Stat Comput (2012) 22:1167–1180
DOI 10.1007/s11222-011-9288-2

Approximate Bayesian computational methods

Jean-Michel Marin · Pierre Pudlo · Christian P. Robert · Robin J. Ryder

Received: 5 January 2011 / Accepted: 5 September 2011 / Published online: 21 October 2011

Abstract Approximate Bayesian Computation (ABC) methods, also known as likelihood-free techniques, have appeared in the past ten years as the most satisfactory approach to intractable likelihood problems, first in genetics then in a broader spectrum of applications. However, these methods suffer to some degree from calibration difficulties that make them rather volatile in their implementation and thus render them suspicious to the users of more traditional Monte Carlo methods. In this survey, we study the various improvements and extensions brought on the original ABC algorithm in recent years.

Keywords Likelihood-free methods · Bayesian statistics · ABC methodology · DIYABC · Bayesian model choice

1 Introduction

Conducting a Bayesian analysis in situations where the likelihood function $\ell(\boldsymbol{\theta}|\mathbf{y})$ is not available raises a computational issue. The likelihood may be unavailable

- for mathematical reasons: it is not available in closed from as a function of $\boldsymbol{\theta}$, or
- for computational reasons: it is too expensive too calculate.

In some specific settings, the likelihood is expressed as an intractable multidimensional integral

$$\ell(\boldsymbol{\theta}|\mathbf{y}) = \int \ell^{\star}(\boldsymbol{\theta}|\mathbf{y}, \mathbf{u})\mathrm{d}\mathbf{u},$$

where $\mathbf{y} \in \mathcal{D} \subseteq \mathbb{R}^n$ is observed, $\mathbf{u} \in \mathbb{R}^p$ a latent vector and $\boldsymbol{\theta} \in \mathbb{R}^d$ the parameter of interest. For instance, when facing coalecent models in population genetics (see, e.g. Tavaré et al. 1997), $\mathbf{y}$ is the genotypes of the present sample, while $\mathbf{u}$ stands for their genealogical tree and the genotypes of their ancestors. In the particular set-up of hierarchical models with partly conjugate priors, it may be that the corresponding conditional distributions can be simulated and this property leads to a Gibbs sampler (Gelfand and Smith 1990). Such a decomposition is not available in general and there is no generic way to implement an MCMC algorithm like the Metropolis–Hastings algorithm (see, e.g., Robert and Casella 2004; Marin and Robert 2007). Typically, the increase in dimension induced by the data augmentation from $\boldsymbol{\theta}$ to $\mathbf{u}$ may be such that the convergence properties of the corresponding MCMC algorithms are too poor for the algorithm to be considered in practice.

In others situations, the normalizing constant of the likelihood $Z_{\boldsymbol{\theta}}$ is unknown

$$\ell(\boldsymbol{\theta}|\mathbf{y}) = \ell_1(\boldsymbol{\theta}|\mathbf{y})/Z_{\boldsymbol{\theta}}.$$

This is typically the case of Gibbs random fields used to model the dependency within spatially correlated data, with applications in epidemiology and image analysis, among

This research was financially supported by the French Agence Nationale de la Recherche grant 'EMILE' ANR-09-BLAN-0145-01, as well as by the Fondation des Sciences Mathématiques de Paris and a GIS scholarship for the fourth author.

J.-M. Marin (✉) · P. Pudlo
I3M, UMR CNRS 5149, Université Montpellier 2, Montpellier, France
e-mail: jean-michel.marin@math.univ-montp2.fr

C.P. Robert · R.J. Ryder
CEREMADE, Université Paris Dauphine and Crest INSEE, Paris, France

others (e.g. Rue and Held 2005). For such models, a solution relying on the simulation of pseudo-samples has been proposed by Møller et al. (2006). However the dependency of this solution on a pseudo-target distribution makes it difficult to calibrate (Cucala et al. 2009; Friel and Pettitt 2008) in general settings.

Bayesian inference thus faces a large class of settings where the likelihood function is not completely known, e.g. $\ell(\boldsymbol{\theta}|\mathbf{y}) = \ell_1(\boldsymbol{\theta}|\mathbf{y})\ell_2(\boldsymbol{\theta})$ with ℓ_2 unknown, and where exact simulation from the corresponding posterior distribution is impractical or even impossible. Such settings call for practical if cruder approximations methods. In the past, Laplace approximations (Tierney and Kadane 1986) and variational Bayes solutions (Jaakkola and Jordan 2000) have been advanced for such problems. However, Laplace approximations require some analytic knowledge of the posterior distribution, while variational Bayes solutions replace the true model with another pseudo-model which is usually much simpler and thus misses some of the features of the original model.

The ABC methodology, where ABC stands for *approximate Bayesian computation*, was mentioned as early as 1984 through a pedagogical and philosophical argument in Rubin (1984). It offers an almost automated resolution of the difficulty with models which are intractable but can be simulated from. It was first proposed in population genetics by Tavaré et al. (1997), who introduced Approximate Bayesian Computation methods as a rejection technique bypassing the computation of the likelihood function via a simulation from the corresponding distribution. The exact version of the method can only be implemented in a small range of cases. Pritchard et al. (1999) produce a generalisation based on an approximation of the target. We study here the foundations as well as the implementation of the ABC method, with illustrations from time series.

This survey describes the genesis of the ABC approach and its justifications (Sect. 2), the calibration of the method (Sect. 3), recent sequential improvements (Sect. 4), post-processing of ABC outputs (Sect. 5), and the specific application of ABC to model choice (Sect. 6). The illustrations of the ABC methodology are based on the posteriors of the MA(2) and MA(1) models for which the true posterior distribution can be computed; the impact of the ABC approximation can thus be assessed. We do not cover the increasingly wide array of applications of ABC here; see Csillèry et al. (2010a) for a survey of implementations of ABC in genomics and ecology. Neither do we address the controversy raised by Templeton (2008, 2010) about the lack of validity of the ABC approach in statistical testing. Answers to those criticisms are provided in Beaumont et al. (2010), Csillèry et al. (2010b), Berger et al. (2010), among others.

2 Genesis of the ABC approach and justifications

Prehistory Rubin (1984) advances a visionary statement that 'Bayesian statistics and Monte Carlo methods are ideally suited to the task of passing many models over one dataset'. Furthermore, he produces in this paper a description of the first ABC algorithm. Followed by Tavaré et al. (1997), the original ABC algorithm is in fact a special case of an accept-reject method (see, e.g., Robert and Casella 2004), where the parameter $\boldsymbol{\theta}$ is generated from the prior $\pi(\boldsymbol{\theta})$ and the acceptance is conditional on the corresponding simulation of a sample being 'almost' identical to the (true) observed sample, which is denoted $\mathbf{y}$ throughout this paper. For the original algorithm given below (and solely for this algorithm), we suppose that $\mathbf{y}$ takes values in a finite or countable set $\mathcal{D}$.

Algorithm 1 Likelihood-free rejection sampler 1

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 $\mathbf{z} = \mathbf{y}$
  set $\boldsymbol{\theta}_i = \boldsymbol{\theta}'$,
end for

It is straightforward to show that the outcome $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_N)$ resulting from this algorithm is an iid sample from the posterior distribution since

$$f(\boldsymbol{\theta}_i) \propto \sum_{\mathbf{z}\in\mathcal{D}} \pi(\boldsymbol{\theta}_i) f(\mathbf{z}|\boldsymbol{\theta}_i) \mathbb{I}_{\mathbf{y}}(\mathbf{z}) = \pi(\boldsymbol{\theta}_i) f(\mathbf{y}|\boldsymbol{\theta}_i)$$
$$\propto \pi(\boldsymbol{\theta}_i|\mathbf{y}).$$

Rubin (1984) does not promote this simulation method in situations where the likelihood is not available but rather exhibits it as an intuitive way to understand posterior distributions from a frequentist perspective, because parameters from the posterior are more likely to be those that could have generated the observed data. (The issue of the zero probability of the exact equality between simulated and observed data in continuous settings is not addressed in the original paper, presumably because the very notion of a 'match' between simulated and observed data is not precisely defined.)

The first ABC In a population genetics setting, Pritchard et al. (1999) extend the above algorithm to the case of continuous sample spaces, producing the first genuine ABC algorithm, defined as follows

Algorithm 2 Likelihood-free rejection sampler 2

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 $\rho\{\eta(\mathbf{z}), \eta(\mathbf{y})\} \leq \varepsilon$
  set $\boldsymbol{\theta}_i = \boldsymbol{\theta}'$,
end for

where the parameters of the algorithm are

- η, a function on $\mathcal{D}$ defining a statistic which most often is not sufficient,
- $\rho > 0$, a distance on $\eta(\mathcal{D})$,
- $\varepsilon > 0$, a tolerance level.

The likelihood-free algorithm above thus samples from the marginal in $\mathbf{z}$ of the joint distribution

$$\pi_\varepsilon(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y}) = \frac{\pi(\boldsymbol{\theta}) f(\mathbf{z}|\boldsymbol{\theta}) \mathbb{I}_{A_{\varepsilon,\mathbf{y}}}(\mathbf{z})}{\int_{A_{\varepsilon,\mathbf{y}} \times \boldsymbol{\theta}} \pi(\boldsymbol{\theta}) f(\mathbf{z}|\boldsymbol{\theta}) \mathrm{d}\mathbf{z}\mathrm{d}\boldsymbol{\theta}}, \tag{1}$$

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

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

The basic idea behind ABC is that using a representative (enough) summary statistic η coupled with a small (enough) tolerance ε should produce a good (enough) approximation to the posterior distribution, namely that

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

Before moving to the extensions of the above algorithm, let us consider a simple dynamic example.

Example The MA(q) process is a stochastic process $(y_k)_{k \in \mathbb{N}^*}$ defined by

$$y_k = u_k + \sum_{i=1}^{q} \theta_i u_{k-i}, \tag{2}$$

where $(u_k)_{k \in \mathbb{Z}}$ is an iid sequence of standard Gaussians $\mathcal{N}(0, 1)$. Even though a Bayesian analysis can handle non-identifiable settings and still estimate properly identifiable quantities (see, e.g., Marin and Robert 2007, Chap. 5), we will impose a standard identifiability condition on this model, namely that the roots of the polynomial

$$\mathcal{Q}(x) = 1 - \sum_{i=1}^{q} \theta_i x^i$$

are all outside the unit circle in the complex plane. A simple prior distribution is therefore the uniform distribution over the corresponding range of θ_i's, especially when q is small and the set of resulting parameters is easy to describe. In the case processed in the figures below for $q = 2$, we obtain the triangle

$$-2 < \theta_1 < 2, \qquad \theta_1 + \theta_2 > -1, \qquad \theta_1 - \theta_2 < 1.$$

Although the prior on $\boldsymbol{\theta}$ is very simple, and despite the Gaussian nature of the random variables, the likelihood associated with a series $(y_k)_{1 \leq k \leq n}$ is more complex because of the need to integrate out $u_{-q+1}, \ldots, u_{-1}, u_0$. (The easier alternative is to condition on $(y_k)_{1 \leq k \leq q}$, see Marin and Robert 2007, even though the general case can also be handled by MCMC simulations as the likelihood is available, at least for small values of n.)

Running one iteration of ABC in this setting then simply requires

(a) simulating the MA(q) coefficients θ uniformly over the acceptable range,
(b) generating an iid sequence $(u_k)_{-q < k \leq n}$,
(c) producing a simulated series $(z_k)_{1 \leq k \leq n}$.

Depending on the focus of the analysis, the distance can be the raw distance between the series

$$\rho^2\{(z_k)_{1 \leq k \leq n}, (y_k)_{1 \leq k \leq n}\} = \sum_{k=1}^{n} (y_k - z_k)^2$$

or the quadratic distance between summary statistics like the first q autocovariances

$$\tau_j = \sum_{k=j+1}^{n} y_k y_{k-j}$$

which is our choice for the illustration provided in Fig. 1. This experiment shows how an ABC sample fits the level sets of the true posterior density for a simulated sample of length 100 using the parameters $(\theta_1, \theta_2) = (0.6, 0.2)$ and a tolerance level equal to the 0.1% quantile of the sample of the distances. (The level sets were computed from the exact likelihood for the MA(2) model and a grid of values of $\boldsymbol{\theta}$ over the acceptable range.) This plot illustrates how the distribution of the sample points departs from true posterior: the approximation does not reconstruct the posterior perfectly. Decreasing ε would lead to a better concentration of the posterior density on the level sets, but at the expense of the size of the resulting sample or at a higher computing cost.

MCMC-ABC In practice, using simulations from a non informative prior distribution $\pi(\cdot)$ is very inefficient because this does not account for the data at the proposal stage and thus leads to proposed values located in low posterior probability regions. As an answer to this problem, Marjoram et al.

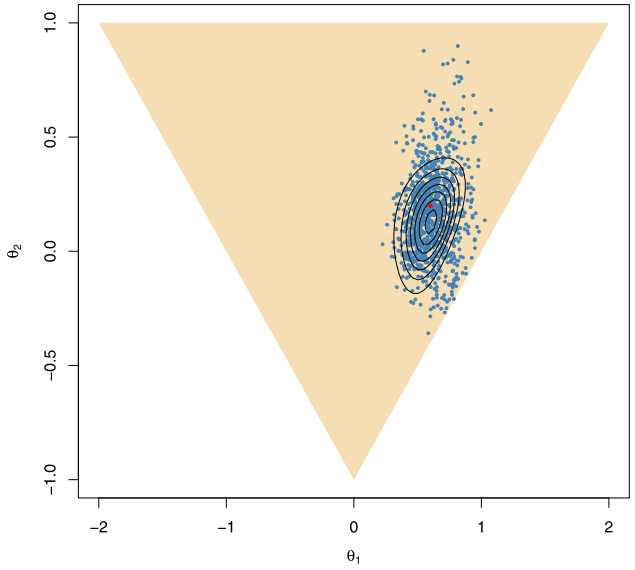

Fig. 1 (Color online) Comparison of the level sets (*in black*) of the true posterior distribution with the scatter plot (*in blue*) of an ABC sample when using autocovariances as summary statistics. The threshold ε is chosen so that 0.1% of the $N = 10^6$ simulated datasets are accepted. The observed dataset has been drawn from an MA(2) model with $n = 100$ epochs and parameter $\boldsymbol{\theta} = (0.6, 0.2)$ (*the red dot*). The *triangle* is the range of acceptable values of $\boldsymbol{\theta}$

(2003) introduce an MCMC-ABC algorithm (Algorithm 3) targeting the approximate posterior distribution π_ε of (1).

Algorithm 3 Likelihood-free MCMC sampler

Use Algorithm 2 to get a realisation $(\boldsymbol{\theta}^{(0)}, \mathbf{z}^{(0)})$ from the ABC target distribution $\pi_\varepsilon(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y})$
for $t = 1$ to N **do**
  Generate $\boldsymbol{\theta}'$ from the Markov kernel $q(\cdot|\boldsymbol{\theta}^{(t-1)})$,
  Generate $\mathbf{z}'$ from the likelihood $f(\cdot|\boldsymbol{\theta}')$,
  Generate u from $\mathcal{U}_{[0,1]}$,
  if $u \leq \frac{\pi(\boldsymbol{\theta}')q(\boldsymbol{\theta}^{(t-1)}|\boldsymbol{\theta}')}{\pi(\boldsymbol{\theta}^{(t-1)})q(\boldsymbol{\theta}'|\boldsymbol{\theta}^{(t-1)})}$ and $\rho\{\eta(\mathbf{z}'), \eta(\mathbf{y})\} \leq \varepsilon$ **then**
    set $(\boldsymbol{\theta}^{(t)}, \mathbf{z}^{(t)}) = (\boldsymbol{\theta}', \mathbf{z}')$
  else
    $(\boldsymbol{\theta}^{(t)}, \mathbf{z}^{(t)}) = (\boldsymbol{\theta}^{(t-1)}, \mathbf{z}^{(t-1)})$,
  end if
end for

The acceptance probability used in Algorithm 3 does not involve the calculation of the likelihood and it thus satisfies ABC requirements. It also produces an MCMC algorithm which exactly targets $\pi_\varepsilon(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y})$ as its stationary distribution. Indeed,

$$\frac{\pi_\varepsilon(\boldsymbol{\theta}', \mathbf{z}'|\mathbf{y})}{\pi_\varepsilon(\boldsymbol{\theta}^{(t-1)}, \mathbf{z}^{(t-1)}|\mathbf{y})} \times \frac{q(\boldsymbol{\theta}^{(t-1)}|\boldsymbol{\theta}')\, f(\mathbf{z}^{(t-1)}|\boldsymbol{\theta}^{(t-1)})}{q(\boldsymbol{\theta}'|\boldsymbol{\theta}^{(t-1)})\, f(\mathbf{z}'|\boldsymbol{\theta}')}$$

$$= \frac{\pi(\boldsymbol{\theta}')\, f(\mathbf{z}'|\boldsymbol{\theta}')\mathbb{I}_{A_{\varepsilon,\mathbf{y}}}(\mathbf{z}')}{\pi(\boldsymbol{\theta}^{(t-1)})\, f(\mathbf{z}^{(t-1)}|\boldsymbol{\theta}^{(t-1)})\mathbb{I}_{A_{\varepsilon,\mathbf{y}}}(\mathbf{z}^{(t-1)})} \times \frac{q(\boldsymbol{\theta}^{(t-1)}|\boldsymbol{\theta}')\, f(\mathbf{z}^{(t-1)}|\boldsymbol{\theta}^{(t-1)})}{q(\boldsymbol{\theta}'|\boldsymbol{\theta}^{(t-1)})\, f(\mathbf{z}'|\boldsymbol{\theta}')}$$

$$= \frac{\pi(\boldsymbol{\theta}')q(\boldsymbol{\theta}^{(t-1)}|\boldsymbol{\theta}')}{\pi(\boldsymbol{\theta}^{(t-1)})q(\boldsymbol{\theta}'|\boldsymbol{\theta}^{(t-1)})}\mathbb{I}_{A_{\varepsilon,\mathbf{y}}}(\mathbf{z}').$$

The initialisation of the MCMC sampler with the rejection sampler (Algorithm 2) can be bypassed since the Markov chain forgets its initial state. The computational cost of the initialisation is then reduced. But then we have to run the MCMC longer to achieve convergence and omit the burn-in first iterations from the output, which also has a computational cost.

As noted above, the ABC approximation depends on tuning parameters (the summary statistic η, the tolerance ε, and the distance ρ) that have to be chosen prior to running the algorithm and the calibration of which is discussed in most of the literature. The tolerance ε is somewhat the easiest aspect of this calibration issue in that, when ε goes to zero, the ABC algorithm becomes exact.

Noisy ABC Wilkinson (2008) proposes to switch perspective, replacing the approximation error resulting from the loose acceptance condition in the above likelihood-free samplers with an exact inference from a controlled approximation of the target, essentially a convolution of the regular target with an arbitrary kernel function. The corresponding ABC target is thus

$$\pi_\varepsilon(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y}) = \frac{\pi(\boldsymbol{\theta})\, f(\mathbf{z}|\boldsymbol{\theta})\, K_\varepsilon(\mathbf{y} - \mathbf{z})}{\int \pi(\boldsymbol{\theta})\, f(\mathbf{z}|\boldsymbol{\theta})\, K_\varepsilon(\mathbf{y} - \mathbf{z})\mathrm{d}\mathbf{z}\mathrm{d}\boldsymbol{\theta}}, \tag{3}$$

where K_ε is a well-chosen kernel parameterised by the bandwidth ε. This perspective is interesting in that the outcome is completely controlled, due to the degree of freedom brought by the choice of the kernel. Wilkinson (2008) makes the valuable point that if the model includes an error term, then taking the distribution of that error term to be K_ε leads to an ABC algorithm which simulates exactly from the error-in-variables posterior. In practice, Wilkinson's (2008) approach requires a modification of the standard ABC algorithms, taking into account the kernel K_ε for the simulation of $\mathbf{z}$. The new algorithm which includes an accept-reject step imposes an upper bound on the convolution kernel K_ε.

This perspective of the "noisy ABC" is also adopted by Fearnhead and Prangle (2010) who study the convergence of ABC based inference. They show that the convolution induced by the kernel representation leads to the true parameter being the maximum of the integrated log-likelihood and thus that a Bayes estimator is converging to the true value when the number of observations goes to infinity and the

tolerance level goes to zero. They also stress the connection with the econometrics approach of indirect inference (Gouriéroux et al. 1993).

ABC filtering Jasra et al. (2011) propose an ABC scheme for filtering when the distribution of the observables conditioned on the hidden state is not available pointwise, related to the convolution particle filter of Campillo and Rossi (2009). It is particularly appealing in that it allows complex (hence realistic) statistical models for filtering. Theoretical arguments are given to prove that the ABC approximation of the filter does not accumulate errors along the sequence of observables, when the model has good mixing properties. Dean et al. (2011) illustrate this implementation in the specific case of hidden Markov (HMM) models, relating the ABC implementation with Wilkinson's (2008) perspective and demonstrating that the pseudo (or noisy) model for which ABC is exact also is an HMM. Using this representation, they further establish ABC consistency. While Dean et al. (2011) establish that ABC leads to an asymptotic bias for a fixed value of the tolerance ε, they also prove that an arbitrary accuracy can be attained with enough data and a small enough ε. We note that the restriction to summary statistics that preserve the HMM structure is paramount for the results in the paper to apply, hence preventing the use of truly summarizing statistics that would not grow in dimension with the size of the HMM series. We also note that the approach of Jasra et al. (2011) has been rediscovered by Calvet and Czellar (2011) and that there are some similar ideas in Gauchi and Vila (2011). Finally, the convergence result central to Dean et al. (2011) is also connected with Fearnhead and Prangle's (2010) version, mentioned above, in that they both rely on pseudo-likelihood consistency arguments.

3 Calibration of ABC

Summary statistics Several authors have considered the fundamental difficulty associated with the choice of the summary statistic, $\eta(\mathbf{y})$, which one would like to consider as a quasi-sufficient statistic. First, for most real problems (a notable exception being found in Grelaud et al. 2009 in the case of Gibbs random fields), it is impossible to find sufficient statistics. Second, the summary statistics of interest are usually determined by the problem at hand and chosen by the experimenters in the field.

Assuming a large collection of summary statistics is available, Joyce and Marjoram (2008) consider the sequential inclusion of those statistics into the ABC target. The inclusion of a new statistic within the set of summary statistics is assessed in terms of a likelihood ratio test, without taking into account the sequential nature of the tests. We have reservations about the method, first and foremost that the construction of the statistics is not discussed, while the method is not independent from parametrisation, and also that the order in which the statistics are considered is paramount for their inclusion/exclusion. A regularisation of the method proposed at the end of the paper is to use a forward-backward selection mechanism to address this last issue. However, this correction does not address another issue, namely the impact of the correlation between the summary statistics. Note at last that Joyce and Marjoram's (2008) method still depends on an approximation factor that needs to be calibrated prior to running the algorithm. In his thesis, Ratmann (2009) proposes a similar examination of the successive inclusion of various statistics.

A related perspective is that of McKinley et al. (2009). They perform a simulation experiment comparing ABC-MCMC and ABC-SMC (discussed below) with regular data augmentation MCMC. The authors test strategies to select the tolerance level, and to choose the distance ρ and the summary statistics. The conclusions are not very surprising, in that

(a) repeating simulations of the data points given one simulated parameter does not seem to contribute to an improved approximation of the posterior by the ABC sample,
(b) the tolerance level does not seem to have a strong influence,
(c) the choice of the distance, of the summary statistics and of the calibration factors are paramount to the success of the approximation, and
(d) ABC-SMC outperforms ABC-MCMC (MCMC remaining the reference).

Fearnhead and Prangle (2010) study the selection of summary statistics with the interesting perspective that ABC is then considered from a purely inferential viewpoint and calibrated for estimation purposes. (This contrasts with most alternative perspectives that envision ABC as a poor man's non-parametric estimation of the posterior distribution.) Fearnhead and Prangle (2010) rely on a randomised version of the summary statistics from which they derive a well-calibrated version of ABC, *i.e.* an algorithm that gives proper predictions of given quantities. The authors consider choices of summary statistics, and establish that the posterior expectations of the parameters of interest are optimal summary statistics, although this follows from their choice of loss function.

Tolerance threshold and ABC approximation error As noted above, the choice of the tolerance level ε is mostly a matter of computational power: smaller ε's are associated with higher computational costs and the standard practice (Beaumont et al. 2002) is to select ε as a small percentile

of the simulated distances $\rho\{\eta(\mathbf{z}), \eta(\mathbf{y})\}$. An alternative described below is to set the ABC algorithm within the non-parametric setting of density estimation, in which case ε is understood as a bandwidth and can be derived from the simulated population. As noted in Fearnhead and Prangle (2010), this perspective implies that the optimal ε is then different from zero.

Standing rather apart from other contributions to the field, Ratmann et al. (2009) provide an intrinsically novel way of looking at the ABC approximation error (and hence at the tolerance). It is presented as a tool assessing the goodness of fit of a given model. The fundamental idea there is to use the tolerance ε as an additional parameter of the model, simulating from a joint posterior distribution

$$f(\boldsymbol{\theta}, \varepsilon|\mathbf{y}) \propto \xi(\varepsilon|\mathbf{y}, \boldsymbol{\theta})\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta})\pi_{\varepsilon}(\varepsilon),$$

where $\xi(\varepsilon|\mathbf{y}, \boldsymbol{\theta})$ plays the role of the likelihood, and $\pi_{\boldsymbol{\theta}}$ and π_{ε} are the corresponding priors on $\boldsymbol{\theta}$ and ε. In this approach, $\xi(\varepsilon|\mathbf{y}, \boldsymbol{\theta})$ is the prior predictive density of $\rho\{\eta(\mathbf{z}), \eta(\mathbf{y})\}$ given $\boldsymbol{\theta}$ and $\mathbf{y}$ when $\mathbf{z}$ is distributed from $f(\mathbf{z}|\boldsymbol{\theta})$. We note here a connection with Wilkinson's (2008) target (3) in that $\pi(\boldsymbol{\theta})f(\mathbf{z}|\boldsymbol{\theta})K_{\varepsilon}(\mathbf{y} - \mathbf{z})$ is identical to the above once we replace $\mathbf{y} - \mathbf{z}$ by ε.

Ratmann et al. (2009) then derive an ABC algorithm they call ABC_{μ} to simulate an MCMC chain targeting this joint distribution, replacing $\xi(\varepsilon|\mathbf{y}, \boldsymbol{\theta})$ with a non-parametric kernel approximation. For each model under comparison, the marginal posterior distribution on the error ε is then used to assess the fit of the model, the logic being that this posterior should include 0 in a reasonable credible interval. While the authors stress they use the data once, they also define the above target by using simultaneously a prior distribution on ε and a conditional distribution on the same ε that they interpret as the likelihood in $(\varepsilon, \boldsymbol{\theta})$. The product is most often defined as a density in $(\varepsilon, \boldsymbol{\theta})$, so it can be simulated from, but the Bayesian interpretation of the outcome is delicate, especially because it seems the prior on ε contributes significantly to the final assessment of the model. As discussed in Robert et al. (2010), some of the choices of Ratmann et al. (2009) can be argued about, in particular the ambivalent role of the approximation error. The most important aspect of the paper is that the original motivation of running ABC for conducting inference on the parameters of a model is replaced by the alternative goal of running ABC for assessing a model; see Ratmann et al.'s (2010) reply to the remarks made by Robert et al. (2010).

Example Returning to the MA(2) model, we study the impact of the choice of the distance and of the tolerance on the approximation. In this example, we simulated a sample of size 50 from a MA(2) model based on the same parameters as above. First, we compare the impact of using the raw distance between the complete datasets instead of the distance between the autocovariances (introduced above). Figure 2 shows that the raw distance between the observed and the simulated time series is inefficient and fairly non-discriminative. For the raw distance, the spread of the parameters accepted after the ABC step is indeed much wider than for the second distance, especially when compared with the level sets of the posterior density. We thus use only the distance between the autocovariances in the remainder of the paper.

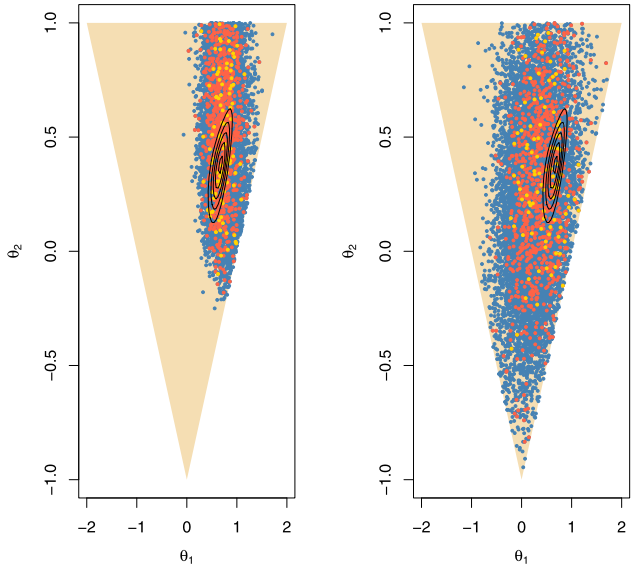

Fig. 2 (Color online) Scattering of two ABC samples when the computations are based on the autocovariance distance (*left*) and the raw distance (*right*), using different quantiles on the simulated distance for ε (1% *in blue*, 1‰ *in red*, and 0.1‰ *in yellow*). The level sets of the posterior density are exhibited *in black*

We now turn to the tolerance ε. Figure 3 shows that decreasing ε along empirical quantiles of the simulated distances $\rho(\eta(\mathbf{z}), \eta(\mathbf{y}))$ improves the approximation, although we never reach the true marginal densities (this is particularly true for the parameter θ_2). The marginal densities of the ABC samples were obtained by the R default density estimator and the true marginal densities by numerical integration.

4 Sequential improvements

Importance sampling Sequential techniques can enhance the efficiency of the ABC algorithm by learning about the target distribution, as in Sisson et al.'s (2007) partial rejection control (PRC) version. The ABC-PRC modification introduced by Sisson et al. (2007) consists of producing samples $(\boldsymbol{\theta}_1^{(t)}, \ldots, \boldsymbol{\theta}_N^{(t)})$ at each iteration $1 \leq t \leq T$ of the algorithm by using a particle filter methodology. Starting with a regular ABC step, the generation of the $\boldsymbol{\theta}_i^{(t)}$'s relies on

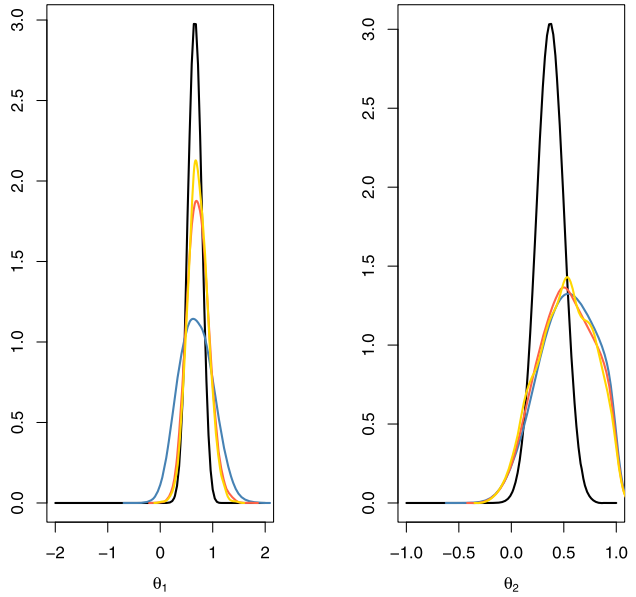

Fig. 3 (Color online) Evolution of the distribution of ABC samples using different quantiles for ε (10% *in blue*, 1% *in red*, and 0.1% *in yellow*) when compared with the true marginal densities. The dataset is the same as in Fig. 2

Markov transition kernels K_t,

$$\boldsymbol{\theta}_i^{(t)} \sim K_t(\boldsymbol{\theta}|\boldsymbol{\theta}^\star),$$

until $\mathbf{z} \sim f(\mathbf{z}|\boldsymbol{\theta}_i^{(t)})$ is such that $\rho(\eta(\mathbf{z}), \eta(\mathbf{y})) \leq \varepsilon$, where $\boldsymbol{\theta}^\star$ is selected at random among the previous $\boldsymbol{\theta}_i^{(t-1)}$'s with probabilities $\omega_i^{(t-1)}$. The probability $\omega_i^{(t)}$ is derived by an importance sampling argument as

$$\omega_i^{(t)} \propto \frac{\pi(\boldsymbol{\theta}_i^{(t)}) L_{t-1}(\boldsymbol{\theta}^\star|\boldsymbol{\theta}_i^{(t)})}{\pi(\boldsymbol{\theta}^\star) K_t(\boldsymbol{\theta}_i^{(t)}|\boldsymbol{\theta}^\star)},$$

where L_{t-1} is an arbitrary transition kernel. While this method is based upon the theoretical work of Del Moral et al. (2006) and their SMC sampler, the application to approximate Bayesian computation results in a bias in the approximation to the posterior, because the likelihood is removed in a standard ABC fashion (Sisson et al. 2009). Replacing the likelihood with the indicator function provides an unbiased estimator of the likelihood that cannot be used as such in the denominator of a Metropolis–Hastings acceptance probability, hence the resulting bias.

An alternative version called ABC-PMC and based on genuine importance sampling arguments, proposed by Beaumont et al. (2009), bypasses this difficulty, in connection with the population Monte Carlo method of Douc et al. (2007). It includes an automatic scaling of the forward kernel. The correction published in Sisson et al. (2009) acknowledges the existence of a bias and suggests a correction essentially identical to the PMC solution of Beaumont et al. (2009).

As illustrated in the pseudo-code below, ABC-PMC constructs a kernel approximation to the target distribution based on earlier simulations and estimates the random walk scale (which is also the kernel bandwidth) from those simulations, using in addition a decreasing sequence of tolerance thresholds $\varepsilon_1 \geq \cdots \geq \varepsilon_T$:

Algorithm 4 Likelihood-free population Monte Carlo sampler

At iteration $t = 1$,
for $i = 1$ to N **do**
  repeat
    Simulate $\boldsymbol{\theta}_i^{(1)} \sim \pi(\boldsymbol{\theta})$ and $\mathbf{z} \sim f(\mathbf{z} \mid \boldsymbol{\theta}_i^{(1)})$
  until $\rho(\eta(\mathbf{z}), \eta(\mathbf{y})) \leq \varepsilon_1$
  Set $\omega_i^{(1)} = 1/N$
end for
Take Σ_1 as twice the empirical variance of the $\boldsymbol{\theta}_i^{(1)}$'s
for $t = 2$ to T **do**
  for $i = 1$ to N **do**
    repeat
      Pick $\boldsymbol{\theta}_i^\star$ from the $\boldsymbol{\theta}_j^{(t-1)}$'s with probabilities $\omega_j^{(t-1)}$
      Generate $\boldsymbol{\theta}_i^{(t)} \sim \mathcal{N}(\boldsymbol{\theta}_i^\star, \Sigma_{t-1})$ and $\mathbf{z} \sim f(\mathbf{z} \mid \boldsymbol{\theta}_i^{(t)})$
    until $\rho(\eta(\mathbf{z}), \eta(\mathbf{y})) \leq \varepsilon_t$
    Set
    $\omega_i^{(t)} \propto \pi(\boldsymbol{\theta}_i^{(t)}) / \sum_{j=1}^N \omega_j^{(t-1)} \varphi\{\Sigma_{t-1}^{-1/2}(\boldsymbol{\theta}_i^{(t)} - \boldsymbol{\theta}_j^{(t-1)})\}$
  end for
  Take Σ_t as twice the weighted variance of the $\boldsymbol{\theta}_i^{(t)}$'s
end for

Another related paper is Toni et al.'s (2009) proposal of a parallel sequential ABC algorithm. Just like ABC-PMC, the ABC-SMC algorithm (an acronym found in several papers) developed therein is based on a sequence of simulated samples, Markov transition kernels, and importance weights rather than SMC justifications. The unavailable likelihood is estimated by the indicator of the tolerance zone or an average of indicators as in Marjoram et al. (2003). The bulk of the paper is dedicated to the analysis of ODEs, using uniform distributions as transition kernels. The adaptivity of the ABC-SMC algorithm is restricted to a progressive reduction of the tolerance, ε_t, since the kernels K_t's remain the same across iterations, in contrast with the ABC-PMC motivation for tuning the K_t's to the target. The paper also contains a comparison with ABC-PRC, which shows a bias in the variance of the ABC-PRC output, in line with Beaumont et al. (2009).

McKinley et al. (2009) have coded the parallel sequential ABC algorithm on an infectious disease model (a recent outbreak of Ebola Haemorrhagic Fever in the Democratic Republic of the Congo—for which there is no known treatment and which is responsible for an 88% decline in observed chimpanzee populations since 2003!). They show that the

ABC-SMC sampler outperforms ABC-MCMC (MCMC remaining the reference). The comparison experiment is based on a single dataset, with fixed random walk variances for the MCMC algorithms; note that the prior used in the simulation might be too highly peaked around the true value (gamma rates of 0.1). Some of the ABC scenarios do produce estimates that are rather far away from the references given by MCMC, for instance CABC-MCMC when the threshold ε is 10 and the number of repeats R is 100.

Backward kernels and SMC Del Moral et al. (2011) exhibit the connection between the ABC algorithm and the foundational SMC paper of Del Moral et al. (2006) that inspired Sisson et al. (2007). As opposed to the latter, and despite a common framework, this ABC-SMC paper properly relies on the idea of using backward kernels L_t to simplify the importance weights and to remove from these weights the dependence on the unknown likelihood. A major assumption of Del Moral et al. (2011) is that the forward kernels K_t are supposed to be invariant against the true target (which is a tempered-like version of the true posterior in sequential Monte Carlo), a choice not explicitely made in Sisson et al. (2007). One of the novelties in the paper is that the authors rely on M repeated simulations of the pseudo-data $\mathbf{z}$ given the parameter, rather than using a single simulation. In that perspective, each simulated parameter gets a non-zero weight that is proportional to the number of accepted $\mathbf{z}$'s. The limiting case $M \to \infty$ brings in an exact simulation from the tempered targets π_{ε_t}'s, so there is a convergence principle and the stabilisation of the approximation could be assessed to calibrate M. The adaptivity in the ABC-SMC algorithm is found in the on-line construction of the thresholds: the thresholds decrease slowly enough to keep a large number of accepted transitions from the previous sample. An important feature is that the update in the importance weights simplifies to the ratio of the proportions of surviving particles, due to the choice of the reversal backward kernels L_t and to the use of invariant transition forward kernels K_t.

In a very related manner, Drovandi and Pettitt (2010) use a combination of particles and of MCMC moves to adapt a proposal to the true target, with acceptance probability

$$\min\left\{1, \frac{\pi(\boldsymbol{\theta}^*)K(\boldsymbol{\theta}_c|\boldsymbol{\theta}^*)}{\pi(\boldsymbol{\theta}^*)K(\boldsymbol{\theta}^*|\boldsymbol{\theta}_c)}\right\}$$

where $\boldsymbol{\theta}^*$ is the proposed value, $\boldsymbol{\theta}_c$ is the current value (picked at random from the particle population), and K is a proposal kernel used to simulate the proposed value. The algorithm is adaptive in that the previous population of particles is used to make the choice of the proposal K, as well as of the tolerance level ε_t. The level of novelty of the method compared with Del Moral et al. (2011) is quite limited, since the paper adapts the tolerance on-line as an α-quantile of the previous particle population. The convergence analysis which is omitted by Drovandi and Pettitt (2010) is perhaps not so standard, mainly because the MCMC is applied only to half of the particle system. The only strong methodological difference between the two papers is that the MCMC steps are now repeated 'numerous times'. However, this partly cancels the appeal of an $O(N)$ order method versus the $O(N^2)$ order ABC-PMC and ABC-SMC methods. An interesting remark there is that advances are needed in cases when simulating the pseudo-observations is very costly, as in Ising models. However, replacing exact simulation by a few steps from a Gibbs sampler as in Grelaud et al. (2009) cannot be very detrimental to the convergence of an approximate algorithm.

5 Post-processing of ABC output

Local linear regression Improvements to the general ABC scheme have been achieved by viewing the problem as a conditional density estimation and developing techniques to allow for larger ε (Beaumont et al. 2002). This is a post-processing scheme in that the simulation process *per se* does not change but the analysis of the ABC output does. The authors endeavour to include all simulated summary statistics, even those far away from the observed summary statistic, by shrinking the corresponding parameters in a linear manner. More specifically, they replace the simulated $\boldsymbol{\theta}$'s with

$$\boldsymbol{\theta}^* = \boldsymbol{\theta} - \{\eta(\mathbf{z}) - \eta(\mathbf{y})\}^{\mathrm{T}}\hat{\beta},$$

where $\hat{\beta}$ is obtained by a weighted least squares regression of $\boldsymbol{\theta}$ on $(\eta(\mathbf{z}) - \eta(\mathbf{y}))$, using weights of the form

$$K_\delta\left\{\rho\{\eta(\mathbf{z}), \eta(\mathbf{y})\}\right\},$$

where K_δ is a non-parametric kernel with bandwidth δ.

Example We implement this correction of Beaumont et al. (2002) in the MA(2) model, again using the first two autocovariances as summary statistic $\eta(\mathbf{z})$, and we apply a non-parametric local regression based on the Epanechnikov kernel (Wasserman 2007). We keep δ equal to the value of the tolerance ε used in the regular ABC scheme. Figures 4 and 5 summarise the results. When using a 0.1% quantile, the two density estimates are identical in the case of the parameter θ_2. The post-processed density estimate of θ_1 is closer to the true posterior. When using a 20% quantile, the impact of the local regression is more spectacular. We recover results close to those obtained with the 0.1% quantile. This exhibits the point that local regression strongly attenuates the impact of the truncation brought by ε.

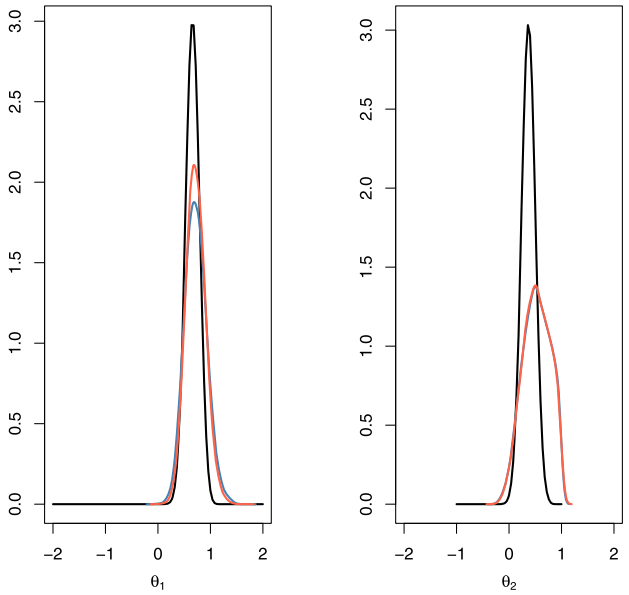

Fig. 4 (Color online) Comparison of the density estimates of the distributions of the parameters using an ABC approximation with ε as the 0.1% quantile on the autocovariance distances (*in blue*) and the Beaumont et al. (2002) correction (*in red*). The *red* and *blue curves* are confounded for the parameter θ_2

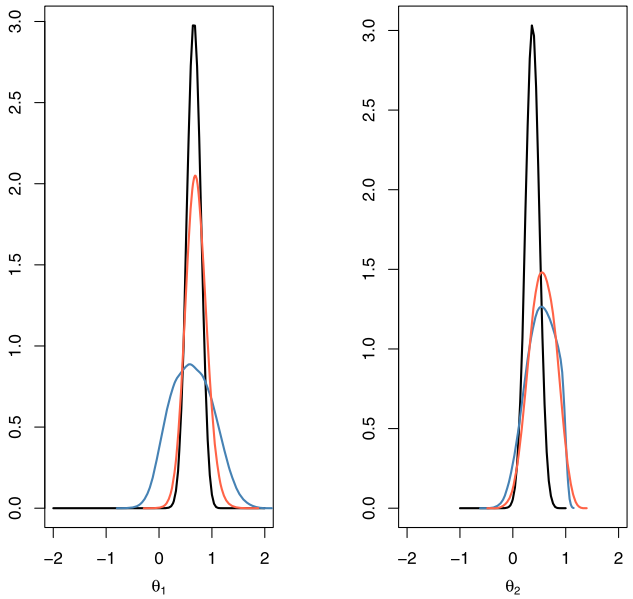

Fig. 5 (Color online) Comparison of the approximate distributions of the parameters using an ABC approximation with ε as the 20% quantile on the autocovariance distance (*in blue*) and the Beaumont et al. (2002) correction (*in red*)

Nonlinear regression Blum and François (2010) propose a generalisation of Beaumont et al.'s (2002) ABC post-processing where the local linear regression of the parameter $\boldsymbol{\theta}$ on the summary statistics $\eta(\mathbf{z})$ is replaced by a nonlinear regression with heteroskedasticity. In this new approach, the nonlinear mean and variance are estimated by a neural net with one hidden layer, using the R package nnet (R Development Core Team 2006). The result is interesting in that it seems to allow for the inclusion of more or even all the simulated pairs $(\boldsymbol{\theta}, \mathbf{z})$, compared with Beaumont et al. (2002). This is somehow to be expected since the nonlinear fit adapts differently to different parts of the space. Therefore, weighting simulated $(\boldsymbol{\theta}, \mathbf{z})$'s by a kernel $K_\delta(\mathbf{z} - \mathbf{y})$ is not very relevant and it is thus not surprising that the bandwith δ is not influential, in contrast with basic ABC and even (Beaumont et al. 2002) where δ has a different meaning. The non-parametric perspective adopted in the paper is nonetheless of the highest importance, as it proves the most fruitful approach to the interpretation of ABC methods. In connection with this paper, Blum (2010) provides a good review of the non-parametric handling of ABC techniques. The true difficulty with the non-parametric perspective lies with the curse of dimensionality. This issue might be addressed by mixing dimension reduction with recycling by shrinking as in Beaumont et al. (2002).

Inverse regression Leuenberger et al. (2010) also relate to the local regression ideas in Beaumont et al. (2002). As in the earlier work by Wilkinson (2008), the approximation to the distribution of the parameters given the observed summary statistics is central to the paper. In opposition to Beaumont et al. (2002), there is no clear shrinkage for summary statistics that are far away from the observed summary statistics: all accepted parameters are weighted similarly in the Gaussian linear approximation to the truncated prior. The other difference with Beaumont et al. (2002) is that the authors model $\mathbf{z}$ given $\boldsymbol{\theta}$ rather than $\boldsymbol{\theta}$ given $\mathbf{z}$, in an inverse regression perspective, followed by a sort of Laplace approximation reminding Rue et al. (2009).

6 ABC and model choice

6.1 Bayesian model choice

Model choice is one particular aspect of Bayesian analysis that involves computational complexity, if only because several models are considered simultaneously (see, e.g., Robert 2001, Marin and Robert 2010). In addition to the parameters of each model, the inference considers the model index $\mathcal{M}$, which is associated with its own prior distribution $\pi(\mathcal{M} = m)$ $(m = 1, \ldots, M)$ as well as a prior distribution on the parameters conditional on the value m of the model index, $\pi_m(\boldsymbol{\theta}_m)$, defined on the parameter space Θ_m. The choice between these models is then driven by the posterior distribution of $\mathcal{M}$, a challenging computational target where ABC brings a straightforward solution. Indeed, once $\mathcal{M}$ is incorporated within the parameters, the ABC approximation to the posterior follows from the same principles

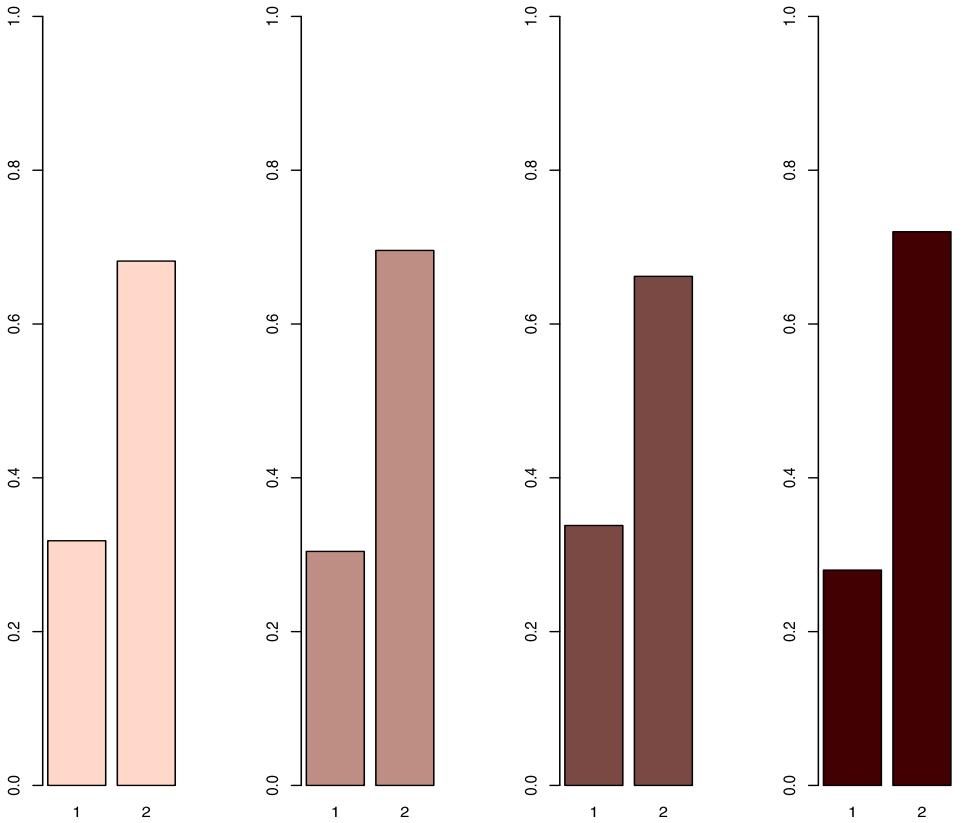

Fig. 6 Boxplots of the evolution [against ε] of ABC approximations to the Bayes factor. The representation is made in terms of frequencies of visits to [accepted proposals from] models MA(1) (*left*) and MA(2) (*right*) during an ABC simulation when ε corresponds to the 10, 1, 0.1, 0.01% quantiles on the simulated autocovariance distances. The data are the same as in Fig. 5. The true Bayes factor B_{21} is equal to 17.71, corresponding to posterior probabilities of 0.05 and 0.95 for the MA(1) and MA(2) models respectively

as regular ABC, as shown by the following pseudo-code, where $\boldsymbol{\eta}(\mathbf{z}) = (\eta_1(\mathbf{z}), \ldots, \eta_M(\mathbf{z}))$ is the concatenation of the summary statistics used for all models (with elimination of duplicates).

Algorithm 5 Likelihood-free model choice sampler (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 $\rho\{\boldsymbol{\eta}(\mathbf{z}), \boldsymbol{\eta}(\mathbf{y})\} < \varepsilon$
  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 acceptance frequency from model m, namely

$$\frac{1}{N}\sum_{i=1}^{N} \mathbb{I}_{m^{(i)}=m}.$$

This also corresponds to the proportion of simulated datasets that are closer to the data $\mathbf{y}$ than the tolerance ε. Cornuet et al. (2008) follow the rationale that led to the local linear regression in Beaumont et al. (2002) and rely on a weighted polychotomous logistic regression to estimate $\pi(\mathcal{M} = m|\mathbf{y})$. This modeling clearly brings some further stability to the above estimate of $\pi(\mathcal{M} = m|\mathbf{y})$ and is implemented in the DIYABC software described in Cornuet et al. (2008).

Example Returning once again to our benchmark MA(2) model, we compare the computation of the model posterior probabilities based on an ABC sample (acceptance frequency within each model) with the true value of the Bayes factor, which was obtained by numerical integration. The dataset used in the experiment is a time-series simulated and we wish to choose between two models: a MA(2) or a MA(1) model. Figure 6 shows our estimates for data simulated from AR MA(2) model. The weight of the MA(2) model increases slightly as ε decreases. However, even for the quantile at 0.01% the estimated posterior probability for the MA(2) model is equal to 0.72 which is far from the true value 0.95. Figure 7 shows a similar phenomenon for data simulated from an MA(1) model.

The discrepancy in the above example shows the limitations of the ABC approximation of Bayes factors exposed in Robert et al. (2011). While we could expect to obtain a

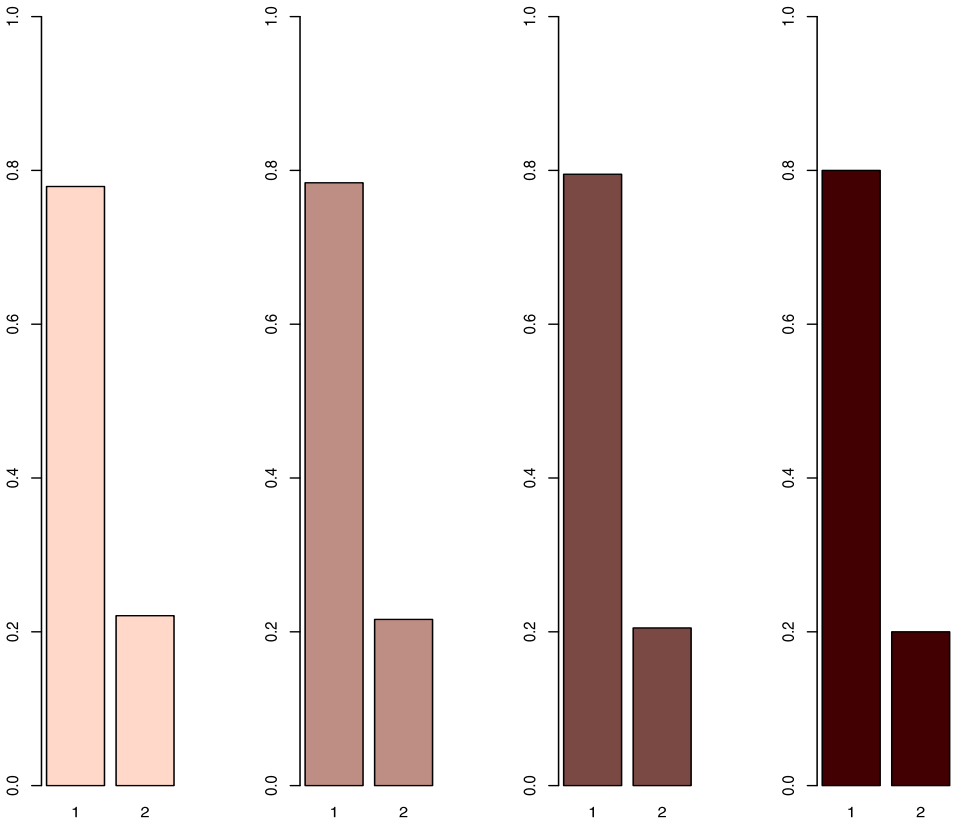

Fig. 7 Boxplots of evolution of Bayes factor approximations in terms of frequencies of visits to models MA(1) (*left*) and MA(2) (*right*) using an ABC approximation with $10, 1, .1, .01\%$ quantiles on the autocovariance distance as ε. The dataset is a sample of 50 points from a MA(1) model with $\theta_1 = 0.6$. The true Bayes factor B_{21} is equal to .004 corresponding to posterior probabilities of 0.996 and 0.004 for the MA(1) and MA(2) models respectively

better approximation with a massive computational effort, it may be that the use of different summary statistics for different models prevents us from converging to the true value. In other words, the concatenation of sufficient statistics for individual models does not always constitute a sufficient statistic for model choice, as discussed in the next paragraph.

6.2 The case of Gibbs random fields

Grelaud et al. (2009) show that, for Gibbs random fields and in particular for Potts models, where the goal is to compare several neighbourhood structures, the computation of the posterior probabilities of the models under competition can be operated by likelihood-free simulation techniques. We recall first that Gibbs random fields are probabilistic models associated with the likelihood function

$$\ell(\boldsymbol{\theta}|\mathbf{y}) = \frac{1}{Z_{\boldsymbol{\theta}}} \exp\{\boldsymbol{\theta}^{\mathrm{T}} \eta(\mathbf{y})\},$$

where $\mathbf{y}$ is a vector of dimension n taking values over a finite set $\mathcal{X}$ (possibly a lattice), $\eta(\cdot)$ is the potential function defining the random field, taking values in $\mathbb{R}^p$, $\boldsymbol{\theta} \in \mathbb{R}^p$ is the associated parameter, and $Z_{\boldsymbol{\theta}}$ is the corresponding normalising constant. A special but important case of Gibbs random fields is associated with a neighbourhood structure denoted by $i \sim i'$ (meaning that i and i' are neighbours), in that

$$\eta(\mathbf{y}) = \sum_{i' \sim i} \mathbb{I}_{\{y_i = y_{i'}\}},$$

where $\sum_{i' \sim i}$ indicates that the summation is over all the pairs of neighbours. In that case, $\boldsymbol{\theta}$ is a scalar.

The central property ensuring an ABC resolution for Gibbs random fields is that, due to their exponential family structure, there exists a sufficient statistic vector that runs across models and which allows for an exact ($\varepsilon = 0$) simulation from the posterior probabilities of the models. Indeed, model choice involves M Gibbs random fields in competition; each field is associated with a potential function η_m ($1 \leq m \leq M$), *i.e.* with the corresponding likelihood

$$\ell_m(\boldsymbol{\theta}_m|\mathbf{y}) = \exp\left\{\boldsymbol{\theta}_m^{\mathrm{T}} \eta_m(\mathbf{y})\right\} \Big/ Z_{\boldsymbol{\theta}_m, m},$$

where $\boldsymbol{\theta}_m \in \Theta_m$ and $Z_{\boldsymbol{\theta}_m, m}$ is the unknown normalising constant. From a Bayesian perspective, considering an extended parameter space $\Theta = \bigcup_{m=1}^{M} \{m\} \times \Theta_m$ that includes the model index $\mathcal{M}$, the computational target is thus the model posterior probability

$$\pi(\mathcal{M} = m|\mathbf{y}) \propto \int_{\Theta_m} \ell_m(\boldsymbol{\theta}_m|\mathbf{y}) \pi_m(\boldsymbol{\theta}_m)\, \mathrm{d}\boldsymbol{\theta}_m \pi(\mathcal{M} = m),$$

i.e. the marginal in $\mathcal{M}$ of the posterior distribution on $(\mathcal{M}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M)$ given $\mathbf{y}$. Each model has its own sufficient statistic $\eta_m(\cdot)$. Then, for each *individual* model, the vector of statistics $\boldsymbol{\eta}(\cdot) = (\eta_1(\cdot), \ldots, \eta_M(\cdot))$ is clearly sufficient. However Grelaud et al. (2009) exposed the fact that $\boldsymbol{\eta}$ is also sufficient for the *joint* parameter $(\mathcal{M}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M)$.

That the concatenation of the sufficient statistics of each model is also a sufficient statistic for the joint parameter across models is clearly a property that is specific to exponential families. As shown by Didelot et al. (2011), ABC-based model choice can process exponential families by creating inter-model sufficient statistics that incorporate the intra-model sufficient statistics as well as possibly the dominating measures for all models. The Gibbs random field above is a specific case of this sufficiency. However, outside exponential families, the possibility of creating a sufficient statistic of a dimension that is much lower than the dimension of the data is impossible, as explained in Robert et al. (2011).

6.3 General issues

Toni et al. (2009) and Toni and Stumpf (2010) review ABC-based model choice, inclusive of the above Gibbs random field example. The authors study in particular the consequences of implementing a sequential algorithm like ABC-PMC in this set-up. The ABC algorithm is modified to incorporate the model index, resorting to the previous assessment of $\pi(\mathcal{M} = m|\mathbf{y})$ to propose the model indices of the next population. The importance sampling features of this setting imply that the posterior probability can be estimated from the importance weights. However, the adaptivity at the core of ABC-PMC and ABC-SMC implies adapting an approximation kernel for each model. As most other perspectives on ABC, Toni and Stumpf (2010) do not question the role of the ABC distance in model choice settings. The Bayes factors are observed to be sensitive to the choice of the prior distributions, of the tolerance levels, and to the variances of the kernels K_t (see Sect. 4), a dependence that should not occur, since this is a simulation parameter that is unrelated with the statistical problem.

It is worth pointing out the remark made by Leuenberger et al. (2010) about model choice and the use of the approximation of the normalising constant resulting from the modelling to get to the marginal likelihood and the computation of the Bayes factor. This relates to earlier comments about the ABC acceptance rate approximating the marginal and a recent paper by Bartolucci et al. (2006) studying ways of computing marginal probabilities by Rao–Blackwellising reversible jump acceptance probabilities.

Grelaud et al. (2009) also make the most of this ABC feature for Ising models, since an exact ABC (corresponding to $\varepsilon = 0$) algorithm is then available for model selection.

A (minor) Bayesian issue mentioned by Ratmann et al. (2009) is the fact that both $\boldsymbol{\theta}$ and ε are taken to be the same across models. In a classical Bayesian perspective, modulo the reparameterisation, $\boldsymbol{\theta}$ cannot be entirely different from one model to the next, but using the same prior on ε over all models under comparison is more of an issue. The paper also considers the impact of testing for the adequacy of a model as testing for the hypothesis H_0: $\varepsilon = 0$, an interesting if controversial stance, since even when the model fits, ε necessarily varies around zero.

At this stage, the most perplexing feature of ABC model choice is the lack of convergence guarantees. As exposed in Robert et al. (2011), most settings where ABC model choice is implemented do not allow for inter-model sufficiency in the selection of the summary statistics, because some models are not within exponential families and because using the whole data is too demanding. As shown by the MA example above. this lack of sufficiency may be quite detrimental to the quality of the ABC approximation of the Bayes factors. There is therefore currently no theoretical support for the use of ABC approximations of Bayes factors and posterior model probabilities, and we thus advise for more empirical assessments in the spirit of Ratmann et al. (2009) that evaluate the model fit within each model without concluding by exact figures of the probabilities of the different models.

7 Discussion

Approximate Bayesian Computation allows for inference from a wide class of models which would otherwise be unavailable. As such, it has spawned interest in both theoretical issues and applications. Recent advances regarding the calibration of the method lead to an approximation that is good enough to be highly useful in many situations. The efficiency of the method can be greatly improved with sequential techniques and post-processing regression on the output.

Nonetheless, ABC is not a silver bullet. In the current state of the art, it can only be used for model choice in a limited range of models. Future advances must at the same time expand further the tools to make ABC useful in a wider class of models, extend pre- and post-processing methods to control the approximation, and establish more clearly in which cases ABC reaches its limitations.

ABC methods are currently under an intense scrutiny by both statisticians and practitioners, hence the object of an unparalleled development. While this rapid development provides answers to some interrogations from the statistical community about the validity of the approach and from the practitioners about a higher efficiency of the method, some issues remain unsolved, among which:

– the convergence results obtained so far are unpractical in that they require either the tolerance to go to zero or the

sample size to go to infinity. Obtaining exact error bounds for positive tolerances and finite sample sizes would bring a strong improvement in both the implementation of the method and in the assessment of its worth. As cleverly suggested by a referee, one could consider the L_1 error bound and analyze the decomposition:

$$\mathbb{E}\left\{\left|\frac{1}{N}\sum_{i=1}^{N}\phi\left(\boldsymbol{\theta}_i\right)-\pi(\phi)\right|\right\}$$
$$\leq \mathbb{E}\left\{\left|\frac{1}{N}\sum_{i=1}^{N}\phi\left(\boldsymbol{\theta}_i\right)-\pi_{\varepsilon}(\phi)\right|\right\}+\left|\pi_{\varepsilon}(\phi)-\pi(\phi)\right|$$

where $\phi : \mathbb{R}^d \to \mathbb{R}$ is some test-function, $\pi(\phi) = \int \phi(\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{y})\mathrm{d}\boldsymbol{\theta}$ and the expectation is taken with respect to the uncertainty generated by the Monte Carlo algorithm which has produced $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N$. For the first term, one can deal with the convergence rate as with any Monte Carlo method and the second term is the deterministic bias brought about by the ABC approximation.

- even though ABC is often presented as a converging method that approximates Bayesian inference, it can also be perceived as an inference technique per se and hence analysed in its own right. Connections with indirect inference have already been drawn, however the fine asymptotic analysis of ABC would be most useful to derive. Moreover, it could indirectly provide indications about the optimal calibration of the algorithm.
- in connection with the above, the connection of ABC-based inference with other approximative methods like variational Bayes inference is so far unexplored. Comparing and interbreeding those different methods should become a research focus as well.
- the construction and selection of the summary statistics is so far highly empirical. An automated approach based on the principles of data analysis and approximate sufficiency would be much more attractive and convincing, especially in non-standard and complex settings.
- the debate about ABC-based model choice is so far inconclusive in that we cannot guarantee the validity of the approximation, while considering that a "large enough" collection of summary statistics provides an acceptable level of approximation. Evaluating the discrepancy by exploratory methods like the bootstrap would shed a much more satisfactory light on this issue.
- the method necessarily faces limitations imposed by large datasets or complex models, in that simulating pseudo-data may itself become an impossible task. Dimension-reducing technique that would simulate directly the summary statistics will quickly become necessary.

Acknowledgements The authors are grateful to J.-M. Cornuet for bringing the problem to their attention and for a highly enjoyable and fruitful collaboration over the past years. Part of this work was conducted while the third author was visiting the Department of Statistics at the Wharton Business School of the University of Pennsylvania, to whom he is most grateful for its support. The authors are also grateful to J.-L. Foulley for several interesting discussions, as well as to a referee for its very relevant suggestions.

References

Bartolucci, F., Scaccia, L., Mira, A.: Efficient Bayes factor estimation from the reversible jump output. Biometrika **93**(1), 41–52 (2006)

Beaumont, M., Zhang, W., Balding, D.: Approximate Bayesian computation in population genetics. Genetics **162**(4), 2025–2035 (2002)

Beaumont, M., Cornuet, J.-M., Marin, J.-M., Robert, C.: Adaptive approximate Bayesian computation. Biometrika **96**(4), 983–990 (2009)

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., Excoffier, L.: In defense of model-based inference in phylogeography. Mol. Ecol. **19**(3), 436–446 (2010)

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

Blum, M.: Approximate Bayesian computation: a non-parametric perspective. J. Am. Stat. Assoc. **105**(491), 1178–1187 (2010)

Blum, M., François, O.: Non-linear regression models for approximate Bayesian computation. Stat. Comput. **20**(1), 63–73 (2010)

Calvet, L., Czellar, V.: State-observation sampling and the econometrics of learning models. Technical Report (2011). arXiv:1105.4519

Campillo, F., Rossi, V.: Convolution particle filter for parameter estimation in general state-space models. IEEE Trans. Aerosp. Electron. Syst. **45**(3), 1063–1072 (2009)

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

Csillèry, K., Blum, M., Gaggiotti, O., François, O.: Approximate Bayesian computation (ABC) in practice. Trends Ecol. Evol. **25**(7), 410–418 (2010a)

Csillèry, K., Blum, M., Gaggiotti, O., François, O.: Invalid arguments against ABC: a reply to A.R. Templeton. Trends Ecol. Evol. **25**(7), 490–491 (2010b)

Cucala, L., Marin, J.-M., Robert, C., Titterington, D.: Bayesian inference in k-nearest-neighbour classification models. J. Am. Stat. Assoc. **104**(485), 263–273 (2009)

Dean, T.A., Singh, S.S., Jasra, A., Peters, G.W.: Parameter estimation for hidden Markov models with intractable likelihoods. Technical Report (2011). arXiv:1103.5399

Del Moral, P., Doucet, A., Jasra, A.: Sequential Monte Carlo samplers. J. R. Stat. Soc. B **68**(3), 411–436 (2006)

Del Moral, P., Doucet, A., Jasra, A.: An adaptive sequential Monte Carlo method for approximate Bayesian computation. Stat. Comput. (2011, to appear)

Didelot, X., Everitt, R., Johansen, A., Lawson, D.: Likelihood-free estimation of model evidence. Bayesian Anal. **6**(1), 48–76 (2011)

Douc, R., Guillin, A., Marin, J.-M., Robert, C.: Convergence of adaptive mixtures of importance sampling schemes. Ann. Stat. **35**(1), 420–448 (2007)

Drovandi, C., Pettitt, A.: Estimation of parameters for macroparasite population evolution using approximate Bayesian computation. Biometrics **67**(1), 225–233 (2010)

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

Friel, N., Pettitt, A.: Marginal likelihood estimation via power posteriors. J. R. Stat. Soc. B **70**(3), 589–607 (2008)

Gauchi, J.-P., Vila, J.-P.: Nonparametric filtering approaches for identification and inference in nonlinear dynamic systems. Technical report. Personal communication (2011)

Gelfand, A., Smith, A.: Sampling based approaches to calculating marginal densities. J. Am. Stat. Assoc. **85**(410), 398–409 (1990)

Gouriéroux, C., Monfort, A., Renault, E.: Indirect inference. J. Appl. Econom. **8**, 85–118 (1993)

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

Jaakkola, T., Jordan, M.: Bayesian parameter estimation via variational methods. Stat. Comput. **10**(1), 25–37 (2000)

Jasra, A., Singh, S.S., Martin, J.S., McCoy, E.: Filtering via approximate Bayesian computation. Stat. Comput. (2011, to appear)

Joyce, P., Marjoram, P.: Approximately sufficient statistics and Bayesian computation. Stat. Appl. Genet. Mol. Biol. **7**(1), 26 (2008)

Leuenberger, C., Wegmann, D., Excoffier, L.: Bayesian computation and model selection in population genetics. Genetics **184**(1), 243–252 (2010)

Marin, J.-M., Robert, C.: Bayesian Core. Springer, New York (2007)

Marin, J.-M., Robert, C.: Importance sampling methods for Bayesian discrimination between embedded models. In: Chen, M.-H., Dey, D., Müller, P., Sun, D., Ye, K. (eds.) Frontiers of Statistical Decision Making and Bayesian Analysis, pp. 513–527. Springer, New York (2010)

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

McKinley, T., Cook, A., Deardon, R.: Inference in epidemic models without likelihoods. Int. J. Biostat. **5**(1), 24 (2009)

Møller, J., Pettitt, A., Reeves, R., Berthelsen, K.: An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants. Biometrika **93**(2), 451–458 (2006)

Pritchard, J., Seielstad, M., Perez-Lezaun, A., Feldman, M.: Population growth of human Y chromosomes: a study of Y chromosome microsatellites. Mol. Biol. Evol. **16**(12), 1791–1798 (1999)

R Development Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna (2006)

Ratmann, O.: ABC under model uncertainty. PhD thesis, Imperial College, London (2009)

Ratmann, O., Andrieu, C., Wiujf, C., Richardson, S.: Model criticism based on likelihood-free inference, with an application to protein network evolution. Proc. Natl. Acad. Sci. **106**(26), 1–6 (2009)

Ratmann, O., Andrieu, C., Wiuf, C., Richardson, S.: Reply to Robert et al.: Model criticism informs model choice and model comparison. Proc. Natl. Acad. Sci. **107**(3), E6 (2010)

Robert, C.: The Bayesian Choice, 2nd edn. Springer, New York (2001)

Robert, C., Casella, G.: Monte Carlo Statistical Methods, 2nd edn. Springer, New York (2004)

Robert, C.P., Mengersen, K., Chen, C.: Model choice versus model criticism. Proc. Natl. Acad. Sci. **107**(3), E5 (2010)

Robert, C.P., Cornuet, J.-M., Marin, J.-M., Pillai, N.: Lack of confidence in ABC model choice. Proc. Natl. Acad. Sci. **108**(37), 15112–15117 (2011)

Rubin, D.: Bayesianly justifiable and relevant frequency calculations for the applied statistician. Ann. Stat. **12**(4), 1151–1172 (1984)

Rue, H., Held, L.: Gaussian Markov Random Fields: Theory and Applications. Monographs on Statistics and Applied Probability, vol. 104. Chapman & Hall, London (2005)

Rue, H., Martino, S., Chopin, N.: Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations. J. R. Stat. Soc. B **71**(2), 319–392 (2009)

Sisson, S.A., Fan, Y., Tanaka, M.: Sequential Monte Carlo without likelihoods. Proc. Natl. Acad. Sci. **104**(6), 1760–1765 (2007)

Sisson, S.A., Fan, Y., Tanaka, M.: Sequential Monte Carlo without likelihoods: Errata. Proc. Natl. Acad. Sci. **106**(39), 16889 (2009)

Tavaré, S., Balding, D., Griffith, R., Donnelly, P.: Inferring coalescence times from DNA sequence data. Genetics **145**(2), 505–518 (1997)

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

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

Tierney, L., Kadane, J.: Accurate approximations for posterior moments and marginal densities. J. Am. Stat. Assoc. **81**(393), 82–86 (1986)

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

Toni, T., Welch, D., Strelkowa, N., Ipsen, A., Stumpf, M.: Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. J. R. Soc. Interface **6**(31), 187–202 (2009)

Wasserman, L.: All of Nonparametric Statistics. Springer, New York (2007)

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