Bayesian Analysis (2011) **6**, Number 1, pp. 49–76

Likelihood-free estimation of model evidence

Xavier Didelot*, Richard G. Everitt†, Adam M. Johansen‡ and Daniel J. Lawson§

Abstract. Statistical methods of inference typically require the likelihood function to be computable in a reasonable amount of time. The class of "likelihood-free" methods termed Approximate Bayesian Computation (ABC) is able to eliminate this requirement, replacing the evaluation of the likelihood with simulation from it. Likelihood-free methods have gained in efficiency and popularity in the past few years, following their integration with Markov Chain Monte Carlo (MCMC) and Sequential Monte Carlo (SMC) in order to better explore the parameter space. They have been applied primarily to estimating the parameters of a given model, but can also be used to compare models.

Here we present novel likelihood-free approaches to model comparison, based upon the independent estimation of the evidence of each model under study. Key advantages of these approaches over previous techniques are that they allow the exploitation of MCMC or SMC algorithms for exploring the parameter space, and that they do not require a sampler able to mix between models. We validate the proposed methods using a simple exponential family problem before providing a realistic problem from human population genetics: the comparison of different demographic models based upon genetic data from the Y chromosome.

1 Introduction

Let x_{obs} denote some observed data. If a model M has parameter θ, with $p(\theta|M)$ denoting the prior and $p(x_{\text{obs}}|M,\theta)$ the likelihood, then the model evidence (also termed marginal likelihood or integrated likelihood) is defined as:

$$p(x_{\text{obs}}|M) = \int p(x_{\text{obs}}|M,\theta)p(\theta|M)\mathrm{d}\theta. \tag{1}$$

To compare two models M_1 and M_2 one may compute the ratio of evidence of two models, called the Bayes Factor (Kass and Raftery 1995; Robert 2001):

$$B_{1,2} = \frac{p(x_{\text{obs}}|M_1)}{p(x_{\text{obs}}|M_2)}. \tag{2}$$

In particular, if we assign equal prior probabilities to the two models M_1 and M_2,

*Department of Statistics, University of Oxford, UK, mailto:didelot@stats.ox.ac.uk
†Department of Mathematics, University of Bristol, UK, mailto:richard.everitt@bristol.ac.uk
‡Department of Statistics, University of Warwick, UK, mailto:a.m.johansen@warwick.ac.uk
§Department of Mathematics, University of Bristol, UK, mailto:dan.lawson@bristol.ac.uk

 DOI:10.1214/11-BA602

then their posterior odds ratio is equal to the Bayes Factor:

$$\frac{p(M_1|x_{\text{obs}})}{p(M_2|x_{\text{obs}})} = \frac{p(x_{\text{obs}}|M_1)}{p(x_{\text{obs}}|M_2)} \frac{p(M_1)}{p(M_2)} = \frac{p(x_{\text{obs}}|M_1)}{p(x_{\text{obs}}|M_2)}. \tag{3}$$

Jeffreys (1961) gave the following qualitative interpretation of a Bayes Factor: *1 to 3 is barely worth a mention, 3 to 10 is substantial, 10 to 30 is strong, 30 to 100 is very strong and over a 100 is decisive evidence in favor of model* M_1*. Values below 1 take the inverted interpretation in favor of model* M_2*.*

The many approaches to estimating Bayes Factors can be divided into two classes: those that estimate a Bayes Factor without computing each evidence independently and those that do involve such an explicit calculation. Without exhaustively enumerating these approaches, it is useful to mention those which are of particular relevance in the present context. In the first category we find the reversible jump technique of Green (1995), as well as the methods of Stephens (2000) and Dellaportas et al. (2002). In the second category we find the harmonic mean estimator of Newton and Raftery (1994) and its variations, the method of Chib (1995), the annealed importance sampling estimator of Neal (2001) and the power posteriors technique of Friel and Pettitt (2008).

Here we present a method for estimating the evidence of a model when the likelihood $p(x_{\text{obs}}|M, \theta)$ is not available in the sense that it either cannot be evaluated or such evaluation is prohibitively expensive. This difficulty arises frequently in a wide range of applications, for example in population genetics (Beaumont et al. 2002) or epidemiology (Luciani et al. 2009).

2 Background

2.1 Approximate Bayesian Computation for Parameter Estimation

Approximate Bayesian Computation is the name given to techniques which avoid evaluation of the likelihood by simulation of data from the associated model. The main focus of ABC has been the estimation of model parameters and we begin with a survey of the basis of these methods and the various computational algorithms which have been developed for their implementation.

Basic ABC algorithm

When dealing with posterior distributions that are sufficiently complex that calculations cannot be performed analytically, it has become common place to invoke Monte Carlo approaches: drawing samples which can be used to approximate the posterior distribution and using that sample approximation to calculate quantities of interest. One of the simplest methods of sampling from a posterior distribution $p(\theta|x_{\text{obs}})$ is to use rejection sampling, drawing samples from the prior distribution and accepting them with probability proportional to their likelihood. This, however, requires the explicit evaluation of the likelihood $p(x_{\text{obs}}|\theta)$ for every simulated parameter value. Representing

the likelihood as a degenerate integral:

$$p(x_{\text{obs}}|\theta) = \int p(x|\theta)\delta_{x_{\text{obs}}}(\mathrm{d}x),$$

suggests that it could be approximated by replacing the singular mass at x_{obs} with a continuous distribution (or a less concentrated discrete distribution in the case of discrete observations) to obtain the approximation:

$$\hat{p}(x_{\text{obs}}|\theta) = \int p(x|\theta)\pi_\epsilon(x|x_{\text{obs}})\mathrm{d}x, \tag{4}$$

where $\pi_\epsilon(x|x_{\text{obs}})$ is a normalized kernel (i.e. a probability density with respect to the same measure as $p(x|\theta)$) centered on x_{obs} and with a degree of concentration determined by ϵ.

The approximation in Equation 4 admits a Monte Carlo approximation that is unbiased (in the sense that no further bias is introduced by the use of this additional step). If $X \sim p(x|\theta)$ then the expectation of $\pi_\epsilon(X|x_{\text{obs}})$ is exactly $\hat{p}(x_{\text{obs}}|\theta)$. One can view this approximation in the following intuitive way:

$$\begin{aligned}
\mathbf{E}_{X\sim p(x|\theta)}(\pi_\epsilon(X|x_{\text{obs}})) &= \int \pi_\epsilon(x|x_{\text{obs}})p(x|\theta)\mathrm{d}x \\
&= \mathbf{E}_{X\sim\pi_\epsilon(x|x_{\text{obs}})}(p(X|\theta)) \\
&\approx p(x_{\text{obs}}|\theta) \text{ when } \epsilon \text{ is small.}
\end{aligned} \tag{5}$$

This approximate equality holds in the sense that under weak regularity conditions, for sufficiently-small, positive ϵ the error due to the approximation is a small and monotonically decreasing function of ϵ which converges as $\epsilon \downarrow 0$. Using this approximation in place of the likelihood in the rejection sampling algorithm above results in the basic Approximate Bayesian Computation (ABC) algorithm:

Algorithm 1.

---

1. Generate $\theta^* \sim p(\theta)$
2. Simulate $x^* \sim p(x|\theta^*)$
3. Accept θ^* with probability proportional to $\pi_\epsilon(x^*|x_{\text{obs}})$ otherwise return to step 1

The ABC algorithm was first described in this exact form by Pritchard et al. (1999) although similar approaches were previously discussed by (Tavaré et al. 1997; Fu and

Li 1997; Weiss and von Haeseler 1998). Here and below we assume that the full data x_{obs} is used in the inference. It is usually necessary in real inference problems to make use of summary statistics (Pritchard et al. 1999) which we discuss in Section 3.2 in a model comparison context.

If $\pi_\epsilon(x|x_{\text{obs}})\mathrm{d}x$ places probability 1 on $\{x_{\text{obs}}\}$ then the algorithm is exact, but the acceptance probability is zero unless the data is discrete. Indeed, the above representation of the ABC procedure only admits the exact case as a limit when dealing with continuous data (the Dirac measure admits no Lebesgue density). Any other choice of kernel results in an algorithm producing samples from an approximation of the posterior distribution $p(\theta|x_{\text{obs}})$. For example, Pritchard et al. (1999) and many later applications used a locally uniform density

$$\pi_\epsilon(x|x_{\text{obs}}) \propto \begin{cases} 1 & \text{if } D(x, x_{\text{obs}}) < \epsilon \\ 0 & \text{otherwise,} \end{cases} \tag{6}$$

where $D(\cdot,\cdot)$ is some metric and ϵ is a (small) tolerance value. Other choices for $\pi_\epsilon(x|x_{\text{obs}})$ are discussed in Beaumont et al. (2002). It is interesting to note that the use of such an approximate kernel π_ϵ in the ABC algorithm can be interpreted as exact sampling under a model where uniform additive error terms exist (Wilkinson 2008).

ABC-MCMC

Markov Chain Monte Carlo (MCMC, Gilks and Spiegelhalter 1996; Robert and Casella 2004) methods are a family of simulation algorithms intended to provide sequences of dependent samples which are marginally distributed according to a distribution of interest. Application of ergodic theory and central limit theorems justifies the use of these sample sequences to approximate integrals with respect to that distribution. MCMC is often considered in situations in which more elementary Monte Carlo techniques, such as rejection sampling, are unable to provide sufficiently efficient simulation. In the ABC context, if the likelihood is sharply peaked relative to the prior, then the rejection sampling algorithm described previously is likely to suffer from an extremely low acceptance rate. MCMC algorithms intended to improve the efficiency of ABC-based approximations have been developed. In particular, Marjoram et al. (2003) proposed the incorporation of the ABC approximation of Equation 4 into an MCMC algorithm.

This algorithm, like any standard Metropolis-Hastings algorithm, requires a mutation kernel Q to propose new values of the parameters given the current values and accepts them with appropriate probability to ensure that the invariant distribution of the Markov chain is preserved. This algorithm can be interpreted as a standard Metropolis-Hastings algorithm on an extended space. It involves simulating a Markov chain over the space of both parameters and data, (θ, x), with an invariant distribution proportional to $p(\theta)p(x|\theta)\mathbf{I}_{D(x_{\text{obs}},x)<\epsilon}$ in the usual way. At stationarity, the marginal distribution of θ is proportional to $p(\theta)\hat{p}(x_{\text{obs}}|\theta)$ in the notation of Equations 4 and 6. Marjoram et al. (2003) demonstrated that the stationary distribution of this MCMC algorithm converges in an appropriate sense to the posterior distribution $p(\theta|x_{\text{obs}})$ as $\epsilon \downarrow 0$.

ABC-SMC

The Sequential Monte Carlo sampler (SMC sampler, Del Moral et al. 2006) is another Monte Carlo technique which can be employed to sample from complex distributions. It can provide an alternative to MCMC in some settings. It employs importance sampling and resampling techniques in order to efficiently produce a (weighted) sample from a distribution or sequence of distributions of interest. It is particularly well suited to situations in which successive members of the sequence of distributions are increasingly concentrated.

In the ABC context, it is natural to consider the use of SMC techniques applied to the joint distribution of (θ, x) in the same way as the ABC-MCMC algorithm. A natural sequence of distributions is obtained by considering a decreasing sequence of values of ϵ. Although such an approach may seem computationally costly, it does not require a successful global exploration of the final distribution in order to characterize it well and hence may outperform MCMC in situations in which it is rather difficult to design fast-mixing transition kernels. However, the need to resimulate data sets from the prior during each iteration reduces the benefit which can be obtained in the ABC setting.

Sisson et al. (2007) proposed the integration of the ABC approximation of Section 2.1 within an SMC sampler in the following manner:

Algorithm 2.

---

1. Set $t = 1$. For $i = 1, \ldots, N$, sample $\theta_1^i \sim p(\theta)$ and set $w_1^i = 1/N$.
2. Increment $t = t + 1$. For $i = 1, \ldots, N$
   (a) Generate $\theta_t^i \sim Q_t(\theta|\theta_{t-1}^i)$,
   (b) Simulate $x^* \sim p(x|\theta_t^i)$
   (c) Compute
   $$w_t^i = \frac{p(\theta_t^i)\pi_{\epsilon_t}(x^*|x_{\text{obs}})}{\sum_{j=1}^N Q_t(\theta_t^i|\theta_{t-1}^j)} \tag{7}$$
3. If $t < T$, resample the particles in population t and return to step 2.

Unlike standard SMC algorithms this approach employs a Monte Carlo estimate of an importance weight defined on only the marginal space at the current iteration. Such strategies (which can be justified via Slutzky's lemma, the delta method and appropriate conditioning arguments — see, for example, Shao 1999) have been previously employed

in particle filtering (Klaas et al. 2005) and come at the cost of increasing the computational complexity from $\mathcal{O}(N)$ to $\mathcal{O}(N^2)$. There is no fundamental need to employ such a marginalization and a more standard SMC algorithm could also be considered — this point was made explicitly by Del Moral et al. (2008) who proposed an $\mathcal{O}(N)$ approach and also developed adaptive versions of the algorithm.

These algorithms can be understood in the framework of Del Moral et al. (2006), with appropriate choices of auxiliary kernel. In the case of algorithm 2, the auxiliary kernel is the sample approximation of the optimal kernel first proposed by Peters (2005). In the case of the algorithm of Del Moral et al. (2008), this is the time reversal kernel associated with the MCMC kernel, with the selection and mutation steps exchanged because the importance weight at time t depends only upon the sample at time $t-1$ when this approximation is employed.

2.2 Existing methods for likelihood-free model selection

The ABC techniques described so far were designed to infer the parameters of a given model. Methods to test the fit of a model without explicit comparison to other models (i.e. Bayesian model criticism) have been proposed by Thornton and Andolfatto (2006) who computed posterior predictive p-values (Meng 1994), and by Ratmann et al. (2009) who extended a model with additional error terms, the posterior distributions of which indicate how good the fit is. Model criticism and assessment of goodness-of-fit are important in their own right, but there are situations in which explicit comparison of the models using Bayes Factors is desirable (Robert et al. 2010) and the idea of using ABC in this context dates back to at least Wilkinson (2007).

When the two models that we wish to compare are nested, the basic ABC algorithm and its MCMC and SMC extensions can be used directly to estimate a Bayes Factor. This is achieved by performing inference under the larger model, but placing half of the prior weight on the subspace of the full parameter space which corresponds to the simpler model. This technique was first used by Pritchard et al. (1999) to compare a population genetics model in which the population size grows exponentially at rate $r > 0$ with the model with $r = 0$.

In order to compute the Bayes Factor of two models M_1 and M_2 with parameters θ_1 and θ_2, Grelaud et al. (2009) considered the model M with parameters (m, θ_1, θ_2) where m is *a priori* uniformly distributed in $\{1, 2\}$, $\theta_1 = 0$ when $m = 2$ and $\theta_2 = 0$ when $m = 1$. In this way, both models M_1 and M_2 are nested within model M and each has equal prior weight 0.5 in model M.

Algorithm 3.

1. Set $M^* = M_1$ with probability 0.5, otherwise set $M^* = M_2$
2. Generate $\theta^* \sim p(\theta|M^*)$
3. Simulate $x^* \sim p(x|\theta^*, M^*)$
4. Accept (M^*, θ^*) if $D(x, x_{\text{obs}}) < \epsilon$ otherwise return to step 1

The ratio of the number of accepted samples for which $M = M_1$ to those for which $M = M_2$ when the above algorithm is run many times is an estimator of the Bayes Factor between models M_1 and M_2. One drawback of this algorithm is that it is based on the ABC rejection sampling algorithm and does not take advantage of the improved exploration of the parameter space available in the ABC-MCMC or ABC-SMC algorithms. Toni et al. (2009) proposed an ABC-SMC algorithm to compute the Bayes Factor of models once again by considering a metamodel in which all models of interest are nested. Here we propose a different approach which is to estimate the evidence of each model separately.

3 Methodology

This section presents an approach to the direct approximation of model evidence, and thus Bayes Factors, within the ABC framework. It is first shown that the standard ABC approach can provide a natural estimate of the normalizing constant that corresponds to the evidence of each model, and then algorithms based around the strengths of MCMC and SMC implementation are presented. The choice of summary statistics when applying ABC-based algorithms to the problem of model selection is then discussed.

3.1 Estimation of model evidence

Just as in the standard parameter estimation problem, the following ABC approach to the estimation of model evidence is based around a simple approximation. This approximation can be dealt with directly via a rejection sampling argument which subject to certain additional constraints leads to the approach advocated by Grelaud et al. (2009). Considering a slightly more general framework and casting the problem as that of estimating an appropriate normalizing constant allows the use of other sampling methods based around the same distributions.

Basic ABC setting

When the likelihood is available, the model evidence can be estimated using importance sampling. Let $q(\theta)$ be a distribution of known density over the parameter θ which dominates the prior distribution and from which it is possible to sample efficiently. Using the standard importance sampling identity, the evidence can be rewritten as follows:

$$\begin{aligned} p(x_{\text{obs}}) &= \int p(x_{\text{obs}}|\theta)p(\theta)\mathrm{d}\theta = \int \frac{p(x_{\text{obs}}|\theta)p(\theta)}{q(\theta)} q(\theta)\mathrm{d}\theta \\ &\approx \frac{1}{N}\sum_{i=1}^{N} \frac{p(x_{\text{obs}}|\theta_i)p(\theta_i)}{q(\theta_i)} \text{ with } \theta_i \sim q(\theta), \end{aligned} \tag{8}$$

where $w(\theta_i) = \frac{p(x_{\text{obs}}|\theta_i)p(\theta_i)}{q(\theta_i)}$ is termed the weight of θ_i and Equation 8 shows that the evidence can be estimated by the empirical mean of the weights obtained by drawing a collection of samples from q. This approach provides an unbiased estimate of the evidence but requires the evaluation of the importance weights including the values of the likelihood.

When the likelihood is not available, we can use the ABC approximation of Equation 4 in place of the likelihood in Equation 8 to obtain the following algorithm:

Algorithm 4.

---

1. `For` $i = 1, \ldots, N$
   (a) `Generate` $\theta_i \sim q(\theta)$
   (b) `Simulate` $x_i \sim p(x|\theta_i)$
   (c) `Compute` $w_i = \frac{\pi_\epsilon(x_i|x_{\text{obs}})p(\theta_i)}{q(\theta_i)}$
2. `Return` $\frac{1}{N}\sum_{i=1}^{N} w_i$

The average of the importance weights is an unbiased estimator of the normalising constant associated with the ABC approximation to the posterior; it is shown in the appendix that this converges as $\epsilon \downarrow 0$ to the marginal likelihood under mild continuity conditions. In principle, the algorithm above can be used with any proposal distribution which dominates the true distribution; in order to control the variance of the importance weights it is desirable that the proposal should have tails at least as heavy as those of the target. One possibility is to use the prior $p(\theta)$ as proposal distribution. In this case, the algorithm above becomes similar to the ABC rejection sampling algorithm and the weights simplify into $w_i = \pi_\epsilon(x_i|x_{\text{obs}})$. If $\pi_\epsilon(x|x_{\text{obs}})$ is taken to be an indicator function

as in Equation 6, then the result of the algorithm above is simply equal to the proportion of accepted values times the normalizing constant of π_ϵ. If this algorithm is applied to two models M_1 and M_2, then the Bayes Factor $B_{1,2}$ is approximated by the ratio of the number of accepted values under each model which is equivalent to algorithm 3.

This approach suffers from the usual problem of importance sampling from a posterior using proposals generated according to a prior distribution (Kass and Raftery 1995). If the posterior is concentrated relative to the prior, most of the weights will be very small. In the ABC context this phenomenon exhibits itself in a particular form: the θ_i will have small probabilities of generating an x_i similar to x_{obs} and therefore most of the weights w_i will be small. Thus the estimate will be dominated by a few larger weights where θ_i happened to be simulated from a region of higher posterior value, and therefore the estimate of the evidence will have a large variance. Such a problem is well known when performing importance sampling generally (Liu 2001). In the scenario in which the likelihood is known this problem can be dealt with by employing an approximation of the optimal proposal distribution (see, for example, Robert and Casella 2004). Unfortunately, it is not straightforward to do so in the ABC context. To avoid this issue, we show how the algorithm above can be applied to take advantage of the improvements in parameter space exploration introduced by ABC-MCMC and ABC-SMC.

Working with an approximate posterior sample

Let $\theta_1, \ldots, \theta_N$ denote a sample approximately drawn from $p(\theta|x_{\text{obs}}, M)$, for example the output from the ABC-MCMC algorithm. Let Q denote a mutation kernel, let θ_i^* be the result of applying Q to θ_i and let $q(\theta)$ denote the resulting distribution of the θ_i^*. Then a Monte Carlo approximation of the unknown marginal proposal distribution, $q(\theta)$, is given by:

$$q(\theta) \approx \frac{1}{N} \sum_{j=1}^{N} Q(\theta|\theta_j). \tag{9}$$

Using this proposal distribution $q(\theta)$ in algorithm 4 together with the estimate above for its density leads to the following algorithm to estimate the evidence $p(x_{\text{obs}}|M)$:

Algorithm 5.

1. For $i = 1, \ldots, N$
   (a) Generate $\theta_i^* \sim Q(\theta|\theta_i)$
   (b) Simulate $x_i^* \sim p(x|\theta_i^*)$
   (c) Compute $$w_i = \frac{p(\theta_i^*)\pi_\epsilon(x_i^*|x_{\text{obs}})}{\frac{1}{N}\sum_{j=1}^{N} Q(\theta_i^*|\theta_j)} \tag{10}$$
2. Return $\frac{1}{N}\sum_{i=1}^{N} w_i$

Equation 10 provides a consistent estimate of the exact importance weight. Therefore algorithm 5 is valid in the sense that under standard regularity conditions, it provides a consistent estimate of the ABC approximation of the evidence discussed in the previous section. The kernel Q should be chosen to have heavy tails in order to have heavy tails in the proposal distribution q and thus prevent the variance of the weights from becoming infinite. Note that algorithm 5 is of complexity $\mathcal{O}(N^2)$. We did not find this to be an issue in our applications. In situations where this is too computationally expensive an alternative would be to choose the proposal distribution $q(\theta)$ to be a standard distribution with parameters determined by the moments of the sample $\theta_1, \ldots, \theta_N$. However, this becomes equivalent to an importance sampler with a fine-tuned proposal distribution, which might perform badly in general.

ABC-SMC setting

The ABC-SMC algorithm produces weighted samples suitable for approximating the posterior $p(\theta|x_{\text{obs}}, M)$. These samples could be resampled and algorithm 5 applied to produce an estimate of the evidence. However, like any SMC sampler, the ABC-SMC algorithm produces a natural estimate of the unknown normalizing constant which in the present case is the quantity which we seek to estimate. An indication of this is given by the fact that algorithm 5 takes a very similar form to one step of the ABC-SMC algorithm.

In particular, the weights estimated in Equation 7 of the ABC-SMC algorithm of Sisson et al. (2007) are of the exact same form as those calculated in Equation 10. It is therefore straightforward to obtain an estimate of the evidence (noting that this differs from the MCMC version slightly in that in the SMC case the distribution of the previous sample was intended to target $\pi_{\epsilon_{t-1}}$ rather than π_{ϵ_t}):

$$p(x_{\text{obs}}|M) \approx \frac{1}{N} \sum_{i=1}^{N} w_T^i. \tag{11}$$

In contrast, the ABC-SMC algorithm of Del Moral et al. (2008) allows for estimation of the normalizing constant via the standard estimator:

$$p(x_{\text{obs}}|M) \approx \prod_{t=1}^{T} \frac{1}{N} \sum_{i=1}^{N} w_t^i. \tag{12}$$

Notice that the estimator in Equation 12 employs all of the samples generated within the SMC process, not just those obtained in the final iteration as does Equation 11. That SMC algorithms can produce unbiased estimates of unknown normalizing constants has been noted before (Del Moral 2004; Del Moral et al. 2006).

3.2 Working with summary statistics

Summary statistics in ABC

The ABC algorithms described in Section 2.1 were written as though the full data x_{obs} was being used and compared to simulated data using π_ϵ. In practice this is not often possible because most data is of high dimensionality, and consequently any simulated data is, with high probability, in some respect different from that which is observed. To deal with this difficulty some summary statistic, $s(x_{\text{obs}})$, is often used in place of the full data x_{obs} in the algorithms of Section 2.1, and compared to the corresponding statistics of the simulated data. A first example of this is found in Pritchard et al. (1999).

Sufficient statistics are ubiquitous in statistics, but when considering model comparison it is important to consider precisely what is meant by sufficiency. A summary statistic s is said to be sufficient *for the model parameters* θ if the distribution of the data is independent of the parameters when conditioned on the statistic:

$$p(x|s(x),\theta) = p(x|s(x)). \tag{13}$$

If s is sufficient in this sense, then substituting $s(x)$ for x in the algorithms of Section 2.1 has no effect on the exactness of the ABC approximation (Marjoram et al. 2003). It remains the case that the approximation error can be controlled to any level by choosing sufficiently small ϵ. If the statistics are not sufficient then it introduces an additional layer of approximation. A compromise is required: the simpler and lower the dimension of s the better the performance of the simulation algorithms (Beaumont et al. 2002) but the more severe the approximation.

Summary statistics in ABC for model choice

The algorithms in Section 3.1 intended for the calculation of Bayes Factors have also been written assuming that the full data x_{obs} is being used. For the same reasons as above, this is not always practical and summary statistics often have to be used. If a summary statistic $s(x_{\text{obs}})$ is substituted for the full data x_{obs} in the algorithms of Section 3.1, the result is that they estimate $p(s(x_{\text{obs}})|M)$ instead of the evidence $p(x_{\text{obs}}|M)$.

As $s(x_{\text{obs}})$ is a deterministic function of x_{obs}, the relationship between these two quantities can be written as follows:

$$p(x_{\text{obs}}|M) = p(x_{\text{obs}}, s(x_{\text{obs}})|M) = p(s(x_{\text{obs}})|M)p(x_{\text{obs}}|s(x_{\text{obs}}), M). \tag{14}$$

Unfortunately the last term in Equation 14 is not readily computable in most models of interest. Here we consider the conditions under which this does not affect the estimate of a Bayes Factor. In general, we have:

$$B_{1,2} = \frac{p(x_{\text{obs}}|M_1)}{p(x_{\text{obs}}|M_2)} = \frac{p(s(x_{\text{obs}})|M_1)}{p(s(x_{\text{obs}})|M_2)} \frac{p(x_{\text{obs}}|s(x_{\text{obs}}), M_1)}{p(x_{\text{obs}}|s(x_{\text{obs}}), M_2)}. \tag{15}$$

We say that a summary statistic s is sufficient *for comparing two models, M_1 and M_2*, if and only if the last term in Equation 15 is equal to one, so that:

$$B_{1,2} = \frac{p(s(x_{\text{obs}})|M_1)}{p(s(x_{\text{obs}})|M_2)}. \tag{16}$$

This definition can be readily generalized to the comparison of more than two models. When Equation 16 holds, the algorithms described in Section 3.1 can be applied using $s(x_{\text{obs}})$ in place of x_{obs} for two models M_1 and M_2 to produce an estimate of the Bayes Factor $B_{1,2}$ without introducing any additional approximation.

As was noted by Grelaud et al. (2009), it is important to realize that sufficiency for M_1, M_2 or both (as defined by Equation 13) does not guarantee sufficiency for comparing them (as defined in Equation 16). For instance, consider $x_{\text{obs}} = (x_1, \ldots, x_n)$ where each component is independent and identically distributed. Grelaud et al. (2009) consider models M_1 where $x_i \sim \text{Poisson}(\lambda)$ and M_2 where $x_i \sim \text{Geom}(\mu)$. In this case $s(x) = \sum_{i=1}^{n} x_i$ is sufficient for both models M_1 and M_2, yet $p(x_{\text{obs}}|s(x_{\text{obs}}), M_1) \neq p(x_{\text{obs}}|s(x_{\text{obs}}), M_2)$ and it is apparent that $s(x)$ is not sufficient for comparing the two models.

Finding a summary statistic sufficient for model choice

A generally applicable method for finding a summary statistic s sufficient for comparing two models M_1 and M_2 is to consider a model M in which both M_1 and M_2 are nested. Then any summary statistic sufficient for M (as defined in Equation 13) is sufficient for comparing M_1 and M_2 (as defined in Equation 16):

$$\begin{aligned}
p(x|M_1) &= \int p(x|\theta, M_1)p(\theta|M_1)\mathrm{d}\theta = \int p(x|\theta, M)p(\theta|M_1)\mathrm{d}\theta \\
&= \int p(x|s(x), \theta, M)p(s(x)|\theta, M)p(\theta|M_1)\mathrm{d}\theta \\
&= p(x|s(x), M)\int p(s(x)|\theta, M_1)p(\theta|M_1)\mathrm{d}\theta \\
&= p(x|s(x), M)p(s(x)|M_1).
\end{aligned} \tag{17}$$

Similarly $p(x|M_2) = p(x|s(x), M)p(s(x)|M_2)$ and therefore:

$$\frac{p(x|M_1)}{p(x|M_2)} = \frac{p(x|s(x), M)p(s(x)|M_1)}{p(x|s(x), M)p(s(x)|M_2)} = \frac{p(s(x)|M_1)}{p(s(x)|M_2)}, \tag{18}$$

which means that Equation 16 holds and therefore s is sufficient for comparing M_1 and M_2.

Note that this approach exploits the fact that under these circumstances the problem of model choice becomes one of parameter estimation, albeit in a context in which the

prior distributions take a particular form which may impede standard approaches to computation. Of course, essentially any model comparison problem can be cast in this form.

Summary statistics sufficient for comparing exponential family models

We now consider the case where comparison is made between two models that are both members of the exponential family. In this case, the likelihood under each model $i = \{1, 2\}$ can be written as:

$$p(x|M_i, \theta_i) \propto \exp(s_i(x)^T \theta_i + t_i(x)), \tag{19}$$

where s_i is a vector of sufficient statistics (in the ordinary sense) for model i, θ_i the associated vector of parameters and $t_i(x)$ captures any intrinsic relationship between model i and its data which is not dependent upon its parameters. The $t_i(x)$ terms are important when comparing members of the exponential family which have different base measures: they capture the interaction between the data and the base measure which is independent of the value of the parameters but is important when comparing models. It is precisely this t_i term which prevents statistics sufficient for each model from being adequate for the comparison of the two models.

Consider the extended model M with parameter $(\theta_1, \theta_2, \alpha_1, \alpha_2)$, where θ_1 and θ_2 are as before and $\alpha_i \in \{0, 1\}$, defined via:

$$\begin{aligned} p(x|M, \theta_1, \theta_2, \alpha) &\propto \exp\left(s_1(x)^T \theta_1 + s_2(x)^T \theta_2 + \alpha_1 t_1(x) + \alpha_2 t_2(x)\right) \\ &\propto \exp\left([s_1(x)^T, s_2(x)^T, t_1(x), t_2(x)] \begin{bmatrix} \theta_1 \\ \theta_2 \\ \alpha_1 \\ \alpha_2 \end{bmatrix} \right). \end{aligned} \tag{20}$$

M reduces to M_1 if we take $\theta_2 = 0, \alpha_1 = 1, \alpha_2 = 0$, and M reduces to M_2 if we take $\theta_1 = 0$ and $\alpha_1 = 0, \alpha_2 = 1$. Thus both M_1 and M_2 are nested within M. It is furthermore clear that the model M is an exponential family model for which $S(x) = [s_1(x), s_2(x), t_1(x), t_2(x)]$ is sufficient. Following the argument of Section 3.2, $S(x)$ is a sufficient statistic for the model choice problem between models M_1 and M_2 (as defined by Equation 16). A special case of this result is that the combination of the sufficient statistics of two Gibbs Random Field models is sufficient for comparing them, as previously noted by Grelaud et al. (2009).

4 Applications

4.1 Toy Example

The problem

It is convenient to first consider a simple example in which it is possible to evaluate the evidence analytically in order to validate and compare the performance of the algorithms described. We turn to the example described by Grelaud et al. (2009) in which the observations are assumed to be independent and identically distributed according to a Poisson(λ) distribution in model M_1 and a Geometric(μ) distribution in model M_2 (cf. Section 3.2). The canonical form of the two models (as defined in Equation 19), with n observations, is:

$$p(x|\theta_1, M_1) \propto \exp\left(\sum_{j=1}^{n} x_j\theta_1 - \sum_{j=1}^{n} \log x_j!\right), \tag{21}$$

$$p(x|\theta_2, M_2) \propto \exp\left(\sum_{j=1}^{n} x_j\theta_2\right), \tag{22}$$

where $\theta_1 = \log\lambda$ and $\theta_2 = \log(1-\mu)$ under the usual parametrization. Hence, we can incorporate both in a model of the form:

$$p(x|\theta, \alpha, M) \propto \exp\left((\theta_1+\theta_2)\sum_j x_j + \alpha\sum_j \log x_j!\right). \tag{23}$$

In this particular case θ_1 and θ_2 can be merged as they both multiply the same statistic. This leads to the conclusion that $(s_1, t_1) = (\sum_j x_j, \sum_j \log x_j!)$ is sufficient for comparing models M_1 and M_2. Here $\sum_j x_j$ is a statistic sufficient for parameter estimation in either model whilst $\sum_j \log x_j!$ captures the differing probabilities of the data under the base measure of the Poisson and geometric distributions.

We assign equal prior probability to each of the two models and complete their definition by assigning an Exponential(1) prior to λ in model M_1 and a Uniform([0,1]) prior to μ in model M_2. These priors are conjugate to the likelihood function in each model, so that it is possible to compute analytically the evidence under each model:

$$p(x|M_1) = \frac{s_1!}{\exp(t_1)(n+1)^{s_1+1}}, \tag{24}$$

$$p(x|M_2) = \frac{n!s_1!}{(n+s_1+1)!}. \tag{25}$$

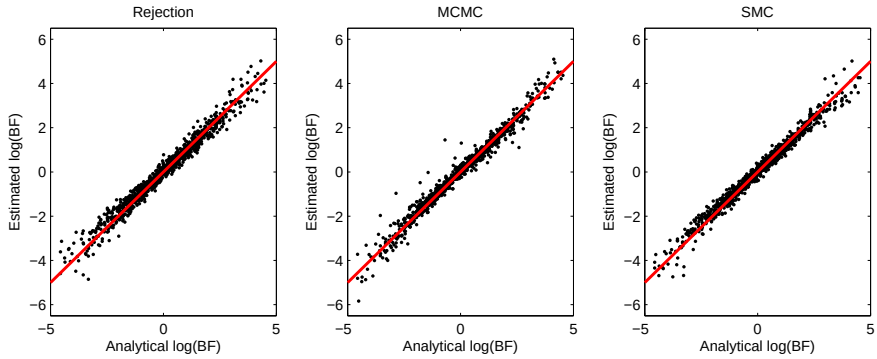

Figure 1: Comparison of the exact and estimated values of the log- Bayes Factor for each of the three estimation schemes for the example of Section 4.1.

Comparison of algorithms

In order to test our approximate method of model choice in this context, we generated datasets of size $n = 100$ made of independent and identically distributed random variables from Poisson(0.5). We generated (using rejection sampling) 1000 such datasets uniformly covering the range of $p_1 = \frac{p(x|M_1)}{p(x|M_1)+p(x|M_2)}$ from 0.01 to 0.99, to ensure that testing is performed in a wide range of scenarios. For each dataset, we estimated the evidence of the two models M_1 and M_2 using three different schemes:

1. The rejection algorithm 4 using the prior for proposal distribution, $N = 30,000$ iterations and tolerance ϵ. This is equivalent to using the algorithm of Grelaud et al. (2009).

2. The MCMC algorithm run for $N = 15,000$ iterations with tolerance ϵ and mutation kernel $x \rightarrow \text{Norm}(x, 0.1)$ followed by algorithm 5 to estimate the evidence using kernel Q equal to Student's t distribution with 4 degrees of freedom.

3. The SMC algorithm 2 run with $N = 10,000$ particles and the sequence of tolerances $\{3\epsilon, 2\epsilon, \epsilon\}$, followed by Equation 11 to estimate the evidence.

Note that each of these three schemes requires exactly 30,000 simulations of datasets, so that if simulation was the most computationally expensive step (as is ordinarily the case when complex models are considered) then each of the three schemes would have the same computational cost. Furthermore, we used the same tolerance $\epsilon = 0.05$ in the three schemes so that they are equally approximate in the sense of Equation 4. The main difference between these three schemes therefore lies in how well they explore this approximate posterior, which directly affects the precision of the evidence estimation.

Figure 1 compares the values of the log- Bayes Factor $B_{1,2} = \frac{p(x|M_1)}{p(x|M_2)}$ computed exactly (using Equations 24 and 25) and estimated using each of the three schemes. All

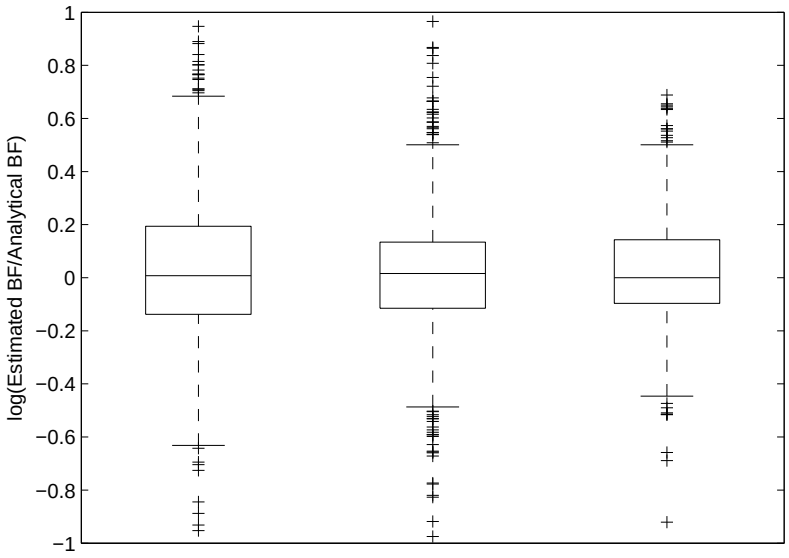

Figure 2: Boxplot of the log-ratio of the exact and estimated values of the Bayes Factor for each of the three estimation schemes for the example of Section 4.1.

three schemes perform best when the Bayes Factor is moderate in either direction. When one model is clearly preferable to the other, all three methods become less accurate because the estimate of the evidence for the unlikely model becomes more approximate. However, as pointed out by Grelaud et al. (2009), precise estimation of the Bayes Factor is typically less important when one model is clearly favored over the other since it does not affect the conclusion of which model is "correct". In cases where it is less clear which of the two models is correct (for example where the log- Bayes Factor is between -2 and 2) the estimation of the Bayes Factor is less accurate using the rejection scheme than using the MCMC or SMC schemes.

Figure 2 shows the log-ratio of the exact and estimated values of the Bayes Factor represented as a boxplot for each of the three estimation schemes. The interquartile ranges are 0.33 for the rejection scheme, 0.24 for the MCMC scheme and 0.23 for the SMC scheme. It is therefore clear that both the MCMC and SMC schemes perform better at estimating the Bayes Factor than the rejection scheme. This difference is explained by the fact that the MCMC and SMC schemes explore the posterior distribution of parameter under each model more efficiently than the rejection sampler, thus resulting in better estimates of the evidence of each parameter and therefore of the Bayes Factor. Because the example we considered here is relatively simple, with only one

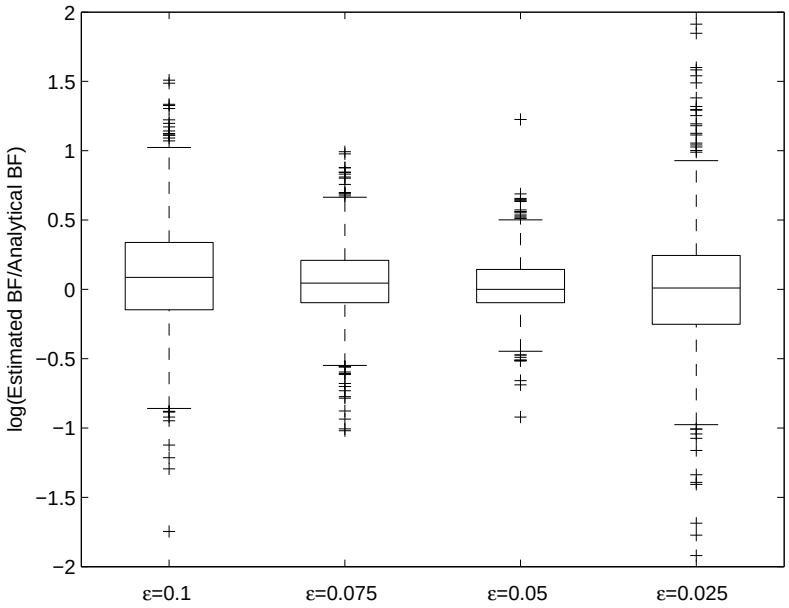

Figure 3: Boxplot of the log-ratio of the exact and estimated values of the Bayes Factor in the SMC scheme with 4 different values of the final tolerance ϵ for the example of Section 4.1.

parameter in each model, the rejection scheme was still able to estimate Bayes Factors reasonably well (Figure 1). But for more complex models where the prior distribution of parameters would be very diffuse relative to their posterior distribution, the acceptance rate of a rejection scheme would become very small for a reasonably small value of the tolerance ϵ (Marjoram et al. 2003; Sisson et al. 2007). In such cases it becomes necessary to improve the sampling of the posterior distribution using MCMC or SMC techniques. We also implemented a scheme based on the algorithm of Del Moral et al. (2008) and Equation 12 which resulted in an improvement over the rejection sampling scheme but which did not perform as well as the other schemes considered. Due to the different form of the estimator used by this algorithm it is not clear that this ordering would be preserved when considering more difficult problems. The question of which sampling scheme provides the best estimates of evidence is highly dependent on the problem and exact implementation details as it is when sampling of parameters is the aim.

Choice of the tolerance ϵ

A key component of any Approximate Bayesian Computation algorithm is the choice of the tolerance ϵ (e.g. Marjoram et al. 2003). If the tolerance is too small then the

acceptance rate is small so that either the posterior is estimated by only a few points or the algorithm would need to be run for longer. On the other hand if the tolerance is too large then the approximation in Equation 4 becomes inaccurate. We found that the choice of the tolerance is also paramount when the aim is to estimate an evidence or a Bayes Factor. Figure 3 shows the log-ratio of the exact and estimated values of the Bayes Factor for the SMC scheme described above, using four different values of the final tolerance ϵ: 0.1, 0.075, 0.05 and 0.025 (similar results were obtained using the rejection and MCMC schemes). As ϵ goes down from 0.1 to 0.05, the estimation of the Bayes Factor improves because each evidence is calculated more accurately thanks to a more accurate sampling of the posterior. However, estimating the Bayes Factor is less accurate when using $\epsilon = 0.025$ than $\epsilon = 0.05$ because the number of particles accepted in each model becomes too small for the approximation in Equation 8 to hold well.

It should be noted that all three techniques produce better estimates with greater simulation effort. Figure 4 shows that $\epsilon = 0.05$ performs best, but this is only true for the number of simulation (30,000) that we allowed. Using a larger number of simulations allows both the use of a smaller ϵ, reducing the bias of the ABC approximation, and the use of a larger number of samples which reduces the Monte Carlo error.

4.2 Application in population genetics

The problem

Pritchard et al. (1999) used an Approximate Bayesian Computation approach to analyze microsatellite data from 8 loci on the Y chromosome and 445 human males sampled around the world (Pérez-Lezaun et al. 1997; Seielstad et al. 1998). This data was also later reanalyzed by Beaumont et al. (2002). The population model assumed by both studies was the coalescent (Kingman 1982a,b,c) with mutations happening at rate μ per locus per generation. A number of mutational models were considered by Pritchard et al. (1999), but here we follow Beaumont et al. (2002) in focusing on the single-step model (Ohta and Kimura 1973). Pritchard et al. (1999) used a model of population size similar to that described by Weiss and von Haeseler (1998), where an ancestral population of previously constant size N_A started to grow exponentially at time t_g generations before the present and at a rate r per generation. Let M_1 denote this model of population size dynamics. Thus if t denotes time in generations before the present, the population size $N(t|M_1)$ at time t follows:

$$N(t|M_1) = \begin{cases} N_A & \text{if } t > t_g, \\ N_A \exp(r(t_g - t)) & \text{if } t \leq t_g. \end{cases} \tag{26}$$

Pritchard et al. (1999) also considered a model where the population size is constant at N_A. This can be obtained by setting $t_g = 0$ in Equation 26. The constant population size model is therefore nested in the above model, which allows to perform model comparison between them directly as described in Section 2.2 by performing inference under the larger model with half of the prior weight placed on the smaller model, i.e. $t_g = 0$.

Pritchard et al. (1999) used this method and found strong support for the exponential growth model, with a posterior probability for the constant model $< 1\%$.

Algorithmic framework

Here we propose to reproduce and extend those results by considering other population size models which are not necessarily nested into one another. Simulation of data under the coalescent with any population size dynamics can be achieved by first simulating a coalescent tree under a constant population size model (Kingman 1982a) and then rescaling time according to the function $N(t)$ of the population size in the past as described by Griffiths and Tavaré (1994); Hein et al. (2005).

We summarize the data using the same three statistics as Pritchard et al. (1999), namely the number of distinct haplotypes n, the mean (across loci) of the variance in repeat numbers $\bar{V}$ and the mean effective heterozygosity $\bar{H}$. For the observed data, we find that $n = 316$, $\bar{V} = 1.1488$ and $\bar{H} = 0.6358$. Beaumont et al. (2002) supplemented these with a number of additional summary statistics but found little improvement. Note that the summary statistics we use are not sufficient either for estimating the parameters of a given model (i.e. in the sense of Equation 13) or for the comparison of two models (i.e. in the sense of Equation 16). We will return to this difficulty in the discussion. We also use the same definition of π_ϵ as Pritchard et al. (1999), namely an indicator function (Equation 6) with a Chebyshev distance.

	μ ($\times 10^{-4}$)	r ($\times 10^{-4}$)	t_g	N_A ($\times 10^3$)
Prior	$\Gamma(10{,}8{\cdot}10^{-5})$ 8 [4;14]	Exp(0.005) 50 [1.3;180]	Exp(1000) 1000 [25;3700]	Log-$\mathcal{N}$(8.5,2) 36 [0.1;250]
Pritchard et al. (1999)	7 [4;12]	75 [22;209]	900 [300;2150]	1.5 [0.1;4.9]
Beaumont et al. (2002)	7.2 [3.5;12]	75 [23;210]	900 [320;2100]	1.5 [0.14;4.4]
This study	7.4 [3.6;12]	76 [22;215]	920 [310;2300]	1.4 [0.08;4.4]

Table 1: Means and 95% credibility intervals for the estimates of the parameters of the model M_1 used by Pritchard et al. (1999) and defined by Equation 26.

Pritchard et al. (1999) used the rejection ABC algorithm to sample from the parameters (μ, r, t_g, N_A) of their model (Equation 26) assuming the priors shown in Table 1. Beaumont et al. (2002) repeated this approach, and found that they get ~ 1600 acceptable simulation when performing 10^6 simulations with $\epsilon = 0.1$. We repeated this approach once again and found that it took $\sim$600000 simulations to get 1000 acceptances, which is in accordance with the acceptance rate reported by Beaumont et al. (2002). To generate this number of simulations took ~ 12 hours on a modern computer.

To reduce this computational cost, we implemented an ABC-SMC algorithm with the sequence of tolerances $\{\epsilon_1 = 0.8, \epsilon_2 = 0.4, \epsilon_3 = 0.2, \epsilon_4 = 0.1\}$, and a requirement of 1000 accepted particles for each generation. The final generation therefore contained 1000 accepted particles for the tolerance $\epsilon = 0.1$, making it comparable to the sample produced by the rejection algorithm, with the difference that it only required $\sim$5% of

the number of simulations needed by the rejection algorithm. The results of this analysis are shown in Table 1 and are in agreement with those of Pritchard et al. (1999) and Beaumont et al. (2002).

Other models

As an alternative to the model M_1 used by Pritchard et al. (1999), we consider the model denoted M_2 of pure exponential growth as used for example by Slatkin and Hudson (1991):

$$N(t|M_2) = N_0 \exp(-rt). \tag{27}$$

This model has three parameters: the mutation rate μ, the current effective population size N_0 and the rate of growth r. We assume the same priors for μ and r as in the model M_1 of Pritchard et al. (1999), and for N_0 use the same diffuse prior as for N_A in M_1.

As a third alternative, we consider the model of sudden expansion (Rogers and Harpending 1992) denoted M_3 where t_g generations back in time the effective population size suddenly increased to its current size:

$$N(t|M_3) = \begin{cases} N_0 & \text{if } t < t_g, \\ N_0 \cdot s & \text{if } t \geq t_g. \end{cases} \tag{28}$$

This model M_3 has four parameters: the mutation rate μ, the current population size N_0, the time t_g when the size suddenly increased and the factor s by which it used to be smaller. The priors for μ, N_0 and t_g were as defined previously for models M_1 and M_2, and for s we followed Thornton and Andolfatto (2006) in using a Uniform([0,1]) prior.

Finally we consider a bottleneck model M_4 as described by Tajima (1989) where the effective population size was reduced by a factor s between time t_g and $t_g + t_b$ before the present:

$$N(t|M_4) = \begin{cases} N_0 & \text{if } t < t_g, \\ N_0 \cdot s & \text{if } t_g \leq t < t_g + t_b, \\ N_0 & \text{if } t \geq t_g + t_b. \end{cases} \tag{29}$$

This model has five parameters: the mutation rate μ, the current population size N_0, the time t_g when the bottleneck finished, its duration t_b and its severity s.

Comparison of models and consequences

For each of the 4 models described above, we computed the evidence using Equation 11 (excluding the multiplicative constant π_ϵ which is the same for all evidences since

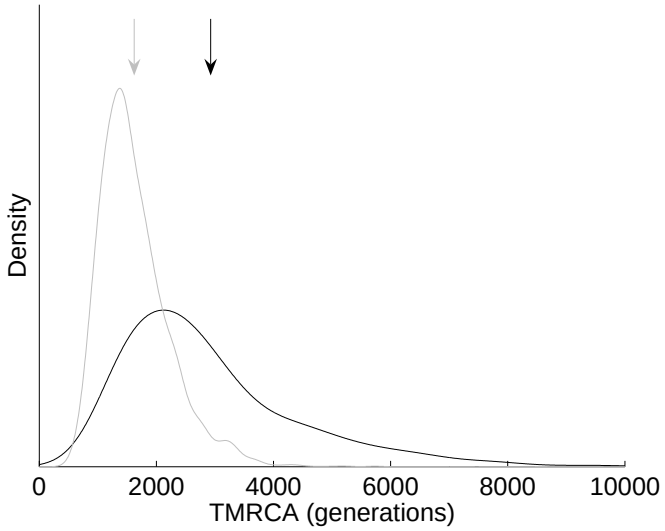

Figure 4: Plots of the posterior densities for the TMRCA under model M_1 (black) and model M_2 (gray). The mean of each distribution is indicated by an arrow of the corresponding color.

the same tolerance and summary statistics were used). The Bayes Factors for the comparison of the 4 models are shown in Table 2. According to the scale of Jeffreys (1961) (cf. Introduction), we have equivalently good fit to the data of models M_1 and M_2, substantial ground to reject model M_3 and very strong evidence to reject model M_4. The fact that models M_1 and M_2 have a Bayes Factor close to 1 means that there is no evidence to support a period during which the effective population size was constant (as assumed in the model of Pritchard et al. 1999) before it started its exponential growth.

	M_1	M_2	M_3	M_4
M_1 (Pritchard et al. 1999)	1.00	0.96	8.54	33.32
M_2 (pure exponential growth)	1.04	1.00	8.92	34.80
M_3 (sudden increase)	0.12	0.11	1.00	3.90
M_4 (bottleneck)	0.03	0.03	0.26	1.00

Table 2: Bayes Factors for the comparison between models M_1, M_2, M_3 and M_4. The value reported on the i-th row and the j-th column is the Bayes Factor $B_{i,j}$ between models M_i and M_j.

We estimated the time to the most recent common ancestor (TMRCA) of the human male population by recording for each model the TMRCAs of each simulation accepted in the last SMC generation. In spite of the fact that they fit equally well to the data, the

models M_1 and M_2 produce fairly different estimates of the TMRCA of the human male population (Figure 4). The pure exponential growth model results in a point TMRCA estimate of 1600 generations which is almost half of the model of Pritchard et al. (1999) with an estimate of 3000 generations. The TMRCA estimate under the pure exponential model is in better agreement with the results based on different datasets of Tavaré et al. (1997) and Thomson et al. (2000).

5 Discussion

We have presented a novel likelihood-free approach to model comparison, based on the independent estimation of the evidence of each model. This has the advantage that it can easily be incorporated within an MCMC or SMC framework, which can greatly improve the exploration of a large parameter space, and consequently results in more accurate estimates of evidence and Bayes Factor for a given computational cost. We also proposed a general method for finding a summary statistic sufficient for comparing two models, and showed how this could be applied in particular to models of the exponential family. Following this method ensures that the only approximation being made comes from the use of the tolerance ϵ, and the advanced sampling techniques that we use allow to reach low values of the tolerance in much less time than would be needed using rejection sampling. We illustrated this point on a toy example where marginal likelihoods can be computed analytically and sufficient statistics are available.

However, for more complex models such as the ones we considered in our population genetics application, sufficient statistics of reasonably low dimensionality (as required for ABC to be efficient) are not available. In such situation one must rely on statistics that are thought to be informative about the model comparison problem. This is analogous to the necessity to use non-sufficient statistic in standard ABC (where sampling of parameters is the aim) when complex model and data are involved (Beaumont et al. 2002; Marjoram et al. 2003). Joyce and Marjoram (2008) have described a method to help find summary statistics that are close to sufficiency in this setting, and given the relationship that we established between sufficiency for model comparison and sufficient for parameter estimation (cf. Section 3.2), these should prove useful also in the likelihood-free model comparison context. A number of sophisticated methodological techniques have been described in recent years and could be directly applied in the model selection context (Peters et al. 2008; Del Moral et al. 2008; Fearnhead and Prangle 2010).

Although the proposed method inherits all of the difficulties of both ABC and Bayesian model comparison based upon a finite collection of candidate models, the results of Section 4 suggest that when these difficulties (particularly the interpretation of the procedure, the selection of appropriate statistics and the choice of prior distributions for the model parameters) can be adequately resolved good results can be obtained by these methods.

Appendix: Convergence of the Evidence Approximation

Our strategy involves two steps:

1. Approximate the joint density of a phantom x and θ under a given model with:

$$\hat{p}(x,\theta) = \frac{p(\theta)p(x|\theta)\pi_\epsilon(x|x_{obs})}{Z_\epsilon},$$

   where π_ϵ is a probability density with respect to the same dominating measure as $p(x|\theta)$.

2. Estimate Z_ϵ numerically for each model using Monte Carlo methods.

Here we demonstrate that the Z_ϵ of the first step approximates the normalising constant of interest, $Z = \int p(x_{obs},\theta)d\theta = p(x_{obs})$. The approximation techniques used in the second step are essentially standard and their convergence follows by standard arguments.

Proposition 1. If $p(x|\theta)$ is continuous for almost every θ (with respect to the prior measure over θ) and either:

1. $\text{supp}\,(\pi_\epsilon(\cdot|x_{obs})) \subset B_\epsilon(x_{obs}) = \{x : |x - x_{obs}| < \epsilon\}$, or,
2. For $p(\theta)p(x|\theta)dxd\theta$-almost every (θ, x): $p(x_{obs}|\theta) \leq M < \infty$ and $\pi_\epsilon(\cdot|x_{obs})$ becomes increasingly and arbitrarily concentrated around x_{obs} for sufficiently small ϵ in the sense that:

$$\forall \epsilon, \alpha > 0 : \exists \varepsilon_\alpha^\star > 0 \text{ such that } \forall \gamma \leq \varepsilon_\alpha^\star : \int_{B_\epsilon(x_{obs})} \pi_\gamma(x|x_{obs})dx > 1 - \alpha \qquad (30)$$

then:

$$\lim_{\epsilon \to 0} Z_\epsilon = p(x_{obs}) = Z.$$

Proof. Consider first the case in which condition 1 holds. For any $\delta > 0$ there exists $\epsilon_\delta > 0$ such that:

$$\forall \epsilon \leq \epsilon_\delta, \forall x \in B_\epsilon(x_{obs}) : |p(x|\theta) - p(x_{obs}|\theta)| < \delta.$$

For given δ, consider $Z_\epsilon - Z$ for $\epsilon \leq \epsilon_\delta$:

$$\begin{aligned} |Z_\epsilon - Z| &= \left| \int p(\theta) \int p(x|\theta)\pi_\epsilon(x|x_{obs})dxd\theta - \int p(\theta)p(x_{obs}|\theta)d\theta \right| \\ &\leq \int p(\theta) \int |p(x|\theta) - p(x_{obs}|\theta)|\, \pi_\epsilon(x|x_{obs})dxd\theta \\ &\leq \int p(\theta)\delta d\theta = \delta \end{aligned}$$

As this holds for arbitrary δ, $\lim_{\epsilon\to 0} Z_\epsilon \to Z$.

The second case, with general π_ϵ, uses similar logic: For any $\delta > 0$, there exists $\epsilon'_\delta > 0$ such that:

$$\forall \epsilon \leq \epsilon'_\delta, \forall x \in B_\epsilon(x_{obs}) : |p(x|\theta) - p(x_{obs}|\theta)| < \delta/2.$$

Furthermore, for any $\delta, \epsilon > 0$, we can find $\varepsilon'(\delta, \epsilon) > 0$ such that:

$$\forall \gamma < \varepsilon'(\delta, \epsilon) : \int_{B_\epsilon(x|x_{obs})} \pi_\gamma(x|x_{obs})dx > 1 - \delta/2M.$$

For any given $\delta > 0$, for any $\varepsilon < \varepsilon'(\delta, \epsilon'_\delta) \wedge \epsilon'_\delta$:

$$\begin{aligned}
|Z_\epsilon - Z| &= \left| \int p(\theta) \int p(x|\theta)\pi_\epsilon(x|x_{obs})dxd\theta - \int p(\theta)p(x_{obs}|\theta)d\theta \right| \\
&\leq \int p(\theta) \left\{ \int_{B_{\epsilon'_\delta}(x_{obs})} |p(x|\theta) - p(x_{obs}|\theta)|\, \pi_\epsilon(x|x_{obs})dx + \right. \\
&\qquad \left. \int_{\overline{B_{\epsilon'_\delta}(x_{obs})}} |p(x|\theta) - p(x_{obs}|\theta)|\, \pi_\epsilon(x|x_{obs})dx \right\} d\theta \\
&\leq \delta/2 + M \cdot \delta/2M = \delta.
\end{aligned}$$

Where the first integral is bounded by a simple continuity argument and the second by bounding the difference between a positive function evaluated at two points by its supremum and noting that the integral of π_ϵ over $\overline{B_{\epsilon'_\delta}(x_{obs})}$ is at most $\delta/2M$. Again, this result holds for any $\delta > 0$ and so Z_ϵ converges to Z as $\epsilon \to 0$.

Comments

Condition *a* holds for any sequence π_{ϵ_i} with compact support that contracts to a point as $\epsilon_i \downarrow 0$ by simple relabelling.

Although the result is reassuring and holds under reasonably weak conditions, verifying these assumptions will often be difficult in practice as ABC is generally used for models which are not analytically well characterised.

Similar arguments would allow rates of convergence to be obtained with the additional assumption of (local) Lipschitz continuity.

The proof in the case of discrete data spaces is direct: for ϵ smaller than some threshold the approximation is exact.

References

Beaumont, M. A., Zhang, W., and Balding, D. J. (2002). "Approximate Bayesian Computation in Population Genetics." *Genetics*, 162(4): 2025–2035.
URL http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve\&db=pubmed\&dopt=Abstract\&list_uids=12524368 50, 52, 59, 66, 67, 68, 70

Chib, S. (1995). "Marginal Likelihood From the Gibbs Output." *Journal of the American Statistical Association*, 90(432): 1313–1321. 50

Del Moral, P. (2004). *Feynman-Kac formulae: genealogical and interacting particle systems with applications*. Probability and Its Applications. New York: Springer. 58

Del Moral, P., Doucet, A., and Jasra, A. (2006). "Sequential monte carlo samplers." *Journal of the Royal Statistical Society Series B*, 68(3): 411–436. 53, 54, 58

— (2008). "An adaptive sequential Monte Carlo method for approximate Bayesian computation." *Technical Report, Imperial College London*.
URL http://www2.imperial.ac.uk/~aj2/smc_abc_arno.pdf 54, 58, 65, 70

Dellaportas, P., Forster, J., and Ntzoufras, I. (2002). "On Bayesian model and variable selection using MCMC." *Statistics and Computing*, 12(1): 27–36. 50

Fearnhead, P. and Prangle, D. (2010). "Semi-automatic Approximate Bayesian Computation." *Arxiv preprint arXiv:1004.1112*. 70

Friel, N. and Pettitt, A. (2008). "Marginal likelihood estimation via power posteriors." *Journal Of The Royal Statistical Society Series B*, 70(3): 589–607. 50

Fu, Y. and Li, W. (1997). "Estimating the age of the common ancestor of a sample of DNA sequences." *Molecular Biology and Evolution*, 14(2): 195–199.
URL http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve\&db=pubmed\&dopt=Abstract\&list_uids=9029798 51

Gilks, W. and Spiegelhalter, D. (1996). *Markov chain Monte Carlo in practice*. Chapman & Hall/CRC. 52

Green, P. (1995). "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination." *Biometrika*, 82(4): 711–732. 50

Grelaud, A., Robert, C., Marin, J., Rodolphe, F., and Taly, J. (2009). "ABC likelihood-free methods for model choice in Gibbs random fields." *Bayesian Analysis*, 4(2): 317–336. 54, 55, 60, 61, 62, 63, 64

Griffiths, R. and Tavaré, S. (1994). "Sampling theory for neutral alleles in a varying environment." *Philosophical Transactions of the Royal Society B*, 344(1310): 403–410. 67

Hein, J., Schierup, M., and Wiuf, C. (2005). *Gene genealogies, variation and evolution: a primer in coalescent theory*. Oxford University Press, USA. 67

Jeffreys, H. (1961). *Theory of probability*. Clarendon Press, Oxford :, 3rd ed. edition. 50, 69

Joyce, P. and Marjoram, P. (2008). "Approximately sufficient statistics and Bayesian computation." *Statistical Applications in Genetics and Molecular Biology*, 7(1). 70

Kass, R. and Raftery, A. (1995). "Bayes factors." *Journal of the American Statistical Association*, 90(430). 49, 57

Kingman, J. F. C. (1982a). "Exchangeability and the Evolution of Large Populations." In Koch, G. and Spizzichino, F. (eds.), *Exchangeability in Probability and Statistics*, 97–112. North-Holland, Amsterdam. 66, 67

— (1982b). "On the genealogy of large populations." *Journal of Applied Probability*, 19A: 27–43. 66

— (1982c). "The coalescent." *Stochastic Processes and their Applications*, 13(235): 235–248. 66

Klaas, M., de Freitas, N., and Doucet, A. (2005). "Toward Practical N2 Monte Carlo: the Marginal Particle Filter." In *Proceedings of the Proceedings of the Twenty-First Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-05)*, 308–315. Arlington, Virginia: AUAI Press.
URL http://uai.sis.pitt.edu/papers/05/p308-klaas.pdf 54

Liu, J. (2001). *Monte Carlo strategies in scientific computing*. Springer Verlag. 57

Luciani, F., Sisson, S., Jiang, H., Francis, A., and Tanaka, M. (2009). "The epidemiological fitness cost of drug resistance in Mycobacterium tuberculosis." *Proceedings of the National Academy of Sciences*, 106(34): 14711–14715. 50

Marjoram, P., Molitor, J., Plagnol, V., and Tavaré, S. (2003). "Markov chain Monte Carlo without likelihoods." *Proceedings of the National Academy of Sciences*, 100(26): 15324–15328.
URL http://www.pnas.org/cgi/content/abstract/100/26/15324 52, 59, 65, 70

Meng, X. (1994). "Posterior predictive p-values." *The Annals of Statistics*, 22(3): 1142–1160. 54

Neal, R. (2001). "Annealed importance sampling." *Statistics and Computing*, 11(2): 125–139. 50

Newton, M. and Raftery, A. (1994). "Approximate Bayesian inference with the weighted likelihood bootstrap." *Journal of the Royal Statistical Society Series B*, 56(1): 3–48. 50

Ohta, T. and Kimura, M. (1973). "A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population." *Genetical Research*, 22(2): 201–204.
URL http://www.hubmed.org/display.cgi?uids=4777279 66

Pérez-Lezaun, A., Calafell, F., Seielstad, M., Mateu, E., Comas, D., Bosch, E., and Bertranpetit, J. (1997). "Population genetics of Y-chromosome short tandem repeats in humans." *Journal of Molecular Evolution*, 45(3): 265–270.
URL http://www.hubmed.org/display.cgi?uids=9302320 66

Peters, G., Fan, Y., and Sisson, S. (2008). "On sequential Monte Carlo, partial rejection control and approximate Bayesian computation." *Arxiv preprint arXiv:0808.3466*. 70

Peters, G. W. (2005). "Topics In Sequential Monte Carlo Samplers." M.Sc., University of Cambridge, Department of Engineering. 54

Pritchard, J., Seielstad, M., Perez-Lezaun, A., and Feldman, M. (1999). "Population growth of human Y chromosomes: a study of Y chromosome microsatellites." *Molecular Biology and Evolution*, 16(12): 1791–1798.
URL http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve\&db=pubmed\&dopt=Abstract\&list_uids=10605120 51, 52, 54, 59, 66, 67, 68, 69, 70

Ratmann, O., Andrieu, C., Wiuf, C., and Richardson, S. (2009). "Model criticism based on likelihood-free inference, with an application to protein network evolution." *Proceedings of the National Academy of Sciences*, 106(26): 10576–10581.
URL http://www.hubmed.org/display.cgi?uids=19525398 54

Robert, C. P. (2001). *The Bayesian Choice*. Springer Texts in Statistics. New York: Springer Verlag, 2nd edition. 49

Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods*. New York: Springer, 2nd edition. 52, 57

Robert, C. P., Mengersen, K., and Chen, C. (2010). "Model choice versus model criticism." *Proceedings of the National Academy of Sciences*, 107(3): E5–E5.
URL http://www.pnas.org/content/107/3/E5.short 54

Rogers, A. R. and Harpending, H. (1992). "Population growth makes waves in the distribution of pairwise genetic differences." *Molecular Biology and Evolution*, 9(3): 552–569.
URL http://www.hubmed.org/display.cgi?uids=1316531 68

Seielstad, M. T., Minch, E., and Cavalli-Sforza, L. L. (1998). "Genetic evidence for a higher female migration rate in humans." *Nature Genetics*, 20(3): 278–280.
URL http://www.hubmed.org/display.cgi?uids=9806547 66

Shao, J. (1999). *Mathematical Statistics*. Springer. 53

Sisson, S. A., Fan, Y., and Tanaka, M. M. (2007). "Sequential Monte Carlo without likelihoods." *Proceedings of the National Academy of Sciences*, 104(6): 1760–1765.
URL http://www.pnas.org/content/104/6/1760.abstract 53, 58, 65

Slatkin, M. and Hudson, R. R. (1991). "Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations." *Genetics*, 129(2): 555–562.
URL http://www.hubmed.org/display.cgi?uids=1743491 68

Stephens, M. (2000). "Bayesian analysis of mixture models with an unknown number of components-an alternative to reversible jump methods." *The Annals of Statistics*, 28(1): 40–74. 50

Tajima, F. (1989). "The effect of change in population size on DNA polymorphism." *Genetics*, 123(3): 597–601.
URL http://www.hubmed.org/display.cgi?uids=2599369 68

Tavaré, S., Balding, D. J., Griffiths, R. C., and Donnelly, P. (1997). "Inferring Coalescence Times From DNA Sequence Data." *Genetics*, 145(2): 505–518.
URL http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve\&db=pubmed\&dopt=Abstract\&list_uids=9071603 51, 70

Thomson, R., Pritchard, J. K., Shen, P., Oefner, P. J., and Feldman, M. W. (2000). "Recent common ancestry of human Y chromosomes: evidence from DNA sequence data." *Proceedings of the National Academy of Sciences*, 97(13): 7360–7365.
URL http://www.hubmed.org/display.cgi?uids=10861004 70

Thornton, K. and Andolfatto, P. (2006). "Approximate Bayesian Inference Reveals Evidence for a Recent, Severe Bottleneck in a Netherlands Population of Drosophila melanogaster." *Genetics*, 172(3): 1607–1619.
URL http://www.genetics.org/cgi/content/abstract/172/3/1607 54, 68

Toni, T., Welch, D., Strelkowa, N., Ipsen, A., and Stumpf, M. (2009). "Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems." *Journal of The Royal Society Interface*, 6(31): 187–202. 55

Weiss, G. and von Haeseler, A. (1998). "Inference of Population History Using a Likelihood Approach." *Genetics*, 149(3): 1539–1546.
URL http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve\&db=pubmed\&dopt=Abstract\&list_uids=9649540 52, 66

Wilkinson, R. (2008). "Approximate Bayesian computation (ABC) gives exact results under the assumption of model error." *Arxiv preprint arXiv:0811.3355*. 52

Wilkinson, R. G. (2007). "Bayesian Estimation of Primate Divergence Times." Ph.D. thesis, University of Cambridge. 54