim \chi_n^2 \tag{17}$$

follow a χ^2-distribution with n degrees of freedom. As a quantification of model assessment we propose to report the Kolmogorov–Smirnov test statistic for the empirical distribution of d_j and the reference χ^2-distribution. (Reporting P-values will be of little use in practice since the null hypothesis does never hold exactly and hence the P-values will become very small due to the large sample size.)

When the summary statistics are created from a general linear model, the fit should be optimal. This is indeed the case as the simulation results in Table 1 show. We performed 200 simulations of randomly created general linear models with $m = 3$ parameters, $n = 4$ summary statistics, and a multivariate normal prior. The observed statistics were also created from the respective models. For each simulated observed statistic and

TABLE 1

Mean and standard deviation of the L_1 distance between inferred and expected posteriors for randomly generated GLMs with $N_P = 3$, $N_S = 4$ [prior $N(0, 0.2^2)$, 200 simulations]

p^a	$d_1(\pi_0, \pi_\epsilon)$	$d_1(\pi_0, \pi_{REG})$	$d_1(\pi_0, \pi_{GLM})$	KS statistics[b]
1.00	0.51 ± 0.22	0.15 ± 0.10	0.01 ± 0.001	0.004 ± 0.001
0.50	0.42 ± 0.19	0.13 ± 0.10	0.02 ± 0.008	0.007 ± 0.003
0.10	0.29 ± 0.18	0.13 ± 0.11	0.03 ± 0.01	0.02 ± 0.01
0.05	0.24 ± 0.16	0.13 ± 0.12	0.03 ± 0.01	0.03 ± 0.01
0.01	0.21 ± 0.17	0.15 ± 0.14	0.05 ± 0.02	0.06 ± 0.02

[a] Acceptance rate as a fraction.
[b] KS statistic describing the linear model fit (see text).

different acceptance rates $p = 1.00$, 0.50, 0.10, 0.05, and 0.01 we calculated the approximate posterior distributions π_ϵ, π_{REG}, and π_{GLM} for the rejection algorithm, ABC–REG, and ABC–GLM, respectively. As the prior is multivariate normal, the true posterior π_0 can be analytically determined. Table 1 contains the means and standard deviations over the 200 simulations of the total variation distances of the approximate posteriors to the true posterior π_0 as well as the mean and standard deviations of the Kolmogorov–Smirnov test statistics for the GLM model fit. As is expected, the model fit is perfect [*i.e.*, the Kolmogorov–Smirnov (KS) statistic is close to 0] for acceptance rate $p = 1$. As the acceptance rate becomes lower, the model fit deteriorates since the truncated model of a GLM is no longer exactly a general linear model. The total variation distance to the true posterior increases slightly as p gets smaller but the improved rejection posterior π_ϵ mostly outbalances the poorer model fit. As is expected in this ideal situation, ABC–GLM and ABC–REG substantially improve the posterior estimation over the pure rejection prior.

To test the other extreme we performed 200 simulations for a nonlinear one-parameter model with uniformly rather than normally distributed error terms; the prior was again a normal distribution. (The details of this toy model are described in APPENDIX B.) As Table 2 shows, the GLM model fit is already poor for an acceptance rate of $p = 1.00$ (KS statistic ~0.10) and further deteriorates as p decreases. Note that the approximate posteriors π_{REG} and π_{GLM} are closer to the true posterior in average than π_ϵ and that both adjustment methods perform similarly. As expected, the accuracy of the posteriors increases with smaller acceptance rates, despite the fact that the model fit within the ϵ-ball decreases. This suggests that the rejection step contributes substantially to the estimation accuracy, especially when the true model is nonlinear. We should mention that in ~30% of the simulations both ABC–GLM and ABC–REG actually increased the distance to the true posterior in comparison to the rejection posterior π_ϵ. As a rule of thumb we suggest that posterior

TABLE 2

Mean and standard deviation of the L_1 distance between inferred and expected posteriors for the uniform errors model (see APPENDIX B) with $N_P = 1$, $N_S = 5$ {prior $N(0, 2^2)$, error Unif[−10, 10], 200 simulations}

p[a]	$d_1(\pi_0, \pi_\epsilon)$	$d_1(\pi_0, \pi_{REG})$	$d_1(\pi_0, \pi_{GLM})$	KS statistics[b]
1.00	0.56 ± 0.24	0.49 ± 0.25	0.46 ± 0.29	0.09 ± 0.01
0.50	0.40 ± 0.30	0.36 ± 0.28	0.37 ± 0.27	0.12 ± 0.01
0.10	0.38 ± 0.28	0.35 ± 0.26	0.34 ± 0.23	0.14 ± 0.03
0.05	0.34 ± 0.29	0.33 ± 0.27	0.32 ± 0.23	0.14 ± 0.02
0.01	0.29 ± 0.23	0.26 ± 0.22	0.26 ± 0.18	0.16 ± 0.03

[a] Acceptance rate as a fraction.
[b] KS statistic describing the linear model fit (see text).

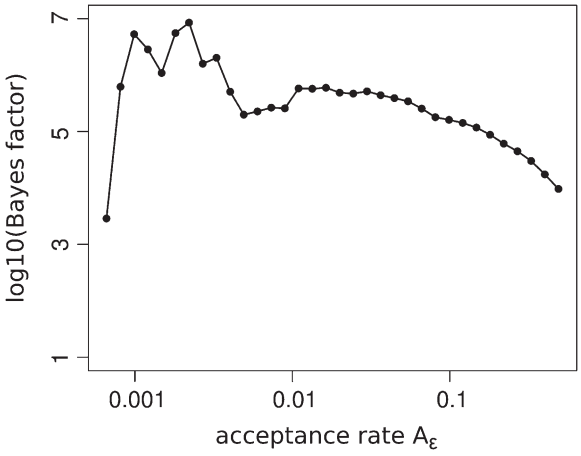

FIGURE 3.—Bayes factor for the island relative to the panmictic population model for different acceptance rates (logarithmic scale). For very low acceptance rates we observe large fluctuations whereas the Bayes factor is quite stable for larger values. Note that $A_\epsilon \leq 0.005$ corresponds to ≤500 simulations, too small a sample size for robust statistical model estimation.

adjustments obtained by ABC–GLM or ABC–REG should not be trusted without further validation if the Kolmogorov–Smirnov statistic for the GLM model fit exceeds a value of, say, 0.10. In that case linear models are not sufficiently flexible to account for effects like nonlinearity in the parameters and nonnormality and heteroscedasticity in the error terms. In the setting of ABC–REG a wider class of models is introduced in BLUM and FRANCOIS (2009), where machine-learning algorithms are applied for the parameter estimations. Whether these extensions can be applied in our context remains to be seen. The advantage of the general linear model is that estimations can be done with ordinary least squares and the important quantities like marginal posteriors and marginal likelihoods can be obtained analytically. For more complex models these quantities will probably be accessible only via numerical integration, Monte Carlo methods, etc.

Application to chimpanzees: In standard taxonomies, chimpanzees, the closest living relatives of humans, are classified into two species: the common chimpanzee (*Pan troglodytes*) and the bonobo (*P. paniscus*). Both species are restricted to Africa and diverged ∼9 MYA (WON and HEY 2005; BECQUET and PRZEWORSKI 2007). The common chimpanzees are further subdivided into three large populations or subspecies on the basis of their separation by geographic barriers. Among them, the western chimpanzees (*P. troglodytes verus*) form the most remote group. Interestingly, recent multilocus studies found consistent levels of gene flow between the western and the central (*P. t. troglodytes*) chimpanzees (WON and HEY 2005; BECQUET and PRZEWORSKI 2007). Nonetheless, a recent study of 310 microsatellites in 84 common chimpanzees supports a clear distinction between the previously labeled populations (BECQUET *et al.* 2007). Using a PCA analysis, indication for substructure within the western chimpanzees was found in the same study.

To demonstrate the applicability of the model selection given in the THEORY section we contrast two different models of the western chimpanzee population with this data set: a model of a single panmictic population with constant size and a finite island model of constant size and constant migration among demes. While we estimated $\theta = 2N_e\mu$, priors were set on N_e and μ separately with $\log_{10}(N_e) \sim \text{Unif}([3, 5])$ and $\mu \sim N(5 \times 10^{-4}, 2 \times 10^{-4})$ truncated on $\mu \in [10^{-4}, 10^{-3}]$. In the case of the finite island model, we had an additional prior $n_{pop} \sim \text{Unif}([10, 100])$ on the number of islands, and individuals were attributed randomly to the different islands.

We obtained genotypes for all 50 individuals reported to be of western chimpanzee origin from the study of BECQUET *et al.* (2007), excluding captive-born hybrids. We checked the proposed (BECQUET *et al.* 2007) mutation pattern for each individual locus, and all alleles not matching the assumed stepwise mutation model were set as missing data. A total of 265 loci were used, after removing the loci on the X and the Y chromosome as well as those being monomorphic among the western chimpanzees. All simulations were performed using the software SIMCOAL2 (LAVAL and EXCOFFIER 2004) and we reproduced the pattern of missing data observed in the data set. Using the software package Arlequin3.0 (EXCOFFIER *et al.* 2005), we calculated two summary statistics on the data set: the average number of alleles per locus, K, and F_{IS}, the fixation index within the western chimpanzees. We performed a total of 100,000 simulations per model.

In Figure 3 we report the Bayes factor of the island model according to (16) for different acceptance rates A_ϵ; see (14). While there is a large variation for very small acceptance rates, the Bayes factor stabilizes for $A_\epsilon \geq 0.005$. Note that $A_\epsilon \leq 0.005$ corresponds to <500 simulations and that the ABC–GLM approach, based on a model estimation and a smoothing step, is expected to

produce poor results since the estimation of the model parameters is unreliable due to the small sample size. The good news is that the Bayes factor is stable over a large range of tolerance values. We may therefore safely reject the panmictic population model in favor of population subdivision among western chimpanzees with a Bayes factor of $B \approx 10^5$.

DISCUSSION

Due to still increasing computational power it is nowadays possible to tackle estimation problems in a Bayesian framework for which analytical calculation of the likelihood is inhibited. In such cases, approximate Bayesian computation is often the choice. A key innovation in speeding up such algorithms was the use of a regression adjustment, termed ABC–REG in this article, which used the frequently present linear relationship between generated summary statistics $\mathbf{s}$ and parameters of the model $\boldsymbol{\theta}$ in a neighborhood of the observed summary statistics $\mathbf{s}_{\text{obs}}$ (Beaumont *et al.* 2002). The main advantage is that larger tolerance values ϵ still allow us to extract reasonable information about the posterior distribution $\pi(\boldsymbol{\theta}|\mathbf{s})$ and hence less simulations are required to estimate the posterior density.

Here we present a new approach to estimate approximate posterior distributions, termed ABC–GLM, similar in spirit to ABC–REG, but with two major advantages: First, by using a GLM to estimate the likelihood function, ABC–GLM is always consistent with the prior distribution. Second, while we do not find the ABC–GLM approach to substantially outperform ABC–REG in standard situations, it is naturally embedded into a standard Bayesian framework, which in turn allows the application of well-known Bayesian methodologies such as model averaging or model selection via Bayes factors. Our simulations show that the rejection step is especially beneficial if the true model is nonlinear for both ABC approaches. ABC–GLM is further compatible with any type of ABC sampler, including likelihood-free MCMC (Marjoram *et al.* 2003) or population Monte Carlo (Beaumont *et al.* 2009). Also, more complicated regression regimes taking nonlinearity or heteroscedacity into account may be envisioned when the GLM is replaced by some more complex model. A great advantage of the current GLM setting is its simplicity, which renders implementation in standard statistical packages feasible.

We showed the applicability of the model selection procedure via Bayes factors by opposing two different models of population structure among the western chimpanzees *P. troglodytes verus*. Our analysis strongly suggests population substructure within the western chimpanzees since an island model is significantly favored over a model of a panmictic population. While none of our simple models is thought to mimic the real setting exactly, we still believe that they capture the main characteristics of the demographic history influencing our summary statistics, namely the number of alleles K and the fixation index F_{IS}. While the observed F_{IS} of 2.6% has been attributed to inbreeding previously (Becquet *et al.* 2007), we propose that such values may easily arise if diploid individuals are sampled in a randomly scattered way over a large, substructured population. While it was almost impossible to simulate the value $F_{\text{IS}} = 2.6\%$ in the model of a panmictic population, it easily falls within the range of values obtained from an island model.

We are grateful to Laurent Excoffier, David J. Balding, Christian P. Robert, and the anonymous referees for their useful comments on a first draft of this manuscript. This work has been supported by grant no. 3100A0-112072 from the Swiss National Foundation to Laurent Excoffier.

LITERATURE CITED

Beaumont, M., 2007 *Simulations, Genetics, and Human Prehistory—A Focus on Islands*. McDonald Institute Monographs, University of Cambridge, Cambridge, UK.

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

Beaumont, M., R. C. Cornuet and J.-M. Marin, 2009 Adaptive approximate Bayesian computation. Biometrika (in press).

Becquet, C., and M. Przeworski, 2007 A new approach to estimate parameters of speciation models with application to apes. Genome Res. **17:** 1505–1519.

Becquet, C., N. Patterson, A. Stone, M. Przeworski and D. Reich, 2007 Genetic structure of chimpanzee populations. Genome Res. **17:** 1505–1519.

Blum, M., and O. Francois, 2009 Non-linear regression models for approximate Bayesian computation. Stat. Comput. (in press).

Estoup, A., M. Beaumont, F. Sennedot, C. Moritz and J. M. Cornuet, 2004 Genetic analysis of complex demographic scenarios: spatially expanding populations of the cane toad, Bufo marinus. Evolution **58:** 2021–2036.

Excoffier, L., G. Laval and S. Schneider, 2005 Arlequin (version 3.0): an integrated software package for population genetics data analysis. Evol. Bioinform. Online **1:** 47–50.

Greene, W., 2003 *Econometric Analysis*, Ed. 5. Pearson Education, Upper Saddle River, NJ.

Hamilton, G., M. Stoneking and L. Excoffier, 2006 Molecular analysis reveals tighter social regulation of immigration in patrilocal populations than in matrilocal populations. Proc. Natl. Acad. Sci. USA **102:** 7476–7480.

Laval, G., and L. Excoffier, 2004 Simcoal 2.0: a program to simulate genomic diversity over large recombining regions in a subdivided population with a complex history. Bioinformatics **20:** 2485–2487.

Lindley, D., and A. Smith, 1972 Bayes estimates for the linear model. J. R. Stat. Soc. B **34:** 1–44.

Marjoram, P., and S. Tavaré, 2006 Modern computational approaches for analysing molecular genetic variation data. Nat. Rev. Genet. **10:** 759–770.

Marjoram, P., J. Molitor, V. Plagnol and S. Tavaré, 2003 Markov chain Monte Carlo without likelihoods. Proc. Natl. Acad. Sci. USA **100:** 15324–15328.

Sisson, S., Y. Fan and M. Tanaka, 2007 Sequential Monte Carlo without likelihoods. Proc. Natl. Acad. Sci. USA **104:** 1760–1765.

Tallmon, D. A., G. Luikart and M. A. Beaumont, 2004 Comparative evaluation of a new effective population size estimator based on approximate Bayesian computation. Genetics **167:** 977–988.

Tavaré, S., D. Balding, R. Griffiths and P. Donnelly, 1997 Inferring coalescence times from DNA sequence data. Genetics **145:** 505–518.
Watterson, G., 1975 Number of segregating sites in genetic models without recombination. Theor. Popul. Biol. **7:** 256–276.
Wegmann, D., C. Leuenberger and L. Excoffier, 2009 Efficient approximate Bayesian computation coupled Markov chain Monte Carlo without likelihood. Genetics **182:** 1207–1218.
Weiss, G., and A. von Haeseler, 1998 Inference of population history using a likelihood approach. Genetics **149:** 1539–1546.
Won, Y., and J. Hey, 2005 Divergence population genetics of chimpanzees. Mol. Biol. Evol. **22:** 297–307.
Zellner, A., 1971 *An Introduction to Bayesian Inference in Econometrics.* Wiley, New York.

APPENDIX A: PROOFS OF THE MAIN FORMULAS

To keep this article self-contained, we present a proof of formulas (11) and (15). The argument is an adaptation from the proof of Lemma 1 in Lindley and Smith (1972). By linearity it clearly suffices to show the formulas for one fixed sampled parameter $\boldsymbol{\theta}^j$. The results then follow.

Theorem. *Suppose that, given the parameter vector* $\boldsymbol{\theta}$*, the distribution of the statistics vector* $\mathbf{s}$ *is multivariate normal,*

$$\mathbf{s} \sim \mathcal{N}(\mathbf{C}\boldsymbol{\theta} + \mathbf{c}_0, \boldsymbol{\Sigma}_s),$$

and, given the fixed parameter vector $\boldsymbol{\theta}^j$*, the distribution of the parameter* $\boldsymbol{\theta}$ *is*

$$\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\theta}^j, \boldsymbol{\Sigma}_\theta).$$

Then:

1. *The distribution of* $\boldsymbol{\theta}$ given $\mathbf{s}$ *is*

$$\boldsymbol{\theta} \,|\, \mathbf{s} \sim \mathcal{N}(\mathbf{T}\mathbf{v}^j, \mathbf{T}),$$

where $\mathbf{T} = \left(\mathbf{C}^t\boldsymbol{\Sigma}_s^{-1}\mathbf{C} + \boldsymbol{\Sigma}_\theta^{-1}\right)^{-1}$ *and* $\mathbf{v}^j = \mathbf{C}^t\boldsymbol{\Sigma}_s^{-1}(\mathbf{s} - \mathbf{c}_0) + \boldsymbol{\Sigma}_\theta^{-1}\boldsymbol{\theta}^j$.

2. *The marginal distribution of* $\mathbf{s}$ *is*

$$\mathbf{s} \sim \mathcal{N}(\mathbf{m}^j, \mathbf{D}),$$

where $\mathbf{m}^j = \mathbf{c}_0 + \mathbf{C}\boldsymbol{\theta}^j$ *and* $\mathbf{D} = \boldsymbol{\Sigma}_s + \mathbf{C}\boldsymbol{\Sigma}_\theta\mathbf{C}^t$.

Proof. By Bayes' theorem

$$\pi(\boldsymbol{\theta} \,|\, \mathbf{s}) \propto f(\mathbf{s} \,|\, \boldsymbol{\theta})\pi(\boldsymbol{\theta}).$$

The product on the right-hand side is of the form $\exp(-\frac{1}{2}Q)$, where

$$\begin{aligned}
Q &= (\mathbf{s} - \mathbf{c}_0 - \mathbf{C}\boldsymbol{\theta})^t\boldsymbol{\Sigma}_s^{-1}(\mathbf{s} - \mathbf{c}_0 - \mathbf{C}\boldsymbol{\theta}) + (\boldsymbol{\theta} - \boldsymbol{\theta}^j)^t\boldsymbol{\Sigma}_\theta^{-1}(\boldsymbol{\theta} - \boldsymbol{\theta}^j) \\
&= \boldsymbol{\theta}^t(\mathbf{C}^t\boldsymbol{\Sigma}_s^{-1}\mathbf{C} + \boldsymbol{\Sigma}_\theta^{-1})\boldsymbol{\theta} - 2((\mathbf{s} - \mathbf{c}_0)^t\boldsymbol{\Sigma}_s^{-1}\mathbf{C}\boldsymbol{\theta} + (\boldsymbol{\theta}^j)^t\boldsymbol{\Sigma}_\theta^{-1})\boldsymbol{\theta} + \ \ldots \\
&\quad \ldots \ + (\mathbf{s} - \mathbf{c}_0)^t\boldsymbol{\Sigma}_s^{-1}(\mathbf{s} - \mathbf{c}_0) + (\boldsymbol{\theta}^j)^t\boldsymbol{\Sigma}_\theta^{-1}\boldsymbol{\theta}^j \\
&= \boldsymbol{\theta}^t\mathbf{T}^{-1}\boldsymbol{\theta} - 2(\mathbf{v}^j)^t\boldsymbol{\theta} + (\mathbf{s} - \mathbf{c}_0)^t\boldsymbol{\Sigma}_s^{-1}(\mathbf{s} - \mathbf{c}_0) + (\boldsymbol{\theta}^j)^t\boldsymbol{\Sigma}_\theta^{-1}\boldsymbol{\theta}^j \\
&= (\boldsymbol{\theta} - \mathbf{T}\mathbf{v}^j)^t\mathbf{T}^{-1}(\boldsymbol{\theta} - \mathbf{T}\mathbf{v}^j) - (\mathbf{v}^j)^t\mathbf{T}\mathbf{v}^j + \ \ldots \\
&\quad \ldots \ + (\boldsymbol{\theta}^j)^t\boldsymbol{\Sigma}_\theta^{-1}\boldsymbol{\theta}^j + (\mathbf{s} - \mathbf{c}_0)^t\boldsymbol{\Sigma}_s^{-1}(\mathbf{s} - \mathbf{c}_0).
\end{aligned}$$

In the last step we completed the square with respect to $\boldsymbol{\theta}$ and used the fact that $\mathbf{T}$ is symmetric. Up to a constant that does not depend on $\boldsymbol{\theta}^j$ we hence get

$$\pi(\boldsymbol{\theta} \,|\, s) \propto c(\boldsymbol{\theta}^j)\exp\left(-\frac{1}{2}((\boldsymbol{\theta} - \mathbf{T}\mathbf{v}^j)^t\mathbf{T}^{-1}(\boldsymbol{\theta} - \mathbf{T}\mathbf{v}^j))\right),$$

where $c(\boldsymbol{\theta}^j) = \exp(-\frac{1}{2}((\boldsymbol{\theta}^j)^t\boldsymbol{\Sigma}_\theta^{-1}\boldsymbol{\theta}^j - (\mathbf{v}^j)^t\mathbf{T}\mathbf{v}^j))$. This proves the first part of the theorem and—by linear superposition—the validity of Equation 11.

To prove the second part of the theorem, observe that $\mathbf{s} = \mathbf{C}\boldsymbol{\theta} + \mathbf{c}_0 + \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \Sigma_s)$ and $\boldsymbol{\theta} = \boldsymbol{\theta}^j + \boldsymbol{\eta}$ with $\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \Sigma_\theta)$. Putting these equalities together, we get

$$\mathbf{s} \sim \mathbf{C}\boldsymbol{\theta}^j + \mathbf{c}_0 + \mathbf{C}\boldsymbol{\eta} + \boldsymbol{\epsilon}.$$

This, being a linear combination of independent multivariate normal variables, is still multivariate normal with mean $\mathbf{C}\boldsymbol{\theta}^j + \mathbf{c}_0$ and its covariance matrix is given by

$$E[(\mathbf{C}\boldsymbol{\eta} + \boldsymbol{\epsilon})(\mathbf{C}\boldsymbol{\eta} + \boldsymbol{\epsilon})^t] = E[\mathbf{C}\boldsymbol{\eta}(\mathbf{C}\boldsymbol{\eta})^t + \boldsymbol{\epsilon}\boldsymbol{\epsilon}^t] = \mathbf{C}E[\boldsymbol{\eta}\boldsymbol{\eta}^t]\mathbf{C}^t + E[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^t] = \mathbf{C}\boldsymbol{\Sigma}_\theta\mathbf{C}^t + \boldsymbol{\Sigma}_s.$$

This proves the second part of the theorem as well as formula (15). ■

APPENDIX B: NONLINEAR TOY MODELS

In this section we describe a class of toy models that are nonlinear in the parameter $\theta \in \mathbb{R}$ and have nonnormal, possibly heteroscedastic error terms. Still their likelihoods are easy to calculate analytically. We set

$$\mathbf{s} = \mathbf{f}(\theta) + \boldsymbol{\epsilon}(\theta) = \begin{pmatrix} f_1(\theta) \\ \vdots \\ f_n(\theta) \end{pmatrix} + \begin{pmatrix} \epsilon_1(\theta) \\ \vdots \\ \epsilon_n(\theta) \end{pmatrix}.$$

Here $f_i(\theta)$ are monotonically increasing continuous functions of θ and $\epsilon_i(\theta)$ are independent, uniformly distributed error terms in the interval $[-u_i(\theta),\ u_i(\theta)] \subseteq \mathbb{R}$, where $u_i(\theta)$ are nondecreasing, continuous functions:

$$\epsilon_i(\theta) \sim \mathrm{Unif}([-u_i(\theta),\ u_i(\theta)]).$$

It is straightforward to check that for a prior $\pi(\theta)$ the posterior distribution of θ given $\mathbf{s} = (s_1, \ldots, s_n)^t$ is (up to a normalizing constant)

$$\pi(\theta \mid \mathbf{s}) \propto \frac{1}{u_1(\theta) \cdot \ldots \cdot u_n(\theta)} \mathrm{Ind}([\theta_{\min}, \theta_{\max}])\pi(\theta),$$

where

$$\theta_{\min} = \max_i \{g_i^{-1}(s_i)\}, \quad \theta_{\max} = \min_i \{h_i^{-1}(s_i)\}$$

and

$$g_i(\theta) = f_i(\theta) + u_i(\theta), \quad h_i(\theta) = f_i(\theta) - u_i(\theta).$$

For the simulations in Table 2 we chose $n = 5$, $f_1(\theta) = \cdots = f_5(\theta) = \theta^3$, and $u_1(\theta) = \cdots = u_5(\theta) \equiv 10$. The prior was $\pi(\theta) \sim \mathcal{N}(0, 4)$.