



Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation

Paul Fearnhead and Dennis Prangle

Lancaster University, UK

[Read before The Royal Statistical Society at a meeting organized by the Research Section on Wednesday, December 14th, 2011, Professor D. M. Titterton in the Chair]

Summary. Many modern statistical applications involve inference for complex stochastic models, where it is easy to simulate from the models, but impossible to calculate likelihoods. Approximate Bayesian computation (ABC) is a method of inference for such models. It replaces calculation of the likelihood by a step which involves simulating artificial data for different parameter values, and comparing summary statistics of the simulated data with summary statistics of the observed data. Here we show how to construct appropriate summary statistics for ABC in a semi-automatic manner. We aim for summary statistics which will enable inference about certain parameters of interest to be as accurate as possible. Theoretical results show that optimal summary statistics are the posterior means of the parameters. Although these cannot be calculated analytically, we use an extra stage of simulation to estimate how the posterior means vary as a function of the data; and we then use these estimates of our summary statistics within ABC. Empirical results show that our approach is a robust method for choosing summary statistics that can result in substantially more accurate ABC analyses than the *ad hoc* choices of summary statistics that have been proposed in the literature. We also demonstrate advantages over two alternative methods of simulation-based inference.

Keywords: Indirect inference; Likelihood-free inference; Markov chain Monte Carlo methods; Simulation; Stochastic kinetic networks

1. Introduction

1.1. Background

Many modern statistical applications involve inference for stochastic models given partial observations. Often it is easy to simulate from the models but calculating the likelihood of the data, even by using computationally intensive methods, is impracticable. In these cases a natural approach to inference is to use simulations from the model for different parameter values, and to compare the simulated data sets with the observed data. Loosely, the idea is to estimate the likelihood of a given parameter value from the proportion of data sets, simulated using that parameter value, that are ‘similar to’ the observed data. This idea dates back at least as far as Diggle and Gratton (1984).

If we replace ‘similar to’ with ‘the same as’ (see for example Tavaré *et al.* (1997)), then this approach would give an unbiased estimate of the likelihood; and asymptotically as we increase the amount of simulation we obtain a consistent estimate. However, in most applications the probability of an exact match of the simulated data with the observed data is negligible, or 0,

Address for correspondence: Paul Fearnhead, Department of Mathematics and Statistics, Fylde College, Lancaster University, Lancaster, LA1 4YF, UK.
E-mail: p.fearnhead@lancaster.ac.uk

so we cannot consider such exact matches. The focus of this paper is how to define ‘similar to’ for these cases.

In this paper we focus on a particular approach: approximate Bayesian computation (ABC). This approach combines an estimate of the likelihood with a prior to produce an approximate posterior, which we shall refer to as the ABC posterior. The use of ABC initially became popular within population genetics, where simulation from a range of population genetic models is possible by using the coalescent (Kingman, 1982), but where calculating likelihoods is impracticable for realistic sized data sets. The first use of ABC was by Pritchard *et al.* (1999), who looked at inference about human demographic history. Further applications include inference for recombination rates (Padhukasahasram *et al.*, 2006), evolution of pathogens (Wilson *et al.*, 2009) and evolution of protein networks (Ratmann *et al.*, 2009). Its increasing importance can be seen by the current range of application of ABC, which has recently been applied within epidemiology (McKinley *et al.*, 2009; Tanaka *et al.*, 2006), inference for extremes (Bortot *et al.*, 2007), dynamical systems (Toni *et al.*, 2009) and Gibbs random fields (Grelaud *et al.*, 2009) among many others. Part of the appeal of ABC is its flexibility; it can easily be applied to any model for which forward simulation is possible. For example Wegmann *et al.* (2009) stated that ABC

‘should allow evolutionary geneticists to reasonably estimate the parameters they are really interested in, rather than require them to shift their interest to problems for which full-likelihood solutions are available’.

Recently software has been developed to help to implement ABC within population genetics (Cornuet *et al.*, 2008; Lopes *et al.*, 2009) and systems biology (Liepe *et al.*, 2010).

1.2. Approximate Bayesian computation algorithms and approximations

Consider analysing n -dimensional data \mathbf{y}_{obs} . We have a model for the data, which depends on an unknown p -dimensional parameter θ . Denote the probability density of the data given a specific parameter value by $\pi(\mathbf{y}|\theta)$, and denote our prior by $\pi(\theta)$. We assume that it is simple to simulate \mathbf{Y} from $\pi(\mathbf{y}|\theta)$ for any θ , but that we do not have an analytic form for $\pi(\mathbf{y}|\theta)$.

We define the ABC posterior in terms of

- a function $S(\cdot)$ which maps the n -dimensional data onto a d -dimensional summary statistic,
- a density kernel $K(\mathbf{x})$ for a d -dimensional vector \mathbf{x} , which integrates to 1, and
- a bandwidth $h > 0$.

density kernel as the weight

Let $\mathbf{s}_{\text{obs}} = S(\mathbf{y}_{\text{obs}})$. If we now define an approximation to the likelihood as

$$p(\theta|\mathbf{s}_{\text{obs}}) = \int \pi(\mathbf{y}|\theta) K[\{S(\mathbf{y}) - \mathbf{s}_{\text{obs}}\}/h] d\mathbf{y},$$

(6)

then the ABC posterior can be defined as

$$\pi_{\text{ABC}}(\theta|\mathbf{s}_{\text{obs}}) \propto \pi(\theta) p(\theta|\mathbf{s}_{\text{obs}}). \quad (1)$$

The idea of ABC is that the ABC posterior will approximate, in some way, the true posterior for θ and can be used for inference about θ . The form of approximation will depend on the choice of $S(\cdot)$, $K(\cdot)$ and h . For example, if $S(\cdot)$ is the identity function then we can view $p(\theta|\mathbf{s}_{\text{obs}})$ as a kernel density approximation to the likelihood. If also $h \rightarrow 0$, then the ABC posterior will converge to the true posterior. For other choices of $S(\cdot)$, the kernel is measuring the closeness of \mathbf{y} to \mathbf{y}_{obs} just via the closeness of $S(\mathbf{y})$ to \mathbf{s}_{obs} . The reason for considering ABC is that we can

small bandwidth h means we use summary statistics $S(\mathbf{y})$ that is close to the observed $\mathbf{s}(\mathbf{y}_{\text{obs}})$ to estimate (\mathbf{y}) i.e. the approximation $p(\cdot|\mathbf{s}_{\text{obs}})$ is close to the real likelihood
窗宽 h 越小, 较远的点获得的权重越小

Table 1. Algorithm 1: importance (and rejection) sampling implementation of ABC

<p><i>Input</i>—a set of data \mathbf{y}_{obs}, and a function $S(\cdot)$; a density kernel $K(\cdot)$, with $\max\{K(\mathbf{x})\} = 1$ and a bandwidth $h > 0$; a proposal density $g(\theta)$, with $g(\theta) > 0$ if $\pi(\theta) > 0$; an integer $N > 0$</p> <p><i>Initialize</i>—define $\mathbf{s}_{\text{obs}} = S(\mathbf{y}_{\text{obs}})$</p> <p><i>Iterate</i>—for $i = 1, \dots, N$:</p> <p>step 1, simulate θ_i from $g(\theta)$; step 2, simulate \mathbf{y}_{sim} from $\pi(\mathbf{y} \theta_i)$, and calculate $\mathbf{s} = S(\mathbf{y}_{\text{sim}})$; step 3, with probability $K\{(\mathbf{s} - \mathbf{s}_{\text{obs}})/h\}$ set $w_i = \pi(\theta_i)/g(\theta_i)$; otherwise set $w_i = 0$</p> <p><i>Output</i>—a set of parameter values $\{\theta_i\}_{i=1}^N$ and corresponding weights $\{w_i\}_{i=1}^N$</p>	
--	--

step 3: if K is uniform, generate u from $\text{Unif}[0,1]$, and accept θ_i and set w_i if $u \leq K\{(\mathbf{s} - \mathbf{s}_{\text{obs}})/h\}$

construct Monte Carlo algorithms which approximate the ABC posterior and only require the ability to simulate from $\pi(\mathbf{y}|\theta)$.

For simplicity of the future exposition, we shall assume that $\max\{K(\mathbf{x})\} = 1$. This assumption imposes no restriction, since, if $K_0(\mathbf{x})$ is a density kernel, then so is $K(\mathbf{x}) = h_0^{-d} K_0(\mathbf{x}/h_0)$ for any $h_0 > 0$. Thus we can choose h_0 so that $\max\{K(\mathbf{x})\} = 1$. Note that a bandwidth λ for kernel $K_0(\mathbf{x})$ is equivalent to a bandwidth $h = \lambda h_0$ for $K(\mathbf{x})$, so the value of h_0 just redefines the unit of the bandwidth h .

One algorithm for approximating the ABC posterior (1), based on importance sampling is given by algorithm 1 (Table 1). For alternative importance sampling approaches, based on sequential Monte Carlo sampling, see Beaumont *et al.* (2009), Sisson *et al.* (2007) and Toni *et al.* (2009). A standard importance sampler for approximating the ABC posterior would repeatedly simulate a parameter value θ' from a proposal $g(\theta)$ and then assign it an importance sampling weight,

$$\frac{\pi(\theta')}{g(\theta')} \int \pi(\mathbf{y}|\theta') K\left\{\frac{S(\mathbf{y}) - \mathbf{s}_{\text{obs}}}{h}\right\} d\mathbf{y}. \quad (2)$$

This is not possible here, as $\pi(\mathbf{y}|\theta)$ is intractable, so algorithm 1 introduces an extra Monte Carlo step. This step first simulates \mathbf{y}_{sim} from $\pi(\mathbf{y}|\theta')$ and then accepts this value with probability $K\{[S(\mathbf{y}_{\text{sim}}) - \mathbf{s}_{\text{obs}}]/h\}$. Accepted values are assigned weights $\pi(\theta')/g(\theta')$. The key thing is that the expected value of such weights is just expression (2), which is all that is required for this to be a valid importance sampling algorithm targeting expression (1).

The output of algorithm 1 is a weighted sample of θ -values, which approximates the ABC posterior of θ . Many of the weights will be 0, and in practice we would remove the corresponding θ -values from the sample. A specific case of algorithm 1 occurs when $g(\theta) = \pi(\theta)$, and then we have a rejection sampling algorithm. The most common implementation of ABC has a deterministic accept–reject decision in step 3, and this corresponds to $K(\cdot)$ being the density of a uniform random variable—the support of the uniform random variable defining the values of $\mathbf{s} - \mathbf{s}_{\text{obs}}$ which are accepted.

An alternative Monte Carlo procedure for implementing ABC is based on Markov chain Monte Carlo (MCMC) sampling (Marjoram *et al.*, 2003; Bortot *et al.*, 2007) and given in algorithm 2 (Table 2). Both this and algorithm 1 target the same ABC posterior (for a proof of the validity of this algorithm see Sisson *et al.* (2010)).

Good implementation of ABC requires a trade-off between the approximation error between the ABC posterior and the true posterior, and the Monte Carlo approximation error of the ABC posterior. The latter of these will also be affected by the algorithm that is used to implement

1. approximation error between ABC posterior and true posterior
2. MC approximation error of ABC posterior: depend on which algo to use(importance?M-H?MCMC?, then depends on the average acceptance probability below

Table 2. Algorithm 2: MCMC sampling implementation of ABC

Input—a set of data \mathbf{y}_{obs} , and a function $S(\cdot)$;
 a density kernel $K(\cdot)$, with $\max\{K(\mathbf{x})\} = 1$ and a bandwidth $h > 0$;
 a transition kernel $g(\cdot|\cdot)$;
 an integer $N > 0$
Initialize—define $\mathbf{s}_{\text{obs}} = S(\mathbf{y}_{\text{obs}})$, and choose or simulate θ_0 and \mathbf{s}_0
Iterate—for $i = 1, \dots, N$:
 step 1, simulate θ from $g(\theta|\theta_{i-1})$;
 step 2, simulate \mathbf{y}_{sim} from $\pi(\mathbf{y}|\theta)$, and calculate $\mathbf{s} = S(\mathbf{y}_{\text{sim}})$;
 step 3, with probability

$$\min \left[1, \frac{K\{(\mathbf{s} - \mathbf{s}_{\text{obs}})/h\}}{K\{(\mathbf{s}_{i-1} - \mathbf{s}_{\text{obs}})/h\}} \frac{\pi(\theta) g(\theta_{i-1}|\theta)}{\pi(\theta_{i-1}) g(\theta|\theta_{i-1})} \right]$$

accept θ and set $\theta_i = \theta$ and $\mathbf{s}_i = \mathbf{s}$; otherwise set $\theta_i = \theta_{i-1}$ and $\mathbf{s}_i = \mathbf{s}_{i-1}$
Output—a set of parameter values $\{\theta_i\}_{i=1}^N$.

ABC, and the specific implementation of that algorithm. We shall consider both of these approximations in turn.

To understand the former approximation, consider the following alternative expression of the ABC posterior as a continuous mixture of posteriors. Consider the random variable $\mathbf{S} = S(\mathbf{Y})$, which is just the summary statistic of the data. Denote the posterior for θ given $\mathbf{S} = \mathbf{s}$ by $\pi(\theta|\mathbf{s})$, and the marginal density for \mathbf{S} by $\pi(\mathbf{s}) := \int \pi(\mathbf{s}|\theta) \pi(\theta) d\theta$. Then

$$\pi_{\text{ABC}}(\theta|\mathbf{s}_{\text{obs}}) = \int \beta(\mathbf{s}) \pi(\theta|\mathbf{s}) d\mathbf{s}, \quad \beta(\mathbf{s}) = \frac{\pi(\mathbf{s}) K\{(\mathbf{s} - \mathbf{s}_{\text{obs}})/h\}}{\int \pi(\mathbf{s}) K\{(\mathbf{s} - \mathbf{s}_{\text{obs}})/h\} d\mathbf{s}}. \quad (3)$$

the mixing weight

The mixing weight for a given value of \mathbf{S} is just the conditional density for such a value given acceptance in step 3 of algorithm 1. If h is small then $\pi_{\text{ABC}}(\theta|\mathbf{s}_{\text{obs}}) \approx \pi(\theta|\mathbf{s}_{\text{obs}})$.

can split the approximation of the posterior by ABC into the approximation of $\pi(\theta|\mathbf{y}_{\text{obs}})$ by $\pi(\theta|\mathbf{s}_{\text{obs}})$, and the further approximation of $\pi(\theta|\mathbf{s}_{\text{obs}})$ by $\pi_{\text{ABC}}(\theta|\mathbf{s}_{\text{obs}})$. The former is controlled by the choice of $S(\cdot)$; the latter by $K(\cdot)$ and h . two sources or approx error: one from how sufficient the summary stat is, and one from how ABC posterior is close to the true posterior

Now consider the Monte Carlo approximation within importance sampling ABC. Consider a scalar function $a(\theta)$, and define the ABC posterior mean of $a(\theta)$ as

$$E_{\text{ABC}}\{a(\theta)|\mathbf{s}_{\text{obs}}\} = \int a(\theta) \pi_{\text{ABC}}(\theta|\mathbf{s}_{\text{obs}}) d\theta.$$

Assuming that this exists then, by the law of large numbers, as $N \rightarrow \infty$

$$\frac{\sum_{i=1}^N w_i a(\theta_i)}{\sum_{i=1}^N w_i} \rightarrow E_{\text{ABC}}\{a(\theta)|\mathbf{s}_{\text{obs}}\}, \quad (4)$$

where the w_i s are the importance sampling weights, and convergence is in probability. For rejection sampling, for large N the variance of this estimator is just

$$\text{var}_{\text{ABC}}\{a(\theta)|\mathbf{s}_{\text{obs}}\} / N_{\text{acc}}, \quad (5)$$

where the numerator is the ABC posterior variance of $a(\theta)$, and the denominator is $N_{\text{acc}} = N \int p(\theta|\mathbf{s}_{\text{obs}}) \pi(\theta) d\theta$, the expected number of acceptances.

Similar calculations for both importance sampling in general and for the MCMC version of ABC are given in Appendix A. In all cases the Monte Carlo error depends on $\int p(\theta|\mathbf{s}_{\text{obs}}) \pi(\theta) d\theta$, the average acceptance probability of the rejection sampling algorithm. The following lemma

characterizes this acceptance probability for small h in this case of continuous summary statistics (similar results can be shown for discrete summary statistics).

Lemma 1. Assume that either

- (a) the marginal distribution for the summary statistics $\pi(\mathbf{s})$ is continuous at $\mathbf{s} = \mathbf{s}_{\text{obs}}$ and that the kernel $K(\cdot)$ has finite support or
- (b) $\pi(\mathbf{s})$ is continuously differentiable, $|\partial\pi(\mathbf{s})/\partial s_i|$ is bounded above for all i and $\int |x_i| K(\mathbf{x}) d\mathbf{x}$ is bounded for all i . Then, in the limit as $h \rightarrow 0$,

$$\int p(\theta|\mathbf{s}_{\text{obs}}) \pi(\theta) d\theta = \pi(\mathbf{s}_{\text{obs}})h^d + o(h^d), \quad (6)$$

where d is the dimension of \mathbf{s}_{obs} .

For a proof of lemma 1, see Appendix B.

This result gives insight into how $S(\cdot)$ and h affect the Monte Carlo error. To minimize Monte Carlo error, we need h^d to be not too small. Thus ideally we want $S(\cdot)$ to be a low dimensional summary of the data that is sufficiently informative about θ that $\pi(\theta|\mathbf{s}_{\text{obs}})$ is close, in some sense, to $\pi(\theta|\mathbf{y}_{\text{obs}})$. The choice of h affects the accuracy of π_{ABC} in approximating $\pi(\theta|\mathbf{s}_{\text{obs}})$, but also the average acceptance probability (6), and hence the Monte Carlo error.

1.3. Approach and outline

Previous justification for ABC has, at least informally, been based around π_{ABC} approximating $\pi(\theta|\mathbf{y}_{\text{obs}})$ globally. This is possible if \mathbf{y}_{obs} is low dimensional, and $S(\cdot)$ is the identity function, or if we have low dimensional sufficient statistics for the data. In these cases we can control the error in the ABC approximation by choosing h sufficiently small. In general applications this is not so. Arguments have been made about trying to choose approximate sufficient statistics (see for example Joyce and Marjoram (2008)). However, the definition of such statistics is not clear, and more importantly it is difficult or impossible to construct a general method for finding such statistics.

We take a different approach and weaken the requirement for π_{ABC} to be a good approximation to $\pi(\theta|\mathbf{y}_{\text{obs}})$. We argue for π_{ABC} to be a good approximation solely in terms of the accuracy of certain estimates of the parameters. We also would like to know how to interpret the ABC posterior and probability statements that are derived from it. To do this we consider a property that we call calibration, which we define formally below. If π_{ABC} is calibrated, then this means that probability statements that are derived from it are appropriate, and in particular that we can use π_{ABC} to quantify uncertainty in estimates. We argue that using such criteria we can construct ABC posteriors which have both good inferential properties, and which can be estimated well by using Monte Carlo methods such as in algorithm 1.

In Section 2 we define formally what we mean by calibration and accuracy. Standard ABC is not calibrated, but a simple modification, which we call *noisy ABC*, is. We further show that if the bandwidth h is small then standard ABC is approximately calibrated. Theoretical results are used to produce recommendations about when standard ABC and when noisy ABC should be used. Then, for a certain definition of accuracy we show that the optimal choice of summary statistics $S(\mathbf{Y})$ are the true posterior means of the parameters. Although these are unknown, simulation can be used to estimate the summary statistics. Our approach to doing this is described in Section 3, together with results that show the advantage of ABC over directly using the summary statistics to estimate parameters. Section 4 gives examples comparing our implementation of ABC with previous methods. The paper ends with a discussion.

2. Calibration and accuracy of approximate Bayesian calibration

First we define what we mean by calibration and introduce a version of ABC that is calibrated. We show how the idea of calibration can be particularly important when analysing multiple data sets. We then discuss our definition of accuracy of the ABC posterior, and how this can be used to guide the choice of summary statistic that we use.

2.1. Calibration and noisy approximate Bayesian calibration

Consider a subset of the parameter space \mathcal{A} . For given data \mathbf{y}_{obs} , the ABC posterior will assign a probability to the event $\theta \in \mathcal{A}$

$$\Pr_{\text{ABC}}(\theta \in \mathcal{A} | \mathbf{s}_{\text{obs}}) = \int_{\mathcal{A}} \pi_{\text{ABC}}(\theta | \mathbf{s}_{\text{obs}}) d\theta.$$

For a given probability q , consider the event $E_q(\mathcal{A})$ that $\Pr_{\text{ABC}}(\theta \in \mathcal{A} | \mathbf{s}_{\text{obs}}) = q$. Then the ABC posterior is calibrated if

$$\Pr\{\theta \in \mathcal{A} | E_q(\mathcal{A})\} = q. \quad (7)$$


The probability is then defined in terms of the density on parameters and data given by our prior and likelihood model, $\pi(\theta) \pi(\mathbf{y} | \theta)$, but ignoring any Monte Carlo randomness. Statement (7) states that, under repeated sampling from the prior, data and summary statistics, events assigned probability q by the ABC posterior will occur with probability q . A consequence of calibration is that the ABC posterior will appropriately represent uncertainty in the parameters: for example we can construct appropriate credible intervals, i.e. calibration means that we can use the ABC posterior as we would any standard posterior distribution.

Standard ABC posteriors (3) are not calibrated in general. Instead we introduce the idea of *noisy ABC* which is calibrated. Noisy ABC involves defining summary statistics which are random. A noisy ABC importance sampling algorithm is obtained by changing the initialization within algorithm 1 or algorithm 2 to 'simulate \mathbf{x} from $K(\mathbf{x})$ and define $\mathbf{s}_{\text{obs}} = S(\mathbf{y}_{\text{obs}}) + h\mathbf{x}$ ', giving algorithm 3. The resulting ABC posterior for a given \mathbf{y}_{obs} is random, and the definition of the probability in equation (7) needs to account for this extra randomness.

Theorem 1. Algorithm 3 produces an ABC posterior that is calibrated.

Proof. The ABC posterior that is derived by algorithm 3 is $\pi_{\text{ABC}}(\theta | \mathbf{s}_{\text{obs}})$ where \mathbf{s}_{obs} is related to the data \mathbf{y}_{obs} by

$$\mathbf{s}_{\text{obs}} = S(\mathbf{y}_{\text{obs}}) + h\mathbf{x}, \quad (8)$$

and \mathbf{x} is the realization of a random variable with density $K(\mathbf{x})$. However, the definition of $\pi_{\text{ABC}}(\theta | \mathbf{s}_{\text{obs}})$ is just that of the true posterior for θ given data \mathbf{s}_{obs} generated by equation (8). It immediately follows that this density is calibrated. 

A related idea was used by Wilkinson (2008) who showed that the ABC posterior is equivalent to the true posterior under an assumption of appropriate model error. In the limit as $h \rightarrow 0$ noisy ABC is equivalent to standard ABC; we discuss the links between the two in more detail in Section 2.4.

2.2. Inference from multiple sources of data

Consider combining data from m independent sources, $\mathbf{y}_{\text{obs}}^{(1)}, \dots, \mathbf{y}_{\text{obs}}^{(m)}$. It is possible to use

individual ABC analyses for each data set, and then to combine these inferences. One sequential approach is to use the ABC posterior after analysing the i th data set as a prior for analysing the $(i+1)$ th data set. Algorithms for such an approach are a special case of those discussed in Wilkinson (2011).

One consequence of calibration is that such inferences will be well behaved in the limit as m becomes large. To see this define the ABC approximation to the i th data set as $p(\theta|\mathbf{s}_{\text{obs}}^{(i)})$, and note that the above approach is targeting the following ABC posterior:

$$\pi_{\text{ABC}}(\theta|\mathbf{s}_{\text{obs}}^{(1)}, \dots, \mathbf{s}_{\text{obs}}^{(m)}) \propto \pi(\theta) \prod_{i=1}^m p(\theta|\mathbf{s}_{\text{obs}}^{(i)}).$$

If we use noisy ABC, where $\mathbf{s}_{\text{obs}}^{(i)}$ is randomly centred on $S(\mathbf{y}_{\text{obs}}^{(i)})$, then by the same argument as for theorem 1 this ABC posterior will be calibrated. Furthermore, we have the following result.

Theorem 2. Let θ_0 be the true parameter value. Consider noisy ABC, where $\mathbf{s}_{\text{obs}} = S(\mathbf{y}_{\text{obs}}) + h\mathbf{x}$, where \mathbf{x} is drawn from $K(\cdot)$. Then the expected noisy ABC log-likelihood,

$$E[\log\{p(\theta|\mathbf{S}_{\text{obs}})\}] = \iint \log[p\{\theta|S(\mathbf{y}) + h\mathbf{x}\}] \pi(\mathbf{y}|\theta_0) K(\mathbf{x}) d\mathbf{y} d\mathbf{x},$$

has its maximum at $\theta = \theta_0$.

Proof. Make the change of variable $\mathbf{s}_{\text{obs}} = S(\mathbf{y}) + h\mathbf{x}$. Then

$$E[\log\{p(\theta|\mathbf{S}_{\text{obs}})\}] = \frac{1}{h^d} \iint \log\{p(\theta|\mathbf{s}_{\text{obs}})\} \pi(\mathbf{y}|\theta_0) K\left\{\frac{\mathbf{s}_{\text{obs}} - S(\mathbf{y})}{h}\right\} d\mathbf{y} d\mathbf{s}_{\text{obs}}.$$

Now by definition $p(\theta|\mathbf{s}_{\text{obs}}) = \int \pi(\mathbf{y}|\theta) K[\{\mathbf{s}_{\text{obs}} - S(\mathbf{y})\}/h] d\mathbf{y}$, and thus we obtain

$$E[\log\{p(\theta|\mathbf{S}_{\text{obs}})\}] = \frac{1}{h^d} \iint \log\{p(\theta|\mathbf{s}_{\text{obs}})\} p(\theta_0|\mathbf{s}_{\text{obs}}) d\mathbf{s}_{\text{obs}}.$$

By Jensen's inequality this has its maximum at $\theta = \theta_0$. □

The importance of this result is that, under the standard regularity conditions (Bernardo and Smith, 1994), the noisy ABC posterior will converge onto a point mass on the true parameter value as $m \rightarrow \infty$. By comparison, if we use standard ABC, then we have no such guarantee. As a simple example assume that $Y^{(i)}$ are independent and identically distributed from a normal distribution with mean 0 and variance σ^2 , and our kernel is chosen to be normal with variance $\tau < \sigma^2$. If we use standard ABC, the ABC posterior given $y^{(1)}, \dots, y^{(m)}$ will converge to a point mass on $\sigma^2 - \tau$ as $m \rightarrow \infty$.

We look at this issue empirically in Section 4.3.

2.3. Accuracy and choice of summary statistics

Calibration itself is not sufficient to define a sensible ABC posterior. For example the prior distribution is always calibrated but will not give accurate estimates of parameters. Thus we also want to maximize the accuracy of estimates based on the ABC posterior. We shall define accuracy in terms of a loss function for estimating the parameters. A natural choice of loss function is quadratic loss. Let θ_0 be the true parameter values, and $\hat{\theta}$ an estimate. Then we shall consider the class of loss functions, which is defined in terms of a $p \times p$ positive definite matrix A ,

$$L(\theta_0, \hat{\theta}; A) = (\theta_0 - \hat{\theta})^T A (\theta_0 - \hat{\theta}).$$



We now consider implementing ABC to minimize this quadratic error loss of estimating the parameters. We consider the limit of $h \rightarrow 0$. This will give results that define the optimal choice of summary statistics, $S(\cdot)$.

For any choice of weight matrix A that is of full rank, the following theorem shows that the optimal choice of summary statistics is $S(\mathbf{y}_{\text{obs}}) = E(\theta|\mathbf{y}_{\text{obs}})$, which is the true posterior mean.

Theorem 3. Consider a $p \times p$ positive definite matrix A of full rank. Given observation \mathbf{y}_{obs} , let Σ be the true posterior variance for θ .

- (a) The minimal possible quadratic error loss $E\{L(\theta, \hat{\theta}; A)|\mathbf{y}_{\text{obs}}\}$ occurs when $\hat{\theta} = E(\theta|\mathbf{y}_{\text{obs}})$ and is $\text{tr}(A\Sigma)$.
- (b) If $S(\mathbf{y}_{\text{obs}}) = E(\theta|\mathbf{y}_{\text{obs}})$ then in the limit as $h \rightarrow 0$ the minimum loss, based on inference using the ABC posterior, is achieved by $\hat{\theta} = E_{\text{ABC}}(\theta|\mathbf{s}_{\text{obs}})$. The resulting expected loss is $\text{tr}(A\Sigma)$.

Proof. Part (a) is a standard result of Bayesian statistics (Bernardo and Smith, 1994). For (b) we just need to show that, in the limit as $h \rightarrow 0$, $E_{\text{ABC}}(\theta|\mathbf{s}_{\text{obs}}) = E(\theta|\mathbf{y}_{\text{obs}})$. By definition, in the limit as $h \rightarrow 0$, $\mathbf{s}_{\text{obs}} = S(\mathbf{y}_{\text{obs}})$ with probability 1, and $\pi_{\text{ABC}}(\theta|\mathbf{s}_{\text{obs}}) = \pi(\theta|\mathbf{s}_{\text{obs}})$. Furthermore

$$\begin{aligned} E_{\text{ABC}}(\theta|\mathbf{s}_{\text{obs}}) &= \int \theta \pi(\theta|\mathbf{s}_{\text{obs}}) d\theta, \\ &= \iint \theta \pi(\theta|\mathbf{y}) \pi(\mathbf{y}|\mathbf{s}_{\text{obs}}) d\mathbf{y} d\theta, \end{aligned}$$

where $\pi(\mathbf{y}|\mathbf{s}_{\text{obs}})$ is the conditional distribution of the data \mathbf{y} given the summary statistic \mathbf{s}_{obs} . Finally by definition, all \mathbf{y} that are consistent with \mathbf{s}_{obs} satisfy $\int \theta \pi(\theta|\mathbf{y}) d\theta = E(\theta|\mathbf{y}_{\text{obs}})$, and hence the result follows. \square

Use of squared error loss leads to ABC approximations that attempt to have the same posterior mean as the true posterior. Using alternative loss functions would mean matching other features of the posterior: for example absolute error loss would result in matching the posterior medians. It is possible to choose other summary statistics that also achieve the minimum expected loss. However, any such statistic with dimension $d > p$ will cause larger Monte Carlo error (see lemma 1 and the discussion below).

2.4. Comparison of standard approximate Bayesian computation and noisy approximate Bayesian computation

Standard and noisy ABC are equivalent in the limit as $h \rightarrow 0$, with the ABC posteriors converging to $E\{\theta|\mathbf{S}(\mathbf{y}_{\text{obs}})\}$. We can further quantify the accuracy of estimates based on standard or noisy ABC for $h \approx 0$. For noisy ABC we have the following result.

Theorem 4. Assume condition (a) of lemma 1, that $\pi\{E(\theta|\mathbf{y}_{\text{obs}})\} > 0$, and the kernel $K(\cdot)$ corresponds to a random variable with mean 0. If $\mathbf{S}(\mathbf{y}_{\text{obs}}) = E(\theta|\mathbf{y}_{\text{obs}})$ then for small h the expected quadratic loss that is associated with $\hat{\theta} = E_{\text{ABC}}(\theta|\mathbf{s}_{\text{obs}})$ is

$$E\{L(\theta, \hat{\theta}; A)|\mathbf{y}_{\text{obs}}\} = \text{tr}(A\Sigma) + h^2 \int \mathbf{x}^T A \mathbf{x} K(\mathbf{x}) d\mathbf{x} + o(h^2).$$

Proof. The idea is that $E_{\text{ABC}}(\theta|\mathbf{s}_{\text{obs}}) = \mathbf{s}_{\text{obs}} + o(h)$, and the squared error loss based on $\hat{\theta} = \mathbf{s}_{\text{obs}}$ is just $\text{tr}(A\Sigma) + h^2 \int \mathbf{x}^T A \mathbf{x} K(\mathbf{x}) d\mathbf{x}$. See Appendix C. \square

A similar result exists for standard ABC (Prangle, 2011), which shows that in this case

$$E\{L(\theta, \hat{\theta}; A) | \mathbf{y}_{\text{obs}}\} = \text{tr}(A\Sigma) + O(h^4).$$

Extensions of these results can be used to give guidance on the choice of kernel. For noisy ABC they suggest a uniform kernel on the ellipse $\mathbf{x}^T A \mathbf{x} < c$, for some c ; for standard ABC they also suggest a uniform kernel on an ellipse, but the form of the ellipse is difficult to calculate in practice. We do not give these results in more detail, as in practice we have found that the choice of kernel has relatively little effect on the accuracy of either ABC algorithm.

These two results also give an insight into the overall accuracy of a Monte Carlo ABC algorithm (see also Blum (2010)) by using the following informal argument. For simplicity consider a rejection sampling algorithm. The Monte Carlo variance is inversely proportional to the acceptance probability. Thus using lemma 1 we have that the Monte Carlo variance is $O(N^{-1}h^{-d})$, where N is the number of proposals. The expected quadratic loss based on estimates from the ABC importance sampler will be increased by this amount. Thus for noisy ABC we want to choose h to minimize

$$\text{tr}(A\Sigma) + h^2 \int \mathbf{x}^T A \mathbf{x} K(\mathbf{x}) d\mathbf{x} + \frac{C_0}{Nh^d},$$

for some constant C_0 . This gives that we want $h = O(N^{-1/(2+d)})$, and the overall expected loss above $\text{tr}(A\Sigma)$ would then decay as $N^{-2/(2+d)}$. For standard ABC a similar argument gives $h = O(N^{-1/(4+d)})$, and the overall expected loss would then decay as $N^{-4/(4+d)}$.

Thus we can see that the choice between using standard ABC and noisy ABC is a choice of a trade-off between accuracy and calibration. Noisy ABC is calibrated, but for small h will give less accurate estimates. As such for the analysis of a single data set where the number of summary statistics is not too large, and hence h is small, we recommend the use of standard ABC. If we wish to combine inferences from ABC analyses of multiple data sets, then, in the light of the discussion in Section 2, we recommend noisy ABC. For all the examples in Section 4 we found that this approach worked well in practice.

Possibly the best approach is to use noisy ABC, but to use **Rao–Blackwellization** ideas to average out the noise that is added to the summary statistics. Such an approach would have the guarantee that the resulting expected quadratic loss for estimating any function of the parameters would be smaller than that from noisy ABC. However, implementing such a Rao–Blackwellization scheme efficiently appears non-trivial, with the only simple approach being to run noisy ABC independently on the same data set, and then to average the estimates across each of these runs.

3. Semi-automatic approximate Bayesian computation

The above theory suggests that we wish to choose summary statistics that are equal to posterior means. Although we cannot use this result directly, as we cannot calculate the posterior means, we can use simulation to estimate appropriate summary statistics.

Our approach is

- (a) to use a pilot run of ABC to determine a region of non-negligible posterior mass,
- (b) to simulate sets of parameter values and data,
- (c) to use the simulated sets of parameter values and data to estimate the summary statistics and
- (d) to run ABC with this choice of summary statistics.

Step (a) of this algorithm is optional. Its aim is to help to define an appropriate training region

of parameter space from which we should simulate parameter values. In applications where our prior distributions are relatively informative, this step should be avoided as we can simulate parameter values from the prior in step (b). However, it is important if we have uninformative priors, particularly if they are improper.

If we implement step (a), we assume that we have arbitrarily chosen some summary statistics to use within ABC. In our implementation below we choose our training region as a hypercube, with the range for each parameter being the range of that parameter observed within our pilot run. Then in step (b) we simulate parameter values from the prior truncated to this training region, and for each choice of parameter value we simulate an artificial data set. We repeat this M times, so that we have M sets of parameter values, each with a corresponding simulated data set.

There are various approaches that we can take for step (c). In practice we found that using linear regression, with appropriate functions of the data as predictors, is both simple and worked well. We also considered using the lasso (Hastie *et al.*, 2001) and canonical correlation analysis (Mardia *et al.*, 1979) but in general neither of these performed better than linear regression (though the lasso may be appropriate if we wish to use a large number of explanatory variables within the linear model).

Our linear regression approach involved considering each parameter in turn. First we introduce a vector-valued function $f(\cdot)$, so that $f(\mathbf{y})$ is a vector of, possibly non-linear, transformations of the data. The simplest choice is $f(\mathbf{y}) = \mathbf{y}$, but in practice including other or different transformations as well may be beneficial. For example, in one application below we found $f(\mathbf{y}) = (\mathbf{y}, \mathbf{y}^2, \mathbf{y}^3, \mathbf{y}^4)$, i.e. a vector of length $4n$ that consists of the data plus all second, third and fourth powers of individual data points, produced a better set of summary statistics.

For the i th summary statistic the simulated values of the i th parameter, $\theta_i^{(1)}, \dots, \theta_i^{(M)}$, are used as the responses; and the transformations of the simulated data, $f(\mathbf{y}^{(1)}), \dots, f(\mathbf{y}^{(M)})$, are used as the explanatory variables. We then fit the model

$$\theta_i = E(\theta_i | \mathbf{y}) + \varepsilon_i = \beta_0^{(i)} + \beta^{(i)} f(\mathbf{y}) + \varepsilon_i,$$

where ε_i is some zero-mean noise, using least squares. The fitted function $\hat{\beta}_0^{(i)} + \hat{\beta}^{(i)} f(\mathbf{y})$ is then an estimate of $E(\theta_i | \mathbf{y})$. The constant terms can be neglected in practice as ABC uses only the difference in summary statistics. Thus the i th summary statistic for ABC is just $\hat{\beta}^{(i)} f(\mathbf{y})$.

Our approach of using a training region means that our models for the posterior means are based on only parameter values that are simulated within this region. We therefore suggest adapting the ABC run in step (d) so that the prior is truncated to lie within this training region (a similar idea was used in Blum and François (2010)). This can be viewed as using, weakly, the information that we have from the pilot ABC run within the final ABC run and has links with composite likelihood methods (Lindsay, 1988). More importantly it makes the overall algorithm robust to problems where $E\{\hat{\beta}^{(i)} f(\mathbf{Y}) | \theta_i\}$ is similar for two dissimilar values of θ_i : one inside the training region and one outside.

In practice below we use roughly a quarter of our total central processor unit time on steps (a) and (b) and half on step (d), with step (c) having negligible central processor unit cost. We call this semi-automatic ABC as the choice of summary statistics is now based on simulation, but there are still choices by the user in terms of fitting the linear model in step (c). This input is in terms of the choice of $f(\mathbf{y})$ to be used. Step (c) is now a familiar statistical problem, and standard model checks can be used to decide whether that choice of $f(\mathbf{y})$ is appropriate and, if not, how it could be improved. Also, repeating step (c) with a different choice of $f(\mathbf{y})$ can be done without any further simulation of data and thus is quick in terms of the central processor unit cost. Furthermore standard model comparison procedures (e.g. using the Bayesian

information criterion BIC) can be used to choose between summary statistics that are obtained from linear regressions using different explanatory variables.

A natural question is whether this approach is better than the current approach to ABC, where summary statistics are chosen arbitrarily. In our implementation we still need to choose summary statistics for step (a), and we also need to choose the set of explanatory variables for the linear model. Thus it could be argued that all we have done is to replace one arbitrary choice with another. However, we believe that our approach is more robust to these choices than standard ABC is. First the choice of summary statistics in step (a) is purely to make step (b) more efficient, and as such the final results depend little on this choice. Secondly, when we choose the explanatory variables we can choose many such variables (of the order of hundreds). As such we are much more likely to include among these some variables which are informative about the parameters of interest than standard ABC is where generally a few summary statistics are used. If many summary statistics are used in ABC, then this will require a large value of h and will often be inefficient because the accept–reject decision is based not only on the informative summary statistics, but also those which are less informative. These issues are demonstrated empirically in the examples that we consider.

Our approach has similarities to that of Beaumont *et al.* (2002) (see also Blum and François (2010)), who used linear regression to correct the output from ABC. The key difference is that our approach uses linear regression to construct the summary statistics, whereas Beaumont *et al.* (2002) used linear regression to reduce the error between $\pi_{\text{ABC}}(\theta|\mathbf{s}_{\text{obs}})$ and $\pi(\theta|\mathbf{s}_{\text{obs}})$. In particular the method of Beaumont *et al.* (2002) assumes that appropriate low dimensional summary statistics have already been chosen. We look at differences between our approach and that of Beaumont *et al.* (2002) empirically in the examples.

3.1. Why use approximate Bayesian computation?

Our approach involves using simulation to find estimates of the posterior mean of each parameter. A natural question is why not use these estimates directly? We think that using ABC has two important advantages over just using these estimates directly. The first is that ABC gives you a posterior distribution, and thus you can quantify uncertainty in the parameters as well as obtain point estimates.

Moreover we have the following result.

Theorem 5. Let $\tilde{\theta} = E\{\theta|S(\mathbf{y}_{\text{obs}})\}$. Then, for any function g ,

$$E\{L(\theta, \tilde{\theta}; A)|S(\mathbf{y}_{\text{obs}})\} \leq E(L[\theta, g\{S(\mathbf{y}_{\text{obs}})\}; A]|S(\mathbf{y}_{\text{obs}})).$$

Furthermore, asymptotically as $h \rightarrow 0$ the ABC posterior mean estimate of θ is optimal among estimates based on $S(\mathbf{y}_{\text{obs}})$.

Proof. The proof of the first part is the standard argument that the mean is the optimal estimator under quadratic loss (Bernardo and Smith, 1994). The second part follows because as $h \rightarrow 0$ the ABC posterior mean tends to $\tilde{\theta}$. \square

Note that this result states that, in the limit as $h \rightarrow 0$, ABC gives estimates that are at least as accurate as or more accurate than any other estimators based on the same summary statistics.

3.1.1. Comparison with indirect inference

Indirect inference (Gourieroux and Ronchetti, 1993) is a method that is similar to ABC in that it uses simulation from a model to produce estimates of the model's parameters. The general

procedure involves first analysing the data under an approximating model and estimating the parameters, called auxiliary parameters, for this model. Then data are simulated for a range of parameter values, and for each simulated data set we obtain an estimate of the auxiliary parameters. Finally we estimate the true parameters on the basis of which parameter values produced estimates of the auxiliary parameters that are closest to those estimated from the true data. (In practice we simulate multiple data sets for each parameter value and obtain an estimate of the auxiliary parameters on the basis of these multiple data sets.) The link to ABC is that the auxiliary parameters in indirect inference are equivalent to the summary statistics in ABC. Both methods then use (different) simulation approaches to produce estimates of the true parameters from the values of the auxiliary parameters (or summary statistics) for the real data.

For many approximating models, the auxiliary parameters depend on a small set of summary statistics of the data; these were called auxiliary statistics in Heggland and Frigessi (2004). In these cases indirect inference is performing inference based on these auxiliary statistics. The above result shows that, in the limit as $h \rightarrow 0$, ABC will be more accurate than an indirect inference method whose auxiliary statistics are the same as the summary statistic that is used for ABC. We investigate this empirically in Section 4.

4. Examples

The performance of semi-automatic ABC was investigated in a range of examples: independent draws from a complex distribution (Section 4.2), a stochastic kinetic network for biochemical reactions (Section 4.3), a partially observed $M/G/1$ -queue (Section 4.5), an ecological population size model (Section 4.4) and a model for the transmission of tuberculosis (Section 4.6). Section 4.1 describes implementation details that are common to all examples. Section 4.3 concerns a data set for which ABC and related methods have not previously been used and highlights the use of noisy ABC in the sequential approach of Section 2.2. The other examples have previous analyses in the literature, and we show that semi-automatic ABC compares favourably against existing methods including indirect inference, the synthetic likelihood method of Wood (2010) and ABC with *ad hoc* summary statistics, with or without the regression correction method of Beaumont *et al.* (2002). We also show that direct use of the linear predictors that are created during semi-automatic ABC can be inaccurate (e.g. Section 4.6). Apart from in Section 4.3 noisy ABC runs are not shown; they are similar to non-noisy semi-automatic ABC but slightly less accurate. The practical details of implementing our method are also explored: in particular how the choice of explanatory variables $f(\mathbf{y})$ is made.

4.1. Implementation details

Apart from the sequential implementation of ABC in Section 4.3, all ABC analyses were performed by using algorithm 2 with a normal transition kernel. The density kernel was uniform on an ellipsoid $\mathbf{x}^T \mathbf{A} \mathbf{x} < c$. This is a common choice in the literature and is close to optimal for runs using semi-automatic ABC summary statistics as discussed in Section 2.4. For summary statistics that are not generated by our method we generally used $\mathbf{A} = \mathbf{I}$. In Section 4.4 a different choice was necessary and is discussed there. For ABC using summary statistics from our method, recall from Section 2.3 that \mathbf{A} defines the relative weighting of the parameters in our loss function. In Sections 4.2 and 4.6 the parameters are on similar scales so we used $\mathbf{A} = \mathbf{I}$. Elsewhere, marginal parameter variances were calculated for the output of each pilot run and the means of these (s_1^2, s_2^2, \dots) taken. A diagonal \mathbf{A} -matrix was formed with i th diagonal entry s_i^{-2} .

Other tuning details that are required by algorithm 2 are the choice of h , the variance matrix of the transition kernel and the starting values of the chain. Where possible, these were chosen by manual experimentation or based on previous analyses (e.g. from pilot runs). Otherwise they were based on a very short ABC rejection sampling analysis. Except where noted otherwise, h was tuned to give an acceptance rate of roughly 1% as this gave reasonable results in the applications that were considered. An alternative would be to use computational methods that try to choose h for each run; see Bortot *et al.* (2007) and Ratmann *et al.* (2007).

In the following examples our method is compared with an ABC analysis with summary statistics based on the existing literature, which is referred to as the 'comparison' analysis. To allow a fair comparison, this uses the same number of simulations as the entire semi-automatic method and a lower acceptance rate: roughly 0.5%.

For simulation studies on multiple data sets, the accuracies of the various analyses were compared as follows. The point estimate for each data set was calculated, and the quadratic loss (9) of each parameter estimate relative to the true parameter value was calculated. We present the mean quadratic losses of the individual parameters. In tables of results we highlight the smaller quadratic losses (all within 10% of the smallest values) by italicizing.

4.2. Inference for g -and- k -distribution

The g -and- k -distribution is a flexibly shaped distribution that is used to model non-standard data through a small number of parameters (Haynes, 1998). It is defined by its inverse distribution function (10), below, but has no closed form density. Likelihoods can be evaluated numerically but this is costly (Rayner and MacGillivray, 2002; Drovandi and Pettitt, 2009). ABC methods are attractive because simulation is straightforward by the inversion method. Here we use the fact that we can calculate the maximum likelihood estimate numerically to compare ABC with a full-likelihood analysis also.

The distribution is defined by

$$F^{-1}(x; A, B, c, g, k) = A + B \left[1 + c \frac{1 - \exp\{-g z(x)\}}{1 + \exp\{-g z(x)\}} \right] \{1 + z(x)^2\}^k z(x) \quad (10)$$

where $z(x)$ is the x th standard normal quantile, A and B are location and scale parameters and g and k are related to skewness and kurtosis. The final parameter c is typically fixed as 0.8, and this is assumed throughout, leaving unknown parameters $\theta = (A, B, g, k)$. The only parameter restrictions are $B > 0$ and $k > -\frac{1}{2}$ (Rayner and MacGillivray, 2002).

Allingham *et al.* (2009) used ABC to analyse a simulated data set of $n = 10^4$ independent draws from the g -and- k -distribution with parameters $\theta_0 = (3, 1, 2, 0.5)$. A uniform prior on $[0, 10]^4$ was used and the summary statistics were the full set of order statistics. We studied multiple data sets of a similar form as detailed below. Our aim is first to show how we can implement semi-automatic ABC in a situation where there are large numbers of possible explanatory variables (just using the order statistics gives 10^4 explanatory variables), and to see how the accuracy of semi-automatic ABC compares with the use of arbitrarily chosen summary statistics in Allingham *et al.* (2009). We also aim to look at comparing semi-automatic ABC with the linear regression correction of Beaumont *et al.* (2002) and with indirect inference.

4.2.1. Comparison of approximate Bayesian computation methods

The natural choice of explanatory variables for this problem is based on the order statistics and also powers of the order statistics. Considering up to the fourth power seems appropriate as informally the four parameters are linked to location, scale, skewness and kurtosis. However,

fitting the linear model with the resulting 4×10^4 explanatory variables is impracticable. As a result we considered using a subset of m evenly spaced order statistics, together with up to l powers of this subset. To choose appropriate values for m and l we fitted linear models with m ranging over a grid of values between 60 and 140, and l ranging between 1 and 4, and we used BIC (averaged across the models for the four parameter values) to choose an appropriate value for m and l . We then used the summary statistics that are obtained from the linear model with this value of m and l in the final run of ABC. For simplicity we did this for the first data set and kept the same value of m and l for analysing all subsequent data sets.

Using subsets of order statistics has computational advantages, as these can be generated efficiently by simulating corresponding standard uniform order statistics by using the exponential spacings method of Ripley (1987) (page 98) and performing inversion by substituting these in equation (10). The cost is linear in the number of order statistics required. Our pilot ABC run used the summary statistics from Allingham *et al.* (2009). Fitting the different linear models added little to the overall computational cost, which is dominated by the simulation of the data sets at the various stages of the procedure.

Our semi-automatic ABC procedure chose $m = 100$ order statistics and $l = 4$. The accuracy of the resulting parameter estimates, measured by squared error loss across implementation of semi-automatic ABC on 50 data sets, is shown in Table 3. For comparison we show results of the ABC method of Allingham *et al.* (2009), implemented to have the same overall computational cost. We also show the accuracy of estimates that were obtained by post-processing the results of Allingham *et al.* (2009) by using the regression correction of Beaumont *et al.* (2002). We could use this regression correction on only 48 of the 50 data sets, as, for the remaining two, there were too few acceptances in the ABC run for the regression correction to be stable (on those two data sets the resulting loss after performing the regression correction was orders of magnitude greater than the original ABC estimates).

Although the analysis of Allingham *et al.* (2009) performed poorly, and hence produced a poor pilot region for semi-automatic ABC, semi-automatic ABC appears to perform well. It has losses that are between a factor of 2 and 100 smaller than for the method of Allingham *et al.* (2009) with or without the regression correction. Using the regression correction does

Table 3. Mean quadratic losses of various ABC analyses of 50 g -and- k data sets with parameters (3,1,2,0.5)[†]

<i>Method</i>	<i>A</i>	<i>B</i>	<i>g</i>	<i>k</i>
Allingham <i>et al.</i> (2009)	0.0059	0.0013	3.85	0.00063
Allingham + regression	0.00040	0.0017	0.28	0.00051
Semi-automatic ABC	<i>0.00016</i>	<i>0.00056</i>	0.044	0.00023
Comparison	0.00025	0.00063	0.0061	0.00041
Comparison + regression	<i>0.00016</i>	<i>0.00055</i>	<i>0.0014</i>	<i>0.00015</i>
Semi-automatic ABC	<i>0.00015</i>	<i>0.00053</i>	<i>0.0014</i>	<i>0.00015</i>
Maximum likelihood estimation	0.00016	0.00055	0.0013	0.00014

[†]The first three rows are based on using the summary statistics of Allingham *et al.* (2009) in ABC, and in the ABC pilot run for semi-automatic ABC. The next three rows use just 100 evenly spaced order statistics. Results based on using the regression correction of Beaumont *et al.* (2002) are denoted ‘regression’. For ‘Allingham + regression’ we give the mean loss for just 48 of the 50 data sets. For the remaining two data sets the number of ABC acceptances was low (about 200), and the regression correction was unstable. For comparison we give the mean quadratic loss of the true maximum likelihood estimates. Mean quadratic losses within 10% of the minimum value for that parameter are italicized.

improve the accuracy of the estimates for three of the four parameters, but to a lesser extent than semi-automatic ABC. For comparison, using the predictors from the linear regression to estimate the parameters directly had similar accuracy for all parameters except g , where the linear predictor's average error was greater by about a third.

To investigate the effect of the pilot run on semi-automatic ABC, and the choice of summary statistics on the regression correction, we repeated this analysis by implementing ABC with 100 order statistics. This greatly improved the performance of ABC, showing the importance of the choice of summary statistics. The improved pilot run also improves semi-automatic ABC but to a much lesser extent. In this case semi-automatic ABC has similar accuracy to that of the comparison ABC run with the regression correction. For a further comparison we also calculated the maximum likelihood estimates for each data set numerically. The two best ABC runs have mean quadratic loss that is almost identical to that of the maximum likelihood estimates.

4.2.2. Comparison with indirect inference

Theorem 5 shows that asymptotically ABC is at least as accurate as other estimators based on the same summary statistics. We tested this with a comparison against indirect inference (see Section 3.1). The semi-automatic ABC analysis was repeated, and, to give a direct comparison, its summary statistics were used as the indirect inference auxiliary statistics.

Initial analysis showed that which method is more accurate depends on the true value of θ and in particular the parameter g ; this is illustrated by Fig. 1. Therefore we studied data sets that were produced from varying g -values; we drew 50 g -values from its prior and for each simulated data sets of n g -and- k -draws conditional on $\theta = (3, 1, g, 0.5)$ for $n = 10^2, 10^3, 10^4$.

Each semi-automatic ABC analysis used a total of 3.1×10^6 simulated data sets. Indirect inference was roughly tuned so that the total number of simulations equalled this. Similar results can be obtained from indirect inference using many fewer simulations, and indirect inference is thus a computationally quicker algorithm. Mean losses are given in Table 4, showing that, although the methods perform similarly for $n = 10^4$, ABC is more accurate for smaller n .

More detail is given in Fig. 1 which plots the true against estimated g -values. Of particular interest is the case $n = 100$ where the g -parameter is very difficult to identify for $g > 3$. It is over this range that ABC outperforms indirect inference most clearly, with estimates from indirect inference being substantially more variable than those for ABC.

Table 4. Mean quadratic losses of semi-automatic ABC and indirect inference analyses of 50 g -and- k data sets with variable g -parameters[†]

n	Method	A	B	g	k
10^4	Pilot	0.0003	0.0008	1.7	0.0004
	Indirect inference	0.0003	0.0022	0.082	0.0063
	Semi-automatic ABC	<i>0.0001</i>	<i>0.0005</i>	<i>0.059</i>	<i>0.0002</i>
10^3	Pilot	0.0031	0.014	4.6	0.0073
	Indirect inference	0.0066	0.014	0.83	0.0053
	Semi-automatic ABC	<i>0.0012</i>	<i>0.0094</i>	<i>0.51</i>	<i>0.0042</i>
10^2	Pilot	0.0089	0.039	4.8	0.057
	Indirect inference	0.018	0.059	5.5	0.067
	Semi-automatic ABC	<i>0.0075</i>	<i>0.046</i>	<i>3.5</i>	<i>0.040</i>

[†]The smallest losses for each parameter and sample size are italicized.

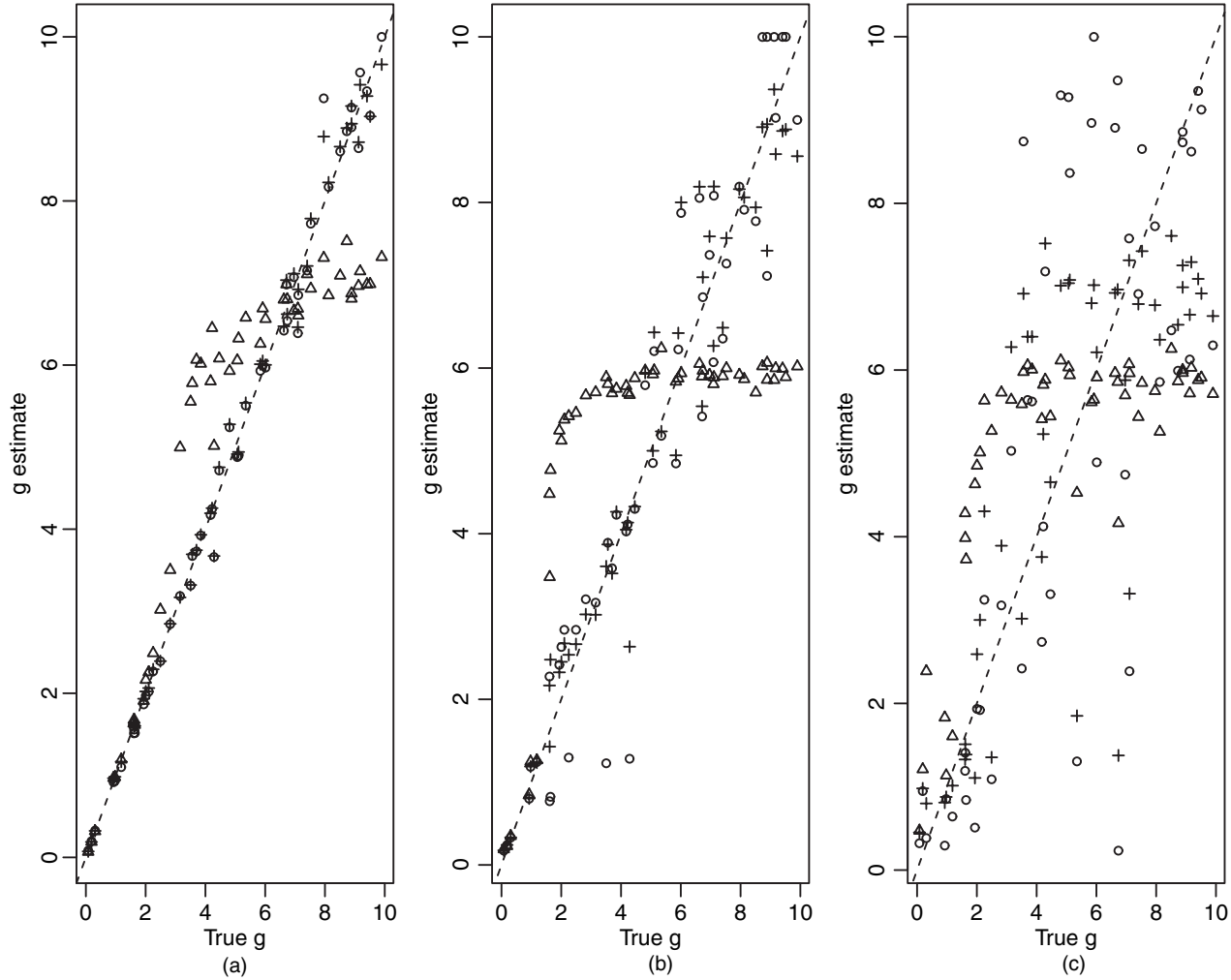


Fig. 1. Estimated g -values from indirect inference (\circ) and semi-automatic ABC pilot (Δ) and final (\times) analyses, plotted against true g -values for 50 g -and- k data sets of sample size n : (a) $n = 10^4$; (b) $n = 10^3$; (c) $n = 10^2$

4.3. Inference for stochastic kinetic networks

Stochastic kinetic networks are used to model biochemical networks. The state of the network is determined by the number of each of a discrete set of molecules and evolves stochastically through reactions between these molecules. See Wilkinson (2009) and references therein for further background.

Inference for these models is challenging, as the transition density of the model is intractable. However, simulation from the models is possible, e.g. by using the algorithm of Gillespie (Gillespie, 1977). As such they are natural applications for ABC methods. Here we focus on a simple example of a stochastic kinetic network: the Lotka–Volterra model of Boys *et al.* (2008). Although the model is simple, Boys *et al.* (2008) show the difficulty of full likelihood inference.

This model has a state which consists of two molecules. Denote the state at time t by $\mathbf{Y}_t = (Y_t^{(1)}, Y_t^{(2)})$. There are three types of reaction: the birth of a molecule of type 1, the death of a molecule of type 2 and a reaction between the two molecules which removes a type 1 molecule and adds a type 2 molecule. (The network is called the Lotka–Volterra model, because of its link with predator–prey models, type 1 molecules being prey and type 2 being predators.)

The dynamics for the model are Markov, and can be specified in terms of transition over a small time interval δt . For positive parameters θ_1 , θ_2 and θ_3 , we have

$$\Pr\{\mathbf{Y}_{t+\delta t} = (z_1, z_2) | \mathbf{Y}_t = (y_1, y_2)\} = \begin{cases} 1 - (\theta_1 y_1 + \theta_2 y_1 y_2 + \theta_3 y_2) \delta t + o(\delta t) & \text{if } z_1 = y_1 \text{ and } z_2 = y_2, \\ \theta_1 y_1 \delta t + o(\delta t) & \text{if } z_1 = y_1 + 1 \text{ and } z_2 = y_2, \\ \theta_2 y_1 y_2 \delta t + o(\delta t) & \text{if } z_1 = y_1 - 1 \text{ and } z_2 = y_2 + 1, \\ \theta_3 y_2 \delta t + o(\delta t) & \text{if } z_1 = y_1 \text{ and } z_2 = y_2 - 1, \\ o(\delta t) & \text{otherwise.} \end{cases}$$

We shall focus on the case of the network being fully observed at discrete time points, and also on just observing the type 2 molecules initially, together with the type 1 molecules at all observation points. All simulations use the parameters from Boys *et al.* (2008), with $\theta_1 = 0.5$, $\theta_2 = 0.0025$ and $\theta_3 = 0.3$, and evenly sampled data collected at time intervals of length τ . We shall perform analysis conditional on the known initial state of the system.

4.3.1. Sequential approximate Bayesian computation analysis

Wilkinson (2011) considers simulation-based approaches for analysing stochastic kinetic

Table 5. Algorithm 4: a sequential ABC sampler for the Lotka–Volterra model

Input—a set of times t_0, \dots, t_n and data values $\mathbf{y}_{t_0}, \dots, \mathbf{y}_{t_n}$;
a number of particles, N , a kernel $K(\cdot)$ and a bandwidth h

Initialize—for $i = 1, \dots, N$ sample $\theta^{(i)} = (\theta_1^{(i)}, \theta_2^{(i)}, \theta_3^{(i)})$ from the prior distribution for the parameters

Iterate—for $j = 1, 2, \dots, n$:

step 1, for $i = 1, \dots, N$, sample a value for the state at time t_j , $\mathbf{y}_{t_j}^{(i)}$, given its value at time t_{j-1} , $\mathbf{y}_{t_{j-1}}^{(i)}$ by using the Gillespie algorithm;

step 2, for $i = 1, \dots, N$, calculate weights

$$w^{(i)} = K\{(\mathbf{y}_{t_j}^{(i)} - \mathbf{y}_{t_j})/h\};$$

step 3, sample N times from a kernel density approximation to a weighted sample of θ -values, $\{\theta^{(i)}, w_i\}_{i=1}^N$ (see for example Liu and West (2001)); denote this sample $\theta^{(1)}, \dots, \theta^{(N)}$

Output—a sample of θ -values.

networks which are based on sequential Monte Carlo methods (see Doucet *et al.* (2000) for an introduction). In some applications, to get these methods to work for reasonable computational cost, Wilkinson (2011) suggests using ABC. A version of such an algorithm (though based on importance sampling rather than MCMC sampling for the sequential update) for the Lotka–Volterra model is given in algorithm 4 (Table 5). There are many approaches to improve the computational efficiency of this algorithm; see for example Doucet *et al.* (2000, 2001) for details.

The discussion following theorem 2 shows that an algorithm like algorithm 4 may give inconsistent parameter estimates, even when ignoring the Monte Carlo error. A simple remedy to this is to implement noisy ABC within this algorithm. This can be done by adding noise to the observed values. The noisy sequential ABC algorithm is given by algorithm 5, which differs from algorithm 4 by replacing step 2 with

step 2, simulate \mathbf{x}_j from $K(\mathbf{x})$; for $i = 1, \dots, N$, calculate weights

$$w^{(i)} = K\{(\mathbf{y}_{t_j}^{(i)} - \mathbf{y}_{t_j} - h\mathbf{x}_j)/h\}.$$

To evaluate the relative merits of the two sequential ABC algorithms, we analysed 100 simulated data sets for the Lotka–Volterra model. For stability of the sequential Monte Carlo algorithm we chose $K(\cdot)$ to be the density function of a bivariate normal random variable, as a uniform kernel can lead to iterations where all weights are 0. We analysed two data scenarios, with $\tau = 0.1$: one with full observations, and one where only the number of type 1 molecules is observed. The sequential ABC algorithms were implemented with $N = 5000$ and $h = \sqrt{\tau}$. The latter was chosen to be a small value for which the sequential algorithms still performed adequately in terms of Monte Carlo performance (as measured by variability of the weights after each iteration).

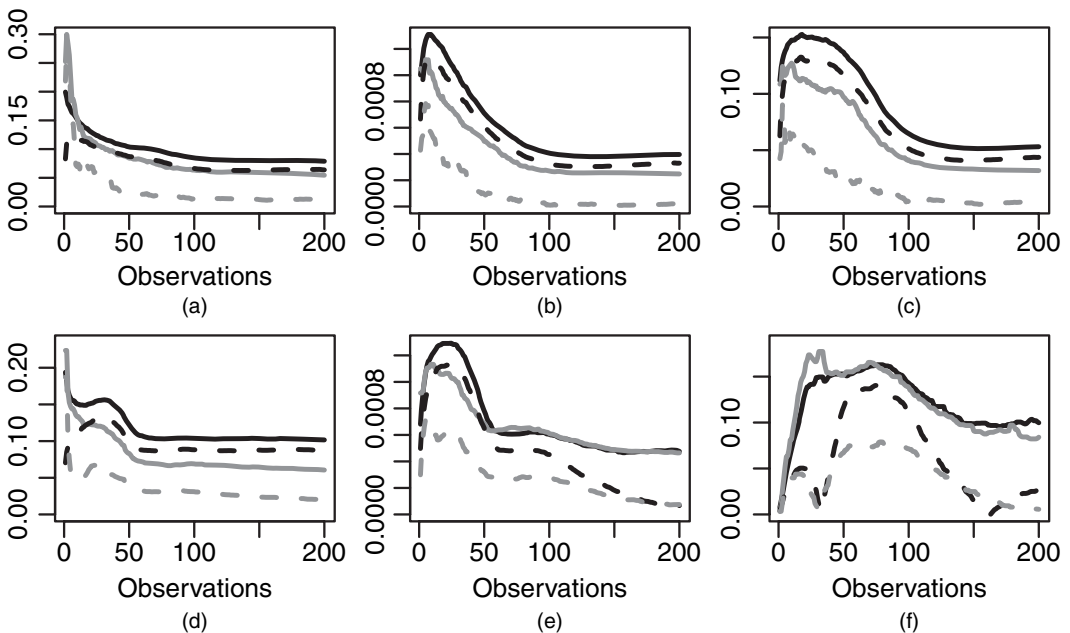


Fig. 2. Plots of root-mean-square error loss (—) and absolute bias (---) for the standard sequential ABC algorithm (—, ---) and the noisy ABC version (—, ---) as a function of the number of observations ($\tau = 0.1$), (a)–(c) observing the number of both molecules and (d)–(f) for observing only the number of type 1 molecules: (a), (d) θ_1 ; (b), (e) θ_2 ; (c), (f) θ_3

Results are shown in Fig. 2. These show both the absolute bias and the root mean quadratic loss for estimating each parameter for both ABC algorithms as the number of observations analysed varies between 1 and 200. For the full observation case, we can see evidence of bias in all three parameters for the standard sequential ABC algorithm. Noisy ABC shows evidence of being asymptotically unbiased as the number of observations increases. Overall, noisy ABC appears more accurate, except perhaps if only a handful of observations are made. When just type 1 molecules are observed, the picture is slightly more complex. We observed evidence of bias for the standard ABC algorithm for θ_1 , and for this parameter noisy ABC is more accurate. For the other parameters, there appears little difference between the accuracy and bias of the two ABC algorithms. The difference between the parameters is likely to be because θ_1 affects only the dynamics of the observed molecule.

a discrete population model

4.4. Inference for Ricker model

The Ricker map is an example of an ecological model with complex dynamics. It updates a population size N_t over a time step by

$$N_{t+1} = rN_t \exp(-N_t + e_t)$$

where the e_t are independent $N(0, \sigma_e^2)$ noise terms. Wood (2010) studied a model in which Poisson observations y_t are made with means ϕN_t . The parameters of interest are $\theta = (\log(r), \sigma_e, \phi)$. The initial state is $N_0 = 1$ and the observations are $y_{51}, y_{52}, \dots, y_{100}$. This is a complex inference scenario in which there is no obvious natural choice of summary statistics. Furthermore Wood (2010) argued that estimating parameters by maximum likelihood is difficult for this model owing to the chaotic nature of the system.

Wood (2010) analysed this model by using a Metropolis–Hastings algorithm to explore a ‘synthetic likelihood’, $L_s(\theta)$. A summary statistic function must be specified as in ABC, and $L_s(\theta)$ is then the density of \mathbf{s}_{obs} under an $N(\mu_\theta, \Sigma_\theta)$ model where μ_θ and Σ_θ are an approximation of the mean and variance of \mathbf{s}_{obs} for the given parameter value obtained through simulation (see Wood (2010) for more details).

We simulated 50 data sets with $\log(r) = 3.8$, $\phi = 10$ and $\log(\sigma_e)$ -values drawn from a uniform distribution on $[\log(0.1), 0]$. (Initial analysis showed that the true $\log(\sigma_e)$ -value affected the performance of the various methods.) These were analysed under our approach and that of Wood (2010) (using the code in that paper’s supplementary material). The distribution just mentioned was used as prior for $\log(\sigma_e)$, and improper uniform priors were placed on $\log(r) \geq 0$ and ϕ . All parameters were assumed independent under the prior. As each parameter had a non-negativity constraint, the MCMC algorithms used log-transforms of θ as the state. Each semi-automatic ABC analysis used 10^6 simulated data sets.

The summary statistics that were used by Wood (2010) were the autocovariances to lag 5, the coefficients of a cubic regression of the ordered differences $\Delta_t = y_t - y_{t-1}$ on those of the observed data, least squares estimates for the model $y_{t+1}^{0.3} = \beta_1 y_t^{0.3} + \beta_2 y_t^{0.6} + \varepsilon_t$, the mean observation \bar{y} and $\sum_{t=51}^{100} \mathbb{I}(y_t = 0)$ (the number of zero observations). We denote this set by E_0 and use it in the semi-automatic ABC pilot runs. The pilot runs used acceptance kernel $A = \Sigma^{-1}$ where Σ is the sample variance matrix of 500 simulated summary statistics vectors for a representative fixed parameter value.

In the regression stage of semi-automatic ABC, training data sets mostly consisting of 0s were fitted poorly. Since the observed data sets had at most 31 0s, any data sets with 45 or more 0s were discarded from our training data, and automatically rejected in the subsequent ABC runs. Summary statistics were constructed for two nested sets of explanatory variables. The smaller

Table 6. Mean quadratic losses for various analyses of 50 simulated Ricker data sets†

<i>Method</i>	<i>log(r)</i>	<i>σ_e</i>	<i>φ</i>
Synthetic likelihood	0.050	0.032	0.66
Comparison	0.039	0.038	0.54
Comparison + regression	0.046	0.041	0.78
Semi-automatic ABC	0.039	0.032	0.36

†Losses within 10% of the smallest values for that parameter are italicized.

set, E1, included E0 and additionally $\sum_{i=51}^{100} \mathbb{I}(y_i = j)$ for $1 \leq j \leq 4$, $\log(\bar{y})$, the logarithm of the sample variance, $\log(\sum_{i=51}^{100} y_i^j)$ for $2 \leq j \leq 6$ and auto-correlations up to lag 5. The larger set, E2, also added $(y_t)_{51 \leq t \leq 100}$ (time-ordered observations), $(y_{(t)})_{1 \leq t \leq 50}$ (magnitude-ordered observations), $(y_t^2)_{51 \leq t \leq 100}$, $(y_{(t)}^2)_{1 \leq t \leq 50}$, $\{\log(1 + y_t)\}_{51 \leq t \leq 100}$, $\{\log(1 + y_{(t)})\}_{1 \leq t \leq 50}$, $(\Delta_t^2)_{52 \leq t \leq 100}$ and $\{\Delta_{(t)}^2\}_{1 \leq t \leq 49}$.

Using set E2 instead of E1 reduced BIC in each linear regression by the order of thousands for all except one data set, suggesting that E2 gives better predictors. Thus we used the summary statistics based on set E2 within the ABC analysis. The results for these ABC analyses (Table 6) show an improvement over the synthetic likelihood for estimating $\log(r)$ and ϕ , and identical performance for estimating σ_e . The semi-automatic ABC analysis also does better than the comparison ABC analysis (based on summary statistics E_0). For this application the linear regression adjustment of Beaumont *et al.* (2002) actually produces worse results than using the raw output of the comparison ABC analyses.

Finally we looked at 95% credible intervals that were constructed from the semi-automatic ABC and synthetic likelihood method. The coverage frequencies of these intervals were 0.86, 0.70 and 0.96 for synthetic likelihood and 0.98, 0.92 and 1 for our method. Whereas the synthetic likelihood intervals appear to have coverage frequencies that are too low for two of the parameters, those from ABC are consistent with 0.95 coverage given a sample size of 50 data sets.

4.5. Inference for M/G/1-queue

Queuing models are an example of stochastic models which are easy to simulate from but often have intractable likelihoods. It has been suggested to analyse such models by using simulation-based procedures, and we shall look at a specific M/G/1-queue that has been analysed by both ABC (Blum and François, 2010) and indirect inference (Heggland and Frigessi, 2004) before. In this model, the service times are uniformly distributed in the interval $[\theta_1, \theta_2]$ and inter-arrival times are exponentially distributed with rate θ_3 . The queue is initially empty and only the interdeparture times y_1, y_2, \dots, y_{50} are observed.

We analysed 50 simulated data sets from this model. The true parameters were drawn from the prior under which $(\theta_1, \theta_2 - \theta_1, \theta_3)$ are uniformly distributed on $[0, 10]^2 \times [0, \frac{1}{3}]$. This choice gives arrival and service times of similar magnitudes, avoiding the less interesting situation where all y_i -values are independent draws from a single distribution.

The analysis of Blum and François (2010) used as summary statistics evenly spaced quantiles of the interdeparture times, including the minimum and maximum. Our semi-automatic ABC pilot analyses replicate this choice, using 20 quantiles. The explanatory variables $f(\mathbf{y})$ that we used to construct summary statistics were the ordered interdeparture times. Adding powers of

Table 7. Mean quadratic losses for various analyses of 50 $M/G/1$ data sets†

<i>Method</i>	θ_1	θ_2	θ_3
Comparison	1.1	2.2	0.0013
Comparison + regression	<i>0.020</i>	1.1	<i>0.0013</i>
Semi-automatic ABC	<i>0.022</i>	1.0	<i>0.0013</i>
Semi-automatic predictors	0.024	1.2	0.0017
Indirect inference	0.18	<i>0.42</i>	0.0033

†Losses within 10% of the smallest values for that parameter are italicized.

these values to $f(\mathbf{y})$ produced only minor improvements so the results are not reported. The analysis of each data set used 10^7 simulated data sets, split in the usual way.

We also applied the indirect inference approach of Heggland and Frigessi (2004). This used auxiliary statistics $(\bar{y}, \min(y_i), \hat{\theta}_2^{\text{ML}})$ where $\hat{\theta}_2^{\text{ML}}$ is the maximum likelihood estimate of θ_2 under an auxiliary model which has a closed form likelihood, namely that corresponding to independent observations from the steady state of the queue. Numerical calculation of $\hat{\theta}_2^{\text{ML}}$ is expensive so indirect inference used many fewer simulated data sets than ABC but had similar run times.

Table 7 shows the results. Semi-automatic ABC outperforms a comparison analysis using 20 quantiles as the summary statistics, but once a regression correction has been applied to the latter the results become very similar. Here the semi-automatic linear predictors are less accurate when used directly rather than in ABC. Indirect inference is more accurate at estimating θ_2 , presumably because of the accuracy of $\hat{\theta}_2^{\text{ML}}$ as an estimate of θ_2 . However, it is still substantially less accurate for the other two parameters. One advantage of indirect inference is that, as it requires fewer simulations to estimate the parameters accurately, it can more easily accommodate summaries that are expensive to calculate, such as $\hat{\theta}_2^{\text{ML}}$.

4.6. Inference of transmission of tuberculosis

Tanaka *et al.* (2006) used ABC to analyse tuberculosis bacteria genotype data sampled in San Francisco over a period from 1991 to 1992. Table 8 shows the data, consisting of 473 bacteria samples split into clusters which share the same genotype on a particular genetic marker. Thus the data consist of 282 bacteria samples that had unique genotypes, 20 pairs of bacteria that had the same genotype, and so on.

The model proposed was based on an underlying continuous time Markov process. Denote the total number of cases at time t by $N(t)$. The process starts at $t = 0$ with $N(0) = 1$. There are three types of event: birth, death (encompassing recovery of the host) and mutation. The rate of each type of event is the product of $N(t)$ and the appropriate parameter: α for birth, δ for death and θ for mutation. It was assumed that each mutation creates a completely new genotype. Cluster data are a simple random sample of 473 cases taken at the first t such that $N(t) = 10000$. The model conditions on such a t existing in the underlying process.

Table 8. Tuberculosis bacteria genotype data

Cluster size	1	2	3	4	5	8	10	15	23	30
Number of clusters	282	20	13	4	2	1	1	1	1	1

These data contain no information on time, so, for $k > 0$, parameter values (α, δ, θ) and $(k\alpha, k\delta, k\theta)$ give the same likelihood. We reparameterize to (a, d, θ) where $a = \alpha/(\alpha + \delta + \theta)$ and $d = \delta/(\alpha + \delta + \theta)$. The likelihood under this parameterization depends only on a and d . To reflect prior ignorance of (a, d) we use the prior density $\pi(a, d, \theta) \propto \pi(\theta) \mathbb{I}(0 \leq d \leq a) \mathbb{I}(a + d < 1)$, where $\pi(\theta)$ is the marginal prior for θ that was used in Tanaka *et al.* (2006). The prior restriction $d \leq a$ avoids the need for simulations in which $N(t) = 10000$ is highly unlikely to occur. The other

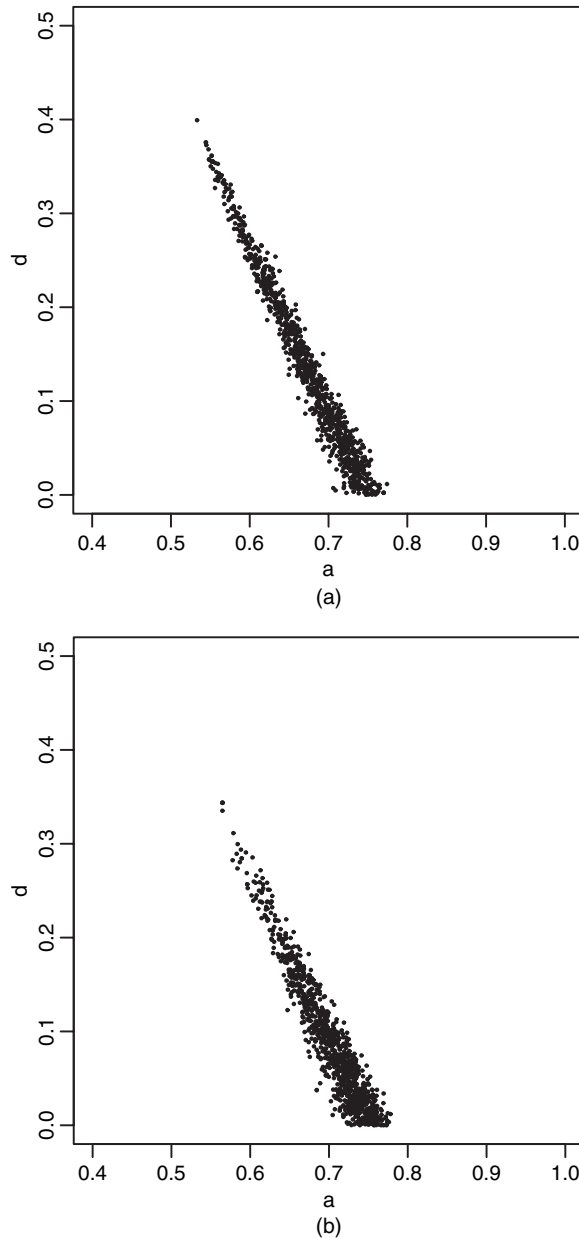


Fig. 3. ABC output for the tuberculosis application (every 1000th state is plotted): (a) comparison; (b) semi-automatic ABC

prior restrictions follow from positivity constraints on the original parameters. Under this prior and parameterization, the marginal posterior of θ is equal to its prior and the problem reduces to inference on a and d .

Tanaka *et al.* (2006) used two summary statistics for their ABC analysis: $g/473$ and $H = 1 - \sum_i (n_i/473)^2$, where g is the number of distinct clusters in the sample, and n_i is the number of observed samples in the i th genotype cluster. We retain this choice for our semi-automatic ABC pilot and the comparison ABC analysis reported below.

As parameters in the pilot output are highly correlated (Fig. 3), we fitted a line to the output by linear regression and made a reparameterization, $(u \ v)^T = M(a \ d)^T$, where M is a rotation matrix chosen so that on the fitted line u is constant. The semi-automatic ABC analysis was continued using (u, v) as the parameters of interest. The explanatory variables $f(\mathbf{y})$ comprised the number of clusters of size i for $1 \leq i \leq 5$, the number of clusters of size above 5, the average cluster size H , the size of the three largest clusters and the squares of all these quantities. The semi-automatic ABC analysis used 4×10^6 simulated data sets in total.

Fig. 3 shows the ABC posteriors, indicating that our methodology places less weight on the high d tail of the ABC posterior, reducing the marginal variances from (0.0029, 0.0088) (comparison) to (0.0017, 0.0048) (semi-automatic ABC).

We further investigated this model through a simulation study. We constructed 50 data sets by simulating parameters from the ABC posterior, and simulating data from the model for each of these pairs. We then compared ABC with the summary statistics of Tanaka *et al.* (2006) to semi-automatic ABC. Averaged across these data sets, semi-automatic ABC reduced the squared error loss by around 20–25% for the two parameters. Here we also found that the regression correction of Beaumont *et al.* (2002) did not change the accuracy of the method of Tanaka *et al.* (2006). Using the linear predictors obtained within semi-automatic ABC to estimate parameters (rather than as summary statistics) gave substantially worse estimates, with mean-squared error for a increasing by a factor of over 3.

5. Discussion

We have argued that ABC can be justified by aiming for calibration of the ABC posterior, together with accuracy of estimates of the parameters. We have introduced a new variant of ABC, called noisy ABC, which is calibrated. Standard ABC can be justified as an approximation to noisy ABC, but one which can be more accurate. Theoretical results suggest that, when a single data set is analysed by using a relatively small number of summary statistics, standard ABC should be preferred.

However, we show that, when we attempt to combine ABC analyses of multiple data sets, noisy ABC is to be preferred. Empirical evidence for this comes from our analysis of a stochastic kinetic network, where using standard ABC within a sequential ABC algorithm leads to biased parameter estimates. This result could be important for ABC analyses of population genetic data, if inferences are combined across multiple genomic regions. We believe that the ideal approach would be to implement a Rao–Blackwellized version of noisy ABC, where we attempt to average out the noise that is added to the summary statistics. If implemented, this would lead to a method which is more accurate than noisy ABC for estimating any function of the parameters. However, at present we are unsure how to implement such a Rao–Blackwellization scheme in a computationally efficient manner.

The main focus of this paper was a semi-automatic approach to implementing ABC. The main idea is to use simulation to construct appropriate summary statistics, with these summary statistics being estimates of the posterior mean of the parameters. This approach is based

on theoretical results which show that choosing summary statistics as the posterior means produces ABC estimators that are optimal in terms of minimizing quadratic loss. We have evaluated our method on several models, comparing both with ABC as it has been implemented in the literature, with indirect inference, and the synthetic likelihood approach of Wood (2010).

The most interesting comparison was between semi-automatic ABC and ABC using the regression correction of Beaumont *et al.* (2002). In several examples, these two approaches gave similarly accurate estimates. However, the semi-automatic ABC seemed to be more robust, with the regression correction actually producing less accurate estimates for the Ricker model, and not improving the accuracy for the tuberculosis model.

There are alternatives to using linear regression to construct the summary statistics, with many dimension reduction techniques being possible (see, for example, Wegmann *et al.* (2009) and Bazin *et al.* (2010)). However, one particular approach motivated by our theory is to use approximate estimates for each parameter. Such an approach has been used in Wilson *et al.* (2009), where estimates of the recombination rate under the incorrect demographic model were used as summary statistics in ABC. It is also the idea behind the approach of Drovandi *et al.* (2011), who summarized the data through parameter estimates under an approximating model. A further alternative is to use sliced inverse regression (Li, 1991) rather than linear regression to produce the summary statistics. The advantage of sliced inverse regression is that it can, where appropriate, generate multiple linear combinations of the explanatory variables, such that the posterior mean is approximately a function of these linear combinations. Each linear combination could then be used as a summary statistic within ABC.

Finally we note that there has been much recent research into the computational algorithms underpinning ABC. As well as the rejection sampling, importance sampling and MCMC algorithms that we have mentioned, there has been work on using sequential Monte Carlo methods (Beaumont *et al.*, 2009; Sisson *et al.*, 2007; Peters *et al.*, 2010) and adaptive methods (Bortot *et al.*, 2007; Del Moral *et al.*, 2012) among others. The ideas in this paper are focusing on a separate issue within ABC, and we think that the semi-automatic approach for implementing ABC that we describe can be utilized within whatever computational method is preferred.

Acknowledgements

We thank Chris Sherlock for many helpful discussions, and the reviewers for their detailed comments.

Appendix A: Variance of importance sampling–approximate Bayesian computation and Markov chain Monte Carlo–approximate Bayesian computation

Calculating the variance of estimates of posterior means by using importance sampling is complicated by the dependence between the importance sampling weights and values of $a(\theta)$ (see Liu (1996)). A common approach to quantifying the accuracy of importance sampling is to use an effective sample size. If the weights are normalized to have mean 1, then the effective sample size N_{eff} is N divided by the mean of the square of the weights. Liu (1996) argued that, for most functions $a(\theta)$, the variance of the estimator in expression (4) will be approximately expression (5) but with N_{acc} replaced by N_{eff} .

For a given proposal distribution $g(\theta)$, the effective sample size is

$$N_{\text{eff}} = N \frac{\int p(\theta | \mathbf{s}_{\text{obs}}) \pi(\theta) d\theta}{E_{\text{ABC}}\{\pi(\theta)/g(\theta)\}}. \quad (11)$$

It can be shown that the optimal proposal distribution, in terms of maximizing N_{eff}/N , is

$$g_{\text{opt}}(\theta|\mathbf{y}_{\text{obs}}) \propto \pi(\theta) p(\theta|\mathbf{s}_{\text{obs}})^{1/2},$$

in which case

$$N_{\text{eff}}^* = N_{\text{acc}} \left[\frac{1 + \text{var}_{\pi} \{p(\theta|\mathbf{s}_{\text{obs}})^{1/2}\}}{E_{\pi} \{p(\theta|\mathbf{s}_{\text{obs}})^{1/2}\}^2} \right],$$

where the variance and expectation on the right-hand side are with respect to $\pi(\theta)$. It is immediate that $N_{\text{eff}}^* \geq N_{\text{acc}}$, with equality only if $p(\theta|\mathbf{s}_{\text{obs}})$ does not depend on θ . The potential gains of importance sampling occur when $p(\theta|\mathbf{s}_{\text{obs}})$ varies greatly.

Analysis of the Monte Carlo error within the MCMC ABC algorithm is more difficult. However, consider fixing a proposal kernel $g(\cdot|\cdot)$, which will fix the type of transitions attempted. The Monte Carlo error then will be primarily governed by the average acceptance probability. For simplicity assume that $K(\cdot)$ is a uniform kernel and that either $g(\cdot|\cdot)$ is chosen to have the prior $\pi(\theta)$ as its stationary distribution or the term $\pi(\theta) g(\theta_{i-1}|\theta)/\pi(\theta_{i-1}) g(\theta|\theta_{i-1}) \approx 1$ and can be ignored. The average acceptance probability at stationarity is

$$\int \int \pi_{\text{ABC}}(\theta|\mathbf{s}_{\text{obs}}) g(\theta'|\theta) p(\theta'|\mathbf{s}_{\text{obs}}) d\theta d\theta' = \int \pi(\theta) p(\theta|\mathbf{s}_{\text{obs}}) d\theta \int \int \pi_{\text{ABC}}(\theta|\mathbf{s}_{\text{obs}}) g(\theta'|\theta) \frac{\pi_{\text{ABC}}(\theta'|\mathbf{s}_{\text{obs}})}{\pi(\theta')} d\theta d\theta'.$$

The integral comes from averaging over the current and proposed values for the MCMC algorithm, with the average acceptance probability for a given proposed value θ' being $p(\theta'|\mathbf{s}_{\text{obs}})$. The right-hand side comes from using expression (1). The first term on the right-hand side is the average acceptance probability of the rejection algorithm. The second term is 1 if we use an independence sampler $g(\theta'|\theta)$ and will be much greater than 1 if the ABC posterior is peaked relative to the prior, and if the transition kernel proposes localized moves.

Appendix B: Proof of lemma 1

Write

$$\begin{aligned} \int p(\theta|\mathbf{s}_{\text{obs}}) \pi(\theta) d\theta &= \int \int K\{(\mathbf{s} - \mathbf{s}_{\text{obs}})/h\} \pi(\mathbf{s}|\theta) \pi(\theta) d\theta d\mathbf{s} \\ &= \int h^d K(\mathbf{x}) \pi(\mathbf{s}_{\text{obs}} + h\mathbf{x}) d\mathbf{x}. \end{aligned}$$

The first equality comes from the definition of $p(\theta|\mathbf{s}_{\text{obs}})$; the second by integrating out θ and making a change of variable $\mathbf{x} = (\mathbf{s} - \mathbf{s}_{\text{obs}})/h$. So

$$\left| h^{-d} \int p(\theta|\mathbf{s}_{\text{obs}}) \pi(\theta) d\theta - \pi(\mathbf{s}_{\text{obs}}) \right| \leq \int K(\mathbf{x}) |\pi(\mathbf{s}_{\text{obs}} + h\mathbf{x}) - \pi(\mathbf{s}_{\text{obs}})| d\mathbf{x}.$$

This bound is shown to be $o(1)$ under either condition.

Consider first condition (a). Define c to be the maximum value of $|\mathbf{x}|$ such that $K(|\mathbf{x}|) > 0$. For any $\varepsilon > 0$, by continuity of $\pi(\mathbf{s})$ at \mathbf{s}_{obs} we have that there is a $\delta > 0$ such that $|\mathbf{x}| < \delta$ implies $|\pi(\mathbf{s}_{\text{obs}} + \mathbf{x}) - \pi(\mathbf{s}_{\text{obs}})| < \varepsilon$. Define $h_{\varepsilon} = \delta/c$. Then for $h < h_{\varepsilon}$ we have

$$\int K(\mathbf{x}) |\pi(\mathbf{s}_{\text{obs}} + h\mathbf{x}) - \pi(\mathbf{s}_{\text{obs}})| d\mathbf{x} \leq \varepsilon.$$

This inequality follows as $h|\mathbf{x}| < \delta$ for all \mathbf{x} where $K(\mathbf{x}) > 0$.

Now consider condition (b). By differentiable continuity, Taylor's theorem gives

$$\pi(\mathbf{s}_{\text{obs}} + h\mathbf{x}) = \pi(\mathbf{s}_{\text{obs}}) + \sum_i h x_i r_i(\mathbf{x}).$$

The remainder factor $r_i(\mathbf{x})$ is $|\partial\pi(\mathbf{z})/\partial s_i|$ for some $\mathbf{z}(\mathbf{x})$, so, by assumption, $|r_i(\mathbf{x})| \leq R$, a finite bound. Thus

$$\int K(\mathbf{x}) |\pi(\mathbf{s}_{\text{obs}} + h\mathbf{x}) - \pi(\mathbf{s}_{\text{obs}})| d\mathbf{x} \leq hR \sum_i \int |x_i| K(\mathbf{x}) d\mathbf{x}.$$

Appendix C: Proof of theorem 4

Rearrangement of the loss function gives

$$E\{L(\theta, \hat{\theta}; A) | \mathbf{y}_{\text{obs}}\} - \text{tr}(A\Sigma) = E\{(\tilde{\theta} - \hat{\theta})^T A(\tilde{\theta} - \hat{\theta}) | \mathbf{y}_{\text{obs}}\},$$

where $\tilde{\theta} = E(\theta | \mathbf{y}_{\text{obs}})$, the mean under the true posterior, which equals $S(\mathbf{y}_{\text{obs}})$ for the summary statistics under discussion. Define $\delta(\mathbf{s}_{\text{obs}}) = \tilde{\theta}(\mathbf{s}_{\text{obs}}) - \mathbf{s}_{\text{obs}}$ and make the change of variables $\mathbf{x} = (\tilde{\theta} - \mathbf{s}_{\text{obs}})/h$. Now

$$\begin{aligned} E\{L(\theta, \hat{\theta}; A) | \mathbf{y}_{\text{obs}}\} - \text{tr}(A\Sigma) - h^2 \int \mathbf{x}^T A \mathbf{x} K(\mathbf{x}) d\mathbf{x} &= -2h \int \mathbf{x}^T A \delta(\tilde{\theta} + h\mathbf{x}) K(\mathbf{x}) d\mathbf{x} \\ &\quad + \int \delta(\tilde{\theta} + h\mathbf{x})^T A \delta(\tilde{\theta} + h\mathbf{x}) K(\mathbf{x}) d\mathbf{x}. \end{aligned}$$

It is required that the modulus of the right-hand side be $o(h^2)$. Let R be the support of K . Since this is finite, it suffices to show that

$$\max_{\mathbf{x} \in R} \{\delta(\tilde{\theta} + h\mathbf{x})\} = o(h).$$

To find an expression for δ , observe that condition (1) holds for noisy ABC. Taking its expectation and making the change of variable $\mathbf{s} = S(\mathbf{y})$ give

$$\hat{\theta}(\mathbf{t}) = \frac{\int \theta \pi(\theta) \pi(\mathbf{s} | \theta) K\{(\mathbf{s} - \mathbf{t})/h\} d\theta d\mathbf{s}}{\int \pi(\mathbf{s}) K\{(\mathbf{s} - \mathbf{t})/h\} d\mathbf{s}}.$$

Note that

$$\int \theta \pi(\theta) \pi(\mathbf{s} | \theta) d\theta = \pi(\mathbf{s}) E(\theta | \mathbf{s}) = \mathbf{s} \pi(\mathbf{s}),$$

where the second equality is due to our choice of S . Thus

$$\delta(\mathbf{t}) = \frac{\int (\mathbf{s} - \mathbf{t}) \pi(\mathbf{s}) K\{(\mathbf{s} - \mathbf{t})/h\} d\mathbf{s}}{\int \pi(\mathbf{s}) K\{(\mathbf{s} - \mathbf{t})/h\} d\mathbf{s}}.$$

Make the change of variables $\mathbf{y} = (\mathbf{s} - \mathbf{t})/h$ and consider the case $\mathbf{t} = \tilde{\theta} + h\mathbf{x}$. Then

$$\delta(\tilde{\theta} + h\mathbf{x}) = \frac{h \int \mathbf{y} \pi\{\tilde{\theta} + h(\mathbf{x} + \mathbf{y})\} K(\mathbf{y}) d\mathbf{y}}{\int \pi\{\tilde{\theta} + h(\mathbf{x} + \mathbf{y})\} K(\mathbf{y}) d\mathbf{y}}.$$

By the argument in Appendix B, continuity of $\pi(\mathbf{s})$ at $\mathbf{s} = \tilde{\theta}$ gives denominator $\pi(\tilde{\theta}) + o(1)$. Consider the i th component of the integral in the numerator,

$$\left| \int y_i \pi\{\tilde{\theta} + h(\mathbf{x} + \mathbf{y})\} K(\mathbf{y}) d\mathbf{y} - \int y_i \pi(\tilde{\theta}) K(\mathbf{y}) d\mathbf{y} \right| \leq \max_{\mathbf{y} \in R} |y_i| \left| \int [\pi\{\tilde{\theta} + h(\mathbf{x} + \mathbf{y})\} - \pi(\tilde{\theta})] K(\mathbf{y}) d\mathbf{y} \right|.$$

This integral is $o(h)$ by the continuity argument just mentioned. Noting that $\int y_i K(\mathbf{y}) d\mathbf{y} = 0$ by assumption, we have

$$\int \mathbf{y} \pi\{\tilde{\theta} + h(\mathbf{x} + \mathbf{y})\} K(\mathbf{y}) d\mathbf{y} = o(1).$$

Combining these results gives the required bound for δ .

References

- Allingham, D., King, R. A. R. and Mengersen, K. L. (2009) Bayesian estimation of quantile distributions. *Statist. Comput.*, **19**, 189–201.
- Bazin, E., Dawson, K. J. and Beaumont, M. A. (2010) Likelihood-free inference of population structure and local adaptation in a bayesian hierarchical model. *Genetics*, **185**, 587–602.
- Beaumont, M. A., Cornuet, J.-M., Marin, J.-M. and Robert, C. P. (2009) Adaptive approximate Bayesian computation. *Biometrika*, **96**, 983–990.
- Beaumont, M. A., Zhang, W. and Balding, D. J. (2002) Approximate Bayesian computation in population genetics. *Genetics*, **162**, 2025–2035.
- Bernardo, J. M. and Smith, A. F. M. (1994) *Bayesian Theory*. Chichester: Wiley.
- Blum, M. G. B. (2010) Approximate bayesian computation: a nonparametric perspective. *J. Am. Statist. Ass.*, **105**, 1178–1187.
- Blum, M. G. B. and François, O. (2010) Non-linear regression models for Approximate Bayesian Computation. *Statist. Comput.*, **20**, 63–73.
- Bortot, P., Coles, S. and Sisson, S. (2007) Inference for stereological extremes. *J. Am. Statist. Ass.*, **102**, 84–92.
- Boys, R. J., Wilkinson, D. J. and Kirkwood, T. B. L. (2008) Bayesian inference for a discretely observed stochastic kinetic model. *Statist. Comput.*, **18**, 125–135.
- Cornuet, J.-M., Santos, F., Beaumont, M. A., Robert, C. P., Marin, J.-M., Balding, D. J., Guillemaud, T. and Estoup, A. (2008) Inferring population history with DIY ABC: a user-friendly approach to approximate Bayesian computation. *Bioinformatics*, **24**, 2713–2719.
- Del Moral, P., Doucet, A. and Jasra, A. (2012) An adaptive sequential Monte Carlo method for approximate Bayesian computation. *Statist. Comput.*, to be published.
- Diggle, P. J. and Gratton, R. J. (1984) Monte Carlo methods of inference for implicit statistical models (with discussion). *J. R. Statist. Soc. B*, **46**, 193–227.
- Doucet, A., de Freitas, J. F. G. and Gordon, N. J. (eds) (2001) *Sequential Monte Carlo Methods in Practice*. New York: Springer.
- Doucet, A., Godsill, S. J. and Andrieu, C. (2000) On sequential Monte Carlo sampling methods for Bayesian filtering. *Statist. Comput.*, **10**, 197–208.
- Drovandi, C. C. and Pettitt, A. N. (2009) Likelihood-free Bayesian estimation of quantile distributions. *Technical Report*. Queensland University of Technology, Brisbane.
- Drovandi, C. C., Pettitt, A. N. and Faddy, M. J. (2011) Approximate Bayesian computation using indirect inference. *Appl. Statist.*, **60**, 317–337.
- Gillespie, D. T. (1977) Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.*, **81**, 2340–2361.
- Gouriéroux, C. and Ronchetti, E. (1993) Indirect inference. *J. Appl. Econometr.*, **8**, s85–s118.
- Grelaud, A., Robert, C., Marin, J. M., Rodolphe, F. and Taly, J. F. (2009) ABC likelihood-free methods for model choice in Gibbs random fields. *Baysn Anal.*, **4**, 317–336.
- Hastie, T., Tibshirani, R. and Friedman, J. (2001) *The Elements of Statistical Learning, Data Mining, Inference, and Prediction*. New York: Springer.
- Haynes, M. (1998) Flexible distributions and statistical models in ranking and selection procedures, with applications. *PhD Thesis*. Queensland University of Technology, Brisbane.
- Heggland, K. and Frigessi, A. (2004) Estimating functions in indirect inference. *J. R. Statist. Soc. B*, **66**, 447–462.
- Joyce, P. and Marjoram, P. (2008) Approximately sufficient statistics and Bayesian computation. *Statist. Applic. Genet. Molec. Biol.*, **7**, article 26.
- Kingman, J. F. C. (1982) The coalescent. *Stoch. Processes Appl.*, **13**, 235–248.
- Li, K.-C. (1991) Sliced inverse regression for dimension reduction. *J. Am. Statist. Ass.*, **86**, 316–327.
- Liepe, J., Barnes, C., Cule, E., Erguler, K., Kirk, P., Toni, T. and Stumpf, M. P. (2010) ABC-SysBioapproximate Bayesian computation in Python with GPU support. *Bioinformatics*, **26**, 1797–1799.
- Lindsay, B. G. (1988) Composite likelihood methods. *Contemp. Math.*, **80**, 221–239.
- Liu, J. S. (1996) Metropolis independent sampling with comparisons to rejection sampling and importance sampling. *Statist. Comput.*, **6**, 113–119.
- Liu, J. and West, M. (2001) Combined parameter and state estimation in simulation based filtering. In *Sequential Monte Carlo in Practice* (eds A. Doucet, J. F. G. de Freitas and N. J. Gordon), pp. 197–223. New York: Springer.
- Lopes, J. S., Balding, D. and Beaumont, M. A. (2009) PopABC, a program to infer historical demographic parameters. *Bioinformatics*, **25**, 2747–2749.
- Mardia, K. V., Kent, J. T. and Bibby, J. M. (1979) *Multivariate Analysis*. London: Academic Press.
- Marjoram, P., Molitor, J., Plagnol, V. and Tavaré, S. (2003) Markov chain Monte Carlo without likelihoods. *Proc. Natn. Acad. Sci. USA*, **100**, 15324–15328.
- McKinley, T., Cook, A. R. and Deardon, R. (2009) Inference in epidemic models without likelihoods. *Int. J. Biostatist.*, **5**, article 24.
- Padhukasahasram, B., Wall, J. D., Marjoram, P. and Nordborg, M. (2006) Estimating recombination rates from single-nucleotide polymorphisms using summary statistics. *Genetics*, **174**, 1517–1528.

- Peters, G. W., Fan, Y. and Sisson, S. A. (2010) On sequential Monte Carlo, partial rejection control and approximate Bayesian computation. To be published.
- Prangle, D. (2011) Summary statistics and sequential methods for approximate Bayesian computation. *PhD Thesis*. Lancaster University, Lancaster.
- Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A. and Feldman, M. W. (1999) Population growth of human Y chromosomes, a study of Y chromosome microsatellites. *Molec. Biol. Evol.*, **16**, 1791–1798.
- Ratmann, O., Andrieu, C., Wiuf, C. and Richardson, S. (2009) Model criticism based on likelihood-free inference, with an application to protein network evolution. *Proc. Natn. Acad. Sci. USA*, **106**, 10576–10581.
- Ratmann, O., Jorgensen, O., Hinkley, T., Stumpf, M., Richardson, S. and Wiuf, C. (2007) Using likelihood-free inference to compare evolutionary dynamics of the protein networks of *h. pylori* and *p. falciparum*. *PLOS Comput. Biol.*, **3**, article e230.
- Rayner, G. D. and MacGillivray, H. L. (2002) Numerical maximum likelihood estimation for the g-and-k and generalized g-and-h distributions. *Statist. Comput.*, **12**, 57–75.
- Ripley, B. (1987) *Stochastic Simulation*. Chichester: Wiley.
- Sisson, S. A., Fan, Y. and Tanaka, M. M. (2007) Sequential Monte Carlo without likelihoods. *Proc. Natn. Acad. Sci. USA*, **104**, 1760–1765; correction, **106** (2009), article 16889.
- Sisson, S. A., Peters, G. W., Briers, M. and Fan, Y. (2010) A note on target distribution ambiguity of likelihood-free samplers. *Preprint*.
- Tanaka, M. M., Francis, A. R., Luciani, F. and Sisson, S. A. (2006) Using approximate Bayesian computation to estimate tuberculosis transmission parameters from genotype data. *Genetics*, **173**, 1511–1520.
- Tavaré, S., Balding, D. J., Griffiths, R. C. and Donnelly, P. (1997) Inferring coalescent times from DNA sequence data. *Genetics*, **145**, 505–518.
- 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. *J. R. Soc. Interface*, **6**, 187–202.
- Wegmann, D., Leuenberger, C. and Excoffier, L. (2009) Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood. *Genetics*, **182**, 1207–1218.
- Wilkinson, D. J. (2009) Stochastic modelling for quantitative description of heterogeneous biological systems. *Nat. Rev. Genet.*, **10**, 122–133.
- Wilkinson, D. J. (2011) Parameter inference for stochastic kinetic models of bacterial gene regulation, a Bayesian approach to systems biology. In *Bayesian Statistics 9* (eds J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West), pp. 679–690. Oxford: Oxford University Press.
- Wilkinson, R. D. (2008) Approximate Bayesian computation (ABC) gives exact results under the assumption of model error. *Preprint arXiv:0811.3355v1*. University of Nottingham, Nottingham.
- Wilson, D. J., Gabriel, E., Leatherbarrow, A. J. H., Cheesebrough, J., Gee, S., Bolton, E., Fox, A., Hart, C. A., Diggle, P. J. and Fearnhead, P. (2009) Rapid evolution and the importance of recombination to the gastro-enteric pathogen *campylobacter jejuni*. *Molec. Biol. Evol.*, **26**, 385–397.
- Wood, S. (2010) Statistical inference for noisy nonlinear ecological dynamic systems. *Nature*, **466**, 1102–1104.

Discussion on the paper by Fearnhead and Prangle

Mark Beaumont (*University of Bristol*)

I congratulate Paul Fearnhead and Dennis Prangle for their highly stimulating paper. An important contribution of this study has been the development of a principled approach to the construction of summary statistics in approximate Bayesian computation (ABC). In addition, for those interested in filtering problems and state space models, the authors highlight the potential importance of bias in standard ABC methods. They show that their ‘noisy ABC’ leads to properly calibrated posterior distributions. In my discussion I shall concentrate only on their method of constructing summary statistics, but I acknowledge the insights that their paper has provided on the importance of bias for some sequential applications of ABC.

The ABC approach is a way of inferring parameters in implicit statistical models (Diggle and Gratton, 1984) within the Bayesian framework. The current interest in ABC stems from developments in population genetics (Tavaré *et al.*, 1997; Pritchard *et al.*, 1999). In population genetics the likelihood can typically only be written in terms of the genealogical history, a high dimensional auxiliary variable that needs to be integrated out. Computational methods involving Monte Carlo sampling have been developed to do this. With these methods the conditional distribution of genealogical histories needs to be sampled, which imposes a constraint. ABC is particularly suited to tackling such problems because, although generated during the simulation, the genealogical history is irrelevant to computation of the summary statistics, which depend on the data alone. Thus, typically, a far larger space of genealogical histories can be explored.

A major contribution of this paper is to define a method for constructing summary statistics. The apparently arbitrary nature of the choice of summary statistics has always been perceived as the Achilles heel

of ABC. Many practical applications have at the outset taken a large number of summary statistics, as is the case in the current study. The motivation has arisen from an informal notion of sufficiency: the more summaries, the closer to joint sufficiency. However, the more summaries, the more difficult it is to match the observations closely (the ‘curse of dimensionality’). Thus it has often been necessary then to reduce their number. One approach, as exemplified by the work of Joyce and Marjoram (2008) and Nunes and Balding (2010), has been to choose subsets of the summaries. An alternative approach has been to define a projection to a lower dimensional representation of the data, via principal components analysis (Itan *et al.*, 2009; Bazin *et al.*, 2010) or partial least squares (Wegmann *et al.*, 2009).

It seems to me that there is quite a strong similarity between the projection methods that have been used in ABC and the method of Fearnhead and Prangle. For example, the first partial least squares component is $T_1 \propto \Sigma_j \widehat{\text{cov}}(X_j, Y)$ (Garthwaite, 1994) for centred predictors X_j and response Y . If the summary statistics are standardized to have the same variance and are orthogonal to each other (admittedly unrealistic given their joint dependence on the model parameters) then the first component of partial least squares as used by Wegmann *et al.* (2009) is proportional to Fearnhead and Prangle’s estimate of the posterior mean. Other projection approaches may also be similar: for example the first component of target projection developed in chemometrics is proportional to the linear predictor for non-orthogonal variables (Kvalheim, 2010). Of course it must be appreciated that the earlier projection approaches in ABC have been entirely intuitive—aimed at retaining the information content in the summaries while addressing the dimensionality problem. By contrast Fearnhead and Prangle propose a principled method based on minimizing expected quadratic loss.

The authors make some revealing comparisons between related methods. Of particular interest to me are the comparisons between the authors’ approach and the regression adjustment of Beaumont *et al.* (2002), in which theirs often outperforms the latter. At first sight this might seem rather puzzling. Performance is quantified through an estimate of expected quadratic loss (mean-square error) using the inferred posterior mean $E(\theta|\mathbf{y}_{\text{obs}})$ for some parameter θ . Both methods use linear regression (locally) to obtain an estimate of $E(\theta|\mathbf{y}_{\text{obs}})$ and therefore should give the same expected loss. However, the approach of Fearnhead and Prangle is then to use the linear predictor as a summary statistic in the final ‘production run’ of the ABC, thereby greatly reducing the volume of the ellipsoid, defined by $\mathbf{x}^T \mathbf{A} \mathbf{x} < c$, for summary statistics \mathbf{x} and weight matrix \mathbf{A} , that is needed to accept a reasonable number of points. As noted by Marin *et al.* (2011) the method of Beaumont *et al.* (2002) is really a post-processing scheme, independent of the method of simulation (e.g. rejection, Markov chain Monte Carlo or sequential Monte Carlo sampling).

In conclusion, the authors are to be highly commended for this paper, which offers a major advance, both conceptually and practically, in the treatment of intractable likelihood problems. It gives me great pleasure to propose the vote of thanks.

Christian P. Robert (*Université Paris Dauphine, Institut Universitaire de France, Paris, and Centre de Recherche en Economie et Statistiques, Malakoff*)

A discussion paper on the fast growing technique of approximate Bayesian computation (ABC) is quite timely, especially when it addresses the important issue of summary statistics that are used by such methods. I thus congratulate the authors on their endeavour.

Although ABC has gradually been analysed from a (mainstream) statistical perspective, this is one of the very first papers performing a decision theoretic analysis of the factors influencing the performances of the method (along with, for example, Dean *et al.* (2010)). Indeed, a very interesting input of the authors is that ABC is considered there from a purely inferential viewpoint and calibrated for estimation purposes. The most important result therein is in my opinion the consistency result in theorem 2, which shows that noisy ABC is a coherent estimation method when the number of observations grows to ∞ . I, however, dispute the generality of the result, as explained below.

In Fearnhead’s and Prangle’s setting, the Monte Carlo error that is inherent in ABC is taken into account through the average acceptance probability, which collapses to 0 when $h \rightarrow 0$, meaning that $h = 0$ is a sub-optimal choice. This is a strong (and valid) point of the paper because this means that the ‘optimal’ value of h is not 0, a point that is repeated later in this discussion. The decomposition of the error into

$$\text{tr}(\mathbf{A}\Sigma) + h^2 \int \mathbf{x}^T \mathbf{A} \mathbf{x} K(\mathbf{x}) d\mathbf{x} + \frac{C_0}{Nh^d}$$

is very similar to error decompositions that are found in (classical) non-parametric statistics. In this respect, I fail to understand the authors’ argument that lemma 1 implies that a summary statistic with larger dimension also has larger Monte Carlo error: given that $\pi(\mathbf{s}_{\text{obs}})$ also depends on h , the appearance

of h^d in equation (6) is not enough of an argument. There actually is a larger issue that I also have against several recent papers on the topic, where the bandwidth h or the tolerance ε is treated as a given or an absolute number whereas it should be calibrated in terms of a collection of statistical and computational factors, the number d of summary statistics being one of them.

When the authors consider the errors that are made in using ABC, balancing the Monte Carlo error due to simulation with the ABC error due to approximation (and non-zero tolerance), they fail to account for ‘the third man’ in the picture, namely the error that is made in replacing the (exact) posterior inference based on \mathbf{y}_{obs} with the (exact) posterior inference based on \mathbf{s}_{obs} , i.e. for the loss of information due to the use of the summary statistics at the centre of the paper. (As shown in Robert *et al.* (2011), this loss may be so extreme that the resulting inference becomes inconsistent.) Although the remarkable (and novel) result in the proof of theorem 3 that

$$\mathbb{E}[\theta | \mathbb{E}[\theta | \mathbf{y}_{\text{obs}}]] = \mathbb{E}[\theta | \mathbf{y}_{\text{obs}}]$$

shows that $\mathbf{s}_{\text{obs}} = \mathbb{E}[\mathbf{y}_{\text{obs}}]$ does not lose any (first-order) information when compared with \mathbf{y}_{obs} , and hence is ‘almost’ sufficient in that weak sense, theorem 3 considers only a specific estimation aspect, rather than full Bayesian inference, and is furthermore parameterization dependent. In addition, the second part of theorem 3 should be formulated in terms of the above identity, as ABC plays no role when $h = 0$.

If I concentrate more specifically on the mathematical aspects of the paper, a point of the utmost importance is that theorem 2 can only hold at best when θ is identifiable for the distribution \mathbf{s}_{obs} . Otherwise, some other values of θ satisfy $p(\theta | \mathbf{s}_{\text{obs}}) = p(\theta_0 | \mathbf{s}_{\text{obs}})$. Considering the specific case of an ancillary statistic \mathbf{s}_{obs} clearly shows that the result cannot hold in full generality. Therefore, vital assumptions are clearly missing to achieve a rigorous formulation of theorem 2. The call to Bernardo and Smith (1994) is thus not really relevant in this setting as the convergence results therein require conditions on the likelihood that are not necessarily verified by the distribution of \mathbf{s}_{obs} . We are thus left with the open question of the asymptotic validation of the noisy ABC estimator—ABC being envisioned as an inference method *per se*—when the summary variables are not sufficient. Obtaining necessary and sufficient conditions on those statistics as done in Marin *et al.* (2011) for model choice is therefore paramount, the current paper obviously containing essential features to achieve this goal.

In conclusion, I find the paper both exciting and bringing both new questions and new perspectives to the forefront of ABC research. I thus unreservedly support the vote of thanks.

The vote of thanks was passed by acclamation.

Richard D. Wilkinson (*University of Nottingham*)

Calibration is a way of assessing probability statements against some notion of truth, usually reality. We are well calibrated if $p\%$ of all predictions reported at probability p are true. This is close to Fearnhead and Prangle’s definition: P_{ABC} is calibrated if

$$p\{\theta \in A | E_q(A)\} = q,$$

i.e., given that A is an event assigned probability q by approximate Bayesian computation (ABC), then we are calibrated if A occurs with probability q . This differs slightly from previous definitions as the base measure (or ‘truth’) is defined not by reality, but by our definition of the prior, likelihood and summary, i.e. the distribution

$$\pi(\theta | s) \propto \int \pi(s | y) \pi(y | \theta) \pi(\theta) dy.$$

In other words, they are not comparing P_{ABC} with reality, but with a modeller-specified distribution.

Calibration, even with reality as the base measure, is not universally accepted by Bayesians as something to strive for (Seidenfeld, 1985). It is even more questionable here as we care about how statements that we make relate to the world, not to a mathematically defined posterior. For example, the fact that the prior is calibrated under this definition should give us pause for thought. Moreover, calibration in this case says that noisy ABC is calibrated with respect to the user-defined posterior only if we were to do the analysis repeatedly. In a particular analysis, nothing can be said, as we generate only one noisy data set (this criticism does not apply to Section 2.2 where noisy ABC is a more natural choice).

I prefer to view ABC as a method that provides exact inference under a different model (Wilkinson, 2008), and to try to choose this alternative model to be of scientific interest. For example, if we believe that there is model or measurement error on the simulator, with distribution $\pi(s^{\text{obs}} | y^{\text{simulator}}) = K[\{s^{\text{obs}} - S(y^{\text{simulator}})\}/h]$ then kernel ABC gives exact inference. If we were now to use noisy ABC we would have added two lots of error.

In summary, if you know that your simulator is imperfect, then I would argue that it is better to attempt to account for the simulator's imperfections in the modelling and inference and to do exact inference, than it is to do an analysis using noisy ABC that is calibrated with respect to some base measure that we know to be meaningless.

Julien Cornebise, Mark Girolami and Ioannis Kosmidis (*University College London*)

Parametric estimation of the summary statistics likelihood

We would like to draw attention to the work of Wood (2010) which is of direct relevance to approximate Bayesian computation (ABC) despite having been largely overlooked in the ABC literature. In Section 1.2 of the current paper the authors note that $K[\{S(\mathbf{y}) - \mathbf{s}_{\text{obs}}\}/h]$ is a Parzen–Rosenblatt density kernel. As has already been suggested in for example Del Moral *et al.* (2012) one can simulate R observations $\mathbf{y}_1, \dots, \mathbf{y}_R$ for a given value of θ and use the corresponding non-parametric kernel density estimate $\Sigma_r K[\{S(\mathbf{y}_r) - \mathbf{s}_{\text{obs}}\}/h]/(Rh^d)$ for $p(\mathbf{s}_{\text{obs}}|\theta)$. Wood (2010) suggested the *synthetic likelihood* by invoking the assumption of multivariate normality such that $\mathbf{s}_{\text{obs}} \sim \mathcal{N}(\boldsymbol{\mu}_\theta, \Sigma_\theta)$. Plug-in estimates of $\boldsymbol{\mu}_\theta$ and Σ_θ are obtained by the empirical mean $\hat{\boldsymbol{\mu}}_\theta^R$ and covariance $\hat{\Sigma}_\theta^R$ using the simulated statistics $S(\mathbf{y}_1), \dots, S(\mathbf{y}_R)$ yielding a parametric density estimate $\mathcal{N}(\mathbf{s}_{\text{obs}}; \hat{\boldsymbol{\mu}}_\theta^R, \hat{\Sigma}_\theta^R)$. This *synthetic likelihood* can then be used in a Markov chain Monte Carlo (MCMC) setting analogous to MCMC–ABC—and can similarly be used in importance sampling–ABC and sequential Monte Carlo–ABC settings. The convergence rate of the variance of the parametric density estimate is independent of the dimension of the summary statistics, which is in contrast with the non-parametric rate which suffers from the curse of dimensionality. This lower variance could improve mixing of the MCMC algorithm underpinning ABC, as already demonstrated in the pseudomarginal approach to MCMC sampling of Andrieu and Roberts (2009).

Of course Wood does not offer an automatic choice of the summary statistics: the user selects a (possibly large) set of summary statistics based on domain knowledge of the problem. This is similar to the way that Section 3 offers to select the ‘transformations’ $f(\mathbf{y})$, which are the first round of summary statistics. However, the relative weighting of each statistic is automatically inferred via the corresponding variance estimate. Could such a feature be of benefit in semi-automatic ABC?

The assumption of multivariate normality on the distribution of the summary statistics plays a critical role in Wood's approach. He justified it by

- (a) choosing polynomial regression coefficients as summary statistics and, most interestingly,
- (b) by using a pilot run to improve the normality of the statistics by quantile regression transformations—a preliminary step that is conceptually similar to the pilot ABC run of Section 3.

We conjecture that such transformations could allow for the use of parametric density estimation within semi-automatic ABC, possibly benefitting from the increased convergence rate and making use of the variance of the sampled statistics. Additionally, we wonder whether theorem 4 could be modified to study the optimality of such transformed Gaussian statistics.

Anthony Lee (*University of Warwick, Coventry*), **Christophe Andrieu** (*University of Bristol*) and **Arnaud Doucet** (*University of Oxford*)

We congratulate the authors on a structured contribution to the practical use of approximate Bayesian computation (ABC) methods. We focus here on the conditional joint density

$$\bar{\pi}_{X,Y|\Theta}(x, y|\theta) = \pi_{Y|\Theta}(x|\theta) \bar{\pi}_{Y|X}(y|x),$$

which is central to all forms of ABC. Here x and y denote the simulated and observed data or summary statistics in ABC and $\bar{\pi}_{X|\Theta} = \pi_{Y|\Theta}$. In the paper, $\bar{\pi}_{Y|X}(y|x) = K\{(y-x)/h\}$ and

$$\bar{\pi}_{Y|\Theta}(y|\theta) = \int \bar{\pi}_{X,Y|\Theta}(x, y|\theta) dx \neq \pi_{Y|\Theta}(y|\theta)$$

leads to the approximation. Although neither $\bar{\pi}_{Y|\Theta}(y|\theta)$ nor $\pi_{Y|\Theta}(x|\theta)$ can be evaluated, the ability to sample according to $\pi_{Y|\Theta}(\cdot|\theta)$ allows for rejection, importance and Markov chain Monte Carlo sampling according to $\bar{\pi}_{\Theta, X|Y}(\cdot|y)$. The calibration of noisy ABC is then immediate. If $\tilde{y} \sim \bar{\pi}_{Y|X}(\cdot|y)$, then marginally $\tilde{y} \sim \bar{\pi}_{Y|\Theta}(\cdot|\theta^*)$ since $y \sim \pi_{Y|\Theta}(\cdot|\theta^*)$ for some $\theta^* \in \Theta$. Inference using $\bar{\pi}$ with $Y = \tilde{y}$ is then consistent with the data-generating process although $\bar{\pi}_{\Theta|Y}(\cdot|\tilde{y})$ may not be closer to $\pi_{\Theta|Y}(\cdot|y)$ than $\bar{\pi}_{\Theta|Y}(\cdot|y)$.

The tractability of $\bar{\pi}_{\Theta, X|Y}$, whose unavailable marginal $\bar{\pi}_{\Theta|Y}(\cdot|y)$ is of interest, puts ABC within the domain of pseudomarginal approaches (Beaumont, 2003; Andrieu and Roberts, 2009), and the grouped independence Metropolis–Hastings (GIMH) algorithm has been used in Becquet and Przeworski (2007). We

Table 9. Algorithm 5: rejuvenating GIMH–ABC

<p>At time t, with $\theta_t = \theta$:</p> <p>step 1, sample $\theta' \sim g(\cdot \theta)$;</p> <p>step 2, sample $z_{1:N} \sim \pi_{Y \Theta}^{\otimes N}(\cdot \theta')$;</p> <p>step 3, sample $x_{1:N-1} \sim \pi_{Y \Theta}^{\otimes(N-1)}(\cdot \theta)$;</p> <p>step 4, with probability</p> $\min \left\{ 1, \frac{\pi(\theta') g(\theta \theta') \sum_{i=1}^N \mathbf{1}_{B_h(z_i)}(y)}{\pi(\theta) g(\theta' \theta) \left(1 + \sum_{i=1}^{N-1} \mathbf{1}_{B_h(x_i)}(y) \right)} \right\}$ <p>set $\theta_{t+1} = \theta'$; otherwise set $\theta_{t+1} = \theta$</p>
--

Table 10. Algorithm 6: one-hit Markov chain Monte Carlo–ABC

<p>At time t, with $\theta_t = \theta$:</p> <p>step 1, sample $\theta' \sim g(\cdot \theta)$;</p> <p>step 2, with probability $1 - \min\{1, \pi(\theta') g(\theta \theta')/\pi(\theta) g(\theta' \theta)\}$, set $\theta_{t+1} = \theta$ and go to time $t+1$;</p> <p>step 3, sample $z_i \sim \pi_{Y \Theta}(\cdot \theta')$ and $x_i \sim \pi_{Y \Theta}(\cdot \theta)$ for $i = 1, \dots$ until $y \in B_h(z_i)$ and/or $y \in B_h(x_i)$;</p> <p>step 4, if $y \in B_h(z_i)$ set $\theta_{t+1} = \theta'$ and go to time $t+1$;</p> <p>step 5, if $y \in B_h(x_i)$ set $\theta_{t+1} = \theta$ and go to time $t+1$</p>

present two novel Markov chain Monte Carlo–ABC algorithms based on recent work (Andrieu *et al.*, 2012), and for simplicity restrict ourselves to the case $\bar{\pi}_{Y|X}(y|x) \propto \mathbf{1}_{B_h(x)}(y)$, where $\mathbf{1}_{B_h(x)}$ is the indicator function of a metric ball of radius h around x . These algorithms define Markov chains solely on Θ .

In the GIMH algorithm with N auxiliary variables, the state of the chain is $(\theta, x_{1:N})$ where $x_{1:N} := (x_1, \dots, x_N)$ and at each iteration we propose new values $(\theta', z_{1:N})$ via $\theta' \sim g(\cdot|\theta)$ and $z_{1:N} \sim \pi_{Y|\Theta}^{\otimes N}(\cdot|\theta')$. Algorithm 5 in Table 9 presents an alternative to the GIMH algorithm with the crucial difference in step 3, where the GIMH algorithm would use the previously simulated values of $x_{1:N}$ instead of sampling $N-1$ new ones. This algorithm can have superior performance to the GIMH algorithm in some cases where the latter becomes ‘stuck’. Algorithm 6 in Table 10 involves a random number of simulations instead of fixed N , adapting the computation in each iteration to the simulation problem at hand. Data are simulated by using both θ and θ' until a ‘hit’ occurs. It can be verified that the invariant distribution of θ is $\bar{\pi}_{\Theta|Y}(\cdot|y)$ for both algorithms. The probability of accepting the move $\theta \rightarrow \theta'$ after step 1 in algorithm 5, as $N \rightarrow \infty$ approaches

$$\min \left\{ 1, \frac{\bar{\pi}_{\Theta|Y}(\theta'|y) g(\theta|\theta')}{\bar{\pi}_{\Theta|Y}(\theta|y) g(\theta'|\theta)} \right\}.$$

For algorithm 6 this probability is exactly

$$\min \left\{ 1, \frac{\pi(\theta') g(\theta|\theta')}{\pi(\theta) g(\theta'|\theta)} \right\} \frac{\bar{\pi}_{Y|\Theta}(y|\theta')}{\bar{\pi}_{Y|\Theta}(y|\theta) + \bar{\pi}_{Y|\Theta}(y|\theta') - c \bar{\pi}_{Y|\Theta}(y|\theta) \bar{\pi}_{Y|\Theta}(y|\theta')},$$

where c is the volume of the ball of radius h . Regarding the ‘automatic’ implementation of ABC, algorithm 5 could automate the use of N processors on a parallel computer or algorithm 6 could be used to adapt computational effort to the target of interest automatically.

Simon R. White, Theodore Kyraios and S. P. Preston (*Medical Research Council Biostatistics Unit, Cambridge*)

We congratulate the authors for a thought-provoking paper. There are two aspects in the paper which are of particular interest to us:

- (a) noisy approximate Bayesian computation (ABC) (Section 2.1) and
- (b) inference from multiple sources of data (Section 2.2).

White *et al.* (2012) proposed an ABC approach that is appropriate to discretely observed Markov models with the aims of reducing the computational cost, and avoiding the use of summary statistics. The Markov property enables the likelihood to be written as

$$\pi(\mathcal{X}|\theta) = \pi(x_1|\theta) \prod_{i=2}^n \pi(x_i|x_{i-1}, \dots, x_1, \theta) = \pi(x_1|\theta) \prod_{i=2}^n \pi(x_i|x_{i-1}, \theta).$$

Then, by multiple applications of Bayes's theorem,

$$\begin{aligned} \pi(\theta|\mathcal{X}) &\propto \pi(\mathcal{X}|\theta) \pi(\theta) = \prod_{i=2}^n \frac{\pi(x_i|x_{i-1}, \theta) \pi(\theta)}{\pi(\theta)} \pi(x_1|\theta) \pi(\theta) \\ &\propto \pi(\theta)^{(1-n)} \pi(\theta|x_1) \prod_{i=2}^n \varphi_i(\theta), \end{aligned} \quad (12)$$

where $\varphi_i(\theta) = c_i^{-1} \pi(x_i|x_{i-1}, \theta) \pi(\theta)$ and c_i is a normalizing constant.

Essentially, in expression (12) the posterior density of the parameters given the full data χ has been decomposed into a product involving densities $\varphi_i(\theta)$, each of which depends only on a pair of data points, $\{x_{i-1}, x_i\}$. The approach is then to apply ABC to draw samples from each $\varphi_i(\theta)$. Compared with 'matching' the whole data set, matching pairs of observations leads to substantially higher acceptance rates. To estimate $\pi(\theta|\mathcal{X})$ we need to estimate the $\varphi_i(\theta)$, which can be done, for example, by using either Gaussian approximations or kernel density estimates.

The factorization (12) essentially amounts to 'inference from multiple sources of data' which Fearnhead and Prangle use to motivate noisy ABC. Theorem 2 in the paper makes a strong case for using noisy ABC in the context of applying expression (12). However, if one can set the tolerance to be very strict—or even zero, as White *et al.* (2012) showed is possible in some settings—then the question of whether to use noisy ABC can be sidestepped.

Howell Tong (*London School of Economics and Political Science*)

- (a) Dr Gavin Ross reminded us in his discussion of Diggle and Gratton (1984) that estimating likelihoods of parameters by using artificially generated data was an old and obvious idea that went back earlier than 1984, e.g. Ross (1972). In fact, the seminal idea is at least as old as Student (1908a, b) which estimated the likelihood of the mean and that of the correlation coefficient via simulation using 3000 criminals!
- (b) George Box tells us that all models are wrong. Thus, instead of studying inferential procedures predicated on a true model, it is practically more relevant to study the *likelihood of a model*, and to develop reliable and likelihood-free estimation of the parameters of a *wrong* model so as to match observed features of the data best. Akaike (1978) and Xia and Tong (2011) are relevant references. In Xia and Tong (2011) inferences for time series data are based on all-step-ahead predictions and others, unlike conventional methods, such as maximum likelihood estimation, which are typically based on just one-step-ahead predictions.
- (c) It seems to me that approximate Bayesian computation and the indirect inference are based on different philosophies, the former being predicated on a true model specification whereas the latter apparently not so. From this perspective, I suggest that the noisy approximate Bayesian computation may be viewed as an attempt at removing the assumption.
- (d) Sections 1.3 and 5 refer to sufficient statistics. Note that sufficiency may be lost if the model is wrong, e.g. Xia and Tong (2011), especially page 23.
- (e) The difficulty of likelihood inference of toy models in the chaos literature is principally due to the singularity and non-uniqueness of the measures on which they live. The difficulty may often be removed by appealing to a well-known Kolmogorov construction (Chan and Tong, 2001).

Christophe Andrieu (*University of Bristol*), **Arnaud Doucet** (*University of Oxford*) and **Anthony Lee** (*University of Warwick, Coventry*)

Exact simulation to tackle intractability in model-based statistical inference has been exploited in recent years for exact inference (Beaumont, 2003; Beskos *et al.*, 2006; Andrieu and Roberts, 2009; Andrieu *et al.*, 2010) (see Gourieroux *et al.* (1993) for earlier work). Approximate Bayesian computation (ABC) is a specialization of this idea to the scenario where the likelihood associated with the problem is intractable but involves an additional approximation. The authors are to be thanked for a useful contribution to the latter aspect. Our remarks are presented in the ABC context but apply equally to exact inference. A simple fact which seems to have been overlooked is that sampling exactly $Y \sim f(y|\theta)$ on a computer most often means

that $Y = \phi(\theta, U)$ where U is a random vector of probability distribution $D(\cdot)$ and $\phi(\cdot, \cdot)$ is a mapping either known analytically or available as a ‘black box’. The vector U may be of random dimension, i.e. $D(\cdot)$ may be defined on an arbitrary union of spaces (e.g. when the exact simulation involves rejections) and is most often known analytically—we suggest taking advantage of this latter fact. In the light of the above we can rewrite the ABC proxy likelihood

$$\tilde{p}(y^*|\theta) = \int_{\mathcal{Y}} K(y, y^*) p(y|\theta) dy$$

in terms of the quantities involved in the exact simulation of Y

$$\tilde{p}(y^*|\theta) = \int_{\mathcal{U}} K\{\phi(\theta, u), y^*\} D(u) du.$$

In a Bayesian context the posterior distribution of interest is therefore

$$\tilde{p}(\theta|y^*) \propto \int_{\mathcal{U}} K\{\phi(\theta, u), y^*\} D(u) du p(\theta).$$

Provided that $D(\cdot)$ is tractable, we are in fact back to the usual, analytically tractable, ‘latent variable’ scenario and any standard simulation method can be used to sample θ and U . Crucially we are in no way restricted to the usual approach where $U_i \sim^{\text{IID}} D(\cdot)$ to approximate the proxy likelihood. In particular, for θ fixed, we can introduce useful dependence between $\phi(\theta, U_1), \phi(\theta, U_2), \dots$, e.g. using Markov chain Monte Carlo (MCMC) sampling of invariant distribution $D(\cdot)$ started at stationarity (Andrieu *et al.*, 2005). The structure of $\tilde{p}(\theta, u|y^*)$ may, however, be highly complex and sophisticated methods may be required. One possible suggestion is the use of particle MCMC methods (Andrieu *et al.*, 2010) to improve sampling on the U -space; for example for a fixed value of θ estimate the proxy likelihood $\int_{\mathcal{U}} K\{y^*, \phi(\theta, u)\} D(u) du$ unbiasedly using a sequential Monte Carlo sampler (Del Moral *et al.*, 2006) targeting a sequence of intermediate distributions between $D(u)$ and $K\{\phi(\theta, u), y^*\} D(u)$ proportional to

$$K_j\{\phi(\theta, u), y^*\} D_j(u)$$

for $\{K_j(\cdot, \cdot), j = 1, \dots, n-1\}$ and $\{D_j(\cdot), j = 1, \dots, n-1\}$ and plug such an estimate in standard MCMC algorithms. Note the flexibility offered by the choice of $\{K_j(\cdot, \cdot)\}$ and $\{D_j(\cdot)\}$ which can allow us to incorporate progressively both the dependence structure on U and the constraint imposed by $K(\cdot, \cdot)$. When $\phi(\cdot, \cdot)$ is known analytically, under sufficient smoothness conditions one can use the infinitesimal perturbation analysis (Pflug, 1996; Andrieu *et al.*, 2005) approach to estimate for example with respect to θ

$$\nabla_{\theta} \int_{\mathcal{U}} K\{\phi(\theta, u), y^*\} D(u) du.$$

Again such ideas equally apply to genuine latent variable models and have the potential to lead to efficient exact inference methods in otherwise apparently ‘intractable’ scenarios.

Kevin J. Dawson (*Rothamsted Research, Harpenden*)

Sufficiency and marginal sufficiency: their relevance to likelihood-free Bayesian computation

I congratulate the authors for their thought-provoking paper. A statistic $T(x)$ is *Bayes sufficient* if it satisfies the condition

$$\pi(\omega|x) = \pi\{\omega|T(x)\}, \quad (13)$$

at all points ω in the parameter space \mathcal{G} , for all priors $\pi(\omega)$ (Kolmogorov, 1942; Yamada and Morimoto, 1992). When we can find such a statistic, it is natural to specify an acceptance region of the form

$$\{x : x \in \mathcal{X}, \|T(x) - T(X_0)\| \leq \delta\}, \quad (14)$$

in the sample space \mathcal{X} . (Here $\|\cdot\|$ denotes the Euclidean norm.) The rejection sampling method would then generate a large sample of observations Ω_i from an approximation to the posterior density $\pi(\omega|X_0)$. The higher the dimension d of the statistic $T(x)$, the lower is the relative volume of the acceptance region (of radius δ).

A sample is a convenient form in which to store information about the density of a high dimensional variable (parameter), such as the posterior $\pi(\omega|X_0)$, because we can easily extract a sample from any low dimensional marginal, $\pi(\theta|X_0) = \pi\{\Theta(\omega)|X_0\}$, simply by taking each observation Ω_i , and applying the mapping $\Theta_i = \Theta(\Omega_i)$. From this perspective, the sample of observations Ω_i is an intermediate step.

We could specify an acceptance region to be used directly in a rejection sampling method for a low dimensional marginal $\pi(\theta|X_0)$. For this, we can make do with a statistic $T(x)$ which satisfies the condition

$$\pi(\theta|x) = \pi\{\theta|T(x)\}, \quad (15)$$

at all points $\theta = \Theta(\omega)$, in the parameter space $\Theta(\mathcal{G})$, for the chosen prior $\pi(\omega)$ (or family of priors). The statistic $T(x)$ is said to be *marginally sufficient* for the parameter $\theta = \Theta(\omega)$, with respect to $\pi(\omega)$ (Raiffa and Schlaifer, 1961, 2000; Basu, 1977).

If we can find such a statistic, then we can again specify an acceptance region of the form (14) and use the rejection method to generate a large sample of observations Ω_i from an approximation to $\pi\{\omega|T(X_0)\}$ (which is not necessarily the posterior density $\pi(\omega|X_0)$, and hence a large sample of observations $\Theta_i = \Theta(\Omega_i)$ from an approximation to $\pi\{\theta|T(X_0)\} = \pi(\theta|X_0)$ (Bazin *et al.*, 2010). We could do this separately for each parameter of interest $\theta = \Theta(\omega)$. We are still left with the problem of finding a statistic $T(x)$ of relatively low dimension, which (approximately) satisfies condition (15). The ideas in Section 2.3 and 3 may help here. However, from this perspective it appears that, when we are approximating the marginal posterior of a (one-dimensional) parameter $\theta = \Theta(\omega)$, we may be wasting dimensions of the summary statistic $T(x)$ if we include estimates of posterior means of other (one-dimensional) parameters. A better strategy may be to use estimates of the posterior moments, cumulants or quantiles, of the parameter θ alone, as components of the statistic $T(x)$.

The following contributions were received in writing after the meeting.

Chris P. Barnes, Sarah Filippi and Michael P. H. Stumpf (*Imperial College London*)

If we assume that the optimal summary statistics are the posterior means we can consider the joint posterior distribution of the summary and parameters

$$\pi_{ABC}(\theta, \beta, \Sigma | \mathbf{s}_{\text{obs}}) \propto \pi(\theta) \pi(\beta, \Sigma) \int \pi\{\mathbf{S}(\mathbf{y}) | \mathbf{y}, \beta, \Sigma\} \pi(\mathbf{y} | \theta) K[\{\mathbf{S}(\mathbf{y}) - \mathbf{s}_{\text{obs}}\} / h] d\mathbf{y},$$

where β and Σ are the additional parameters due to the regression of \mathbf{y} on θ .

Obviously, a fully Bayesian approach to jointly estimating the summaries and parameters is possible but would defeat the purpose of using summary statistics. The approach that Fearnhead and Prangle take is to perform a linear regression using a set of simulations taken from a region of reasonable posterior probability, and then to replace β and Σ by their maximum likelihood values, $\hat{\beta}$ and $\hat{\Sigma}$, obtained by fitting the model

$$S(\mathbf{y}) = Y_{\mathbf{y}}\beta + \varepsilon \quad \varepsilon \sim N(0, \Sigma),$$

where $Y_{\mathbf{y}}$ is the design matrix and can contain any choices of $f(\mathbf{y})$. They constrain Σ to be diagonal though perhaps this is an unnecessary restriction.

When viewed from this perspective we can see two possible extensions to Fearnhead and Prangle's approach.

- The uncertainty in the linear model is currently neglected and could be approximated by sampling from the multivariate normal distribution associated with the linear regression, i.e. by replacing the deterministic function $S(\mathbf{y}) = Y_{\mathbf{y}}\hat{\beta}$ with the distribution $S(\mathbf{y}) \sim \text{MVN}(Y_{\mathbf{y}}\hat{\beta}, \hat{\Sigma})$.
- It should be possible to incorporate the summary statistic estimation or choice step in any existing Markov chain Monte Carlo and sequential Monte Carlo algorithms, thereby simultaneously estimating the summary statistics and π_{ABC} . For Markov chain Monte Carlo approaches selection of summaries can be performed most straightforwardly during the burn-in (when time reversibility is not required); otherwise it is feasible to construct Metropolis–Hastings acceptance schemes which also fulfil detailed balance. In a sequential Monte Carlo algorithm we have a new set of simulations in each population, which we can use to estimate β and Σ ; (importance) weights need to be chosen with care, however, as they will depend implicitly on the set of chosen statistics.

Simon Barthelmé (*Berlin University of Technology*), **Nicolas Chopin** (*Centre de Recherche en Economie et Statistiques, Malakoff, and Ecole Nationale de la Statistique et de l'Administration Economique, Paris*), **Ajay Jasra** (*National University of Singapore*) and **Sumeetpal S. Singh** (*University of Cambridge*)

We strongly believe that the main difficulty with approximate Bayesian computation (ABC) type methods is the choice of summary statistics. Although introducing summary statistics may be sometimes beneficial (Wood, 2010), in most cases this induces a bias which is challenging to quantify. We thus welcome this important work on automatically choosing summary statistics. The fact remains that the optimality criterion that is proposed in the paper is a little limiting; we want to approximate a full posterior distribution, not simply the posterior expectation. In addition, the approach proposed does not offer a way to monitor the bias that is induced by the optimal set of summary statistics, except by numerically comparing many alternative summary statistics, which is potentially tedious.

It is perhaps useful to note that there are now ABC methods that do not use summary statistics, at least for certain classes of models. The expectation propagation (EP)–ABC algorithm of Barthelmé and Chopin (2011) is a fast approximation scheme for ABC posteriors based on constraints of the form $\|y_i - y_i^*\| < \varepsilon$. It is typically orders of magnitude faster than Monte-Carlo-based ABC algorithms, while, in some scenarios, featuring an approximation error that is smaller, due to the absence of summary statistics. It is currently limited, however, to models such that the y_i may be simulated sequentially by using some chain rule decomposition.

For hidden Markov models, ‘exact’ ABC inference (i.e. not relying on either summary statistics or an approximation scheme) may be achieved as well, via the hidden Markov model (HMM)–ABC approach of Dean *et al.* (2010) and Dean and Singh (2011) (see also McKinley *et al.* (2010)), which show that an ABC posterior may be reinterpreted as the posterior of an artificial HMM, where the observations are corrupted with noise. This interpretation makes the remark of Wilkinson (2008) even more compelling: without summary statistics, an ABC posterior may be interpreted as the correct posterior of a model where the *actual data* (as opposed to the summary statistics) are corrupted with noise. For instance, the Ricker model example, and with some adaptation for the Lokta–Volterra example of the paper.

These two approaches already cover many ABC applications and could be applied directly to three examples of the paper: g -and- k -distributions (EP–ABC), Lokta–Volterra processes (EP–ABC, HMM–ABC with a slight modification) and the Ricker model (HMM–ABC). We are currently working on extending this work in other dependence structures for the observations and we hope that others will also join us in this effort to remove summary statistics in ABC.

M. G. B. Blum and O. François (*Université Joseph Fourier, Grenoble*)

The main focus of Fearnhead and Prangle is to construct optimal summary statistics by approximating the conditional mean of the model parameters given the data. The examples suggest the use of linear regression where the predictors are powers of order statistics, and the number of predictors is chosen by the Bayesian information criterion BIC.

In previous work, we introduced similar ideas where summary statistics were constructed through non-linear regression using adaptive basis functions coupled with regularization and cross-validation (Blum and François, 2010). A motivation of our approach was to reduce the dimensionality of the original set of summary statistics via internal projections on lower dimensional subspaces. The non-linear regression results improved on the linear correction results of Beaumont *et al.* (2002) in several examples.

We compared results of simulations on the gk -model by using the R package `abc` (Csilléry *et al.*, 2012) and the R package `gk` written by Dennis Prangle. A simulation involved the generation of 100 evenly spaced order statistics obtained from 10000 independent draws from the g -and- k -distribution. We performed a total of 10000 simulations using a uniform distribution over (0,10) for each parameter. We evaluated the mean-squared error of point estimates using 50 data sets generated with the parameter values $(A, B, g, k) = (3, 1, 2, 0.5)$ and implemented a tolerance rate of 0.5%. With the exception of g , we found that using an adaptive basis reduced estimation errors when compared with a rejection algorithm using the summary statistics provided by Fearnhead and Prangle (Table 11).

Polynomial regressions are known to suffer from non-local behaviour where the value of the response corresponding to a particular value of the predictor may have a large and undesirable influence on the predicted response for a very different value of the predictor (Magee, 1998). More generally, the projection method of Fearnhead and Prangle could be improved by considering other types of predictors than polynomial functions.

Table 11. Relative errors of point estimates with respect to a rejection algorithm that uses 100 order statistics†

<i>Method</i>	<i>A (%)</i>	<i>B (%)</i>	<i>g (%)</i>	<i>k (%)</i>
Non-linear adjustment	–81	–82	–51	–80
Semi-automatic ABC	–40	–16	–77	–63

†The relative values for semi-automatic ABC are taken from Table 3 of the paper.

Table 12. Mean quadratic losses for various analysis of 50 $M/G/1$ data sets with $\theta_1 = 1$, $\theta_2 = 7$ and $\theta_3 = 0.1$

<i>Method</i>	θ_1	θ_2	θ_3
Semi-automatic ABC	0.092	2.0	0.0082
Artificial neural network + ABC	0.035	1.5	0.00099

Yining Chen (*University of Cambridge*)

I thank the authors for their stimulating contribution to approximate Bayesian computation. I shall limit my comments to a brief discussion of the authors' choice of summary statistics.

The dimension of the summary statistics

Theorem 4 of the paper states a theoretical result by taking $S(\mathbf{y}) = E(\theta|\mathbf{y})$. However, in practice this optimal choice of summary statistics is unknown. In view of this, even in the limit as $h \rightarrow 0$, an extra bias term needs to be taken into consideration in the decomposition of the quadratic loss, namely

$$E[L\{\theta, E(\theta|\mathbf{s}_{\text{obs}}); A\}|\mathbf{y}_{\text{obs}}] - E[L\{\theta, E(\theta|\mathbf{y}_{\text{obs}}); A\}|\mathbf{y}_{\text{obs}}].$$

In many cases, if $S(\mathbf{y})$ is not a sufficient statistic for θ , the above quantity is strictly greater than 0 and commonly decreases as the dimensionality of the summary statistics increases. Therefore, the choice of d is a balance between the information loss and the Monte Carlo error. Setting d to be the dimension of the parameter space by default might be suboptimal.

Beyond the linear regression approach

The linear regression approach is an innovative attempt made by the authors to estimate $E(\theta|\mathbf{y})$. The performance of this approach depends somewhat on the quality of the pilot run. Indeed, this approach works well if all the simulated observations are concentrated in a small area so that linear models can nicely summarize the local behaviour of $E(\theta|\mathbf{y})$. Otherwise, if $E(\theta|\mathbf{y})$ is far from linear, considerable effort is required to model $E(\theta|\mathbf{y})$. To make the procedure more flexible, we suggest using an artificial neural network instead of the linear regression in step (c) to estimate the summary statistics. In fact, other non-parametric methods may also work. As an illustration, we reran the $M/G/1$ -queue example in Section 4.5. Most of the settings follow the original paper, except that we implemented algorithm 1 without the pilot run and used only 2×10^6 simulated data sets. For simplicity, we applied the artificial neural network with one hidden layer containing 15 nodes. The true parameters that were used in the simulation are $\theta_1 = 1$, $\theta_2 = 7$ and $\theta_3 = 0.1$. Results are reported in Table 12. Clearly, our new estimator can be an improvement, though admittedly it is computationally more expensive.

David Draper (*University of California, Santa Cruz*)

I approach the discussion of this interesting paper from the perspective of a potential user of the method proposed here who has no previous experience working on problems like those posed in the paper (in which simulating from the model is easy but calculating likelihood values is difficult; call this problem class (*)), and I offer the authors three questions from that perspective.

- Given that the sets of tricks and tweaks that led to the best performance for semi-automatic approximate Bayesian computation (ABC) varied across the examples of Section 4, how should a new user proceed?
- Let us define as *automatic* an analysis approach that produces accurate results without any need for user-chosen inputs or user-driven adaptive choice of such inputs. How close to *automatic* is semi-automatic ABC on a completely new problem,
 - for the authors of this paper?
 - for other experienced Bayesian researchers who have no previous exposure to semi-automatic ABC?
 - for researchers with limited experience in Bayesian data analysis?
- If I wish to use the authors' method on a new {data set D , model M }, from problem class (*) but, unlike the examples in Section 4, it would seem that a good course of action would be as follows:

- (i) find, by some means (how?), a set of parameter values θ^* that generates data sets D^* from M that are in some sense similar to D ;
- (ii) conduct a small programme of original research, something like what the authors have done in Section 4, aimed at identifying a version of semi-automatic ABC that performs well across many simulated D^* replicates; and only then
- (iii) use the approach derived in (ii) to solve my original scientific problem.

Is this approximately the state of play, if I wish to have some assurance that the authors' method will perform well in my problem, or have I missed something? (It may be that the course of action in (c) is a decent answer to my questions (a) and (b), but perhaps the authors have a better answer that they can provide in their rejoinder. I ask all these questions in a friendly spirit; I am genuinely interested to know the authors' views on how close to automatic their method is at present, and how we can become closer to automatic in the future.)

Christopher C. Drovandi and Antony N. Pettitt (*Queensland University of Technology, Brisbane*)

The authors present an approach that promises to deliver something that every exponent of approximate Bayesian computation (ABC) wishes for: a low dimensional close-to-sufficient summary statistic in some automated and computationally convenient way. For this the authors should be congratulated.

It is of general interest whether the semi-automatic method may be sensitive to the first-stage algorithm, in terms of the tolerance and the initial choice of summary statistics. A lower tolerance can be obtained

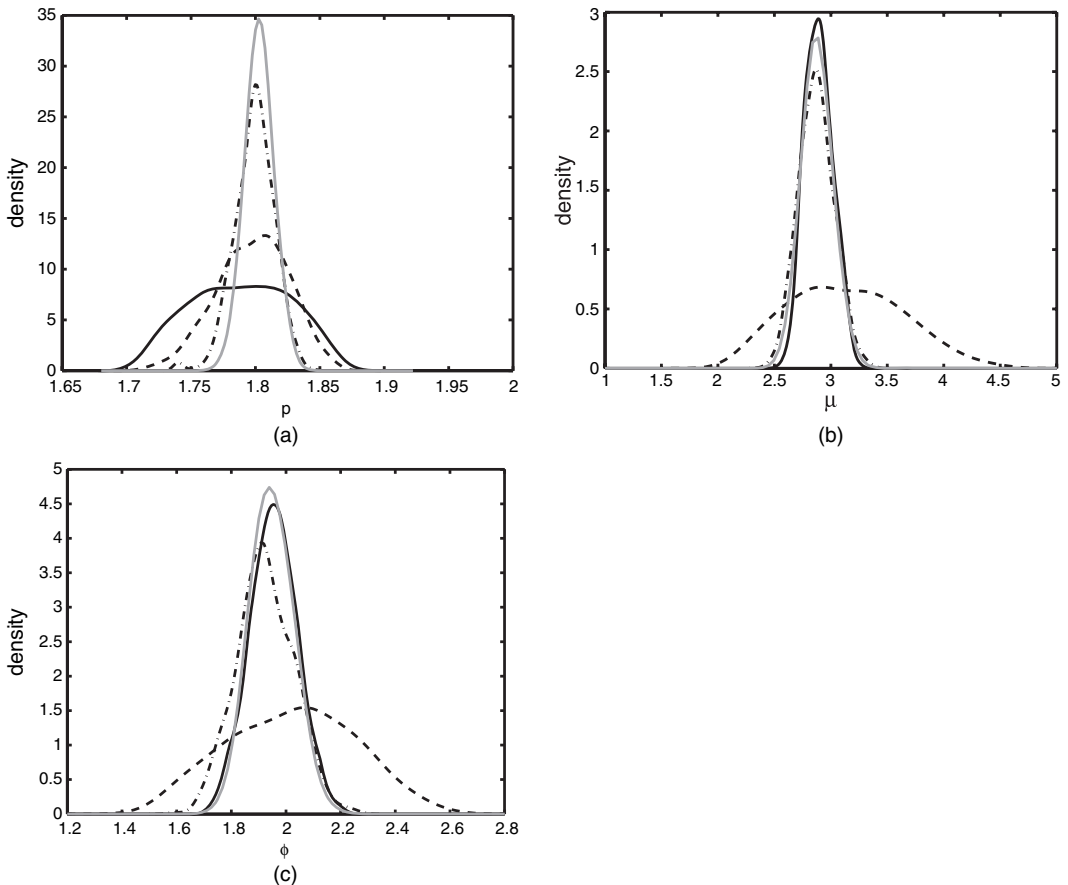


Fig. 4. Posterior distributions for the Tweedie model (— — —, full; · — · —, automatic; —, semi-automatic; —, true): (a) posterior for p ; (b) posterior for μ ; (c) posterior for ϕ

with sequential Monte Carlo (SMC) ABC (e.g. Drovandi and Pettitt (2011a)). An initial set of summaries could be obtained via Nunes and Balding (2010), which is an automatic algorithm to choose the optimal subset of summaries out of a larger set of summaries.

The second part of this discussion is to compare the semi-automatic and Nunes and Balding (2010) approaches on the Tweedie model (a generalized linear model with $E[Y] = \mu$ and $\text{var}(y) = \phi\mu^p$) where $1 < p < 2$, $\mu > 0$ and $\phi > 0$. Having $1 < p < 2$ produces a mixture of 0s and continuous response. 1000 observations are drawn with $p = 1.8$, $\mu = 3$ and $\phi = 2$. Our priors are uniform. The likelihood is analytically intractable but a suitable approximation has been proposed by Dunn and Smyth (2005). Simulation is straightforward for some of the parameter space. The initial summaries considered are the number of 0s and, for the non-zero component of the data, minimum, maximum, the octiles and robust versions of scale, skewness and kurtosis in Drovandi and Pettitt (2011b). A uniform weighting function is used with a Euclidean distance discrepancy on the log-summaries.

SMC ABC was applied with the initial set of summaries (referred to as ‘full’), to obtain the training region for semi-automatic ABC. We estimated the true posterior via importance sampling (‘true’).

For the semi-automatic ABC, a regression was applied where the predictors were every 100th-order statistic and the square of these. The observed summaries estimated from the regression are close to the true posterior means. We then performed SMC ABC with these summaries and the updated prior (‘semi-automatic’).

The Nunes and Balding (2010) algorithm selected the number of 0s, the second, third, sixth and seventh octile, as well as the robust measures of scale, skewness and kurtosis summaries. SMC ABC is used to refine these particles to a lower tolerance (referred to as ‘automatic’).

Fig. 4 shows the results. Semi-automatic ABC does well for μ and ϕ but overestimates the posterior variance for p . To improve the approximation it may be of interest to include additional features of the posterior distribution (e.g. the variance) via more regressions and to include these extra summary statistics.

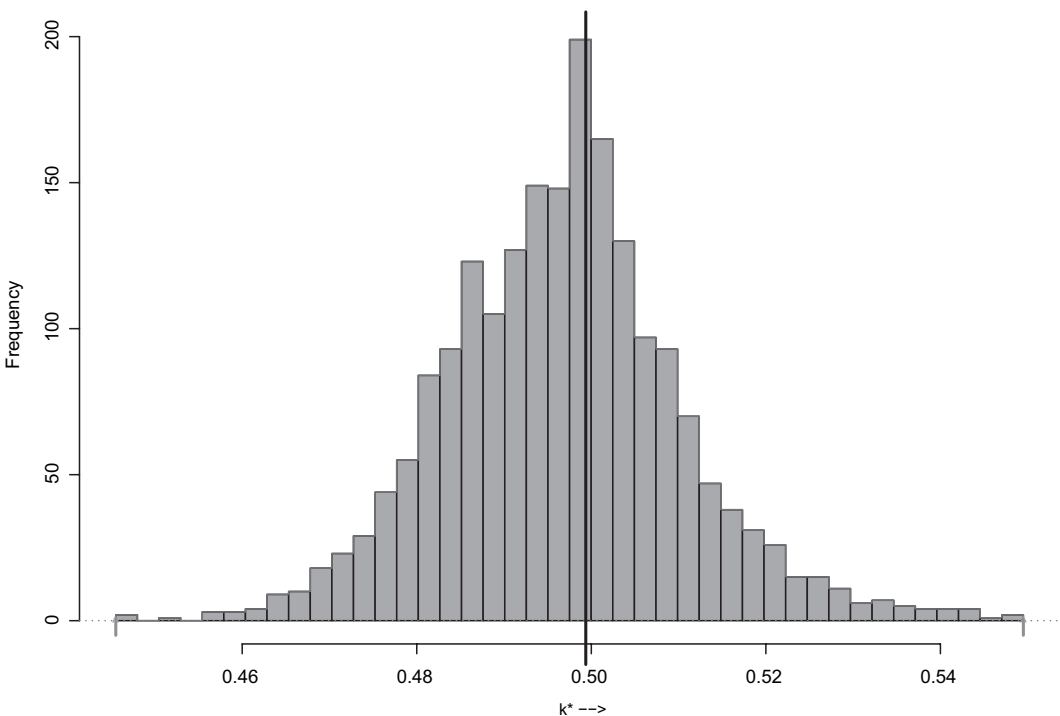


Fig. 5. 1000 parametric bootstrap replications of the maximum likelihood estimate for k in the g -and- k -distribution, Section 4.2: mean 0.497; standard deviation 0.014

As an aside, we are interested in the author's comments on how the semi-automatic approach might be extended to accommodate models with covariates (see for example Drovandi *et al.* (2011)).

Bradley Efron (*Stanford University*)

The authors make a nice case for approximate Bayesian computation (ABC), in both its old and its new formulations, at least in situations where there is genuine prior information. I am less convinced though when prior information is absent, and $\pi(\theta)$ is assigned according to some conventional 'uninformative' method. Such applications, which are now common in the literature, are not fully Bayesian and need to compete with frequentist approaches.

The g -and- k -distribution of Section 4.2 provides a simple example. Following the stipulations of Allingham *et al.* (2009) in the third paragraph, I carried out a parametric bootstrap analysis based on a good approximation to the maximum likelihood estimator $\hat{\theta} = (\hat{A}, \hat{B}, \hat{g}, \hat{k})$, i.e. I sampled data sets of size $n = 10^4$ according to $F^{-1}(x; \hat{A}, \hat{B}, c = 0.8, \hat{g}, \hat{k})$, each time obtaining a bootstrap replication θ^* . $B = 1000$ such replications took 90 s of computer time, using off-the-shelf maximum likelihood estimator algorithms.

The 1000 bootstraps of \hat{k} are shown in Fig. 5 The histogram has a long-tailed normal shape, with bootstrap standard error estimate $\hat{\sigma}_{\text{boot}} = 0.014$. This is close to the semi-automatic ABC estimate 0.015 ($= 0.00023^{1/2}$), in the third row of Table 3. (The only notable disagreement was with \hat{g} , where the ABC standard error was about seven times too large.)

As the authors emphasize in Section 3.1 we need to quantify the uncertainty in our point estimates. The bootstrap histogram gives such quantification at least to a first order of accuracy. To go further we could make the kinds of 'BCa' -corrections in Efron (1987). For instance, 57% of the 1000 \hat{k}^* -values were less than \hat{k} , indicating a substantial downward 'median bias' that would need correction to achieve second-order accuracy.

ABC methods also employ parametric resampling. The difference is that they do so by sampling from a range of possible θ -values, whereas the parametric bootstrap resamples only from a $\theta = \hat{\theta}$. The bootstrap-ABC comparison seems natural in this respect.

Matteo Fasiolo, Natalya Pya and Simon N. Wood (*University of Bath*)

We are impressed by the key idea that the best statistics are estimates of the parameters. To investigate whether such statistics would improve the synthetic likelihood approach (Wood, 2010), we tried to replicate and extend Table 6. We found two difficulties. Firstly, it was not easy to get the semi-automatic approximate Bayesian computation (ABC) regression step right for a model as non-linear as the Ricker model, and to obtain reasonable results we had to tune the size of the parameter space region over for the linear regression step carefully. Secondly, the results for all methods were each dominated by 1–3 spectacular failures, making the mean quadratic loss over the replicates too volatile a measure to use for comparison: different runs of the whole study gave us substantially different answers about which method performed best.

For the Ricker model, the difficulty in getting the regression step right may reflect the chaotic nature of the dynamics, and it might be better to use simple auto-regressions to estimate the parameters. We weaken the characterization of the statistics as estimates of the posterior means by doing this, but it seems implausible that this will really much degrade the information content of such statistics. Logging the Ricker model suggests the linear regression

$$\log(Y_{t+1}/Y_t) = \log(r) - Y_t/\phi + d_t$$

for \hat{r} and $\hat{\phi}$ (there will be many missing data). d_t contains variability from the original process noise and the Poisson variability, suggesting that σ_e can be estimated by the outrageous expedient of averaging the finite values of

$$\max[0, d_t^2 - \{Y_{t+1}^{-1} + (\hat{\phi}^{-1} - Y_t^{-1})^2 Y_t\}]$$

we refer to \hat{r} and $\hat{\phi}$ and $\hat{\sigma}_e$ as 'Fearnhead–Prangle (FP)' statistics below.

To reduce volatility, we modified the study by using a uniform grid of 50 $\log(\sigma_e)$ -values on the interval $[\log(0.1), 0]$. Priors were uniform on $\log\{\log(r)\}$ -, $\log(\sigma_e)$ - and $\log(\phi)$ -scales, with truncations $1 < \log(r) < 10$, $0.02 < \sigma_e < 2$ and $1 < \phi < 20$. For a loss measure that gives more repeatable results between replicates of the study, we took the squared difference between the true parameter values and the corresponding mean from the converged chain, but used the *geometric mean* of this across the 50 replicates (downweighting the influence of the extreme failures). We obtain the following results. Confidence interval coverage probabilities were around 80–90% for all methods (Table 13). (Note that the Ricker model has a rather special structure that actually allows direct Markov chain Monte Carlo sampling using N_t as an auxiliary variable, so we include this also.)

Table 13. Comparative results from alternative methods

<i>Method</i>	$\log(r)$	σ_e	ϕ
Synthetic likelihood	0.016	0.0082	0.18
Synthetic likelihood + FP statistics	0.014	0.0057	0.15
Direct Markov chain Monte Carlo sampling	0.009	0.0054	0.09
ABC	0.013	0.0087	0.244
ABC + FP statistics	0.009	0.0091	0.127
ABC + regression	0.009	0.007	0.088
Semi-automatic ABC	0.012	0.0079	0.102

So adding FP statistics to synthetic likelihood improves results. A possible advantage of using FP statistics with synthetic likelihood rather than ABC is that most effort is devoted to obtaining scientifically meaningful statistics, rather than tuning, and model comparison comes ‘for free’, but of course the computational cost of this may be high.

Sarah Filippi, Chris P. Barnes and Michael P. H. Stumpf (*Imperial College London*)

Fearnhead and Prangle construct summary statistics by using the quadratic loss function. Minimizing this criterion leads to choosing a summary statistic that is equal to the posterior mean, which may be estimated through linear regression. As they argue,

‘the use of the least square error loss leads to an ABC approximation that attempts to have the same posterior mean as the true posterior’.

But instead of the posterior mean we may be interested in other properties of the posterior distribution. For example, we may aim to obtain an approximation of the posterior probability distribution which enables us to obtain predictions as well as if they were coming from the true distribution. From this perspective the loss function of interest is to measure how well the statistic captures the information on the parameter θ that is relevant for predicting data X . In information theory, a statistic $S(\cdot)$ is said to be sufficient if the mutual information between the random variables Θ and X , which is denoted by

$$I(\Theta; X) = \iint p(\theta, x) \log \left\{ \frac{p(\theta, x)}{\pi(\theta)p(x)} \right\} d\theta dx,$$

is equal to $I\{\Theta; S(X)\}$. We denote by $\pi(\theta)$ the prior distribution and by $p(x)$ the evidence. A loss function could then be $I(\Theta; X) - I\{\Theta; S(x)\}$. To determine a subset of statistics which minimizes this loss function given a set of summary statistics, one could simulate the parameters values and data by using importance sampling. Let us denote by $\{\theta^{(i)}, x^{(i)}\}_{1 \leq i \leq N}$ the resulting N -sample and by $\{\omega^{(i)}\}_{1 \leq i \leq N}$ the corresponding weights. From these the mutual information is easily estimated:

$$I\{\Theta; S(X)\} \approx \sum_{i=1}^N \omega^{(i)} \log \left\{ \frac{\omega^{(i)}}{\pi(\theta^{(i)}) p(x^{(i)})} \right\},$$

where $p(x^{(i)})$ may, for example, be estimated by using a kernel approximation, e.g.

$$p(x^{(i)}) \approx \sum_{k=1}^N \kappa(x^{(i)} | x^{(k)})$$

where $x^{(k)} \sim f(\cdot | \theta^{(k)})$ and $\theta^{(k)} \sim \pi(\theta)$; $\kappa(\cdot | \cdot)$ is a transition kernel. In practice this may be computationally expensive since it requires the computation of the value of the loss function for all subsets of the set of statistics at hand.

Barnes *et al.* (2011) also use an information theory framework for constructing a minimal set of summary statistics which is sufficient (for parameter estimation and model selection). However, they take into account the observed value y and then look for statistics such that the information contained in $S(y)$ about the parameter θ is equal to the information contained in y . Doing so minimizes the Kullback–Leibler divergence between the posterior distribution probabilities $p\{\Theta | S(y)\}$ and $p(\Theta | y)$. Fearnhead and Prangle focus on the construction of summary statistics without considering the observed value; the approach outlined above does the equivalent, but by minimizing $I(\Theta; X) - I\{\Theta; S(X)\}$.

Andrew Gelman (Columbia University, New York)

I like the idea of using graphical and numerical summaries of closeness of simulated to real data. Once you are doing this, can you also check the fit of your model? The methods in the paper under discussion seem to work on the basis of the assumption that the model is correct. But, if none of the simulations look like a given data set, perhaps it would be useful to record this lack of fit.

Mark Girolami and Julien Cornebise (University College London)

We congratulate the authors for their excellent paper. We would like to suggest consideration of cases where it would make sense, from the perspective of statistical inference, to focus directly on $p(\theta|\mathbf{s})$, i.e. to base inferences on the preprocessed summarized data \mathbf{s} , rather than on the raw data \mathbf{y}_{obs} . Such a practice is standard in fields such as statistical discriminant analysis, pattern recognition, machine learning and computer vision, where preprocessing such as feature extraction (see for example Lowe (2004)), edge detection and thresholding are routine, or in medical signal processing (e.g. magnetic resonance imaging), where inference occurs on preprocessed output of the medical instrument. Wood (2010) focused on qualitative descriptors of noisy chaotic dynamic systems presenting strong dependence on the initial conditions, with applications to ecological models: the primary interest for the user of these models is the characteristics of the trajectory (regularity, pseudoperiod, maxima, extinction of the population, ...), not its actual path.

Statistically speaking, as illustrated in the directed acyclic graph of Fig. 6, this is nothing but shifting the model one layer down the hierarchical model, permuting the role of \mathbf{y} and \mathbf{s} as auxiliary simulation variable and variable of interest, with the advantage of removing the proxy approximation: the summary statistics are not an approximation anymore, but the actual focus of interest. This is reminiscent of discriminative–generative modelling (see for example Xue and Titterton (2010) and Hopcroft *et al.* (2010)). The choice of those statistics then becomes either a modelling problem based on domain-specific expertise or, drawing further on the comparison with computer vision, a matter of sparse base construction as recently developed in compressed sensing (Candès and Wakin, 2008).

The only remaining layer of approximation is that of density estimation by the kernel K . Unfortunately, this kernel density estimation is only *asymptotically* unbiased and is biased for *finite* sample size; the Metropolis–Hastings ratio in approximate Bayesian computation–Markov chain Monte Carlo sampling

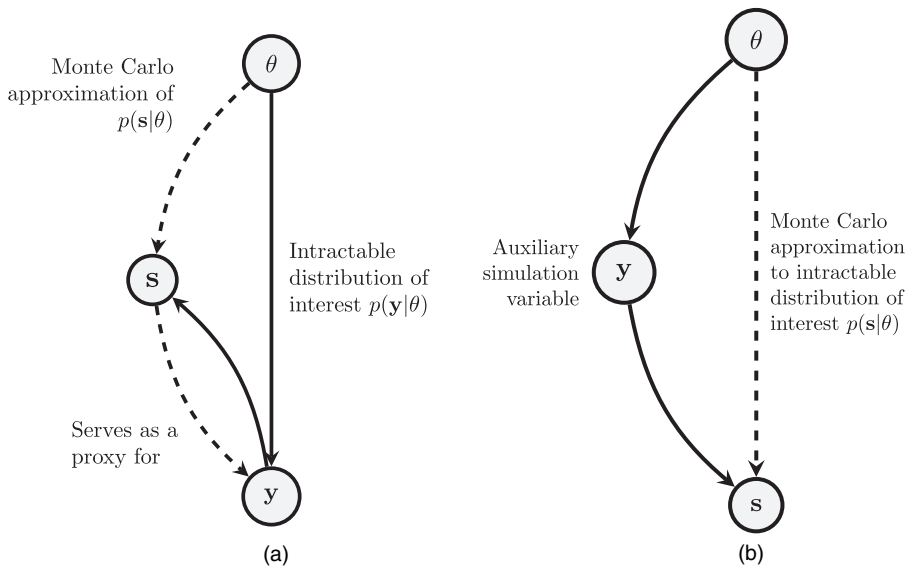


Fig. 6. Graphical representation of the two possible uses of approximate Bayesian computation (the roles of the data \mathbf{y} and of the summary \mathbf{s} are inverted; plain arrows represent distributions from which it is easy to sample; annotated broken arrows represent logical relations): (a) classical use (inference based on the raw data \mathbf{y} ; the summary statistics \mathbf{s} serve to compute a Monte Carlo estimate of $p(\mathbf{s}|\theta)$ as a proxy for the intractable likelihood $p(\mathbf{y}|\theta)$); (b) possible complementary use (inference based on the summarized data \mathbf{s} ; the raw data \mathbf{y} serve as an intermediate simulation step)

(Table 2) cannot be cast in the expected auxiliary variable of Andrieu *et al.* (2007) extending Andrieu and Roberts (2009), not yet available but summarized in Andrieu *et al.* (2010), section 5.1.

Florian Hartig (*Helmholtz Centre for Environmental Research, Leipzig*)

I greatly enjoyed reading this paper. A general method for selecting appropriate summary statistics is not only a substantial theoretical advance but is also likely to increase the appeal of approximate Bayesian computation (ABC) for consumers of inferential methodology. My comments will be concerned mostly with the latter: what are the properties of semi-automatic ABC (SABC) in typical, applied settings?

Many previous approaches, including indirect inference, have based the choice of summary statistics in some way on the observed data. I perceive it as an interesting shift that Fearnhead and Prangle, albeit using the observed data for selecting the relevant parameter space, determine optimal summary statistics entirely based on simulations from the hypothesized model. This is elegant, and in line with the general philosophy of ABC. However, although my sympathy is completely with this approach, I fear that deriving all inferential decisions (mean effects, error model and summary statistics) from the hypothesized model may leave SABC particularly exposed to structural errors in this model. *A priori*, it is not clear to me whether optimal summary statistics for the simulated data remain optimal or at least good when applied to observed data that have different properties from those of the simulated data. Clearly, it would be unfair to criticize a method for not being able to deal with gross model misspecifications, but I feel that a better understanding of the robustness of SABC towards typical levels of structural error would be helpful.

My second remark concerns Section 2.4, in which standard and ‘noisy’ ABC are compared. The authors remark earlier on the work of Wilkinson (2008), who proposed to interpret the ‘noise’ $h\mathbf{x}$ (equation (8)) as the result of an additional error model that is placed on top of the actual ABC model. This idea, however, is not revived in Section 2.4, which evolves around minimizing quadratic loss. Although I fully acknowledge the importance of the latter, I think that the ideas of Wilkinson (2008) deserve further attention in this context. Observation errors are often quite small in genetic applications, but they can be large and well known in other fields that are still to be conquered by ABC. Thus, there could be a good chance that $h\mathbf{x}$ can be fixed in many practical applications.

Ajay Jasra (*National University of Singapore*)

I congratulate the authors on a thought-provoking paper. My main remark is on the trade-off between the bias of the approximate Bayesian computation (ABC) approximation and the accuracy of the Monte Carlo approximation; I restrict my comments to standard ABC, but one can easily extend them to noisy ABC. Perhaps the most natural way to treat the error of a Monte-Carlo-based estimate is as follows, as noted in Marin *et al.* (2011) given a test function $\varphi: \Theta \rightarrow \mathbb{R}$ we would like to study, for $p \geq 1$,

$$\mathbb{E} \left[\left| \sum_{i=1}^N \tilde{w}_i \varphi(\theta_i) - \pi(\varphi|\mathbf{s}_{\text{obs}}) \right|^p \right]^{1/p}$$

where \mathbb{E} is the expectation with respect to the randomness generated by Monte Carlo sampling (Markov chain Monte Carlo and sequential Monte Carlo (SMC) sampling etc.), $\pi(\varphi|\mathbf{s}_{\text{obs}}) := \int_{\Theta} \varphi(\theta) \pi(\theta|\mathbf{s}_{\text{obs}}) d\theta$ and \tilde{w}_i are normalized weights (e.g. $1/N$ for Markov chain Monte Carlo sampling). Then we have

$$\mathbb{E} \left[\left| \sum_{i=1}^N \tilde{w}_i \varphi(\theta_i) - \pi(\varphi|\mathbf{s}_{\text{obs}}) \right|^p \right]^{1/p} \leq \mathbb{E} \left[\left| \sum_{i=1}^N \tilde{w}_i \varphi(\theta_i) - \pi_{\text{ABC}}(\varphi|\mathbf{s}_{\text{obs}}) \right|^p \right]^{1/p} + |\pi_{\text{ABC}}(\varphi|\mathbf{s}_{\text{obs}}) - \pi(\varphi|\mathbf{s}_{\text{obs}})|. \quad (16)$$

This is a decomposition into Monte Carlo error, which can be studied by the quantitative bounds in the appropriate area (e.g. Del Moral (2004) for SMC sampling) and the bias, which can be studied for example by expression (3) in the paper. For hidden Markov models (HMMs), where the summary statistic is the identity (to facilitate inference see Jasra *et al.* (2012) and McKinley *et al.* (2009)), I have considered this issue in Martin *et al.* (2012) for ABC smoothing via SMC sampling by using the approach in Del Moral *et al.* (2010). In that context, one can deal with the SMC error using ideas in Del Moral *et al.* (2010) and the bias by using the backward-in-time representation of the HMM (i.e. ideas from the literature on HMMs). Martin *et al.* (2012) show that inequality (16) has an SMC error that falls with h and a bias that increases linearly with h ; there is also a linear dependence on the time parameter. I would expect that the analysis of the error associated with ABC methods lies in the domain of specialists in the specific statistical model under study and the associated Monte Carlo algorithm.

Theodore Kypraios (*University of Nottingham*)

I congratulate the authors for a thought-provoking paper on deriving summary statistics for approximate Bayesian computation (ABC) algorithms. They take the approach to weaken the requirement that the ABC posterior (π_{ABC}) is a good approximation to the *true* posterior $\pi(\theta|\mathbf{y}_{\text{obs}})$. This enables them to derive their choice of summary statistics which is equal to the expectation $E[\theta|\mathbf{y}_{\text{obs}}]$. Since this expectation is usually unavailable in practice, they propose an approach (Section 3) which consists of these three steps:

- (a) an optional pilot run to determine a region of non-negligible posterior mass;
- (b) simulating sets of parameter values and data;
- (c) use the simulated sets of parameter values and data to estimate the summary statistics.

Although step (a) is optional, can the authors elaborate on what effect the training region has, i.e. being too informative (or uninformative)? Is there a way to assess the error of this step?

In step (c) the authors found that using linear regression with appropriate functions of the data as predictors worked well. Although I appreciate that such an approach is simple and computationally cheap, one would have to make arbitrary choices for the predictors. Therefore, I wonder whether the authors had considered implementing a non-parametric regression instead.

Finally, what if in some situations we have some intuitive and meaningful summary statistics that could be useful for inferring the parameters? Could the authors' approach be used in conjunction with some user-specified summary statistics? And, will the theoretical results in the paper still be valid?

K. V. Mardia (*University of Leeds*) and **T. Hamelryck** (*University of Copenhagen*)

We found the paper very stimulating. There seems to be a link with the *reference ratio method* (Hamelryck *et al.*, 2010). Let $f(\mathbf{x})$ be the probability density function (PDF) of \mathbf{X} which is unknown but the

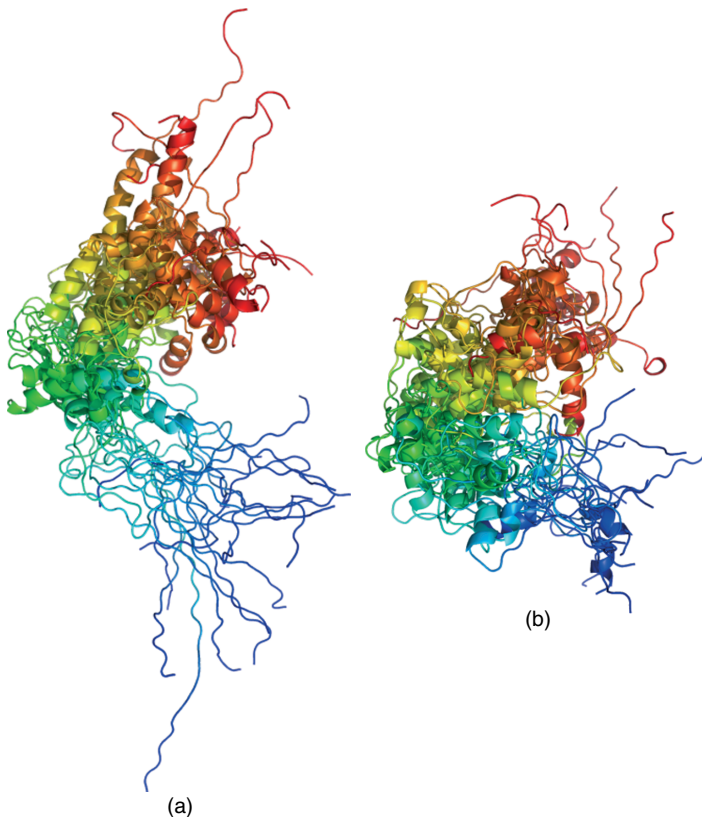


Fig. 7. 20 structures sampled by radius of gyration, (a) without the reference ratio method and (b) with the reference ratio method: the conformations in (b) have the desired distribution of the radius of gyration whereas the conformations in (a) are not sufficiently compact (Mardia *et al.*, 2011)

PDF $g(\mathbf{x})$ which is ‘approximately’ close to $f(\mathbf{x})$ is known (Mardia *et al.*, 2011); $g(\cdot)$ is specified. Also, the true PDF of $Y = m(\mathbf{X})$ from $f(\mathbf{x})$ is known, say $f_1(y)$ where $m(\cdot)$ is given. In physics, \mathbf{X} is called the *fine-grained variable* and Y the *coarse-grained variable*; we have taken Y as a univariate variable but it could be in any subspace of \mathbf{X} . Let $g_1(y)$ be the PDF of y for $g(\cdot)$ and write $\mathbf{x} = (y, \mathbf{z})$. An $f(\cdot)$ which minimizes the Kullback–Leibler divergence to $g(\mathbf{x})$ and which also has the desired marginal for y is given by

$$\hat{f}(\mathbf{x}) = \hat{f}(y, \mathbf{z}) = f_1(y) g_2(\mathbf{z}|y),$$

where $g_2(\mathbf{z}|y)$ is the conditional probability density function of \mathbf{z} given y .

A good example is given in Hamelryck *et al.* (2010) related to protein structure prediction. A protein can be described in terms of two dihedral angles. Let $f(\mathbf{x})$ be the unknown distribution of the dihedral angles $\{(\phi_i, \psi_i), i = 1, \dots, n\}$ for a protein with known sequence of n amino acids. We can generate a distribution of $\{(\phi_i, \psi_i)\}$ from a hidden Markov model that captures protein structure on a local length scale, but that does not capture long-range interactions. We can use the reference ratio method to combine this distribution with a distribution over a coarse-grained variable that describes aspects of long-range structure. We use a coarse-grained variable y corresponding to the radius of gyration with $f_1(y)$, say $N(22 \text{ \AA}, \sqrt{2} \text{ \AA})$. Using the sequence of n amino acids, generate $g(\mathbf{x})$ from the hidden Markov model and obtain $g_1(y)$ through sampling from $g(\mathbf{x})$. Hence, $\hat{f}(\mathbf{x})$ can be obtained. Fig. 7 gives an example where $\hat{f}(\mathbf{x})$ generates conformations that are quite compact compared with the desired distribution over the radius of gyration unlike $g(\mathbf{x})$. We are given $f_1(y)$ and we can only sample efficiently from $g(\mathbf{x})$. From a generative point of view, we want

- (a) to sample y from $f_1(y)$ and
- (b) to sample \mathbf{x} from $g(\mathbf{x}|y)$.

The problem lies in step (b): there is no efficient way to sample from $g(\mathbf{x}|y)$ —we can only sample efficiently from $g(\mathbf{x})$. One could use rejection sampling or the approximate Bayesian computation method for step (b), but that would be very inefficient. Hamelryck *et al.* (2010) used Metropolis–Hastings steps with acceptance probability $\min(1, p)$, $p = f_1(y')/f_1(y) \times g_1(y')/g_1(y)$ where y' denotes new and y old. Now, if accepted, draw a new \mathbf{x}' from $g(\mathbf{x}|y')$.

Jean-Michel Marin (*Université Montpellier 2*) and **Christian P. Robert** (*Université Paris-Dauphine, Institut Universitaire de France, Paris, and Centre de Recherche en Economie et Statistiques, Malakoff*)
Fearnhead and Prangle do not follow the usual perspective of looking at approximate Bayesian computation (ABC) as a converging (both in N and h) approximation to the true posterior density (Marin *et al.*, 2012). Instead, they consider a randomized (or noisy) version of the summary statistics

$$s_{\text{obs}} = S(y_{\text{obs}}) + hx, \quad x \sim K(x),$$

and they derive a calibrated version of ABC, i.e. an algorithm that gives ‘proper’ predictions, but only for the (pseudo)posterior based on this randomized version of the summary statistics. This randomization, however, conflicts with the Bayesian paradigm in that it seems to require adding pure noise to (and removing information from) the observation to conduct inference. Furthermore, theorem 2 is valid for any value of h . We thus wonder at the overall statistical meaning of *calibration*, since even the prior distribution (corresponding to $h = \infty$) is calibrated, whereas the most informative (or least randomized) case (ABC) is not necessarily calibrated. Nonetheless, the interesting aspect of this switch in perspective is that the kernel K used in the acceptance probability, with bandwidth h ,

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

need not behave like an estimate of the true sampling density since it appears in the (randomized) pseudo-model.

As clearly stated in the paper, the ABC approximation is a kernel convolution approximation. This type of approximation has been studied in the approximation theory literature. Typically, Light (1993) introduced a technique for generating an approximation to a given continuous function by using convolution kernels. Also, in Levesley *et al.* (1996), a class of continuous integrable functions is constructed to serve as kernels associated with convolution operators that produce approximations to arbitrary continuous functions. It could eventually be promising to adapt some of the techniques introduced in these papers.

Overall, we remain somewhat sceptical about the ‘optimality’ resulting from this choice of summary statistics as

- (a) practice—at least in population genetics (Cornuet *et al.*, 2008)—shows that proper approximation to genuine posterior distributions stems from using a number of summary statistics that is (much) larger than the dimension of the parameter;
- (b) the validity of the approximation to the optimal summary statistics used as the actual summary statistics ultimately depends on the quality of the pilot run and hence on the choice of the summary statistics therein (this approximation is furthermore susceptible to deterioration as the size of the pilot summary statistics grows) and
- (c) important inferential issues like model choice are not covered by this approach and recent results of ours (Marin *et al.*, 2011) show that estimating statistics are likely to bring inconsistent solutions in this context; those results imply furthermore than a naive duplication of theorem 3, namely based on the Bayes factor as a candidate summary statistic, would be most likely to fail.

In conclusion, we congratulate the authors for their original approach to this major issue in ABC design and, more generally, for bringing this novel and exciting inferential method to the attention of the readership.

Jorge Mateu and Ahmed Arafat (*University Jaume I, Castellón*)

The authors are to be congratulated on a valuable contribution in the context of inference for complex stochastic models. In a wide variety of modern statistical applications, we face the problem that likelihoods are not in closed forms with very complicated expressions. However, it is still possible to simulate from them. In these cases, and to perform statistical inference, we can resort to approximate Bayesian computation (ABC) methods. We would like to bring the authors' attention to a particular problem that could be benefit from this strategy.

Inhomogeneous neural spike train models can be derived by combining a renewal model with a one-to-one transformation function to relate the variable of the renewal probability density to the spike times and the stimulus. We generate inhomogeneous models in which the spike trains have Markov dependence. Let $(0, T]$ denote the observation interval and $0 \leq t_0 < t_1 < \dots < t_k < t_{k+1} < \dots < t_n = T$ be the spikes recorded from a given neuron. The next outcome of the spike activity could be predicted by a probability density of the form

$$p_{\hat{\theta}(t^{k-1})}(t_k | t_{k-1}) = f_t(t_k | t_{k-1}) = \frac{\gamma \lambda(t_k)}{\Gamma(\gamma)} \left\{ \gamma \int_{t_{k-1}}^{t_k} \lambda(u) du \right\}^{\gamma-1} \exp \left\{ -\gamma \int_{t_{k-1}}^{t_k} \lambda(u) du \right\}$$

where $\hat{\theta}(t^{k-1})$ represents the parameter estimator of the prediction strategy given the $(k-1)$ -state process, γ is a parameter and $\lambda(t)$ is a strictly positive intensity function. We could consider for $\lambda(t)$ a spatiotemporal model for place cell spiking in which the intensity of the stimulus response is modelled as a Gaussian function

$$\lambda\{t|x(t), \mu, \alpha, W\} = \exp\left\{\alpha - \frac{1}{2}(x(t) - \mu)'W^{-1}(x(t) - \mu)\right\}$$

where α is a location intensity parameter and W is a scale matrix. Or we could also assume that the intensity is a mixture of Gaussian components (ϕ) of the form $\lambda(t|\theta_l) = \sum_{j=1}^l p_j \phi(t|\mu_j, \Sigma_j)$. Thus we have a model for the data $\mathbf{y} = (t_0, \dots, t_n)$ depending on an unknown p -dimensional parameter $\theta = (\gamma, \mu, \Sigma, \dots)$ (the dots refer to other possible (realistic) parameters coming from the intensity function). Following the same notation as the authors', the ABC posterior can be defined as

$$\pi_{\text{ABC}}(\theta | \mathbf{s}_{\text{obs}}) \propto \pi(\theta) p(\theta | \mathbf{s}_{\text{obs}}) = \pi(\theta) \int \pi(\mathbf{y} | \theta) K[\{S(\mathbf{y}) - \mathbf{s}_{\text{obs}}\} / h] d\mathbf{y}$$

where

$$\pi(t_0, \dots, t_n | \theta) = \prod_{k=1}^n f_t(t_k | t_{k-1}) = \prod_{k=1}^n \frac{\gamma \lambda(t_k)}{\Gamma(\gamma)} \left\{ \gamma \int_{t_{k-1}}^{t_k} \lambda(u) du \right\}^{\gamma-1} \exp \left\{ -\gamma \int_{t_{k-1}}^{t_k} \lambda(u) du \right\}.$$

Finding appropriate random summary statistics to obtain a calibrated and noisy version of the ABC approximation is now possible along the lines indicated by the authors.

We can also resort to estimating the likelihood on the basis of earlier samples by using a sequential Monte Carlo algorithm (Sisson *et al.*, 2007; Beaumont *et al.*, 2009) in which we produce samples $(\theta_1^{(l)}, \dots, \theta_N^{(l)})$ by using at iteration $l=1$ a regular ABC step and, at each iteration $l=2, \dots, L$, Markov transition kernels K_l of the form $K_l(\theta_k^{(l)} | \theta_k^{(l-1)}) = \tau_k^{-1} \varphi\{\tau_k^{-1}(\theta_k^{(l)} - \theta_k^{(l-1)})\}$ so that the l th iteration sample is produced from the proposal distribution $\hat{\pi}_l(\theta^{(l)}) \propto \sum_{j=1}^N \omega_j^{(l-1)} K_l(\theta^{(l)} | \theta_j^{(l-1)})$. It would be interesting to see a

comparison between a noisy ABC version and a sequential Monte Carlo algorithm with this particular set-up.

Jesper Møller (*Aalborg University*)

The authors' construction of summary statistics for approximate Bayesian computation (ABC) is indeed relevant and appreciated, but how do we construct 'good' kernels K depending on the parameter θ ?

Consider a Gibbs random field; Section 1.1 mentions Gibbs random fields, with a reference to Grelaud *et al.* (2009) for the use of ABC. Then a natural (and often low dimensional) sufficient statistic $S(\mathbf{y})$ is given, namely the potential. More generally, consider a density

$$\pi(\mathbf{y}|\theta) = h\{S(\mathbf{y})|\theta\} / Z_\theta$$

where $h\{S(\mathbf{y})|\theta\}$ is easy to compute but Z_θ is an intractable normalizing constant. Here the *auxiliary variable method* (AVM) from Møller *et al.* (2006) applies for simulation from the posterior distribution of θ . As shown below, there are various choices of auxiliary variable distributions in the AVM; for one choice, which does not depend on θ , the AVM is effectively ABC; but other choices where the auxiliary variable density depends on θ should be used instead.

Specifically, the AVM is given by Table 2 if

- (a) K is replaced by a density $f(\cdot|\theta, \mathbf{y}_{\text{obs}})$ which may depend on both the data \mathbf{y}_{obs} and θ , and which is easy to compute and
- (b) the acceptance probability $\min\{1, H\}$ has Hastings ratio

$$H = \frac{f(\mathbf{y}_{\text{sim}}|\theta, \mathbf{y}_{\text{obs}})}{f(\mathbf{y}_{i-1}|\theta, \mathbf{y}_{\text{obs}})} \frac{\pi(\theta) g(\theta_{i-1}|\theta)}{\pi(\theta_{i-1}) g(\theta|\theta_{i-1})} I$$

where

$$I = \frac{h(\mathbf{s}_{\text{obs}}|\theta) h(\mathbf{s}_{i-1}|\theta_{i-1})}{h(\mathbf{s}_{\text{obs}}|\theta) h(\mathbf{s}|\theta)}.$$

ABC corresponds to the special case of the AVM with I replace by 1 and with $f(\mathbf{y}_{\text{sim}}|\theta, \mathbf{y}_{\text{obs}}) = K\{(\mathbf{s} - \mathbf{s}_{\text{obs}})/h\}$ not depending on θ . Whereas ABC is 'likelihood free', the AVM is only 'normalizing free'.

Møller *et al.* (2006) discussed the choice of auxiliary variable density $f(\mathbf{y}|\theta, \mathbf{y}_{\text{obs}})$. Ideally $f(\mathbf{y}|\theta, \mathbf{y}_{\text{obs}}) = \pi(\mathbf{y}|\theta)$, but then H depends on $Z_\theta/Z_{\theta_{i-1}}$, so they argued that instead $f(\mathbf{y}|\theta, \mathbf{y}_{\text{obs}})$ should be a good approximation to $\pi(\mathbf{y}|\theta)$. For Gibbs random fields and Gibbs point processes, a partially ordered Markov model (POMM) can be used to approximate $\pi(\mathbf{y}|\theta)$. Berthelsen and Møller (2008) demonstrated this, where the POMM depends on both θ and \mathbf{y}_{obs} .

This indicates that for ABC it might be useful to let K depend on θ . Incidentally, in the cases of Gibbs random fields and Gibbs point processes,

- (a) the AVM together with a POMM approximation should be preferred instead of ABC and
- (b) when simulating from $\pi(\mathbf{y}|\theta)$, either a perfect ('exact') sampler (e.g. Propp and Wilson (1996) and Kendall and Møller (2000)) or, if perfect sampling is too slow (as it often happens to be), a sufficiently long run of a Markov chain Monte Carlo algorithm may be used. (In the latter case, the AVM is also 'approximate'.)

Finally, the AVM has been extended in various ways, including the exchange algorithm in Murray *et al.* (2006).

Pablo Nigra (*University of Castilla La Mancha Toledo*) and **Emilio Porcu** (*University of Castilla La Mancha, Ciudad Real*)

Approximate Bayesian computation (ABC) is a computational method for drawing inference from complex stochastic models, whose likelihoods cannot be calculated directly. To avoid this hindrance, simulations of artificial data in parameter space are carried out, to compare their collected summary statistics with the summary statistics of observed data later. In this paper the authors develop a strategy to collect summary statistics for ABC in a semi-automatic way. First, they demonstrate theoretically that the optimal summary statistic is given by the posterior means of the parameters. These posterior means cannot be obtained analytically. Consequently, the authors dedicate a stage of the simulation to estimate the posterior means for using them later within ABC.

- (a) On page 424, it is shown how to make inference from multiple sources of data. The assumption is that the sources generate independent data. Can noisy ABC be applicable to correlated data? Models for spatial data are usually correlated.
- (b) This paper is well written, with a clear exposition of the new semi-automatic ABC. The theoretical background is suitably presented in a small number of theorems. The performance of the new approach is compared with that of classical ABC in several models of interest, showing clearly the new method's strengths and weaknesses. For example, classic ABC is preferable to semi-automatic ABC only when a small number of summary statistics is used.

D. J. Nott (*National University of Singapore*) and **Y. Fan and S. A. Sisson** (*University of New South Wales, Sydney*)

This paper introduces several novel ideas, including a method of deriving summary statistics s for functions of model parameters in a way that minimizes the variability of the posterior mean of those functions. In this comment, we make a connection between this approach to summary statistic choice and Bayes linear analysis (Goldstein and Wooff, 2007). Bayes linear analysis can be viewed as optimal linear estimation of a parameter vector θ where an estimator of the form $a + Bs$ is constructed for a p -dimensional vector a and a $p \times d$ matrix B minimizing

$$E\{(\theta - a - Bs)^T(\theta - a - Bs)\}, \quad (17)$$

where s is a d -vector of data (i.e. summary statistics). The expectation is with respect to the joint prior distribution of s and θ . The optimal linear estimator is given by

$$E_s(\theta) = E(\theta) + \text{cov}(\theta, s) \text{var}(s)^{-1}\{s - E(s)\}.$$

The estimator $E_s(\theta)$ is referred to as the adjusted expectation of θ given s . A Monte Carlo approximation to expression (17) based on $(\theta^{(m)}, s^{(m)}) \sim p(s|\theta)p(\theta)$, $i = 1, \dots, M$, is a least squares criterion for a linear regression of the simulated parameters on the summary statistics.

Thus, for large M , the semi-automatic summary statistics of Fearnhead and Prangle can be viewed as Bayes linear estimates of the posterior means. When computations are performed on a restricted parameter space (as per Section 3, point (a)), this interpretation still holds under a truncated prior for θ . Finally, the Bayes linear interpretation also holds for more flexible regression models, by considering suitable basis expansions involving functions of s , assuming that transformations of θ to maintain homoscedasticity are available. The links between regression methods in approximate Bayesian computation and Bayes linear analysis are discussed further in Nott *et al.* (2011).

Our final comment relates to the identification of a single summary statistic per posterior parameter of interest. In Nott *et al.* (2011), we propose to improve the accuracy of the joint posterior sample from any ABC method by firstly independently estimating the marginal posteriors $p(\theta_i | s_{\text{obs}})$, $i = 1, \dots, p$. Estimating marginal posteriors is easier than estimating the joint posterior owing to the lower dimensionality. We then replace the margins of the joint posterior with the more precisely estimated marginal distributions, thereby providing a more precise estimate of the true posterior distribution. This marginal adjustment strategy will be very efficient if highly informative but low dimensional and identifiable summary quantities are available for each marginal parameter.

As such, we propose that our marginal adjustment strategy using the semi-automatic summary statistics of Fearnhead and Prangle, following a standard approximate Bayesian computation analysis using the same statistics, would potentially provide even more precise estimates of the true posterior distribution. This approach is less affected by the increase in dimensionality of θ , than for regular approximate Bayesian computation analyses.

Mohammed A. Sedki (*Université Montpellier 2*) and **Pierre Pudlo** (*Institut National de Recherche Agronomique and Université Montpellier 2*)

We congratulate the authors for their interesting and stimulating paper on approximate Bayesian computation (ABC). Our attention was drawn to the regression building the new statistics in Section 3. Fearnhead and Prangle point out similarities with the post-processing proposed by Beaumont *et al.* (2002). But they defend their algorithm on its ability to select an efficient subset of summary statistics. The main idea here is certainly to bypass the curse of dimensionality. For example, in population genetics, a large number of populations commonly induce more than 100 summary statistics with the DIYABC software of Cornuet *et al.* (2008).

Apart from Blum (2010), the widely used post-processing of Beaumont *et al.* (2002) has been little studied theoretically, although it significantly improves the accuracy of the ABC approximation. Actually,

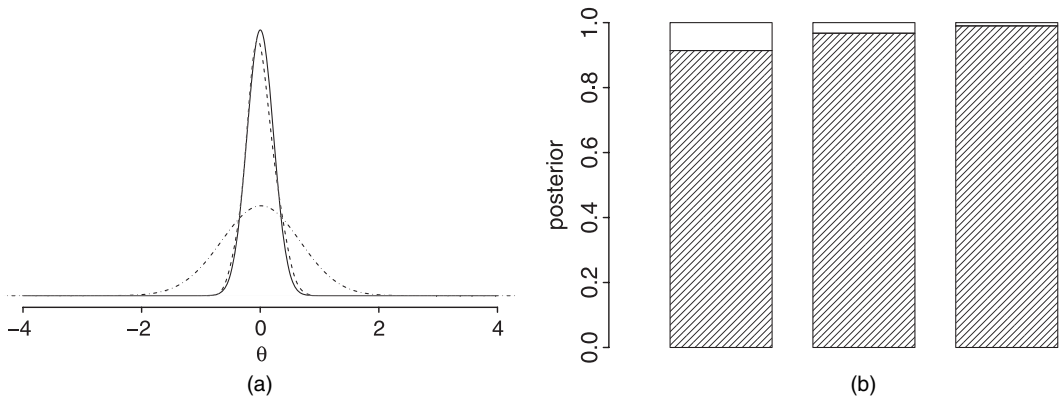


Fig. 8. (a) Posterior density estimates in the first example (the prior over θ is $\text{Unif}(-5, 5)$, and X is a Gaussian vector of dimension 20, with independent components, $X_i|\theta \sim \mathcal{N}(\theta, 1)$; the summary statistics are $S_1 = \text{mean}(X_{1:20})$, $S_2 = \text{median}(X_{1:20})$, $S_3 \sim \text{Unif}(-5, 5)$ and $S_4 \sim \mathcal{N}(0, 1)$; applying the BIC here improves the posterior density estimates by removing S_3 and S_4 ; —, true; ---, ABC without the BIC; ·····, ABC after the BIC) and (b) the model choice problem which is described in Robert *et al.* (2011) might be summed up in the following way: considering three populations, we must decide whether population 3 diverged from population 1 (\square , model 2) or 2 (\boxtimes , model 1); among 24 summary statistics, the BIC selects the two summary statistics LIK31 and LIK32 (see Table S1 of Robert *et al.* (2011)) which estimates genetic similarities between population 3 and the two other populations

Beaumont *et al.* (2002) replaced the θ s kept in the rejection algorithm with the residuals of a regression learning θ on the summary statistics. In the model choice settings (see, for example, Robert *et al.* (2011)), this post-processing uses a logistic regression predicting the model index; see Beaumont (2008). In both cases, it attempts to correct the discrepancy between the observed data set and the simulated data sets accepted by the ABC algorithm. We were intrigued by what would happen when postponing the variable selection criterion that is proposed in this paper until this post-processing.

Although a more detailed study is needed, we implemented two experiments:

- (a) one with a parameter estimation in the Gaussian family and
- (b) one with a model choice in the first population genetics example of Robert *et al.* (2011).

We ran the classical ABC algorithm and used a Bayesian information criterion (BIC) during the local linear regression to select the relevant statistics. Then, we scanned once again the whole reference table drawn from the prior to find the nearest particles to the observation, considering only the subset of statistics selected by the BIC. We ended with a local linear regression on this new set of particles. Numerical results are given in Fig. 8 and show that applying the BIC during the post-processing of Beaumont *et al.* (2002) is a promising idea.

S. A. Sisson and Y. Fan (*University of New South Wales, Sydney*)

This paper proposes a way of deriving summary statistics for functions of model parameters in a way that minimizes the variability of the posterior mean of those functions. Based on samples of (θ, s) in a truncated region of the prior (as per Section 3, point (a)), one fits a regression model, e.g. $\theta = \alpha + \beta F(s)$, where $F(s) = (f(s), \dots, f(s))$. The summary statistic proposed is the regression mean response, $\beta F(s)$, with precisely one statistic for each function of interest. Although using $\beta F(s)$ rather than s allows more precise estimation of the marginal posterior means of the functions of interest, it seems credible that posterior expectations of certain other quantities may be estimated less precisely under $\beta F(s)$ than s . This implies that *all* posterior expectations of interest in any analysis must be handled in this manner to guarantee the best possible precision.

However, in some approximate Bayesian computation (ABC) applications, such as extreme value theory (e.g. Bortot *et al.* (2007) and Erhardt and Smith (2012)), interest is typically in a large number (or even all) posterior quantiles (point estimates and credible intervals) above some high threshold. Our question is how does one mechanistically handle a very large (or even infinite) number of posterior functions of interest within the framework proposed?

In principle, we could use the proposed process directly, and regress all $p' \gg p$ posterior quantities of interest against $f(s)$, and using the resulting $\beta F(s)$ as the relevant summary statistics. However, as p' becomes large (or even as $p' \rightarrow \infty$, for example, where interest is in all posterior quantiles), this means that the accuracy of the resulting ABC posterior approximation will fall dramatically, compared with when using just s , given the increased dimension of the vector of summary statistics $\beta F(s)$. This comes in addition to the required increase in the number of (θ, s) samples that are required to perform the regression. Alternatively, we could repeatedly perform many separate implementations of the procedure proposed, each one aiming to estimate different (lower dimensional) aspects of the posterior as precisely as possible. Of course this approach raises questions of computational overheads, whether the separately estimated quantities would be consistent with each other, and which combinations of functions of interest to include in each analysis, e.g. all posterior parameter means and one function of interest, or some other combination.

Our final comment notes that the performance of regression-based ABC procedures, such as that of Beaumont *et al.* (2002), is sensitive to multicollinearity and large numbers of uninformative summary statistics, and as such may ‘overadjust’ the (θ, s) sample and thereby poorly estimate the posterior mean. As the dimension of $f(s)$ would increase rapidly with p' , this naturally raises the question about how the proposed semi-automatic framework would perform in the case of large p' with a potentially unreliable regression component.

Yingcun Xia (National University of Singapore)

I thank the authors for an interesting approach to tackling one of the most challenging problems in statistics. I am particularly interested in its application to the estimation of dynamical models. The work of Wood (2010) was referred to many times by Fearnhead and Prangle. I would thus like to consider the blowfly data that were studied in Wood (2010). The model used is a discrete stochastic version of that of Gurney *et al.* (1980): $dN_t/dt = PN_{t-\tau} \exp(-N_{t-\tau}|N_0) - \delta N_t$. The performances of three estimation methods including maximum likelihood estimation, the Markov chain Monte Carlo (MCMC) method of Wood (2010) and the catch-all method by Xia and Tong (2011) are shown in Fig. 9 in which broken curves

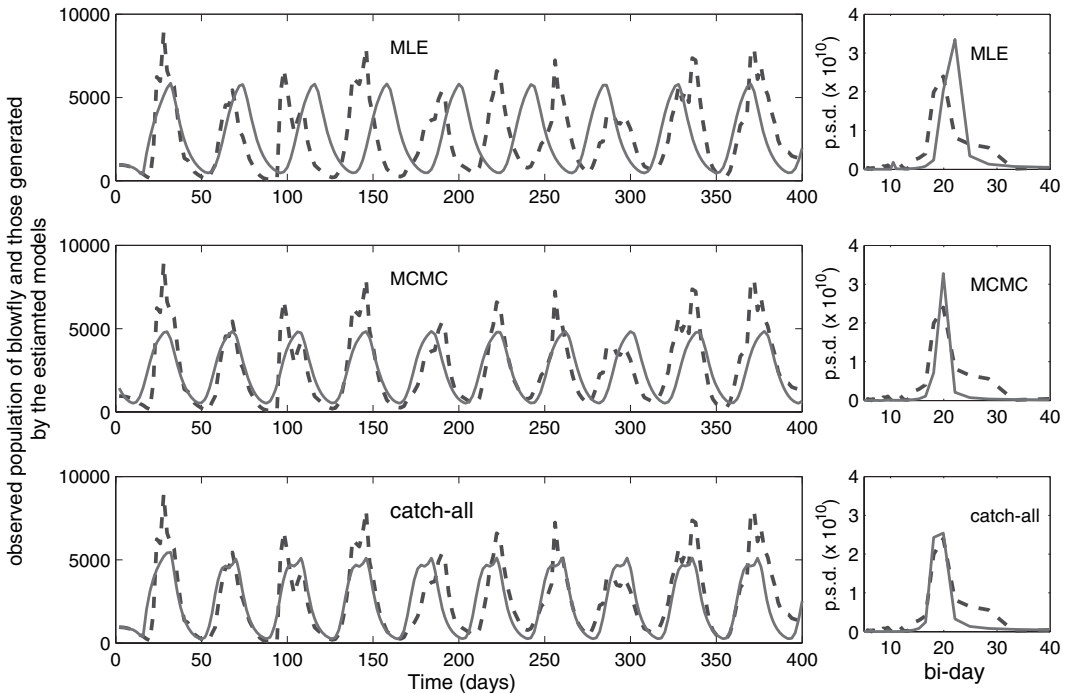


Fig. 9. Comparison of various estimation methods for the blowfly data: the broken curves in the left-hand panels are the observed time series, and those in the right-hand panels the power spectral densities; the full curves in the left-hand panels are the time series of the estimated skeletons, and those in the right-hand panels their power spectral densities

represent observed data and full curves represent the estimated skeleton. It is known that the dynamical cycles are the most important feature and must be captured by the estimated model. In this sense, both MCMC and catch-all methods outperform the conventional maximum likelihood estimation methods by comparing the dynamics in the left-hand panels and the power spectral density in the right-hand panels. The catch-all method can even capture the upside-down U-shape at the peak of the dynamics, which is also reflected in the power spectral density, but the MCMC algorithm fails. My question is can the method proposed refine the estimator of Wood (2010) and capture the shape in the dynamics?

The **authors** replied later, in writing, as follows.

We thank all the people who took the time to read and comment on our paper. We found the comments to be insightful and stimulating, and help to show what an exciting research area likelihood-free methods currently are. We have organized our response under some general themes.

Methods for fitting summary statistics

The comments include numerous questions, and ideas, about improving on the use of linear regression to find summary statistics. We think that one of the main contributions of the paper is to motivate a different view of how you choose summary statistics, namely that aiming for sufficient statistics is unrealistic in most applications, and instead we should aim for implementing approximate Bayesian computation (ABC) so that it can produce accurate inferences for parameters and features that we care about. Our main theoretical result (theorem 3) then motivates using summary statistics that are good estimators of the quantities (or parameters) of interest. We found that using linear regression is a simple, but effective, way of constructing such estimators from simulated data. However, we are not surprised if such a method can be improved on, e.g. in the challenging setting described by Sisson and Fan. The ideas of using principal component analysis, partial least squares or target projection (all mentioned by Beaumont), artificial neural networks (Chen) or non-linear regression methods (Blum and François) are all very sensible suggestions. In fact the comments of Chen and of Blum and François give empirical evidence of improvements that are possible over linear regression. As shown by Fasiolo, Pya and Wood the idea of wanting summaries that are parameter estimators, together with an understanding of the underlying model, can naturally motivate good summary statistics in some cases. This is an excellent example of how intuition can be used to help to guide choice (Kypriaos).

Another interesting idea is to estimate marginal posteriors for parameters of interest. Dawson proposes the possibility of using marginal sufficient statistics, but we remain sceptical about how easy it would be to find such sufficient statistics in challenging applications. It is possible to make use again of the ideas of theorem 3 by, as Nott, Fan and Sisson suggest, using an estimator of the parameter of interest or, as Dawson suggests, estimators of a number of functions of the parameter. Results in Nott *et al.* (2011) indicate that such an approach may work well in some situations and, if so, this would be a simple approach to applying ABC to models with a large number of nuisance parameters.

Finally, some comments (Drovandi and Pettitt, and Filippi, Barnes and Stumpf) suggested methods that search over sets of possible summary statistics (e.g. Joyce and Marjoram (2008) and Nunes and Balding (2010)). The problem that we see with these methods is that, unless you have a relatively small set of summary statistics to choose from, such search strategies are likely to be computationally prohibitive.

How many summary statistics?

A related issue is how many summary statistics should be used (Chen, Marin and Robert). We agree with the comment of Robert that our argument, based on lemma 1, that you want few summary statistics is rather vague. Perhaps a clearer argument comes from the effect that the dimension of the summary statistic has on the rate of convergence of the error in simple cases (Section 2.4; see also Blum (2010)). (We are hopeful that the ideas of Jasra can be used to produce more rigorous and general results on the error in ABC.)

Even if you accept that you want as few summary statistics as possible, do you really want to have just one summary statistic per parameter or feature of the model of interest (Chen, Marin and Robert) as suggested by theorem 3? The main issue here is that theorem 3 is based on having the true posterior expectations as summary statistics, whereas in practice we can only use approximations of these. As we mention in Section 5, the use of sliced inverse regression (Li, 1991) as a method to choose summary statistics may be useful here: it enables you potentially to find two or more summaries for one parameter such that the posterior mean of the parameter is well approximated by a non-linear function of these summaries.

Sisson and Fan raise a related question: what if there are a large number of features of a model that are of interest? They give an example where interest is in a range of quantiles of an extreme value model.

They also point out issues that will arise if the chosen features are highly correlated with one another. In practice, the choice of the number of features of interest will need to be limited by the maximum number of summary statistics you wish to use. As such we would recommend choosing a subset of features: aiming for those that are most important, and which avoids highly dependent features. A further important point is that if the features of interest are truly collinear then there will be a subset of features or parameters which if estimated accurately would give accurate estimates for all the features of interest. For example if you are interested in quantiles of a distribution from a scale–location family, then estimates of the location parameter and an appropriate scale parameter should be sufficient.

How automatic is semi-automatic approximate Bayesian computation?

Draper asks the question of just how automatic semi-automatic is, and whether it can be made truly automatic. We can give a little more insight into this from some recent application of semi-automatic ABC (Blum *et al.*, 2012). As part of a comparison of different ABC approaches we applied semi-automatic ABC to three different models: in each case we were only given a description of the model, together with a large set of samples from the model for different parameter values. For two of the three models, a first implementation of semi-automatic ABC gave good results. For the third, a more complex model with more parameters, a simple implementation of semi-automatic ABC was not as successful. However, problems with the summary statistics that were chosen were evident when fitting the linear model. A look at the output from the linear model not only made clear that some parameters were being poorly estimated, but also how the choice of explanatory variables and summary statistics could be changed to obtain more accurate results. As such a key feature of semi-automatic ABC is that problems can be picked up and corrected before running the final ABC algorithm itself. (It may also be that using some of the alternative non-linear regression models that were suggested in the discussion to construct summary statistics would give a more automated approach than the one that we have implemented.)

Marin and Robert, and Kypriaos ask the related question of how robust the method is to the accuracy of the pilot run. We present some results on the robustness of our method to the pilot run within the paper. The pilot run does have an effect on the accuracy of the ABC method; but the method can recover to some extent from a poor pilot run. It may be that if other approaches to finding summary statistics, e.g. based on local linear regression, were used then this dependence would be reduced. Also, the idea of Barnes, Filippi and Stumpf of an adaptive ABC algorithm, which tries to learn appropriate summary statistics within a Markov chain Monte Carlo or sequential Monte Carlo implementation of ABC, is one potential way of avoiding a pilot run. Such adaptive ABC methods are one promising direction for moving towards more automated procedures.

Links between approximate Bayesian computation and other likelihood-free methods

We thank Tong for pointing out the much longer history of likelihood-free methods than we had realized. Several discussants (Beaumont, Cornebise, Girolami and Kosmidis, Barthelmé, Chopin, Jasra and Singh, Mardia and Hamelryck, and Tong) ask about, and describe, links between ABC and other likelihood-free methods. Tong suggests that ABC and indirect inference are based on different philosophies. That may be true, but there are strong links between the two approaches. Both have an approach to choosing summaries of the data (in indirect inference this is by fitting an approximating model), and then use simulation from the true model to perform inference based on these summaries. A fruitful area of research will be to investigate these links, and the differences, between different methods. For example, we feel it is likely that some of the asymptotic results for indirect inference may apply to ABC methods as well (as we observed the two methods gave identical estimates for large sample sizes in the g -and- k -example). Also, as noted by Drovandi *et al.* (2011), you can use the idea of an approximating model to construct the summaries that you then use within ABC.

Sedki and Pudlo ask about combining semi-automatic ABC with the post-processing approach of Beaumont *et al.* (2002). In most of the applications that we have considered, post-processing semi-automatic ABC produced negligible improvements in accuracy; here it seems that ABC can extract much of the information content of our summary statistics, leaving little simple structure to be exploited by post-processing. However, an exception is the most complex example of Blum *et al.* (2012), where post-processing semi-automatic ABC substantially improved accuracy.

Making approximate Bayesian computation robust

Another issue, which was brought up by many discussants, was whether you can make ABC robust to model misspecification. This is currently an important, yet open, question. The combination of choosing summaries, and not requiring a perfect match of simulated summaries with the observed summaries,

suggests that it should be possible to make ABC a robust procedure, though how to do this unclear. It would be interesting if the ideas in Xia and Tong (2011), where you fit a model on the basis of how well it matches the ‘important’ aspects of the data, or the related methods described by Girolami and Cornebise can be used within ABC.

Wilkinson and Hartig mention the idea of relating the ‘noise’ (associated with the acceptance kernel; see equation (8)) that is used in ABC to the error in the model (Wilkinson, 2008). We think that this is an important idea if viewed in the correct way. If you have an appropriate description, or model, for the error in your simulator, then Wilkinson (2008) shows that there is an implementation of ABC which samples from the true posterior. However, just implementing ABC and then hoping that the noise kernel within ABC is a good model for error in your simulator is not sensible. As was shown in Section 4.3, the effect of not requiring an exact match between observed and simulated summaries can affect your inferences in ways that do not relate to the expected effect of model uncertainty (biasing estimates, rather than just increasing uncertainty).

As a first approach to dealing with inaccuracies in the model that is used you can check the fit of your model (Gelman). Ideas of how to do this within ABC are described in Ratmann *et al.* (2009) and Robert *et al.* (2010). Formal model choice is also possible by ABC methods (Toni *et al.*, 2009; Didelot *et al.*, 2011), and Marin and Robert ask about the application of our approach here. We are currently working on this subject and believe that a version of the semi-automatic method can be used which retains our theoretical results and also has good properties under the asymptotic regime of Marin *et al.* (2011).

Calibration and noisy approximate Bayesian computation

Wilkinson and Marin and Robert question how meaningful calibration is. Firstly, as pointed out in the paper and also by Marin and Robert, calibration on its own is not useful: the prior is calibrated, but it uses no information from the data. This is why we have argued for both calibration and accuracy. We agree with Wilkinson that ideally our inferences would be calibrated against real life, but we do not see how this can be achieved in most applications. The benefit of our definition of calibration is that it corresponds to calibration of the Bayesian posterior, in that the probabilities are correct under repeated sampling from the joint distribution of parameters and data (see also Cook *et al.* (2006), who used this idea to test software for Bayesian inference).

We argue that the main advantage of noisy ABC is when we combine data from multiple sources. In answer to Nigra and Porcu’s comment, this could be in applications with independent and identically distributed data, or for state space models (see Section 4.3, and comments by White, Kypraios and Preston, and Mateu and Arafat; also the method described by Barthelmé, Chopin, Jasra and Singh), and this may be a worthwhile approach to the application that Drovandi and Pettitt ask about. In these situations theorem 2 gives us some reassurance about the properties of ABC estimators. However, as pointed out by Robert, properties such as consistency also require other conditions of the model that is being analysed. In many situations where you combine data from multiple sources you can avoid having any ABC approximation at all (White, Kypraios and Preston; see also Wilkinson (2011)).

Other links

Several discussants drew out parallels between ABC and the pseudomarginal approach (Andrieu and Roberts, 2009), or the auxiliary variable method (Møller *et al.*, 2006). These related methods are versions of Markov chain Monte Carlo methods which can be used when you can obtain unbiased estimates of the likelihood. We were interested to see the novel idea for implementing ABC that was suggested by Lee, Andrieu and Doucet and the ideas of thinking about these and related algorithms in terms of their dependence on underlying random seeds (Andrieu, Doucet and Lee). The latter have the potential of providing smoother approximations, by using the same seeds for different parameter values (see also the coupled ABC method of Neal (2012)). Finally Møller suggests how links with these methods could be used to choose the kernel in ABC—something that is still an open problem. We also wonder whether the error distribution in the fitted linear model could be used to help to choose an appropriate kernel (see the discussion by Barnes, Filippi and Stumpf) and value of h (using results in Section 2.4 and Prangle (2011)).

Efron suggests the use of bootstrap methods, rather than ABC, for the g -and- k -example, and he draws interesting links between ABC and the bootstrap in terms of quantifying uncertainty. However the g -and- k -example is a particularly simple application, and one where numerical methods can be used to calculate the likelihood and hence the maximum likelihood estimates. Really in such an application there is no need for the complexity of ABC (as is shown by Efron). We are unclear whether the bootstrap could be so easily

implemented in more challenging applications where the likelihood really is not attainable, and it is such applications where the power of ABC, or other likelihood-free methods, is more apparent.

References in the discussion

- Akaike, H. (1978) On the likelihood of a time series model. *Statistician*, **27**, 217–235.
- Allingham, D., King, R. A. R. and Mengersen, K. L. (2009) Bayesian estimation of quantile distributions. *Statist. Comput.*, **19**, 189–201.
- Andrieu, C., Doucet, A. and Holenstein, R. (2010) Particle Markov chain Monte Carlo methods (with discussion). *J. R. Statist. Soc. B*, **72**, 269–342.
- Andrieu, C., Doucet, A. and Lee, A. (2012) Active particles and locally adaptive Markov chain Monte Carlo. To be published.
- Andrieu, C., Doucet, A. and Roberts, G. O. (2007) The expected auxiliary variable method for Monte Carlo simulation. *Technical Report*. Department of Mathematics, University of Bristol, Bristol.
- Andrieu, C., Doucet, A. and Tadić, V. B. (2005) On-line parameter estimation in general state-space models. In *Decision and Control*, pp. 332–337. New York: Institute of Electrical and Electronics Engineers.
- Andrieu, C. and Roberts, G. O. (2009) The pseudo-marginal approach for efficient Monte Carlo computations. *Ann. Statist.*, **37**, 697–725.
- Barnes, C., Filippi, S., Stumpf, M. and Thorne, T. (2011) Considerate approaches to achieving sufficiency for abc model selection. *Arxiv Preprint*.
- Barthelmé, S. and Chopin, N. (2011) Expectation-propagation for summary-less, likelihood-free inference. *Preprint arXiv:1107.5959*.
- Basu, D. (1977) On the elimination of nuisance parameters. *J. Am. Statist. Ass.*, **72**, 355–366.
- Bazin, E., Dawson, K. J. and Beaumont, M. A. (2010) Likelihood-free inference of population structure and local adaptation in a Bayesian hierarchical model. *Genetics*, **185**, 1411–1423.
- Beaumont, M. A. (2003) Estimation of population growth or decline in genetically monitored populations. *Genetics*, **164**, 1139.
- Beaumont, M. A. (2008) Joint determination of topology, divergence time and immigration in population trees. In *Simulation, Genetics and Human Prehistory* (eds S. Matsumura and P. Forsten), pp. 134–154. McDonald Institute.
- Beaumont, M. A., Cornuet, J.-M., Marin, J.-M. and Robert, C. P. (2009) Adaptive approximate Bayesian computation. *Biometrika*, **96**, 983–990.
- Beaumont, M. A., Zhang, W. Y. and Balding, D. J. (2002) Approximate Bayesian computation in population genetics. *Genetics*, **162**, 2025–2035.
- Becquet, C. and Przeworski, M. (2007) A new approach to estimate parameters of speciation models with application to apes. *Gen. Res.*, **17**, 1505–1519.
- Bernardo, J. and Smith, A. (1994) *Bayesian Theory*. New York: Wiley.
- Berthelsen, K. K. and Møller, J. (2008) Non-parametric Bayesian inference for inhomogeneous Markov point processes. *Aust. New Zeal. J. Statist.*, **50**, 627–649.
- Beskos, A., Papaspiliopoulos, O., Roberts, G. O. and Fearnhead, P. (2006) Exact and computationally efficient likelihood-based estimation for discretely observed diffusion processes (with discussion). *J. R. Statist. Soc. B*, **68**, 333–382.
- Blum, M. G. B. (2010) Approximate bayesian computation: a non-parametric perspective. *J. Am. Statist. Ass.*, **105**, 1178–1187.
- Blum, M. G. B. and François, O. (2010) Non-linear regression models for Approximate Bayesian Computation. *Statist. Comput.*, **20**, 63–73.
- Blum, M. G. B., Nunes, M., Prangle, D. and Sisson, S. A. (2012) A comparative review of dimension reduction methods in approximate Bayesian computation. To be published.
- Bortot, P., Coles, S. G. and Sisson, S. A. (2007) Inference for stereological extremes. *J. Am. Statist. Ass.*, **102**, 84–92.
- Candès, E. J. and Wakin, M. B. (2008) An introduction to compressive sampling. *IEEE Signal Process. Mag.*, **25**, 21–30.
- Chan, K. S. and Tong, H. (2001) *Chaos: a Statistical Perspective*. New York: Springer.
- Cook, S., Gelman, A. and Rubin, D. (2006) Validation of software for Bayesian models using posterior quantiles. *J. Computat. Graph. Statist.*, **15**, 675–692.
- Cornuet, J.-M., Santos, F., Beaumont, M. A., Robert, C. P., Marin, J.-M., Balding, D. J., Guillemaud, T. and Estoup, A. (2008) Inferring population history with DIYABC: a user-friendly approach to Approximate Bayesian Computation. *Bioinformatics*, **24**, 2713–2719.
- Csilléry, K., François, O. and Blum, M. G. B. (2012) abc: an R package for approximate Bayesian computation (ABC). *Meth. Ecol. Evol.*, to be published, doi 10.1111/j.2041-210X.2011.00179.x.
- Dean, T. A. and Singh, S. (2011) Asymptotic behaviour of approximate Bayesian estimators. *Preprint arXiv:1105.3655*.

- Dean, T. A., Singh, S. S., Jasra, A. and Peters, G. W. (2010) Parameter estimation for hidden Markov models with intractable likelihoods. *Preprint arXiv:1103.5399*. Cambridge University Engineering Department, Cambridge.
- Del Moral, P. (2004) *Feynman-Kac Formulae: Genealogical and Interacting Particle Systems with Applications*. New York: Springer.
- Del Moral, P., Doucet, A. and Jasra, A. (2006) Sequential Monte Carlo samplers. *J. R. Statist. Soc. B*, **68**, 411–436.
- DelMoral, P., Doucet, A. and Jasra, A. (2012) An adaptive sequential Monte Carlo method for approximate Bayesian computation. *Statist. Comput.*, to be published, doi 10.1007/s11222-011-9271-y.
- Del Moral, P., Doucet, A. and Singh, S. (2010) A backward interpretation of Feynman-Kac formulae. *Math. Modling Numer. Anal.*, **44**, 947–975.
- Didelot, X., Everitt, R., Johansen, A. and Lawson, D. (2011) Likelihood-free estimation of model evidence. *Baysn Anal.*, **6**, 49–76.
- Diggle, P. J. and Gratton, R. J. (1984) Monte Carlo methods of inference for implicit statistical models (with discussion). *J. R. Statist. Soc. B*, **46**, 193–227.
- Drovandi, C. C. and Pettitt, A. N. (2011a) Estimation of parameters for macroparasite population evolution using approximate Bayesian computation. *Biometrics*, **67**, 225–233.
- Drovandi, C. C. and Pettitt, A. N. (2011b) Likelihood-free Bayesian estimation of multivariate quantile distributions. *Computnl Statist. Data Anal.*, **55**, 2541–2556.
- Drovandi, C. C., Pettitt, A. N. and Faddy, M. J. (2011) Approximate Bayesian computation using indirect inference. *Appl. Statist.*, **60**, 317–337.
- Dunn, P. K. and Smyth, G. K. (2005) Series evaluation of Tweedie exponential dispersion model densities. *Statist. Comput.*, **15**, 267–280.
- Efron, B. (1987) Better bootstrap confidence intervals (with comments). *J. Am. Statist. Ass.*, **82**, 171–200.
- Erhardt, R. J. and Smith, R. L. (2012) Approximate Bayesian computing for spatial extremes. *Computnl Statist. Data Anal.*, **56**, 1468–1481.
- Garthwaite, P. H. (1994) An interpretation of partial least-squares. *J. Am. Statist. Ass.*, **89**, 122–127.
- Goldstein, M. and Wooff, D. (2007) *Bayes Linear Statistics: Theory and Methods*. Chichester: Wiley.
- Gourieroux, C., Monfort, A. and Renault, E. (1993) Indirect inference. *J. Appl. Econometr.*, **8**, S85–S118.
- Grelaud, A., Robert, C., Marin, J. M., Rodolphe, F. and Taly, J. F. (2009) ABC likelihood-free methods for model choice in Gibbs random fields. *Baysn Anal.*, **4**, 317–336.
- Gurney, W. S. C., Blythe, S. P. and Nisbet, R. M. (1980) Nicholson's blowflies revisited. *Nature*, **287**, 17–21.
- Hamelryck, T., Borg, M., Paluszewski, M., Paulsen, J., Frellsen, J., et al. (2010) Potentials of mean force for protein structure prediction vindicated, formalized and generalized. *PLOS ONE*, **5**, no. 11, article e13714.
- Hopcroft, L. E. M., McBride, M. W., Harris, K. J., Sampson, A. K., McClure, J. D., Graham, D., Young, G., Holyoake, T. L., Girolami, M. A. and Dominiczak, A. F. (2010) Predictive response-relevant clustering of expression data provides insights into disease processes. *Nucleic Acids Res.*, **38**, 6831.
- Itan, Y., Powell, A., Beaumont, M. A., Burger, J. and Thomas, M. G. (2009) The origins of lactase persistence in Europe. *PLOS Computnl Biol.*, **5**, no. 8, article e1000491.
- Jasra, A., Singh, S., Martin, J. and McCoy, E. (2012) Filtering via approximate Bayesian computation. *Statist. Comput.*, to be published.
- Joyce, P. and Marjoram, P. (2008) Approximately sufficient statistics and Bayesian computation. *Statist. Applic. Genet. Molec. Biol.*, **7**, article 26.
- Kendall, W. S. and Møller, J. (2000) Perfect simulation using dominating processes on ordered spaces, with application to locally stable point processes. *Adv. Appl. Probab.*, **32**, 844–865.
- Kolmogorov, A. N. (1942) Determination of the centre of dispersion and degree of accuracy for a limited number of observation. *Izv. Akad. Nauk USSR Ser. Mat.*, **6**, 3–32.
- Kvalheim, O. M. (2010) Interpretation of partial least squares regression models by means of target projection and selectivity ratio plots. *J. Chemetr.*, **24**, 496–504.
- Levesley, J., Yuan, X., Light, W. and Cheney, W. (1996) Convolution operators for radial basis approximation. *SIAM J. Math. Anal.*, **27**, 286–304.
- Li, K.-C. (1991) Sliced inverse regression for dimension reduction. *J. Am. Statist. Ass.*, **86**, 316–327.
- Light, W. (1993) Techniques for generating approximations via convolution kernels. *Numer. Alg.*, **5**, 247–261.
- Lowe, D. G. (2004) Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Visn.*, **60**, 91–110.
- Magee, L. (1998) Nonlocal behaviour in polynomial regressions. *Am. Statistn*, **52**, 20–22.
- Mardia, K. V., Frellsen, J., Borg, M., Ferkinghoff-Borg, J. and Hamelryck, T. (2011) A statistical view of the reference ratio method. In *LASR Proc.* (eds K. V. Mardia, A. Gusnanto, A. D. Riley and J. Voss), pp. 56–61. Leeds: University of Leeds.
- Marin, J., Pillai, N., Robert, C. and Rousseau, J. (2011) Relevant statistics for Bayesian model choice. *Preprint arXiv:1111.4700*.
- Marin, J.-M., Pudlo, P., Robert, C. P. and Ryder, R. J. (2012) Approximate Bayesian computational methods. *Statist. Comput.*, to be published.
- Martin, J., Jasra, A., Singh, S., Whiteley, N. and McCoy, E. (2012) Approximate Bayesian computation for smoothing, to be published.

- McKinley, J., Cook, A. and Deardon, R. (2009) Inference for epidemic models without likelihoods. *Int. J. Biostatist.*, **5**.
- Møller, J., Pettitt, A. N., Berthelsen, K. K. and Reeves, R. W. (2006) An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants. *Biometrika*, **93**, 451–458.
- Murray, I., Ghahramani, Z. and MacKay, D. J. C. (2006) MCMC for doubly-intractable distributions. In *Proc. 22nd A. Conf. Uncertainty in Artificial Intelligence* (eds R. Dechter and T. S. Richardson), pp. 359–366. Cambridge: Association for Uncertainty in Artificial Intelligence Press.
- Neal, P. (2012) Efficient likelihood-free bayesian computation for household epidemics. *Statist. Comput.*, to be published.
- Nott, D. J., Fan, Y., Marshall, L. and Sisson, S. A. (2011) Approximate Bayesian computation and Bayes linear analysis: towards high-dimensional ABC. *Preprint*. (Available from <http://arxiv.org/abs/1112.4755>.)
- Nunes, M. A. and Balding, D. J. (2010) On optimal selection of summary statistics for approximate Bayesian computation. *Statist. Applic. Genet. Molec. Biol.*, **9**, no. 1, article 34.
- Pflug, G. C. (1996) *Optimization of Stochastic Models: the Interface between Simulation and Optimization*. Boston: Kluwer Academic.
- Prangle, D. (2011) Summary statistics and sequential methods for approximate Bayesian computation. *PhD Thesis*. Lancaster University, Lancaster.
- Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A. and Feldman, M. W. (1999) Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molec. Biol. Evoln.*, **16**, 1791–1798.
- Propp, J. G. and Wilson, D. B. (1996) Exact sampling with coupled Markov chains and applications to statistical mechanics. *Rand. Struct. Alg.*, **9**, 223–252.
- Raiffa, H. and Schlaifer, R. (1961) *Applied Statistical Decision Theory*. Cambridge: Harvard University Press.
- Raiffa, H. and Schlaifer, R. (2000) *Applied Statistical Decision Theory*, new edn. New York: Wiley.
- Ratmann, O., Andrieu, C., Wiuf, C. and Richardson, S. (2009) Model criticism based on likelihood-free inference, with an application to protein network evolution. *Proc. Natn. Acad. Sci. USA*, **106**, 10576–10581.
- Robert, C. P., Cornuet, J.-M., Marin, J.-M. and Pillai, N. (2011) Lack of confidence in approximate bayesian computation model choice. *Proc. Natn. Acad. Sci. USA*, **108**, 15112–15117.
- Robert, C. P., Mengersen, K. and Chen, C. (2010) Model choice versus model criticism. *Proc. Natn. Acad. Sci. USA*, **107**, no. 3, article E5.
- Ross, G. J. S. (1972) Stochastic model-fitting by evolutionary operation. In *Mathematical Models in Ecology* (ed. J. N. R. Jeffers), pp. 297–308. Oxford: Blackwell Scientific.
- Seidenfeld, T. (1985) Calibration, coherence, and scoring rules. *Philos. Sci.*, **52**, 274–294.
- Sisson, S. A., Fan, Y. and Tanaka, M. M. (2007) Sequential Monte Carlo without likelihoods. *Proc. Natn. Acad. Sci. USA*, **104**, 1760–1765; correction, **106** (2009), article 16889.
- Student (1908a) Probable error of a mean. *Biometrika*, **6**, 1–25.
- Student (1908b) Probable error of a correlation coefficient. *Biometrika*, **6**, 302–310.
- Tavaré, S., Balding, D. J., Griffiths, R. C. and Donnelly, P. (1997) Inferring coalescence times from DNA sequence data. *Genetics*, **145**, 505–518.
- 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. *J. R. Soc. Interface*, **6**, no. 31, 187–202.
- Wegmann, D., Leuenberger, C. and Excoffier, L. (2009) Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood. *Genetics*, **182**, 1207–1218.
- White, S. R., Preston, S. P. and Kypraios, T. (2010) Fast approximate Bayesian computation for discretely observed Markov models using a factorised posterior distribution. To be published.
- Wilkinson, D. J. (2011) Parameter interface for stochastic kinetic models of bacterial gene regulation: a Bayesian approach to systems biology. In *Bayesian Statistics 9* (eds J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West), pp. 679–690. Oxford: Oxford University Press.
- Wilkinson, R. D. (2008) Approximate Bayesian computation (ABC) gives exact results under the assumption of model error. *Preprint arXiv:0811.3355v1*. University of Nottingham, Nottingham.
- Wood, S. (2010) Statistical inference for noisy nonlinear ecological dynamic systems. *Nature*, **466**, 1102–1104.
- Xia, Y. and Tong, H. (2011) Feature matching in time series modeling (with discussion). *Statist. Sci.*, **26**, 21–81.
- Xue, J. H. and Titterton, D. M. (2010) Joint discriminative-generative modelling based on statistical tests for classification. *Pattn Recogn Lett.*, **31**, 1048–1055.
- Yamada, S. and Morimoto, H. (1992) Sufficiency. In *Current Issues in Statistical Inference: Essays in Honor of D. Basu* (eds M. Gosh and P. K. Pathak), pp. 86–98. Hayward: Institute of Mathematical Statistics.