Annual Review of Statistics and Its Application

Approximate Bayesian Computation

Mark A. Beaumont

School of Biological Sciences, University of Bristol, Bristol BS8 1TQ, United Kingdom; email: m.beaumont@bristol.ac.uk

Annu. Rev. Stat. Appl. 2019. 6:2.1–2.25

The *Annual Review of Statistics and Its Application* is online at statistics.annualreviews.org

https://doi.org/10.1146/annurev-statistics-030718-105212

Keywords

Monte Carlo, intractable likelihood, Bayesian

Abstract

Many of the statistical models that could provide an accurate, interesting, and testable explanation for the structure of a data set turn out to have intractable likelihood functions. The method of approximate Bayesian computation (ABC) has become a popular approach for tackling such models. This review gives an overview of the method and the main issues and challenges that are the subject of current research.

1. INTRODUCTION

In recent years the advent of machine learning has placed into a deeper focus the aims of statistical inference, particularly the relative roles of prediction and explanation. An explanation for the data typically involves uncovering the structure and parameterization of a mechanistic model. Diggle & Gratton (1984) distinguished two forms of statistical model: those that are prescribed in terms of known distributions, with known likelihood functions, and those that are implicit, from which we can simulate samples but do not have access to an explicit expression for the likelihood. These latter models are often described as generative models. Simulations from implementations of generative models have increasingly been used to give training data sets for supervised machine learning purposes, potentially bridging the two cultures of Breiman (2001). In turn, interest in uncovering the generative model from a machine learning perspective has overlapped with the tradition of likelihood-free inference methods (Diggle & Gratton 1984), of which approximate Bayesian computation (ABC) forms a part. There have been many thorough reviews of ABC in both the statistical and more applied literature over the past 10 years, and this article aims to briefly introduce the method, review areas of recent activity, and make connections with the machine learning literature where appropriate.

The basic outline of what subsequently became known as ABC was introduced by Pritchard et al. (1999) for solving an application in population genetics. The method addresses the problem of finding the posterior distribution of parameters in a model that explains a potentially rich and complex data set. Such data sets typically consist of n observations $y_{\text{obs}} = (y_{\text{obs},1}, \ldots y_{\text{obs},n})$ where each $y_{\text{obs},i}$ may be of high dimension. The standard ABC approach is to use a mapping $s(y)$ to a lower dimensional and simpler set of summary statistics s. The model implies the existence of a density $f_n(s|\theta)$, but we have no straightforward access to it. The target of inference in ABC is

$$p_\epsilon(\theta, s|s_{\text{obs}}) \propto \pi(\theta) f_n(s|\theta) K_\epsilon(\|s - s_{\text{obs}}\|),$$

where $s_{\text{obs}} = s(y_{\text{obs}})$, $\pi(\theta)$ is the prior, $K_\epsilon(x)$ is a kernel function with scaling parameter (bandwidth) ϵ, and $\|\cdot\|$ is a distance metric, which is usually Euclidean. The ABC posterior for θ is the marginal

$$p_\epsilon(\theta|s_{\text{obs}}) = \int p_\epsilon(\theta, s|s_{\text{obs}}) \mathrm{d}s.$$

The motivation behind ABC is the notion that it is straightforward to devise Monte Carlo algorithms to sample from $p_\epsilon(\theta|s_{\text{obs}})$ without needing an explicit expression for the likelihood function $f_n(s|\theta)$. A typical simple algorithm is the following:

Algorithm 1.

1. Sample $\theta_i \sim \pi(\theta)$.
2. Simulate s_i from the generative model having implicit density $f_n(s|\theta_i)$.
3. Reject with probability proportional to $K_\epsilon(\|s_i - s_{\text{obs}}\|)$.
4. Repeat steps 1–3 until a sufficiently large sample of size M is obtained.

The resulting accepted θ_i are drawn from the ABC posterior $p_\epsilon(\theta|s_{\text{obs}})$, which converges to $p(\theta|s_{\text{obs}})$ as $\epsilon \to 0$. Most implementations of ABC use a uniform kernel, corresponding to the use of an indicator function at the rejection step:

$$\mathbb{I}\{\|s - s_{\text{obs}}\| \leq \epsilon\}.$$

The algorithm is illustrated schematically in **Figure 1**.

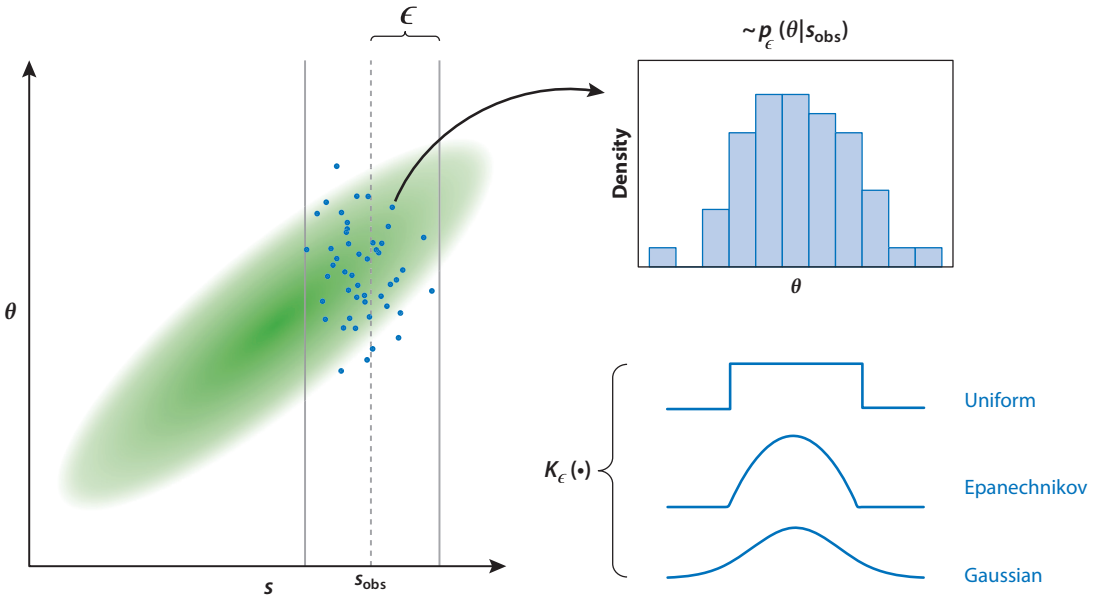

Figure 1

This figure illustrates the joint distribution $p(\theta, s)$ for univariate θ and s. Algorithm 1 samples points from $p_\epsilon(\theta, s|s_{obs})$, leading to summaries of samples from the ABC posterior $p_\epsilon(\theta|s_{obs})$ such as the histogram at the upper right. At the bottom right are shown the commonest kernels $K(\cdot)$. Note that in the case of the Gaussian ϵ corresponds to the standard deviation, leading to a more tapered rejection region than shown here.

Note that Algorithm 1 describes an online procedure, with a fixed ϵ and M, necessarily leading to an uncertain number of simulations. In fact, many ABC algorithms choose an initial number N of simulations and retain all sampled points, choosing the value of ϵ as the empirical quantile corresponding to $\Pr(\|s_i - s_{obs}\| < \alpha)$ for some proportion α. If we assume that the parameter vector and summary statistic vector have p and r elements, respectively, then the set of sampled points forms a reference table with N rows and $p + r$ columns (**Figure 2**). Before rejection, steps 1 and 2

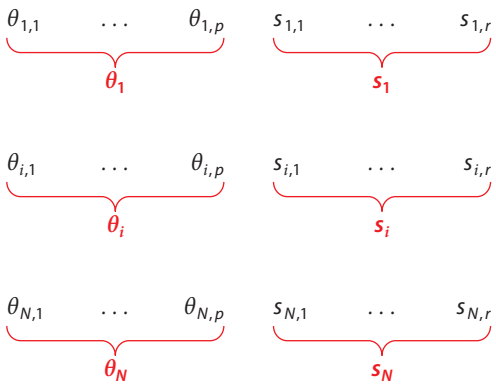

Figure 2

Structure of the reference table that is commonly mentioned in ABC analyses.

of Algorithm 1 jointly sample from $p(\theta, s)$, and step 2, considered alone, samples from the marginal $p(s)$, which is the prior predictive distribution of s.

In the outline above, θ is generally a vector, and it is straightforward to extend the approach to include an indicator m_j, $j = 1, \ldots, K$, for K different models, in which case the length of θ may vary among models, and the table illustrated in **Figure 2** will be a ragged array. Monte Carlo methods, for example, the same rejection algorithm above, can be used to report marginal probabilities for the m_j (described in more detail in Section 5).

The types of model to which ABC may be most usefully applied typically involve high-dimensional latent variables over which we wish to marginalize. The advantage of ABC is that traversing through the space of latent variables, often a reason for slow convergence of many Markov chain Monte Carlo (MCMC) algorithms, is not a constraint. Typically, latent variables are not kept in the reference table and are discarded during simulation. The range of applications of ABC tends to reflect this feature, and it is now widely used in a number of different fields. Examples include population genetics (Sjödin et al. 2012), ecology (Jabot & Lohier 2016), epidemiology (McKinley et al. 2018), systems biology (Liepe et al. 2014), anthropology (Kandler & Powell 2018), psychology (Turner et al. 2013), environmental modeling (Cui et al. 2018), climate modeling (Holden et al. 2018), and astronomy (Hahn et al. 2017). Many of the methods discussed in the present article are implemented in software, particularly as R packages (summarized in Kousathanas et al. 2018).

Research on the ABC method falls naturally into a number of themes, which are discussed further below. From a theoretical perspective, major topics have been the sensitivity to choice of summary statistics and also the convergence properties as $\epsilon \to 0$. The accuracy of ABC model choice and asymptotic behavior of ABC as $n \to \infty$ have also been investigated. Research has also focused on developing computational approaches to improve efficiency—that is, to make inferences based on smaller values of ϵ than is feasible with pure rejection (Algorithm 1). Many of the Monte Carlo computational approaches developed for Bayesian inference translate straightforwardly to ABC. However, some more ABC-specific algorithms have also been developed, which are based on modeling the joint distribution $p_\epsilon(\theta, s|s_{\text{obs}})$. This review first covers these postsampling adjustment methods because they are widely applicable and also relevant to the choice of summary statistics and the examination of the convergence and asymptotic behavior of ABC.

2. REGRESSION-ADJUSTMENT TECHNIQUES

An early technique in ABC has been regression-adjustment, which has been shown to often give improved convergence, for a given ϵ, to the ideal target $p(\theta|s_{\text{obs}})$ (Li & Fearnhead 2018a, Blum 2010). The method can typically be applied for any Monte Carlo method that gives samples from $p_\epsilon(\theta, s|s_{\text{obs}})$. The basis of the approach is that, given a sample $i = 1, \ldots, M$ of

$$\{\theta_i, s_i\} \sim p_\epsilon(\theta, s|s_{\text{obs}}),$$

we can use regression to obtain an estimate of $E(\theta|s)$ and then adjust each sampled θ_i as

$$\theta_i^* = \theta_i - \hat{E}(\theta|s) + \hat{E}(\theta|s_{\text{obs}}).$$

Beaumont et al. (2002) originally suggested using weighted linear regression with weights from an Epanechnikov kernel (alternatively, unweighted, if using Algorithm 1 above, where the kernel weights are included in the rejection algorithm). Blum & François (2010) introduced nonlinear regression using a one-layer neural network and modified the regression-adjustment algorithm to include a correction for heteroscedasticity. This latter is obtained by an additional regression

step on the residuals to obtain an estimate of the standard deviation of residuals as a function of θ, $\hat{\sigma}(\theta|s)$. The modified adjustment step is then

$$\theta_i^* = \frac{\hat{\sigma}(\theta|s_{\text{obs}})}{\hat{\sigma}(\theta|s)}(\theta_i - \hat{E}(\theta|s)) + \hat{E}(\theta|s_{\text{obs}}).$$

Blum & François (2010) show that both changes often lead to improvement over the original method of Beaumont et al. (2002), as measured by squared error. The adjustment is applied to each component of the parameter vector individually, which can be justified in terms of the asymptotic behavior of ABC (Frazier et al. 2018).

When the observations are not well explained by the model, the results obtained from regression-adjustment are potentially more misleading than the use of standard rejection (van der Vaart et al. 2015, Frazier et al. 2017). In this case s_{obs} may be an outlier in the distribution $p_\epsilon(\theta, s|s_{\text{obs}})$, and regression-adjustment extrapolates rather than interpolates. Model-checking methods (see Section 6) are useful for identifying such problems and potentially suggesting solutions. A further issue is that the regression-adjustment may yield values of θ_i^* that are outside the support of the prior or implicit likelihood function (for example, giving negative parameter values in models that do not allow this, or parameter values outside the range of a prior). This latter problem is partially addressed by transforming the simulated values of θ prior to regression-adjustment (Csilléry et al. 2012), although back-transformation will then lead to biased estimates. The potential for regression-adjustment to give problematic θ_i^* has prompted the use of methods that target the implicit likelihood $f_n(s|\theta)$, using multivariate regression methods (Leuenberger & Wegmann 2010, Fan et al. 2013). The method of Leuenberger & Wegmann (2010) is a standard component of the ABCtoolbox package (Wegmann et al. 2010) that is widely used in population genetics.

Another regression-based method is that of Nakagome et al. (2013), who use kernel ridge-regression to introduce nonlinearity (note that the term "kernel" here does not refer to a density kernel as in Algorithm 1). The basis of the approach is to note that, for a summary statistic vector of length r and N simulated samples (see **Figure 2**), regularized regression can be performed as a function of $(s^T s + \lambda I_r)^{-1}$ or $(ss^T + \lambda I_N)^{-1}$. The former involves inverting a $r \times r$ matrix, whereas the latter involves inverting a $N \times N$ matrix. The entries in the Gram matrix $G = ss^T$ are inner products, $\langle s_i, s_k \rangle = \sum_{j=1}^{r} s_{i,j} s_{k,j}$, for two vectors of summary statistics, s_i and s_k. The kernel trick relies on the fact that a suitably smooth function (a kernel) applied to each term of the Gram matrix corresponds to the inner product of a highly nonlinear transformation $\phi(\cdot)$ of the original coordinates:

$$\kappa(\langle s_i, s_k \rangle) = \langle \phi(s_i), \phi(s_k) \rangle.$$

The form of $\phi(\cdot)$ is unknown in general, but not required. Nakagome et al. (2013) use the radial basis kernel function

$$\kappa(\langle s_i, s_k \rangle) = \exp(-\|s_i - s_k\|^2/\sigma^2),$$

with a chosen bandwidth σ. (Although this form of $\kappa(\cdot)$ does not appear to involve an inner product, it can be shown to be a function of an infinite sum of terms $\langle s_i, s_k \rangle^n / n!$.) Ridge regression is then used to compute an estimate of $\mathrm{E}(\theta|s_{\text{obs}})$. An advantage of the method is that it can work efficiently for very large numbers of summary statistics. A potential constraint is that it is restricted to a relatively small reference table because of the difficulties of inverting an $N \times N$ matrix when the number of simulated samples, N, is large. Also, there is no standard method for choosing the kernel function, although the radial basis function above is popular. The kernel parameters (i.e., σ above) and regularization parameter are typically chosen using cross-validation. However, for the

examples studied, Nakagome et al. (2013) show improved performance over standard local linear regression and also the semiautomatic ABC method (Fearnhead & Prangle 2012), described in Section 3.2 below.

3. SUMMARY STATISTICS, DIMENSION REDUCTION, AND TOLERANCE INTERVAL

The performance of ABC depends on the choice of summary statistics (Prangle 2018). The motivation for mapping raw data to summary statistics is primarily to make comparison between the observations and the simulated data more efficient. Typically, raw data sets contain many exchangeable elements: The individual $y_{\text{obs},i}$ are exchangeable, and often so are the elements of which each is composed. Simple distance metrics on the raw data, such as the Euclidean metric, take no account of the exchangeability and are inefficient (Sousa et al. 2009). The challenge is to develop methods that overcome this problem (Chan et al. 2018). An advantage of most summary statistics is that the functions to compute them are typically invariant to permutations of exchangeable elements. But then ABC can appear subjective and arbitrary, inviting concern that the results of a study are dependent on the choice of summary statistics. Two main approaches have been taken to address this problem: One is to choose subsets of summary statistics that satisfy some optimality criterion (Joyce & Marjoram 2008, Nunes & Balding 2010); an alternative approach has been to find an optimal projection of a set of summary statistics onto a lower dimensional map (Wegmann et al. 2009, Fearnhead & Prangle 2012). Implicit in both approaches is the assumption that, provided a large enough initial set of summaries is chosen, the resulting subset or projection will not be sensitive to the initial composition of statistics.

3.1. Optimal Subsets of Summary Statistics

The study by Joyce & Marjoram (2008) introduced the concept of approximate sufficiency (AS) in ABC. The general idea is that, having initially identified a set of trial summary statistics, S, some approximately sufficient subset $s \subseteq S$ can be found. The log likelihood for a vector of summary statistics can be written as

$$\log f(s_1, \ldots, s_k|\theta) = \log f(s_1|\theta) + \log f(s_2|s_1, \theta) + \cdots + \log f(s_k|s_1, \ldots, s_{k-1}, \theta).$$

If $s_1, \ldots, s_{k-1}$ are sufficient, then $f(s_k|s_1, \ldots, s_{k-1}, \theta)$ is independent of θ. The basis of their method is to develop a score

$$\delta_k = \sup_\theta \left\{\log f(s_k|s_1, \ldots, s_{k-1}, \theta)\right\} - \inf_\theta \left\{\log f(s_k|s_1, \ldots, s_{k-1}, \theta)\right\},$$

and then test whether δ_k is less than some threshold. The summary statistics are deemed approximately sufficient if the difference in log-likelihood for any θ is less than or equal to the threshold.

Since the ratio of posterior densities is proportional to the ratio of likelihood functions,

$$e^{\delta_k} = \frac{\sup_\theta R_k(\theta)}{\inf_\theta R_k(\theta)},$$

where

$$R_k(\theta) = \frac{p(\theta|s_1, s_2, \ldots, s_k)}{p(\theta|s_1, s_2, \ldots, s_{k-1})}.$$

In the context of ABC, Joyce & Marjoram consider univariate θ and use Algorithm 1 to draw samples of θ from $p_\epsilon(\theta|s_{\text{obs}})$. They assume that the length of the summary statistic vector is initially $k_{\max}$, and the method first creates a reference table of draws of θ and $s_1, \ldots, s_{k_{\max}}$ from the joint distribution $p(s, \theta)$ (steps 1 and 2 of Algorithm 1). The values of θ are binned, and the numbers of sampled values in each bin are proportional to the posterior density. With the reference table, it is straightforward to compute the estimate $\hat{R}_k(\theta)$ for any k and sequence of summary statistics, and estimate

$$\hat{\delta}_k = \max_{j,l}[\log \hat{R}_k(\theta_j) - \log \hat{R}_k(\theta_l)].$$

It is necessary to choose a sequence of summary statistics to test and then stop adding statistics once $\hat{\delta}_k$ is less than some threshold. A drawback of the approach, as with stepwise methods in regression, is that the order of testing summary statistics will matter. They suggest selecting s_k randomly from among the remaining statistics, and then, if it is included, systematically testing to drop any of the other statistics in the current set. They apply the method to a number of test data sets to illustrate the performance of the approach. In an example with a known sufficient statistic, this was always chosen across a number of test data sets, but in other examples, different subsets were chosen for different test data sets. This latter observation is not surprising since the Pitman-Koopman-Darmois theorem shows that only for models in the exponential family is the number of summary statistics bounded (equal to the number of parameters in the model) irrespective of sample size. Thus, for most intractable models that ABC is applied to, there is unlikely to be a specific set of summary statistics sufficient for θ that is the same for all data.

An alternative method selecting a subset of summary statistics $s \subseteq S$ has been proposed by Nunes & Balding (2010) based on finding a subset that minimizes the entropy of the posterior distribution:

$$H = \mathbb{E}[-\log p(\theta|s_{\text{obs}})].$$

They consider the situation where they have samples of $\theta_i \sim p_\epsilon(\theta|s_{\text{obs}})$, where $s_{\text{obs}} \subseteq S$ is evaluated for the observations. The method uses the nearest-neighbor method of Singh et al. (2003) to estimate entropy for the kth nearest neighbor:

$$H_k = \frac{p}{M}\sum_{i=1}^{M} \log \|\theta_i - \theta_{i,k}\| + K, \qquad 1.$$

where H_k is an estimate of H, k is the kth neighbor (they use $k = 4$), p is the length of the parameter vector, M is the number of accepted samples from the rejection algorithm, $\|\theta_i - \theta_{i,k}\|$ is the Euclidean distance of the kth nearest neighbor from the focal θ_i sampled from the ABC rejection algorithm, and K is a constant that depends on M, k, and p. One of the algorithms they present is similar to that of Joyce & Marjoram (2008) but uses the minimum of Equation 1 as the criterion for choosing summary statistics, where the θ_i are computed for different subsets of summary statistics. The optimal set, S_{ME}, is obtained by searching through all subsets to find that which minimizes Equation 1. They demonstrate modestly improved performance over the Joyce & Marjoram method for similar example data sets and models. The minimum entropy approach favors a narrow posterior distribution but does not include a measure of the accuracy of the point estimate. Addressing this, they also present a two-stage procedure in which they first find the optimal set of summary statistics for the target data set, $S_{ME,\text{obs}}$, and then identify n_s simulated data sets that have the smallest Euclidean distance to the target, using the first-stage optimal set of summary statistics $S_{ME,\text{obs}}$. For the jth data set, with a known θ_j, rejection ABC is applied, which

generates samples θ_i from the posterior $p_\epsilon(\theta|s_{\text{obs}})$. The square root of the mean sum of squared errors, R, can be computed as

$$R(j) = \left(\frac{1}{n} \sum_{i=1}^{n} \|\theta_i - \theta_j\|^2 \right)^{1/2},$$

which is then averaged over the n_s simulated data sets

$$MR = \frac{1}{n_s} \sum_{j=1}^{n_s} R(j). \qquad 2.$$

Using this as the optimality criterion, they find a new subset S_2 by searching all subsets of the original set of summary statistics. This latter approach, while computationally much more demanding, shows substantially improved performance over the minimum-entropy method (and the method of Joyce & Marjoram). Thus, from a heuristic perspective, including criteria that minimize entropy and minimize integrated squared error seems desirable. However, the AS method of Joyce & Marjoram (2008) is theoretically well motivated, and the overall performance of AS may be enhanced by improving the implementation details.

3.2. Projection

As observed by Joyce & Marjoram (2008), and in accord with the Pitman-Koopman-Darmois theorem, Nunes & Balding (2010) find that the best choice of summary statistics varies across data sets. In view of the difficulties in implementing methods based on AS, an alternative approach is to project the summary statistics onto a lower-dimensional space in some optimal way.

3.2.1. Semiautomatic ABC. Assume that the ABC method yields a point estimate, $\hat{\theta}$, for parameter θ. Fearnhead & Prangle (2012) prove that the summary statistic that minimizes quadratic loss, defined by

$$L = (\theta - \hat{\theta})^{\mathrm{T}} A (\theta - \hat{\theta}),$$

for suitable matrix A, is when $s_{\text{obs}} = E(\theta|y_{\text{obs}})$. That is, the optimal summary statistic, in terms of minimizing quadratic loss, for inferring a parameter θ, is the true posterior mean, given the data. Although this may seem somewhat circular because we do not have access to the true posterior mean, it implies that a suitable summary statistic is an estimate of the posterior mean $\hat{E}(\theta|y)$, or $\hat{E}(\theta|s(y))$, as in regression-adjusted ABC (Section 2). Thus, the method involves projection of the high-dimensional data onto a single dimension for each component of the parameter vector. Fearnhead & Prangle (2012) propose that the summary statistic function $s(\cdot)$ is potentially an identity function of y, or a nonlinear transformation. However, as noted earlier, exchangeability of the components of y may make it preferable to choose summary statistics. In the examples given by Fearnhead & Prangle (2012), where the data are directly used, these are not exchangeable, but are order statistics or measurements from time series. An outline of the proposed method is as follows:

1. Choose a vector-valued function $s(\cdot)$.
2. Compute s_{obs} for the observed data.
3. Optionally, use a pilot simulation to determine a bounded region within which to sample θ proportional to the prior $\pi(\theta)$; alternatively, sample from the prior directly.

4. Simulate N parameter vectors and corresponding summary statistics, giving a reference table as in **Figure 2**. Typically the summary statistics are centered and scaled to have unit variance.
5. For each component of the parameter vector, $\theta_{\cdot,j}$, perform linear regression to obtain an estimate of $E(\theta_{\cdot,j}|y)$ as $[\beta^{(j)}]^T(1,s)$.
6. Perform ABC (e.g., as in Algorithm 1) using the projection $[\beta^{(j)}]^T(1,s)$ for components $j = 1,\ldots,p$ of the parameter vector.

Thus, the approach is to learn the optimal summary statistics by obtaining the linear predictor of $E(\theta_{\cdot,j}|y)$ using simulations from the joint distribution $p(\theta,s)$. The linear predictors for each component of the parameter vector form the columns of a matrix C, allowing the mapping of a simulated summary statistic vector s_i with r components to $s_i^T C$, with p components and observations $s_{\text{obs}}^T C$. It should be noted that since the intercept term is the same for s_i and s_{obs}, it can be discarded, so that C has dimensions $r \times p$. Fearnhead & Prangle (2012) show that the approach performs very well across a wide range of examples. The method is related to the widely used partial least squares (PLS) approach of Wegmann et al. (2009). Both methods give a projection of a high-dimensional set of summary statistics to a lower dimension using simulations from $p(\theta,s)$. However, in the case of PLS, the projection is orthogonal, and the elements of the projection matrix are not straightforwardly related to the least-squares linear predictors. Wegmann et al. (2009) suggest using cross-validation to choose the number of projected summary statistics, whereas in the case of semiautomatic ABC, one summary statistic is used for each parameter that is inferred. A recent example of the use of the PLS method in population genetics for improving the efficiency of ABC, explicitly in the context of AS, is described by Kousathanas et al. (2016).

The implication in Fearnhead & Prangle (2012) that it is optimal to use one statistic for each parameter is further strengthened by the study of Li & Fearnhead (2018b), which examines the asymptotic behavior of ABC algorithms as the sample size $n \to \infty$. They show that if $r > p$ then it always possible to find a projection of the original summary statistics down to p dimensions, which has asymptotic variance lower than or equal to the original r summary statistics. Although the theorem in Li & Fearnhead (2018b) does not imply a particular form for the projection, the results in Fearnhead & Prangle (2012) suggest a useful approach for finding it.

3.2.2. Nonlinear projection. Influenced by the results of Fearnhead & Prangle (2012), several studies have used neural networks and deep learning to model the relationship between parameter values and summary statistics. The method of Jiang et al. (2017) aims to find an optimal projection, using the posterior mean as a summary statistic following Fearnhead & Prangle (2012), by connecting the full set of raw data to the input layer of a deep neural network (DNN) with three hidden layers. They note the tendency of DNNs to overfit and suggest that this is best addressed by simulating training sets that are many times larger than the number of parameters. They examine regularization methods for the neural network but do not obtain significant improvement. In the example data sets they investigate, they show that a DNN leads to improvement over linear fits to the raw data. As in Fearnhead & Prangle (2012), they chose a test data set where exchangeability is not an issue. Creel (2017) also uses a deep learning network to find the posterior mean, which is then used as a summary statistic. In this case, the method is based on a set of predefined summary statistics, rather than using the full set of data.

The study of Chan et al. (2018) develops an all-encompassing approach to likelihood-free inference suitable for exchangeable data, where a neural network framework takes the raw data as input and returns a representation of the posterior as a function, which, in their example, is a softmax classifier whose output is interpreted as a posterior probability. Their test case example is

identifying recombination hot spots from DNA sequence data, where the network is trained to return the posterior probability that a section of sequence is a recombination hot spot. To tackle the problem of exchangeability among the elements of data, for example, among the rows of a multivariate data matrix, the network is designed so that the initial layers provide a mapping of the data, ignoring the exchangeability, to which a symmetric function is then applied. Thus, the output of the initial layers of the network can be regarded as a set of functions $\Phi_i(y_1), \ldots, \Phi_i(y_n)$ of the input. A symmetric function g can then be applied. In their example they suggest the element-wise maximum:

$$g := \max(\Phi_i(y_1), \ldots, \Phi_i(y_M)).$$

This then yields a set of summary statistics that are acted on by further layers of the neural network, and the weights for all the edges in the network are updated using a stochastic optimization technique. In their example application, the method appears to perform very strongly in comparison with a leading composite-likelihood approach.

Marin et al. (2016) have introduced a method based on the random forests algorithm to combine summary statistics (the random forests method can also be applied for model choice; Section 5). They develop an approach to obtain an estimate of the posterior expectation $E(\theta|y_{\text{obs}})$ from multiple summary statistics and then, rather than using regression-adjustment methods, estimate quantiles of the posterior directly. The method is illustrated with a complex population genetic example of human demographic history using a large-scale single nucleotide polymorphism data set.

Following Fearnhead & Prangle (2012), Mitrovic et al. (2016) have used kernel ABC (Section 2) to perform nonlinear regression in a two-step procedure. They first generate candidate summary statistics from a kernel ridge-regression of parameter values against data, and then they use the parameter estimates as summary statistics. They are able to show in example data sets that they achieve improved performance over semiautomatic ABC. Mitrovic et al. (2016) point out that using the two-step approach of Fearnhead & Prangle (2012) overcomes a potential limitation of the kernel method, which is restricted by the size of the Gram matrix that needs to be inverted, because the second step (ABC with the projected summary statistics) can be performed with an arbitrarily large number of samples.

3.3. A Comparison of Methods

Blum et al. (2013) provide a detailed empirical comparison of different methods for choosing summary statistics, based on data simulated under three different models, motivated by practical applications. Their study uses *MR*, the square root of the mean sum of squared errors averaged over test data sets, as used by Nunes & Balding (2010) (Equation 2), to compare methods relative to that of plain rejection ABC. They show that regression-adjustment often gives improved performance compared with plain rejection, and then they include this in the procedures they examine. They find that, for their examples, all methods they compared had generally improved performance over standard rejection. The results are variable across the models they examine and illustrate that it is difficult, on the basis of a relatively small set of empirical example-based analyses, to strongly favor one method over another. The computationally expensive two-stage method of Nunes & Balding (2010) and the projection method of Fearnhead & Prangle (2012) performed well. The semiautomatic ABC method outperformed the PLS projection method for many components of the parameter vector. The method of Joyce & Marjoram (2008) performed generally quite poorly. A reasonable conclusion based on these results is that the method of Fearnhead & Prangle (2012) is a safe and quickly implemented option, often yielding substantial improvement

gains over plain rejection, although there is clearly scope for further improvements in methods for choosing summary statistics.

3.4. Rejection Kernel and Bandwidth

Key components in the computation of the ABC posterior

$$p_\epsilon(\theta, s|s_{\text{obs}}) \propto \pi(\theta) f_n(s|\theta) K_\epsilon(\|s - s_{\text{obs}}\|)$$

are the kernel function $K_\epsilon(\|s - s_{\text{obs}}\|)$ and the choice of ϵ. From the perspective of computational ease, also supported by theoretical results on asymptotic efficiency (Li & Fearnhead 2018b), an appropriate choice of ϵ can be made indirectly via the proportion of simulated points that are accepted (Beaumont et al. 2002). The two most commonly used kernel functions are the uniform kernel and the Epanechnikov kernel, and the most commonly used distance metric is Euclidean. There has, however, been much research on different methods of scaling the summary statistics. For a vector s_i, corresponding to a row of s, the squared Euclidean distance from $s_{i,\text{obs}}$ can be written as

$$(s_i - s_{i,\text{obs}})^T A (s_i - s_{i,\text{obs}}),$$

with $A = \text{diag}(1, \ldots, 1)$, the identity matrix. In this case, standard rejection defines an ellipse:

$$\mathbb{I}\{(s_i - s_{i,\text{obs}})^T A (s_i - s_{i,\text{obs}}) < \epsilon\} \qquad 3.$$

(Fearnhead & Prangle 2012). Written in this way, it can be seen that there is some level of duality between the projection method that is used and the choice of scaling for the distance metric. One of the simplest and widely used scaling approaches is to divide the summary statistics by their estimated standard deviations in the sampled prior predictive distribution, giving the projection $A = \text{diag}(1/\hat{\sigma}_1^2, \ldots, 1/\hat{\sigma}_r^2)$ resulting in a rejection ellipsoid. A robust alternative is to use the median absolute deviation (Csilléry et al. 2012).

As noted by Prangle (2017) there are many choices available for A, giving generalized rejection ellipsoids, such as, for example, the estimated precision matrix from the prior predictive distribution, giving a Mahalanobis distance. Similarly, the projection method of Fearnhead & Prangle (2012), yielding a matrix of coefficients A, leads to an ellipsoid

$$\mathbb{I}\{(s_i - s_{i,\text{obs}})^T A A^T (s_i - s_{i,\text{obs}}) < \epsilon\},$$

which has the same form as Equation 3 above. Thus, it can be seen the choice of projection and the choice of scaling for the distance metric are closely bound together, with different choices of A leading to different shapes of the acceptance envelope around S_{obs}. Prangle (2017) shows that there are significant improvements to ABC inference when the elements, w_i, of the diagonal scaling matrix $A = \text{diag}(1/w_1^2, \ldots, 1/w_r^2)$ are learned in a sequential algorithm (Section 4.2). It would be of interest to see how the elements of the general matrix A can be learned through an iterative ABC algorithm, although as Prangle (2017) points out, controlling the stability of such an algorithm and proving convergence may be challenging.

4. COMPUTATIONAL TECHNIQUES

A variety of computational methods have been proposed to improve the efficiency of ABC inference (Sisson & Fan 2018). The rejection method outlined in Algorithm 1 assumes that the

parameter values are sampled from the prior. However, if the data are informative and so the posterior distribution has a more concentrated density than the prior, this basic algorithm is not efficient. One general approach (Fearnhead & Prangle 2012) is to use some chosen proposal distribution for importance sampling and reweight accordingly. Another solution is to use regression-adjustment methods (Section 2). However, most applications of ABC use one of two widely used methodologies that were introduced within 10 years of the original ABC algorithm of Pritchard et al. (1999), often in conjunction with regression-adjustment.

4.1. Markov Chain Monte Carlo

The first MCMC version of ABC was introduced by Marjoram et al. (2003). The ABC-MCMC algorithm and its variants have been widely used. The algorithm follows that of a typical Metropolis-Hastings algorithm, and the ABC counterpart of the likelihood ratio is the accept/reject step $\mathbb{I}\{\|s - s_{\text{obs}}\| \leq \epsilon\}$.

Algorithm 2.

1. Choose a value for ϵ, start with $t = 1$, and choose an initial value for $\theta^{(1)}$ (e.g., $\theta^{(1)} \sim \pi(\theta)$).
2. Propose a new value of θ from a Metropolis-Hastings kernel $\theta' \sim q(\cdot|\theta^{(t)})$.
3. Simulate $s \sim f_n(s|\theta')$.
4. With probability

$$\min\left(1, \frac{\pi(\theta')q(\theta^{(t)}|\theta')}{\pi(\theta^{(t)})q(\theta'|\theta^{(t)})}\mathbb{I}\{\|s - s_{\text{obs}}\| \leq \epsilon\}\right), \qquad 4.$$

   $\theta^{(t+1)} = \theta'$; $\theta^{(t+1)} = \theta^t$ otherwise.
5. Increment $t = t + 1$.
6. Repeat from step 2 until convergence.

From an efficiency perspective, the value of $\mathbb{I}\{\|s - s_{\text{obs}}\| \leq \epsilon\}$ is typically tested first in Equation 4 although, using the result of Peskun (1973), it is possible to split Equation 4 into two steps and move to the second step with probability

$$\min\left(1, \frac{\pi(\theta')q(\theta^{(t)}|\theta')}{\pi(\theta^{(t)})q(\theta'|\theta^{(t)})}\right),$$

before moving to step 5 of the algorithm with probability

$$\mathbb{I}\{\|s - s_{\text{obs}}\| \leq \epsilon\}.$$

Although this splitting of the Metropolis-Hastings step is less efficient generally (Peskun 1973), this variant may be more efficient for ABC if the generative model is expensive to simulate. An additional modification is to view

$$\mathbb{I}\{\|s - s_{\text{obs}}\| \leq \epsilon\}$$

as a Monte Carlo estimate based on a sample size of $N = 1$ from the likelihood

$$\int f_{(\epsilon, s_{\text{obs}})}(s|\theta,)\mathrm{d}s.$$

Thus, with larger sample sizes, the algorithm becomes an ABC version of the pseudomarginal algorithm (Beaumont 2003, Andrieu & Roberts 2009). With the pseudomarginal algorithm, there

is typically an optimal value of $N \gg 1$. By contrast, Bornn et al. (2017) demonstrated that for ABC, it is generally the case that a sample of size 1 in the MCMC setting is most efficient.

An attractive feature of ABC-MCMC is that it can straightforwardly be used with a flat improper prior. Because the acceptance rate is higher when in regions of parameter space with high likelihood, the ABC-MCMC algorithm has a lower acceptance rate when in the tails of the posterior distribution, which can lead to poor mixing (Baragatti et al. 2013). To accommodate this, a tempering approach can be used (Ratmann et al. 2007) in which, during the burn-in phase, a larger value of ϵ is chosen and progressively reduced to $\epsilon_{\min}$. Statistics from the posterior are then computed from parameter values simulated with $\epsilon_{\min}$. Another method is to treat ϵ as a parameter of the model (Bortot et al. 2007, Ratmann et al. 2009) and then compute posterior quantities conditional on $\mathbb{I}\{\|s - s_{\text{obs}}\| \leq \epsilon_{\min}\}$.

4.2. Sequential Monte Carlo

As with MCMC, the motivation for the application of sequential approaches to ABC is to improve efficiency of the proposal distribution and allow for a smaller bandwidth in the acceptance kernel. There are two forms of the ABC sequential Monte Carlo (ABC-SMC) algorithm that are widely used. One group of methods (Beaumont et al. 2009, Sisson et al. 2009, Toni et al. 2009), introduced by Sisson et al. (2007), can be regarded as an ABC version of population Monte Carlo (PMC) (Cappé et al. 2004) and is based on sequential importance sampling. An alternative approach, analogous to standard particle-filtering algorithms, was introduced by Del Moral et al. (2012).

The rationale of the PMC approach is to successively fit, at the tth iteration, an approximating kernel density $K_t(\cdot)$ to the samples from the posterior generated at each step. This approximating kernel is then used as the proposal distribution for the next step. An importance weight corrects for sampling from the proposal distribution rather than the prior (step 1 of Algorithm 1). The initial proposal density $q_1(\theta)$ is often taken to be the prior $\pi(\theta)$. However, some other initial density for $q_1(\theta)$ can be specified (noting also that the weight for the ith particle, w_i, can be multiplied by some constant). Subsequent ($t > 1$) proposal distributions have density

$$q_t(\theta) = \left(\sum_{i=1}^{N} w_i^{t-1} K_t(\theta | \theta_i^{t-1}) \right) / \sum_{i=1}^{N} w_i^{t-1}.$$

The ABC-PMC algorithm is then:

Algorithm 3*a*.

1. Start with $t = 1$.
2. Repeat steps 1–4 of Algorithm 1, sampling from $q_t(\theta)$ rather than $\pi(\theta)$, until M particles are obtained.
3. For $i = 1, \ldots, M$ set importance weight $w_i^t = \pi(\theta_i^t)/q_t(\theta_i^t)$.
4. Set $t = t + 1$; repeat until termination criterion is reached.

The proposal kernel is often taken to be

$$K_t(\theta | \theta') = N(\theta', 2\Sigma_{t-1}),$$

where Σ_{t-1} is the empirical covariance matrix computed from the weighted particles at iteration $t - 1$. The stopping criterion can be based on choosing a succession of ϵ_t, either in advance or adaptively while running the algorithm (Drovandi & Pettitt 2011).

The ABC-SMC algorithm of Del Moral et al. (2012) differs from ABC-PMC in that it uses a Metropolis-Hastings proposal kernel for regenerating particles. Additionally, the algorithm uses

resampling from the particle weights, as in a bootstrap particle filter. The particle weights are given by the kernel $K_\epsilon(\|s - s_{\text{obs}}\|)$, which is taken to be $\mathbb{I}\{\|s - s_{\text{obs}}\| \leq \epsilon\}$. An optional feature of the algorithm is that for each particle, θ_i, D data sets can be simulated, and therefore weights can be computed as

$$w_i^t = \frac{1}{D}\left\{\sum_j^D \mathbb{I}\{\|s_j - s_{\text{obs}}\| \leq \epsilon\}\right\}.$$

The algorithm then shares features with the MCMC variants discussed in Section 4.1. However, for ease of explication, it is assumed that $D = 1$ in the example algorithm below. The aim is to choose values of ϵ_t adaptively such that a proportion α of the particles that are accepted within the tolerance ϵ_{t-1} are also accepted within ϵ_t, i.e.,

$$\alpha \sum_{j=1}^N \mathbb{I}\{\|s - s_{\text{obs}}\| \leq \epsilon_{t-1}\} = \sum_{j=1}^N \mathbb{I}\{\|s - s_{\text{obs}}\| \leq \epsilon_t\}.$$

Initially $\epsilon_0 = \infty$. The ABC-SMC algorithm of Del Moral et al. (2012) can be given in simplified form as:

Algorithm 3*b*.

1. Initialize, with $t = 0$, $i = 1, \ldots, N$ sample $\{\theta_i^{(t)}, s_i^{(t)}\}$ with weight $w_i = 1/N$.
2. Set $t = t + 1$.
3. Compute ϵ_t using α, as described.
4. For all $i \in (1, \ldots, N)$ such that $\|s - s_{\text{obs}}\| \leq \epsilon_t$, set $w_i = 0$.
5. Renormalize weights $\sum_i w_i = 1$ and compute the effective sample size $\text{ESS} = \left(\sum_{i=1}^N w_i^2\right)^{-1}$.
6. If $\text{ESS} < N/2$, resample N particles according to weight w_i.
7. Use a Metropolis-Hastings kernel to perturb all particles with $w_i > 0$:

$$\theta_i' \sim q_t(\cdot|\theta_i^{t-1}),\ s_i\prime \sim f_n(s|\theta_i).$$

8. Apply step 4 of Algorithm 2.
9. Repeat from step 2 until the stopping rule $\epsilon_t < \epsilon_T$.

Control of the approach to the target value of the tolerance ϵ_T is via the parameter α. Exploration of suitable values for ϵ_T and α is important for the efficient use of the algorithm.

The package `EasyABC` (Jabot et al. 2013) implements both approaches to sequential ABC given in Algorithms 3*a* and 3*b*. A comparison of the approaches is given by Daly et al. (2017), who compare a number of different rejection kernels, motivated by a concern to accommodate model error (Wilkinson 2013). They find that the PMC version of the algorithm is more sensitive to the form of the rejection kernel that they use.

4.3. ABC and Big Data

Recently there has been interest in methods for combining Monte Carlo inferences from multiple data sets (Scott et al. 2016). There are many contexts in which the need to combine data sets may arise. For example, in population genomics, it may be infeasible to make inferences for whole genomes and more efficient to make inferences on sections of genome; after these components are computed by a cluster, the results can be combined. Thus, we assume that the data can be

broken into L components (sites), and the likelihood factorizes:

$$p(y|\theta) = \prod_{j=1}^{L} p(y_j|\theta),$$

so that

$$p(\theta|y) \propto \pi(\theta) \prod_{j=1}^{L} p(y_j|\theta)$$

or

$$p(\theta|y) \propto \prod_{j=1}^{L} p(y_j|\theta)\pi(\theta)^{1/L}$$

(Scott et al. 2016). Monte Carlo methods, if applied in this way, typically yield random draws from the subposterior (Scott 2017), i.e., the ith particle sampled from the jth site is sampled

$$\theta_{i,j} \sim p(\theta|y_j)\pi(\phi)^{1/L}.$$

The challenge is to design methods that will combine the information from these individual particles. The most common assumption is that in a big data setting, the Bernstein–von Mises theorem holds, and the target distribution can be approximated by a multivariate Gaussian. The consensus Monte Carlo method (Scott et al. 2016) proposes to fit multivariate Gaussians to the sampled $\theta_{i,j}$ from each subposterior and multiply the densities together.

The consensus Monte Carlo approach is strongly related to the EP-ABC method (Barthelmé & Chopin 2014, Barthelmé et al. 2018), which uses expectation propagation (EP) (Minka 2001). The aim of EP is to find a solution in terms of matching factors:

$$\tilde{p}(\theta|y) = \prod_{j=0}^{L} g_j(\theta).$$

The parameters of the $g_j(\theta)$ are initialized and then fitted in a series of sweeps through the data. For the jth site,

1. The cavity distribution is formed:

$$g_{-j}(\theta) = \left(\prod_{k=0}^{L} g_k(\theta)\right) / g_j(\theta).$$

2. The tilted distribution is

$$\propto g_{-j}(\theta)p(y_j|\theta).$$

3. A new $g'_j(\theta)$ is found that minimizes the Kullback-Leibler divergence between the tilted distribution and $g_{-j}(\theta)g'_j(\theta)$.

If the $g_j(\theta)$ are from the exponential family, minimization of Kullback-Leibler divergence is equivalent to choosing moments of $g_{-j}(\theta)g'_j(\theta)$ to be the same as those of the tilted distribution. In the standard EP algorithm, numerical methods such as quadrature are used to achieve this. The

algorithm is repeatedly applied to all n sites until convergence. An example parallel EP-ABC algorithm, from Barthelmé et al. (2018), is:

Algorithm 4.

1. Initialize natural parameters $\lambda_0, \ldots, \lambda_L$.
2. $\lambda = \sum_{j=0}^{L} \lambda_j,\ \tau = t(\lambda)$.
3. For $i = 1, \ldots, M$, sample particles $\theta_i \sim \mathcal{N}(\theta|\tau),\ s_i \sim f(s|\theta_i)$.
4. For sites $j = 1, \ldots, L$:
   (*a*) Weight particles $i = 1, \ldots, M$ by $\mathcal{N}(\theta_{j,i}|\tau_{-j})/\mathcal{N}(\theta_{j,i}|\tau)K_\epsilon(\|s_j - s_{j,\text{obs}}\|)$.
   (*b*) Use weighted particles to compute empirical parameters τ_j and transform to natural parameters $\lambda_j = t^{-1}(\tau_j)$.
   (*c*) Resimulate $\{\theta_i, s_i\}$ as in step 3 when the effective number of particles becomes too small.
5. Stop when the change in λ is small enough.

This algorithm can be unpacked as follows. Distributions in the exponential family can be parameterized by their mean parameter, which, for a Gaussian, can be taken as

$$\tau = \{\mu, \Sigma\},$$

with mean vector μ and covariance matrix Σ, or natural parameter

$$\lambda = \left\{\mu\Sigma^{-1}, \Sigma^{-1}\right\},$$

with functions $t(\cdot), t^{-1}(\cdot)$ mapping between them. For distributions of the same member of the exponential family, the natural parameter of the product of density functions is the sum of the natural parameters of each density. It is convenient to use the shorthand $\lambda_{-j} = \sum_{i\neq j}^{L} \lambda_i,\ \tau_{-j} = t(\lambda_{-j})$. In the algorithm, λ_0 is the natural parameter vector for the prior and remains fixed. Arbitrary initial points are chosen for remaining λ_j. Step 3 of the algorithm simulates particles from the current approximation of the posterior predictive distribution $p_{s_{\text{obs}}}(\theta, s)$. Step 4*a* applies an importance weight $\mathcal{N}(\theta_{j,i}|\lambda_{-j})/\mathcal{N}(\theta_{j,i}|\lambda)$ to correct for the fact that $\theta_{j,i}$ is not simulated from the cavity distribution. In the case of the EP-ABC algorithm and other Monte Carlo EP algorithms (Gelman et al. 2014), the moments estimated in step 4*b* have an appreciable variance in comparison with standard EP, which is based on an analytical solution or quadrature. The EP algorithm is a fixed-point recursion and potentially unstable in the face of Monte Carlo noise, and Hasenclever et al. (2017) show that a modification leads to a stochastic gradient descent algorithm that is generally more efficient in the face of noisy moment estimates. An advantage of using the EP approach with ABC is that it shares characteristics with ABC-SMC algorithms, in that it allows refinement of the proposal distribution for site j and converges on the EP approximation $\propto p(\theta|s)/p(\theta|s_j)$, potentially allowing for refinement of the bandwidth ϵ. The parallel version of the algorithm also has the advantage that the reference table (step 3 of Algorithm 4) need only be simulated infrequently. Using this approach, Barthelmé et al. (2018) were able to achieve up to a 100-fold increase in computation speed for some problems in comparison with a standard ABC algorithm.

5. MODEL CHOICE

From its earliest introduction, ABC has been widely used to compare models in a Bayesian framework (Pritchard et al. 1999). In this case, we may consider a series of models, labeled by index $1, \ldots, K$, with a sampled model indicator $m_i \in \{m_1, \ldots, m_K\}$.

Algorithm 5.

1. Sample $m_i \sim \pi(m)$.
2. Sample $\theta_i \sim \pi(\theta|m_i)$.
3. Simulate s_i from the generative model having implicit density $f_n(s|\theta_i, m_i)$.
4. Reject with probability proportional to $K_\epsilon(\|s_i - s_{\text{obs}}\|)$.
5. Repeat steps 1–3 until M acceptances are obtained.

As with parameter estimation, variants of the MCMC and SMC versions of ABC can also be used for model choice. It is also possible to use a regression-based variant of ABC to estimate the posterior probability of each model using multinomial logistic regression (Beaumont 2008) or logistic regression with a neural network (Blum & François 2010).

The approximation inherent in ABC depends on the choice of summary statistics, and this potentially affects the accuracy of model choice (Didelot et al. 2011, Robert et al. 2011). These authors note that, since the summary statistic vector is a deterministic function of the data,

$$p(x_{\text{obs}}|m) = p(x_{\text{obs}}, s_{\text{obs}}|m) = p(s_{\text{obs}}|m)p(x_{\text{obs}}|s_{\text{obs}}, m).$$

This implies that the ABC method will only give accurate estimates of the marginal likelihood ratio

$$p(s_{\text{obs}}|m_1)/p(s_{\text{obs}}|m_2)$$

if

$$p(x_{\text{obs}}|s_{\text{obs}}, m_1)/p(x_{\text{obs}}|s_{\text{obs}}, m_2) = 1,$$

which will only be the case if s_{obs} either is sufficient for both models (Grelaud et al. 2009) or gives the same departure from sufficiency. This observation could also apply to any ratio of posterior densities $p(s_{\text{obs}}|\theta_1)/p(s_{\text{obs}}|\theta_2)$ and is a necessary consequence of the approximation inherent in ABC. Didelot et al. (2011) show that $p(x_{\text{obs}}|s_{\text{obs}}, m_1)/p(x_{\text{obs}}|s_{\text{obs}}, m_2) = 1$ holds in the case where the models m_1 and m_2 are nested submodels of model m, for which s_{obs} is sufficient. However, Robert et al. (2011) argue this property will not hold more generally and give examples where ABC fails to converge to the true model as the sample size increases. Marin et al. (2014) show that a necessary condition for an ABC model choice algorithm to converges on the true model is that, as the sample size increases, the mean of the posterior predictive distribution of the summary statistic vector converge to different values under the different models. This has motivated approaches to identify summary statistics that are able to discriminate between models. Marin et al. (2014) propose to simulate samples from the posterior predictive distribution and show that the mean of the summary statistic vector is different under the two models. The method of Prangle et al. (2014b) follows directly from Fearnhead & Prangle (2012). They show that, given a summary statistic vector s, an optimal summary statistic $T(s)$ is the vector of posterior probabilities $T(s) = \{T_1(s), \ldots, T_{K-1}(s)\}$ for models $m_1, \ldots, m_{k-1}$, and use logistic regression to estimate $T(s)$ from pilot simulations. A related approach, suggested by Estoup et al. (2012), is to use linear discriminant analysis to project the summary statistic matrix into $K - 1$ orthogonal vectors that maximize the separability of the K models. A machine learning method using the random forests algorithm has been proposed by Pudlo et al. (2015). This gives a classifier based on a weighted set of decision trees derived from the summary statistics. Although the classification probability is not Bayesian, the method appears to perform favorably in many cases. It would appear that pathological model choice behavior of ABC typically arises when a small number of summary statistics are used relative to the number of parameters in the competing models. With a range of summary

statistics, and application of methods to reduce the dimensionality, it may be possible to achieve a satisfactory level of approximation.

6. MODEL-CHECKING

As suggested by Gelman et al. (2013), model-checking should form a natural part of the Bayesian approach to hypothesis testing and falsification. Model-checking within the ABC framework arises naturally as a consequence of the ready availability of samples from the prior predictive distribution in the standard formulation of ABC (Ratmann et al. 2009), which involves drawing samples from the joint distribution of parameters and summary statistics and then finding the marginal density

$$p(s) = \int p(s|\theta)\pi(\theta)\mathrm{d}\theta.$$

It is also straightforward to obtain samples from the posterior predictive distribution (Nott et al. 2018):

$$p_{s_{\mathrm{obs}}}(s) = \int p(s|\theta)p(\theta|s_{\mathrm{obs}})\mathrm{d}\theta.$$

Posterior predictive checks typically involve the computation of the empirical p-value of discriminatory summary statistics under a predictive distribution simulated from repeated draws of θ from the posterior (Gelman et al. 2013). Again, the ABC framework is helpful here because summaries are a natural part of the method. Gelman et al. (1996) suggest defining a discrepancy function $D(y, \theta)$, for example, based on the deviation of y from its expectation under θ. Nott et al. (2018) note that the advantage of a single discrepancy function is that it reduces the test to a univariate one. However, such a function may not be straightforwardly available in the case of intractable likelihood functions, and an empirical p-value based on the Mahalanobis distance of the observed summary statistic vector from the simulated posterior predictive mean vector may be a suitable alternative. Typically, these p-values are not well calibrated because of the induced association between simulated and observed summary statistics if using parameters sampled from the posterior (Rubin-Delanchy & Lawson 2014, Nott et al. 2018). A standard approach to calibration (Hjort et al. 2006) is to estimate posterior predictive p-values $(p_1, \ldots, p_Q)$ with Q pseudoobserved data sets (PODs) drawn from the prior predictive distribution, and then approximate the calibrated p-value as the fraction of p_j that are less than p_{obs}. This can be expensive because each POD requires computation of a separate posterior distribution followed by estimation of a posterior-predictive p-value. However, regression-adjusted ABC lends itself to this approach because the same simulated reference table $\{\theta_i, s_i\} \sim \pi(\theta) f_n(s|\theta)\ i = 1, \ldots, M$ can be reused for each POD. Nott et al. (2018) propose a method based on the regression-adjustment method of Blum & François (2010). Nott et al. show that with an ecological population-dynamic model, they achieve similar accuracy in a shorter time than a full-likelihood method.

6.1. Calibration and Coverage

Model-checking also requires an examination of the coverage properties of posterior distributions, introduced above in the context of calibration of posterior inferences. Calibration, in this context, is the property that under repeated sampling of true parameter values θ_i from the prior, the probability of observing θ_i to be in some region $\mathcal{A}$ of the posterior is $\int_{\mathcal{A}} p(\theta|s_{\mathrm{obs}})\mathrm{d}\theta$ (Cook et al. 2006). The need for an acceptance kernel as an integral part of the ABC method generally leads to

posterior distributions that are not perfectly calibrated (Fearnhead & Prangle 2012, Wilkinson 2013, Rodrigues et al. 2018). This can lead to bias, particularly for interval estimates. For example, in the case of a univariate normal model with variance σ^2 and kernel bandwidth ϵ, the ABC estimate will converge to $\sigma^2 - \epsilon$ with increasing n. This bias can become important in sequential algorithms, including filtering algorithms, because the bias can be propagated (Fearnhead & Prangle 2012, Dean et al. 2014). One alternative for avoiding this is a slight modification of the ABC algorithm to compute a noisy value for the observed summary statistic $s'_{\text{obs}} = s_{\text{obs}} + x$ where $x \sim K_\epsilon(x)$, which is perfectly calibrated (Fearnhead & Prangle 2012). In the limit as $\epsilon \to 0$, the standard and noisy versions of ABC converge to the same posterior. This corresponds to the viewpoint suggested by Wilkinson (2013) that ABC can be regarded as exact Monte Carlo in which the rejection kernel corresponds to the stochastic observation equation of the model—i.e., it is exact for a different model. There is a variance/bias tradeoff in the application of the noisy versus standard versions of ABC, and the use of the noisy version is only recommended for longer iterations of an SMC algorithm (Dean et al. 2014, Yıldırım et al. 2015).

Although standard ABC is not perfectly calibrated, measuring the degree of departure as a function of bandwidth ϵ is a useful part of model-checking. One general approach for Monte Carlo methods (Cook et al. 2006) is to examine posterior coverage of the true parameter value, drawn from the prior, $\theta^* \sim \pi(\theta)$. If the method is perfectly calibrated, the empirical p-value $\Pr(\theta \leq \theta^* | s_{\text{obs}})$ should be uniform across many draws of θ^* (Wegmann et al. 2009). However, as noted by Prangle et al. (2014a), θ^* should also give uniform p-values if the algorithm simply returned the prior $\pi(\theta)$, so the test is potentially conservative. In fact, values θ^* drawn from any region $A \propto p(\theta^*, s_{\text{obs}}) I\{s_{\text{obs}} \in A\}$ should be calibrated, and, assuming the test is motivated around a target real data set, Prangle et al. (2014b) suggest choosing A to be in the vicinity of the real target, which, in general, will give a different marginal distribution for θ^* than if the prior was used. Their algorithm for model choice is similarly motivated. In this case, they suggest that under perfect coverage, the model label associated with a data set giving posterior probability z_A for a reference model A should follow a Bernoulli distribution with probability z_A. Thus, a test for coverage would be to see if there is a linear relationship with slope 1 and intercept 0 between the predicted and observed probability of belonging to model A.

Rodrigues et al. (2018) suggest that the method for assessing calibration can also be used in a postprocessing step to transform samples θ_i so that they are approximately calibrated. From the distribution of $\theta_i \sim p_\epsilon(\theta | s_{\text{obs}})$ it is possible to estimate the distribution function $F_{\text{obs}}(\theta)$. Furthermore, conditional on each s_i jointly simulated with the θ_i in the standard ABC prior predictive distribution $\{\theta_i, s_i\} \sim \pi(\theta) f_n(s|\theta)$, it is also possible to estimate an ABC posterior with distribution function $F_i(\theta)$. Rodrigues et al. (2018) propose the transformation $\theta'_i = F^{-1}_{\text{obs}}(p_i)$ where the p_i are calculated from the simulated distribution functions as $p_i = F_i(\theta)\{\theta_i\}$. Rodrigues et al. (2018) offer various methods for improving the computational efficiency of the approach.

7. RELATED METHODS

A relative of the ABC approach is the method of indirect inference introduced by Gourieroux et al. (1993) and Heggland & Frigessi (2004), which is typically used in the context of maximum likelihood estimation. An approximating model is developed for the problem, with tractable likelihood. The parameters of this model can be estimated by standard maximum likelihood. Denoting by $\hat{\phi}_{\text{obs}}$ and $\hat{\phi}_{\text{sim}}$ the maximum likelihood estimation for the observed and simulated data, $\hat{\theta}$ for the target modeled is estimated by minimizing some measure of discrepancy between $\hat{\phi}_{\text{obs}}$ and $\hat{\phi}_{\text{sim}}$. Comparisons and discussions of ABC and indirect inference are given by Fearnhead & Prangle (2012) and Drovandi et al. (2015). For intractable models that are also very expensive to simulate,

an alternative to ABC is the method of Bayesian emulation (Kennedy & O'Hagan 2001), which aims to fit a more tractable approximating model using a set of pilot simulations. A comparison and implementation in an ABC context is given by Jabot et al. (2014) (see also Holden et al. 2018).

The method of synthetic likelihood (Wood 2010), as with the ABC approach of Leuenberger & Wegmann (2010), aims to model the likelihood $f_n(s|\theta)$ as a multivariate normal. This is applied in an MCMC framework. Instead of the standard rejection kernel $\mathbb{I}\{\|s - s_{\text{obs}}\| \leq \epsilon\}$ used in step 4 of Algorithm 2, a direct estimate of $f_n(s|\theta')$ for proposed θ' is made from J simulations $s_j \sim f_n(s|\theta')$ and by estimating the sample mean and covariance:

$$\hat{\mu} = (1/J)\sum_{j=1}^{J} s_j$$

$$\hat{\sigma} = (1/(j-1))\sum_{j=1}^{J}(s_j - \hat{\mu})(s_j - \hat{\mu})^T.$$

An estimate of the likelihood can then be obtained from the multivariate density:

$$\hat{f}(s|\theta') := \text{MVN}_{\theta'}(\hat{\mu}, \hat{\Sigma}).$$

Example applications of synthetic likelihood in ecology, including comparisons with ABC, are described by Fasiolo & Wood (2018). The use of unbiased estimates of the mean and covariance does not lead to unbiased estimates of the density. However, Price et al. (2017) use the theory of Ghurye & Olkin (1969) to obtain unbiased density estimates. The resulting algorithm corresponds to a pseudomarginal MCMC sampler (Andrieu & Roberts 2009) in the case that the likelihood is multivariate normal. Although implemented in an MCMC algorithm, in principle, the synthetic likelihood approach can also be used to sample from the posterior, as in a standard Monte Carlo rejection algorithm (Price et al. 2017). In that case it can be seen that, in contrast with ABC, which uses a fixed kernel, the synthetic likelihood method uses a rejection kernel that is adaptively a function of θ. However, the synthetic likelihood method is potentially costly because typically $J \gg 1$, and Wilkinson (2014) and Meeds & Welling (2014) suggest different methods for improving efficiency by recycling estimates during the algorithm using a Gaussian process approximation.

8. CONCLUSIONS

It should be apparent from this review that the original methodology introduced by Pritchard et al. (1999) has evolved into a family of likelihood-free statistical methods that carry the umbrella term of ABC. The basic philosophy behind these approaches is that by constructing and simulating from generative models, one can understand the target system better, and also make better predictions and test hypotheses.

This review has attempted to give some overview of current developments in ABC, although, since it is still a rapidly evolving field of research, it is difficult to pick out the major strands. There have recently been advances in the theoretical justification of the ABC approach, particularly with regard to the apparent uncertainty arising from the choice of summary statistics. A recent theme has been the utilization of methods, some of which derive from the machine learning community, to improve performance and efficiency of ABC, particularly when using large data sets. For large-scale modeling, the application of ABC with emulation methods seems promising (Jabot et al. 2014, Holden et al. 2018). An area that seems worthy of further investigation is the behavior of ABC methods under model misspecification (Frazier et al. 2017, Ratmann et al. 2009). Overall,

however, it is clear that ABC is a promising method for tackling a variety of scientific problems and that the theoretical developments and availability of software appear to be keeping pace with the expansion in its range of applications.

DISCLOSURE STATEMENT

The author is not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

The author would like to thank the editor and anonymous reviewer for their queries and suggestions, which improved this review.

LITERATURE CITED

Andrieu C, Roberts GO. 2009. The pseudo-marginal approach for efficient Monte Carlo computations. *Ann. Stat.* 37:697–725

Baragatti M, Grimaud A, Pommeret D. 2013. Likelihood-free parallel tempering. *Stat. Comput.* 23:535–49

Barthelmé S, Chopin N. 2014. Expectation propagation for likelihood-free inference. *J. Am. Stat. Assoc.* 109:315–33

Barthelmé S, Chopin N, Cottet V. 2018. Divide and conquer in ABC: expectation-propagation algorithms for likelihood-free inference. In *Handbook of Approximate Bayesian Computation*, ed. SA Sisson, Y Fan, MA Beaumont, pp. 415–34. Boca Raton, FL: CRC

Beaumont MA. 2003. Estimation of population growth or decline in genetically monitored populations. *Genetics* 164:1139–60

Beaumont MA. 2008. Joint determination of topology, divergence time and immigration in population trees. In *Simulations, Genetics and Human Prehistory*, ed. S Matsumura, P Forster, C Renfrew, pp. 135–54. Cambridge, UK: McDonald Inst.

Beaumont MA, Cornuet JM, Marin JM, Robert CP. 2009. Adaptive approximate Bayesian computation. *Biometrika* 96:983–90

Beaumont MA, Zhang W, Balding DJ. 2002. Approximate Bayesian computation in population genetics. *Genetics* 162:2025–35

Blum MG. 2010. Approximate Bayesian computation: a nonparametric perspective. *J. Am. Stat. Assoc.* 105:1178–87

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

Blum MG, Nunes MA, Prangle D, Sisson SA, et al. 2013. A comparative review of dimension reduction methods in approximate Bayesian computation. *Stat. Sci.* 28:189–208

Bornn L, Pillai NS, Smith A, Woodard D. 2017. The use of a single pseudo-sample in approximate Bayesian computation. *Stat. Comput.* 27:583–90

Bortot P, Coles SG, Sisson SA. 2007. Inference for stereological extremes. *J. Am. Stat. Assoc.* 102:84–92

Breiman L. 2001. Statistical modeling: the two cultures (with comments and a rejoinder by the author). *Stat. Sci.* 16:199–231

Cappé O, Guillin A, Marin JM, Robert CP. 2004. Population Monte Carlo. *J. Comput. Graph. Stat.* 13:907–29

Chan J, Perrone V, Spence JP, Jenkins PA, Mathieson S, Song YS. 2018. A likelihood-free inference framework for population genetic data using exchangeable neural networks. arXiv:1802.06153 [cs.LG]

Cook SR, Gelman A, Rubin DB. 2006. Validation of software for Bayesian models using posterior quantiles. *J. Comput. Graph. Stat.* 15:675–92

Creel M. 2017. Neural nets for indirect inference. *Econom. Stat.* 2:36–49

Csilléry K, François O, Blum MG. 2012. abc: an R package for approximate Bayesian computation (ABC). *Methods Ecol. Evol.* 3:475–79

Cui T, Peeters L, Pagendam D, Pickett T, Jin H, et al. 2018. Emulator-enabled approximate Bayesian computation (ABC) and uncertainty analysis for computationally expensive groundwater models. *J. Hydrol.* 564:191–207

Daly AC, Cooper J, Gavaghan DJ, Holmes C. 2017. Comparing two sequential Monte Carlo samplers for exact and approximate Bayesian inference on biological models. *J. R. Soc. Interface* 14:20170340

Dean TA, Singh SS, Jasra A, Peters GW. 2014. Parameter estimation for hidden Markov models with intractable likelihoods. *Scand. J. Stat.* 41:970–87

Del Moral P, Doucet A, Jasra A. 2012. An adaptive sequential Monte Carlo method for approximate Bayesian computation. *Stat. Comput.* 22:1009–20

Didelot X, Everitt RG, Johansen AM, Lawson DJ. 2011. Likelihood-free estimation of model evidence. *Bayesian Anal.* 6:49–76

Diggle PJ, Gratton RJ. 1984. Monte Carlo methods of inference for implicit statistical models. *J. R. Stat. Soc. B* 46:193–227

Drovandi CC, Pettitt AN. 2011. Estimation of parameters for macroparasite population evolution using approximate Bayesian computation. *Biometrics* 67:225–33

Drovandi CC, Pettitt AN, Lee A. 2015. Bayesian indirect inference using a parametric auxiliary model. *Stat. Sci.* 30:72–95

Estoup A, Lombaert E, Marin JM, Guillemaud T, Pudlo P, et al. 2012. Estimation of demo-genetic model probabilities with approximate Bayesian computation using linear discriminant analysis on summary statistics. *Mol. Ecol. Resourc.* 12:846–55

Fan Y, Nott DJ, Sisson SA. 2013. Approximate Bayesian computation via regression density estimation. *Stat* 2:34–48

Fasiolo M, Wood SN. 2018. ABC in ecological modelling. In *Handbook of Approximate Bayesian Computation*, ed. SA Sisson, Y Fan, MA Beaumont, pp. 597–622. Boca Raton, FL: CRC

Fearnhead P, Prangle D. 2012. Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation. *J. R. Stat. Soc. B* 74:419–74

Frazier DT, Martin GM, Robert CP, Rousseau J. 2018. Asymptotic properties of approximate Bayesian computation. *Biometrika* 105:593–607

Frazier DT, Robert CP, Rousseau J. 2017. Model misspecification in ABC: consequences and diagnostics. arXiv:1708.01974 [math.ST]

Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. 2013. *Bayesian Data Analysis*. Boca Raton, FL: CRC

Gelman A, Meng XL, Stern H. 1996. Posterior predictive assessment of model fitness via realized discrepancies. *Stat. Sinica* 6:733–60

Gelman A, Vehtari A, Jylänki P, Robert C, Chopin N, Cunningham JP. 2014. Expectation propagation as a way of life. arXiv:1412.4869 [stat.CO]

Ghurye S, Olkin I. 1969. Unbiased estimation of some multivariate probability densities and related functions. *Ann. Math. Stat.* 40:1261–71

Gourieroux C, Monfort A, Renault E. 1993. Indirect inference. *J. Appl. Econom.* 8:S85–118

Grelaud A, Robert CP, Marin JM, Rodolphe F, Taly JF. 2009. ABC likelihood-free methods for model choice in Gibbs random fields. *Bayesian Anal.* 4:317–36

Hahn C, Vakili M, Walsh K, Hearin AP, Hogg DW, Campbell D. 2017. Approximate Bayesian computation in large-scale structure: constraining the galaxy–halo connection. *Mon. Notices R. Astron. Soc.* 469:2791–805

Hasenclever L, Webb S, Lienart T, Vollmer S, Lakshminarayanan B, et al. 2017. Distributed Bayesian learning with stochastic natural gradient expectation propagation and the posterior server. *J. Mach. Learn. Res.* 18:3744–80

Heggland K, Frigessi A. 2004. Estimating functions in indirect inference. *J. R. Stat. Soc. B* 66:447–62

Hjort NL, Dahl FA, Steinbakk GH. 2006. Post-processing posterior predictive p-values. *J. Am. Stat. Assoc.* 101:1157–74

Holden PB, Edwards NR, Hensman J, Wilkinson RD. 2018. ABC for climate: dealing with expensive simulators. In *Handbook of Approximate Bayesian Computation*, ed. SA Sisson, Y Fan, MA Beaumont, pp. 569–95. Boca Raton, FL: CRC

Jabot F, Faure T, Dumoulin N. 2013. EasyABC: performing efficient approximate Bayesian computation sampling schemes using R. *Methods Ecol. Evol.* 4:684–87

Jabot F, Lagarrigues G, Courbaud B, Dumoulin N. 2014. A comparison of emulation methods for approximate Bayesian computation. arXiv:1412.7560 [q-bio.QM]

Jabot F, Lohier T. 2016. Non-random correlation of species dynamics in tropical tree communities. *Oikos* 125:1733–42

Jiang B, Wu T, Zheng C, Wong WH. 2017. Learning summary statistic for approximate Bayesian computation via deep neural network. *Stat. Sinica* 27:1595–1618

Joyce P, Marjoram P. 2008. Approximately sufficient statistics and Bayesian computation. *Stat. Appl. Genet. Mol. Biol*. **https://doi.org/10.2202/1544-6115.1389**

Kandler A, Powell A. 2018. Generative inference for cultural evolution. *Phil. Trans. R. Soc. B* 373:20170056

Kennedy MC, O'Hagan A. 2001. Bayesian calibration of computer models. *J. R. Stat. Soc. B* 63:425–64

Kousathanas A, Duchen P, Wegmann D. 2018. A guide to general-purpose ABC software. In *Handbook of Approximate Bayesian Computation*, ed. SA Sisson, Y Fan, MA Beaumont, pp. 369–413. Boca Raton, FL: CRC

Kousathanas A, Leuenberger C, Helfer J, Quinodoz M, Foll M, Wegmann D. 2016. Likelihood-free inference in high-dimensional models. *Genetics* 203:893–904

Leuenberger C, Wegmann D. 2010. Bayesian computation and model selection without likelihoods. *Genetics* 184:243–52

Li W, Fearnhead P. 2018a. Convergence of regression adjusted approximate Bayesian computation. *Biometrika* 105:301–18

Li W, Fearnhead P. 2018b. On the asymptotic efficiency of approximate Bayesian computation estimators. *Biometrika* 105:285–99

Liepe J, Kirk P, Filippi S, Toni T, Barnes CP, Stumpf MP. 2014. A framework for parameter estimation and model selection from experimental data in systems biology using approximate Bayesian computation. *Nat. Protoc.* 9:439

Marin JM, Pillai NS, Robert CP, Rousseau J. 2014. Relevant statistics for Bayesian model choice. *J. R. Stat. Soc. B* 76:833–59

Marin JM, Raynal L, Pudlo P, Ribatet M, Robert CP. 2016. ABC random forests for Bayesian parameter inference. arXiv:1605.05537 [stat.ME]

Marjoram P, Molitor J, Plagnol V, Tavaré S. 2003. Markov chain Monte Carlo without likelihoods. 100:15324–28

McKinley TJ, Vernon I, Andrianakis I, McCreesh N, Oakley JE, et al. 2018. Approximate Bayesian computation and simulation-based inference for complex stochastic epidemic models. *Stat. Sci.* 33:4–18

Meeds E, Welling M. 2014. GPS-ABC: Gaussian process surrogate approximate Bayesian computation. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, ed. N Zhang, J Tian, pp. 593–602. Arlington, VA: AUAI

Minka TP. 2001. Expectation propagation for approximate Bayesian inference. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, ed. J Breese, D Koller, pp. 362–69. San Francisco: Morgan Kaufmann

Mitrovic J, Sejdinovic D, Teh YW. 2016. DR-ABC: approximate Bayesian computation with kernel-based distribution regression. In *Proceedings of The 33rd International Conference on Machine Learning*, pp. 1482–91. Brookline, MA: Microtome

Nakagome S, Fukumizu K, Mano S. 2013. Kernel approximate Bayesian computation in population genetic inferences. *Stat. Appl. Genet. Mol. Biol.* 12:667–78

Nott DJ, Drovandi CC, Mengersen K, Evans M. 2018. Approximation of Bayesian predictive *p*-values with regression ABC. *Bayesian Anal.* 13:59–83

Nunes MA, Balding DJ. 2010. On optimal selection of summary statistics for approximate Bayesian computation. *Stat. Appl. Genet. Mol. Biol*. **https://doi.org/10.2202/1544-6115.1576**

Peskun PH. 1973. Optimum Monte-Carlo sampling using Markov chains. *Biometrika* 60:607–12
Prangle D. 2017. Adapting the ABC distance function. *Bayesian Anal.* 12:289–309
Prangle D. 2018. Summary statistics in approximate Bayesian computation. In *Handbook of Approximate Bayesian Computation*, ed. SA Sisson, Y Fan, MA Beaumont, pp. 125–52. Boca Raton, FL: CRC
Prangle D, Blum M, Popovic G, Sisson S. 2014a. Diagnostic tools for approximate Bayesian computation using the coverage property. *Aust. N. Z. J. Stat.* 56:309–29
Prangle D, Fearnhead P, Cox MP, Biggs PJ, French NP. 2014b. Semi-automatic selection of summary statistics for ABC model choice. *Stat. Appl. Genet. Mol. Biol.* 13:67–82
Price LF, Drovandi CC, Lee A, Nott DJ. 2017. Bayesian synthetic likelihood. *J. Comput. Graph. Stat.* **https://doi.org/10.1080/10618600.2017.1302882**
Pritchard JK, Seielstad MT, Perez-Lezaun A, Feldman MW. 1999. Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Mol. Biol. Evol.* 16:1791–98
Pudlo P, Marin JM, Estoup A, Cornuet JM, Gautier M, Robert CP. 2015. Reliable ABC model choice via random forests. *Bioinformatics* 32:859–66
Ratmann O, Andrieu C, Wiuf C, Richardson S. 2009. Model criticism based on likelihood-free inference, with an application to protein network evolution. *PNAS* 106:10576–81
Ratmann O, Jørgensen O, Hinkley T, Stumpf M, Richardson S, 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:e230
Robert CP, Cornuet JM, Marin JM, Pillai NS. 2011. Lack of confidence in approximate Bayesian computation model choice. *PNAS* 108:15112–17
Rodrigues G, Prangle D, Sisson S. 2018. Recalibration: a post-processing method for approximate Bayesian computation. *Comput. Stat. Data Anal.* 126:53–66
Rubin-Delanchy P, Lawson DJ. 2014. Posterior predictive p-values and the convex order. arXiv:1412.3442 [math.ST]
Scott SL. 2017. Comparing consensus Monte Carlo strategies for distributed Bayesian computation. *Braz. J. Probability Stat.* 31:668–85
Scott SL, Blocker AW, Bonassi FV, Chipman HA, George EI, McCulloch RE. 2016. Bayes and big data: the consensus Monte Carlo algorithm. *Int. J. Manag. Sci. Eng. Manag.* 11:78–88
Singh H, Misra N, Hnizdo V, Fedorowicz A, Demchuk E. 2003. Nearest neighbor estimates of entropy. *Am. J. Math. Manag. Sci.* 23:301–21
Sisson SA, Fan Y. 2018. ABC samplers. In *Handbook of Approximate Bayesian Computation*, ed. SA Sisson, Y Fan, MA Beaumont, pp. 87–123. Boca Raton, FL: CRC
Sisson SA, Fan Y, Tanaka MM. 2007. Sequential Monte Carlo without likelihoods. *PNAS* 104:1760–65
Sisson SA, Fan Y, Tanaka MM. 2009. Correction: sequential Monte Carlo without likelihoods. *PNAS* 106:16889–89
Sjödin P, E. Sjöstrand A, Jakobsson M, Blum MG. 2012. Resequencing data provide no evidence for a human bottleneck in Africa during the penultimate glacial period. *Mol. Biol. Evol.* 29:1851–60
Sousa VC, Fritz M, Beaumont MA, Chikhi L. 2009. Approximate Bayesian computation without summary statistics: the case of admixture. *Genetics* 181:1507–19
Toni T, Welch D, Strelkowa N, Ipsen A, Stumpf MP. 2009. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *J. R. Soc. Interface* 6:187–202
Turner BM, Dennis S, Van Zandt T. 2013. Likelihood-free Bayesian analysis of memory models. *Psychol. Rev.* 120:667–78
van der Vaart E, Beaumont MA, Johnston AS, Sibly RM. 2015. Calibration and evaluation of individual-based models using approximate Bayesian computation. *Ecol. Model.* 312:182–90
Wegmann D, Leuenberger C, Excoffier L. 2009. Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood. *Genetics* 182:1207–18
Wegmann D, Leuenberger C, Neuenschwander S, Excoffier L. 2010. ABCtoolbox: a versatile toolkit for approximate Bayesian computations. *BMC Bioinformat.* 11:116
Wilkinson RD. 2013. Approximate Bayesian computation (ABC) gives exact results under the assumption of model error. *Stat. Appl. Genet. Mol. Biol.* 12:129–41

Wilkinson RD. 2014. Accelerating ABC methods using Gaussian processes. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, ed. S Kaski, J Corander, pp. 1015–23. Brookline, MA: Microtome

Wood SN. 2010. Statistical inference for noisy nonlinear ecological dynamic systems. *Nature* 466:1102–4

Yıldırım S, Singh SS, Dean T, Jasra A. 2015. Parameter estimation in hidden Markov models with intractable likelihoods using sequential Monte Carlo. *J. Comput. Graph. Stat.* 24:846–65