AN OVERVIEW ON APPROXIMATE BAYESIAN COMPUTATION [*]

Meïli Baragatti[1] and Pierre Pudlo[2]

Abstract. Approximate Bayesian computation techniques, also called likelihood-free methods, are one of the most satisfactory approach to intractable likelihood problems. This overview presents recent results since its introduction about ten years ago in population genetics.

Résumé. Les méthodes bayésiennes approchées constituent l'un des outils majeurs d'inférence statistique en dimension finie lorsque la vraisemblance du modèle paramétrique considéré n'est pas accessible. Nous présentons quelques résultats récents qui ont permis d'augmenter significativement l'efficacité de ces techniques depuis leurs introductions dans le domaine de la génétique des populations il y a maintenant plus d'une dizaine année.

1. Introduction

When conducting a parametric Bayesian analysis, Monte Carlo methods aims at approximating the posterior via the empirical distribution of a simulated sample on the parameter space, Θ say. More precisely, in the regular case, the posterior has density

$$\pi(\theta \,|\, x_{\mathrm{obs}}) \propto \pi(\theta) f(x_{\mathrm{obs}} \,|\, \theta), \tag{1}$$

where $\pi(\theta)$ is the prior density on Θ and $f(x \,|\, \theta)$ the density of a stochastic model. Most of the Monte Carlo algorithms require numerical evaluations of the likelihood $f(x_{\mathrm{obs}} \,|\, \theta)$ at many values of θ. But the likelihood may be intractable when the stochastic model is based on a latent process $u \in \mathscr{U}$, $i.e.$,

$$f(x \,|\, \theta) = \int_{\mathscr{U}} f(x, u \,|\, \theta) \mathrm{d}u,$$

or when the likelihood is known up to a normalising constant depending on θ, $i.e.$,

$$f(x \,|\, \theta) = \frac{g(x, \theta)}{Z_\theta}, \quad \text{where } Z_\theta = \int g(x, \theta) \mathrm{d}x.$$

And some models such as hidden Markov random fields suffer from both difficulties, see, e.g., Everitt (2012).

Since their introduction by Tavaré et al. (1997), approximate Bayesian computation (ABC) methods have been widely used with intractable likelihoods. The basic ABC samplers are presented in Section 2, for parameter estimation in Paragraph 2.1 and for model choice in Paragraph 2.4. Actually, ABC is based on a hierarchy of approximations of the posterior discussed in Paragraphs 2.2 and 2.3. Calibration of ABC schemes are discussed in Section 3. And other samplers focused on speeding up the computation are given in Section 4.

2. Basic samplers and their targets

The basic idea underlying ABC is as follow. Using simulations from the stochastic model, we can produce a simulated sample from the joint distribution

$$\pi(\theta)f(x \,|\, \theta). \tag{2}$$

The posterior distribution is then the conditional distribution of (2) knowing that the data x is equal to the observation x_{obs}. From the joint, simulated sample, ABC derives approximations of the conditional density $\pi(\theta \,|\, x_{\mathrm{obs}})$ and of other functional of the posterior such as moments or quantiles.

2.1. Acceptation-rejection sampler

The previous elegant idea suffers from two shortcomings. First, the algorithm might be time consuming if simulating from the stochastic model is not straightforward. We postponed presentation of other ABC algorithms that speed up the computation in Section 4. But the most profound problem is that estimating the conditional distribution of θ knowing $x = x_{\mathrm{obs}}$ requires that some simulations fall into the neighbourhood of x_{obs}. If the data space $\mathscr{X}$ is not of very small dimension, we face the curse of dimensionality, namely that it is almost impossible to simulate a data set near the observed one. To solve the problem, ABC schemes perform a (non linear) projection of the (observed and simulated) data sets on a space of reasonable dimension d via some summary statistics $\eta : \mathscr{X} \to \mathbb{R}^d$ and set a metric $\delta[s, s']$ on $\mathbb{R}^d$.

Algorithm 1 ABC acceptation-rejection algorithm with a given threshold ε

1: **for** $i = 1 \to N_{\mathrm{prior}}$ **do**
2:     **draw** θ_i from the prior $\pi(\theta)$
3:     **draw** x from the likelihood $f(x \,|\, \theta_i)$
4:     **compute** $s_i = \eta(x)$
5:     **store** the particle (θ_i, s_i)
6: **end for**
7: **for** $i = 1 \to N_{\mathrm{prior}}$ **do**
8:     **compute** the distance $\delta[s_i, s_{\mathrm{obs}}]$
9:     **reject** the particle if $\delta(s_i, s_{\mathrm{obs}}) > \varepsilon$
10: **end for**

Note that the first loop between lines 1– 6 does not depend on the data set x_{obs}. Therefore, the simulated particles might be reused to analyze other data sets. In the second loop (lines 7– 10), the particles might be weighted by $w_i = K_\varepsilon(\delta[s_i, s_{\mathrm{obs}}])$, where K_ε is some smoothing kernel with bandwidth ε. Then, the acceptation-rejection algorithm corresponds to $K_\varepsilon(\delta) = \mathbf{1}\{\delta \leq \varepsilon\}$.

2.2. Target of the algorithm and approximations

The holy grail of the ABC scheme is the posterior distribution (1). To bypass the curse of dimensionality, ABC introduces the non linear projection $\eta : \mathscr{X} \to \mathbb{R}^d$. Hence, we cannot recover anything better than the conditional distribution of θ knowing that the simulated summary statistics $\eta(x)$ is equal to $s_{\mathrm{obs}} = \eta(x_{\mathrm{obs}})$, *i.e.*,

$$\pi\big(\theta \,|\, \eta(x) = s_{\mathrm{obs}}\big). \tag{3}$$

Moreover, the event $\eta(x) = s_{\text{obs}}$ might have very low probability, if not null. Hence, the output of the ABC algorithm is a sample from the distribution conditioned by the larger event $A_\varepsilon = \{\delta[\eta(x), s_{\text{obs}}] \leq \varepsilon\}$, namely

$$\pi\big(\theta \,\big|\, \delta[\eta(x), s_{\text{obs}}] \leq \varepsilon\big). \tag{4}$$

The distribution (4) tends to (3) when $\varepsilon \to 0$. But, if the user want to keep the size of the output constant, decreasing the threshold ε might be problematic in terms of processing time: indeed, the algorithm requires an average number of simulations $N_{\text{prior}} = N/\pi(A_\varepsilon)$ and $\pi(A_\varepsilon)$ can decrease very fast toward 0 when $\varepsilon \to 0$. Actually, the error we commit be estimating the density of the output of the algorithm rather than computing explicitly (3) has been widely studied in nonparametric statistics since the seminal paper of Rosenblatt (1969).

2.3. **Post-processing the output of the sampler**

In order to provide a better approximate of (3) with the ABC sample, Beaumont et al. (2002) proposed a post-process substituting the output (θ_i, s_i), $i = 1, \ldots, N$, of the algorithm with the sample $\theta_i^\star = \theta_i - \hat{\beta}(s_i - s_{\text{obs}})$, $i = 1, \ldots, N$, where $\hat{\beta}$ is computed via a local linear regression of θ given s. The rationale behind that computation is as follow. Conditionally on A_ε, if the distribution of $(\theta, \eta(x))$ were driven by the linear model

$$\theta = \beta\eta(x) + u,$$

with u independent of $\eta(x)$ and $\mathbb{E}(u) = 0$, then $\beta s_{\text{obs}} + u$ were distributed as θ in (3). Hence, apart from the fact that β is replaced with its estimate, the $\theta_i^\star$ were distributed according to (3). During the post-process, the accepted particles are weighted by $w_i = K_\varepsilon(\delta[s_i, s_{\text{obs}}])$ where K_ε is a nonparameteric kernel with bandwidth ε. The coefficient β is then estimated via a weighted least square regression. This post-process is easy to implement and thus very popular among ABC practitioners. It should be seen as a trick to decrease ε without increasing the computation time.

It has been theoretically studied by Blum (2010) when the aim is a posterior density estimation. Alternatives using non-linear regression have also been proposed by Blum and François (2010). The principle is to model the distribution of $(\theta, \eta(x))$ by

$$\theta = g(\eta(x)) + u,$$

and then to estimate g by a neural net with one hidden layer.

Sadly, the error we commit by replacing (1) with (3) is the most difficult to control. Of course, distributions (1) and (3) are equal when the statistic $\eta(x)$ is sufficient.

2.4. **ABC model choice**

We can also perform a Bayesian model choice with ABC (Grelaud et al., 2009). Consider that we have a finite set of models, $\mathscr{M}_k$, $k = 1, \ldots, m$, and that each of them is parametrised by some vector $\theta \in \Theta_k$. Each parameter space (whose dimension can depend on k) is equipped with a prior distribution $\pi_k(\theta)$. And the set of models is equipped with a (discrete) prior distribution $\pi(k)$. But the non-linear projection $\eta : \mathscr{X} \to \mathbb{R}^d$ via the summary statistics should not depend on k. Denoting by $f_k(x \,|\, \theta)$ the likelihood of model $\mathscr{M}_k$, the ABC model choice is given in Algorithm 2.

The accepted particles (k_i) at the end of the algorithm are distributed according to

$$\pi\big(k \,\big|\, \delta[s, s_{\text{obs}}] \leq \varepsilon\big).$$

And

$$\hat{\pi}(k \,|\, x_{\text{obs}}) = \frac{\sum \mathbf{1}\{k_i = k, \ \delta[s_i, s_{\text{obs}}] \leq \varepsilon\}}{\sum \mathbf{1}\{\delta[s_i, s_{\text{obs}}] \leq \varepsilon\}}$$

is an estimate of the posterior probability of model $\mathscr{M}_k$.

Algorithm 2 ABC model choice

1: **for** $i = 1 \rightarrow N_{\mathrm{prior}}$ **do**
2:     **draw** k_i from the prior $\pi(k)$
3:     **draw** $\theta \in \Theta_{k_i}$ from $\pi_{k_i}(\theta)$
4:     **draw** x from the likelihood $f_{k_i}(x \,|\, \theta)$ of model $\mathscr{M}_{k_i}$
5:     **compute** $s_i = \eta(x)$
6:     **store** the particle (k_i, s_i)
7: **end for**
8: **for** $i = 1 \rightarrow N_{\mathrm{prior}}$ **do**
9:     **compute** the distance $\delta[s_i, s_{\mathrm{obs}}]$
10:     **reject** the particle (k_i, s_i) if $\delta[s_i, s_{\mathrm{obs}}] > \varepsilon$
11: **end for**

There has been a debate whether we can trust ABC (Robert et al., 2011). The first important fact one should be aware is that the concatenation of sufficient statistics for each model does not give a sufficient statistic for the model choice problem. For example, assume that the data set is composed of iid Gaussian drawings with mean θ and that we have two models at hand, the first one being $\mathscr{N}((\theta, \ldots, \theta), I)$ and the second one being $\mathscr{N}((\theta, \ldots, \theta), 2I)$, where I is the identity matrix. Then, $\eta(x) = \sum x_i$ is a well known sufficient statistic (θ is the expected value of iid drawings of Gaussian distributions) for both models. But $\eta(x)$ will not be able to discriminate the models. Actually, the covariance of the data set is a sufficient statistics for the above model choice. The debate on confidence of ABC model choice has been closed by Marin et al. (2013), who provide conditions on the summary statistics $\eta(x)$ to ensure consistency of ABC model choice when the size of the data set increases.

3. Calibration

3.1. Acceptation threshold

When the Bayesian analysis requires ABC, it is hard or event impossible to set the acceptation threshold ε in advance, *i.e.*, before any simulation. Most often, the user sets the computational effort that he is prepared to face, *i.e.*, N_{prior}, the number of simulations from the model and sets the size of the output, N, for the accuracy of the sample estimates (kernel density estimates, quantile estimates, etc.). Then the threshold ε appears as a Monte Carlo approximation of the quantile of order $\alpha = N/N_{\mathrm{prior}}$ of the distances $d(\eta(x), s_{\mathrm{obs}})$ when $x \sim \int \pi(\theta) f(x \,|\, \theta) \mathrm{d}\theta$.

Because this calibration procedure of ε depends on the simulated data sets, its theoretical analysis is complex. Actually, it can be seen then as an N-nearest neighbour procedure among a simulated sample of size N_{prior}. This is the point of view adopted by Biau et al. (2012), who prove asymptotic properties of the posterior density estimation. Their theoretical results favour a calibration of N such that $N \approx N_{\mathrm{prior}}^{(m+4)/(m+d+4)}$ when m is the dimension of θ and $d > 4$ is the dimension of $\eta(x)$.

Another more original way to calibrate ε has been proposed by Ratmann et al. (2013) in order to obtain first order exact approximations. The authors assume that the data set might be decomposed into n iid blocks, and that the vector of summary statistics has the same structure. They take advantage of that last structure to calibrate ε and the number of independent blocks in the simulated data sets.

As in Wilkinson (2013), one might also consider that the observed summary statistics are noisy, either because the (intractable) likelihood comes from a model that does not fit to the reality or because their data collection process has added some noise. Hence, one might be appealed to set a model on the summary statistics space such that

$$s_{\mathrm{noisy}} = \eta(x) + \gamma, \quad \text{where } x \sim f(x \,|\, \theta) \text{ and } \gamma \sim K_\varepsilon(\gamma). \tag{5}$$

Then the threshold ε (or the bandwidth of the kernel K_ε on the summary statistics space that weights the accepted particles in the algorithm) is calibrated by an evaluation of the error and ABC becomes exact in that framework.

3.2. **Summary statistics**

Let us go back to our initial problem, which is computing (1) with a Monte Carlo approximation of (4). The major part of the difference is controlled by the choice of the summary statistics $\eta(x)$, which is paramount to the success of the approximation. In many concrete situations, the statistics are chosen by the practitioner, depending on the problem at hand. Several studies provide guidance in this choice.

Joyce and Marjoram (2008) have proposed an inclusion-exclusion scheme assuming we have at our disposal a large set of univariate summary statistics. This scheme is related to the simulation experiments performed by McKinley et al. (2009). A review discussing dimension reduction methods in ABC has been written by Blum et al. (2012). The most satisfactory method to assess the ability of a given summary statistic to capture information on θ is based on the post-process presented above. A first numerical study has been performed by Sedki and Pudlo (2012). Aeschbacher et al. (2012) advocate inclusion/exclusion via boosting, a machine-learning algorithm combining a set of weaker learners to construct a strong learner. Each candidate is selected based on its influence on the strong learner locally, in some neighbourhood of the observed data.

Fearnhead and Prangle (2012) have adopted a different approach in order to construct a vector of summary statistics almost from scratch. When the aim of ABC is estimating θ, their theoretical results show that the posterior mean of θ is the best summary statistics. The proposed scheme relies then on a first run of the ABC algorithm to derive a predictor $\hat{\eta}(x)$ of this posterior mean, depending on the (simulated or observed) data x trough a large collection of summary statistics. In the same vein, Prangle et al. (2013) focus on ABC model choice, proving that the vector of posterior probabilities of each model is indeed a sufficient statistic for the model choice problem.

Barthelmé and Chopin (2012) show that Expectation Propagation approximation inference technique, can be adapted to the likelihood-free context. The resulting algorithm does not require summary statistics, is an order of magnitude faster than existing techniques.

4. Speeding up the computations

The acceptation-rejection algorithm given above might be very time consuming. But, because of the independence of each proposed particle, the computations can be easily divided to be carried out by many CPU, *i.e.*, the algorithm is embarrassingly parallel. In contrast, the sequential schemes presented below are not that easy to parralelise on multi-core computers or clusters. See Marin et al. (2012b). But sequential schemes get their efficiency from their ability to sample the parameter space in a clever way, avoiding areas of low posterior probability.

4.1. **MCMC-ABC**

Indeed, when seeing the ABC acceptation-rejection algorithm as a sampler from(4), drawing particles from the prior distribution $\pi(\theta)$ is inefficient, in particular when the target is much more concentrated that the prior. The prior indeed leads to proposed values in low posterior probability areas of the parameter space. To answer this problem, Marjoram et al. (2003) have introduced an MCMC-ABC algorithm targeting (4). The update of the current state (θ_t, s_t) is based on a proposal (θ^*, s^*), where

$$\theta^* \sim q(\theta_t, \cdot), \quad x^* \sim f(x \,|\, \theta^*) \quad \text{and} \quad s^* = \eta(x^*).$$

The proposal is accepted with accepted with probability

$$1 \wedge \frac{\pi(\theta^*)q(\theta^*, \theta_t)}{\pi(\theta_t)q(\theta_t, \theta^*)}\mathbf{1}\{\delta[s^*, s_{\text{obs}}] \leq \varepsilon\}.$$

Hence the equilibrium distribution of the chain is

$$\pi\big(\theta,\, s \,\big|\, \delta[s, s_{\mathrm{obs}}] \leq \varepsilon\big).$$

The major drawback of the MCMC-ABC algorithm is than the threshold ε should be given in advance. If the chosen value of ε is too large, the approximation error between (4) and (3) in not negligible anymore. But if the chosen value of ε is too small, the Markov chain has poor mixing properties. The chain is indeed a Metropolis-Hastings chain accepting the proposed value of θ' if one new simulation $s = \eta(x)$ from $f(x\,|\,\theta')$ falls into the ε-neighbourhood of the observation s_{obs}.

4.2. **Sequential improvements**

In order to bypass both drawbacks of the MCMC-ABC algorithm, several population techniques have been proposed by Sisson et al. (2007), Sisson et al. (2009), Beaumont et al. (2009), Toni et al. (2009), Drovandi and Pettitt (2011), Del Moral et al. (2012) and Sedki et al. (2012), based on a decreasing sequence of thresholds $\varepsilon_1 > \varepsilon_2 > \cdots > \varepsilon_T$. We refer the reader to the review of Marin et al. (2012a) for a description of the main sequential algorithms. But we shall note here that the algorithm of Sedki et al. (2012) is a self-calibrated version of the likelihood-free SMC sampler of Del Moral et al. (2006). At each iteration, a new threshold ε_t is tuned to achieve a trade-off between good mixing properties of the MCMC and a small difference between (4) and (3).

4.3. **Parallel tempering**

Baragatti et al. (2012) have proposed to use a ABC-parallel tempering (ABC-PT) scheme to answer to the major weakness of MCMC-ABC, namely poor mixing properties of the Markov chain when the threshold ε is small — and it should be small so that (4) approximates (3). Their algorithm runs several MCMC-ABC chains in parallel, each of them corresponding to a given value of the threshold. And they allow fair swaps between two chains as in standard parallel tempering. If the thresholds of the K chains are $\varepsilon_1 < \cdots < \varepsilon_K$, then the i-th Markov chain has marginal equilibrium distribution

$$\pi\big(\theta, s \,\big|\, \delta(\eta(x), s_{\mathrm{obs}}) \leq \varepsilon\big).$$

And the equilibrium law of the joint chain is the tensor product of these marginals. To achieve this goal, after each internal update of the K chains, a pair of chains, $i < j$ say, is chosen uniformly. A swap between the i-th and the j-th chain is then accepted if $d(s_j, s_{\mathrm{obs}}) \leq \varepsilon_i$. This exchange move satisfies ABC requirements as there is no need to compute any likelihood, and moreover, the acceptance rate of the exchange depends only the tolerance thresholds. Though not adaptive in the calibration of the thresholds, the authors rightly recommend ABC-PT when the posterior is multi-modal.

5. Conclusion

Approximate Bayesian computation has proved itself useful to handle intractable likelihood in a quite general context, requiring only one's ability to perform simulation from the model. This explains its popularity in various fields such as population genetics, latent diffusion models, epidemiology... The literature on ABC is increasing rapidly. Because of the non-linear projection with the summary statistics, ABC can also handle large data sets. In comparable settings, Laplace approximations or variational Bayes solutions have also been advanced. The old Laplace approximation, see, e.g. Tierney and Kadane (1986), is based on some analytic computations that cannot be conducted in all situations. Meanwhile, the variational Bayes schemes have often few theoretical supports. We can also mention that Bayesian computation via empirical likelihood Mengersen et al. (2013) is a clever scheme to handle large data sets on which the "true" posterior distribution will never be known.

Acknowledgments.

References

Aeschbacher, S., Beaumont, M. A., and Futschik, A. (2012). A novel approach for choosing summary statistics in approximate Bayesian computation. *Genetics*, 192(3):1027–1047.

Baragatti, M., Grimaud, A., and Pommeret, D. (2012). Likelihood-free parallel tempering. *Statistics and Computing*, pages 1–15.

Barthelmé, S. and Chopin, N. (2011) ABC-EP: Expectation Propagation for likelihood-free Bayesian computation. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 289–296. ACM.

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

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

Biau, G., Cérou, F., and Guyader, A. (2012). New insights into approximate bayesian computation. Technical report, arXiv preprint arXiv:1207.6461.

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

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

Blum, M., Nunes, M., Prangle, D., and Sisson, S. (2012). A comparative review of dimension reduction methods in approximate Bayesian computation. *Statistical Science*, 28:189–208.

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

Del Moral, P., Doucet, A., and Jasra, A. (2012). An adaptive sequential monte carlo method for approximate bayesian computation. *Statistics and Computing*, 22(5):1009–1020.

Drovandi, C. C. and Pettitt, A. N. (2011). Estimation of parameters for macroparasite population evolution using approximate bayesian computation. *Biometrics*, 67(1):225–233.

Everitt, R. G. (2012). Bayesian parameter estimation for latent Markov random fields and social networks. *Journal of Computational and Graphical Statistics*, 21(4):940–960.

Fearnhead, P. and Prangle, D. (2012). Semi-automatic Approximate Bayesian Computation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(3):419–474.

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

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

Marin, J.-M., Pillai, N., Robert, C. P., and Rousseau, J. (2013). Relevant statistics for Bayesian model choice. Technical report, arXiv preprint arXiv:1110.4700v3.

Marin, J.-M., Pudlo, P., Robert, C. P., and Ryder, R. J. (2012a). Approximate bayesian computational methods. *Statistics and Computing*, 22(6):1167–1180.

Marin, J.-M., Pudlo, P., and Sedki, M. (2012b). Optimal parallelization of a sequential approximate Bayesian computation algorithm. In *Simulation Conference (WSC), Proceedings of the 2012 Winter*, pages 1–7. IEEE.

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

McKinley, T., Cook, A., and Deardon, R. (2009). Inference in epidemic models without likelihoods. *The International Journal of Biostatistics*, 5(1):24.

Mengersen, K. L., Pudlo, P., and Robert, C. P. (2013). Bayesian computation via empirical likelihood. *Proc. Natl. Acad. Sci. USA*, 110(4):1321–1326.

Prangle, D., Fearnhead, P., Cox, M. P., Biggs, P. J., and French, N. P. (2013). Semi-automatic selection of summary statistics for ABC model choice. Technical report, arXiv preprint arXiv:1302.5624.

Ratmann, O., Camacho, A., Meijer, A., and Donker, G. (2013). Statistical modelling of summary values leads to accurate approximate bayesian computations. Technical report, arXiv preprint arXiv:1305.4283.

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

Rosenblatt, M. (1969). Conditional probability density and regression estimators. *Multivariate analysis II*, 25:31.

Sedki, M., Pudlo, P., Marin, J.-M., Robert, C. P., and Cornuet, J.-M. (2012). Efficient learning in ABC algorithms. Technical report, arXiv preprint arXiv:1210.1388v2.

Sedki, M. A. and Pudlo, P. (2012). Discussion of D. Fearnhead and D. Prangle's Read Paper "Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation". *J. Roy. Statist. Soc. Ser. B*, 74(3):466–467.

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

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

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

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

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

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