Lack of confidence in ABC model choice

Christian P. Robert, [* † ‡] Jean-Marie Cornuet, [§] Jean-Michel Marin, [¶] and Natesh S. Pillai [‖]

[*]Université Paris Dauphine, CEREMADE, Paris, France,[†]Institut Universitaire de France,[‡]CREST, Paris, France,[§]CBGP, INRA, Montpellier, France,[¶]I3M, UMR CNRS 5149, Université Montpellier 2, France, and [‖]Department of Statistics, Harvard University, Cambridge, MA

Approximate Bayesian computation (ABC) have become a essential tool for the analysis of complex stochastic models. Earlier, Grelaud et al. (2009) advocated the use of ABC for Bayesian model choice in the specific case of Gibbs random fields, relying on a inter-model sufficiency property to show that the approximation was legitimate. Having implemented ABC-based model choice in a wide range of phylogenetic models in the DIY-ABC software (Cornuet et al., 2008), we now present theoretical background as to why a generic use of ABC for model choice is ungrounded, since it depends on an unknown amount of information loss induced by the use of insufficient summary statistics. The approximation error of the posterior probabilities of the models under comparison may thus be unrelated with the computational effort spent in running an ABC algorithm. We then conclude that additional empirical verifications of the performances of the ABC procedure are necessary to conduct model choice. [1]

likelihood-free methods | Bayes factor | DIYABC | Bayesian model choice | sufficiency

Inference on population genetic models such as coalescent trees is one representative example of cases when statistical analyses like Bayesian inference cannot easily operate because the likelihood function associated with the data is not completely known, i.e. cannot be computed in a manageable time (Tavaré et al., 1997; Beaumont et al., 2002; Cornuet et al., 2008). The fundamental reason for this impossibility is that the statistical model associated with coalescent data needs to integrate over trees of high complexity.

In such settings, traditional approximation tools based on Monte Carlo simulation (Robert and Casella, 2004) from the Bayesian posterior distribution are unavailable for all practical purposes. Indeed, due to the complexity of the latent structures defining the likelihood (such as the coalescent tree), simulation of those structures is too unstable to be trusted to bring a reliable approximation in a manageable time. Such complex models call for a practical if cruder approximation method, the ABC methodology (Tavaré et al., 1997; Pritchard et al., 1999). This rejection technique bypasses the computation of the likelihood function via simulations from the corresponding distribution (see Beaumont, 2010 and Lopes and Beaumont, 2010 for recent surveys). The wide and successful array of applications based on implementations of ABC in genomics and ecology is covered by Csillèry et al. (2010a).

In the following, we argue that ABC is a valid approximation method for conducting Bayesian inference in complex stochastic models, barring the limitation that it cannot be trusted to discriminate between those complex stochastic models when based on summary statistics. In essence, we highlight the fact that, since ABC is conducting model choice based on insufficient statistics, the resulting inference is flawed in that the loss of information may be severe to the point of inconsistency: ABC model selection may easily fail to recover the true model, even with an infinite amount of observation and of computation. We demonstrate this inconsistency in a limiting (and most favourable) case.

Our conclusion are to opt for a cautionary approach when using ABC in model choice. The level of approximation resulting from this algorithm cannot be evaluated, except via Monte Carlo evaluations of the performances of the method. More empirical measures such as those proposed in the DIY-ABC software (Cornuet et al., 2008), in Ratmann et al. (2009) and in Drovandi et al. (2011) thus seem to be the only available solution at the current time for conducting model comparison.

We stress here that, while Templeton (2008, 2010) repeatedly expressed reservations about the formal validity of the ABC approach in statistical testing, those criticisms were addressed at the Bayesian paradigm *per se* rather than at the approximation method. Quite clearly, Templeton's criticisms got rebutted in Beaumont et al. (2010); Csillèry et al. (2010b); Berger et al. (2010) and are not relevant for the current paper.

Methods

The ABC algorithm. The setting in which ABC operates is the approximation of a simulation from the posterior distribution $\pi(\boldsymbol{\theta}|\mathbf{y}) \propto \pi(\boldsymbol{\theta})f(\mathbf{y}|\boldsymbol{\theta})$ when distributions associated with both the prior π and the likelihood f can be simulated (the later being unavailable in closed form). The first ABC algorithm was introduced by Pritchard et al. (1999) as follows: given a sample $\mathbf{y}$ from a sample space $\mathcal{D}$, a sample $(\theta_1, \dots, \theta_M)$ is produced by

Algorithm 1: ABC sampler

> **for** $i = 1$ to N **do**
> > **repeat**
> > > Generate $\boldsymbol{\theta}'$ from the prior distribution $\pi(\cdot)$
> > > Generate $\mathbf{z}$ from the likelihood $f(\cdot|\boldsymbol{\theta}')$
> > **until** $\rho\{\eta(\mathbf{z}), \eta(\mathbf{y})\} \leq \epsilon$
> > set $\boldsymbol{\theta}_i = \boldsymbol{\theta}'$,
> **end for**

The parameters of the ABC algorithm are the statistic η, the distance $\rho\{\cdot,\cdot\}$, and the tolerance level $\epsilon > 0$. The approximation of the posterior distribution provided by the ABC sampler is such that it instead samples from the marginal in $\boldsymbol{\theta}$ of the joint distribution

$$\pi_\epsilon(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y}) = \frac{\pi(\boldsymbol{\theta})f(\mathbf{z}|\boldsymbol{\theta})\mathbb{I}_{A_{\epsilon,\mathbf{y}}}(\mathbf{z})}{\int_{A_{\epsilon,\mathbf{y}} \times \Theta} \pi(\boldsymbol{\theta})f(\mathbf{z}|\boldsymbol{\theta})\mathrm{d}\mathbf{z}\mathrm{d}\boldsymbol{\theta}},$$

where $\mathbb{I}_B(\cdot)$ denotes the indicator function of B and

$$A_{\epsilon,\mathbf{y}} = \{\mathbf{z} \in \mathcal{D}|\rho\{\eta(\mathbf{z}), \eta(\mathbf{y})\} \leq \epsilon\}.$$

The basic justification of the ABC approximation is that, when using a sufficient statistic η and a small (enough) tolerance ϵ, we have

$$\pi_\epsilon(\boldsymbol{\theta}|\mathbf{y}) = \int \pi_\epsilon(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y})\mathrm{d}\mathbf{z} \approx \pi(\boldsymbol{\theta}|\mathbf{y}),$$

the genuine posterior distribution.

In practice, the statistic η is insufficient and the approximation then converges to $\pi(\boldsymbol{\theta}|\eta(\mathbf{y}))$ when ϵ goes to zero. This loss of information is a necessary price to pay for the access to computable quantities. While acknowledging the gain brought by ABC in handling Bayesian inference in complex models, we demonstrate here that the loss due to the ABC approximation may be arbitrary in the specific setting of Bayesian model choice.

ABC model choice. The standard Bayesian tool for model comparison is the marginal likelihood (Jeffreys, 1939)

$$w(\mathbf{y}) = \int_{\Theta} \pi(\boldsymbol{\theta}) f(\mathbf{y}|\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta},$$

which leads to the Bayes factor for comparing the evidences of models with likelihoods $f_1(\mathbf{y}|\boldsymbol{\theta}_1)$ and $f_2(\mathbf{y}|\boldsymbol{\theta}_2)$,

$$B_{12}(\mathbf{y}) = \frac{w_1(\mathbf{y})}{w_2(\mathbf{y})} = \frac{\int_{\Theta_1} \pi_1(\boldsymbol{\theta}_1) f_1(\mathbf{y}|\boldsymbol{\theta}_1) \, \mathrm{d}\boldsymbol{\theta}_1}{\int_{\Theta_2} \pi_2(\boldsymbol{\theta}_2) f_2(\mathbf{y}|\boldsymbol{\theta}_2) \, \mathrm{d}\boldsymbol{\theta}_2}.$$

As detailed in Beaumont et al. (2010); Berger et al. (2010), this ratio provides a valid criterion for model comparison that is naturally penalised for model complexity.

Bayesian model choice proceeds by creating a probability structure across M models (or likelihoods). It introduces the model index $\mathcal{M}$ as an extra unknown parameter, associated with its prior distribution, $\pi(\mathcal{M} = m)$ $(m = 1, \ldots, M)$, while the prior distribution on the parameter is conditional on the value m of the $\mathcal{M}$ index, denoted by $\pi_m(\boldsymbol{\theta}_m)$ and defined on the parameter space Θ_m. The choice between those models is then driven by the posterior distribution of $\mathcal{M}$,

$$\mathbb{P}(\mathcal{M} = m|\mathbf{y}) = \frac{\pi(\mathcal{M} = m) w_m(\mathbf{y})}{\sum_k \pi(\mathcal{M} = k) w_k(\mathbf{y})}$$

where $w_k(\mathbf{y})$ denotes the marginal likelihood for model k.

While this posterior distribution is well-defined and straightforward to interpret, it offers a challenging computational conundrum in Bayesian analysis. When the likelihood is not available, ABC represents the almost unique solution. Pritchard et al. (1999) describe the use of model choice based on ABC for distinguishing between different mutation models. The justification behind the method is that the average ABC acceptance rate associated with a given model is proportional to the posterior probability corresponding to this approximative model, when identical summary statistics, distance, and tolerance level are used over all models. In practice, an estimate of the ratio of marginal likelihoods is given by the ratio of observed acceptance rates. Using Bayes formula, estimates of the posterior probabilities are straightforward to derive. This approach has been widely implemented in the literature (see, e.g., Estoup et al., 2004, Miller et al., 2005, Pascual et al., 2007, and Sainudiin et al., 2011).

A highly representative illustration of the use of an ABC model choice approach is given by Miller et al. (2005) which analyses the European invasion of the western corn rootworm, which is North America's most destructive corn pest. Because this pest was initially introduced in Central Europe, it was believed that subsequent outbreaks in Western Europe originated from this area. Based on this ABC model choice analysis of the genetic variability of the rootworm, the authors conclude that this belief is false: There have been at least three independent introductions from North America during the past two decades.

An improvement to the above estimate is due to Fagundes et al. (2007), via a regression regularisation. In this approach, model indices are processed as categorical variables in a formal multinomial (polychotomous) regression. For instance, when comparing two models, this leads to a standard logistic regression. Rejection-based approaches were lately introduced by Cornuet et al. (2008), Grelaud et al. (2009) and Toni et al. (2009), in a Monte Carlo perspective simulating model indices as well as model parameters. Those more recent extensions are already widely in use by the population genetics community, as exemplified by Belle et al. (2008); Cornuet et al. (2010); Excoffier et al. (2009); Ghirotto et al. (2010); Guillemaud et al. (2009); Leuenberger and Wegmann (2010); Patin et al. (2009); Ramakrishnan and Hadly (2009); Verdu et al. (2009), or Wegmann and Excoffier (2010). Another illustration of the popularity of this approach is given by the availability of three softwares implementing an ABC model choice methodology:

- ABC-SysBio, which relies on a SMC-based ABC for inference in system biology, including model-choice (Toni et al., 2009).
- DIYABC, which relies on a regularised ABC-MC algorithm on population history using molecular markers (Cornuet et al., 2008).
- PopABC, which relies on a regular ABC-MC algorithm for genealogical simulation (Lopes et al., 2009).

As exposed in e.g. Grelaud et al. (2009), Toni and Stumpf (2010), or Didelot et al. (2011), once $\mathcal{M}$ is incorporated within the parameters, the ABC approximation to its posterior follows from the same principles as in regular ABC. The corresponding implementation is as follows, using for the summary statistic a statistic $\boldsymbol{\eta}(\mathbf{z}) = \{\eta_1(\mathbf{z}), \ldots, \eta_M(\mathbf{z})\}$ that is the concatenation of the summary statistics used for all models (with an obvious elimination of duplicates).

Algorithm 2: ABC-MC
> **for** $i = 1$ to N **do**
> > **repeat**
> > > Generate m from the prior $\pi(\mathcal{M} = m)$
> > > Generate $\boldsymbol{\theta}_m$ from the prior $\pi_m(\boldsymbol{\theta}_m)$
> > > Generate $\mathbf{z}$ from the model $f_m(\mathbf{z}|\boldsymbol{\theta}_m)$
> > > **until** $\rho\{\boldsymbol{\eta}(\mathbf{z}), \boldsymbol{\eta}(\mathbf{y})\} \leq \epsilon$
> > Set $m^{(i)} = m$ and $\boldsymbol{\theta}^{(i)} = \boldsymbol{\theta}_m$
> **end for**

The ABC estimate of the posterior probability $\pi(\mathcal{M} = m|\mathbf{y})$ is then the frequency of acceptances from model m in the above simulation

$$\widehat{\pi(\mathcal{M} = m}|\mathbf{y}) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}_{m^{(i)} = m}.$$

This also corresponds to the frequency of simulated pseudo-datasets from model m that are closer to the data $\mathbf{y}$ than the tolerance ϵ. In order to improve the estimation by smoothing, Cornuet et al. (2008) follow the rationale that motivated the use of a local linear regression in Beaumont et al. (2002) and rely on a weighted polychotomous regression to estimate $\pi(\mathcal{M} = m|\mathbf{y})$ based on the ABC output. This modelling is implemented in the DIYABC software.

The difficulty with ABC-MC

There is a fundamental discrepancy between the genuine Bayes factors (or the corresponding) posterior probabilities) and the approximations resulting from ABC-MC.

The ABC approximation to a Bayes factor, B_{12} say, resulting from Algorithm 2 is

$$\widehat{B_{12}}(\mathbf{y}) = \frac{\pi(\mathcal{M} = 2)}{\pi(\mathcal{M} = 1)} \frac{\sum_{i=1}^{N} \mathbb{I}_{m^{(i)} = 1}}{\sum_{i=1}^{N} \mathbb{I}_{m^{(i)} = 2}}$$

An alternative representation is given by

$$\widehat{B_{12}}(\mathbf{y}) = \frac{\pi(\mathcal{M}=2)}{\pi(\mathcal{M}=1)} \frac{\sum_{t=1}^{T} \mathbb{I}_{m^t=1}\,\mathbb{I}_{\rho\{\boldsymbol{\eta}(\mathbf{z}^t),\boldsymbol{\eta}(\mathbf{y})\}\leq\epsilon}}{\sum_{t=1}^{T} \mathbb{I}_{m^t=2}\,\mathbb{I}_{\rho\{\boldsymbol{\eta}(\mathbf{z}^t),\boldsymbol{\eta}(\mathbf{y})\}\leq\epsilon}},$$

where the pairs (m^t, z^t) are simulated from the (joint) prior and T is the total number of simulations that are necessary for N acceptances in Algorithm 2. In order to study the limiting behaviour of this approximation, we first let T go to infinity. (For simplification purposes and without loss of generality, we choose a uniform prior on the model index.) The limit of $\widehat{B_{12}}(\mathbf{y})$ is then

$$
\begin{aligned}
B_{12}^{\epsilon}(\mathbf{y}) &= \frac{\mathbb{P}[\mathcal{M}=1, \rho\{\boldsymbol{\eta}(\mathbf{z}),\boldsymbol{\eta}(\mathbf{y})\}\leq\epsilon]}{\mathbb{P}[\mathcal{M}=2, \rho\{\boldsymbol{\eta}(\mathbf{z}),\boldsymbol{\eta}(\mathbf{y})\}\leq\epsilon}\\
&= \frac{\iint \mathbb{I}_{\rho\{\boldsymbol{\eta}(\mathbf{z}),\boldsymbol{\eta}(\mathbf{y})\}\leq\epsilon}\pi_1(\boldsymbol{\theta}_1)f_1(\mathbf{z}|\boldsymbol{\theta}_1)\,\mathrm{d}\mathbf{z}\,\mathrm{d}\boldsymbol{\theta}_1}{\iint \mathbb{I}_{\rho\{\boldsymbol{\eta}(\mathbf{z}),\boldsymbol{\eta}(\mathbf{y})\}\leq\epsilon}\pi_2(\boldsymbol{\theta}_2)f_2(\mathbf{z}|\boldsymbol{\theta}_2)\,\mathrm{d}\mathbf{z}\,\mathrm{d}\boldsymbol{\theta}_2}\\
&= \frac{\iint \mathbb{I}_{\rho\{\boldsymbol{\eta},\boldsymbol{\eta}(\mathbf{y})\}\leq\epsilon}\pi_1(\boldsymbol{\theta}_1)f_1^{\boldsymbol{\eta}}(\boldsymbol{\eta}|\boldsymbol{\theta}_1)\,\mathrm{d}\boldsymbol{\eta}\,\mathrm{d}\boldsymbol{\theta}_1}{\iint \mathbb{I}_{\rho\{\boldsymbol{\eta},\boldsymbol{\eta}(\mathbf{y})\}\leq\epsilon}\pi_2(\boldsymbol{\theta}_2)f_2^{\boldsymbol{\eta}}(\boldsymbol{\eta}|\boldsymbol{\theta}_2)\,\mathrm{d}\boldsymbol{\eta}\,\mathrm{d}\boldsymbol{\theta}_2},
\end{aligned}
$$

where $f_1^{\boldsymbol{\eta}}(\boldsymbol{\eta}|\boldsymbol{\theta}_1)$ and $f_2^{\boldsymbol{\eta}}(\boldsymbol{\eta}|\boldsymbol{\theta}_2)$ denote the densities of $\boldsymbol{\eta}(\mathbf{z})$ when $\mathbf{z} \sim f_1(\mathbf{z}|\boldsymbol{\theta}_1)$ and $\mathbf{z} \sim f_2(\mathbf{z}|\boldsymbol{\theta}_2)$, respectively. By L'Hospital formula, if we let ϵ go to zero, the above converges to

$$B_{12}^{\boldsymbol{\eta}}(\mathbf{y}) = \frac{\int \pi_1(\boldsymbol{\theta}_1)f_1^{\boldsymbol{\eta}}(\boldsymbol{\eta}(\mathbf{y})|\boldsymbol{\theta}_1)\,\mathrm{d}\boldsymbol{\theta}_1}{\int \pi_2(\boldsymbol{\theta}_2)f_2^{\boldsymbol{\eta}}(\boldsymbol{\eta}(\mathbf{y})|\boldsymbol{\theta}_2)\,\mathrm{d}\boldsymbol{\theta}_2},$$

which is exactly the Bayes factor for testing model 1 versus model 2 based on the sole observation of $\boldsymbol{\eta}(\mathbf{y})$. This result is completely coherent with the current perspective on ABC, namely that the inference derived from the ideal ABC output when $\epsilon = 0$ only uses the information contained in $\boldsymbol{\eta}(\mathbf{y})$. Thus, in the limiting case, i.e. when the ABC algorithm uses an infinite computational power, the ABC odds ratio does not take into account the features of the data besides the value of $\boldsymbol{\eta}(\mathbf{y})$, which is why the limiting Bayes factor only depends on the distribution of $\boldsymbol{\eta}$ under both models.

In contrast with point estimation, the loss of information resulting from considering solely $\boldsymbol{\eta}$ seriously impacts the resulting inference on which model is best supported by the data. Indeed, as exhibited in a special counterexample by Grelaud et al. (2009), the information contained in $\boldsymbol{\eta}(\mathbf{y})$ is almost always lesser than the information contained in $\mathbf{y}$ and this even in the case $\boldsymbol{\eta}(\mathbf{y})$ is a sufficient statistic for *both models*. In other words, $\boldsymbol{\eta}(\mathbf{y})$ *being sufficient for both* $f_1(\mathbf{y}|\boldsymbol{\theta}_1)$ *and* $f_2(\mathbf{y}|\boldsymbol{\theta}_2)$ *does not usually imply that* $\boldsymbol{\eta}(\mathbf{y})$ *is sufficient for* $\{m, f_m(\mathbf{y}|\boldsymbol{\theta}_m)\}$. To see why this is the case, consider the most favourable case, namely when $\boldsymbol{\eta}(\mathbf{y})$ is a sufficient statistic for both models. We then have by the factorisation theorem (Lehmann and Casella, 1998) that $f_i(\mathbf{y}|\boldsymbol{\theta}_i) = g_i(\mathbf{y})f_i^{\boldsymbol{\eta}}(\boldsymbol{\eta}(\mathbf{y})|\boldsymbol{\theta}_i)$ $(i = 1, 2)$, therefore that

$$
\begin{aligned}
B_{12}(\mathbf{y}) &= \frac{w_1(\mathbf{y})}{w_2(\mathbf{y})}\\
&= \frac{\int_{\Theta_1} \pi(\boldsymbol{\theta}_1)g_1(\mathbf{y})f_1^{\boldsymbol{\eta}}(\boldsymbol{\eta}(\mathbf{y})|\boldsymbol{\theta}_1)\,\mathrm{d}\boldsymbol{\theta}_1}{\int_{\Theta_2} \pi(\boldsymbol{\theta}_2)g_2(\mathbf{y})f_2^{\boldsymbol{\eta}}(\boldsymbol{\eta}(\mathbf{y})|\boldsymbol{\theta}_2)\,\mathrm{d}\boldsymbol{\theta}_2}\\
&= \frac{g_1(\mathbf{y})\int \pi_1(\boldsymbol{\theta}_1)f_1^{\boldsymbol{\eta}}(\boldsymbol{\eta}(\mathbf{y})|\boldsymbol{\theta}_1)\,\mathrm{d}\boldsymbol{\theta}_1}{g_2(\mathbf{y})\int \pi_2(\boldsymbol{\theta}_2)f_2^{\boldsymbol{\eta}}(\boldsymbol{\eta}(\mathbf{y})|\boldsymbol{\theta}_2)\,\mathrm{d}\boldsymbol{\theta}_2}\\
&= \frac{g_1(\mathbf{y})}{g_2(\mathbf{y})}B_{12}^{\boldsymbol{\eta}}(\mathbf{y}). \qquad\qquad [1]
\end{aligned}
$$

Therefore, unless $g_1(\mathbf{y}) = g_2(\mathbf{y})$, the two Bayes factors differ by this ratio, $g_1(\mathbf{y})/g_2(\mathbf{y})$, which is only equal to one in a very small number of known cases. This decomposition is a straightforward proof that a model-wise sufficient statistic is usually not sufficient across models, i.e. for model comparison. An immediate corollary is that the ABC-MC approximation does not converge to the exact Bayes factor.

The discrepancy between the limiting ABC inference and the genuine Bayesian inference does not completely come as a surprise, because ABC is indeed an approximation method. Users of ABC algorithms are therefore prepared for some degree of imprecision in their final answer, a point stressed by Wilkinson (2008) or Fearnhead and Prangle (2010) when they qualify ABC as exact inference on a wrong model. However, the magnitude of the difference between $\widehat{B_{12}}(\mathbf{y})$ and $B_{12}^{\boldsymbol{\eta}}(\mathbf{y})$ expressed by [1] is such that there is no direct connection between both answers. In a general setting, if $\boldsymbol{\eta}$ has the same dimension as one component of the n components of $\mathbf{y}$, the ratio $g_1(\mathbf{y})/g_2(\mathbf{y})$ is equivalent to a density ratio for a sample of size $O(n)$, hence it can be arbitrarily small or arbitrarily large when n grows. Contrastingly, the Bayes factor $B_{12}^{\boldsymbol{\eta}}(\mathbf{y})$ is based on an equivalent to a single observation, hence does not necessarily converge with n, as shown by the Poisson and normal examples below and in SI. The conclusion derived from the ABC-based Bayes factor may therefore completely differ from the conclusion derived from the exact Bayes factor and there is no possibility of a generic agreement between both, or even of a manageable correction factor.

For this reason, we warn the community of ABC users that the ABC approach cannot be blindly used for model choice, with the exception of Gibbs random fields as explained in the next section. In all cases when $g_1(\mathbf{y})/g_2(\mathbf{y})$ is different from one, no inference on the true Bayes factor can be made based on the ABC-MC approximation without further information on the ratio $g_1(\mathbf{y})/g_2(\mathbf{y})$, which is most often unavailable in settings where ABC is necessary.

Didelot et al. (2011) also derived this relation between both Bayes factors in their formula [18]. They surprisingly conclude on advocating the use of ABC in complex models, when there is no useful sufficient statistic. We disagree with this perspective for the above reason that no garantee can be given on the validity of the ABC approximation to the Bayes factor.

As a final remark, we note that Sousa et al. (2009) resort to full allelic distributions in an ABC framework, instead of chosing summary statistics. They show that it is possible to apply ABC using allele frequencies to draw inferences in cases where it is difficult to select a set of suitable summary statistics (and when the complexity of the model or the size of dataset makes it computationally prohibitive to use full-likelihood methods). In such settings, were we to consider a model choice problem, the divergence exhibited in the current paper would not occur because the measure of distance does not rely on a reduction of the sample.

Results

The special case of Gibbs random fields. Grelaud et al. (2009) showed that, for Gibbs random fields, the computation of the posterior probabilities of the models under competition can be operated by ABC techniques, since they enjoy a converging approximation to the true Bayes factor. The reason for this property is that, in the above ratio [1] and for this special model, $g_1(\mathbf{y}) = g_2(\mathbf{y})$.

The property that validates an ABC resolution for the comparison of Gibbs random fields is that, due to their specific structure, there exists a sufficient statistic vector that runs across models and that allows for an exact (when $\epsilon = 0$) simulation from the posterior probabilities of the models. Each Gibbs random field model has its own sufficient statistic $\eta_m(\cdot)$

and Grelaud et al. (2009) exposed the fact that the vector of statistics $\boldsymbol{\eta}(\cdot) = (\eta_1(\cdot), \ldots, \eta_M(\cdot))$ is also sufficient for the joint parameter $(\mathcal{M}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M)$.

Didelot et al. (2011) point out that this specific property of Gibbs random fields can be extended to any exponential family (hence to any setting enjoying sufficient statistics, see e.g. Casella and Berger, 2001). Their argument is based on an encompassing property: by including all sufficient statistics and all dominating measure statistics in an encompassing model, models under comparison become submodels of the encompassing model. They then conclude that the concatenation of those statistics is jointly sufficient across models. While this encompassing principle holds in full generality, in particular when comparing models that are already embedded, we think it leads to a biased perspective about the merits of ABC for model choice: in practice, most complex models do not enjoy sufficient statistics (if only because they are not exponential families). As demonstrated in the next section and in the normal example in SI, there is more than a mere loss of information due to the use of insufficient statistics. Looking at what happens in the limiting case when one relies on a common model-wise sufficient statistic is a formal but useful study since it brings light on the potentially huge discrepancy between the ABC-based Bayes factor and the true Bayes factor. To develop a solution to the problem in the formal case of the exponential families does not help in the understanding of the discrepancy in non-exponential models.

Arbitrary ratios. The difficulty with the discrepancy between $B_{12}(\mathbf{y})$ and $B_{12}^{\boldsymbol{\eta}}(\mathbf{y})$ is that this discrepancy is impossible to evaluate in a general setting, while there is no reason to expect a reasonable agreement between both quantities. A first illustration was produced by Marin et al. (2011) in the case of MA(q) time series.

A simple illustration of the discrepancy due to the use of a model-wise sufficient statistic is the setting when a sample $\mathbf{y} = (y_1, \ldots, y_n)$ could come from either a Poisson $\mathcal{P}(\lambda)$ distribution or from a geometric $\mathcal{G}(p)$ distribution, already introduced in Grelaud et al. (2009) as a counter-example to Gibbs random fields and later reprocessed in Didelot et al. (2011) to support their sufficiency argument. In this case, the sum

$S = \sum_{i=1}^{n} y_i = \boldsymbol{\eta}(\mathbf{y})$ is a sufficient statistic for both models but not across models. The distribution of the sample given S is a multinomial $\mathcal{M}(S, 1/n, \ldots, 1/n)$ distribution when the data is Poisson, since S is then a Poisson $\mathcal{P}(n\lambda)$ variable, while it is the uniform distribution with constant probability

$$\frac{1}{\binom{n+S-1}{S}} \mathbb{I}_{\sum_i y_i = S} = \frac{S!(n-1)!}{(n+S-1)!} \mathbb{I}_{\sum_i y_i = S}$$

in the geometric case, since S is then a negative binomial $\mathcal{N}eg(n, p)$ variable. The discrepancy ratio is therefore

$$\frac{g_1(\mathbf{y})}{g_2(\mathbf{y})} = \frac{S! n^{-S} / \prod_i y_i!}{1/\binom{n+S-1}{S}}$$

When simulating n Poisson or geometric variables and using prior distributions

$$\lambda \sim \mathcal{E}(1), \quad p \sim \mathcal{U}(0,1),$$

on the respective models, the exact Bayes factor can be evaluated and the range and distribution of the discrepancy are therefore available. Figure 1 gives the range of $B_{12}(\mathbf{y})$ versus $B_{12}^{\boldsymbol{\eta}}(\mathbf{y})$, showing that $B_{12}^{\boldsymbol{\eta}}(\mathbf{y})$ is in this case absolutely unrelated with $B_{12}(\mathbf{y})$: the values produced by both approaches have nothing in common. As noted above, the approximation $B_{12}^{\boldsymbol{\eta}}(\mathbf{y})$ based on the sufficient statistic S is producing figures of the magnitude of a *single* observation, while the true Bayes factor is of the order of the sample size.

The discrepancy between both Bayes factors is in fact increasing with the sample size, as shown by the following result:

Lemma 1. *Consider model selection between model 1: $\mathcal{P}(\lambda)$ with prior distribution $\pi_1(\lambda)$ equal to an $\mathcal{E}(1)$ distribution and model 2: $\mathcal{G}(p)$ with a uniform prior distribution π_2 when the observed data $\mathbf{y}$ consists of iid observations with $\mathbb{E}[y_i] = \theta_0 > 0$. Then $S(\mathbf{y}) = \sum_{i=1}^{n} y_i$ is the minimal sufficient statistic*
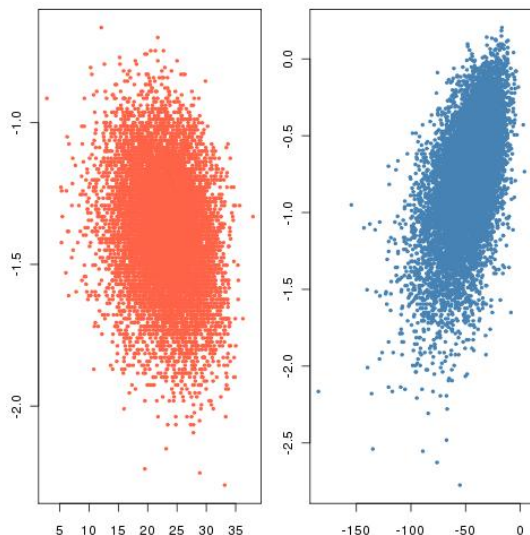
Fig. 1. Comparison between the true log-Bayes factor *(first axis)* for the comparison of a Poisson model versus a negative binomial model and of the log-Bayes factor based on the sufficient statistic $\sum_i y_i$ *(second axis)*, for Poisson *(left)* and negative binomial *(left)* samples of size $n = 50$, based on $T = 10^4$ replications
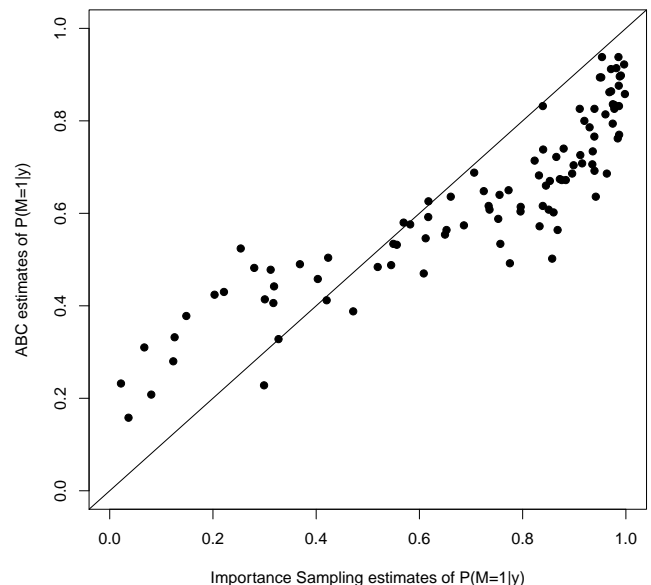
Fig. 2. Comparison of importance sampling and ABC estimates of the posterior probability of scenario 1 in the first population genetic experiment, using 24 summary statistics

for both models and the Bayes factor based on the sufficient statistic $S(\mathbf{y})$, $B_{12}^{\boldsymbol{\eta}}(\mathbf{y})$, satisfies

$$\lim_{n\to\infty} B_{12}^{\boldsymbol{\eta}}(\mathbf{y}) = \frac{(\theta_0+1)^2}{\theta_0} e^{-\theta_0} \quad a.s.$$

Therefore, the Bayes factor based on the sufficient statistic $S(\mathbf{y})$ is *not* consistent; it converges to a non-zero, finite value almost surely.

In this specific setting, Didelot et al. (2011) show that adding $P = \prod_i y_i!$ to the sufficient statistic S induces a statistic (S, P) that is sufficient across both models. While this is a mathematically correct observation, it is not helpful for the understanding of the behaviour of ABC-model choice in realistic settings: outside formal examples as the one above and well-structured although complex exponential families like Gibbs random fields, it is not possible to come up with completion mechanisms that ensure sufficiency across models, It is therefore more fruitful to consider the diverging behaviour of the ABC approximation as given, rather than attempting at solving the problem in a specific case.

Population genetics. We recall that ABC has first been introduced by population geneticists (Beaumont et al., 2002; Marjoram et al., 2003; Pritchard et al., 1999) for statistical inference about the evolutionary history of species, because no likelihood-based approach existed apart from very simple and hence unrealistic situations. This approach has since been used in an increasing number of biological studies (Estoup et al., 2004; Estoup and Clegg, 2003; Fagundes et al., 2007), most of them including model choice. It is therefore crucial to get insights in the validity of such studies, particularly when they deal with species of economical or ecological importance (see, e.g., Lombaert et al., 2010). To this end, we need to compare ABC-based estimates of model posterior probabilities to reliable likelihood-based estimates. Combining different modules based on Stephens and Donnelly (2000), it is possible to approximate the likelihood of population genetic data through importance sampling (IS) in complex scenarios (Cornuet et al., 2009). In order to evaluate the potential discrepancy between ABC-based and likelihood-based posterior probabilities of evolutionary scenarios, we set up two experiments using simulated data and choosing situations in which the choice of a scenario can be problematic. This choice is made in order to provide a wide enough set of intermediate values of model posterior probabilities, so that we better evaluate the divergence between ABC and likelihood estimates.

In the first experiment, we consider two populations (1 and 2) having diverged at a fixed time in the past and a third population (3) having diverged from one of those two populations (scenarios 1 and 2 respectively). Times are set to 60 generations for the first divergence and to 30 generations for the second divergence. One hundred pseudo observed datasets have been simulated, represented by 15 diploid individuals per population genotyped at five independent microsatellite loci. These loci are assumed to evolve according to the strict Stepwise Mutation model (SMM), i.e. when a mutation occurs, the number of repeats of the mutated gene increases or decreases by one unit with equal probability. The mutation rate, common to all five loci, has been set to 0.005 and effective population sizes to 30. In this experiment, both scenarios have a single parameter: the effective population size, assumed to be identical for all three populations. We chose a uniform prior $U[2, 150]$ for this parameter (the true value being 100). The IS algorithm was performed using 100 coalescent trees per particle. The marginal likelihood of both scenarios has been computed for the same set of 1000 particles and they provide
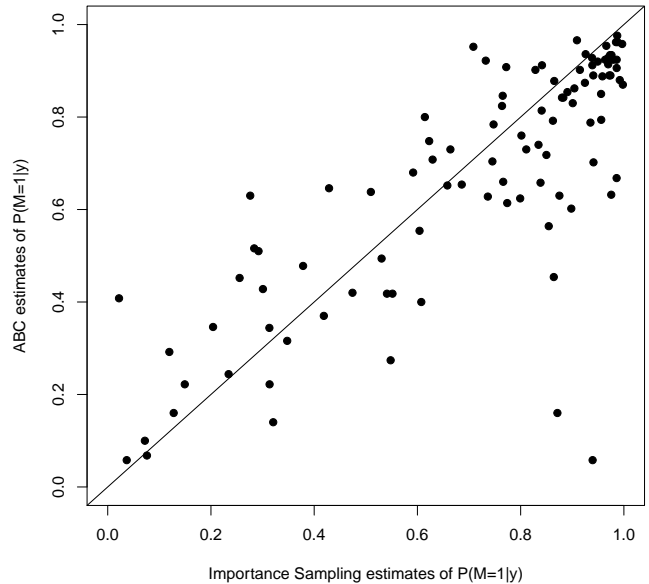
Fig. 3. Same caption as Figure 2 when using 15 summary statistics
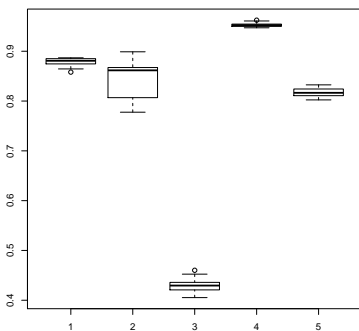
the posterior probability of each scenario. The ABC computations have been performed with DIYABC (Cornuet et al., 2008). A reference table of 2 million datasets has been simulated using 24 usual summary statistics (provided in SI, Table S1) and the posterior probability of each scenario has been estimated as their proportion in the 500 simulated datasets closest to the pseudo observed one.
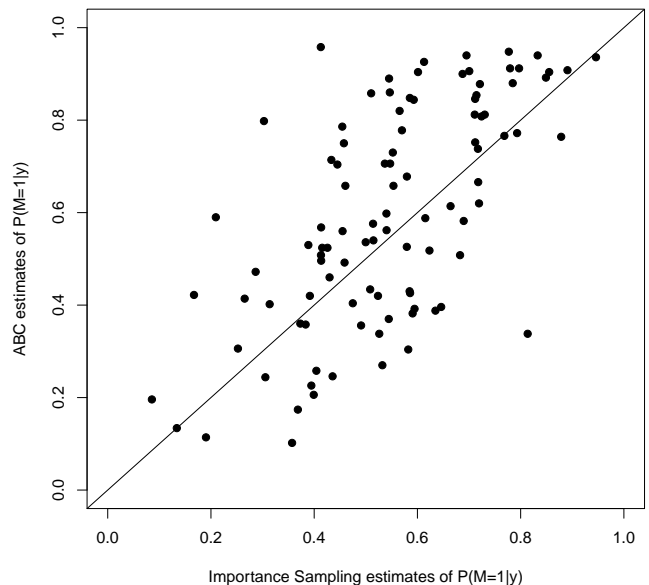
Fig. 4. Comparison of importance sampling and ABC estimates of the posterior probability of scenario 1 in the second population genetic experiment

In the second experiment, we also considered two scenarios including three populations, two of them having diverged 100 generations ago and the third one resulting of a recent admixture between the first two populations (scenario 1) or simply diverging from population 1 (scenario 2) at the same time of 5 generations in the past. In scenario 1, the admixture rate is 0.7 from population 1. Pseudo observed datasets (100) of the same size as in experiment 1 (15 diploid individuals per population, 5 independent microsatellite loci) have been generated assuming an effective population size of 1000 and mutation rates of 0.0005. In contrast with experiment 1, analyses have included the following 6 parameters (provided with the corresponding priors): admixture rate ($U[0.1, 0.9]$), three effective population sizes ($U[200, 2000]$), the time of admixture/second divergence ($U[1, 10]$) and the time of the first divergence ($U[50, 500]$). To account for the higher complexity of the scenarios, the IS algorithm has been performed with
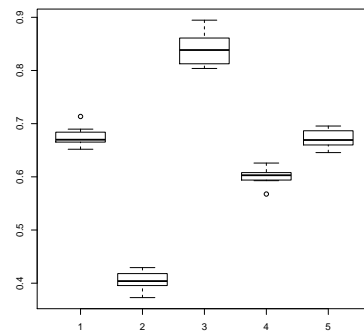
10,000 coalescent trees per particle. Apart from this change, both ABC and likelihood analyses have been performed in the same way as experiment 1.

Figure 2 shows a reasonable fit between the exact posterior probability of model 1 (evaluated by IS) and the ABC approximation in the first experiment on most of the 100 simulated datasets, even though the ABC approximation is almost always biased towards 0.5. When using 0.5 as a boundary for chosing between model 1 and model 2, there is hardly any discrepancy between both approaches, demonstrating that model choice based on ABC can be trusted in this case. Figure 3 considers the same setting when moving from 24 to 15 summary statistics (given in SI, Table S1): the fit degrades quite noticeably. In particular, the number of opposite conclusions in the model choice moves to 12%. In the more complex setting of the second experiment, the discrepancy worsens, as shown on Figure 4. The number of opposite conclusions reaches 26% and the fit between both versions of the posterior probabilities is considerably degraded, with an correlation coefficient of 0.643 between those approximations.

The validity of the importance sampling approximation can obviously be questioned in both experiments, however Figures 5 and 6 in SI display a strong stability of the posterior probability IS approximation across 10 independent runs for 5 different datasets and gives proper confidence in this approach.

Name	Subset	Definition
NAL1	yes	average number of alleles in pop. 1
NAL2	yes	average number of alleles in pop. 2
NAL3	yes	average number of alleles in pop. 3
HET1	yes	average heterozygothy n pop. 1
HET2	yes	average heterozygothy n pop. 2
HET3	yes	average heterozygothy n pop. 3
VAR1	yes	average variance of the allele size in pop. 1
VAR2	yes	average variance of the allele size in pop. 2
VAR3	yes	average variance of the allele size in pop. 3
MGW1	no	Garza-Williamson M in pop. 1
MGW2	no	Garza-Williamson M in pop. 2
MGW3	no	Garza-Williamson M in pop. 3
FST1	no	average FST in pop. 1
FST2	no	average FST in pop. 2
FST3	no	average FST in pop. 3
LIK12	no	probability that sample 1 is from pop. 1
LIK13	no	probability that sample 1 is from pop. 3
LIK21	no	probability that sample 2 is from pop. 1
LIK23	no	probability that sample 2 is from pop. 3
LIK31	no	probability that sample 3 is from pop. 1
LIK32	no	probability that sample 3 is from pop. 2
DAS12	yes	shared allele distance between pop. 1 and 2
DAS13	yes	shared allele distance between pop. 1 and 3
DAS23	yes	shared allele distance between pop. 2 and 3
DM212	yes	distance $(\delta\mu)^2$ between pop. 1 and 2
DM213	yes	distance $(\delta\mu)^2$ between pop. 1 and 3
DM223	yes	distance $(\delta\mu)^2$ between pop. 2 and 2

Discussion

Since its introduction by Tavaré et al. (1997) and Pritchard et al. (1999), ABC has been extensively used in several areas involving complex likelihoods, primarily in population genetics, both for point estimation and testing of hypotheses. In realistic settings, with the exception of Gibbs random fields that satisfy a resilience property with respect to their sufficient statistics, the conclusions drawn on model comparison cannot be trusted *per se* but require further analyses as to the pertinence of the (ABC) Bayes factor based on the summary statistics. This paper has only examined in details the case when the summary statistics are sufficient for both models, while practical situations imply the use of insufficient statistics. We managed to present a realistic if costly comparison of the ABC approximation to the exact Bayes factor in two population genetic settings, showing once more the possibility of significant discrepancies. While this is only a possibility, it nonetheless appears to us an urgent duty to warn the ABC community about the dangers of this approximation, espe-

Fig. 5. Boxplots of the posterior probabilities evaluated over 10 independent Monte Carlo evaluations, for five independent simulated datasets in the first population genetic experiment

Fig. 6. Boxplots of the posterior probabilities evaluated over 10 independent Monte Carlo evaluations, for five independent simulated datasets in the second population genetic experiment

cially when considering the rapidly increasing number of applications estimating posterior probabilities by ABC for conducting model choice.

Further research is needed for producing trustworthy approximations to the posterior probabilities of models. At this stage, unless the whole data is involved in the ABC approximation as in Sousa et al. (2009), our conclusion on ABC-based model choice is to exploit the approximations in an exploratory manner as measures of discrepancies rather than genuine posterior probabilities. This direction relates with the analyses found in Ratmann et al. (2009) and in Drovandi et al. (2011). Furthermore, a version of this exploratory analysis is already provided in the DIY-ABC software of (Cornuet et al., 2008). An option in this software allows for the computation of a Monte Carlo evaluation of false allocation rates resulting from using the ABC posterior probabilities in selecting a model as the most likely. For instance, in the setting of both our population genetic experiments, DIY-ABC gives false allocation rates equal to 20% (under scenarios 1 and 2) and 14.5% and 12.5% (under scenarios 1 and 2), respectively. This evaluation obviously shifts away from the performances of ABC as an aproximation to the posterior probability towards the performances of the whole Bayesian apparatus for selecting a model, but this nonetheless represents a useful and manageable quality assessment for practitionners.

References

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

Beaumont, M., Nielsen, R., Robert, C., Hey, J., Gaggiotti, O., Knowles, L., Estoup, A., Mahesh, P., Coranders, J., Hickerson, M., Sisson, S., Fagundes, N., Chikhi, L., Beerli, P., Vitalis, R., Cornuet, J.-M., Huelsenbeck, J., Foll, M., Yang, Z., Rousset, F., Balding, D., and Excoffier, L. (2010). In defense of model-based inference in phylogeography. *Molecular Ecology*, 19(3):436–446.

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

Belle, E., Benazzo, A., Ghirotto, S., Colonna, V., and Barbujani, G. (2008). Comparing models on the genealogical relationships among Neandertal, Cro-Magnoid and modern Europeans by serial coalescent simulations. *Heredity*, 102(3):218–225.

Berger, J., Fienberg, S., Raftery, A., and Robert, C. (2010). Incoherent phylogeographic inference. *Proc. Nat. Acad. Sci. USA*, 107(41):E57.

Casella, G. and Berger, R. (2001). *Statistical Inference*. Wadsworth, Belmont, CA, second edition.

Cornuet, J.-M., Marin, J.-M., Mira, A., and Robert, C. (2009). Adaptive multiple importance sampling. arXiv:0907.1254.

Cornuet, J.-M., Ravigné, V., and Estoup, A. (2010). Inference on population history and model checking using DNA sequence and microsatellite data with the software DIYABC (v1.0). *BMC Bioinformatics*, 11:401.

Cornuet, J.-M., Santos, F., Beaumont, M. A., Robert, C. P., Marin, J.-M., Balding, D. J., Guillemaud, T., and Estoup, A. (2008). Inferring population history with DIYABC: a user-friendly approach to Approximate Bayesian Computation. *Bioinformatics*, 24(23):2713–2719.

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

Csillèry, K., Blum, M., Gaggiotti, O., and François, O. (2010b). Invalid arguments against ABC: A reply to A.R. Templeton. *Trends in Ecology and Evolution*, 25:490–491.

Didelot, X., Everitt, R., Johansen, A., and Lawson, D. (2011). Likelihood-free estimation of model evidence. *Bayesian Analysis*, 6(1):1–28.

Drovandi, C., Pettitt, A., and Faddy, M. (2011). Approximate Bayesian computation using indirect inference. *J. Royal Statist. Society Series A*, 60(3):503–524.

Estoup, A., Beaumont, M., Sennedot, F., Moritz, C., and Cornuet, J. (2004). Genetic analysis of complex demographic scenarios: spatially expanding populations of the cane toad, Bufo Marinus. *Evolution*, 58(9):2021–2036.

Estoup, A. and Clegg, S. (2003). Bayesian inferences on the recent island colonization history by the bird Zosterops lateralis lateralis. *Mol. Ecol.*, 12:657–674.

Excoffier, C., D., L., and L., W. (2009). Bayesian computation and model selection in population genetics. arXiv:0901.2231.

Fagundes, N., Ray, N., Beaumont, M., Neuenschwander, S., Salzano, F., Bonatto, S., and Excoffier, L. (2007). Statistical evaluation of alternative models of human evolution. *Proc. Nat. Acad. Sci. USA*, 104(45):17614–17619.

Fearnhead, P. and Prangle, D. (2010). Semi-automatic approximate Bayesian computation. arXiv:1004.1112.

Ghirotto, S., Mona, S., Benazzo, A., Paparazzo, F., Caramelli, D., and Barbujani, G. (2010). Inferring genealogical processes from patterns of bronze-age and modern DNA variation in Sardinia. *Mol. Biol. Evol.*, 27(4):875–886.

Grelaud, A., Marin, J.-M., Robert, C., Rodolphe, F., and Tally, F. (2009). Likelihood-free methods for model choice in Gibbs random fields. *Bayesian Analysis*, 3(2):427–442.

Guillemaud, T., Beaumont, M., Ciosi, M., Cornuet, J.-M., and Estoup, A. (2009). Inferring introduction routes of invasive species using approximate Bayesian computation on microsatellite data. *Heredity*, 104(1):88–99.

Jeffreys, H. (1939). *Theory of Probability*. The Clarendon Press, Oxford, first edition.

Lehmann, E. and Casella, G. (1998). *Theory of Point Estimation (revised edition)*. Springer-Verlag, New York.

Leuenberger, C. and Wegmann, D. (2010). Bayesian computation and model selection without likelihoods. *Genetics*, 184(1):243–252.

Lombaert, E., Guillemaud, T., Cornuet, J.-M., Malausa, T., Facon, B., and Estoup, A. (2010). Bridgehead effect in the worldwide invasion of the biocontrol Harlequin Ladybird. *PloS ONE*, 5(3):e9743.

Lopes, J. and Beaumont, M. (2010). ABC: a useful Bayesian tool for the analysis of population data. *Infection, Genetics and Evolution*, 10(6):825–832.

Lopes, J. S., Balding, D., and Beaumont, M. A. (2009). PopABC: a program to infer historical demographic parameters. *Bioinformatics*, 25(20):2747–2749.

Marin, J., Pudlo, P., Robert, C., and Ryder, R. (2011). Approximate Bayesian computational methods. arXiv:1011:0955.

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

Miller, N., Estoup, A., Toepfer, S., Bourguet, D., Lapchin, L., Derridj, S., Kim, K. S., Reynaud, P., Furlan, L., and Guillemaud, T. (2005). Multiple transatlantic introductions of the Western corn rootworm. *Science*, 310(5750):992.

Pascual, M., Chapuis, M., Balanyà, J., Huey, R., Gilchrist, G., Serra, A., and Estoup, A. (2007). Introduction history of Drosophila subobscura in the New World: a microsatellite-based survey using ABC methods. *Molecular Ecology*, 16:3069–3083.

Patin, E., Laval, G., Barreiro, L., Salas, A., Semino, O., Santachiara-Benerecetti, S., Kidd, K., Kidd, J., Van Der Veen, L., Hombert, J., et al. (2009). Inferring the demographic history of African farmers and pygmy hunter-gatherers using a multilocus resequencing data set. *PLoS Genetics*, 5(4):e1000448.

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*, 16:1791–1798.

Ramakrishnan, U. and Hadly, E. (2009). Using phylochronology to reveal cryptic population histories: review and synthesis of 29 ancient DNA studies. *Molecular Ecology*, 18(7):1310–1330.

Ratmann, O., Andrieu, C., Wiujf, C., and Richardson, S. (2009). Model criticism based on likelihood-free inference, with an application to protein network evolution. *Proc. Nat. Acad. Sci. USA*, 106:1–6.

Robert, C. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer-Verlag, New York, second edition.

Sainudiin, R., Thornton, K., Harlow, J., Booth, J., Stillman, M., Yoshida, R., Griffiths, R., McVean, G., and Donnelly, P. (2011). Experiments with the site frequency spectrum. *Bulletin of Mathematical Biology*. (To appear.).

Sousa, V., Fritz, M., Beaumont, M., and Chikhi, L. (2009). Approximate Bayesian computation without summary statistics: the case of admixture. *Genetics*, 181(4):1507–1519.

Stephens, D. and Donnelly, P. (2000). Inference in population genetics (with discussion). *J. Royal Statist. Society Series B*, 62:602–655.

Tavaré, S., Balding, D., Griffith, R., and Donnelly, P. (1997). Inferring coalescence times from DNA sequence data. *Genetics*, 145:505–518.

Templeton, A. (2008). Statistical hypothesis testing in intraspecific phylogeography: nested clade phylogeographical analysis vs. approximate Bayesian computation. *Molecular Ecology*, 18(2):319–331.

Templeton, A. (2010). Coherent and incoherent inference in phylogeography and human evolution. *Proc. Nat. Acad. Sci. USA*, 107(14):6376–6381.

Toni, T. and Stumpf, M. (2010). Simulation-based model selection for dynamical systems in systems and population biology. *Bioinformatics*, 26(1):104–110.

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.

Verdu, P., Austerlitz, F., Estoup, A., Vitalis, R., Georges, M., Théry, S., Froment, A., Le Bomin, S., Gessain, A., Hombert, J.-M., Van der Veen, L., Quintana-Murci, L., Bahuchet, S., and Heyer, E. (2009). Origins and genetic diversity of pygmy hunter-gatherers from western central africa. *Current Biology*, 19(4):312–318.

Wegmann, D. and Excoffier, L. (2010). Bayesian inference of the demographic history of chimpanzees. *Molecular Biology and Evolution*, 27(6):1425–1435.

Wilkinson, R. D. (2008). Approximate Bayesian computation (ABC) gives exact results under the assumption of model error. arXiv:0811.3355.

Appendix

SI Results

A normal illustration. The following reproduces the Poisson geometric illustration in a normal model. If we look at a fully normal $\mathcal{N}(\mu, \sigma^2)$ setting, we have

$$f(\mathbf{y}|\mu) \propto \exp\left\{-n\sigma^{-2}(\bar{y}-\mu)^2/2 - \sigma^{-2}\sum_{i=1}^{n}(y_i-\bar{y})^2/2\right\}\sigma^{-n}$$

hence

$$f(\mathbf{y}|\bar{y}) \propto \exp\left\{-\sigma^{-2}\sum_{i=1}^{n}(y_i-\bar{y})^2/2\right\}\sigma^{-n}\mathbb{I}_{\sum y_i = n\bar{y}}.$$

If we reparameterise the observations into $\mathbf{u} = (y_1 - \bar{y}, \ldots, y_{n-1} - \bar{y}, \bar{y})$, we do get

$$
\begin{aligned}
f(\mathbf{u}|\mu) \propto\ & \sigma^{-n}\exp\left\{-n\sigma^{-2}(\bar{y}-\mu)^2/2\right\} \\
& \times \exp\left\{-\sigma^{-2}\sum_{i=1}^{n-1}u_i^2/2 - \sigma^{-2}\left[\sum_{i=1}^{n-1}u_i\right]^2/2\right\}
\end{aligned}
$$

since the Jacobian is 1. Hence

$$f(\mathbf{u}|\bar{y}) \propto \exp\left\{-\sigma^{-2}\sum_{i=1}^{n-1}u_i^2/2 - \sigma^{-2}\left[\sum_{i=1}^{n-1}u_i\right]^2/2\right\}\sigma^{-n}$$

Considering both models

$$y_1,\ldots,y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu,\sigma_1^2) \quad \text{and} \quad y_1,\ldots,y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu,\sigma_2^2),$$

the discrepancy ratio is then given by

$$\frac{\sigma_2^{n-1}}{\sigma_1^{n-1}}\exp\left\{\frac{\sigma_2^{-2}-\sigma_1^{-2}}{2}\left(\sum_{i=1}^{n-1}(y_i-\bar{y})^2 + \left[\sum_{i=1}^{n-1}(y_i-\bar{y})\right]^2\right)\right\}$$

and is connected with the lack of consistency of the Bayes factor:

Lemma 2. *Consider model selection between model 1: $\mathcal{N}(\mu,\sigma_1^2)$ and model 2: $\mathcal{N}(\mu,\sigma_2^2)$, σ_1 and σ_2 being given, with prior distributions $\pi_1(\mu) = \pi_2(\mu)$ equal to a $\mathcal{N}(0,a^2)$ distribution and when the observed data $\mathbf{y}$ consists of iid observations with finite mean and variance. Then $S(\mathbf{y}) = \sum_{i=1}^{n} y_i$ is the minimal sufficient statistic for both models and the Bayes factor based on the sufficient statistic $S(\mathbf{y})$, $B_{12}^{\eta}(\mathbf{y})$, satisfies*

$$\lim_{n\to\infty} B_{12}^{\eta}(\mathbf{y}) = 1 \quad a.s.$$

Figure 7 illustrates the behaviour of the discrepancy ratio when $\sigma_1 = 0.1$ and $\sigma_2 = 10$, for datasets of size $n = 15$ simulated according to both models. The discrepancy (expressed on a log scale) is once again dramatic, in concordance with the above lemma.

If we now turn to an alternative choice of sufficient statistic, using the pair $(\bar{y}, S^2)$ with

$$S^2 = \sum_{i=1}^{n}(y_i-\bar{y})^2,$$

we follow the solution of Didelot et al. (2011). Using a conjugate prior $\mu \sim \mathcal{N}(0,a^2)$, the true Bayes factor is equal to the Bayes factor based on the corresponding distributions of the pair $(\bar{y}, S^2)$ in the respective models. However, this coincidence does not bring any intuition on the behaviour of the ABC approximations in realistic settings.
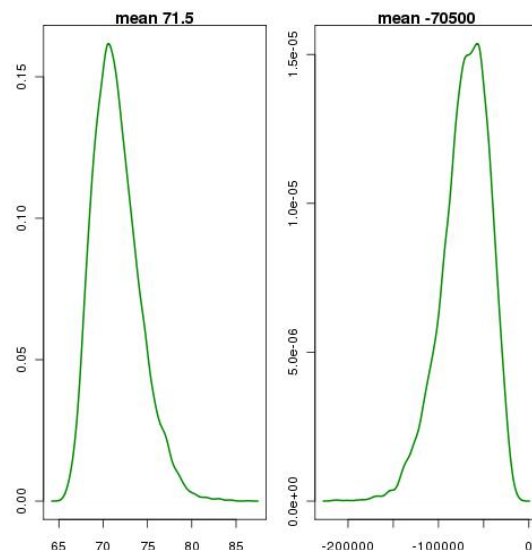
Fig. 7. Empirical distributions of the log discrepancy $\log g_1(\mathbf{y})/g_2(\mathbf{y})$ for datasets of size $n = 15$ simulated from $\mathcal{N}(\mu,\sigma_1^2)$ *(left)* and $\mathcal{N}(\mu,\sigma_2^2)$ *(right)* distributions when $\sigma_1 = 0.1$ and $\sigma_2 = 10$, based on 10^4 replications and a flat prior