

On Maximum Intractable Likelihood Estimation

F. J. Rubio and Adam M. Johansen

{F.J.Rubio,A.M.Johansen}@warwick.ac.uk

University of Warwick, Department of Statistics, Coventry, CV4 7AL, UK.

March 8, 2012

Abstract

Approximate Bayesian Computation (ABC) may be viewed as an analytic approximation of an intractable likelihood coupled with an elementary simulation step. Considering the first step as an explicit approximation of the likelihood allows, also, maximum-likelihood (or maximum-a-posteriori) inference to be conducted, approximately, using essentially the same techniques. Such an approach is developed here and the convergence of this class of algorithms is characterised theoretically. The use of non-sufficient summary statistics is considered. Applying the proposed method to three problems demonstrates good performance. The proposed approach provides an alternative for approximating the maximum likelihood estimator (MLE) in complex scenarios.

Keywords: Approximate Bayesian Computation; Density Estimation; Maximum Likelihood Estimation; Monte Carlo Methods.

1 Introduction

Modern applied statistics must deal with many settings in which the pointwise evaluation of the likelihood function, even up to a normalising constant, is impossible or computationally infeasible. Areas such as financial modelling, genetics, geostatistics and stochastic dynamical systems provide numerous examples of this (see e.g. Pritchard et al. (1999) and Toni et al. (2009)). It is consequently difficult to perform any inference (classical or Bayesian) about the parameters of the model. Various approaches to overcome this difficulty have been proposed. For instance, Composite Likelihood methods (Cox and Reid, 2004), for approximating the likelihood function, and Approximate Bayesian Computational methods (ABC; Pritchard et al., 1999; Beaumont et al., 2002), for approximating the posterior distribution, have been extensively studied in the statistical literature. Here, we study the use of ABC methods, under an appropriate choice of instrumental prior distribution, to approximate the maximum likelihood estimator.

It is well-known that ABC produces a sample approximation of the posterior distribution (Beaumont et al., 2002) in which there exists a genuine approximation error in addition to Monte Carlo variability. Using this sample approximation to characterise the mode of the posterior would in principle allow

(approximate) maximum *a posteriori* (MAP) estimation. Furthermore, using a uniform prior distribution over any set which contains the MLE will lead to a MAP estimate which coincides with the MLE. In low-dimensional problems if we have a sample from the posterior distribution of the parameters, we can estimate its mode by using nonparametric estimators of the density. Therefore, in contexts where the likelihood function is intractable we can use these results to obtain an approximation of the MLE. We will denote the estimator obtained with this approximation AMLE.

The quality of the approximation to the posterior and theoretical properties of the estimators obtained with ABC have been studied in Wilkinson (2008); Marin et al. (2011); Dean et al. (2011) and Fearnhead and Prangle (2012). The use of ABC posterior samples for conducting model comparison was studied in Didelot et al. (2011) and Robert et al. (2011).

Although Marjoram et al. (2003) noted that “It [ABC] can also be used in frequentist applications, in particular for maximum-likelihood estimation” this idea does not seem to have been developed. A method based around maximisation of a non-parametric estimate of the log likelihood function was proposed by Diggle and Gratton (1984) in the particular case of simple random samples; their approach involved sampling numerous replicates of the data for each parameter value and estimating the density in the data space. de Valpine (2004) proposes an importance sampling technique, rather closer in spirit to the approach developed here, by which a smoothed kernel estimation of the likelihood function up to a proportionality constant can be obtained in the particular case of state space models provided that techniques for sampling from the joint distribution of unknown parameters and latent states are available — not a requirement of the more general ABC technique developed below. The same idea was applied and analysed in the context of the estimation of location parameters, with particular emphasis on symmetric distributions, by Jaki and West (2008). The particular case of online parameter estimation in hidden Markov models was also investigated by Dean et al. (2011), who relied upon the specific structure of that problem. To the best of our knowledge neither MAP estimation nor maximum likelihood estimation in general, implemented directly via the “ABC approximation” combined with maximisation of an estimated density, have been studied in the literature.

The estimation of the mode of nonparametric kernel density estimators which may seem, at first, to be a hopeless task has also received a lot of attention (see e.g. Parzen, 1962; Konakov, 1973; Romano, 1988; Abraham et al., 2003; Bickel and Frürwirth, 2006). Alternative nonparametric density estimators which could also be considered within the AMLE context have been proposed recently in Cule et al. (2010); Jing et al. (2012).

In Section 2, we present a brief description of ABC methods. In Section 3 we describe how to use these methods to approximate the MLE and present theoretical results to justify such use of ABC methods. In Section 4, we present simulated and real examples to illustrate the use of the proposed MLE approximation. Section 5 concludes with a discussion of both the developed techniques and the likelihood approximation obtained via ABC in general.

2 Approximate Bayesian Computation

We assume throughout this and the following section that all distributions of interest admit densities with respect to an appropriate version of Lebesgue measure, wherever this is possible, although this assumption can easily be relaxed. Let $\mathbf{x} = (x_1, \dots, x_n)$ be a sample from a distribution $f(\cdot|\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^d$; $\mathcal{L}(\boldsymbol{\theta}; \mathbf{x})$ be the corresponding likelihood function, $\pi(\boldsymbol{\theta})$ be a prior distribution over the parameter $\boldsymbol{\theta}$ and $\pi(\boldsymbol{\theta}|\mathbf{x})$ the corresponding posterior distribution. Consider the following approximation to the posterior

$$\widehat{\pi}_\varepsilon(\boldsymbol{\theta}|\mathbf{x}) = \frac{\widehat{f}_\varepsilon(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int_{\Theta} \widehat{f}_\varepsilon(\mathbf{x}|\mathbf{t})\pi(\mathbf{t})d\mathbf{t}}, \quad (1)$$

where

$$\widehat{f}_\varepsilon(\mathbf{x}|\boldsymbol{\theta}) = \int_{\mathbb{R}^n} K_\varepsilon(\mathbf{x}|\mathbf{y})f(\mathbf{y}|\boldsymbol{\theta})d\mathbf{y}, \quad (2)$$

is an approximation of the likelihood function and $K_\varepsilon(\mathbf{x}|\mathbf{y})$ is a normalised Markov kernel. It's clear that (2) is a smoothed version of the true likelihood and it has been argued that the maximisation of such an approximation can in some circumstances lead to better performance than the maximisation of the likelihood itself (Ionides, 2005), providing an additional motivation for the investigation of MLE via this approximation. The approximation can be further motivated by noting that under weak regularity conditions, the distribution $\widehat{\pi}_\varepsilon(\boldsymbol{\theta}|\mathbf{x})$ is close (in some sense) to the true posterior $\pi(\boldsymbol{\theta}|\mathbf{x})$ when ε is sufficiently small. The simplest approach to ABC samples directly from (1) by the rejection sampling approach presented in Algorithm 1.

Algorithm 1 The basic ABC algorithm.

- 1: Simulate $\boldsymbol{\theta}'$ from the prior distribution $\pi(\cdot)$.
 - 2: Generate \mathbf{y} from the model $f(\cdot|\boldsymbol{\theta}')$.
 - 3: Accept $\boldsymbol{\theta}'$ with probability $\propto K_\varepsilon(\mathbf{x}|\mathbf{y})$ otherwise return to step 1.
-

Now, let $\eta : \mathbb{R}^n \rightarrow \mathbb{R}^m$ be a summary statistic, $\rho : \mathbb{R}^m \rightarrow \mathbb{R}^+$ be a metric and $\varepsilon > 0$. The simplest ABC algorithm can be formulated in this way using the kernel

$$K_\varepsilon(\mathbf{x}|\mathbf{y}) \propto \begin{cases} 1 & \text{if } \rho(\eta(\mathbf{x}), \eta(\mathbf{y})) < \varepsilon, \\ 0 & \text{otherwise.} \end{cases} \quad (3)$$

The ABC rejection algorithm of Pritchard et al. (1999) can be obtained simply by setting $\eta(\mathbf{x}) = \mathbf{x}$. Several improvements to the ABC method have been proposed in order to increase the acceptance rate, see Beaumont et al. (2002), Marjoram et al. (2003) and Sisson (2007) for good surveys of these. An exhaustive summary of these developments falls outside the scope of the present paper.

3 Maximising Intractable Likelihoods

3.1 Algorithm

Point estimation of θ , by MLE and MAP estimation in particular, has been extensively studied (Lehmann and Casella, 1998). Recall that the MLE, $\hat{\theta}$, and the MAP estimator $\tilde{\theta}$ are the values of θ which maximise the likelihood or posterior density for the realised data.

These two quantities coincide when the prior distribution is constant (e.g. a uniform prior $\pi(\theta)$ on a suitable (necessarily compact) set \mathbf{D} which contains $\hat{\theta}$). Therefore, if we use a suitable uniform prior, it is possible to approximate the MLE by using ABC methods to generate an approximate sample from the posterior and then approximating the MAP using this sample. In a different context in which the likelihood can be evaluated pointwise, simulation-based MLEs which use a similar construction have been shown to perform well (see, e.g., Gaetan and Yao, 2003, Lele et al., 2007 and Johansen et al., 2008). In the present setting the optimisation step can be implemented by estimating the posterior density of θ using a nonparametric estimator (e.g. a kernel density estimator) and then maximising this function: Algorithm 2.

Algorithm 2 The AMLE Algorithm

- 1: Obtain a sample $\theta_{m,\varepsilon}^* = (\theta_{m,\varepsilon,1}^*, \dots, \theta_{m,\varepsilon,m}^*)$ from $\hat{\pi}_\varepsilon(\theta|\mathbf{x})$.
 - 2: Using the sample $\theta_{m,\varepsilon}^*$ construct a nonparametric estimator $\hat{\varphi}$ of the density $\hat{\pi}_\varepsilon(\theta|\mathbf{x})$.
 - 3: Calculate the maximum of $\hat{\varphi}$, $\tilde{\theta}_{m,\varepsilon}$. This is an approximation of the MLE $\hat{\theta}$.
-

Note that the first step of this algorithm can be implemented rather generally by using essentially

any algorithm which can be used in the standard ABC context. It is not necessary to obtain an iid sample from the distribution $\hat{\pi}_\varepsilon$: provided the sample is appropriate for approximating that distribution it can in principle be employed in the AMLE context (although correlation between samples obtained using MCMC techniques and importance weights and dependence arising from the use of SMC can complicate density estimation, it is not as problematic as might be expected (Sköld et al., 2003)).

A still more general algorithm could be implemented: using any prior which has mass in some neighbourhood of the MLE and maximising the product of the estimated likelihood and the reciprocal of this prior (assuming that the likelihood estimate has lighter tails than the prior, not an onerous condition when density estimation is performed) will also provide an estimate of the likelihood maximiser, an approach which was exploited by de Valpine (2004) (who provided also an analysis of the smoothing bias produced by this technique in their context). In the interests of parsimony we do not pursue this approach here, although we note that such an extension eliminates the requirement that a compact set containing a maximiser of the likelihood be identified in advance.

One obvious concern is that the approach could not be expected to work well when the parameter space is of high dimension: it is well known that density estimators in high-dimensional settings converge very slowly. Three things mitigate this problem in the present context:

- Many of the applications of ABC have been to problems with extremely complex likelihoods which have only a small number of parameters (such as the financial example considered below).
- When the parameter space is of high dimension one could employ composite likelihood techniques with low-dimensional components estimated via AMLE. Provided appropriate parameter subsets are selected, the loss of efficiency will not be too severe in many cases.
- In certain contexts, as discussed below Proposition 2, it may not be necessary to employ the density estimation step at all.

Finally, we note that direct maximisation of the smoothed likelihood approximation (2) can be interpreted as a pseudo-likelihood technique (Besag, 1975), with the Monte Carlo component of the AMLE algorithm providing an approximation to the maximum of this pseudo-likelihood.

3.2 Asymptotic Behaviour

In this section we provide some theoretical results which justify the approach presented in Section 3.1 under similar conditions to those used to motivate the standard ABC approach.

We begin by showing pointwise convergence of the posterior (and hence likelihood) approximation under reasonable regularity conditions.

Proposition 1 *Let $\mathbf{x} = (x_1, \dots, x_n)$ be a sample from a continuous distribution $f(\cdot|\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^d$; $\mathbf{D} \subset \mathbb{R}^d$ be a compact set that contains $\hat{\boldsymbol{\theta}}$, the MLE of $\boldsymbol{\theta}$; K_ε be the densities of a family of symmetric Markov kernels, indexed by $\varepsilon > 0$, which become increasingly concentrated as ε decreases such that*

$$\int_{\mathcal{B}_\varepsilon(\mathbf{x})} K_\varepsilon(\mathbf{x}|\mathbf{y})d\mathbf{y} = \int_{\mathcal{B}_\varepsilon(\mathbf{x})} K_\varepsilon(\mathbf{y}|\mathbf{x})d\mathbf{y} = 1, \quad \forall \varepsilon > 0.$$

where $\mathcal{B}_\varepsilon(\mathbf{x}) := \{\mathbf{z} : |\mathbf{z} - \mathbf{x}| \leq \varepsilon\}$. Suppose that

$$\sup_{(\mathbf{t}, \boldsymbol{\theta}) \in \mathcal{B}_\varepsilon(\mathbf{x}) \times \mathbf{D}} f(\mathbf{t}|\boldsymbol{\theta}) < \infty,$$

for some $\varepsilon > 0$, and employ a uniform prior on \mathbf{D} for $\boldsymbol{\theta}$. Then, for each $\boldsymbol{\theta} \in \mathbf{D}$

$$\lim_{\varepsilon \rightarrow 0} \hat{\pi}_\varepsilon(\boldsymbol{\theta}|\mathbf{x}) = \pi(\boldsymbol{\theta}|\mathbf{x}).$$

Proof. From the assumptions it follows that

$$\hat{f}_\varepsilon(\mathbf{x}|\boldsymbol{\theta}) = \int_{\mathcal{B}_\varepsilon(\mathbf{x})} K_\varepsilon(\mathbf{x}|\mathbf{y})f(\mathbf{y}|\boldsymbol{\theta})d\mathbf{y}.$$

Furthermore, for each $\boldsymbol{\theta} \in \mathbf{D}$

$$|\hat{f}_\varepsilon(\mathbf{x}|\boldsymbol{\theta}) - f(\mathbf{x}|\boldsymbol{\theta})| \leq \int_{\mathcal{B}_\varepsilon(\mathbf{x})} d\mathbf{y} K_\varepsilon(\mathbf{x}|\mathbf{y}) |f(\mathbf{y}|\boldsymbol{\theta}) - f(\mathbf{x}|\boldsymbol{\theta})| \leq \sup_{\mathbf{y} \in \mathcal{B}_\varepsilon(\mathbf{x})} |f(\mathbf{y}|\boldsymbol{\theta}) - f(\mathbf{x}|\boldsymbol{\theta})| \quad (4)$$

which converges to 0 as $\varepsilon \rightarrow 0$ by continuity. Therefore $\hat{f}_\varepsilon(\mathbf{x}|\boldsymbol{\theta}) \xrightarrow{\varepsilon \rightarrow 0} f(\mathbf{x}|\boldsymbol{\theta})$.

Now, by bounded convergence (noting that boundedness of $\hat{f}_\varepsilon(\mathbf{x}|\boldsymbol{\theta})$, for $\varepsilon < \varepsilon_0$, follows from that of f itself), we have that:

$$\lim_{\varepsilon \rightarrow 0} \int_{\mathbf{D}} \hat{f}_\varepsilon(\mathbf{x}|\boldsymbol{\theta}')\pi(\boldsymbol{\theta}')d\boldsymbol{\theta}' = \int_{\mathbf{D}} f(\mathbf{x}|\boldsymbol{\theta}')\pi(\boldsymbol{\theta}')d\boldsymbol{\theta}'. \quad (5)$$

The result follows by combining (4) and (5), whenever $\pi(\boldsymbol{\theta}|\mathbf{x})$ is itself well defined. ■

This result can be strengthened by noting that it is straightforward to obtain bounds on the error introduced at finite ε under Lipschitz continuity conditions. Unfortunately, such conditions are not typically

verifiable in problems of interest. The following result, in which we show that whenever a sufficient statistic is employed the ABC-rejection algorithm produces a sample from a distribution that converges pointwise to the posterior distribution, follows as a simple corollary to the previous proposition. However, we provide an explicit proof based on a slightly different argument in order to emphasize the role of sufficiency.

Corollary 1 *Let $\mathbf{x} = (x_1, \dots, x_n)$ be a sample from a distribution $f(\cdot|\boldsymbol{\theta})$ over \mathbb{R}^q , $\eta : \mathbb{R}^{n-q} \rightarrow \mathbb{R}^m$ be a sufficient statistic for $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^d$, $\rho : \mathbb{R}^m \times \mathbb{R}^m \rightarrow \mathbb{R}_+$ be a metric and suppose that the density $f^\eta(\cdot|\boldsymbol{\theta})$ is ρ -continuous. Let $\mathbf{D} \subset \mathbb{R}^d$ be a compact set, suppose that*

$$\sup_{(\mathbf{t}, \boldsymbol{\theta}) \in \mathcal{B}_\epsilon \times \mathbf{D}} f^\eta(\mathbf{t}|\boldsymbol{\theta}) < \infty,$$

where $\mathcal{B}_\epsilon = \{\mathbf{t} \in \mathbb{R}^m : \rho(\eta(\mathbf{x}), \mathbf{t}) < \epsilon\}$ for some $\epsilon > 0$ fixed and consider a uniform prior on \mathbf{D} for $\boldsymbol{\theta}$.

Then, for each $\boldsymbol{\theta} \in \mathbf{D}$ and the kernel (3)

$$\lim_{\epsilon \rightarrow 0} \hat{\pi}_\epsilon(\boldsymbol{\theta}|\mathbf{x}) = \pi(\boldsymbol{\theta}|\mathbf{x}).$$

Proof. Using the integral Mean Value Theorem we have that for $\boldsymbol{\theta} \in \mathbf{D}$ and any $\epsilon \in (0, \epsilon)$:

$$\hat{f}_\epsilon(\mathbf{x}|\boldsymbol{\theta}) \propto \int I(\rho(\eta(\mathbf{y}), \eta(\mathbf{x})) < \epsilon) f(\mathbf{y}|\boldsymbol{\theta}) d\mathbf{y} = \int_{\mathcal{B}_\epsilon} f^\eta(\eta'|\boldsymbol{\theta}) d\eta' = \lambda(\mathcal{B}_\epsilon) f^\eta(\xi(\boldsymbol{\theta}, \mathbf{x}, \epsilon)|\boldsymbol{\theta}),$$

for some $\xi(\boldsymbol{\theta}, \mathbf{x}, \epsilon) \in \mathcal{B}_\epsilon$, where λ is the Lebesgue measure. Then

$$\hat{\pi}_\epsilon(\boldsymbol{\theta}|\mathbf{x}) = \frac{f^\eta(\xi(\boldsymbol{\theta}, \mathbf{x}, \epsilon)|\boldsymbol{\theta}) \pi(\boldsymbol{\theta})}{\int_{\mathbf{D}} f^\eta(\xi(\boldsymbol{\theta}', \mathbf{x}, \epsilon)|\boldsymbol{\theta}') \pi(\boldsymbol{\theta}') d\boldsymbol{\theta}'}$$

As this holds for any sufficiently small $\epsilon > 0$, we have by ρ -continuity of $f^\eta(\cdot|\boldsymbol{\theta})$:

$$\lim_{\epsilon \rightarrow 0} f^\eta(\xi(\boldsymbol{\theta}, \mathbf{x}, \epsilon)|\boldsymbol{\theta}) = f^\eta(\eta(\mathbf{x})|\boldsymbol{\theta}). \quad (6)$$

Using the Dominated Convergence Theorem we have

$$\lim_{\epsilon \rightarrow 0} \int_{\mathbf{D}} f^\eta(\xi(\boldsymbol{\theta}', \mathbf{x}, \epsilon)|\boldsymbol{\theta}') \pi(\boldsymbol{\theta}') d\boldsymbol{\theta}' = \int_{\mathbf{D}} f^\eta(\eta(\mathbf{x})|\boldsymbol{\theta}') \pi(\boldsymbol{\theta}') d\boldsymbol{\theta}'. \quad (7)$$

By the Fisher-Neyman factorization Theorem we have that there exists a function $h : \mathbb{R}^{n-q} \rightarrow \mathbb{R}_+$ such

that

$$f(\mathbf{x}|\boldsymbol{\theta}) = h(\mathbf{x})f^\eta(\eta(\mathbf{x})|\boldsymbol{\theta}) \quad (8)$$

The result follows by combining (6), (7) and (8). \square

With only a slight strengthening of the conditions, Proposition 1 allows us to show convergence of the mode as $\varepsilon \rightarrow 0$ to that of the true likelihood. It is known that pointwise convergence together with equicontinuity on a compact set implies uniform convergence (Rudin, 1976; Whitney, 1991). Therefore, if in addition to the conditions in Proposition 1 we assume equicontinuity of $\hat{\pi}_\varepsilon(\cdot|\mathbf{x})$ on \mathbf{D} , a rather weak additional condition, then the convergence to $\pi(\cdot|\mathbf{x})$ is uniform and we have the following direct corollary to Proposition 1:

Corollary 2 *Let $\hat{\pi}_\varepsilon(\cdot|\mathbf{x})$ achieve its global maximum at $\tilde{\boldsymbol{\theta}}_\varepsilon$ for each $\varepsilon > 0$ and suppose that $\pi(\cdot|\mathbf{x})$ has unique maximiser $\tilde{\boldsymbol{\theta}}$. Under the conditions in Proposition 1; if $\hat{\pi}_\varepsilon(\cdot|\mathbf{x})$ is equicontinuous, then*

$$\lim_{\varepsilon \rightarrow 0} \hat{\pi}_\varepsilon(\tilde{\boldsymbol{\theta}}_\varepsilon|\mathbf{x}) = \pi(\tilde{\boldsymbol{\theta}}|\mathbf{x}).$$

Using these results we can show that for a sample $\boldsymbol{\theta}_{m,\varepsilon}^* = (\boldsymbol{\theta}_{m,\varepsilon,1}^*, \dots, \boldsymbol{\theta}_{m,\varepsilon,m}^*)$ from the distribution $\hat{\pi}_\varepsilon(\cdot|\mathbf{x})$ with mode at $\tilde{\boldsymbol{\theta}}_\varepsilon$ and an estimator $\tilde{\boldsymbol{\theta}}_{m,\varepsilon}$, based on $\boldsymbol{\theta}_{m,\varepsilon}^*$, of $\tilde{\boldsymbol{\theta}}_\varepsilon$, such that $\tilde{\boldsymbol{\theta}}_{m,\varepsilon} \rightarrow \tilde{\boldsymbol{\theta}}_\varepsilon$ almost surely when $m \rightarrow \infty$, we have that for any $\xi > 0$ there exists $\varepsilon > 0$ such that

$$\lim_{m \rightarrow \infty} \left| \hat{\pi}_\varepsilon(\tilde{\boldsymbol{\theta}}_{m,\varepsilon}|\mathbf{x}) - \pi(\tilde{\boldsymbol{\theta}}|\mathbf{x}) \right| \leq \xi, \quad \text{a.s.}$$

That is, in the case of a sufficiently well-behaved density estimation procedure, using the simple form of the ABC estimator (Algorithm 1) we have that for any level of precision ξ , the maximum of the AMLE approximation will, for large enough ABC samples, almost surely be ξ -close to the maximum of the posterior distribution of interest, which coincides with the MLE under the given conditions. A simple continuity argument suffices to justify the use of $\tilde{\boldsymbol{\theta}}_{m,\varepsilon}$ to approximate $\tilde{\boldsymbol{\theta}}$ for large m and small ε .

The convergence shown in the above results depends on the use of a sufficient statistic. In contexts where the likelihood is intractable, this may not be available. In the ABC literature, it has become common to employ summary statistics which are not sufficient in this setting. Although it is possible to characterise the likelihood approximation in this setting, it is difficult to draw useful conclusions from such a characterisation. The construction of appropriate summary statistics remains an active research

area (see e.g. Peters et al., 2010 and Fearnhead and Prangle, 2012).

We finally provide one result which provides some support for the use of certain non-sufficient statistics when there is a sufficient quantity of data available. In particular we appeal to the large-sample limit in which it can be seen that for a class of summary statistics the AMLE estimate can almost surely be made arbitrarily close to the true parameter value if a

sufficiently small value of ϵ can be used. This is, of course, an idealisation, but provides some guidance on the properties required for summary statistics to be suitable for this purpose and it provides some reassurance that the use of such statistics can in principle lead to good estimation performance. In this result we assume that the AMLE algorithm is applied with the summary statistics filling the role of the data and hence the ABC kernel is defined directly on the space of the summary statistics:

Proposition 2 *Let $\mathbf{x} = (x_1, x_2, \dots)$ denote a sample with joint measure $\mu(\cdot|\theta)$ for some $\theta \in \mathbf{D} \subset \Theta$. Let $\pi(\theta)$ denote a prior density over \mathbf{D} . Let $\eta_n(\mathbf{x}) = \eta_n(x_1, \dots, x_n)$ denote a sequence of d_η -dimensional summary statistics with distributions $\mu^{\eta_n}(\cdot|\theta)$. Allow η_n^* to denote an observed value of the sequence of statistics obtained from the model with $\theta = \theta^*$.*

Assume that:

- i. $\lim_{n \rightarrow \infty} \eta_n(\mathbf{x}) \stackrel{a.s.}{=} g(\theta)$ for π - a.e. θ
- ii. $g : \Theta \rightarrow \mathbb{R}^{d_\eta}$ is an injective mapping. Letting $H = g(\mathbf{D}) \subset \mathbb{R}^{d_\eta}$ denote the image of the feasible parameter space under g , $g^{-1} : H \rightarrow \Theta$ is an α -Lipschitz continuous function.
- iii. The ABC kernel, $K_\epsilon^\eta(\cdot|\eta')$ is concentrated within a ball of radius ϵ for all ϵ , i.e. $\text{supp } K_\epsilon(\cdot|\eta') \subseteq \mathcal{B}_\epsilon(\eta')$
- iv. The nonparametric estimator used always provides an estimate of the mode which lies within the convex hull of the sample.

Then:

- (a) $\text{supp } \lim_{n \rightarrow \infty} \pi_\epsilon(\theta|\eta_n^*) \subseteq \mathcal{B}_{\alpha\epsilon}(\theta^*)$ for $\mu(\cdot|\theta^*)$ -almost every η^* for π -almost every θ^* .
- (b) The AMLE estimate of the MLE lies within $\mathcal{B}_{\alpha\epsilon}(\theta^*)$ almost surely.

Proof. Allowing $f_\epsilon^{\eta_n}(\eta|\theta)$ to denote the ABC approximation of the density of η_n given θ , we have:

$$\lim_{n \rightarrow \infty} f_\epsilon^{\eta_n}(\eta|\theta) = \lim_{n \rightarrow \infty} \int \mu^{\eta_n}(d\eta'|\theta) K_\epsilon(\eta|\eta') \stackrel{a.s.}{=} \int \delta_{g(\theta)}(d\eta') K_\epsilon(\eta|\eta') = K_\epsilon(\eta|g(\theta))$$

where with the final equality following from assumption i (noting that almost sure convergence of η_n to $g(\theta)$ implies convergence in distribution of η_n to a degenerate random variable taking the value $g(\theta)$).

From which it is clear that $\text{supp } \lim_{n \rightarrow \infty} f_\varepsilon^{\eta_n}(\cdot|\theta) \subseteq \mathcal{B}_\varepsilon(g(\theta))$ by assumption iii.

And the ABC approximation to the posterior density of θ , $\lim_{n \rightarrow \infty} \pi_\varepsilon(\cdot|\eta_n)$, may be similarly constrained:

$$\lim_{n \rightarrow \infty} \pi_\varepsilon(\theta|\eta_n) > 0 \Rightarrow \lim_{n \rightarrow \infty} \|\eta_n - g(\theta)\| \stackrel{a.s.}{\leq} \varepsilon \Rightarrow \lim_{n \rightarrow \infty} \|g^{-1}(\eta_n) - \theta\| \stackrel{a.s.}{\leq} \alpha\varepsilon$$

using assumption ii. And by assumption i, ii and the continuous mapping theorem we have that $g^{-1}(\eta_n^*) \xrightarrow{a.s.} \theta^*$ giving result (a); result (b) follows immediately from assumption iv. ■

Some interpretation of the assumptions of this proposition seems appropriate. The first tells us simply that the summary statistics converge to some function of the parameters in the large sample limit, a mild requirement which is clearly necessary to allow recovery of the parameters from the statistics. The second condition strengthens this slightly, requiring that the limiting values of the statistics and parameters exist in one-to-one correspondence and that this correspondence is regular in a Lipschitz-sense. The remaining conditions simply characterise the behaviour of the ABC approximation and the AMLE algorithm.

It is also noteworthy that this proposition suggests that, at least in the large sample limit, one can use any estimate of the mode which lies within the convex hull of the sampled parameter values. The posterior mean would satisfy this requirement and thus for large enough data sets it is not necessary to employ the nonparametric density estimator at all in order to implement AMLE. This is perhaps an unsurprising result and seems a natural consequence of the usual Bayesian consistency results but it does have implications for implementation of AMLE in settings with large amounts of data for which the summary statistics are with high probability close to their limiting values.

Example 1 (Location-Scale Families and Empirical Quantiles) Consider a simple random sample from a location-scale family, in which we can write the distribution functions in the form:

$$F(x_i|\mu, \sigma) = F_0((x_i - \mu)/\sigma)$$

Allow $\eta_n^1 = \hat{F}^{-1}(q_1)$ and $\eta_n^2 = \hat{F}^{-1}(q_2)$ to denote to empirical quantiles. By the Glivenko-Cantelli

theorem, these empirical quantiles converge almost-surely to the true quantiles:

$$\lim_{n \rightarrow \infty} \begin{pmatrix} \eta_n^1 \\ \eta_n^2 \end{pmatrix} \stackrel{a.s.}{=} \begin{pmatrix} F^{-1}(q_1 | \mu, \sigma) \\ F^{-1}(q_2 | \mu, \sigma) \end{pmatrix}$$

In the case of the location-scale family, we have that:

$$F^{-1}(q^i | \mu, \sigma) = \sigma F_0^{-1}(q^i) + \mu$$

and we can find explicitly the mapping g^{-1} :

$$g^{-1}(\eta_n^1, \eta_n^2) = \begin{pmatrix} \frac{\eta_n^1 - \eta_n^2}{F_0^{-1}(q_1) - F_0^{-1}(q_2)} \\ \eta_n^1 - \frac{\eta_n^1 - \eta_n^2}{F_0^{-1}(q_1) - F_0^{-1}(q_2)} F_0^{-1}(q_1) \end{pmatrix} \stackrel{a.s.}{\rightarrow} \begin{pmatrix} \sigma \\ \mu \end{pmatrix}$$

provided that $F_0^{-1}(q_1) \neq F_0^{-1}(q_2)$ which can be assured if F_0 is strictly increasing and $q_1 \neq q_2$. In this case we even obtain an explicit form for α .

3.3 Use of kernel density estimators

In this section we demonstrate that the simple Parzen estimator can be employed within the AMLE context with the support of the results of the previous section.

Definition 1 (Parzen, 1962) Consider the problem of estimating a density with support on \mathbb{R}^n from m independent random vectors $(\mathbf{Z}_1, \dots, \mathbf{Z}_m)$. Let K be a kernel, h_m be a bandwidth such that $h_m \rightarrow 0$ when $m \rightarrow \infty$, then a kernel density estimator is defined by

$$\hat{\varphi}_m(\mathbf{z}) = \frac{1}{mh_m^n} \sum_{j=1}^m K\left(\frac{\mathbf{z} - \mathbf{Z}_j}{h_m}\right).$$

Under the conditions $h_m \rightarrow 0$ and $mh_m^n / \log(m) \rightarrow \infty$ together with Theorem 1 from Abraham et al. (2003), we have that $\tilde{\theta}_m \xrightarrow{a.s.} \tilde{\theta}$ as $m \rightarrow \infty$. Therefore, the results presented in the previous section apply to the use of kernel density estimation. This demonstrates that this simple non-parametric estimator is adequate for approximation of the MLE via the AMLE strategy, at least asymptotically.

This is, of course, just one of many ways in which the density could be estimated and more sophisticated techniques could very easily be employed and justified in the AMLE context.

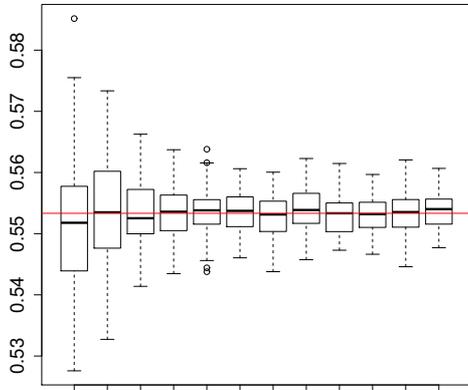
4 Examples

We present two toy examples which illustrate the performance of the algorithm in simple scenarios in which the solution is known before comparing the proposed algorithm to simple ABC in the context of estimation of the parameters of an α -stable distribution. In all three examples the simple ABC rejection algorithm was used, together with ABC kernel (3). For the second and third example, kernel density estimation is conducted using the R command ‘kde’ together with the bandwidth matrix obtained via the smoothed cross validation approach of Duong and Hazelton (2005) using the command ‘Hscv’ from the R package ‘ks’ (Duong, 2011). R source code for these examples is available from the first author upon request.

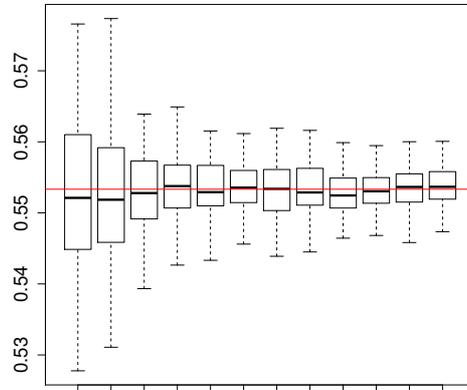
4.1 Binomial Model

Consider a sample of size 30 simulated from a Binomial(10, 0.5) with $\bar{x} = 5.53$. Using the prior $\theta \sim \text{Unif}(0, 1)$, a tolerance $\varepsilon = 0.1$, a sufficient statistic $\eta(\mathbf{x}) = \bar{x}$ and the Euclidean metric we simulate an ABC sample of size 10,000 which, together with Gaussian kernel estimation of the posterior, gives the AMLE estimator $\tilde{\theta} = 0.552$.

There are three quantities affecting the precision in the estimation of $\hat{\theta}$: D , n and ε . Figure 1 illustrates the effect of varying $n \in \{30, 100, 1000, 2000, \dots, 10000\}$ for a fixed ε , two different choices of D and an ABC sample of size 10,000. Boxplots were obtained using 100 replications of the (stochastic) AMLE algorithm. It is apparent that D does not much affect the precision of the estimation, but the choice $D = (0.45, 0.65)$ produces an acceptance rate about 5 times greater than the choice $D = (0, 1)$. Figure 2 shows the effect of $\varepsilon \in \{1, 0.9, \dots, 0.1, 0.05, 0.01\}$ for a fixed n and two different choices of D . In this case we can note that the effect of ε on the precision is significant. Again, the choice of D seems to affect only the acceptance rate.

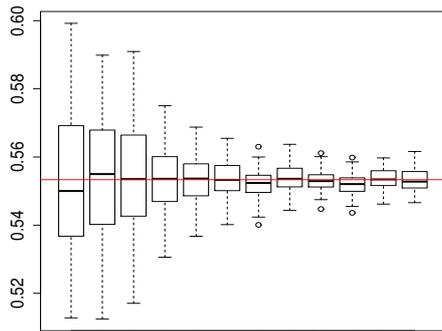


(a) $D = (0.45, 0.65)$

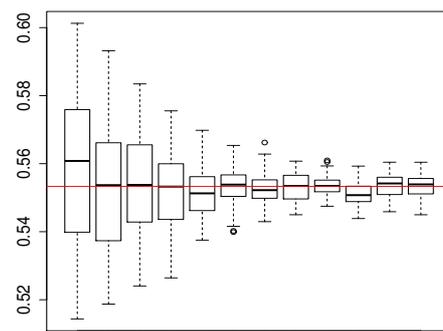


(b) $D = (0, 1)$

Figure 1: Effect of $n \in \{30, 100, 1000, 2000, \dots, 10000\}$ for $\varepsilon = 0.05$. The continuous red line represents the true MLE value.



(a)



(b)

Figure 2: Effect of $\varepsilon \in \{1, 0.9, \dots, 0.1, 0.05, 0.01\}$ for $n = 10000$: (a) $D = (0.45, 0.65)$; (b) $D = (0, 1)$. The continuous red line represents the true MLE value

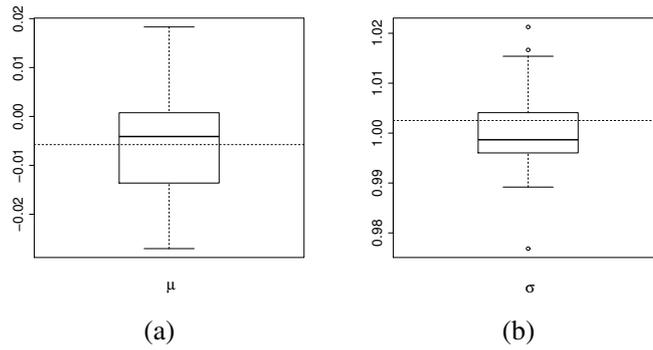


Figure 3: Monte Carlo variability of the AMLE: (a) μ ; (b) σ . The dashed lines represent the true MLE value.

4.2 Normal Model

Consider a sample of size 100 simulated from a $\text{Normal}(0, 1)$ with sample mean $\bar{x} = -0.005$ and sample variance $s^2 = 1.004$. Suppose that both parameters (μ, σ) are unknown. The MLE of (μ, σ) is simply $(\hat{\mu}, \hat{\sigma}) = (-0.005, 1.002)$.

Consider the priors $\mu \sim \text{Unif}(-0.25, 0.25)$ and $\sigma \sim \text{Unif}(0.75, 1.25)$, a tolerance $\varepsilon = 0.01$, a sufficient statistic $\eta(\mathbf{x}) = (\bar{x}, s)$, the Euclidean metric, an ABC sample of size 5,000, and Gaussian kernel estimation of the posterior. Figure 3 illustrates Monte Carlo variability of the AMLE of (μ, σ) . Boxplots were obtained using 50 replicates of the algorithm.

4.3 Stable distributions

Logarithmic daily return prices are typically modelled using Lévy processes. For this reason, it is necessary to model the increments (logarithmic returns) using an infinitely divisible distribution. It has been found empirically that these observations have tails heavier than those of the normal distribution, and therefore an attractive option is the use of the 4-parameter $(\alpha, \beta, \mu, \sigma)$ α -stable family of distributions, which can account for this behaviour. It is well known that maximum likelihood estimation for this family of distributions is difficult. Various numerical approximations of the MLE have been proposed (see e.g. McCulloch, 1986; Nolan, 2001). From a Bayesian perspective, Peters et al. (2010) proposed the use of ABC methods to obtain an approximate posterior sample of the parameters. They propose six summary statistics that can be used for this purpose.

Here, we analyse the logarithmic daily returns using the closing price of IBM ordinary stock from Jan. 1 2009 to Jan. 1 2012. Figure 4 shows the corresponding histogram. For this data set, the MLE using McCulloch's quantile method implemented in the R package 'fBasics' (Wuertz et al., 2010) is

$$(\hat{\alpha}, \hat{\beta}, \hat{\mu}, \hat{\sigma}) = (1.4930, -0.0780, -0.0007, 0.0073).$$

Given the symmetry observed and in the spirit of parsimony, we consider the skewness parameter β to be 0 in order to calculate the AMLE of the parameters (α, μ, σ) . Based on the interpretation of these parameters (shape, location and scale) and the data we use the priors

$$\alpha \sim U(1, 2), \quad \mu \sim U(-0.1, 0.1), \quad \sigma \sim U(0.0035, 0.0125).$$

For the (non-sufficient) summary statistic we use the proposal S_4 in Peters et al. (2010), which consists of the values of the empirical characteristic function evaluated on an appropriate grid. We use the grid $t \in \{-250, -200, -100, -50, -10, 10, 50, 100, 200, 250\}$, an ABC sample of size 5,000, a tolerance $\varepsilon = 0.065$ and Gaussian kernel density estimation. Figure 5 illustrates Monte Carlo variability of the AMLE of (α, μ, σ) . Boxplots were obtained using 25 replicates of the AMLE procedure. A simulation study considering several simulated data sets produced with common parameter values (results not shown) suggest that the sampling variability in McCulloch's estimator exceeds the difference between that estimator and the AMLE estimator based upon S_4 . In general, considerable care must of course be taken in the selection of statistics — it is noteworthy that the quantiles used in McCulloch's own estimator satisfy most of the requirements of Proposition 2, although it is not clear that it is possible to demonstrate the Lipschitz-continuity of g^{-1} in this case.

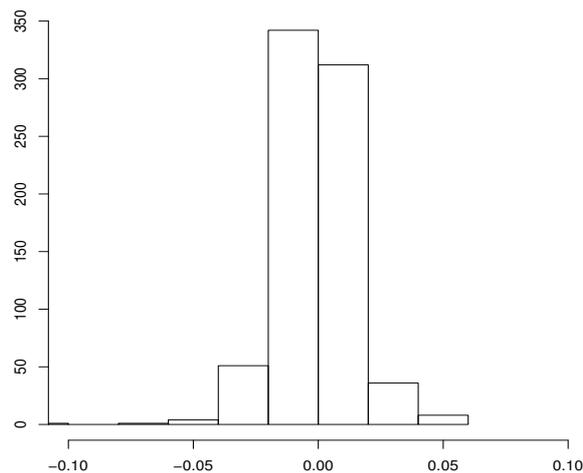


Figure 4: Logarithmic daily returns using the closing price of IBM ordinary stock Jan. 1 2009 to Jan. 1 2012.

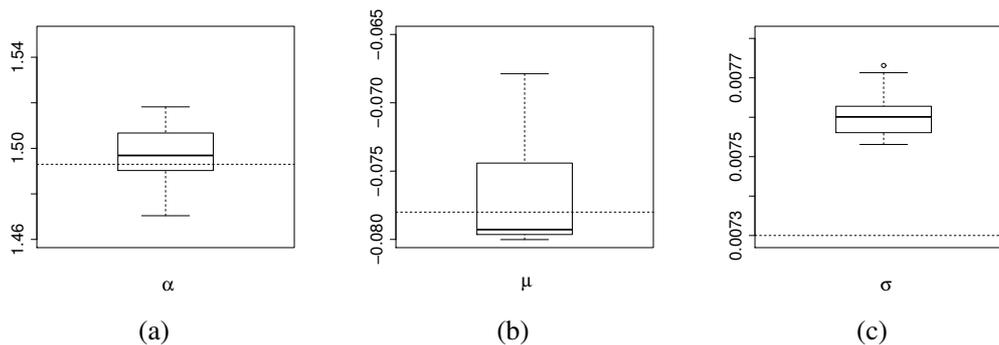


Figure 5: Monte Carlo variability of the AMLE: (a) α ; (b) μ ; (c) σ . Dashed lines represent McCulloch’s estimator produced by the R package ‘fBasics’.

5 Discussion

This paper presents a simple algorithm for conducting maximum likelihood estimation via simulation in settings in which the likelihood cannot (readily) be evaluated and provides theoretical and empirical support for that algorithm. This adds another tool to the “approximate computation” toolbox. This allows the (approximate) use of the MLE in most settings in which ABC is possible: desirable both in itself and because it is unsatisfactory for the approach to inference to be dictated by computational considerations. Furthermore, even in settings in which one wishes to adopt a Bayesian approach to inference it may be interesting to obtain also a likelihood-based estimate as agreement or disagreement between the approaches. Naturally, both ABC and AMLE being based upon the same approximation, the difficulties and limitations of ABC are largely inherited by AMLE. Selection of statistics in the case in which sufficient statistics are not available remains a critical question. There has been considerable work on this topic in recent years (see e.g. Fearnhead and Prangle, 2012).

A side-effect of the AMLE algorithm is an approximate characterisation of the likelihood surface, or in Bayesian settings of the posterior surface. We would strongly recommend that this surface be inspected whenever ABC or related techniques are used as even in settings in which the original likelihood contains strong information about the parameters it is possible for a poor choice of summary statistic to lead to the loss of this information. Without explicit consideration of the approximation, perhaps combined with prior sensitivity analysis, this type of issue is difficult to detect.

Acknowledgements

AMJ gratefully acknowledges support from EPSRC grant EP/I017984/1. FJR acknowledges support from Conacyt, México.

References

- Abraham, C., Biau, G. and Cadre, B. (2003). Simple estimation of the mode of a multivariate density. *The Canadian Journal of Statistics* 31: 23–34.
- Beaumont, M. A., Zhang, W. and Balding, D. J. (2002). Approximate Bayesian computation in population genetics. *Genetics* 162: 2025–2035.
- Besag, J. (1975). Statistical Analysis of Non-Lattice Data. *The Statistician* 24:179–195.
- Bickel, D. R. and Früwirth, R. (2006). On a fast, robust estimator of the mode: Comparisons to other robust estimators with applications. *Computational Statistics & Data Analysis* 50: 3500–3530.
- Cox, D. R. and Reid, N. (2004). A note on pseudolikelihood constructed from marginal densities. *Biometrika* 91: 729–737.
- Cule, M. L., Samworth, R. J. and Stewart, M. I. (2010). Maximum likelihood estimation of a multi-dimensional log-concave density. *Journal Royal Statistical Society B* 72: 545–600.
- Dean, T. A., Singh, S. S., Jasra A. and Peters G. W. (2011). Parameter estimation for hidden Markov models with intractable likelihoods. Arxiv preprint arXiv:1103.5399v1.
- de Valpine, P. (2004) Monte Carlo state space likelihoods by weighted posterior kernel density estimation. *Journal of the American Statistical Society* 99: 523–536.
- Didelot, X., Everitt, R. G., Johansen, A. M. and Lawson, D. J. (2011). Likelihood-free estimation of model evidence. *Bayesian Analysis* 6: 49–76.
- Diggle, P. J. and Gratton, R. J. (1984) Monte Carlo Methods of Inference for Implicit Statistical Models. *Journal of the Royal Statistical Society. Series B (Methodological)* 46:193–227.
- Duong, T. and Hazleton, M. L. (2005). Cross-validation bandwidth matrices for multivariate kernel density estimation. *Scandinavian Journal of Statistics* 32:485–506.
- Duong, T. (2011). *ks: Kernel smoothing*. R package version 1.8.5. <http://CRAN.R-project.org/package=ks>
- Fearnhead, P. and Prangle, D. (2012). Constructing Summary Statistics for Approximate Bayesian Computation: Semi-automatic ABC (with discussion). *Journal of the Royal Statistical Society Series B (Methodology) in press*.
- Gaetan, C. and Yao, J. F. (2003). A multiple-imputation Metropolis version of the EM algorithm. *Biometrika* 90: 643–654.
- Ionides, E. (2005). Maximum Smoothed Likelihood Estimation. *Statistica Sinica* 15: 1003–1014.
- Jaki, T. and West, R. W. (2008). Maximum Kernel Likelihood Estimation. *Journal of Computational and Graphical Statistics* 17: 976.
- Jing, J., Koch, I. and Naito, K (2012). Polynomial Histograms for Multivariate Density and Mode Estimation. *Scandinavian Journal of Statistics* 39:75–96.
- Johansen, A. M., Doucet, A. and Davy, M. (2008). Particle methods for maximum likelihood parameter estimation in latent variable models. *Statistics and Computing* 18: 47–57.

- Konakov, V. D. (1973). On asymptotic normality of the sample mode of multivariate distributions. *Theory of Probability and its Applications* 18: 836–842.
- Lehmann, E. and Casella, G. (1998). *Theory of Point Estimation* (revised edition). Springer-Verlag, New York.
- Lele, S. R., Dennis, B. and Lutscher, F. (2007). Data cloning: easy maximum likelihood estimation for complex ecological models using Bayesian Markov chain Monte Carlo methods. *Ecology Letters* 10: 551–563.
- McCulloch, J. H. (1986). Simple consistent estimators of stable distribution parameters. *Communications in Statistics: Simulation and Computation* 15: 1109–1136.
- Marin, J., Pudlo, P., Robert, C. P. and Ryder, R. (2011). Approximate Bayesian Computational methods. *Statistics and Computing in press*.
- Marjoram, P., Molitor, J., Plagnol, V. and Tavaré, S. (2003). Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences USA*: 15324–15328.
- Nolan, J. P. (2001). Maximum likelihood estimation and diagnostics for stable distributions. In: O.E. Barndorff-Nielsen, T. Mikosh, and S. Resnick, Eds., *Lévy Processes*, Birkhauser, Boston, 379–400.
- Parzen, E. (1962). On estimation of a probability density function and mode. *Annals of Mathematical Statistics* 33: 1065–1076.
- Peters, G. W., Sisson, S. A. and Fan, Y. (2010). Likelihood-free Bayesian inference for α -stable models. *Computational Statistics & Data Analysis in press*. <http://dx.doi.org/10.1016/j.csda.2010.10.004>
- Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A., and Feldman, M. T. (1999). Population Growth of Human Y Chromosomes: A Study of Y Chromosome Microsatellites. *Molecular Biology and Evolution* 16: 1791–1798.
- Robert, C. P. (2007). *The Bayesian Choice* (2nd ed.). New York: Springer.
- Robert, C. P., Cornuet, J., Marin, J. and Pillai, N. S. (2011). Lack of confidence in ABC model choice. *Proceedings of the National Academy of Sciences of the United States of America* 108: 15112–15117.
- Romano, J. P. (1988). On weak convergence and optimality of kernel density estimates of the mode. *The Annals of Statistics* 16: 629–647.
- Rudin, W. (1976). *Principles of Mathematical Analysis*. New York: McGraw-Hill.
- Sisson, S. A., Fan, Y., and Tanaka, M. M. (2007). Sequential Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 104: 1760–1765.
- Sköld, M. and Roberts, G. O. (2003). Density estimation for the Metropolis–Hastings algorithm. *Scandinavian Journal of Statistics* 30: 699–718.
- Tavaré, S., Balding, D., Griffith, R. and Donnelly, P. (1997). Inferring coalescence times from DNA sequence data. *Genetics* 145: 505–518.
- Toni, T.; Welch, D.; Strelkowa, N.; Ipsen, A.; Stumpf, M.P.H. (2009). Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface* 6: 187–202.
- Whitney, K. N. (1991). Uniform Convergence in probability and stochastic equicontinuity. *Econometrica* 59: 1161–1167.
- Wilkinson, R. D. (2008). Approximate Bayesian computation (ABC) gives exact results under the assumption of error model. Arxiv preprint arXiv:0811.3355.
- Wuertz, D. and core team members R (2010). *fBasics: Rmetrics - Markets and Basic Statistics*. R package version 2110.79. <http://CRAN.R-project.org/package=fBasics>