Approximate Bayesian Computation, a survey on recent results

Christian P. Robert

Abstract Approximate Bayesian Computation (ABC) methods have become a "mainstream" statistical technique in the past decade, following the realisation that they were a form of non-parametric inference, connected as well with the econometric technique of indirect inference. In this survey of ABC methods, we focus on the recent literature, following our earlier survey in Marin et al. (2011). Given the recent paradigm shift in the perception and practice of ABC model choice, we particularly insist on this aspect of ABC techniques, including in addition convergence results. Most sections are edited versions of posts published on xianblog.wordpress.com between 2011 and 2015.

1 Mudmap: ABC at a glance

While statistical (probabilistic) models are always to be held at a critical distance, being at best approximations of real phenomena (Box, 1959?, Gelman et al., 2013), and while more complex statistical models are not necessarily better representations of those phenomena, it is becoming increasingly the case that the complexity of models is a barrier to the most common tools in statistical analysis. Complexity might stem from many possible causes, from an elaborate description of the randomness behind the phenomenon to be analysed, to the handling of massive amounts of observations, to imperatives for real-time analysis, to considerable percentages of missing data in an otherwise regular model. All those reasons contribute to make computing the likelihood function a formidable task.

Example 1. Kingman's coalescent Inference in population genetic relies on models such as Kingman's coalescent trees. This is a representative example of cases when

Christian P. Robert
CEREMADE, Université Paris-Dauphine, & Department of Statistics, University of Warwick e-mail: xian@ceremade.dauphine.fr

the likelihood function associated with the data 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.

Kingman's coalescent trees are probabilistic models representing the evolution from an unobserved common ancestor to a collection of N (observed) populations, described by the frequencies of genetic traits such as alleles of loci from the genome in single nucleotide polymorphism (SNP) datasets. For two populations j_1 and j_2 and a given locus, at current time, with allele sizes x_{j_1} and x_{j_2}, a binary tree has for root the most recent time in the past for which they have a common ancestor, defined as the coalescence time τ_{j_1,j_2}. The two copies are thus separated by a branch of gene genealogy of total length $2\tau_{j_1,j_2}$. As explained in Slatkin (1995), according to Kingman's coalescent process, during that duration $2\tau_{j_1,j_2}$, the number of mutations is a random variable distributed from a Poisson distribution with parameter $2\mu\tau_{j_1,j_2}$. Aggregating all populations by pairing the most recently diverged pairs and repeating the pairing on the most recent common ancestors of those pairs produces a binary tree which root is the most recent common ancestor of the collection of populations and with as many branches as there are populations. For a given tree topology, such as the one provided in Figure 1, inferring the tree parameters (coalescent times and mutation rate) is a challenge, because the likelihood of the observations is not available, while the likelihood of the completed model involves missing variables, namely the $2(N-1)$ mutations along all edges of the graph. While this is not a large dimension issue in the case of Figure 1, building an efficient completion mecchanism such as data augmentation or importance sampling proves to be quite complicated (Stephens and Donnelly, 2000). ◀

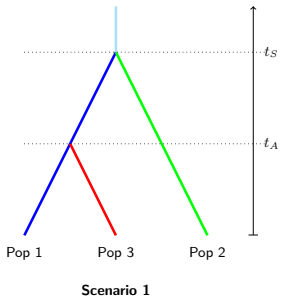

Fig. 1 Possible model of historical relationships between three populations of a given species *(Source: Pudlo et al. (2014), with permission.*

Example 2. Lotka–Volterra prey-predator model The Lotka-Volterra model (Wilkinson, 2006) describes interactions between a first species, referred to as the prey

species, and a second species, referred to as the predator species, in terms of population sizes, $x_1(t)$ and $x_2)t)$. Given the parameter $\theta = (\theta_1, \theta_2)$, the model on the respective population sizes is driven by a system of differential equations (ODEs):

$$\frac{\mathrm{d}x_1}{\mathrm{d}t} = \theta_1 x_1 - x_1 x_2,$$
$$\frac{\mathrm{d}x_2}{\mathrm{d}t} = \theta_2 x_1 x_2 - x_2.$$

with initial values $(x_1(0), x_2(0))$. Typically, one does not observe the entire curve $\mathbf{x}(t)$, but only at a finite number of times $t_1, ..., t_R$. Furthermore, the $\mathbf{x}(t_i)$'s are measured with error,

$$\mathbf{y}(t_i) = \mathbf{x}(t_i) + \nu(t_i),$$

where $\nu(t_i) \overset{\text{i.i.d.}}{\sim} \mathcal{N}_2(0, \Sigma_\nu)$.

While the likelihood associated with this collection of observations is intractable, the process $\mathbf{y}$ can be simulated, which means particle MCMC solutions exist, as in Golightly and Wilkinson (2011), even though an ABC implementation is more straightforward (Toni et al., 2009). ◀

[REALISTIC EXAMPLES HERE]

In such situations, statisticians will try to provide answers by

- modifying the original model towards a more manageable version (see, e.g., the variational Bayes literature as in Jaakkola and Jordan, 2000);
- using only relevant aspects of the model (see, e.g., the many versions of the method of moments in econometrics, Gouriéroux *et al.*, 1993, Heggland and Frigessi, 2004, Gallant and Tauchen, 1996);
- constructing new tools (with numerous examples ranging from the EM algorithm, Dempster et al., 1977, to machine learning, Breiman et al., 1984, Hastie et al., 2001).

The Approximate Bayesian computation (ABC) method covered by this survey is a mix of these different solutions in that it does not return an answer to the original question (namely, to deliver the posterior distribution associated with the original dataset and the postulated model(s)), only uses some summaries of the data (even though it most often requires a constructive definition of the model), and construct a simulation tool that is to some extent novel (albeit a rudimentary version of accept-reject algorithm, Robert and Casella, 2004).

Without yet attempting to justify the ABC method from a theoretical perspective, let me provide here a quick and handwaving description of its *modus vivendi* (operational mode). Its goal is to substitute a Monte Carlo approach to an intractable posterior distribution, while preserving the Bayesian aspect of the statistical analysis. The ABC method is based on the intuition that if a parameter value θ produces one or several simulated datasets $\boldsymbol{x}(\theta)$ that are resembling[1] the observed data set $\boldsymbol{x}(\theta^0)$

[1] The notation $\boldsymbol{x}(\theta)$ is stressing the point that the simulated data is a random transform of the parameter θ, $\boldsymbol{x}(\theta) \sim f(\cdot|\theta)$. From a programming perspective, $\boldsymbol{x}(\theta)$ actually is a function of θ

then θ must be close to the value of the parameter that stands behind the data, θ^0. And, conversely, if the simulated dataset $\boldsymbol{x}(\theta)$ differs a lot from the observed data $\boldsymbol{x}(\theta^0)$, then θ is presumably different from θ. ABC goes beyond this intuition by quantifying the resemblance and estimating the bound for the datasets to be "close enough". Starting from a collection of parameter values usually generated from the prior, ABC first associates with each value a simulated dataset of the same nature as the observed one and then rejects all values of θ for which the simulated dataset is too far from the observed one. The surviving parameter values are subsequently used as if they were generated from the posterior distribution, even though they are not, due to several reasons discussed below. The major reason for failing to accomodate for this difference is that the approximation effect is difficult (and costly) to evaluate.

While ABC is rarely fast, due to the reason that many simulated samples need to be produced and that the underlying statistical model is complex enough to lead to costly generations, it often is the unique answer to settings where regular Monte Carlo methods (including MCMC, Robert and Casella, 2004, and particle filters, Doucet et al., 1999). The method is easily parallelisable as well as applicable to sequential settings, due to its rudimentary nature. Furthermore, once the (massive) collection of pairs $(\theta, \boldsymbol{x}(\theta))$ is produced, it can be exploited multiple times, which makes it paradoxically available for some real time applications.

2 Introduction

2.1 ABC Basics

2.1.1 Intractable likelihoods

Although it has now spread to a wide range of application domains, Approximate Bayesian Computation (ABC) was first introduced in population genetics (Tavaré et al., 1997, Pritchard et al., 1999) to handle models with intractable likelihoods Beaumont (2010). By *intractable*, we mean models where the likelihood function $\ell(\theta|y)$

- is completely defined by the probabilistic model, $y \sim f(y|\theta)$;
- is available nor in closed form, neither by numerical derivation;
- cannot easily be either completed or demarginalised (Tanner and Wong, 1987, Robert and Casella, 2004);
- cannot be estimated by an unbiased estimator (Andrieu and Roberts, 2009).

This intractability prohibits the direct implementation of a generic MCMC algorithm like Gibbs or Metropolis–Hastings schemes. Examples of latent variable models of high dimension abound, primarily in population genetics, but more generally in

and of a random variable or sequence, o: $\boldsymbol{x}(\theta, \mathrm{o})$. By extension, assuming the postulated model is correct, there exists a "true" value of the parameter, θ^o, such that the observed data writes as $\boldsymbol{x}(\theta^0)$.

models including combinatorial structures (e.g., trees, graphs), intractable normalising constants as in $f(y|\theta) = g(y|\theta)/Z(\theta)$ (e.g. Markov random fields, exponential graphs) and other missing (or latent) variables, i.e. when

$$f(y|\theta) = \int_{\mathscr{G}} f(y|G,\theta) f(G|\theta) \mathrm{d}G$$

cannot be computed in a manageable way (while $f(y|G,\theta)$ and $f(G|\theta)$ are available).

Example 3. As an intuitive (or pedestrian!) entry to untractable likelihoods, consider the "case of the single socks", as proposed by Rasmus Bååth on his blog. At the end of one's laudry, the 11 first socks extracted from the washing machine are single (meaning that no complete pair is recovered). *What is the posterior distribution on the number of socks? and on the number of pairs?*

This sounds like an impossible task, but it can be solved by setting a prior on the number of socks, n_s, chosen to be a negative binomial $\mathscr{N}eg(N,\rho)$ random variable based on the size of the family, with mean 30 and standard deviation 15, and on the proportion of pairs in the laundry, chosen to derived from a Beta $p \sim \mathscr{B}e(15,2)$ weight to reflect on the low proportion of single socks, namely $n_p = \lceil p n_s/2 \rceil$. Given (n_s, n_p), it is then straightforward to generate a laundry sequence of 11 socks by a simple sampling without replacement from the population of socks. *A contrario*, it is much more involved to express the distribution of the number of single socks in those 11 random draws (albeit possible, see below). ◄

The idea of the approximation behind ABC is both surprisingly simple and fundamentally related to the very nature of statistics, as solving an inverse problem (Stigler, 1986). Indeed, ABC relies on the feasibility of producing simulated (parameters and) data from the inferred model or models, as it evaluates the unavailable likelihood by the proximity of this simulated data to the observed data. In other words, it relies on the natural assumption that the *forward* step induced by the probabilistic model—from model to data—is reasonably easy to implement in contrast with the *backward* step—from data to model.

2.1.2 An exact Bayesian computation

> "Bayesian statistics and Monte Carlo methods are ideally suited to the task of passing many models over one dataset." D. Rubin, 1984

Not so coincidentally, Rubin (1984), quoted above, used this representation as a mostly non-algorithmic motivation for conducting Bayesian analysis (as opposed to other forms of inference). This paper indeed details the accept-reject algorithm (Robert and Casella, 2004) at the core of the ABC algorithm. Namely, the following algorithm

Algorithm 1 Accept-reject for Bayesian analysis

repeat
  Generate $\theta \sim \pi(\theta)$;
  Generate $x \sim f(x|\theta)$;
  Accept θ if $x = x_0$
until acceptance
return the accepted value of θ

returns as accepted value an output *exactly* generated from the posterior distribution, $\pi(\theta|x_0)$.

Example 4. If we return to the socks example 3, running Algorithm 1 means running the following steps

repeat
  Generate $n_s \sim \mathcal{N}\rceil\}(N,\rho)$ and $p \sim \mathcal{B}e(15,2)$
  Set $n_p = \lceil p n_s/2 \rceil$
  Sample 11 terms from $\{o_{11}, o_{12}, \ldots, o_{n_p 1}, o_{n_p 2}, s_1, \ldots, s_{n_s - 2n_p}\}$
  Accept (n_s, n_p) if there is no pair (o_{i1}, o_{i2}) in the sample
until acceptance
return the accepted value of (n_s, n_p)

and this loop will produce an output from the posterior distribution of (n_s, n_p), that is, conditional on the event that no pair occurred out of 11 draws without replacement. Running the implementation of the above algorithm as done in Bååth's R code leads to Fig. 2. The number of proposals in the above loop was 10^5, resulting in above 10^4 acceptances.

As mentionned above, the probability that the 11 socks all come from different pairs can be computed by introducing a latent variable, namely the number k of orphan socks in the 11 socks, out of the $n_s - 2n_p$ existing orphans. Integrating out this latent variable k (and using Feller, 1970, Chap. 2, Exercise 26 for the number of different pairs in the remaining $11-k$ socks), leads to

$$\sum_{k=0}^{11} \frac{\binom{n_s-2n_p}{k}\binom{2n_p}{11-k}}{\binom{n_s}{11}} \frac{2^{11-k}\binom{n_p}{11-k}}{\binom{2n_p}{11-k}}$$

as the probability of returning no complete pair out of the 11 draws. If we discretise the Beta $\mathcal{B}e(15,2)$ distribution to obtain the probability mass function of n_p, we are therefore endowed with a closed-form posterior distribution

$$\pi(n_s, n_p|\mathcal{D}) \propto \binom{n_s+N-1}{n_s} \rho^{n_s} \int_{2n_p/n_s}^{2(n_p+1)/n_s} p^{16}(1-p)^3 \,\mathrm{d}\, p \sum_{k=0}^{11} \frac{\binom{n_s-2n_p}{k}\binom{2n_p}{11-k}}{\binom{n_s}{11}} \frac{2^{11-k}\binom{n_p}{11-k}}{\binom{2n_p}{11-k}}$$

that we can compare with the ABC output (even though this ABC output is garanteed to correspond to simulations from the true posterior distribution[2]). As demonstrated in Fig. 3, there is indeed a perfect fit bewtween the simulations and the target.

[2] This feature means that the 'A' in 'ABC' is superfluous in this special case!

Obviously, a slight extension of this setup with, say, a second type of socks or a different rule for pulling the socks out of the washing machine (or out of different laundry bags) could sufficiently complicate the likelihood associated with the observations to reach intractability. See for instance Arratia and DeSalvo (2012) for an interesting alternative of Feller's (1970) shoes cupboard problem. ◄

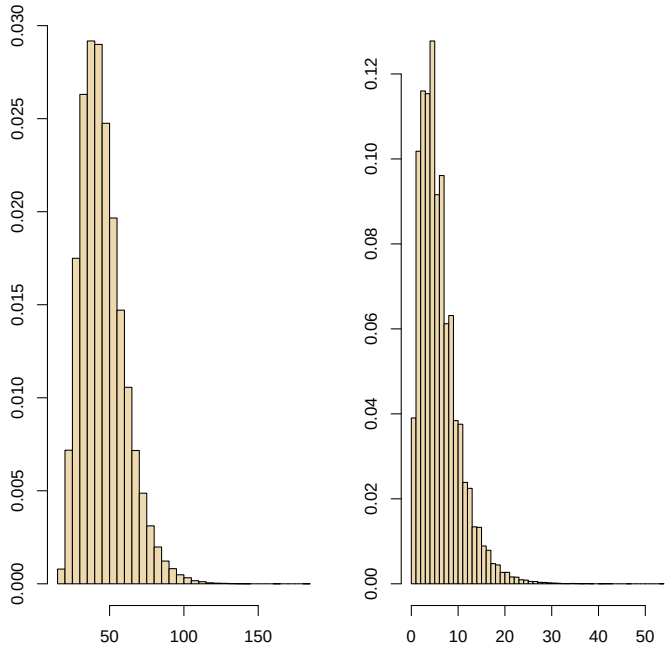

Fig. 2 ABC based simulation of posterior distributions on *(left)* the number of socks, n_s, and *(right)* the number of odd socks, n_o, relying on 10^6 proposals (n_s, n_p) simulated from the prior. *(R code kindly provided by Rasmus Bååth)*

2.1.3 Enters the approximation

Now, ABC proceeds one step further in the approximation, replacing the acceptance step with the tolerance condition

$$d(x, x_0) < \varepsilon$$

in order to handle continuous (and large finite) sampling spaces, $\mathfrak{X}$, but this early occurrence in Rubin (1984) is definitely worth signalling. It is also relevant that

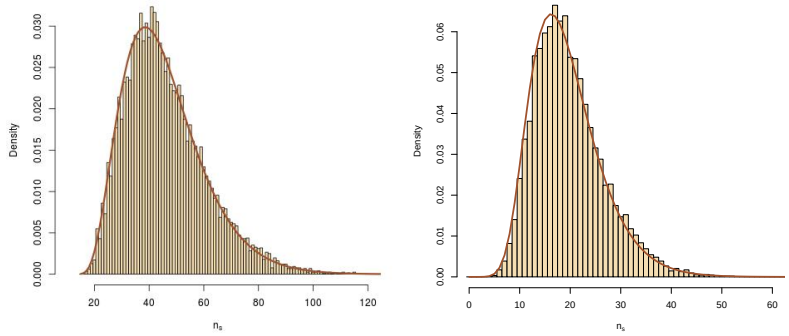

Fig. 3 Same graphs as in Fig. 2 with true posterior marginal distributions.

Rubin does not promote this simulation method in situations where the likelihood is not available but rather as an intuitive way to understand posterior distributions from a frequentist perspective, because θ's from the posterior are those that could have generated the observed data. (The issue of the zero probability of the exact equality between simulated and observed data is not dealt with in the paper, maybe because the notion of a "match" between simulated and observed data is not clearly defined.) Another (just as) early occurrence of an ABC-like algorithm was proposed by Diggle and Gratton (1984).

Algorithm 2 ABC (basic version)

for $t = 1$ to N **do**
  repeat
    Generate θ^* from the prior $\pi(\cdot)$
    Generate $\boldsymbol{x}^*$ from the model $f(\cdot|\theta^*)$
    Compute the distance $\rho(\mathbf{x}^0, \mathbf{x}^*)$
    Accept θ^* if $\rho(\mathbf{x}^0, \mathbf{x}^*) < \varepsilon$
  until acceptance
end for
return N accepted values of θ^*

The ABC method is formally implemented as in Algorithm 2, which requires calibrating the objects $\rho(\cdot,\cdot)$, called the *distance* or *divergence* measure, N, number of accepted simulations, and ε, called the *tolerance*. Algorithm 2 is exact (in the sense of Algorithm 1) when $\varepsilon = 0$. However, this is at best a formal remark since this ideal setting cannot be found in most problems where ABC is needed (see Grelaud et al. (2009) for a specific counterexample) and a positive tolerance is required in

practical settings[3]. While several approaches are found in the literature, we follow here the practice of selecting ε as a quantile of the simulated distances $\rho(\mathbf{x}^0, \mathbf{x}^*)$, which turns out to express ABC as a k-nn method, as pointed out by Biau et al. (2014) and discussed in Section 3.1.

This algorithm is easy to call when checking the performances of the ABC methods on toy examples where the exact posterior distribution is known, in order to test the impact of the various calibration parameters. See for instance Marin et al. (2011) with the case of the MA(2) model. We illustrate the behaviour of the algorithm in a slightly more challenging setting.

Example 5. A surprisingly complex probability density (and hence likelihood function) is the one associated with the empirical mean $\bar{x}_n$ of a Student's t sample.[4] Indeed, if

$$(x_1, \ldots, x_n) \overset{\text{i.i.d.}}{\sim} \mathfrak{T}(\nu, \mu, \tau)\,,$$

the resulting $\bar{x}_n$ has no standard distribution, even though it is also a location scale distribution with parameters μ and τ. To see this, consider that $x_i = \mu + \tau\xi_i$, with $\xi_i \sim \mathfrak{T}_\nu$. Then

$$\bar{x}_n = \mu + \tau\bar{\xi}_n\,, \tag{1}$$

with $\bar{\xi}_n$ distributed from a density that cannot be expressed otherwise than as an $(n-1)$-convolution of t's.

If we observe $p \geq 1$ realisations of $\bar{x}_n$, denoted $\bar{x}^1, \ldots, \bar{x}^p$, Algorithm 2 may be the solution to handling the corresponding implicit likelihood. When the prior on (μ, τ) is the normal-gamma prior

$$\tau^{-1} \sim \mathscr{G}a(1,1)\,,\ \mu|\tau \sim \mathscr{N}(0, 2\tau)\,,$$

Algorithm 2 consists in

1. generating a large sample of (μ, τ) from this prior (the sample is often called a *reference table*, then
2. generating a pseudo-sample $(\bar{x}^1, \ldots, \bar{x}^p)$ for each pair (μ, τ) in the reference table, and
3. deriving the distances ρ between pseudo- and true samples.

The reference table is then post-processed by keeping the parameter values that lead to the $(100\varepsilon)\%$ smallest distances. The choice of the distance is arbitrary, it could for instance be the average squared error

$$\rho(\mathbf{x}_0, \mathbf{x}_1) = \sum_{i=1}^{p} (\bar{x}_0^i - \bar{x}_1^i)^2\,.$$

[3] There even are arguments, see e.g. Fearnhead and Prangle (2012), to justify positive values of ε as preferable.

[4] We are aware that there exist two differing definitions for the $\mathfrak{T}(\nu, \mu, \tau)$ distribution. One is considering μ and τ as location and scale parameters: this is the interpretation chosen in this example. Another one starts from a $\mathscr{N}(\mu, \tau)$ variate and divides it by a normalised χ_ν variate, which leads to a non-standard density for even a single variable.

Figure 4 compares the parameter values selected by Algorithm 2 with the entire reference table, based on 10^5 simulations. The concentration of the parameter values near the true value $(0,1)$ is noticeable, albeit with a skewness towards the smalled values of τ that may reflect the strong skewness in the prior distiribution on τ. ◄

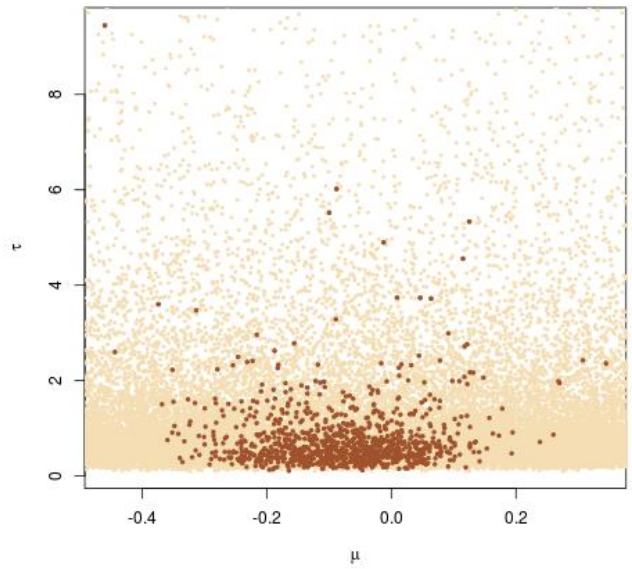

Fig. 4 Sample of $1,000$ simulations from Algorithm 2 when the data is made of $10\,t$ averages with the same sample size $n = 21$ and when the 10^5 ABC simulations in the background constitute the reference table, taken from the normal-gamma prior. *(Note: the true value of* (μ,τ) *is* $(0,1)$*.)*

A further difficulty arises when the prior on θ is improper and hence cannot be simulated. It is then impossible to use directly Algorithm 2. Instead, we can proceed by either

1. using a proper prior with a very large variance, à la BUGS (Lunn et al., 2010), but this is a very inefficient and wasteful proposal, both *per se* and in this particular setting, since most values generated from this prior will be fully incompatible with the data; or
2. replacing the prior simulation by a simulation from a pseudo-posterior, based on the data and mimicking to some extent the true posterior distribution, and weighting the outcome by the ratio of the prior over the pseudo-posterior. In the case of Example 5, we could instead use the posterior distribution associated with a normal sample, that is, by pretending the sample of the $\bar{x}_n$'s is made of normal observations with mean μ and variances $(\nu/\nu-2)(\tau^2/n)$; or

3. using part of the data to build a simpler but achievable posterior distribution. This solution is not available for Example 5, since even a single $\bar{x}^i$ is associated with a complex density; or
4. introducing latent variables to recover a closed form conditional posterior. In the setting of Example 5, it would prove very dear, since this requires producing n pairs $(y_j, z_j) \sim \mathcal{N}(0,1) \times \chi^2_\nu$ to decompose $\bar{x}_{n_i}$ as

$$\bar{x}_n = \mu + \tau^{1/2} \frac{1}{n} \sum_{j=1}^{n} y_j / \sqrt{z_j/\nu}\,.$$

Example 6. Lettuce thus consider the more (when compared with Example 5) challenging case where (a) we observed independently $\bar{x}_{n_1}, \ldots, \bar{x}_{n_p}$ that all are averages of Student's t samples with different sample sizes $n_1, \ldots, n_p$, and (b) the prior on (μ, τ) is the reference prior $\pi(\mu, \tau) = 1/\tau$.

We select the second solution proposed above, namely to rely on a normal approximation for the distribution of the observations, in order to build the following proposal:

$$(\mu, \tau^{-2}) | \bar{x}_{n_1}, \ldots, \bar{x}_{n_p} \sim \mathcal{N}\left(\bar{\bar{x}}, \nu\tau^2/(\nu-2)\sum n_i\right) \times \mathcal{G}\left(1 + p/2, (\nu-2)s^2/2\nu\right)$$

that serves as a proxy generator in the first step of the above algorithm.

If we apply Algorithm 2 to this problem, due to the representation (1), we can follow the next steps:

for $t = 1$ to N **do**
  Generate θ^* from the pseudo-posterior $\pi^*(\cdot|\boldsymbol{x})$
  Create a sample $(\bar{\xi}^t_{n_1}, \ldots, \bar{\xi}^t_{n_p})$
  Derive the transform $\boldsymbol{x}^t = (\bar{x}^t_{n_1}, \ldots, \bar{x}^t_{n_p})$
  Compute the distance $\rho(\mathbf{x}^0, \mathbf{x}^t)$
  Accept θ^* if $\rho(\mathbf{x}^0, \mathbf{x}^t) < \varepsilon$
end for
return N accepted values of θ^* along with importance weights $\omega^* \propto \pi(\theta^*)/\pi^*(\theta^*|\boldsymbol{x})$
or resample those θ^* with replacement according to the corresponding ω^*'s

where the distance is again arbitrarily chosen as the sum of the weighted squared differences.

Following this algorithm for a dataset of 10 averages simulated from central t distributions (i.e., with $\mu = 0$, $\tau = 1$), we obtain an ABC sample displayed on Fig. 5, which shows a reasonable variability of the sample around the true value $(0,1)$. The 10^3 points indicated on this picture are the output of a weighted resampling. If we compare the θ^*'s simulated from the pseudo-posterior $\pi^*(\cdot|\boldsymbol{x})$ with those finally sampled, the difference is quite limited, as exhibited in Fig. 6. The selected points do remain in a close neighbourhood of the mode. This behaviour remains constant through the choice of ε, so we can attribute it to (at least) two possible reasons. The first explanation is that the likelihood associated with $\bar{x}_{n_i}$ should be quite close to a normal likelihood, hence that the pseudo-posterior provides a fairly accurate representation of the true posterior distribution. The second explanation is a *contrario* that the ABC output reflects a lack of discrimination in the condition $\rho(\mathbf{x}^0, \mathbf{x}^t) < \varepsilon$, even for small values of ε and hence corresponds to simulations from a pseudo-

posterior that differs from the posterior. ◀

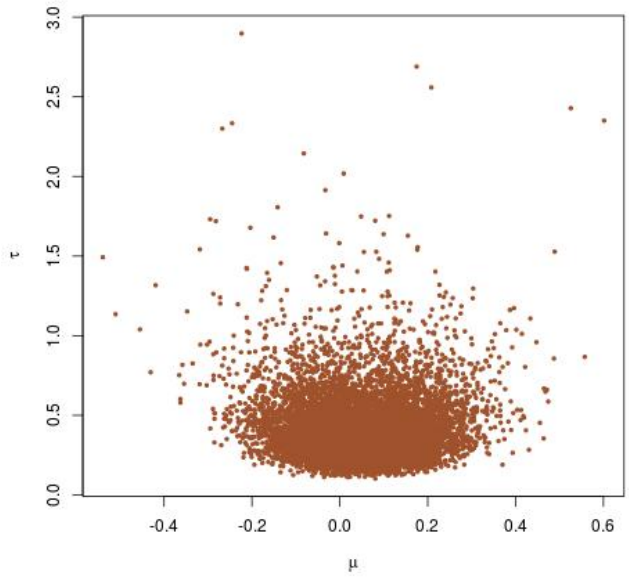

Fig. 5 Sample of $1,000$ simulations from Algorithm 2 when the data is made of $10\,t$ averages with respective sample sizes $n_i = 9,8,8,11,10,5,4,3,5,3$, and when the 10^6 ABC simulations are taken from a pseudo-posterior. *(Note: the true value of* (μ,τ) *is* $(0,1)$*.)*

2.1.4 Enter the summaries

In realistic settings, Algorithm 2 is almost never ever used as such, due to the curse of dimensionality. Indeed, the data $\boldsymbol{x}^0$ is generally complex enough for the proximity $\rho(\mathbf{x}^0,\mathbf{x}^*)$ to be far from small. As illustrated on the time series (toy) example of Marin et al. (2011), the signal-to-noise ratio produced by $\rho(\mathbf{x}^0,\mathbf{x}^*) < \varepsilon$ falls dramatically as the dimension (of the data) increases. This means a corresponding increase in either the total number of simulations N_{ref} or in the tolerance ε is required to preserve a positive acceptance rate. In other words, we are aiming at the parameters to be close rather than the observations themselves. In practice, it is thus paramount to first summarise the data (and decrease the dimension) in a so-called *summary statistic*

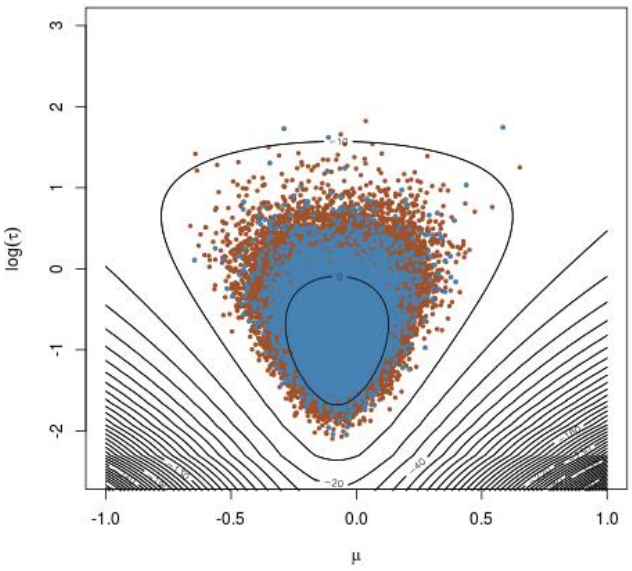

Fig. 6 Same legend as Fig. 5, representing the ABC posterior sample *(in blue)* along with the reference table *(in brown)* and with the level sets of the pseudo-posterior density *(in log scale)*.

before computing a proximity index. Thus enters the notion of *summary statistics*[5] that is central to operational ABC algorithms, as well as the subject of much debate, as discussed in Marin et al. (2011), Blum et al. (2013) and below. A more realistic version of the ABC algorithm is produced in Algorithm 3, where $S(\cdot)$ denotes the summary statistic.

Algorithm 3 ABC (version with summary)

for $t = 1$ to N_{ref} **do**
  Generate $\theta^{(t)}$ from the prior $\pi(\cdot)$
  Generate $\boldsymbol{x}^{(t)}$ from the model $f(\cdot|\theta^{(t)})$
  Compute $d_t = \rho(S(\mathbf{x}^0), S(\mathbf{x}^{(t)}))$
end for
Order distances $d_{(1)} \leq d_{(2)} \leq \ldots \leq d_{(N_{ref})}$
return the values $\theta^{(t)}$ associated with the k smallest distances

[5] While, for a statistician, a statistic is by nature a *summary* of the data, hence making the term redundant, the non-statisticians who introduced this notion in the ABC algorithms felt the need to stress this aspect.

Example 7. Getting back to the Student's t setting of Example 6, the p averages $\bar{x}_{n_i}$ contain information about the parameters μ and τ, but also exhibit variability that is not relevant to the approximation of the posterior probability. It thus makes sense to explore the impact of considering solely the summaries

$$S(\bar{x}_{n_1},\ldots,\bar{x}_{n_p}) = (\bar{\bar{x}}, s^2)$$

already used in the construction of the pseudo-posterior. Algorithm 3 then implies generating pseudo-samples and comparing the values of their summary statistics through a distance. A major issue often overlooked in ABC applications is that the distance needs to be scaled, i.e., the plain sum of squares

$$(\bar{\bar{x}}_1 - \bar{\bar{x}}_2)^2 + (s_1^2 - s_1^2)^2$$

is not appropriate because both components are commensurable. Instead, we suggest using a normalised version like

$$(\bar{\bar{x}}_1-\bar{\bar{x}}_2)^2/\text{mad}(\bar{\bar{x}})^2 + (s_1^2-s_1^2)^2/\text{mad}(s^2)^2$$

where the median absolute deviation (MAD)

$$\text{mad}(S(x)) = \text{median}\,|S(x) - \text{median}(S(x))|$$

is estimated from the (prior) reference table simulated in the ABC algorithm. Running Algorithm 3 with this calibration produces an outcome summarised in Figures 7 and 8. The difference with Figure 4 is striking: while using the same prior, the outcome is not centred around the true value of the parameter in the former case while it is much more accurate in the second case. ◀

The choice of the summary statistics is definitely essential to ensure ABC produces a reliable approximation to the true posterior distribution $\pi(\theta|\mathbf{x}^0)$. A first important remark is that, at best, the outcome of Algorithm 3 will approximate simulation from $\pi(\theta|S(\mathbf{x}^0))$. If the later strongly differs from $\pi(\theta|\mathbf{x}^0)$, there is no way ABC can recover from this. Obviously, when $S(\cdot)$ is a sufficient statistic, there is no loss encurred but this is almost never the case, as exponential families very rarely call for the use of an ABC processing (see Grelaud et al. (2009) for an exception in the setting of the Ising model). A second remark is that, due to the nature of the ABC algorithm, namely the simulation of a huge reference table, followed by the selection of the "closest" parameters, several collections of summaries can be compared at a reasonable computational cost (assuming storing the entire pseudo-datasets a large number of time is feasible).

Example 8. Consider the most standard setting of a normal sample $x_1,\ldots,x_n \sim \mathcal{N}(\mu,\sigma^2)$ under a conjugate prior

$$\mu \sim \mathcal{N}(0,\tau^2)\,, \sigma^{-2} \sim \mathcal{G}a(a,b)\,.$$

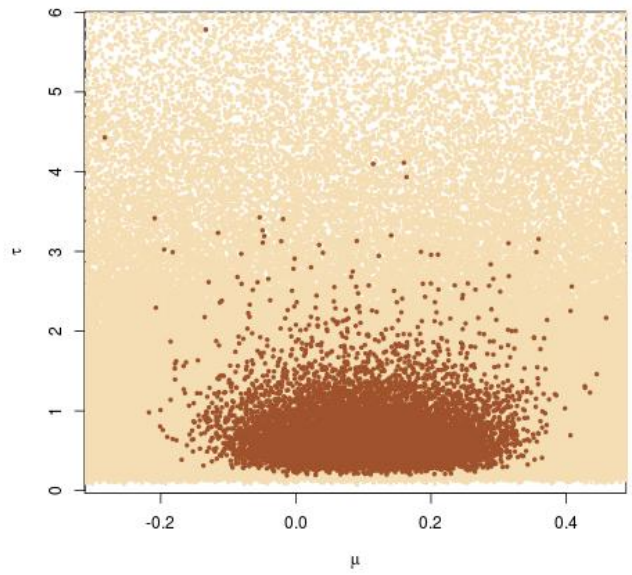

Fig. 7 Sample of $1,000$ simulations from Algorithm 3 when the data is made of $10\,t$ averages with sample sizes $n_i = 21$ and when the 10^6 ABC simulations are taken from the prior. The summary statistics are the empirical mean and variance, while the distance is normalised by the MAD. *(Note: the true value of* (μ,τ) *is* $(0,1)$*.)*

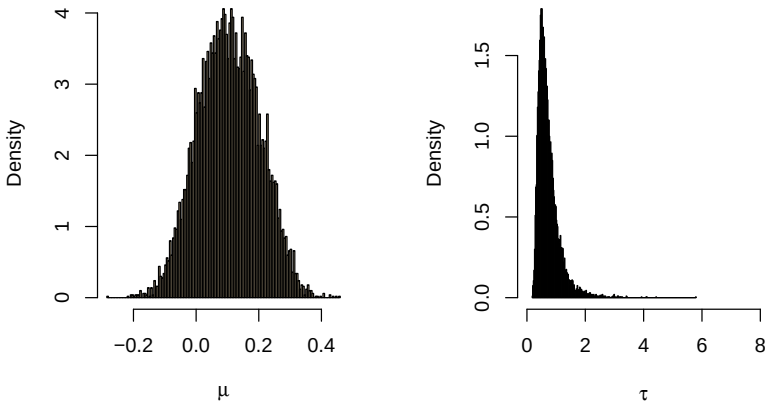

Fig. 8 Marginal histograms of a sample of $1,000$ simulations as in Fig. 7.

If we decide to use the summary statistic $(\bar{x}_n, s_n^2)$, the (true) posterior will not change when compared with using the entire data, since this statistic is sufficient. On the other hand, if we use the pair $(\text{med}(x_1,\ldots,x_n), \text{mad}(x_1,\ldots,x_n))$, it is not sufficient and the (true) posterior will differ. Note that, in this second setting, this true posterior is not available as the joint distribution of the pair $(\text{med}(x_1,\ldots,x_n), \text{mad}(x_1,\ldots,x_n))$ is not available in closed form. Although there is no particular incentive to operate inference conditional on this pair, it provides a most simple illustration of a case when ABC must be used.

In this setting, in order to eliminate scaling effects due to some summaries varying more than others, we once more propose scaling by the mad statistics of those summaries,

$$\rho(S(\mathbf{x}^0), S(\mathbf{x}^{(t)})) = \sum_{i=1}^{2} |S_i(\mathbf{x}^0) - S_i(\mathbf{x}^{(t)}))| \big/ \text{mad}(S_i)$$

where $\text{mad}(S_i)$ is thus based on the reference table.

When implementing ABC based on either of those pairs of summary statistics and a normal dataset of $5,000$ observations, Figure 9 shows that the outcome is identical! Furthermore, a comparison with the genuine output exhibits a significant difference, meaning that the impact of the tolerance is quite large in this case. ◀

2.1.5 Wikipedia entry

For a further introduction to ABC methods, I refer the reader to our earlier survey (Marin et al., 2011). I further recommend Sunnåker et al. (2013), the publication constituting the original version of the Wikipedia page on ABC (Wikipedia, 2014). As a referee of this entry for *PLoS One*, I supported the presentation made in that page as comprehensive and correct, rightly putting stress on the most important aspects of the method. The authors also properly warn about the need to assess assumptions behind and calibrations of the method. (Comments of both referees are included in the original paper, available on-line.)

Note that the ABC method was *not* introduced for conducting model choice, even though this implementation may currently constitute the most frequent application of the method, and the derivation of ABC model choice techniques appeared rather recently (Grelaud et al., 2009, Toni et al., 2009). In almost every setting where ABC is used, there is no non-trivial sufficient summary statistic. Relying on an insufficient statistic then implies a loss of statistical information, as discussed futher below, and I appreciate very much that the authors advertise our warning (Robert et al., 2011) about the potential lack of validity when using an ABC approximation to a Bayes factor for model selection. I also like the notion of "quality control". And the pseudo-example is quite fine as an introduction, while it could be supplemented with the outcome resulting from a large n, to be compared with the true posterior distribution. The section "Pitfalls and remedies" is remarkable in that it details the necessary steps for validating a ABC implementation: the only entry I would remove is the one about

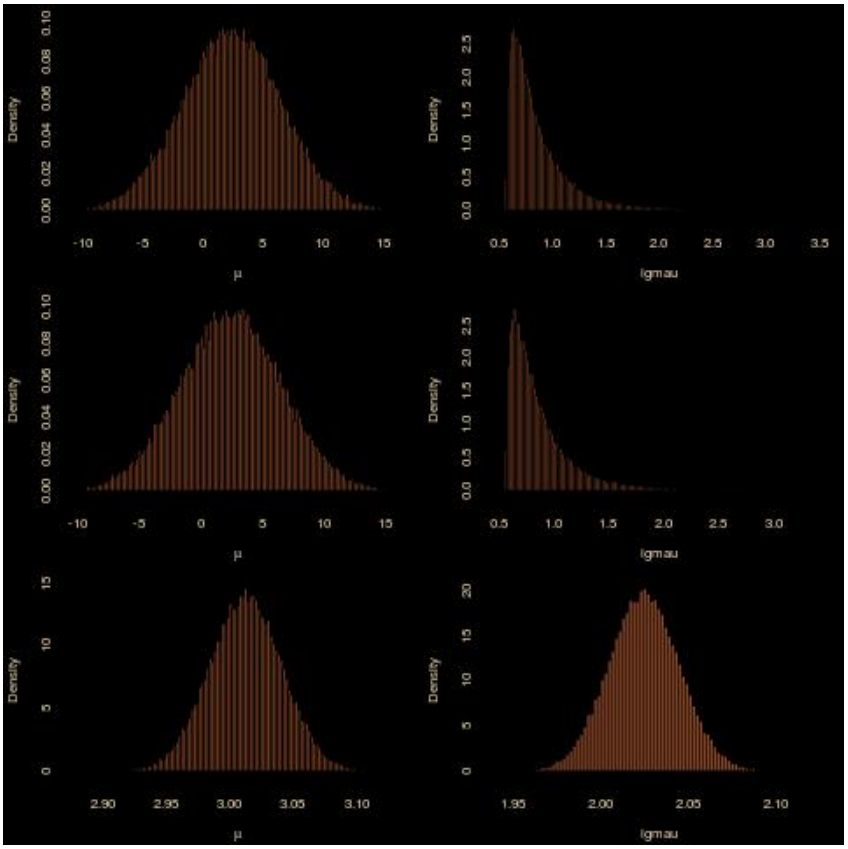

Fig. 9 Marginal histograms in μ and σ^2 *(left and right)* based on two ABC algorithms *(top and middle)* and on the (non-ABC) corresponding Gibbs sampler, for a sample of 5,000 normal $\mathcal{N}(2,4)$ observations, 10^6 ABC and Gibbs iterations, a subsampling rate of 5% for all algorithms and the use of the summary statistics $S(x_1,\ldots,x_n) = (\bar{x}_n, s_n^2)$ *(top)* and $S(x_1,\ldots,x_n) = (\text{med}(x_1,\ldots,x_n), \text{mad}(x_1,\ldots,x_n))$ *(middle)*.

"Prior distribution and parameter ranges", in that this is not a problem inherent to ABC. A last comment is that the section on the non-zero tolerance could emphasise more strongly the fact that this tolerance ε *should not be zero*. (This recommendation may sound paradoxical, but from a practical perspective, $\varepsilon = 0$ can only be achieved with an infinite computing power.)

2.1.6 What does ABC stand for?

An important[6] question that arises in the wake of defining this approximative algorithm is whether or not it constitutes a valid approximation to the posterior distribution $\pi(\theta|S(y_0))$, if not to the original $\pi(\theta|y_0)$. And in case it does not, whether or not it constitutes a proper form of Bayesian inference. Answers to the latter vary according to one's perspective:

– asymptotically, an infinite computing power allows for a zero tolerance, hence for a proper posterior conditioning on $S(y_0)$;
– the outcome of Algorithm 3 is an exact posterior distribution when assuming an error-in-variable model with scale ε (Wilkinson, 2008, 2013);
– it is also an exact posterior distribution once data has been randomised at scale ε (Fearnhead and Prangle, 2012);
– it remains a formal Bayesian procedure albeit applied to an estimated likelihood.

Those answers are not fully satisfactory, in particular because using ABC implies an *ad hoc* modification to the sampling model, but they are also illuminating about the tension that exists between information and precision in complex models. ABC indeed provides a worse approximation of the posterior distribution when the dimension of the summary statistics increase, at a given computational cost. This may sound paradoxical from a purely statistical perspective but it is *in fine* a consequence of the curse of dimensionality and of the fact that the signal-to-noise ratio may be higher in a low dimension statistic than in the raw data. While $\pi(\theta|S(y_0))$ is less concentrated than the original $\pi(\theta|y_0)$, the ABC versions of those two posteriors,

$$\pi(\theta|d(S(Y),S(y_0)) \le \varepsilon_\eta) \quad \text{and} \quad \pi(\theta|d(Y,y_0) \le \varepsilon)\,,$$

may exhibit the opposite feature. (In the above, we introduce the tolerance ε_η to stress the fundamental dependence in the choice of the tolerance on the summary statistics.) A related difficulty with ABC is that the approximation error–of using $\pi(\theta|d(S(Y),S(y_0)) \le \varepsilon_\eta)$ instead of $\pi(\theta|S(y_0))$ or the original $\pi(\theta|y_0)$–is unknown unless one is ready to run costly simulation experiments.

2.2 ABC in the Front Pages

Since their introduction in the late 90's, ABC methods have been implemented in many fields and have helped in validating scientific scenarios and in taking political decisions. Two examples are provided here: First, the science leaflet of *Le Monde* dated October 28, 2012, interviewed my co-author Arnaud Estoup for a work on the multi-coloured Asian ladybird (HA), establishing "that the recent burst of worldwide invasions of HA followed a bridgehead scenario, in which an invasive population in eastern North America acted as the source of the colonists that invaded the European,

[6] Important at least in my opinion!

South American and African continents, with some admixture with a biocontrol strain in Europe".

The Asian ladybird dataset aimed at making inference about the introduction pathway of this insect for the first recorded outbreak of this species in eastern North America. It was first analysed in Lombaert et al. (2011) and Estoup et al. (2012) and includes five natural and biocontrol populations (18 to 35 individuals per sample) genotyped at 18 microsatellite markers. The problem considered—discriminating between evolutionary pathways—required the formalisation and comparison of 10 complex competing scenarios corresponding to various possible routes of introduction (see Analysis 1 in Lombaert et al. (2011) for details).

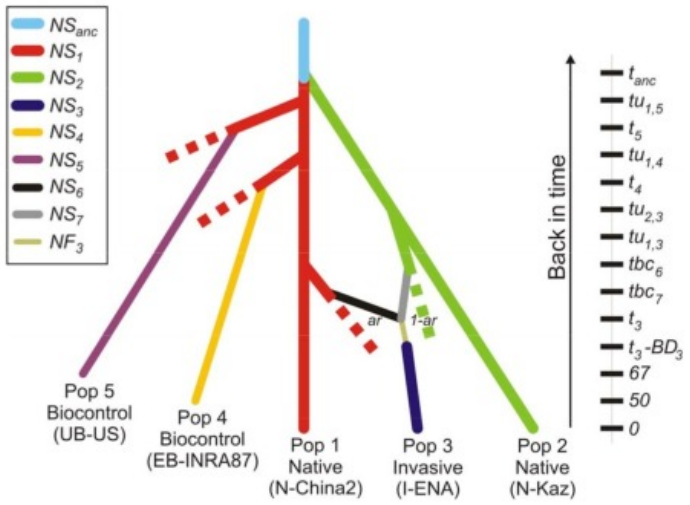

Fig. 10 Topology of the most likely evolutationary tree linking the European Asian ladybird with other populations (*Source: ????*).

> "We predict that control of local badger populations and hence control of environmental transmission will have a relatively limited effect on all measures of bovine TB incidence." E. Brooks-Pollock et al., 2014

The second example relates to a *Nature* paper (Brooks-Pollock et al., 2014) by University of Warwick researchers (including Gareth Roberst) on the modelling of bovine tuberculosis (TB) dynamics in Britain and on the impact of control measures. The data came from the Cattle Tracing System and the VetNet national testing database. The mathematical model was based on a stochastic process and its six parameters were estimated by sequential ABC (SMC-ABC) (Beaumont et al., 2009). The summary statistics chosen in the model were the number of infected farms per county per year and the number of reactors (cattle failing a test) per county per year.

This advanced modelling of a comprehensive dataset on TB in Britain quickly got into a high media profile (see, e.g., *The Guardian* headline of "tuberculosis threat requires mass cull of cattle") as it addressed the definitely controversial culling of badgers (who also carry TB) advocated by the British Government. The study concluded that "only generic measures such as more national testing, whole herd culling or vaccination that affect all routes of transmission are effective at controlling the spread of bovine TB", while the elimination of badgers from the English countryside would have a limited effect. Unsurprisingly, the study was immediately rejected by the UK farming minister: not only did he object to the herd culling solution for economic reasons, but he could not "accept the paper's findings", hopefully not because it relied on ABC techniques.

3 ABC Consistency

While ABC was first perceived with suspicion by the mainstream statistical community (as well as some population geneticists, see Templeton (2008, 2010), Beaumont et al. (2010), Berger et al. (2010), representations of the ABC posterior distribution as a true posterior distribution (Wilkinson, 2013) and of ABC as an auxiliary variable method (Wilkinson, 2013), as a non-parametric technique (Blum, 2010, Blum and François, 2010), connected with both indirect inference (Drovandi et al., 2011) and k-nearest neighbour estimation (Biau et al., 2014) helped to turn ABC into an acceptable component of Bayesian computational methods, albeit requiring caution and calibration (Wikipedia, 2014). The following entries cover some of the advances made in the statistical analysis of the method.

3.1 ABC as knn

Biau et al. (2014) made a significant contribution to the statistical foundations of ABC. It analyses the convergence properties of the ABC algorithm the way it is truly implemented (as in DIYABC (Cornuet et al., 2008) for instance), i.e. with a tolerance bound ε that is determined as a quantile of the simulated distances as in Algorithm 3, say the 10% or the 1% quantile. This means in particular that the interpretation of ε as a non-parametric density estimation bandwidth, while interesting and prevalent in the literature (see, e.g., Blum (2010) and Fearnhead and Prangle (2012)), is only an approximation of the actual practice.

The focus of Biau et al. (2014) is on the mathematical foundations of this practice, an advance obtained by (re)analysing ABC as a k-nearest neighbour (knn) method. Using generic knn results, they derive a consistency property for the ABC algorithm by imposing some constraints upon the rate of decrease of the quantile k as a function of n. More specifically, provided

$$k_N/\log\log N \longrightarrow \infty \quad \text{and} \quad k_N/N \longrightarrow 0$$

when $N \to \infty$, for almost all s_0 (with respect to the distribution of $S(Y)$), with probability 1, convergence occurs, i.e.

$$\frac{1}{k_N}\sum_{j=1}^{k_N} \varphi(\theta_j) \longrightarrow \mathbb{E}[\varphi(\theta_j)|S = s_0]$$

(The setting is restricted to the use of sufficient statistics or, equivalently, to a distance over the whole sample. The issue of summary statistics is not addressed by the paper.) The paper also contains a rigorous proof of the convergence of ABC when the tolerance ε goes to zero. The mean integrated square error consistency of the conditional kernel density estimate is established for a generic kernel (under usual assumptions). Further assumptions (both on the target and on the kernel) allow the authors to obtain precise convergence rates (as a power of the sample size), derived from classical k-nearest neighbour regression, like

$$k_N \approx N^{(p+4)/(m+p+4)}$$

in dimensions m larger than 4 (where N is the simulation size). The paper is theoretical and highly mathematical (with 25 pages of proofs!), but this work clearly constitutes a major reference for the justification of ABC. The authors also mention future work in that direction: I would suggest they consider the case of the insufficient summary statistics from this knn perspective.

3.2 Optimality of Kernels

Filippi et al. (2013) is in the lineage of our (Beaumont et al., 2009) ABC-PMC (population Monte Carlo) paper. The paper focuses on the impact of the transition kernel in our PMC scheme: while we used component-wise adaptive proposals, the paper studies multivariate adaptivity with a covariance matrix adapted from the whole population, or locally or from an approximation to the information matrix. The simulation study run in the paper shows that, even when accounting for the additional cost due to the derivation of the matrix, the multivariate adaptation can improve the acceptance rate by a fair amount. So this is an interesting and positive sequel to our paper.

The main limitation about the paper is that the selection of the tolerance sequence is not done in an adaptive way, while it could, given the recent developments of Del Moral et al. (2012) and of Drovandi and Pettitt (2010) (as well as Marin et al. (2011)). Even though the target is the same for all transition kernels, thus the comparison still makes sense as is, the final product is to build a complete adaptive scheme that comes as close as possible to the genuine posterior.

This paper also raised a new question: there is a slight distinction between the Kullback-Leibler divergence we used and the Kullback-Leibler divergence the authors use here. (In fact, we do not account for the change in the tolerance.) Now, since what only matters is the distribution of the current particles, while the distribution on the past particles is needed to compute the double integral leading to the divergence, there is a complete freedom in the choice of this past distribution. As in Del Moral et al. (2012), the backward distribution $L(\theta_{t-1}|\theta_t)$ could therefore be chosen towards an optimal acceptance rate or something akin.

3.3 Convergence Rates

Dean et al. (2014) addresses ABC consistency in the special setting of hidden Markov models. It relates to Fearnhead and Prangle (2012) discussed below in that those authors also establish ABC consistency for the noisy ABC, given in Algorithm 4, where $\mathfrak{h}(\cdot)$ is a kernel bounded by one (as for instance the unnormalised normal density).

Algorithm 4 ABC (noisy version)

Compute $S(\boldsymbol{x}^0)$ and generate $\tilde{S}^0 \sim \mathfrak{h}(\{s - S(\boldsymbol{x}^0)\}/\varepsilon)$
for $t = 1$ to N **do**
  repeat
    Generate θ^* from the prior $\pi(\cdot)$
    Generate x^* from the model $f(\cdot|\theta^*)$
    Accept θ^* with probability $\mathfrak{h}(\{\tilde{S}^0 - S(\boldsymbol{x})\}/\varepsilon)$
  until acceptance
end for
return N accepted values of θ^*

The authors construct an ABC scheme such that the ABC simulated sequence remains an HMM, the conditional distribution of the observables given the latent Markov chain being modified by the ABC acceptance ball. This means that conducting maximum likelihood (or Bayesian) estimation based on the ABC sample is equivalent to exact inference under the perturbed HMM scheme. In this sense, this equivalence bridges with Wilkinson (2013) and Fearnhead and Prangle (2012) perspectives on "exact ABC". While the paper provides asymptotic bias for a fixed value of the tolerance ε, it also proves that an arbitrary accuracy can be attained with enough data and a small enough ε. The authors of the paper show in addition (as in Fearnhead's and Prangle's) that an ABC inference based on noisy observations $y_1 + \varepsilon z_1, \ldots, y_n + \varepsilon z_n$ with the same tolerance ε, is equivalent to a regular inference based on the original data $y_1, \ldots, y_n$, hence the asymptotic consistence of Algorithm 4. Furthermore, the asymptotic variance of the ABC version is proved to always be greater than the asymptotic variance of the standard MLE, and decreasing as ε^2. The paper also contains an illustration on an HMM with α-stable observables. (Of course,

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 summarising statistics that would not grow in dimension with the size of the HMM series.) In conclusion, a paper that validates noisy ABC without non-parametric arguments and which makes me appreciate even further the idea of noisy ABC: at first, I liked the concept but found the randomisation it involved rather counter-intuitive from a Bayesian perspective. Now, I perceive it as a duplication of the randomness in the data that brings the simulated model closer to the observed model.

Bornn et al. (2014) is a note on the convergence properties of ABC, when compared with acceptance-rejection or regular MCMC. Unsurprisingly, ABC does worse in both cases. What is central to this note is that ABC can be (re)interpreted as a pseudo-marginal method where the data comparison step acts like an unbiased estimator of the true ABC target (not of the original ABC target). From there, it is mostly an application of Andrieu and Vihola (2014) in this setup. The authors also argue that using a single pseudo-data simulation per parameter value is the optimal strategy (as compared with using several), when considering asymptotic variance. This makes sense in terms of simulating in a larger dimensional space but there may be a trade-off when considering the cost of producing those pseudo-datasets against the cost of producing a new parameter. There are a few (rare) cases where the datasets are much cheaper to produce.

Barber et al. (2013) is essentially theoretical and establishes the optimal rate of convergence of the MSE—for approximating a posterior moment—at a rate of $2/(q+4)$, where q is the dimension of the summary statistic, associated with an optimal tolerance in $n^{-1/4}$. I was first surprised at the role of the dimension of the summary statistic, but rationalised it as being the dimension where the non-parametric estimation takes place. There are obviously links with earlier convergence results found in the literature: for instance, Blum (2010) relates ABC to standard kernel density non-parametric estimation and finds a tolerance (bandwidth) of order $n^{-1/q+4}$ and an MSE of order $2/(q+4)$ as well. Similarly, Biau et al. (2014) obtain precise convergence rates for ABC interpreted as a k-nearest-neighbour estimator (Section 3.1). And, as detailed in Section 5.1, Fearnhead and Prangle (2012) derive rates similar to Blum's with a tolerance of order $n^{-1/q+4}$ for the regular ABC and of order $n^{-1/q+2}$ for the noisy ABC.

3.4 Accelerated and Geometric Convergence

Picchini and Lyng Forman (2013) relates to earlier ABC works (and the MATLAB abc-sde package) by the first author. Among other things, it proposes an acceleration device for ABC-MCMC: when simulating from the proposal, the Metropolis-Hastings acceptance probability can be computed and compared with a uniform prior to simulating pseudo-data. In case of rejection, the pseudo-data does not need to be simulated. In case of acceptance, it is compared with the observed data as usual. This is interesting for two reasons: first it always speeds up the algorithm.

Second, it shows the strict limitations of ABC-MCMC, since the rejection takes place without incorporating the information contained in the data. (Even when the proposal incorporates this information, the comparison with the prior does not go this way.) This also relates to one important open problem, namely how to simulate directly summary statistics without simulating the whole pseudo-dataset.

Another thing (related with acceleration) is that the authors use a simulated subsample rather than the simulated sample in order to gain time: this worries me somehow as the statistics corresponding to the observed data is based on the whole observed data. I thus wonder how both statistics could be compared, since they have different distributions and variabilities, even when using the same parameter value. Or is this a sort of pluggin/bootstrap principle, the true parameter being replaced with its estimator based on the whole data? Maybe this does not matter in the end (when accounting for the several levels of approximation).

Lee and Latuszynski (2014) compares four algorithms, from the standard (#1) ABC-MCMC (with N replicates of the pseudo-data) to versions involving simulations of those replicates repeated at the subsequent step (#2), use of a stopping rule in the generation of the pseudo data (#3), and an "ideal" algorithm based on the (unavailable) measure of the ε ball around the data (#4). They recall a result by Tweedie and Roberts (1996), also used in Mengersen and Tweedie (1996), namely that the chain cannot be geometrically ergodic when there exist almost absorbing/sticky states. From there, they derive that (under their technical assumptions) versions #1 and #2 cannot be variance bounding (i.e. the spectral gap is zero), while #3 and #4 can be both variance bounding and geometrically ergodicity under conditions on the prior and the above ball measure. It is thus interesting if a wee bit mysterious that simulating a *random* number of auxiliary variables is sufficient to achieve geometric ergodicity.

3.5 Checking ABC Convergence

Prangle et al. (2013) is a paper on diagnostics for ABC validation via coverage diagnostics. Getting valid approximation diagnostics for ABC is clearly and badly needed. When simulation time is not an issue (!), the DIYABC (Cornuet et al., 2008) software does implement a limited coverage assessment by computing the type I error, i.e. by simulating data under the null model and evaluating the number of time it is rejected at the 5% level (see sections 2.11.3 and 3.8 in the documentation). The current paper builds on a similar perspective.

The core idea is that a (Bayesian) credible interval at a given credible level α should have a similar confidence level (at least asymptotically and even more for matching priors) and that simulating pseudo-data with a known parameter value allows for a Monte-Carlo evaluation of the credible interval "true" coverage, hence for a calibration of the tolerance. The delicate issue is about the generation of those "known" parameters. For instance, if the pair (θ, y) is generated from the joint distribution prior x likelihood, and if the credible region is also based on

the true posterior, the average coverage is the nominal one. On the other hand, if the credible interval is based on a poor (ABC) approximation to the posterior, the average coverage should differ from the nominal one. Given that ABC is always wrong, however, this may fail to be a powerful diagnostic. In particular, when using insufficient (summary) statistics, the discrepancy should make testing for uniformity harder, shouldn't it?

I was puzzled by the coverage property found on page 7:

> "Let $g(\theta|y)$ be a density approximating the univariate posterior $\pi(\theta|y)$, and G_y be the corresponding distribution function. Consider a function $B(\alpha')$ [taking values in the set of Borel sets of] $[0,1]$ defined on $[0,1]$ such that the resulting set has Lebesgue measure α'. Let $C(y,\alpha') = G_y^{-1}(B(\alpha'))$ and $H(\theta,y0)$ be the distribution function for (θ_0,y_0). We say g satisfies the coverage property with respect to distribution $H(\theta_0,y0)$ if for every function B and every α' in $[0,1]$, the probability that θ_0 is in C(y0,α ') is α."

as the probability that belongs to $C(y_0,\alpha')$ is the probability that $G_{y_0}(\theta_0)$ belongs to $B(\alpha')$, which means the conditional of $H(\theta,y_0)$ has to be G_{y_0} if the probability is conditional on y_0. However, I then realised the paper does consider coverage in frequentist terms, which means that the probability is on the pair (θ_0,y_0). In this case, the coverage property will be satisfied for any distribution on y_0 if the conditional is $g(\theta|y)$. This covers both Result 1 and Result 2 (and it seems to relate to ABC being "well-calibrated" for every value of the tolerance, even infinity). I actually find the whole section 2.1 vaguely confusing both because of the use of double indexing ((θ_0,y_0) vs. (θ,y)) and because of the apparent lack of relevance of the posterior $\pi(\theta|y)$ in the discussion (all that matters is the connection between G and H). In their implementation, the authors end up approximating the p-value $P(\theta_0 < \theta)$ and checking for uniformity.

As duly noted in the paper, things get more delicate when $\mathfrak{m}$ the model index itself is involved in this assessment. When integrating the parameters out, the posterior distribution on the model index is a mixture of point masses. Giving e.g. masses 0.7, 0.2, and 0.1 to the three possible values of $\mathfrak{m}$. I thus fail to understand how this translates into [non-degenerate] intervals: I would not derive from these figures that the posterior gives a "70% credible interval that $\mathfrak{m} = 1$" as there is no interval involved. The posterior probability is a number, uniquely defined given the data, without an associated variability in the Bayesian sense. Now, the definition found in the paper is once again one of calibration of the ABC distributions, already discussed in Fearnhead and Prangle (2012). (The paper actually makes not mention of calibration.) At last, I am also lost as to why the calibration condition (5) on the posterior distribution of the model index is a testable one: there is a zero probability to observe again a given value of the posterior probability $g(\mathfrak{m}|y)$ when generating a new y. In the following diagnostic, the authors use instead a test that the (generated) model index is an outcome from a Bernoulli with parameter the posterior probability,

3.6 Threshold Schedules

Silk et al. (2013) attack a typical problem with SMC (and PMC, and even MCMC!) methods, namely the ability to miss (global) modes of the target due to a poor initial exploration. So, if a proposal is built on previous simulations and if those simulations have missed an important mode, it is quite likely that this proposal will concentrate on other parts of the space and miss the important mode even more. This is also why simulated annealing and other stochastic optimisation algorithms are so fickle: a wrong parameterisation (like the temperature schedule) induces the sequence to converge to a local optimum rather than to the global optimum. Since sequential ABC is a form of simulated annealing, the decreasing tolerance (or threshold) playing the part of the temperature, it is no surprise that it suffers from this generic disease.

The method proposed in the paper aims at avoiding this difficulty by looking at sudden drops in the acceptance rate curve (as a function of the tolerance ε),

$$\aleph_t(\varepsilon) = \int p_t(x)\mathbb{I}(\Delta(x,x^\star) \leq \varepsilon)\,\mathrm{d}x,$$

suggesting for threshold the value of ε that maximises the second derivative of this curve. Now, before getting to the issue of turning this principle into a practical algorithm, let me linger at the motivation for it:

> "To see this, one can imagine an ε-ball expanding about the true data; at first the ball only encompasses a small number of particles that were drawn from very close to the global maximum, corresponding to the low gradient at the foot of the shape. Once ε is large enough we are able to accept the relatively large number of particles sitting in the local maximum, which causes the increase in gradient." D. Silk et al., 2013

Thus, the argument for looking at values of ε preceding fast increases in the acceptance rate $\aleph$ is that we are thus avoiding flatter and lower regions of the posterior support corresponding to a local maximum. It clearly encompasses the case studied by the authors of a global highly concentrated global mode, surrounded by flatter lower modes, but it seems to me that this is not necessarily the only possible reason for changes in the shape of the acceptance probability $\aleph$. First, we are looking at an ABC acceptance rate, not at a genuine likelihood. Second, this acceptance rate function depends on the current (and necessarily rough) approximate to the genuine predictive, p_t. Furthermore, when taking into account this rudimentary replacement of the true likelihood function, it is rather difficult to envision how it impacts the correspondence between a proximity in the data and a proximity in the parameter space. (The toy example is both illuminating and misleading, because it considers a setting where the data is a deterministic transform of the parameter.) I thus think the analysis should refer more strongly to the non-parametric literature and in particular to the k-nearest-neighbour approach reformulated by Biau et al. (2014): there is no reason to push the tolerance ε all the way down to zero as this limit does not correspond to the optimal value of the tolerance. If one does not use a non-parametric determination of the "right" tolerance, the lingering question is when

and why stopping the sequence of ABC simulations. The acceptance rate function ℵ is approximated using an unscented transform that escapes me.

In Sedki et al. (2012), we build on the sequential ABC scheme of Del Moral et al. (2012), where the tolerance level at each step is adapted from the previous iterations as a quantile of the distances. (The quantile level is based on a current effective sample size.) In a "systematic" step, the particles that are closest to the observations are preserved and duplicated, while those farther away are sampled randomly. The resulting population of particles is then perturbed by an adaptive (random walk) kernel and the algorithm stops when the tolerance level does not decrease any longer or when the acceptance rate of the Metropolis step is too low. Pierre Pudlo and Mohammed Sedki experimented a parallel implementation of the algorithm, with an almost linear improvement in the number of cores. It is less clear the same would work on a GPU because of the communication requirements. Overall, the new algorithm brings a significant improvement in computing time when compared with earlier versions, mainly because the number of simulations from the model is minimised. (It was tested on a rather complex population scenario retracing the invasion of honeybees in Europe.)

3.7 ABC for big data

> "The results in this paper suggest that ABC can scale to large data, at least for models with a xed number of parameters, under the assumption that the summary statistics obey a central limit theorem." W. Li and P. Fearnhead, 2015

Li and Fearnhead (2015) propose a different lecture of ABC convergence, in that they see it as a big data issue. I somewhat disagree with this choice in that the paper does not address the issue of big or tall data *per se*, e.g., the impossibility to handle the whole data at once and to reproduce it by simulation, but rather asymptotics of ABC. The setting is not dissimilar to the earlier Fearnhead and Prangle (2012) Read Paper. The central theme of this theoretical paper is to study the connection between the number N of Monte Carlo simulations and the tolerance value ε when the number of observations n goes to infinity. A main result of Li and Fearnhead (2015) is that the ABC posterior mean can have the same asymptotic distribution as the MLE when $\varepsilon = o(n^{-1/4})$. This is however of no direct use in practice as the second main result that the Monte Carlo variance is well-controlled only when $\varepsilon = O(n^{-1/2})$. However, as pointed out by the authors (comments on xianblog.wordpress.com), the Monte Carlo variance can be kept under control by a "sensible" choice of importance function, even though it is hard to imagine a universal strategy in this respect.

It may thus seem unrealistic to envision the construction of an importance sampling function of the form $f_{\text{ABC}}(s|\theta)^{\alpha}$ when, obviously, this function cannot be used for simulation purposes. The authors acknowledge this fact, but still build an argument about the optimal choice of α, namely away from 0 and 1, for instance $\frac{1}{2}$. Actually, *any* value different from 0 and 1, is sensible, meaning that the range of acceptable importance functions is wide, which is the key message there (see

comments). Most interestingly, the paper constructs an iterative importance sampling ABC in a spirit similar to Beaumont et al. (2009) ABC-PMC. Even more interestingly, the $\frac{1}{2}$ factor amounts to updating the scale of the proposal as twice the scale of the target, just as in PMC.

Another aspect of the analysis I do not catch is the reason for keeping the Monte Carlo sample size to a fixed value N, while setting a sequence of acceptance probabilities (or of tolerances) along iterations. This is a very surprising result in that the Monte Carlo error does remain under control and does not dominate the overall error!

> "Whilst our theoretical results suggest that point estimates based on the ABC posterior have good properties, they do not suggest that the ABC posterior is a good approximation to the true posterior, nor that the ABC posterior will accurately quantify the uncertainty in estimates." W. Li and P. Fearnhead, 2015

Overall, this is clearly a paper worth reading for understanding the convergence issues related with ABC. With more theoretical support than the earlier Fearnhead and Prangle (2012). However, it does not provide guidance into the construction of a sequence of Monte Carlo samples nor does it discuss the selection of the summary statistic, which has obviously a major impact on the efficiency of the estimation. And to relate to the earlier warning, it does not cope with "big data" in that it still reproduces the original simulation of the n sized sample.

4 Improvements, implementations, and applications

4.1 ABC for State-Space Models

As described in the survey written by Jasra (Jasra, 2014) on the use of ABC methods in a rather general class of time-series models, those methods allow to handle setting where the likelihood of the current observation conditional on the past observations and on a latent (discrete-time) process cannot be computed. Jasra makes the preliminary useful remark that, in most cases, the probabilistic structure of the model (e.g., an hidden Markov type of dependence) is lost within the ABC representation. An exception Jasra and other authors (Calvet and Czellar, 2014, Dean et al., 2014, Ehrlich et al., 2014, Jasra et al., 2014, 2012, Martin et al., 2014, McKinley et al., 2014) exploited quite thoroughly is when the difference between the observed data and the simulated pseudo-data is operated time step by time step, as in

$$\prod_{t=1} \mathbb{I}_{d(y_t, y_t^0) \leq \varepsilon}$$

where $y^0 = (y_1^0, \ldots, y_T^0)$ is the actual observation. The ABC approximation indeed retains the same likelihood structure and allows for derivations of consistency properties (in the number of observations) of the ABC estimates. In particular, using such a distance in the algorithm allows for the approximation to converge to the genuine

posterior when the tolerance ε goes to zero (Biau et al., 2014). This is the setting where (Fearnhead and Prangle, 2012, Theorem 2, see also Dean et al. (2014)) show that noisy ABC is well-calibrated, i.e. has asymptotically proper convergence properties. Most of the results obtained by Jasra and co-authors are dedicated to specific classes of models, from iid models (Dean et al., 2014, Fearnhead and Prangle, 2012, Jasra et al., 2014) to "observation-driven times-series" (Jasra et al., 2014) to other forms of HMM (Dean et al., 2014, Ehrlich et al., 2014, Martin et al., 2014) mostly for MLE consistency results. The constraint above leads to computational difficulties as the acceptance rate quickly decreases with n (unless the tolerance ε is increasing with n). Jasra et al. (2014) then suggest to raise the number of pseudo-observations to average indicators in the above product and to make it random in order to ensure a fixed number of acceptances. Moving to SMC within MCMC, Jasra et al. (2013) establish unbiasedness and convergence within this framework, in connection with the alive particle filter (Le Gland and Oudjane, 2006).

4.2 Accelerated ABC

Richard Wilkinson (2014) starts with a link to the synthetic likelihood approximation of Wood (2010). Wilkinson presents the generalised ABC as a kernel-based acceptance probability, using a kernel $\pi(y|x)$, when y is the observed data and $x = x(\theta)$ the simulated one. He proposes a Gaussian process modelling for the log-likelihood (at the observed data y), with a quadratic (in θ) mean and Matérn covariance matrix. Hence the connection with Wood's synthetic likelihood. Another connection is with QMC (Niederreiter, 1992): the θ's are chosen following a Sobol sequence "in order to minimise the number of design points". Which requires a reparameterisation to $[0,1]^p$. I find this "uniform" exploration of the whole parameter space delicate to envision in complex parameter spaces and realistic problems, since the likelihood is highly concentrated on a tiny subregion of the original $[0,1]^p$. Not mentioning the issue of the spurious mass on the boundaries of the hypercube possibly induced by the change of variable. Wilkinson's sequential algorithm also attempts at eliminating implausible zones of the parameter space. i.e. zones where the likelihood is essentially zero. My questions about this interesting notion are that (a) the early Gaussian process approximations may be poor and hence exclude zones they should not; (b) all Gaussian process approximations at all iterations must be saved; (c) the Sobol sequences apply to the whole $[0,1]^p$ at each iteration but the non-implausible region shrinks at each iteration, which induces a growing inefficiency in the algorithm. The Sobol sequence should be restricted to the previous non-implausible zone.

Overall, this inclusion of Gaussian processes clearly is an interesting proposal that would need more prodding to understand whether or not it is robust to poor initialisation and complex structures. And a proposal belonging to the estimated likelihood branch of ABC, which makes use of the final Gaussian process approximation to run an MCM algorithm. Without returning to pseudo-data simulation, replacing it with log-likelihood simulation.

> "These algorithms sample space randomly and naïvely and do not learn from previous simulations." R. Wilkinson, 2014

The above criticism is moderated in a footnote about ABC-SMC using the "current parameter value to determine which move to make next [but] parameters visited in previous iterations are not taken into account". I still find it excessive in that SMC algorithms and in particular ABC-SMC algorithms are completely free to use the whole past to build the new proposal. This was clearly enunciated in our earlier population Monte Carlo papers. For instance, the complete collection of past particles can be recycled by weights computing through our AMIS algorithm (Cornuet et al., 2012).

4.3 ABC-SubSim

Chiachio et al. (2014) developed a new ABC algorithm, called *ABC-SubSim*. The *SubSim* stands for subset simulation and corresponds to an approach developed by one of the authors for rare-event simulation. This approach looks somewhat similar to the cross-entropy method of Rubinstein and Kroese (2004), in that successive tail sets are created towards reaching a very low probability tail set. Simulating from the current subset increases the probability to reach the following and less probable tail set. The extension to the ABC setting is done by looking at the acceptance region (in the augmented space) as a tail set and by defining a sequence of tolerances. The paper could also be connected with nested sampling (Skilling, 2006) in that constrained simulation through MCMC occurs there as well. Following the earlier paper, the MCMC implementation therein is a random-walk-within-Gibbs algorithm. This is somewhat the central point in that the sample from the previous tolerance level is used to start a Markov chain aiming at the next tolerance level. (Del Moral et al. (2012) use instead a particle filter, which could easily be adapted to the modified Metropolis move considered in the paper.) The core difficulty with this approach, not covered in the paper, is that the MCMC chains used to produce samples from the constrained sets have to be stopped at some point, especially since the authors run those chains in parallel. The stopping rule is not provided (see, e.g., Algorithm 3) but its impact on the resulting estimate of the tail probability could be far from negligible, in particular because there is no burning. The authors re-examined the MA(2) toy benchmark we had used in Marin et al. (2011), reproducing as well the graphical representation on the simplex as shown above.

4.4 Adaptive ABC

Lenormand et al. (2013) develop a refinement of the ABC-PMC algorithm of ours (Beaumont et al., 2009). The authors state in their introduction that ABC-PMC

> "...presents two shortcomings which are particularly problematic for costly to simulate complex models. First, the sequence of tolerance levels $\varepsilon_1, \ldots, \varepsilon_T$ has to be provided to the ABC algorithm. In practice, this implies to do preliminary simulations of the model, a step which is computationally costly for complex models. Furthermore, a badly chosen sequence of tolerance levels may inflate the number of simulations required to reach a given precision as we will see below. A second shortcoming of the PMC-ABC algorithm is that it lacks a criterion to decide whether it has converged. The final tolerance level ε_T may be too large for the ABC approach to satisfactorily approximate the posterior distribution of the model. Inversely, a larger ε_T may be sufficient to obtain a good approximation of the posterior distribution, hence sparing a number of model simulations." Lenormand et al., 2013

shortcomings which I thought were addressed by the ABC-SMC algorithm of Del Moral et al. (2012), the similar algorithm of Drovandi and Pettitt (2010), and our recent paper (Sedki et al., 2012). It is correct that we did not address the choice of the ε_t's in the original paper, even though we already used an on-line selection as a quantile of the current sample of distances. In essence, given the fundamentally non-parametric nature of ABC, the tolerances ε_t should always be determined from the simulated samples, as regular bandwidths. The paper essentially proposes the same scheme as in Del Moral et al. (2012), before applying it to the toy example of Sisson et al. (2007) and to a more complex job dynamic model in central France.

4.5 ABC with Empirical Likelihood

Mengersen et al. (2013) is a paper on ABC using empirical likelihood (EL) that was started by me listening to Brunero Liseo's tutorial in O'Bayes-2011 in Shanghai. Brunero mentioned empirical likelihood as a semi-parametric technique w/o much Bayesian connections and this got me thinking of a possible recycling within ABC. The details about empirical likelihood can be found in Owen (2001), a comprehensive entry, The core idea of empirical likelihood is to use a maximum entropy discrete distribution supported by the data and constrained by estimating equations related with the parameters of interest or of the whole model. Given a dataset $\boldsymbol{x}$ made of n independent replicates $\boldsymbol{x} = (x_1, \ldots, x_n)$ of a rv $X \sim F$, and a collection of generalized moment conditions that identify the parameter (of interest) ϕ

$$\mathbb{E}_F\left[h(X, \phi)\right] = 0$$

where h is a known function, the induced empirical likelihood (Owen, 1988) is defined as

$$L_{\mathrm{el}}(\phi|x) = \max_p \prod_{i=1}^{n} p_i$$

where the maximum is taken on for all p's on the simplex of $\mathbb{R}^n$ such that

$$\sum_i p_i h(x_i, \phi) = 0.$$

As such, it is a non-parametric approach in the sense that the distribution of the data does not need to be specified, only some of its characteristics. Econometricians have been quite busy at developing this kind of approach over the years, see e.g. Gouriéroux and Monfort (1995). However, this empirical likelihood technique can also be seen as a convergent approximation to the likelihood and hence exploited in cases when the exact likelihood cannot be derived. For instance, as a substitute to the exact likelihood in Bayes' formula, as sketched in Algorithm 5.

Algorithm 5 ABC (with empirical likelihood)

for $i = 1 \to N$ **do**
  Generate ϕ_i from the prior distribution $\pi(\cdot)$
  Set the weight $\omega_i = L_{\text{el}}(\phi_i|x_{\text{obs}})$
end for
return (ϕ_i, ω_i), $i = 1, \ldots, N$
Use weighted sample as in importance sampling

Our paper thus examines the consequences of using an empirical likelihood in ABC contexts. Although we called the derived algorithm *ABCel*, it differs from genuine ABC algorithms in that it does not simulate pseudo-data.[7] We had indeed started looking at a simulated data version, but the performances were rather poor, and we thus opted for an importance sampling version where the parameters are simulated from an importance distribution (e.g., the prior) and then weighted by the empirical likelihood (times a regular importance factor if the importance distribution is not the prior).

The difficulty with the method is in connecting the parameters of the assumed distribution with moments of the (iid) data. While this operates rather straightforwardly for quantile distributions (Allingham et al., 2009), it is less clear for dynamic models like ARCH and GARCH (Bollerslev et al., 1992), where we have to reconstruct the underlying iid process. (In this setting, ABCel clearly improves upon ABC for the GARCH(1,1) model but remains less informative than a regular MCMC analysis. And it is even harder for population genetic models, where parameters like divergence dates, effective population sizes, mutation rates, &tc., cannot be expressed as moments of the distribution of the sample at a given locus. In particular, the data-points are not iid. Pierre Pudlo then had the brilliant idea to resort instead to a composite likelihood, approximating the intra-locus likelihood by a product of pairwise likelihoods over all pairs of genes in the sample at a given locus. Indeed, in Kingman's coalescent theory, the pairwise likelihoods can be expressed in closed form, hence we can derive the pairwise composite scores. The comparison with optimal ABC outcomes shows an improvement brought by ABCel in the approximation, at an overall computing cost that is negligible against ABC (i.e., it takes minutes to produce the ABCel outcome, compared with hours for ABC.) We are now looking

[7] The final accronym in Mengersen et al. (2013) thus became BC_{el}, to stress the difference with "standard" ABC.

for extensions and improvements of ABCel, both at the methodological and at the genetic levels, and we would of course welcome any comment at this stage.

Example 9. Quantile distributions constitute a special class of parametric distributions that are (only) defined through their quantile function $F^{-1}(p;\theta)$, in the sense that they have no closed form density function (Haynes et al., 1997, Gilchrist, 2000). They thus constitute an excellent benchmark for ABC algorithms since the corresponding likelihood function is untractable but they can be simulated by the inverse cdf result (Robert and Casella, 2004, Chapter 2). See (Marjoram et al. (2003), McVinish (2012)) for illustrations.

The four-parameter g-and-k distribution is associated with the quantile function

$$Q(r;A,B,g,k) = A + B\left(1 + c\frac{1-\exp(-gz(r))}{1+\exp(-gz(r))}\right)\left(1+z(r)^2\right)^k z(r)$$

where $z(r)$ is the rth normal $\mathcal{N}(0,1)$ quantile. Mengersen et al. (2013) evaluated our BC_{el} algorithm on this distribution using different ranges of quantiles to define the empirical likelihood. Figure 11 illustrates the result of this evaluation for several sets of quantiles and two sample sizes. As described in Mengersen et al. (2013), the performances of BC_{el} are overall comparable with those obtained by Allingham et al. (2009) when using MCMC for the same distribution. However, the much improved speed of BC_{el} compared with competing ABC algorithms means that higher performances could be achieved by increasing the number of simulations and using more efficient simulations than sampling from the prior, as suggested by (Drovandi and Pettitt, 2011) for ABC. ◀

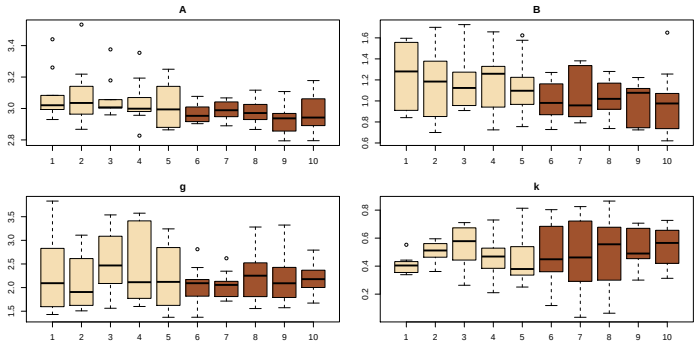

Fig. 11 Boxplots of the variations of the posterior means of the four parameters of the g-and-k distribution, based on BC_{el} approximations, for $n = 100$ observations (1 to 5) and $n = 500$ observations (6 to 10), using $M = 10^4$ simulations and 10 replications. The moment conditions used in the BC_{el} algorithm are quantiles of order $(0.2, 0.5, 0.8)$ (1 and 6), $(0.2, 0.4, 0.6, 0.8)$ (2 and 7), $(0.1, 0.4, 0.6, 0.9)$ (3 and 8), $(0.1, 0.25, 0.5, 0.75, 0.9)$ (4 and 9), and $(0.1, 0.2, \ldots .0.9)$ (5 and 10). *(Source: Mengersen et al., 2013.)*

In the revision of this paper, the main issue raised by the referees was that the potential use of the empirical likelihood (EL) approximation is much less widespread than the possibility of simulating pseudo-data, because EL essentially relies on an *iid* sample structure, plus the availability of parameter defining moments. This is indeed the case to some extent and also the reason why we used a compound likelihood for our population genetic model. There are in fact many instances where we simply cannot come up with a regular EL approximation... However, the range of applications of straight EL remains wide enough to be of interest, as it includes most dynamical models like hidden Markov models. To illustrate this point further, we added in the revision an example borrowed from Cox and Kartsonaki (2012), which proposes a frequentist alternative to ABC based on fractional design. This model ended up being fairly appealing wrt our perspective: while the observed data is dependent in a convoluted way, being a superposition of N renewal processes with gamma waiting times, it is possible to recover an iid structure at the same cost as a regular ABC algorithm by using the pseudo-data to recover an iid process. This revision was thus quite beneficial to our perception of ABC in that (a) it is indeed not as universal as regular ABC and this restriction should be spelled out (the advantage being that, when it can be implemented, it usually runs much much faster!), and (b) in cases where the pseudo-data must be simulated, EL provides a reference/benchmark for the ABC output that comes for free.

Zhu et al. (2014) proposes an alternative to our paper, based on Davison et al. (1992) bootstrap likelihood, which relies on a double-bootstrap to produce a non-parametric estimate of the distribution of a given estimator of the parameter θ. Including a smooth curve-fitting algorithm step, for which not much description is available from the paper.

> "...in contrast with the empirical likelihood method, the bootstrap likelihood doesn't require any set of subjective constrains taking advantage from the bootstrap methodology. This makes the algorithm an automatic and reliable procedure where only a few parameters need to be specified."

The spirit is indeed quite similar to ours in that a non-parametric substitute plays the role of the actual likelihood, with no correction for the substitution. Both approaches are convergent, with similar or identical convergence speeds. While the empirical likelihood relies on a choice of parameter identifying constraints, the bootstrap version starts directly from the [subjectively] chosen estimator of θ. For it indeed needs to be chosen. And computed.

> "Another benefit of using the bootstrap likelihood (...) is that the construction of bootstrap likelihood could be done once and not at every iteration as the empirical likelihood. This leads to significant improvement in the computing time when different priors are compared."

This is an improvement that could apply to the empirical likelihood approach, as well, once a large enough collection of likelihood values has been gathered. But only in small enough dimensions where smooth curve-fitting algorithms can operate. The same criticism applying to the derivation of a non-parametric density estimate for the distribution of the estimator of θ. Critically, the paper only processes examples with a few parameters.

In the comparisons between BCel and BCbl that are produced in the paper, the gain is indeed towards BCbl. Since this paper is mostly based on examples and illustrations, not unlike ours, I would like to see more details on the calibration of the non-parametric methods and of regular ABC, as well as on the computing time. And the variability of both methods on more than a single Monte Carlo experiment.

I am however uncertain as to how the authors process the population genetic example. They refer to the composite likelihood used in our paper to set the moment equations. Since this is not the true likelihood, how do the authors select their parameter estimates in the double-bootstrap experiment? The inclusion of Crackel and Flegal (2014) bivariate Beta, is somewhat superfluous as this example sounds to me (see above) like an artificial setting.

In the case of the Ising model, maybe the pre-processing step in Moores et al. (2014) could be compared with the other algorithms. In terms of BCbl, how does the bootstrap operate on an Ising model, i.e. (a) how does one subsample pixels and (b)what are the validity guarantees?

A test that would be of interest is to start from a standard ABC solution and use this solution as the reference estimator of θ, then proceeding to apply BCbl for that estimator. Given that the reference table would have to be produced only once, this would not necessarily increase the computational cost by a large amount...

Li and Jiang (2014) is connected with an interrogation of ours on the manner to extend our empirical likelihood ABC (Mengersen et al., 2013) to model choice. The current paper is of a theoretical nature, considering a moment defined model

$$\mathbb{E}[g(D,\theta)] = 0,$$

where D denotes the data, as the dimension p of the parameter θ grows with n, the sample size. The approximate model is derived from a prior on the parameter θ and of a Gaussian quasi-likelihood on the moment estimating function $g(D,\theta)$. Examples include single index longitudinal data, quantile regression and partial correlation selection. The model selection setting is one of variable selection, resulting in 2^p models to compare, with p growing to infinity. Which makes the practical implementation rather delicate to conceive. And the probability one of hitting the right model a fairly asymptotic concept.

4.6 ABC via Regression Density Estimation

Fan et al. (2013) argue that one could take advantage of the joint simulation of the pair parameter/sample to derive a non-parametric estimate of the conditional distribution of the summary statistic given the parameter, i.e. the sampling distribution. While most or even all regular ABC algorithms do implicitly or explicitly rely on some level of non-parametric estimation, from non-parametric regression (Beaumont et al., 2002) to direct derivations on non-parametric convergence speeds on the kernel bandwidths (Blum and François, 2010, Fearnhead and Prangle, 2012, Biau et al., 2014), this

paper centres on the idea to use those same simulations ABC relies upon to build an estimate of the sampling distribution, to be used afterwards as the likelihood in either Bayesian or frequentist inference. Rather than relying on traditional kernel estimates, the adopted approach merges mixtures of experts, namely normal regression mixtures with logit weights (Jordan and Jacobs, 1994) for the marginals, along with a copula representation of the joint distribution (of the summary statistics).

So this is a new kid on the large block of ABC methods! In terms of computing time, it sounds roughly equivalent to regular ABC algorithms in that it relies on the joint simulation of the pair parameter/sample. Plus a certain number of mixtures/mixtures of experts estimations. I have no intuition on how greedy those estimations are. In their unique illustration, the authors report density estimation in dimension 115, which is clearly impressive. I did not see any indication of respective computing times. In terms of inference and connection with the Bayesian exact posterior, I see a few potential caveats: first, the method provides an approximation of the conditional density of the summary statistics given the parameters, while the Bayesian approach considers the opposite. This could induce inefficiencies when the prior is vague and leads to a very wide spread for the values of the summary statistics. Using a neighbourhood of the observed statistics to restrict the range of the simulated statistics thus seems appropriate. Second, the use of mixtures of experts assume some linear connection between the parameters and the summary statistics: while this reflects Fearnhead and Prangle (2012) strategy, this is not necessarily appropriate in settings where those parameters cannot be expressed directly as expectations of the summary statistics (see, e.g., the case of population genetics). Third, the approximation proposed by the paper is a pluggin estimate, whose variability and imprecision are not accounted for in the inference process. Maybe not a major issue, as other solutions also rely on pluggin estimates. And I note the estimation is done once for all, when compared with, e.g., our empirical likelihood solution that requires a (fast) optimisation for each new value of the parameters. Fourth, once the approximation is constructed, a new MCMC run is required and since the (approximated) target is a black box the calibration of the MCMC sampler may turn to be rather delicate, as in the 115 dimensional example.

Li et al. (2015) extends on the above through two central ideas: (i) estimate marginal posterior densities for the components of the model parameter by non-parametric means; and (ii) consider all pairs of components to deduce the correlation matrix R of the Gaussian (pdf) transform of the pairwise rank statistic. From those two low-dimensional estimates, the authors derive a joint Gaussian-copula distribution by using inverse pdf transforms and the correlation matrix R, to end up with a meta-Gaussian representation

$$f(\theta) = \frac{1}{|R|^{1/2}} \exp\{\eta'(I - R^{-1})\eta/2\} \prod_{i=1}^{p} g_i(\theta_i)$$

where the η's are the Gaussian transforms of the inverse-cdf transforms of the θ's,that is,

$$\eta_i = \Phi^{-1}(G_i(\theta_i))$$

or rather

$$\eta_i = \Phi^{-1}(\hat{G}_i(\theta_i))$$

given that the g's are estimated. This is obviously an approximation of the joint in that, even in the most favourable case when the g's are perfectly estimated, and thus the components perfectly Gaussian, the joint is not necessarily Gaussian. But it sounds quite interesting, provided the cost of running all those transforms is not overwhelming. For instance, if the g's are kernel density estimators, they involve sums of possibly a large number of terms.

One thing that bothers me in the approach, albeit mostly at a conceptual level for I realise the practical appeal is the use of *different* summary statistics for approximating *different* uni- and bi-dimensional marginals. This makes for an incoherent joint distribution, again at a conceptual level as I do not see immediate practical consequences. Those local summaries also have to be identified, component by component, which adds another level of computational cost to the approach, even when using a semi-automatic approach as in Fernhead and Prangle (2012). Although the whole algorithm relies on a single reference table. As pointed out by the authors in comments on xianblog.wordpress.com, the impact of using a subset of the whole vector of summary statistics can be checked against incoherences, at least on principle as the cost may get quickly huge. They also note existing work in density estimation on such approaches.

The examples in the paper are (i) the banana shaped "Gaussian" distribution of Haario et al. (1999) that we used in our PMC papers, with a twist; and (ii) a g-and-k quantile distribution. The twist in the banana (!) is that the banana distribution is the prior associated with the mean of a Gaussian observation. In that case, the meta-Gaussian representation seems to hold almost perfectly, even in $p = 250$ dimensions. (If I remember correctly, the hard part in analysing the banana distribution was reaching the tails, which are extremely elongated in at least one direction.) For the g-and-k quantile distribution, the same holds, even for a regular ABC. What seems to be of further interest would be to exhibit examples where the meta-Gaussian is clearly an approximation. If such cases exist.

4.7 ABC via emulators

Jabot et al. (2014) run a comparison of so-called *emulation methods* for ABC. The idea therein is to bypass costly simulations of pseudo-data by running cheaper simulations from a pseudo-model (or emulator) constructed via a preliminary run of the original and costly model. To borrow from the paper introduction, ABC-Emulation runs as follows:

- design a small number n of parameter values covering the parameter space;
- generate n corresponding realisations from the model and store the corresponding summary statistics;
- build an emulator (model) based on those n values;

– run ABC using the emulator in lieu of the original model.

A first emulator proposed in the paper is absed on local regression, as in Beaumont et al. (2002), except that it goes the *reverse* way: the regression model predicts a summary statistics given the parameter value. The second emulator relies on Gaussian processes, as in Wilkinson (2014) as well as Meeds and Welling (2014). The comparison of the above emulators is based on an ecological community dynamics model. The results are that the stochastic version is superior to the deterministic one, but overall not very useful when implementing the Beaumont et al. (2002) correction. The paper however does not define what deterministic and what stochastic mean.

> "We therefore recommend the use of local regressions instead of Gaussian processes." Jabot et al., 2015

While I find the conclusions of the paper somewhat over-optimistic given the range of the experiment and the limitations of the emulator options (like non-parametric conditional density estimation), it seems to me that this is a direction to be pursued as we need to be able to simulate directly a vector of summary statistics instead of the entire data process, even when considering an approximation to the distribution of those summaries.

4.8 Hamiltonian ABC

Meeds et al. (2015) manages the *tour de force* of associating antagonistic terms, since ABC is intended for complex and mostly intractable likelihoods, while Hamiltonian Monte Carlo requires a lot from the target, in order to compute gradients and Hessians.

Somewhat obviously (it is always obvious, ex-post!), the paper suggests to use Hamiltonian dynamics on ABC approximations of the likelihood. They compare a Gaussian kernel version

$$\frac{1}{S}\sum_{s=1}^{S}\varphi(y^{\text{obs}}-x_s(\theta);\varepsilon^2)$$

with the synthetic Gaussian likelihood version of Wood (2010)

$$\varphi(y^{\text{obs}}-\mu(\theta);\sigma(\theta)^2+\varepsilon^2)$$

where both mean and variance are estimated from the simulated data. If ε is taken as an external quantity and driven to zero, the second approach is much more stable. But ε is never driven to zero in ABC, or fixed at $\varepsilon = 0.37$[8]: It is instead considered as a kernel bandwidth and hence estimated from the simulated data. Hence ε is commensurable with $\sigma(\theta)$. And this makes me wonder at the relevance of the conclusion that synthetic is better than kernel for Hamiltonian ABC. More globally, I

[8] In a personnal comment, the authors explain that this value was chosen as a small fraction of the simulator noise at the MAP θ.

wonder at the relevance of better simulating from a still approximate target when the true goal is to better approximate the genuine posterior.

Some of the paper covers separate issues like handling gradient by a fast algorithm à la Spall (2003) and incorporating the random generator as part of the Markov chain. And using S common random numbers in computing the gradients for all values of θ. (Although I am not certain all random generators can be represented as a deterministic transform of a parameter θ and of a fixed number of random uniforms. But the authors may consider a random number of random uniforms when they represent their random generators as deterministic transform of a parameter θ and of the random seed. I am also uncertain about the distinction between common, sticky, and persistent random numbers!) As pointed out by the authors in their comments on xianblog.wordpress.com, they bypass finite differences, Hessian computations, and only require a small number of simulations, which may make this approach of considerable interest if explored further in high dimensional settings

4.9 ABC via population annealing

> "We are recommended to try a number of annealing schedules to check the influence of the schedules on the simulated data (...) As a whole, the simulations with the posterior parameter ensemble could, not only reproduce the data used for parameter inference, but also capture and predict the data which was not used for parameter inference." Y. Murakami, 2014

Population annealing is a notion introduced by Iba, the very same Iba (2000) who introduced the notion of population Monte Carlo that we studied in subsequent papers (Cappé et al., 2004, Douc et al., 2007, Cappé et al., 2008) . It reproduces the setting found in many particle filter papers of a sequence of (annealed or rather tempered) targets ranging from an easy (i.e., almost flat) target to the genuine target, and of an update of a particle set by MCMC moves and reweighing. I actually have trouble perceiving the difference with other sequential Monte Carlo schemes as those exposed in Del Moral et al. (2006). And the same is true of the ABC extension covered in Murakami (2014), where the annealed intermediate targets correspond to larger tolerances. This sounds like a traditional ABC-SMC algorithm. Without the adaptive scheme on the tolerance ε found e.g. in Del Moral et al. (2006), since the sequence is set in advance. [However, the discussion about the implementation includes the above quote that suggests a vague form of cross-validated tolerance construction]. The approximation of the marginal likelihood also sounds standard, the marginal being approximated by the proportion of accepted pseudo-samples. Or more exactly by the sum of the SMC weights at the end of the annealing simulation. This actually raises several questions: (a) this estimator is always between 0 and 1, while the marginal likelihood is not restricted [but this is due to a missing $1/\varepsilon$ in the likelihood estimate that cancels from both numerator and denominator]; (b) seeing the kernel as a non-parametric estimate of the likelihood led me to wonder why different ε could not be used in different models, in that the pseudo-data used

for each model under comparison differs. If we were in a genuine non-parametric setting the bandwidth would be derived from the pseudo-data.

"Thus, Bayesian model selection by population annealing is valid." Y. Murakami, 2014

The discussion about the use of ABC population annealing somewhat misses the point of using ABC, which is to approximate the genuine posterior distribution, to wit the above quote: that the ABC Bayes factors favour the correct model in the simulation does not tell anything about the degree of approximation wrt the original Bayes factor. [The issue of non-consistent Bayes factors does not apply here as there is no summary statistic applied to the few observations in the data.] Further, the magnitude of the variability of the values of this Bayes factor as ε varies, from 1.3 to 9.6, mostly indicates that the numerical value is difficult to trust. (I also fail to explain the huge jump in Monte Carlo variability from 0.09 to 1.17 in Table 1.) That this form of ABC-SMC improves upon the basic ABC rejection approach is clear. However it needs to build some self-control to avoid arbitrary calibration steps and reduce the instability of the final estimates.

"The weighting function is set to be large value when the observed data and the simulated data are "close", small value when they are "distant", and constant when they are "equal"." Y. Murakami, 2014

The above quote is somewhat surprising as the estimated likelihood $f(x^{\text{obs}}|x^{\text{obs}},\theta)$ is naturally constant when $x^{\text{obs}} = x^{\text{sim}}$. I also failed to understand how the model intervened in the indicator function used as a default ABC kernel.

4.10 Integrated Likelihood via ABC

Grazian and Liseo (2014) rely on ABC for marginal density estimation. The idea in the paper is to produce an *integrated likelihood* approximation in intractable problems via the ratio

$$L(\psi|x) \propto \frac{\pi(\psi|x)}{\pi(\psi)}$$

both terms in the ratio being estimated from simulations,

$$\hat{L}(\psi|x) \propto \frac{\hat{\pi}^{\text{ABC}}(\psi|x)}{\hat{\pi}(\psi)}$$

(with possible closed form for the denominator). Although most of the examples processed in the paper (Poisson means ratio, Neyman–Scott's problem, g-&-k quantile distribution (Allingham et al., 2009), semi-parametric regression) rely on summary statistics, hence *de facto* replacing the numerator above with a pseudo-posterior conditional on those summaries, the approximation remains accurate (for those examples). In the g-&-k quantile example, Grazian and Liseo (2014) compare our ABC-MCMC algorithm with the one of Allingham et al. (2009): the later does better

by not replicating values in the Markov chain but instead proposing a new value until it is accepted by the usual Metropolis step. (An amazing feature is that both approaches are simultaneously correct!) As noted by the authors, "the main drawback of the present approach is that it requires the use of proper priors", unless the marginalisation of the prior can be done analytically. (This opens an interesting computational question: how can one provide an efficient approximation to a marginal density of a σ-finite measure, assuming this density exists.)

4.11 ABC for MRFs

Everitt (2014) provides a fairly interesting comparison of ABC and MCMC algorithms applied to the cases of MRFs observed with MRF errors (latent MRF models) and of exponential random graphs with errors as those used for social network modelling. The MCMC algorithm is a combination of SMC, of particle MCMC à la Andrieu et al. (2011) and of the exchange algorithm of Murray et al. (2006) that improves upon the single auxiliary variable method of Møller et al. (2006), which can also be reinterpreted à la Andrieu and Roberts (2009). Recall that the exchange algorithm provides a direct evaluation of the ratio of the normalising constants based on a running pair of parameters (hence the possible "exchange"). The issue of simulating exactly from an MRF is bypassed by validating an MCMC algorithm based on a finite number of iterations (under strong conditions). The SMC sampler for MRFs mixes hot coupling (based on a clique completion of a spanning tree of the true graph) and tempering. The ABC algorithm uses the same approach as ours (in Grelaud et al. (2009)) through the summary (sufficient!) statistics, plus the ABC-SMC sampler of Del Moral et al. (2012). The comparison is run on a small 10×10 Ising model and on the Florentine family network Yves Atchadé used in our Wang-Landau paper (Atchadé et al., 2013).

Now, comparing ABC with MCMC is not a thing that would come naturally to me and my answer to the question about their relative merits is to say that one only uses ABC when MCMC cannot work. This study shows a bit more depth in the analysis: First, ABC managed to pick the major features of the posterior in both cases, while a regular MCMC got either stuck in one region or fairly inefficient. Second, the involved fusion algorithm constructed by Richard managed to overcome those difficulties and provided a richer sample than ABC in the same number of runs (as it should, ABC being a slow learner.)

4.12 ABC for copula estimation

Clara Grazian and Brunero Liseo (di Roma) have just arXived a note on a method merging copulas, ABC, and empirical likelihood. The approach is rather hybrid and thus not completely Bayesian, but this must be seen as a consequence of an ill-

posed problem. Indeed, as in many econometric models, the model there is not fully defined: the marginals of iid observations are represented as being from well-known parametric families (and are thus well-estimated by Bayesian tools), while the joint distribution remains uncertain and hence so does the associated copula. The approach in the paper is to proceed stepwise, i.e., to estimate correctly each marginal, well correctly enough to transform the data by an estimated cdf, and then only to estimate the copula or some aspect of it based on this transformed data. Like Spearman's ρ. For which an empirical likelihood is computed and aggregated to a prior to make a BCel weight. (If this sounds unclear, each BEel evaluation is based on a random draw from the posterior samples, which transfers some uncertainty in the parameter evaluation into the copula domain. Thanks to Brunero and Clara for clarifying this point for me!)

At this stage of the note, there are two illustrations revolving around Spearman's ρ. One on simulated data, with better performances than a nonparametric frequentist solution. And another one on a Garch (1,1) model for two financial time-series.

I am quite glad to see an application of our BCel approach in another domain although I feel a tiny bit uncertain about the degree of arbitrariness in the approach, from the estimated cdf transforms of the marginals to the choice of the moment equations identifying the parameter of interest like Spearman's ρ. Especially if one uses a parametric copula which moments are equally well-known. While I see the practical gain in analysing each component separately, the object created by the estimated cdf transforms may have a very different correlation structure from the true cdf transforms. Maybe there exist consistency conditions on the estimated cdfs... Maybe other notions of orthogonality or independence could be brought into the picture to validate further the two-step solution...

4.13 ABC for Bivariate Betas

Crackel and Flegal (2014) is running ABC for inference on the parameters of two families of bivariate betas. I however wonder why ABC was that necessary to handle the model. The said bivariate betas are defined from

$$V_1 = (U_1+U_5+U_7)/(U_3+U_6+U_8)\,,$$
$$V_2 = (U_2+U_5+U_8)/(U_4+U_6+U_7)$$

when

$$U_i \sim \mathcal{G}a(\delta_i,1)$$

and

$$X_1 = V_1/(1+V_1)\,,\; X_2 = V_2/(1+V_2)$$

This makes each term in the pair Beta and both components dependent. This construct was proposed by Arnold and Ng (2011). (The five-parameter version cancels the gammas for i=3,4,5.)

Since the pdf of the joint distribution is not available in closed form, Crackel and Flegal (2014) zoom on ABC-MCMC as the method of choice and discuss simulation experiments. (The choice of the tolerance ε as an absolute rather than relative value, ε=0.2,0.0.6,0.8, puzzles me, esp. since the distance between the summary statistics is not scaled.) I however wonder why other approaches are impossible. (Or why it is necessary to use this distribution to model correlated betas. Unless I am confused copulas were invented to this effect.) First, this is a latent variable model, so latent variables could be introduced inside an MCMC scheme. A wee bit costly but feasible. Second, several moments of those distributions are known so an empirical likelihood approach could be considered.

4.14 Transdimensional ABC

Transdimensional ABC including inference on invasive species models is the theme of Chkrebtii et al. (2013). It attracted my attention for at least two reasons: (a) it brings a new perspective on Bayesian inference in varying dimension models (or in multiple models and model comparison); (b) the application is about invasive species, as in our ABC paper on tracing pathways for the Asian beetle invasion in Europe.

After reading the paper, I however remain unconvinced that a direct duplication of Green's reversible jump MCMC algorithm (Green, 1995) is relevant in this ABC setting: this is indeed the central idea of the authors, namely to apply the reversible jump construct in an ABC-MCMC algorithm, with exactly the same validation as the usual ABC-MCMC algorithm where the indicator of a small enough distance between the observed and the simulated data acts as a (biased) estimator of the likelihood function. There is thus no doubt about the validity of the method. What leaves me somehow lukewarm about this idea is the same feature in the accelerated ABC paper by Picchini and Lyng Forman (2013), that is, the fact that the acceptance step actually occurs at the prior level, the Metropolis-Hastings acceptance probability being the ratio of the priors over the ratio of the proposals. (Plus a second acceptance step induced by the distance between the observed and the simulated data.)

The application to the invasion of European earthworms in northern Alberta is quite interesting, from the fact that those worms did not crawl their way up there but instead hitch-hiked!, to the modelling of the number of introductions by a Poisson spatial process, to the fact that the ABC algorithm can run with infinite precision! This last point makes me wonder whether or not a regular MCMC algorithm is unattainable for this problem. (However, the authors rely on a two-dimensional summary statistic for each pair (g,r), which helps in picking an ε equal to zero.)

4.15 ABC for Design

Hainy et al. (2013) relies on ABC to the construction of optimal designs. Remember that Müller (1999) uses a natural simulated annealing that is quite similar to our MAP [SAME] algorithm (Doucet et al., 2002), relying on multiple versions of the data set to get to the maximum. The paper also builds upon our 2006 paper (Amzal et al., 2006), that took advantage of the then emerging particle methods to improve upon a static horizon target. While our method is sequential in that it pursues a moving target, it does not rely on the generic methodology developed by Del Moral et al. (2006), where a backward kernel brings more stability to the moves. The paper also implements a version of our population Monte Carlo ABC algorithm (Beaumont et al., 2009), as a first step before an MCMC simulation. Overall, the paper sounds more like a review than like a strongly directive entry into ABC based design in that it remains quite generic. I somewhat doubt a realistic implementation (as opposed to the linear model used in the paper) could do without a certain amount of calibration.

4.16 Interacting Particles ABC

Albert et al. (2014) provides a new perspective on ABC. It relates to ABC-MCMC and to ABC-SMC in different ways, but the major point is to propose a sequential schedule for decreasing the tolerance that ensures convergence. Although there exist other proofs of convergence in the literature, this one is quite novel in that it connects ABC with the cooling schedules of simulated annealing. (The fact that the sample size does not appear as in Fearnhead and Prangle (2012) and their non-parametric perspective can be deemed less practical, but I think this is simply another perspective on the problem) The corresponding ABC algorithm is a mix of MCMC and SMC in that it lets a population of N particles evolve in a quasi-independent manner, the population being only used to update the parameters of the independent (normal) proposal and those of the cooling tolerance. Each particle in the population moves according to a Metropolis-Hastings step, but this is not an ABC-MCMC scheme in that the algorithm works with a population at all times, and this is not an ABC-SMC scheme in that there is no weighting and no resampling.

4.17 Preprocessing ABC

Moores et al. (2014) proposes to cut down on the cost of running an ABC experiment by removing the simulation of a humongous state-space vector, as in Potts and hidden Potts model, and replacing it by an approximate simulation of the 1-d sufficient (summary) statistics. In that case, we used a division of the 1-d parameter interval to simulate the distribution of the sufficient statistic for each of those parameter values and to compute the expectation and variance of the sufficient statistic. Then

the conditional distribution of the sufficient statistic is approximated by a Gaussian with these two parameters. And those Gaussian approximations substitute for the true distributions within an ABC-SMC algorithm à la Del Moral et al. (2012).

Across 20 125 $\times$ 125 pixels simulated images, Matt Moores' algorithm took an average of 21 minutes per image for between 39 and 70 SMC iterations, while resorting to pseudo-data and deriving the genuine sufficient statistic took an average of 46.5 hours for 44 to 85 SMC iterations. On a realistic Landsat image, with a total of 978,380 pixels, the precomputation of the mapping function took 50 minutes, while the total CPU time on 16 parallel threads was 10 hours 38 minutes. By comparison, it took 97 hours for 10,000 MCMC iterations on this image, with a poor effective sample size of 390 values. Regular SMC-ABC algorithms cannot handle this scale: It takes 89 hours to perform a single SMC iteration! (Note that path sampling also operates in this framework, thanks to the same precomputation: in that case it took 2.5 hours for 10^5 iterations, with an effective sample size of 10^4.)

4.18 Lazy Version

""A more automated approach would be useful for lazy versions of ABC SMC algorithms."
D. Prangle, 2014

Prangle (2014) is based upon the notion of cutting down massively on the generation of pseudo-samples that are "too far" from the observed sample. This is formalised through a stopping rule that puts the estimated likelihood to zero with a probability $1-\alpha(\theta,x)$ and otherwise divide the original ABC estimate by $\alpha(\theta,x)$. Which makes the modification unbiased when compared with basic ABC. The efficiency appears when $\alpha(\theta,x)$ can be computed much faster than producing the entire pseudo-sample and its distance to the observed sample. When considering an approximation to the asymptotic variance of this modification, Prangle derives a optimal (in the sense of the effective sample size) if formal version of the acceptance probability $\alpha(\theta,x)$, conditional on the choice of a "decision statistic" $\varphi(\theta,x)$. And of an importance function $g(\theta)$. This approach requires to estimate

$$\mathbb{P}(d(S(Y),S(y^o)) < \varepsilon|\varphi)$$

as a function of φ: I would have thought (non-parametric) logistic regression a good candidate towards this estimation, but Dennis is rather critical of this solution.

I added the quote above as I find it somewhat ironical: at this stage, to enjoy laziness, the algorithm has first to go through a massive calibration stage, from the selection of the subsample [to be simulated before computing the acceptance probability $\alpha(\theta,x)$] to the construction of the (somewhat mysterious) decision statistic $\varphi(\theta,x)$ to the estimation of the terms composing the optimal $\alpha(\theta,x)$. The most natural choice of $\varphi(\theta,x)$ seems to be involving subsampling, still with a wide range of possibilities and ensuing efficiencies. (The choice found in the application is somehow anticlimactic in this respect.)

A slight point of perplexity about this "lazy" proposal, namely the static role of ε, which is impractical because not set in stone. As stressed many times, the tolerance is a function of many factors including all the calibration parameters of the lazy ABC, rather than an absolute quantity. It seems to me that playing with a large collection of tolerances may be too costly in this setting.

4.19 ABC vs. EP

> "It seems quite absurd to reject an EP-based approach, if the only alternative is an ABC approach based on summary statistics, which introduces a bias which seems both larger (according to our numerical examples) and more arbitrary, in the sense that in real-world applications one has little intuition and even less mathematical guidance on to why $p(\theta|s(y))$ should be close to $p(\theta|y)$ for a given set of summary statistics s." S. Barthelmé and N. Chopin, 2014

Barthelmé and Chopin (2014) is selling expectation-propagation as quick and dirty version of ABC, avoiding the selection of summary statistics by using the constraint

$$||y_i - y_i^\star|| \leq \varepsilon$$

on each component of the simulated pseudo-data vector y* being the actual data. Expectation-propagation is a variational technique and it consists in replacing the target with the "closest" member from an exponential family, like the Gaussian distribution. The expectation-propagation approximation is found by including a single "observation" at a time, using the other approximations for the prior, and finding the best Gaussian in this pseudo-model. In addition, expectation-propagation provides an approximation of the evidence. In the "likelihood-free" setting, this means computing empirical mean and empirical variance, one observation at a time, under the above tolerance constraint.

Unless I am confused, the expectation-propagation approximation to the posterior distribution is a [sequentially updated] Gaussian distribution, which means that it will only be appropriate in cases where the posterior distribution is approximately Gaussian. Since the three examples processed in the paper are of this kind, e.g. the above reproduction, I wonder at the performances of the expectation-propagation method in less smooth cases, such as ridge-like or multimodal posteriors. The authors mention two limitations: "First, it [EP] assumes a Gaussian prior; and second, it relies on a particular factorisation of the likelihood, which makes it possible to simulate sequentially the data-points", but those seem negligible. I thus remain unconvinced by the concluding sentence quoted above. (The current approach to ABC is to consider $p(\theta|s(y))$ as a target per se, not as an approximation to $p(\theta|y)$.) Nonetheless, expectation-propagation constitutes a quick approximation method that can always used as a reference against other approximations.

4.20 Data-cloning ABC

> "By accepting of having obtained a poor approximation to the posterior, except for the location of its main mode, we switch to maximum likelihood estimation." U. Picchini, 2015

Picchini (2015) is merging ABC with *prior feedback* (Robert and Soubiran, 1993, rechristened *data cloning* in Lele et al. (2007), where a maximum likelihood estimate is produced by an ABC-MCMC algorithm, in the case of state-space models. This relates to an earlier paper by Rubio and Johansen (2013), who also suggested using ABC to approximate the maximum likelihood estimate. Here, the idea is to use an increasing number of replicates of the latent variables, as in our SAME algorithm, to spike the posterior around the maximum of the (observed) likelihood. An ABC version of this posterior returns a mean value as an approximate maximum likelihood estimate.

> "This is a so-called "likelihood-free" approach [Sisson and Fan, 2011], meaning that knowledge of the complete expression for the likelihood function is not required." U. Picchini, 2015

The above remark is sort of inappropriate in that it applies to a non-ABC setting where the latent variables are simulated from the exact marginal distributions, that is, unconditional on the data, and hence their density cancels in the Metropolis-Hastings ratio. This pre-dates ABC by a few years, since this was an early version of particle filter.

> "In this work we are explicitly avoiding the most typical usage of ABC, where the posterior is conditional on summary statistics of data $S(y)$, rather than y." U. Picchini, 2015

Another point I find rather negative is the abve in that, for state-space models, using the entire time-series as a "summary statistic" is unlikely to produce a good approximation.

The discussion on the respective choices of the ABC tolerance δ and on the prior feedback number of copies K is quite interesting, in that Umberto Picchini suggests setting δ first before increasing the number of copies. However, since the posterior gets more and more peaked as K increases, the consequences on the acceptance rate of the related ABC algorithm are unclear. Another interesting feature is that the underlying MCMC proposal on the parameter θ is an independent proposal, tuned during the warm-up stage of the algorithm. Since the tuning is repeated at each temperature, there are some loose ends as to whether or not it is a genuine Markov chain method (unless, as pointed by the author in comments on xianblog.wordpress.com, the adaptation is only done over a long burn-in). The same question arises when considering that additional past replicas need to be simulated when K increases. (Although they can be considered as virtual components of a vector made of an infinite number of replicas, to be used when needed.)

The simulation study involves a regular regression with 101 observations, a stochastic Gompertz model studied by Donnet et al. (2010) with 12 points and a simple Markov model, again with 12 points. While the ABC-DC solutions are close enough to the true MLEs whenever available, a comparison with the cheaper ABC Bayes estimates would have been of interest as well.

5 Summary Statistics, the ABC Conundrum

The main focus of the recent ABC literature has been on the selection and evaluation of summary statistics, including a Royal Statistical Society Read Paper (Fearnhead and Prangle, 2012) that set a reference and gave prospective developments in the discussion section. Reducing the data into a small dimension but sufficienlt informative statistics constitutes a fundamental difficulty when there is no non-trivial sufficient statistic and when the summary statistics are not already provided by the software (like DIYABC, Cornuet et al. (2008)) or imposed by experimenters in the field. This choice has to balance a loss of statistical information a gain in ABC precision, with little available on the amounts of error and information loss involved in the ABC substitution.

5.1 The Read Paper

Fearnhead and Prangle (2012) proposed an original approach to ABC, where ABC is considered from a purely inferential viewpoint and calibrated for estimation purposes. Fearnhead and Prangle do not follow the "traditional" perspective of looking at ABC as a converging approximation to the true posterior density. As Wilkinson (2013) (first posted in 2008), they take instead a randomised/noisy version of the summary statistics and derive a calibrated version of ABC, i.e. an algorithm that gives proper predictions, the drawback being that it is for the posterior given this randomised version of the summary statistics. The paper also contains an important result in the form of a consistency theorem that shows that noisy ABC is a convergent estimation method when the number of observations or datasets grows to infinity. The most interesting aspect in this switch of perspective is that the kernel $\mathfrak{h}$ used in the acceptance probability

$$\mathfrak{h}((s - s_{\text{obs}})/h)$$

does not have to act as an estimate of the true sampling density, since it appears in the (randomised) pseudo-model. (Everything collapses to the true model when the bandwidth h goes to zero.) The Monte Carlo error is taken into account through the average acceptance probability, which collapses to zero when h goes to zero, therefore a suboptimal choice!

A form of tautology stems from the comparison of ABC posteriors via a loss function

$$(\theta_0 - \hat{\theta})^{\mathrm{T}} A (\theta_0 - \hat{\theta})$$

that ends up with the "best" asymptotic summary statistic being

$$\mathbb{E}[\theta | y_{\text{obs}}].$$

This result indeed follows from the very choice of the loss function rather than from an intrinsic criterion. Using the posterior expectation as the summary statistics

still makes sense, especially when the calibration constraint implies that the ABC approximation has the same posterior mean as the true (randomised) posterior. Unfortunately this result is parameterisation dependent and unlikely to be available in settings where ABC is necessary. In the semi-automatic implementation proposed by Fearnhead and Prangle (2012), the authors suggest to use a pilot run of ABC to approximate the above statistics. I wonder at the resulting cost since a simulation experiment must be repeated for each simulated dataset (or sufficient statistic). The simplification in the paper follows from a linear regression on the parameters, thus linking the approach with Beaumont et al. (2002).

Using the same evaluation via a posterior loss, the authors show that the "optimal" kernel is uniform over a region

$$x^{\mathrm{T}}Ax < c$$

where c makes a ball of volume 1. A significant remark is that the error evaluated by Fearnhead and Prangle is

$$\mathrm{tr}(A\Sigma) + h^2\mathbb{E}_{\mathfrak{h}}[x^{\mathrm{T}}Ax] + \frac{C_0}{h^d}$$

which means that, due to the Monte Carlo error, the "optimal" value of h is not zero but akin to a non-parametric optimal speed in $2/2+d$. There should thus be a way to link this decision-theoretic approach with the one of Ratmann et al. (2009) since the latter take h to be part of the parameter vector.

As exposed in my discussion (Robert, 2012), I remain skeptical about the "optimality" resulting from the choice of summary statistics in the paper, partly because practice shows that proper approximation to genuine posterior distributions stems from using a (much) larger number of summary statistics than the dimension of the parameter (albeit unachievable at a given computing cost), partly because the validity of the approximation to the optimal summary statistics depends on the quality of the pilot run, and partly because there are some imprecisions in the mathematical derivation of the results (Robert, 2012). Furthermore, important inferential issues like model choice are not covered by this approach. But, nonetheless, the paper provides a way to construct default summary statistics that should come as a supplement to summary statistics provided by the experts, or even as a substitute.

The paper is also connecting to the computing cost and stressing the relevance of the indirect inference literature (Gouriéroux et al., 1993). A clear strength of the paper remains with Section 4 which provides a major simulation experiment. My only criticism on this section would be about the absence of a phylogeny example that would relate to the models that launched ABC methods. This is less of a mainstream statistics example, but it would be highly convincing to those primary users of ABC.

5.2 Another Review

> "What is very apparent from this study is that there is no single 'best' method of dimension reduction for ABC." M. Blum, M. Nunes, D. Prangle, and S. Sisson, 2012

Blum et al. (2013) offers a detailed review of dimension reduction methods in ABC, along with a comparison on three specific models. Given that, as put above, the choice of the vector of summary statistics is presumably the most important single step in an ABC algorithm and keeping in mind that selecting too large a vector is bound to fall victim of the dimension curse, this constitutes a reference for the ABC literature. Therein, the authors compare regression adjustments à la Beaumont et al. (2002), subset selection methods, as in Joyce and Marjoram (2008), and projection techniques, as in Fearnhead and Prangle (2012). They add to this impressive battery of methods the potential use of AIC and BIC. An argument for using AIC/BIC is that either provides indirect information about the approximation of $p(\theta|y)$ by $p(\theta|s(y))$, even though this does not seem obvious to me.

The paper also suggests a further regularisation of Beaumont et al. (2002) by ridge regression, although L_1 penalty à la Lasso would be more appropriate in my opinion for removing extraneous summary statistics. (I must acknowledge never being a big fan of ridge regression, esp. in the ad hoc version à la Hoerl and Kennard (1970), i.e. in a non-decision theoretic approach where the hyperparameter λ is derived from the data by cross-validation, since it then sounds like a poor man's version of Bayes' and Stein' estimators, just like BIC is a first order approximation to regular Bayes factors). Unsurprisingly, ridge regression does better than plain regression in the comparison experiment when there are many almost collinear summary statistics, but an alternative conclusion could be that regression analysis is not that appropriate with many summary statistics. Indeed, summary statistics are not quantities of interest but data summarising tools towards a better approximation of the posterior at a given computational cost. (I do not get the final comment about the relevance of summary statistics for MCMC or SMC algorithms: the criterion should be the best approximation of $p(\theta|y)$ which does not depend on the type of algorithm.)

5.3 Accurate ABC

Ratmann et al. (2013) introduced the notion of accurate ABC. The central idea is that, if the distribution of the summary statistics is known and if replicas of those summary statistics are available for the true data (and less problematically for the generated data), then a classical statistical test can be turned into a natural distance measure for each statistics and even "natural" bounds can be found on that distance, to the point of recovering most properties of the original posterior distribution... A first worry is this notion that the statistical distribution of a collection of summary statistics is available in closed form: this sounds unrealistic even though it may not constitute a major contention issue. Indeed, replacing a tailored test with a distribution-free test of identical location parameter could not hurt that much. The paper also insists on

sufficiency, which I fear is a lost cause. In my current understanding of ABC, the loss of some amount of information contained in the data should be acknowledged and given a write-off as a Big Data casualty. (See, e.g., Lemma 1.)

Another worry is that the rephrasing of the acceptance distance as the maximal difference for a particular test relies on an elaborate calibration, incl. α, c+, τ+, &tc. (I am not particularly convinced by the calibration in terms of the power of the test being maximised at the point null value.) When cumulating tests and aiming at a nominal α level, the orthogonality of the test statistics in Theorem 1(iii) is puzzling and I think unrealistic.

The notion of accuracy that is central to the paper and its title corresponds to the power of every test being maximal at the true value of the parameter. And somehow to the ABC approximation being maximises at the true parameter, even though I am lost by then [i.e. around eqn (18)] about the meaning of ρ*... The major result in the paper is however that, under the collection of assumptions made therein, the ABC MLE and MAP versions are equal to their exact counterparts. And that these versions are also unbiased. This Theorem 3 sounds fantastic but makes me uneasy: unbiasedness is a sparse property that is rarely found in statistical problems. Change the parameterisation and you loose unbiasedness. And even the possibility to find an unbiased estimator. Since this difficulty does not appear in the paper, I would conclude that either the assumptions are quite constraining or the result holds in a weaker sense... (Witness the use of "essentially unbiased" in Fig. 4.)

The paper seems to imply that the summary statistics are observed repeatedly over the true sample. Unless $n = 1$, this does not seem realistic. (I do not understand everything in Example 1, in particular the complaint that the ABC solutions were biased for finite values of n. That sounds like an odd criticism when applied to Bayesian estimators. Now, it seems the paper is very intent on achieving unbiasedness. So maybe it should be called the aAnsBC algorithm for "not-so-Bayes!) I am also puzzled by the distinction between summary values and summary statistics. This sounds like insisting on having a large enough iid dataset. Or the discussion that the summary parameters are replaced by estimates seems out of context because this adds an additional layer of notation to the existing summary "stuff"... With the additional difficulty that Lemma 1 assumes reparameterisation of the model in terms of those summary parameters. I also object to the point null hypotheses being written in terms of a point estimate, i.e. of a quantity depending on the data x: it sounds like confusing the test [procedure] with the test [problem]. Another example: I read several times Lemma 5 about the calibration of the number of ABC simulations m but cannot fathom what this m is calibrated against. It seems only a certain value of m achieves the accurate correspondence with the genuine posterior, which sounds counter-intuitive.

5.4 ABC with Indirect Summary Statistics

After reading Drovandi et al. (2011), I checked the related Gleim and Pigorsch (2013) about indirect summary statistics. The setting is indeed quite similar to the above, with a description of three ways of connecting indirect inference with ABC, albeit with a different range of illustrations. This preprint states most clearly its assumption that the generating model is a particular case of the auxiliary model, which sounds anticlimactic since the auxiliary model is precisely used because the original one is mostly out of reach! This certainly was the original motivation for using indirect inference.

The part of the paper that I find the most intriguing is the argument that the indirect approach leads to sufficient summary statistics, in the sense that they "are sufficient for the parameters of the auxiliary model and (...) sufficiency carries over to the model of interest". Looking at the details in the Appendix, I found that the argument is lacking, because the likelihood as a functional is shown to be a (sufficient) statistic, which seems both a tautology and irrelevant because this is different from the likelihood considered at the (auxiliary) MLE, which is the summary statistic used in fine.

> "...we expand the square root of an innovation density h in a Hermite expansion and truncate the infinite polynomial at some integer K which, together with other tuning parameters of the SNP density, has to be determined through a model selection criterion (such as BIC). Now we take the leading term of the Hermite expansion to follow a Gaussian GARCH model." A. Gleim and C. Pigorsch, 2013

As in Drovandi et al. (2011), the performances of the ABC-I schemes are tested on a toy example, which is a very basic exponential iid sample with a conjugate prior and a gamma model as auxiliary. The authors use a standard ABC based on the first two moments as their benchmark, however they do not calibrate those moments in the distance and end up with poor performances of ABC (in a setting where there is a sufficient statistic!). The best choice in this experiment appears as the solution based on the score, but the variances of the distances are not included in the comparison tables. The second implementation considered in the paper is a rather daunting continuous-time non-Gaussian Ornstein-Uhlenbeck stochastic volatility model à la Barndorff-Nielsen and Shephard (2001). The construction of the semi-nonparametric (why not semi-parametric?) auxiliary model is quite involved as well, as illustrated by the quote above. The approach provides an answer, with posterior ABC-IS distributions on all parameters of the original model, which kindles the question of the validation of this answer in terms of the original posterior. Handling simultaneously several approximation processes would help in this regard.

5.5 ABC with Score Functions

Ruli et al. (2013) advocate the use of composite score functions for ABC. While the paper provides a survey of composite likelihood methods, the core idea of the paper

is to use the score function (of the composite likelihood) as the summary statistic,

$$\frac{\partial\, c\ell(\theta;y)}{\partial\,\theta},$$

when evaluated at the maximum composite likelihood at the observed data point. In the specific (but unrealistic) case of an exponential family, an ABC based on the score is asymptotically (i.e., as the tolerance ε goes to zero) exact. The choice of the composite likelihood thus induces a natural summary statistics and, as in our empirical likelihood paper, where we also use the score of a composite likelihood, the composite likelihoods that are available for computation are usually quite a few, thus leading to an automated choice of a summary statistic..

An interesting (common) feature in most examples found in this paper is that comparisons are made between ABC using the (truly) sufficient statistic and ABC based on the pairwise score function, which essentially relies on the very same statistics. So the difference, when there is a difference, pertains to the choice of a different combination of the summary statistics or, somehow equivalently to the choice of a different distance function. One of the examples starts from our MA(2) toy-example (Marin et al., 2011). The composite likelihood is then based on the consecutive triplet marginal densities. As shown in the paper, the composite version improves to some extent upon the original ABC solution using three autocorrelations. The overall difficulty with this ABC-cs proposal is that the composite likelihood need to be constructed afresh for every new problem. It thus requires some expertise from the user that precludes its implementation by practitionners from other fields, as was the case for the original ABC algorithm developped by population geneticists (albeit the original ABC algorithm does require the collection of enough summary statistics).

A suggestion I would have about a refinement of the proposed method deals with the distance utilised in the paper, namely the sum of the absolute differences between the statistics. Indeed, this sum is not scaled at all, neither for regular ABC nor for composite ABC, while the composite likelihood perspective provides in addition to the score a natural metric through the matrix $A(\theta)$. So I would suggest comparing the performances of the methods using instead this rescaling since, in my opinion and in contrast with a remark on page 13, it is relevant in some (many?) settings where the amount of information brought by the composite model widely varies from one parameter to the next.

In a related vein, our paper (Martin et al., 2014) offers a new perspective on ABC based on pseudo-scores. For one thing, it concentrates on the selection of summary statistics from a more econometrics than usual point of view, defining asymptotic sufficiency in this context and demonstrated that both asymptotic sufficiency and Bayes consistency can be achieved when using maximum likelihood estimators of the parameters of an auxiliary model as summary statistics. In addition, the proximity to (asymptotic) sufficiency yielded by the MLE is replicated by the score vector. Using the score instead of the MLE as a summary statistics allows for huge gains in terms of speed. The method is then applied to a continuous time state space model, using as auxiliary model an augmented unscented Kalman filter. We also found in

the various state space models tested therein that the ABC approach based on the marginal [likelihood] score was performing quite well, including wrt Fearnhead and Prangle (2012) approach. I strongly support the idea of using such a generic object as the unscented Kalman filter for state space models, even when it is not a particularly accurate representation of the true model. Another appealing feature of the paper is in the connections made with indirect inference.

6 ABC Model Choice

While ABC is a substitute for a proper—possibly MCMC based—Bayesian inference, and thus pertains to all aspects of Bayesian inference, including testing and model checking, the special issue of comparing models via ABC is highly delicate and concentrated most of the criticisms addressed against ABC (Templeton, 2008, 2010). The implementation of ABC model choice follows by treating the model index $\mathfrak{m}$ as an extra parameter with an associated prior, as detailed in the following algorithm:

Algorithm 6 ABC (model choice)

for $i = 1$ to N **do**
  repeat
    Generate $\mathfrak{m}$ from the prior $\pi(\mathscr{M} = m)$
    Generate $\theta_{\mathfrak{m}}$ from the prior $\pi_{\mathfrak{m}}(\theta_{\mathfrak{m}})$
    Generate $\boldsymbol{z}$ from the model $f_{\mathfrak{m}}(\boldsymbol{z}|\theta_{\mathfrak{m}})$
  until $\rho\{S(\boldsymbol{z}), S(\boldsymbol{y})\} \leq \varepsilon$
  Set $\mathfrak{m}^{(i)} = \mathfrak{m}$ and $\theta^{(i)} = \theta_{\mathfrak{m}}$
end for
return the values $\mathfrak{m}^{(i)}$ associated with the k smallest distances

Improvements upon returning raw model indexe frequencies as ABC estimates have been proposed in Fagundes et al. (2007), via a regression regularisation. In this approach, indices are processed as categorical variables in a formal multinomial regression, using for instance logistic regression. Rejection-based approaches as in Algorithm 6 were introduced in 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 versions are widely used 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), Wegmann and Excoffier (2010). As described in the following sections, this adoption may be premature or over-optimistic, since caution and cross-checking are necessary to completely validate the output.

6.1 ABC Model Criticism

Ratmann et al. (2009) is a very original approach to ABC model criticism and thus indirectly to ABC model choice. It is about the use of the ABC approximation error ε in an altogether different way, namely as a tool assessing the goodness of fit of a given model. The fundamental idea is to process ε as an additional parameter of the model, simulating from a joint posterior distribution

$$f(\theta, \varepsilon|x_0) \propto \xi(\varepsilon|x_0, \theta) \times \pi_\theta(\theta) \times \pi_\varepsilon(\varepsilon)$$

where x_0 is the data and $\xi(\varepsilon|x_0, \theta)$ plays the role of the likelihood. (The π's are obviously the priors on θ and ε.) In fact, $\xi(\varepsilon|x_0, \theta)$ is the prior predictive density of $\rho(S(x), S(x_0))$ given θ and x_0 when x is distributed from $f(x|\theta)$. The authors then derive an ABC algorithm they call ABCμ to simulate an MCMC chain targeting this joint distribution, replacing $\xi(\varepsilon|x_0, \theta)$ 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 of it being that this posterior should include 0 in a reasonable credible interval. (Contrary to other ABC papers, ε can be negative and multidimensional in this paper.)

As written above, Ratmann et al. (2009) is a very interesting paper, full of innovations, that should span new directions in the way one perceives ABC. It is also quite challenging, partly due to the frustrating constraints PNAS imposes on the organisation (and submission) of papers. The paper thus contains a rather sketchy main part, a Materials and Methods addendum, and a Supplementary Material file. Flipping back and forth between those files certainly does not improve reading. I have never understood why PNAS was is so rigid about a format that does not suit non-experimental sciences.

Given the wealth of innovations contained in the paper, let me add here that, while the authors stress they use the data once (a point always uncertain to me), 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 (ε, θ). The product being most often defined as a density in (ε, θ), it can be simulated from, but I have trouble seeing this as a regular Bayesian problem, especially because it seems the prior on ε significantly contributes to the final assessment (but is not particularly discussed in the paper, except in the §1.10 section).

Another Bayesian conundrum is the fact that both θ and ε are taken to be the same across models. In a sense, I presume θ can be completely different, but using the same prior on ε over all models under comparison is more of an issue. Further and better developped criticisms were published as Robert et al. (2010), along with a reply by the authors (Ratmann et al., 2010). Let me stress one more time how original this paper is and deplore a lack of follow-up in the literature for a practical method that should be implemented on existing ABC softwares.

6.2 A Clear Lack of Confidence

In Robert et al. (2011), we came to the conclusion, shocking to us, that ABC approximations to posterior probabilities cannot be uniformly trusted. Approximating posterior probabilities by an ABC algorithm, ie by using the frequencies of acceptances of simulations from those models (assuming the use of a common summary statistic to define the distance to the observations). Rather obviously (a posteriori!), we ended up with the limiting behaviour being ruled by a true Bayes factor, except it is the one based on the distributions of the summary statistics under both models.

At first, this did not sound a particularly novel and fundamental result, since all ABC approximations rely on the posterior distributions based on those summary statistics, rather than on the whole dataset. However, while this approximation only has consequences in terms of the precision of the inference for most inferential purposes, it induces a dramatic arbitrariness in the Bayes factor. To illustrate this arbitrariness, consider the case of using a sufficient statistic $S(x)$ for both models. Then, by the factorisation theorem, the true likelihoods factorise as

$$\ell_1(\theta_1|x) = g_1(x)p_1(\theta_1|S(x)) \quad \text{and} \quad \ell_2(\theta_2|x) = g_2(x)p_2(\theta_2|S(x))$$

resulting in a true Bayes factor equal to

$$B_{12}(x) = \frac{g_1(x)}{g_2(x)} B^S_{12}(x)$$

where the last term is the limiting ABC Bayes factor. Therefore, in the favourable case of the existence of a sufficient statistic, using only the sufficient statistic induces a difference in the result that fails to converge with the number of observations or simulations. On the opposite, it may diverge one way or another as the number of observations increases. Again, this is in the favourable case of sufficiency. In the realistic setting of using summary statistics, things deteriorate further! This practical situation indeed implies a wider loss of information compared with the exact inferential approach, hence a wider discrepancy between the exact Bayes factor and the quantity produced by an ABC approximation. It thus appeared to us as an urgent duty to warn the community about the dangers of this approximation, especially when considering the rapidly increasing number of applications using ABC for conducting model choice and hypothesis testing. Furthermore, we unfortunately did not see at the time an immediate and generic alternative for the approximation of Bayes factor.

The paper stresses what I consider a fundamental or even foundational distinction between ABC point (and confidence) estimation and ABC model choice, namely that the problem was at another level for Bayesian model choice (using posterior probabilities). When doing point estimation with in-sufficient summary statistics, the information content is poorer, but unless one uses very degraded summary statistics, inference is converging. The posterior distribution is still different from the true posterior in this case but, at least, gathering more observations brings more

information about the parameter (and convergence when the number of observations goes to infinity). For model choice, this is not guaranteed if we use summary statistics that are not inter-model sufficient, as shown by the Poisson and normal examples. Furthermore, except for very specific cases such as Gibbs random fields (Grelaud et al., 2009), it is almost always impossible to derive inter-model sufficient statistics, beyond the raw sample.

Example 10. Another example is described in the introduction of the "sequel" by Marin et al. (2014), to be discussed below. The setting is one of a comparison between a normal $\boldsymbol{y} \sim \mathcal{N}(\theta_1, 1)$ model and a double exponential $\boldsymbol{y} \sim \mathcal{L}(\theta_2, 1/\sqrt{2})$ model[9] The summary statistics used in the corresponding ABC algorithm are the sample mean, the sample median and the sample variance. Figure 12 exhibits the absence of discrimination between the two models, since the posterior probability of the normal model converges to a central value around 0.5-0.6 when the sample size grows, irrelevant of the true model behind the simulated datasets! ◀

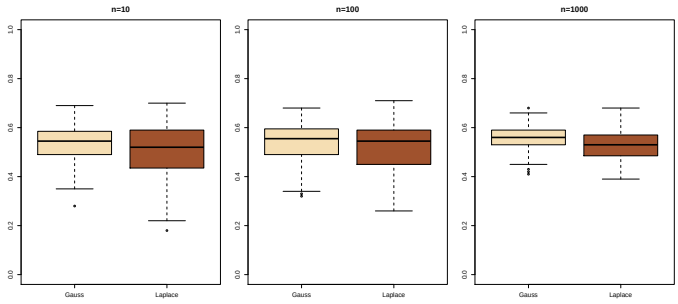

Fig. 12 Comparison of the range of the ABC posterior probability that data is from a normal model *(and not from a Laplace model)* with unknown mean θ when the data is made of $n = 10, 100, 1000$ observations *(left, center, right, resp.)* either from a Gaussian *(lighter)* or Laplace distribution *(darker)* and when the ABC summary statistic is made of the empirical mean, median, and variance. The ABC algorithm generates 10^4 simulations ($5,000$ for each model) from the prior $\theta \sim \mathcal{N}(0,4)$ and selects the tolerance ε as the 1% distance quantile over those simulations. *(Source: Marin et al. (2014).)*

The paper includes a realistic population genetic illustration, where two scenarios including three populations were compared, two populations 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)

[9] The double exponential distribution is also called the Laplace distribution, hence the notation $\mathcal{L}(\theta_2, 1/\sqrt{2})$, with mean θ_2 and variance one.

have been generated assuming an effective population size of 1000 and mutation rates of 0.0005. There are six 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]). Although this is rather costly in computing time, the posterior probability can nonetheless be estimated by importance sampling, based on 1000 parameter values and 1000 trees per parameter value, based on the modules of Stephens and Donnelly (2000). The ABC approximation is obtained from DIYABC (Cornuet et al., 2008), using a reference sample of two million parameters and 24 summary statistics. The result of this experiment is shown above, with a clear divergence in the numerical values despite stability in both approximations. Taking the importance sampling approximation as the reference value, the error rates in using the ABC approximation to choose between scenarios 1 and 2 are 14.5% and 12.5% (under scenarios 1 and 2), respectively. Although a simpler experiment with a single parameter and the same 24 summary statistics shows a reasonable agreement between both approximations, this result comes an additional support to our warning about a blind use of ABC for model selection. The corresponding simulation experiment was quite intense, as, with 50 markers and 100 individuals, the product likelihood suffers from such an enormous variability that 100,000 particles and 100 trees per locus have trouble to address (despite a huge computing cost of more than 12 days on a powerful cluster).

A quite related if less pessimistic paper is Didelot et al. (2011), also concerned with the limiting behaviour for the ratio,

$$B_{12}(x) = \frac{g_1(x)}{g_2(x)} B^S_{12}(x).$$

Indeed, the authors reach the opposite conclusion from ours, namely that the problem can be solved by a sufficiency argument. Their point is that, when comparing models within exponential families (which is the natural realm for sufficient statistics), it is always possible to build an encompassing model with a sufficient statistic that remains sufficient across models. This construction is correct from a mathematical perspective, as seen for instance in the Poisson versus geometric example we first mentioned in Grelaud et al. (2009): adding

$$\prod_{i=1}^{n} x_i!$$

to the sum of the observables into a large sufficient statistic produces a ratio g_1/g_2 that is equal to 1, hence avoids any discrepancy..

Nonetheless, we do not think this encompassing property has a direct impact on the performances of ABC model choice. In practice, complex models do not enjoy sufficient statistics (if only because the overwhelming majority of them are not exponential families, with the notable exception of Gibbs random fields where the above agreement graph is derived). There is therefore a strict loss of information in using ABC model choice, due to the call both to insufficient statistics and to non-zero tolerances. Looking at what happens in the limiting case when one is

relying on a common sufficient statistic is a formal study that brings light on the potentially huge discrepancy between the ABC-based Bayes factor and the true Bayes factor. This is why we consider that finding a solution in this formal case—while a valuable extension of the Gibbs random fields case—does not directly help towards the understanding of the discrepancy found in non-exponential complex models.

6.3 Validating Summaries for ABC Model Choice

Our answer to the (well-received) above warning is provided in Marin et al. (2014), which deals with the evaluation of summary statistics for Bayesian model choice. Even though the idea of separating the mean behaviour of the statistics under both model came rather early, establishing a complete theoretical framework that validated this intuition took quite a while and the assumptions changed a few times around the summer. The simulations associated with the paper were straightforward in that (a) the setup had been suggested to us by a referee (Robert et al., 2011): as detailed in Example 10, they consist in comparing normal and Laplace distributions with different summary statistics (inc. the *median absolute deviation*, which is the median of the absolute deviation from the median, $\text{med}(|\boldsymbol{x} - \text{med}(\boldsymbol{x})|)$), (b) the theoretical results told us what to look for, and (c) they did very clearly exhibit the consistency and inconsistency of the Bayes factor/posterior probability predicted by the theory. Both boxplots shown on Figures 12 and 13 exhibit this agreement: when using (empirical) mean, median, and variance to compare normal and Laplace models, the posterior probabilities do not select the "true" model but instead aggregate near a fixed value. hence ABC based on those summary statistics is not discriminative. When using instead the median absolute deviation as summary statistic, the posterior probabilities concentrate near one or zero depending on whether or not the normal model is the true model. Hence, this summary statistic is highly discriminant for the comparison of the two models. From an ABC perspective, this means that using the median absolute deviation is then satisfactory, as opposed to the above three statistics.

The above example illustrates very clearly the major result of this paper, namely that the mean behaviour of the summary statistic $S(\boldsymbol{y})$ under both models under comparison is fundamental for the convergence of the Bayes factor, i.e. of the Bayesian model choice based on $S(\boldsymbol{y})$. This result, described in the next section, thus brings an almost definitive answer to the question raised in Robert et al. (2011) about the validation of ABC model choice.

The main result in Marin et al. (2014) is that the mean behaviour of the summary statistic $S(\boldsymbol{y})$ under both models under comparison is fundamental for the convergence of the Bayes factor, i.e. of the Bayesian model choice based on $S(\boldsymbol{y})$. This work thus brings an almost definitive answer to the question raised in Robert et al. (2011) about the validation of ABC model choice.

More precisely, Marin et al. (2014) states that, under some "heavy-duty" Bayesian asymptotics assumptions, (a) if the "true" mean of the summary statistic can be

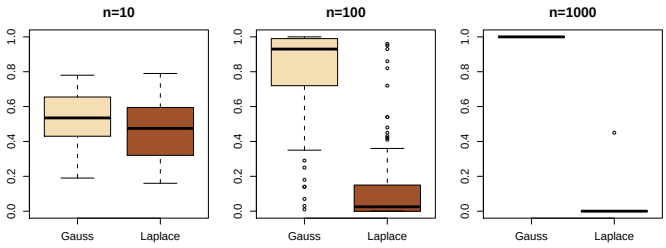

Fig. 13 Comparison of the distributions of the posterior probabilities that the data is from a normal model *(as opposed to a Laplace model)* with unknown mean θ when the data is made of $n = 10, 100, 1000$ observations *(left, center, right, resp.)* either from a Gaussian or Laplace distribution with mean equal to zero and when the summary statistic in the ABC algorithm is the median absolute deviation. The ABC algorithm uses a reference table of 10^4 simulations (5,000 for each model) from the prior $\theta \sim \mathcal{N}(0,4)$ and selects the tolerance ε as the 1% distance quantile over those simulations.

recovered for both models under comparison, then the Bayes factor is of order

$$\mathrm{O}\left(n^{-(d_1-d_2)/2}\right),$$

where d_i is the intrinsic dimension of the parameters driving the summary statistic in model $i = 1, 2$, irrespective of which model is true. (Precisely, the dimensions d_i are the dimensions of the asymptotic mean of the summary statistic under both models.) Therefore, the Bayes factor always asymptotically selects the model having the smallest effective dimension and cannot be consistent. (b) if, instead, the "true" mean of the summary statistic cannot be represented in the other model, then the Bayes factor is consistent. This means that, somehow, the best statistics to be used in an ABC approximation to a Bayes factor are ancillary statistics with different mean values under both models. Else, the summary statistic must have enough components to prohibit a parameter under the "wrong" model to meet the "true" mean of the summary statistic.

One of the referee's comments on the paper was that maybe Bayes factors were not appropriate for conducting model choice, thus making the whole derivation irrelevant. This is a possible perspective but it can be objected that Bayes factors and posterior probabilities are used in conjunction with ABC in dozens of genetic papers. Further arguments are provided in the various replies to both of Templeton's radical criticisms (Templeton, 2008, 2010). That more empirical and model-based assessments also are available is quite correct, as demonstrated in the multicriterion approach of Ratmann et al. (2009). This is simply another approach, not followed by most geneticists so far.

Another criticism was that the paper is quite theoretical and the mathematical assumptions required to obtain the convergence theorems are rather overwhelming. Meaning that in practical cases they cannot truly be checked. However, I think we can eventually address those concerns for two distinct reasons: first, the paper comes

as a third step in a series of papers where we first identified a sufficiency property, then realised that this property was actually quite a rare occurrence, and finally made a theoretical advance as to when is a summary statistic enough (i.e. "sufficient" in the standard sense of the term!) to conduct model choice, with a clear answer that the mean ranges of the summary statistic under each model could not intersect. Second, my own personal view is that those assumptions needed for convergence are not of the highest importance for statistical practice (even though they are needed in the paper!) and thus that, from a methodological point of view, only the conclusion should be taken into account. It is then rather straightforward to come up with (quick-and-dirty) simulation devices to check whether a summary statistic behaves differently under both models, taking advantage of the reference table already available (instead of having to run Monte Carlo experiments with ABC basis). The final version of the paper (Marin et al., 2014) includes a χ^2 check about the relevance of a given summary statistics.

At last, we could not answer in depth a query about the different speeds of convergence of the posterior probabilities under the Gaussian and Laplace distributions. This was a most interesting question in that the marginal likelihoods do indeed seem to converge at different speeds. However, the only precise information we can derive from our result (Theorem 1) is when the Bayes factor is not consistent. Otherwise, we only have a lower bound on its speed of convergence (under the correct model). Getting precise speeds in this case sounds beyond our reach...

6.4 Sufficient and Insufficient Statistics

Barnes et al. (2012) also consider the selection of sufficient statistics towards ABC model choice. It builds on our earlier warning (Robert et al., 2011) about (unfounded) ABC model selection to propose a selection of summary statistics that partly alleviates the original problem. (The part about the discrepancy with the true posterior probability remains to be addressed. As does the issue of whether or not the selected collection of statistics provides a convergent model choice inference, solved in Marin et al. (2014).) Their section "Resuscitating ABC model choice" states quite clearly the goal of the paper:

– "this [use of inadequate summary statistics] mirrors problems that can also be observed in the parameter estimation context,
– for many important, and arguably the most important applications of ABC, this problem can in principle be avoided by using the whole data rather than summary statistics,
– in cases where summary statistics are required, we argue that we can construct approximately sufficient statistics in a disciplined manner,
– when all else fails, a change in perspective, allows us to nevertheless make use of the flexibility of the ABC framework."

The driving idea in the paper is to use an entropy approximation to measure the lack of information due to the use of a given set of summary statistics. The corresponding algorithm then proceeds from a starting pool of summary statistics to build sequentially a collection of the most informative summary statistics (which, in a sense, reminded me of a variable selection procedure based on Kullback-Leibler, we developed with Costas Goutis and Jérôme Dupuis). It is a very interesting advance in the issue of ABC model selection, even though it cannot eliminate all stumbling blocks. The interpretation that ABC should be processed as an inferential method on its own rather than an approximation to Bayesian inference is clearly appealing.

While the information theoretic motivation is attractive, I do not see [as a Bayesian?] the point of integrating over the data space (Result 1 and 2) since the expectation should be only against the parameter and not against the data. If $S = S(X)$ is sufficient, then almost surely, the posterior given $X = x$ is the same as the posterior given $S(x) = s(x)$. Checking for the expectation in X of the log divergence between both posteriors to be zero is unnecessary. So, in the end, this makes me wonder whether (mutual) information theory is the right approach to the problem... Or rather to motivate the use of the Kullback-Leibler divergence, as I fully support the use of this measure of divergence! Also, what is the exact representation used in the paper for computing the Kullback-Leibler divergence KL and for evaluating the posterior densities from an ABC output in the log divergence?

Of course, and as clearly stated in the paper, the whole method relies on the assumption that there is a reference collection of summary statistics that is somehow sufficient. Which is rather unlikely in most realistic settings (this is noted in the discussion of Fearnhead and Prangle (2012) as well as in Robert et al. (2011)). So the term sufficient should not be used as in Figure 3 for instance. Overall, the method of statistic selection [approximately] provides the subset of the reference collection with the same information content as the whole collection. So, its main impact is to exclude irrelevant summary statistics from a given collection. Which is already a very interesting outcome. What would be even more interesting in my opinion would be to evaluate the Kullback-Leibler distance to the true posterior.

Figure 1 of the paper compares the ABC outcome when using four different statistics, empirical mean, empirical variance, minimum and maximum, for a normal sample with imprecise size and unknown mean. The comment that only the empirical mean recovers the true posterior is both correct and debatable because the minimum and maximum observations also contain information about the unknown mean, albeit at a lower convergence rate. This leads to the issue raised by one referee of our PNAS paper about the [lack of] worth in distinguishing between estimation and testing. At a mathematical level, it is correct that a wrong choice of summary statistic (like the empirical variance above) may provide no information for estimation as well as testing. At a methodological level, we now agree that different statistics should be used for testing and for estimation. Minor point: I find it surprising that the tolerance is the same for all collections of summary statistics. Using a log transform is certainly not enough to standardise the thing.

The quite interesting conclusion about the population genetic study states that one model requires more statistics than another one. This is when considering estimation

separately for each model. From a model choice perspective, this cannot be the case: all models must involve the same collection of summary statistics for the posterior probability to be correctly defined. This issue has been puzzling me for years about ABC: a proper ABC approximation is model dependent however one needs the "same" statistics to run the comparison.

Stoehr et al. (2014) consider summary statistics for ABC model choice in hidden Gibbs random fields. The move to a hidden Markov random field means that our original approach (Grelaud et al., 2009) does not apply: there is no dimension-reduction sufficient statistics in that case... The authors introduce a small collection of focussed statistics to discriminate between Potts models. They further define a novel misclassification, the predictive error rate discussed below. In a simulation experiment, the paper shows that the predictive error rate decreases quite a lot by including 2 or 4 geometric summary statistics on top of the no-longer-sufficient concordance statistics.

> "[the ABC posterior probability of index m] uses the data twice: a first one to calibrate the set of summary statistics, and a second one to compute the ABC posterior." J. Stoehr et al., 2014

It took me a while to understand the above quote. If we consider ABC model choice as we did in our original paper, it only and correctly uses the data once. However, if we select the vector of summary statistics based on an empirical performance indicator resulting from the data then indeed the procedure does use the data twice! Is there a generic way or trick to compensate for that, apart from cross-validation?

6.5 Optimal Choice of Summary Statistics

Prangle et al. (2014) offers another study of the selection of summary statistics for ABC model choice. The crux of the analysis is that the Bayes factor is a good type of summary when comparing two models, this result extending to more model by considering instead the vector of evidences. As in the initial Read Paper (Fearnhead and Prangle, 2012), there is no true optimality in using the Bayes factor or vector of evidences, strictly speaking, besides the fact that the vector of evidences is minimal sufficient for the marginal models (integrating out the parameters). The implementation of the principle is similar to this Read Paper setting as well: run a pilot ABC simulation, estimate the vector of evidences, and re-run the main ABC simulation using this estimate as the summary statistic. The paper contains a simulation study using some of our examples (Marin et al., 2011), as well as an application to genetic bacterial data.

That the Bayes factor was acceptable as a statistic was quite natural in terms of our consistency result (Marin et al., 2014) as it is converging to 0 and to ∞ depending from which model the data is generated. The paper is well-written and clear enough to understand how the method is implemented. It also provides a very fair coverage of our own paper. However, I do not understand several points. For one thing, given

that the vector of evidence is the target, I do not see why the vector of Bayes factors for all pairs is used instead, leading to a rather useless inflation in the dimension of the summary statistic. Using a single model for the denominator would be enough (and almost sufficient).

Somehow in connection with the above, the use of the logistic regularisation for computing the posterior probability (following an idea of Marc Beaumont in the mid 2000's) is interesting but difficult to quantify. Using a logistic regression based on the training sample sounds like the natural solution to compute the sufficient statistic, however the construction of the logistic regression by regular variable selection techniques means that different transforms of the data are used to compare different models, an issue that worries me (see again below). Obviously, the overall criticism on the Read Paper, namely that the quality of the outcome ultimately depends on the choice of the first batch of statistics, still applies: too many statistics and there is no reason to believe in the quality of the ABC, too few statistics and there is no reason to trust the predictive power of the logistic regression.

The authors also introduce a different version of the algorithm where they select a subregion of the parameter space(s) during the pilot run and replace the prior with the prior restricted to that region during the main run. The paper claims significant improvements brought by this additional stage, but it makes me somewhat uneasy: For one thing, it uses the data twice, with a risk of over-concentration. For another, I do not see how the restricted region could be constructed, esp. in large dimensions (an issue I had when using HPD regions for harmonic mean estimators), apart from the maybe inefficient hypercube. For yet another (maybe connected with the first thing!), a difference between models is induced by this pilot run restriction, which amounts to changing the prior weights of the models under comparison.

A side remark in the conclusion suggests using different vectors of statistics in a pairwise comparison of models. While I have also been tempted by this solution, because it produces a huge reduction in dimension, I wonder at its validation, as it amounts to comparing models based on different (transforms of) observations, so the evidences are not commensurable. I however agree with the authors that using a set of summary statistics to run ABC model comparisons and another one to run ABC estimation for a given model sounds like a natural approach, as it fights the curse of dimensionality.

6.6 Towards Estimating Posterior Probabilities

Stoehr et al. (2014) attack the recurrent problem of selecting summary statistics for ABC in a hidden Markov random field, where is no fixed dimension sufficient statistics. The paper provides a very broad overview of the issues and difficulties related with ABC model choice, which has been the focus of some advanced research only for a few years. Most interestingly, the authors define a novel, local, and somewhat Bayesian misclassification rate, an error that is conditional on the observed value and derived from the ABC reference table. It is the posterior predictive error

rate

$$\mathbb{P}^{\text{ABC}}(\hat{m}(y^{\text{obs}}) \neq m | S(y^{\text{obs}}))$$

integrating in both the model index m and the corresponding random variable Y (and the hidden intermediary parameter) given the observation. Or rather given the transform of the observation by the summary statistic S. The authors even go further to define the error rate of a classification rule based on a first (collection of) statistic, conditional on a second (collection of) statistic (see Definition 1). A notion rather delicate to validate on a fully Bayesian basis. And they advocate the substitution of the unreliable (estimates of the) posterior probabilities by this local error rate, estimated by traditional non-parametric kernel methods. Methods that are calibrated by cross-validation. Given a reference summary statistic, this perspective leads (at least in theory) to select the optimal summary statistic as the one leading to the minimal local error rate. Besides its application to hidden Markov random fields, which is of interest per se, this paper thus opens a new vista on calibrating ABC methods and evaluating their true performances conditional on the actual data. The advocated abandonment of the estimation of all posterior probabilities could almost justify the denomination of a paradigm shift. This is also the approach advocated in Pudlo et al. (2014).

However, the above posterior predictive error rate is the conditional expected value of a misclassification loss when conditioning on the data (or more precisely some summaries of the data) being what it is. Hence, when integrating this conditional error over the marginal distribution of the summaries of the data, we recover the misclassification error integrated over the whole prior space. This quantity differs from the posterior (predictive) error rate computed in an initial version of Pudlo et al. (2014), which involves an expectation over the predictive distribution given the observed data and thus, a second integral over the data space. As a consequence, the conditional error rates of Stoehr et al. (2014) is on the same ground as the posterior probabilities.

Pudlo et al. (2014) offers the central arguments that (a) using random forests is a good tool for choosing the most appropriate model, (b) evaluating the posterior misclassification error is available via standard ABC arguments, and (c) estimating the posterior probability of the selected model is possible via further random forests. The call to the machine-learning tool of a random forest (Breiman, 2001), traditionally used in classification, may sound at first at odds with a Bayesian approach, but it becomes completely justified once one sets the learning set as generated from the prior predictive distribution. A random forest is then a randomised version of a non-parametric predictor of the model index given the data. Note that Pham et al. (2014) also use random forests for ABC parameter estimation.

Let us briefly recall that a random forest aggregates classification trees, CART, (Breiman et al., 1984) by introducing for each tree a randomisation step represented in Algorithm 7 and consisting in bootstrapping the original sample and subsampling the summary statistics at each node of the tree. A CART is a binary classification tree that partitions the covariate space towards a prediction of the class index. Each node of this tree consists in a rule of the form $S_j < t_j$, where S_j is one of the covariates

and t_j is chosen towards minimising an heterogeneity index (Hastie et al., 2009). In ABC model choice, a CART tree is calibrated from the reference table and it returns a model index for the observed summary statistic s^{obs}, following a path according to these binary rules.

Algorithm 7 Randomised CART

start the tree with a single root
repeat
  pick a non-homogeneous tip v such that $Q(v) \neq 1$
  attach to v two daughter nodes v_1 and v_2
  draw a random subset of covariates of size n_{try}
  for all covariates X_j in the random subset **do**
    find the threshold t_j in the rule $S_j < t_j$ that minimises $N(v_1)Q(v_1)+N(v_2)Q(v_2)$
  end for
  find the rule $S_j < t_j$ that minimises $N(v_1)Q(v_1)+N(v_2)Q(v_2)$ in j **and set** this best rule to node v
until all tips v are homogeneous ($Q(v) = 0$)
set the labels of all tips

Reproduced with permission of the authors from Pudlo et al. (2014).

Pudlo et al. (2014) then selects the most likely model among a collection of models, based on a random forest classifier made of several hundreds CARTs as illustrated below, as a majority vote decision, i.e., the most fequently allocated model among the trees.

Algorithm 8 RF for classification

for $b = 1$ **to** B **do**
  draw a bootstrap sub-sample Z^* of size N_{boot} from the training data
  grow a tree T_b trained on Z^* with Algorithm 7
end for
output the ensemble of trees $\{T_b, b = 1 \dots B\}$

Reproduced with permission of the authors from Pudlo et al. (2014).

A first approach envisioned random forests as a mere filter applied to a large set of summary statistics in order to produce a relevant subset of significant statistics, with the additional appeal of an associated distance between datasets induced by the forest itself. However, we later realised that (a) further ABC steps were counterproductive, once the model was selected by the random forest; (b) including more summary statistics was always beneficial to the performances of the forest; and (c) the connections between (i) the true posterior probability of a model, (ii) the ABC version of this probability, (iii) the random forest frequency approximating the above, were at best very loose. While the random forest approach offers the advantage of incorporating all available summary statistics and not imposing a preliminary selection among those, it obviously weights the most discriminating ones more heavily. For instance,

in Pudlo et al. (2014), the linear discriminant analysis (LDA) components are among the most often used. Experimentts in Pudlo et al. (2014) show that the frequencies of the various models produced by Algorithm 6 are however not directly related with their posterior probabilities.

Exploiting the approach of Stoehr et al. (2014), Pudlo et al. (2014) still managed to produce a reliable estimate of those. Indeed, the posterior expected error associated with the 0–1 loss (Robert, 2001)

$$\mathbb{I}(\hat{\mathfrak{m}}(s^{\text{obs}}) \neq \mathfrak{m}) \tag{2}$$

where $\hat{\mathfrak{m}}(s^{\text{obs}})$ is the model selection procedure, can be shown to satisfy (Pudlo et al., 2014)

$$\mathbb{E}[\mathbb{I}(\hat{\mathfrak{m}}(s^{\text{obs}}) \neq \mathfrak{m})|s^{\text{obs}}] = 1 - \mathbb{P}[\mathfrak{m} = \hat{\mathfrak{m}}(s^{\text{obs}})|s^{\text{obs}}]\,.$$

This expected loss is thus the complement to the posterior probability that the true model is the MAP. While it is not directly available, it can be estimated from the reference table as a regression of $\mathfrak{m}$ or more exactly $\mathbb{I}(\hat{\mathfrak{m}}(s) \neq \mathfrak{m})$ over s^{obs}. A natural solution in this context is to use another random forest, producing a function $\rho(s)$ that estimates $\mathbb{P}[\mathfrak{m} \neq \hat{\mathfrak{m}}(s)|s]$ and to apply this function to the actual observations to deduce $1-\rho(s^{\text{obs}})$ as an estimate of $\mathbb{P}[\mathfrak{m} = \hat{\mathfrak{m}}(s^{\text{obs}})|s^{\text{obs}}]$.

7 Conclusion

This survey reflects upon the diversity and the many directions of progress in the field of ABC research. The overall take-home message is that the on-going research in this area has led both to consider ABC as part of the statistical toolbox and to envision different approaches to statistical modelling, where a complete representation of the whole world is no always feasible. Following the evolution of ABC in the past fifteen years we have thus moved from constructing approximate methods to accepting working with approximate models, a positive move in my opinion. This document being based on blog posts written when the initial version of the paper, they may appear overcritical of papers later smoothed into journal articles. I am thus welcoming any reply or discussion to be included in later versions of the survey.

References

Albert, C., Künsch, H., and Scheidegger, A. (2014). A simulated annealing approach to approximate Bayes computations. *Statistics and Computing*, pages 1–16.

Allingham, D., King, R., and Mengersen, K. (2009). Bayesian estimation of quantile distributions. *Statistics and Computing*, 19:189–201.

Amzal, B., Bois, F., Parent, E., and Robert, C. (2006). Bayesian-optimal design via interacting particle systems. *J. American Statist. Assoc.*, 101(474):773–785.

Andrieu, C., Doucet, A., and Holenstein, R. (2011). Particle Markov chain Monte Carlo (with discussion). *J. Royal Statist. Society Series B*, 72 (2):269–342.

Andrieu, C. and Roberts, G. (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *Ann. Statist.*, 37(2):697–725.

Andrieu, C. and Vihola, M. (2014). Establishing some order amongst exact approximations of MCMCs. *ArXiv e-prints*.

Arnold, B. and Ng, T. (2011). Flexible bivariate beta distributions. *Journal of Multivariate Analysis*, 102(8):1194–1202.

Arratia, R. and DeSalvo, S. (2012). On the Random Sampling of Pairs, with Pedestrian examples. *ArXiv e-prints*.

Atchadé, Y., Lartillot, N., and Robert, C. (2013). Bayesian computation for statistical models with intractable normalizing constants. *Brazilian Journal of Probability and Statistics*, 27(4):416–436.

Barber, S., Voss, J., and Webster, M. (2013). The Rate of Convergence for Approximate Bayesian Computation. *ArXiv e-prints*.

Barndorff-Nielsen, O. E. and Shephard, N. (2001). Non-Gaussian Ornstein-Uhlenbeck-based models and some of their uses in financial economics. *J. Royal Statist. Society Series B*, 63(2):167–241.

Barnes, C., Filippi, S., Stumpf, M., and Thorne, T. (2012). Considerate approaches to constructing summary statistics for ABC model selection. *Statistics and Computing*, 22(6):1181–1197.

Barthelmé, S. and Chopin, N. (2014). Expectation propagation for likelihood-free inference. *Journal of the American Statistical Association*, 109(505):315–333.

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

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

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. National Academy Sciences*, 107(41):E57.

Biau, G., Cérou, F., and Guyader, A. (2014). New insights into Approximate Bayesian Computation. *Annales de l'IHP (Probability and Statistics)*.

Blum, M. (2010). Approximate Bayesian Computation: a non-parametric perspective. *J. American Statist. Assoc.*, 105(491):1178–1187.

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

Blum, M. G. B., Nunes, M. A., Prangle, D., and Sisson, S. A. (2013). A comparative review of dimension reduction methods in Approximate Bayesian computation. *Statistical Science*, 28(2):189–208.

Bollerslev, T., Chou, R., and Kroner, K. (1992). ARCH modeling in finance. a review of the theory and empirical evidence. *J. Econometrics*, 52:5–59.

Bornn, L., Pillai, N., Smith, A., and Woodard, D. (2014). A Pseudo-Marginal Perspective on the ABC Algorithm. *ArXiv e-prints*.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Brooks-Pollock, E., Roberts, G., and Keeling, M. (2014). A dynamic model of bovine tuberculosis spread and control in Great Britain. *Nature*, 511:228—-231.

Calvet, C. and Czellar, V. (2014). Accurate methods for approximate Bayesian computation filtering. *J. Econometrics*. (to appear).

Cappé, O., Douc, R., Guillin, A., Marin, J.-M., and Robert, C. (2008). Adaptive importance sampling in general mixture classes. *Statist. Comput.*, 18:447–459.

Cappé, O., Guillin, A., Marin, J.-M., and Robert, C. (2004). Population Monte Carlo. *J. Comput. Graph. Statist.*, 13(4):907–929.

Chiachio, M., Beck, J. L., Chiachio, J., and Rus, G. (2014). Approximate Bayesian Computation by Subset Simulation. *ArXiv e-prints*.

Chkrebtii, O. A., Cameron, E. K., Campbell, D. A., and Bayne, E. M. (2013). Transdimensional Approximate Bayesian Computation for Inference on Invasive Species Models with Latent Variables of Unknown Dimension. *ArXiv e-prints*.

Cornuet, J.-M., Marin, J.-M., Mira, A., and Robert, C. (2012). Adaptive multiple importance sampling. *Scandinavian Journal of Statistics*, 39(4):798–812.

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

Cox, D. R. and Kartsonaki, C. (2012). The fitting of complex parametric models. *Biometrika*, 99(3):741–747.

Crackel, R. C. and Flegal, J. M. (2014). Approximate Bayesian computation for a flexible class of bivariate beta distributions. *ArXiv e-prints*.

Davison, A. C., Hinkley, D. V., and Worton, B. (1992). Bootstrap likelihoods. *Biometrika*, 79(1):113–130.

Dean, T., Singh, S., Jasra, A., and Peters, G. (2014). Parameter inference for hidden Markov models with intractable likelihoods. *Scand. J. Statist.* (to appear).

Del Moral, P., Doucet, A., and Jasra, A. (2006). Sequential Monte Carlo samplers. *J. Royal Statist. Society Series B*, 68(3):411–436.

Del Moral, P., Doucet, A., and Jasra, A. (2012). An adaptive strategy for sequential Monte Carlo methods. *Bernoulli*, 18(1):252–278.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Royal Statist. Society Series B*, 39:1–38.

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

Diggle, P. and Gratton, R. (1984). Monte Carlo methods of inference for implicit statistical models. *J. Royal Statist. Society Series B*, 46:193–227.

Donnet, S., Foulley, J.-L., and A., S. (2010). Bayesian analysis of growth curves using mixed models defined by stochastic differential equations. *Biometrics*, 66(3):733–741.

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

Doucet, A., de Freitas, N., and Gordon, N. (1999). *Sequential MCMC in Practice*. Springer-Verlag.

Doucet, A., Godsill, S., and Robert, C. (2002). Marginal maximum a posteriori estimation using Markov chain Monte Carlo. *Statistics and Computing*, 12:77–84.

Drovandi, C. and Pettitt, A. (2010). Estimation of Parameters for Macroparasite Population Evolution Using Approximate Bayesian Computation. *Biometrics*. (To appear.).

Drovandi, C. and Pettitt, A. (2011). Likelihood-free bayesian estimation of multivariate quantile distributions. *Computational Statistics and Data Analysis*, 55:2541–2556.

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

Ehrlich, E., Jasra, A., and Kantas, N. (2014). Gradient free parameter estimation for hidden markov models with intractable likelihoods. *Method. Comp. Appl. Probab.* (to appear).

Estoup, A., Lombaert, E., Marin, J.-M., Robert, C., Guillemaud, T., Pudlo, P., and Cornuet, J.-M. (2012). Estimation of demo-genetic model probabilities with Approximate Bayesian Computation using linear discriminant analysis on summary statistics. *Molecular Ecology Ressources*, 12(5):846–855.

Everitt, R. G. (2014). Bayesian Parameter Estimation for Latent Markov Random Fields and Social Networks. *J. Comput. Graph. Statist.* (to appear).

Excoffier, C., Leuenberger, D., and Wegmann, L. (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. *Proceedings of the National Academy of Sciences*, 104(45):17614–17619.

Fan, Y., Nott, D., and Sisson, S. (2013). Approximate Bayesian computation via regression density estimation. *Stat*, 2(1):34–48.

Fearnhead, P. and Prangle, D. (2012). Constructing summary statistics for Approximate Bayesian Computation: semi-automatic Approximate Bayesian Computation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(3):419–474. (With discussion.).

Feller, W. (1970). *An Introduction to Probability Theory and its Applications*, volume 1. John Wiley, New York.

Filippi, S., Barnes, C., Cornebise, J., and Stumpf, M. (2013). On optimality of kernels for approximate Bayesian computation using sequential Monte Carlo. *Statistical Applications in Genetics and Molecular Biology*, 12(1):87—107.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., and Rubin, D. (2013). *Bayesian Data Analysis*. Chapman and Hall, New York, New York, third edition.

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.

Gilchrist, W. (2000). *Statistical Modelling with Quantile Functions*. Chapman and Hall.

Gleim, A. and Pigorsch, C. (2013). Approximate Bayesian computation with indirect summary statistics. Technical report, Universität Jena, Germany.

Golightly, A. and Wilkinson, D. (2011). Bayesian parameter inference for stochastic biochemical network models using particle MCMC. *Interface Focus*, 1(6):807–820.

Gouriéroux, C. and Monfort, A. (1995). *Simulation Based Econometric Methods*. CORE Lecture Series, CORE, Louvain.

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

Grazian, C. and Liseo, B. (2014). Approximate Integrated Likelihood via ABC methods. *ArXiv e-prints*.

Green, P. (1995). Reversible jump MCMC computation and Bayesian model determination. *Biometrika*, 82(4):711–732.

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.

Haario, H., Saksman, E., and Tamminen, J. (1999). Adaptive proposal distribution for random walk Metropolis algorithm. *Computational Statistics*, 14(3):375–395.

Hainy, M., Müller, W. G., and Wagner, H. (2013). Likelihood-free Simulation-based Optimal Design. *ArXiv e-prints*.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer-Verlag, New York.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning. Data mining, inference, and prediction.* Springer Series in Statistics. Springer-Verlag, New York, 2 edition.

Haynes, M. A., MacGillivray, H. L., and Mengersen, K. L. (1997). Robustness of ranking and selection rules using generalised g-and-k distributions. *J. Statist. Plann. Inference*, 65(1):45–66.

Hoerl, A. and Kennard, R. (1970). Ridge regression: biased estimators for non-orthogonal problems. *Technometrics*, 12:55–67.

Iba, Y. (2000). Population-based Monte Carlo algorithms. *Trans. Japanese Soc. Artificial Intell.*, 16(2):279–286.

Jaakkola, T. and Jordan, M. (2000). Bayesian parameter estimation via variational methods. *Statistics and Computing*, 10:25–37.

Jabot, F., Lagarrigues, G., Courbaud, B., and Dumoulin, N. (2014). A comparison of emulation methods for Approximate Bayesian Computation. *ArXiv e-prints*.

Jasra, A. (2014). Approximate Bayesian Computation for a Class of Time Series Models. *ArXiv e-prints*.

Jasra, A., Kantas, N., and Ehrlich, E. (2014). Approximate inference for observation driven time series models with intractable likelihoods. *TOMACS*. (to appear).

Jasra, A., Lee, A., Yau, C., and Zhang, X. (2013). The Alive Particle Filter. *ArXiv e-prints*.

Jasra, A., Singh, S., Martin, J., , and McCoy, E. (2012). Filtering via Approximate Bayesian computation. *Statist. Comp.*, 22:1223–1237.

Jordan, M. and Jacobs, R. (1994). Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, 6:181–214.

Joyce, P. and Marjoram, P. (2008). Approximately sufficient statistics and Bayesian computation. *Statistical Applications in Genetics and Molecular Biology*, 7(1):article 26.

Le Gland, F. and Oudjane, N. (2006). *A sequential particle algorithm that keeps the particle system alive*, volume 337 of *Lecture Notes in Control and Information Sciences*, pages 351–389. Springer, Berlin.

Lee, A. and Latuszynski, K. (2014). Variance bounding and geometric ergodicity of Markov chain Monte Carlo kernels for Approximate Bayesian computation. *Biometrika*, 101(3):655–671.

Lele, S., Dennis, B., and Lutscher, F. (2007). Data cloning: easy maximum likelihood estimation for complex ecological models using Bayesian Markov chain Monte Carlo methods. *Ecology Letters*, 10:551–563.

Lenormand, M., Jabot, F., and Deffuant, G. (2013). Adaptive approximate Bayesian computation for complex models. *Computational Statistics*, 28:2777–2796.

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

Li, C. and Jiang, W. (2014). Model Selection for Likelihood-free Bayesian Methods Based on Moment Conditions: Theory and Numerical Examples. *ArXiv e-prints*.

Li, J., Nott, D. J., Fan, Y., and Sisson, S. A. (2015). Extending approximate Bayesian computation methods to high dimensions via Gaussian copula. *ArXiv e-prints*.

Li, W. and Fearnhead, P. (2015). Behaviour of ABC for Big Data. *ArXiv e-prints*.

Lombaert, E., Guillemaud, T., and et al., C. T. (2011). Inferring the origin of populations introduced from a genetically structured native range by Approximate Bayesian Computation: case study of the invasive ladybird Harmonia axyridis. *Molecular Ecology*, 20:4654–4670.

Lunn, D., Thomas, A., Best, N., and Spiegelhalter, D. (2010). *The BUGS Book: A Practical Introduction to Bayesian Analysis*. Chapman & Hall/CRC Press.

Marin, J., Pillai, N., Robert, C., and Rousseau, J. (2014). Relevant statistics for Bayesian model choice. *J. Royal Statist. Society Series B*. (to appear).

Marin, J., Pudlo, P., Robert, C., and Ryder, R. (2011). Approximate Bayesian computational methods. *Statistics and Computing*, pages 1–14.

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

Martin, G. M., McCabe, B. P. M., Maneesoonthorn, W., and Robert, C. P. (2014). Approximate Bayesian Computation in State Space Models. *ArXiv e-prints*.

Martin, J., Jasra, A., Singh, S., Whiteley, N., Del Moral, P., and McCoy, E. (2014). Approximate Bayesian computation for smoothing. *Stoch. Anal. Appl.* (to appear).

McKinley, T., Ross, J., Deardon, R., and Cook, A. (2014). Simulation-based Bayesian inference for epidemic models. *Computational Statistics and Data Analysis*, 71:434–447.

McVinish, R. (2012). Improving ABC for quantile distributions. *Statistics and Computing*, In Press.

Meeds, E., Leenders, R., and Welling, M. (2015). Hamiltonian ABC. *ArXiv e-prints*.

Meeds, E. and Welling, M. (2014). GPS-ABC: Gaussian Process Surrogate Approximate Bayesian Computation. *ArXiv e-prints*.

Mengersen, K., Pudlo, P., and Robert, C. (2013). Bayesian computation via empirical likelihood. *Proceedings of the National Academy of Sciences*, 110(4):1321–1326.

Mengersen, K. and Tweedie, R. (1996). Rates of convergence of the Hastings and Metropolis algorithms. *Ann. Statist.*, 24:101–121.

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

Moores, M. T., Drovandi, C. C., Mengersen, K., and Robert, C. P. (2014). Preprocessing for approximate Bayesian computation in image analysis. *Statistics and Computing*. (to appear).

Müller, P. (1999). Simulation based optimal design. In Bernardo, J., Berger, J., Dawid, A., and Smith, A., editors, *Bayesian Statistics*, volume 6, pages 459–474, New York. Springer–Verlag.

Murakami, Y. (2014). Bayesian parameter inference and model selection by population annealing in systems biology. *PLoS ONE*, 9(8):e104057.

Murray, I., Ghahramani, Z., , and MacKay, D. (2006). Mcmc for doubly-intractable distributions. In *Uncertainty in Artificial Intelligence*. UAI-2006.

Niederreiter, H. (1992). *Random Number Generation and Quasi-Monte Carlo Methods*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA.

Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75:237–249.

Owen, A. B. (2001). *Empirical Likelihood*. Chapman & Hall.

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.

Pham, K. C., Nott, D. J., and Chaudhuri, S. (2014). A note on approximating ABC-MCMC using flexible classifiers. *Stat*, 3(1):218–227.

Picchini, U. (2015). Approximate maximum likelihood estimation using data-cloning ABC. *ArXiv e-prints*.

Picchini, U. and Lyng Forman, J. (2013). Accelerating inference for diffusions observed with measurement error and large Csample sizes using Approximate Bayesian Computation. *ArXiv e-prints*.

Prangle, D. (2014). Lazy ABC. *ArXiv e-prints*.

Prangle, D., Blum, M. G. B., Popovic, G., and Sisson, S. A. (2013). Diagnostic tools of approximate Bayesian computation using the coverage property. *ArXiv e-prints*.

Prangle, D., Fearnhead, P., Cox, M., Biggs, P., and French, N. (2014). Semi-automatic selection of summary statistics for abc model choice. *Statistical Applications in Genetics and Molecular Biology*, 13(1):67–82.

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

Pudlo, P., Marin, J.-M., Estoup, A., Cornuet, J.-M., Gautier, M., and Robert, C. P. (2014). Reliabe ABC model choice via random forests. *ArXiv e-prints*.

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., Wiuf, C., and Richardson, S. (2010). Reply to robert et al.: Model criticism informs model choice and model comparison. *Proceedings of the National Academy of Sciences*, 107(3):E6–E7.

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. Natl. Acad. Sciences USA*, 106:1–6.

Ratmann, O., Camacho, A., Meijer, A., and Donker, G. (2013). Statistical modelling of summary values leads to accurate Approximate Bayesian Computations. *ArXiv e-prints*.

Robert, C. (2001). *The Bayesian Choice*. Springer-Verlag, New York, second edition.

Robert, C. (2012). Discussion of "constructing summary statistics for Approximate Bayesian Computation" by P. Fernhead and D. Prangle. *J. Royal Statist. Society Series B*, 74(3):447–448.

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

Robert, C., Cornuet, J.-M., Marin, J.-M., and Pillai, N. (2011). Lack of confidence in ABC model choice. *Proceedings of the National Academy of Sciences*, 108(37):15112–15117.

Robert, C., Mengersen, K., and Chen, C. (2010). Model choice versus model criticism. *Proceedings of the National Academy of Sciences*, 107(3):E5.

Robert, C. and Soubiran, C. (1993). Estimation of a mixture model through Bayesian sampling and prior feedback. *TEST*, 2:125–146.

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

Rubinstein, R. Y. and Kroese, D. P. (2004). *The Cross-Entropy Method*. Springer-Verlag, New York.

Rubio, F. J. and Johansen, A. M. (2013). A Simple Approach to Maximum Intractable Likelihood Estimation. *ArXiv e-prints*.

Ruli, E., Sartori, N., and Ventura, L. (2013). Approximate Bayesian Computation with composite score functions. *ArXiv e-prints*.

Sedki, M., Pudlo, P., Marin, J.-M., Robert, C. P., and Cornuet, J.-M. (2012). Efficient learning in ABC algorithms. *ArXiv e-prints*.

Silk, D., Filippi, S., and Stumpf, M. (2013). Optimizing threshold-schedules for sequential approximate Bayesian computation: applications to molecular systems. *Stat. Appl. Genet. Mol. Biol.*, 12(5):603–618.

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

Skilling, J. (2006). Nested sampling for general Bayesian computation. *Bayesian Analysis*, 1(4):833–860.

Slatkin, M. (1995). A measure of population subdivision based on microsatellite allele frequencies. *Genetics*, 139(1):457–462.

Spall, J. C. (2003). *Introduction to Stochastic Search and Optimization*. John Wiley, New York.

Stephens, M. and Donnelly, P. (2000). Inference in molecular population genetics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):605–635.

Stigler, S. (1986). *The History of Statistics*. Belknap, Cambridge.

Stoehr, J., Pudlo, P., and Cucala, L. (2014). Adaptive ABC model choice and geometric summary statistics for hidden Gibbs random fields. *Statistics and Computing*, pages 1–13.

Sunnåker, M., Busetto, A., Numminen, E., Corander, J., Foll, M., and Dessimoz, C. (2013). Approximate Bayesian computation. *PLoS Comput. Biol.*, 9(1):e1002803.

Tanner, M. and Wong, W. (1987). The calculation of posterior distributions by data augmentation. *J. American Statist. Assoc.*, 82:528–550.

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. National Academy of Sciences*, 107(14):6376–6381.

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.

Wikipedia (2014). Approximate Bayesian computation — Wikipedia, The Free Encyclopedia.

Wilkinson, D. (2006). *Stochastic Modelling for Systems Biology*. Chapman & Hall/CRC Press, Boca Raton, Florida.

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

Wilkinson, R. (2013). Approximate Bayesian computation (ABC) gives exact results under the assumption of model error. *Statistical Applications in Genetics and Molecular Biology*, 12(2):129–141.

Wilkinson, R. D. (2014). Accelerating ABC methods using Gaussian processes. *ArXiv e-prints*.

Wood, S. (2010). Statistical inference for noisy nonlinear ecological dynamic systems. *Nature*, 466:1102—1104.

Zhu, W., Diazaraque, J. M. M., and Leisen, F. (2014). A bootstrap likelihood approach to Bayesian computation. Technical report, Universidad Carlos III, Departamento de Estadística y Econometría.