Electronic Journal of Statistics
Vol. 7 (2013) 1632–1654
ISSN: 1935-7524
DOI: 10.1214/13-EJS819

A simple approach to maximum intractable likelihood estimation

F. J. Rubio* **and Adam M. Johansen†**

University of Warwick, Department of Statistics, Coventry, CV4 7AL, UK
e-mail: francisco.rubio@warwick.ac.uk; a.m.johansen@warwick.ac.uk

Abstract: Approximate Bayesian Computation (ABC) can be viewed as an analytic approximation of an intractable likelihood coupled with an elementary simulation step. Such a view, combined with a suitable instrumental prior distribution permits maximum-likelihood (or maximum-a-posteriori) inference to be conducted, approximately, using essentially the same techniques. An elementary approach to this problem which simply obtains a nonparametric approximation of the likelihood surface which is then maximised is developed here and the convergence of this class of algorithms is characterised theoretically. The use of non-sufficient summary statistics in this context is considered. Applying the proposed method to four problems demonstrates good performance. The proposed approach provides an alternative for approximating the maximum likelihood estimator (MLE) in complex scenarios.

Keywords and phrases: Approximate Bayesian Computation, density estimation, maximum likelihood estimation, Monte Carlo methods.
AMS 2000 subject classifications: 62E17, 62F10, 62F12, 62G07, 65C05.

Received January 2013.

Contents

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

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, neurophysiology and stochastic dynamical systems provide numerous examples of this (see e.g. Cox and Smith, 1954; 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 the 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 deterministic approximation error in addition to Monte Carlo variability. The quality of the approximation to the posterior and theoretical properties of the estimators obtained with ABC have been studied in Wilkinson (2008); Blum (2010); 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). 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, under the parameterisation of interest, over any set which contains the MLE will lead to a MAP estimate which coincides with the MLE. This is an immediate consequence of Bayes' Theorem. In low-dimensional problems if we have a sample from the posterior distribution of the parameters, we can estimate its mode by using either nonparametric estimators of the density or another mode–seeking technique such as the *mean-shift* algorithm (Fukunaga and Hostetler, 1975). 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.

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).

Bretó et al. (2009) proposed the *plug-and-play* technique which permits conducting likelihood-based inference, despite the complexity of the corresponding likelihood function, on time series models which allow for simulating realisations at any parameter values. The particular case of parameter estimation in hidden Markov models was also investigated by Dean et al. (2011), whose approach relies upon the specific structure of Markov models (essentially the standard particle filtering estimate of the likelihood for a modified model is employed) and an attempt to numerically optimise the likelihood using these Monte Carlo point estimates is made. They note that even in simple univariate models the simulation cost can be rather high. Their approximation (denoted ABC MLE) can be interpreted as a maximum likelihood estimation of a misspecified model. At the cost of some loss of efficiency the bias introduced by the use of finite tolerance can be eliminated by a *noisy ABC* (Fearnhead and Prangle, 2012) argument.

Another approach to estimation in intractable models is provided by the *indirect inference* approach of Gouriéroux et al. (1993), but this approach requires the introduction of an explicit proxy model and a relationship between the parameters of the original model and its proxy to be specified.

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. However, there has been a lot of interest in this type of problem using different approaches (Cox and Kartsonaki, 2012; Ehrlich et al., 2012; Fan et al., 2012; Mengersen et al., 2013; and Biau et al., 2012 who establish a number of results including one closely related to our Proposition 1 in settings in which a particular post-simulation approach to the specification of the tolerance parameter is adopted) since we completed the first version of this work (Rubio and Johansen, 2012).

The use of the mode of a nonparametric kernel density estimate to estimate the mode of a density, 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 Früwirth, 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).

The remainder of this paper is organised as follows. 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 $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^{q \times n}$ be a sample with joint distribution $f(\cdot|\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^d$; $\mathcal{L}(\boldsymbol{\theta}; \mathbf{x})$ 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

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

where

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

is an approximation of the likelihood function and $K_\varepsilon(\mathbf{x}|\mathbf{y})$ is a normalised Markov kernel. $K_\varepsilon(\cdot|\mathbf{y})$ is typically concentrated around $\mathbf{y}$ with ε acting as a scale parameter. It is 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.

Let $\rho : \mathbb{R}^{q \times n} \times \mathbb{R}^{q \times n} \to \mathbb{R}^+$ be a metric and $\varepsilon > 0$. The simplest ABC algorithm — the rejection algorithm of Pritchard et al. (1999) — can be formulated in this way using the kernel

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

In the literature, summary statistics are often used in place of the original data. Introducing such a statistic, $\boldsymbol{\eta} : \mathbb{R}^{q \times n} \to \mathbb{R}^m$, and defining the kernel on the space of these summary statistics allows these methods to be recovered within the same framework. When these statistics are not sufficient for the inferential task at hand there is an inevitable loss of efficiency (at best). Below

we provide some results which characterise the large sample behaviour and the small ε behaviour.

Several modifications to the ABC method have been proposed in order to improve the computational efficiency, see Beaumont et al. (2002), Marjoram et al. (2003) and Sisson et al. (2007) for examples 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 $\boldsymbol{\theta}$, by MLE and MAP estimation in particular, has been extensively studied (Lehmann and Casella, 1998). Recall that the MLE, $\widehat{\boldsymbol{\theta}}$, and the MAP estimator, $\tilde{\boldsymbol{\theta}}$, are the values of $\boldsymbol{\theta}$ 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 $\pi(\boldsymbol{\theta})$ on some set $\mathbf{D}$ which contains $\widehat{\boldsymbol{\theta}}$). Therefore, if we use a suitable uniform prior (which must be over a bounded set as we require a proper prior from which to sample) then 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 $\boldsymbol{\theta}$ using a nonparametric estimator (e.g. a kernel density estimator) and then maximising this function: Algorithm 2.

We have not here considered similar simulation-based approaches to the direct optimisation of the likelihood function due to the associated computational cost and also because the proposed method has the additional advantages that it fully characterises the likelihood surface and can be conducted concurrently with Bayesian analysis with no additional simulation effort.

Algorithm 2 The AMLE Algorithm

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

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 *i.i.d.* sample from the distribution $\widehat{\pi}_\varepsilon$: 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 used to obtain that estimate) 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, and throughout the remainder of this document we assume that a uniform prior over some set D which includes the MLE is used, 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:

(i) 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 examples considered below).
(ii) 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. Alternatively, a different *mode-seeking* algorithm could be employed (Fukunaga and Hostetler, 1975).
(iii) 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 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 assume throughout that the MLE exists in the model under consideration but that the likelihood is intractable; in the case of non-compact parameter spaces, for example, this may require verification on a case-by-case basis.

We begin by showing pointwise convergence of the posterior (and hence likelihood) approximation under reasonable regularity conditions. It is convenient first to introduce the following concentration condition on the class of ABC kernels which are employed:

Condition K A family of symmetric Markov kernels with densities K_ε indexed by $\varepsilon > 0$ is said to satisfy the concentration condition provided that its members become increasingly concentrated as ε decreases such that

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

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

As the user can freely specify K_ε this is not a problematic condition. It serves only to control the degree of smoothing which the ABC approximation of precision ε can effect.

Proposition 1. *Let $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^{q\times n}$ be a sample with a continuous joint distribution $f(\cdot|\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^d$; $\pi(\boldsymbol{\theta})$ be a bounded prior distribution with support contained in Θ; and let K_ε be the densities of a family of symmetric Markov kernels, which satisfies the concentration condition (**K**).*

Suppose that

$$\sup_{(\mathbf{z},\boldsymbol{\theta})\in\mathcal{B}_\epsilon(\mathbf{x})\times\Theta} f(\mathbf{z}|\boldsymbol{\theta}) < \infty,$$

for some $\epsilon > 0$. Then, for each $\boldsymbol{\theta} \in \Theta$

$$\lim_{\varepsilon\to 0} \widehat{\pi}_\varepsilon(\boldsymbol{\theta}|\mathbf{x}) = \pi\left(\boldsymbol{\theta}|\mathbf{x}\right).$$

Proof. It follows from the concentration condition that:

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

Furthermore, for each $\boldsymbol{\theta} \in \Theta$

$$\begin{aligned}|\widehat{f}_\varepsilon(\mathbf{x}|\boldsymbol{\theta}) - f(\mathbf{x}|\boldsymbol{\theta})| &\leq \int_{\mathcal{B}_\varepsilon(\mathbf{x})} K_\varepsilon(\mathbf{x}|\mathbf{y}) \left|f(\mathbf{y}|\boldsymbol{\theta}) - f(\mathbf{x}|\boldsymbol{\theta})\right| d\mathbf{y} \\ &\leq \sup_{\mathbf{y}\in\mathcal{B}_\varepsilon(\mathbf{x})} |f(\mathbf{y}|\boldsymbol{\theta}) - f(\mathbf{x}|\boldsymbol{\theta})|\end{aligned}$$

due to the symmetry of K_ε which allows us to treat $K_\varepsilon(x|y)$ as a probability density over y. The right hand side of this inequality converges to 0 as $\varepsilon \to 0$ by continuity. Therefore

$$\widehat{f}_\varepsilon(\mathbf{x}|\boldsymbol{\theta}) \xrightarrow{\varepsilon\to 0} f(\mathbf{x}|\boldsymbol{\theta}). \tag{4}$$

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

$$\lim_{\varepsilon\to 0} \int_\Theta \widehat{f}_\varepsilon(\mathbf{x}|\boldsymbol{\theta}')\pi(\boldsymbol{\theta}')d\boldsymbol{\theta}' = \int_\Theta f(\mathbf{x}|\boldsymbol{\theta}')\pi(\boldsymbol{\theta}')d\boldsymbol{\theta}'. \tag{5}$$

The result follows by substitution of (4) and (5) into (1), whenever $\pi(\boldsymbol{\theta}|\mathbf{x})$ is itself well defined. □

Note that the assumption of boundedness of $f(\mathbf{z}|\boldsymbol{\theta})$ over $(\mathbf{z}, \boldsymbol{\theta}) \in \mathcal{B}_\epsilon(\mathbf{x}) \times \Theta$ is a rather mild condition if we restrict the parameter space to a bounded set $\mathbf{D} \subset \mathbb{R}^d$. In the context of this paper this is immediate as we assume that we make use of a uniform instrumental prior on a suitable bounded set $\mathbf{D}$.

Remark 1. Using a similar argument we can show that the result also applies to discrete sampling models since, for ε small enough, $\mathbf{y} \in \mathcal{B}_\varepsilon(\mathbf{x})$ is equivalent to $\mathbf{y} = \mathbf{x}$.

Remark 2. The same result applies mutatis mutandis in the case in which summary statistics are used, but in this context one finds that as ε tends to zero the approximating posterior distribution converges to the approximate distribution of the parameters conditional upon the summary statistics (i.e. that $\lim_{\varepsilon \to 0} \widehat{\pi}_\varepsilon(\boldsymbol{\theta}|\eta(\mathbf{x})) = \pi(\boldsymbol{\theta}|\eta(\mathbf{x}))$), where $\eta : \mathbb{R}^{q \times n} \to \mathbb{R}^m$ denotes the summary statistic. This distribution coincides with the full data posterior (as can be established via the factorisation of Lehmann and Casella (1998, Theorem 6.5), say) only when the statistics are sufficient for inference about the parameters of interest.

This result can be strengthened by noting that it is straightforward to obtain bounds on the error introduced at finite ε if we assume Lipschitz continuity of the true likelihood. 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 simple ABC approximation 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 emphasise the role of sufficiency.

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

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

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

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

Proof. Using the integral Mean Value Theorem (as used in a similar context by Dean et al. (2011, Equation 6)) we find that for $\boldsymbol{\theta} \in \mathbf{D}$ and any $\varepsilon \in (0, \epsilon)$:

$$\begin{aligned}\widehat{f}_\varepsilon(\mathbf{x}|\boldsymbol{\theta}) &\propto \int I(\rho(\eta(\mathbf{y}), \eta(\mathbf{x})) < \varepsilon) f(\mathbf{y}|\boldsymbol{\theta}) d\mathbf{y} \\ &= \int_{\mathcal{B}_\varepsilon} f^\eta(\eta'|\boldsymbol{\theta}) d\eta' = \lambda(\mathcal{B}_\varepsilon(\eta(\mathbf{x}))) f^\eta(\xi(\boldsymbol{\theta}, \mathbf{x}, \varepsilon)|\boldsymbol{\theta}),\end{aligned}$$

for some $\xi(\boldsymbol{\theta}, \mathbf{x}, \varepsilon) \in \mathcal{B}_\varepsilon(\eta(\mathbf{x}))$, where λ is the Lebesgue measure and $I(\cdot)$ is the indicator function. Then

$$\widehat{\pi}_\varepsilon(\boldsymbol{\theta}|\mathbf{x}) = \frac{f^\eta\left(\xi(\boldsymbol{\theta}, \mathbf{x}, \varepsilon)|\boldsymbol{\theta}\right)\pi(\boldsymbol{\theta})}{\int_{\mathbf{D}} f^\eta\left(\xi(\boldsymbol{\theta}', \mathbf{x}, \varepsilon)|\boldsymbol{\theta}'\right)\pi(\boldsymbol{\theta}')d\boldsymbol{\theta}'}.$$

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

$$\lim_{\varepsilon \to 0} f^\eta\left(\xi(\boldsymbol{\theta}, \mathbf{x}, \varepsilon)|\boldsymbol{\theta}\right) = f^\eta\left(\eta(\mathbf{x})|\boldsymbol{\theta}\right). \tag{6}$$

Using the Dominated Convergence Theorem we have

$$\lim_{\varepsilon \to 0} \int_{\mathbf{D}} f^\eta\left(\xi(\boldsymbol{\theta}', \mathbf{x}, \varepsilon)|\boldsymbol{\theta}'\right)\pi(\boldsymbol{\theta}')d\boldsymbol{\theta}' = \int_{\mathbf{D}} f^\eta\left(\eta(\mathbf{x})|\boldsymbol{\theta}'\right)\pi(\boldsymbol{\theta}')d\boldsymbol{\theta}'. \tag{7}$$

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

$$f\left(\mathbf{x}|\boldsymbol{\theta}\right) = h(\mathbf{x}) f^\eta\left(\eta(\mathbf{x})|\boldsymbol{\theta}\right). \tag{8}$$

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

The result also holds for discrete and mixed continuous-discrete models with the obvious changes to the proof.

With only a slight strengthening of the conditions, Proposition 1 allows us to show convergence of the mode as $\varepsilon \to 0$ 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 of Proposition 1 we assume equicontinuity of $\widehat{\pi}_\varepsilon(\cdot|\mathbf{x})$ 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 $\widehat{\pi}_\varepsilon(\cdot|\mathbf{x})$ achieve its global maximum at $\boldsymbol{\theta}_\varepsilon$ for each $\varepsilon > 0$ and suppose that $\pi(\cdot|\mathbf{x})$ has unique maximiser $\boldsymbol{\theta}_0$. Under the conditions of Proposition 1; if $\widehat{\pi}_\varepsilon(\cdot|\mathbf{x})$ is equicontinuous, then*

$$\lim_{\varepsilon \to 0} \widehat{\pi}_\varepsilon(\boldsymbol{\theta}_\varepsilon|\mathbf{x}) = \pi(\boldsymbol{\theta}_0|\mathbf{x}).$$

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

$$\lim_{k \to \infty} \left|\widehat{\pi}_{k,\varepsilon}\left(\tilde{\boldsymbol{\theta}}_\varepsilon|\mathbf{x}\right) - \pi\left(\boldsymbol{\theta}_0|\mathbf{x}\right)\right| \leq \gamma, \text{ a.s.}$$

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, γ, the maximum of the AMLE approximation will,

for large enough ABC samples, almost surely be $\gamma-$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 $\tilde{\boldsymbol{\theta}}_\varepsilon$ to approximate $\boldsymbol{\theta}_0$ for large k and small ε.

The convergence shown in the above results depends on the use of a sufficient statistic in order to guarantee convergence to the MLE. In contexts where the likelihood is intractable, such a statistic 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 present 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 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.

In order to establish this result, we require that, allowing $\eta_n(\mathbf{x}) = \eta_n(x_1, \ldots, x_n)$ to denote a sequence of m-dimensional summary statistics, the following four conditions hold:

- **S.i** There exists some function $g : \Theta \to \mathbb{R}^m$ such that, for $\mathbf{x}$ a simple random sample, $\lim_{n\to\infty} \eta_n(\mathbf{x}) \overset{a.s.}{=} g(\boldsymbol{\theta})$ for $\pi -$ a.e. $\boldsymbol{\theta}$.
- **S.ii** The function $g : \Theta \to \mathbb{R}^m$ is an injective mapping. Letting $H = g(\mathbf{D}) \subset \mathbb{R}^m$ denote the image of the feasible parameter space under g, $g^{-1} : H \to \Theta$ is an α-Lipschitz continuous function for some $\alpha \in \mathbb{R}_+$.
- **S.iii** The ABC kernels, defined in the space of the summary statistics, satisfy condition **K**, i.e. $K_\varepsilon^\eta(\cdot|\eta')$ it is concentrated within a ball of radius ε for all ε: $\operatorname{supp} K_\varepsilon(\cdot|\eta') \subseteq \mathcal{B}_\varepsilon(\eta')$ and for any fixed $\varepsilon > 0$ we require that $\sup_{\eta,\eta'} K_\varepsilon(\eta'|\eta) < \infty$.
- **S.iv** The nonparametric estimator used always provides an estimate of the mode which lies within the convex hull of the sample.

Some interpretation of these conditions 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.

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

Assume that conditions ***S.i***–***S.iv*** *hold. Then, for any* $\varepsilon > 0$*:*

(a) $\operatorname{supp} \lim_{n\to\infty} \widehat{\pi}_\varepsilon(\boldsymbol{\theta}|\eta_n^\star) \subseteq \mathcal{B}_{\alpha\varepsilon}(\boldsymbol{\theta}^\star)$ *for the statistics,* $\eta^\star$*, associated with* $\mu(\cdot|\boldsymbol{\theta}^\star)$*-almost every collection of observations for* π*-almost every* $\boldsymbol{\theta}^\star$.
(b) The AMLE approximation of the MLE lies within $\mathcal{B}_{\alpha\varepsilon}(\boldsymbol{\theta}^\star)$ *almost surely.*

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

$$\lim_{n\to\infty} f_\varepsilon^{\eta_n}(\eta|\boldsymbol{\theta}) = \lim_{n\to\infty} \int \mu^{\eta_n}(d\eta'|\boldsymbol{\theta}) K_\varepsilon(\eta|\eta') \overset{a.s.}{=} K_\varepsilon(\eta|g(\boldsymbol{\theta}))$$

with the final equality following from **S.i** and **S.iii** (noting that almost sure convergence of η_n to $g(\boldsymbol{\theta})$ together with the boundedness of $K_\varepsilon(\eta|\cdot)$ yields directly that $K_\varepsilon(\eta|\eta_n) \overset{a.s.}{\to} K_\varepsilon(\eta|g(\theta))$). From which it is clear that $\operatorname{supp} \lim_{n\to\infty} f_\varepsilon^{\eta_n}(\cdot|\boldsymbol{\theta}) \subseteq \mathcal{B}_\varepsilon(g(\boldsymbol{\theta}))$ by **S.iii**.

And the ABC approximation to the posterior density of $\boldsymbol{\theta}$, $\lim_{n\to\infty} \widehat{\pi}_\varepsilon(\cdot|\eta_n)$, may be similarly constrained:

$$\begin{aligned} \lim_{n\to\infty} \widehat{\pi}_\varepsilon(\boldsymbol{\theta}|\eta_n) > 0 \Rightarrow \quad \lim_{n\to\infty} f_\varepsilon^{\eta_n}(\eta_n|\boldsymbol{\theta}) > 0 &\Rightarrow \quad \lim_{n\to\infty} ||\eta_n - g(\boldsymbol{\theta})|| \overset{a.s.}{\leq} \varepsilon \\ &\Rightarrow \quad \lim_{n\to\infty} ||g^{-1}(\eta_n) - \boldsymbol{\theta}|| \overset{a.s.}{\leq} \alpha\varepsilon \end{aligned}$$

using **S.ii**. And by assumptions **S.i** and **S.ii** together with the continuous mapping theorem we have that $g^{-1}(\eta_n^\star) \overset{a.s.}{\to} \boldsymbol{\theta}^\star$ giving result (a); result (b) follows immediately from **S.iv**. □

It is 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 close in spirit to the result of Marin et al. (2013) in the model selection context, although their argument is quite different, 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.

We conclude this section with some simple examples of situations in which these conditions are met. The first is a simple, concrete example which illustrates that sufficient statistics fit directly into this framework and also that many other statistics have the required properties. A second more abstract example illustrates the general principle.

Example 1 (Binomial Sampling Model)**.** If $\mathbf{x}$ is a simple random sample from a Binomial model with known size r and unknown success probability p, the familiar sufficient statistic $\eta_n(\mathbf{x}) = \frac{1}{nr}\sum_{i=1}^n x_i$ satisfies the required conditions:

S.i is satisfied with $g = \mathrm{Id}$ on $[0,1]$, where Id denotes the identity function, by the strong law of large numbers.

S.ii is satisfied: Id is injective on $[0,1]$ and has 1-Lipschitz inverse.

While the remaining conditions are easily satisfied:

S.iii Is immediate with the simple kernel $K_\varepsilon(\eta'|\eta) = \mathbf{1}_{\mathcal{B}_\varepsilon(\eta)}(\eta')/2\varepsilon$. and **S.iv** is readily satisfiable.

However, sufficiency of the statistic is not required, one could instead use the proportion of the samples which take the value zero, $\widetilde{\eta}_n(\mathbf{x}) = \frac{1}{n}|\{i \in 1,\ldots,n : x_i = 0\}|$ and then $\widetilde{\eta}_n \overset{a.s.}{\to} (1-p)^r$.

S.i is satisfied with $g(p) = (1-p)^r$ by the strong law of large numbers.

S.ii is satisfied as g is injective on $[0,1]$ and $g^{-1}(\widetilde{\eta}) = 1 - \widetilde{\eta}^{1/r}$ has derivative absolutely bounded by $1/r$ on $[0,1]$.

Some care is needed in the selection of statistics, however, even in simple models such as this one: the sample variance of the binomial model also converges almost surely — to $rp(1-p)$ — but using this statistic would not satisfy the requirements of Proposition 2 as it is not an injective mapping and one would not be able to differentiate between parameter values of p and $1-p$.

Example 2 (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 $\eta_n^1 = \widehat{F}^{-1}(q_1)$ and $\eta_n^2 = \widehat{F}^{-1}(q_2)$ to denote two empirical quantiles. By the Glivenko-Cantelli theorem, these empirical quantiles converge almost-surely to the true quantiles:

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

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}:

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

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 $(\mathbf{Z}_1, \ldots, \mathbf{Z}_m)$. Let K be a kernel, h_m be a bandwidth such that $h_m \to 0$ when $m \to \infty$, then a kernel density estimator is defined by

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

Under the conditions $h_m \to 0$ and $m h_m^n / \log(m) \to \infty$ together with Theorem 1 from Abraham et al. (2003), we have that $\tilde{\boldsymbol{\theta}}_m \xrightarrow{a.s.} \tilde{\boldsymbol{\theta}}$ as $m \to \infty$. 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.

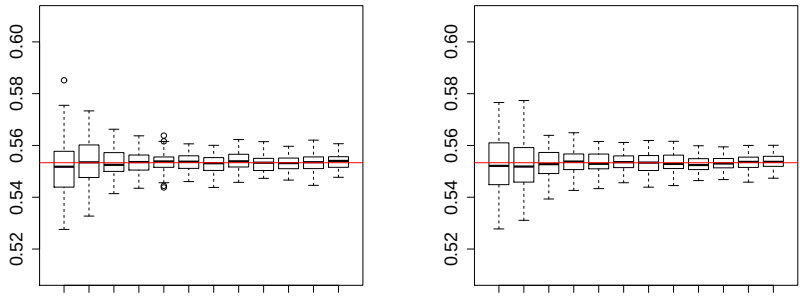
We note that we have focussed upon the more challenging setting of continuous parameter spaces. Naturally, the AMLE approach can be implemented in cases where the parameter space is either continuous, discrete or a combination of discrete and continuous. This will be illustrated in the next section through some examples.

4. Examples

We present four examples in order of increasing complexity. The first two examples illustrate the performance of the algorithm in simple scenarios in which the solution is known; the third compares the algorithm with a numerical method in a setting which has recently been studied using ABC and the final example demonstrates performance on a challenging estimation problem which has recently attracted some attention in the literature. In all the examples the simple ABC rejection algorithm was used, together with ABC kernel (3) and the Euclidean norm. For the second, third and fourth examples, 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.

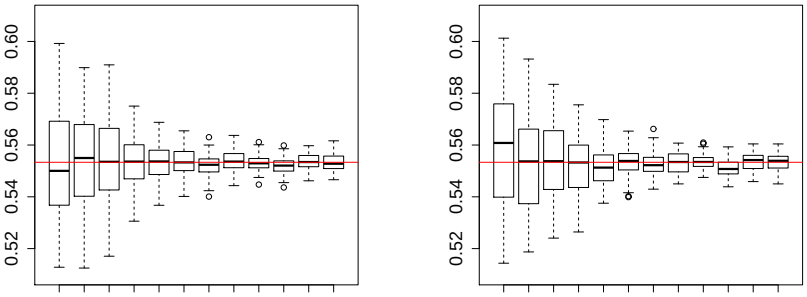
4.1. Binomial model

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 $\eta(\mathbf{x}) = \bar{x}$



(a) $D = (0.45, 0.65)$ (b) $D = (0, 1)$

FIG 1. *Effect of* $k \in \{30, 100, 1,000, 2,000, \ldots, 10,000\}$ *for* $\varepsilon = 0.05$. *The continuous red line represents the true MLE value.*

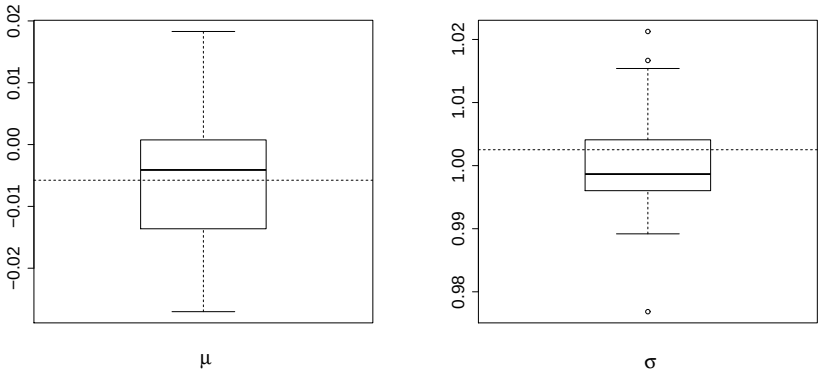


(a) (b)

FIG 2. *Effect of* $\varepsilon \in \{1, 0.9, \ldots, 0.1, 0.05, 0.01\}$ *for* $k = 10,000$*: (a)* $D = (0.45, 0.65)$*; (b)* $D = (0, 1)$. *The continuous red line represents the true MLE value*

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 $\tilde{\theta} = 0.552$.

There are three quantities affecting the precision in the estimation of $\widehat{\theta}$: D, k and ε. Figure 1 illustrates the effect of varying $k \in \{30, 100, 1,000, 2,000, \ldots, 10,000\}$ 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 AMLE algorithm. This demonstrates that, although unsurprising, the acceptance rate and hence computational efficiency is improved when some D which is relatively concentrated around the MLE is available (the choice $D = (0.45, 0.65)$ produces an acceptance rate about 5 times greater than the choice $D = (0, 1)$), the precise range of D has little effect on the final estimate unless ε is very large as discussed below. Figure 2 shows the effect of $\varepsilon \in \{1, 0.9, \ldots, 0.1, 0.05, 0.01\}$ for a



(a) (b)

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

fixed k and two different choices of D. In this case we can note that the effect of ε on the precision is significant. The estimation precision is similar when ε is small for both choices of D. However, we can see that larger values of ε produce the acceptance of more extreme values of θ when combined with the choice $D = (0, 1)$. This is reflected, for instance, in the median of the boxplots corresponding to $\varepsilon = 1$ in Figure 2. This interaction arises because ε is similar in scale to the diameter of D and would not be expected to produce a noticeable effect in realistic settings in which D would typically be much larger than ε — smoothing on a similar scale to the total prior uncertainty in the range of the parameters is unlikely to be desirable.

4.2. Normal model

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

Consider the priors $\mu \sim \text{Unif}(-0.25, 0.25)$ and $\sigma \sim \text{Unif}(0.75, 1.25)$ (crude estimates of location and scale can often be obtained from data, justifying such a choice; using broader prior support here increases computational cost but does not prevent good estimation), a tolerance $\varepsilon = 0.01$, a sufficient statistic $\eta(\mathbf{x}) = (\bar{\mathbf{x}}, s)$, 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. α-stable logarithmic daily returns model

Logarithmic daily return prices are typically modelled using Lévy processes. For this reason, it is necessary to model the increments (logarithmic returns)

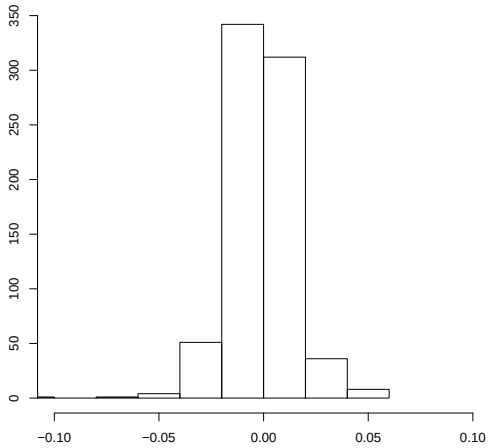


FIG 4. *Logarithmic daily returns using the closing price of IBM ordinary stock Jan. 1 2009 to Jan. 1 2012.*

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 $(\alpha, \beta, \mu, \sigma)$ $\alpha-$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. 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.

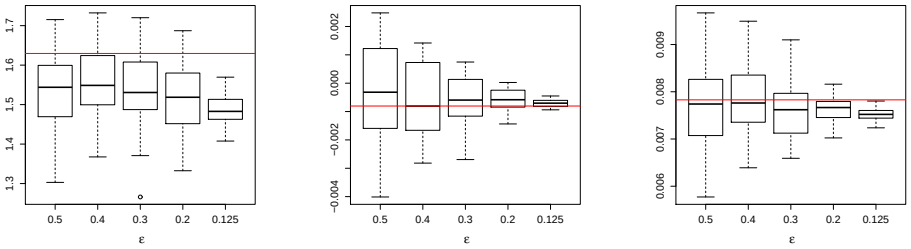
Here, we analyse the logarithmic daily returns using the closing price of IBM ordinary stock from January 1 2009 to January 1 2012. Figure 4 shows the corresponding histogram. For this data set, the MLE obtained using the numerical method implemented in the R package 'fBasics' (Wuertz et al., 2010) is $(\hat{\alpha}, \hat{\beta}, \hat{\mu}, \hat{\sigma}) = (1.6295, -0.05829, -0.0008, 0.0078)$.

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

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

which due to the scale of the data may appear concentrated but are, in fact, rather uninformative, allowing a location parameter essentially anywhere within the convex hull of the data, scale motivated by similar considerations and any value of the shape parameter consistent with the problem at hand.

For the (non-sufficient) summary statistic we use proposal S_4 of Peters et al. (2010), which consists of the values of the empirical characteristic func-



(a) (b) (c)

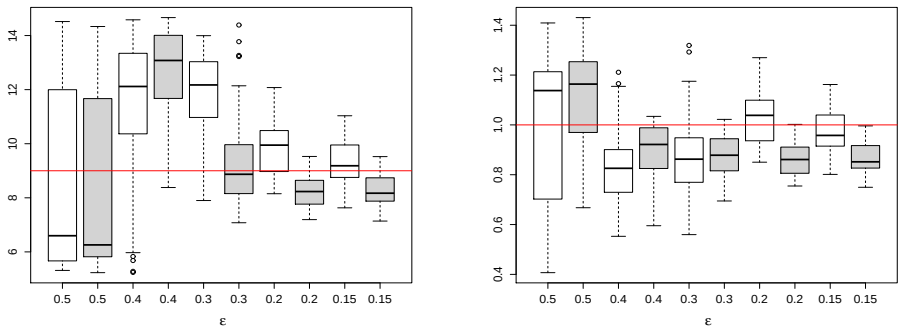
FIG 5. *Monte Carlo variability of the AMLE: (a) α; (b) μ; (c) σ. Horizontal lines represent MLE estimator produced by the R package 'fBasics'.*

tion 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 2,500, a tolerance $\varepsilon \in \{0.5, 0.4, 0.3, 0.2, 0.125\}$ and Gaussian kernel density estimation. Figure 5 illustrates Monte Carlo variability of the AMLE of (α, μ, σ). Boxplots were obtained using 50 replicates of the AMLE procedure. In general, considerable care must of course be taken in the selection of statistics.

4.4. Superposed gamma point processes

The modelling of an unknown number of superposed gamma point processes provides another scenario with intractable likelihoods which is currently attracting some attention (Cox and Kartsonaki, 2012; Mengersen et al., 2013). Intractability of the likelihood in this case is a consequence of the dependency between the observations, which complicates the construction of their joint distribution. Superposed point processes have applications in a variety of areas, for instance Cox and Smith (1954) present an application of this kind of processes in the context of neurophysiology. In this example we consider a simulated sample of size 88 of $N = 2$ superposed point processes with inter-arrival times identically distributed as a gamma random variable with shape parameter $\alpha = 9$ and rate parameter $\beta = 1$ observed in the interval $(0, t_0)$, with $t_0 = 420$. This choice is inspired by the simulated example presented in Cox and Kartsonaki (2012).

In order to make inference on the parameters (N, α, β) using the AMLE approach, we implement two ABC samplers using the priors $N \sim \text{Unif}\{1, 2, 3, 4, 5\}$, $\alpha \sim \text{Unif}(5, 15)$, $\beta \sim \text{Unif}(0.25, 1.5)$, tolerances $\varepsilon \in \{0.5, 0.4, 0.3, 0.2, 0.15\}$ and two sets of summary statistics. The first set of summary statistics, proposed in Cox and Kartsonaki (2012) and subsequently used in Mengersen et al. (2013), consists of the mean rate of occurrence, the coefficient of variation of the intervals between successive points, the sum of the first five autocorrelations of the intervals, the mean of the intervals, and the Poisson indices of dispersion, variance divided by mean, for intervals of length 1, 5, 10 and 20. Cox and Kartsonaki (2012) mention that summary statistics based on the intervals between



(a) (b)

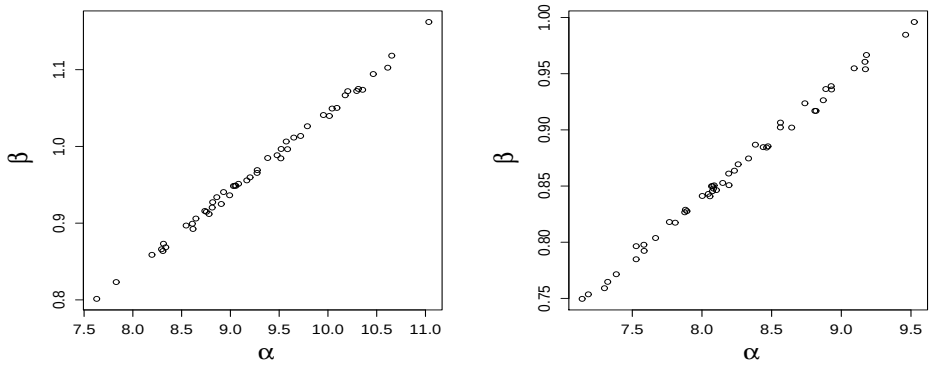
FIG 6. *Effect of* $\varepsilon \in \{0.5, 0.4, 0.3, 0.2, 0.15\}$ *for* $k = 5,000$*: (a)* α*; (b)* β*. The AMLE samples with* 8 *and* 9 *summary statistics are presented in white and gray boxplots, respectively. The continuous red line represents the true value of the parameter.*

TABLE 1
Replicate study with a single data realisation. Empirical distribution of $\widehat{N}$ *for different values of* ε

	8 summary statistics				9 summary statistics			
ε	1	2	3	4	1	2	3	4
0.5	29	0	1	20	33	0	15	2
0.4	5	0	35	10	0	0	50	0
0.3	0	4	46	0	0	37	13	0
0.2	0	50	0	0	0	50	0	0
0.15	0	50	0	0	0	50	0	0

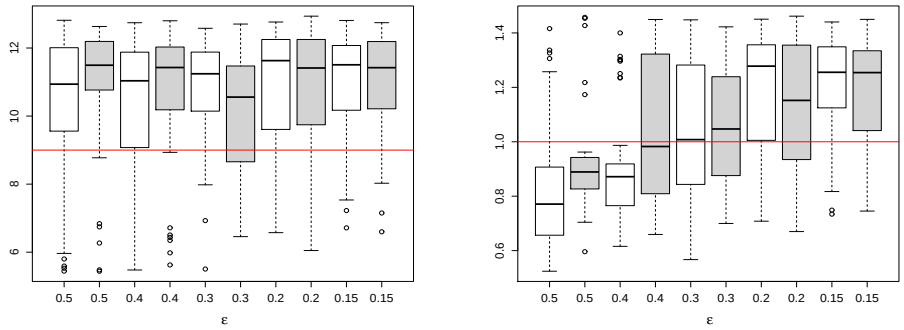
successive points are likely to be useful when N is small, therefore we consider a second set of summary statistics by adding a ninth quantity based on the third moment: the sample skewness of the intervals between successive points $\sum_{j=1}^{n}(x_j - \bar{x})^3/(\sum_{j=1}^{n}(x_j - \bar{x})^2/n)^{3/2}$. The summary statistics of the simulated data are $(0.210, 0.669, -0.355, 4.74, 0.910, 0.476, 0.268, 0.200, 0.493)$.

The joint posterior distribution of (N, α, β) is singular, in the sense that it is neither discrete nor continuous. The AMLE approach is still applicable in this context given that the maximisation step can be conducted by noting that the conditional distribution $\alpha, \beta | N$ is continuous. Therefore, given an ABC sample of these parameters, we can calculate kernel density estimators for the continuous parameters for each value of N and find the maximiser of each. We then multiply the maximum for each N by the number of samples obtained for that value of N and take the largest of these. Figure 6 shows the Monte Carlo variability, estimated by using 50 AMLE samples, for each of the two AMLE approaches based on ABC samples of size $5,000$. We can notice that the precision in the estimation of (α, β) increases faster, as the tolerance decreases, when using 9 summary statistics. We can observe the same phenomenon from Table 1 in the estimation of N. (Note that the horizontal line shows the parameters used to generate the data *not* the true value of the MLE). Figure 7 shows



(a) (b)

FIG 7. *AMLE estimators of* β vs. *AMLE estimators of* α*: (a) 8 summary statistics; (b) 9 summary statistics.*



(a) (b)

FIG 8. *Effect of* $\varepsilon \in \{0.5, 0.4, 0.3, 0.2, 0.15\}$ *for* $k = 5{,}000$*: (a)* α*; (b)* β*. The AMLE samples with* 8 *and* 9 *summary statistics are presented in white and gray boxplots, respectively. The continuous red line represents the true value of the parameter.*

scatter plots of the AMLE estimators of β and α for $\varepsilon = 0.15$ and both sets of summary statistics. This scatterplot demonstrates that the mean (α/β) of the gamma distribution is much more tightly constrained by the data than the shape parameter, leading to a nearly-flat ridge in the likelihood surface. The variability in the estimated value of α/β is, in fact, rather small; while the variability in estimation of the shape parameter reflects the lack of information about this quantity in the data and the consequent flatness of the likelihood surface.

To show the variability of the estimator with different data, we also compare the variability of the estimators obtained using 50 different data sets. For each data set we obtain the corresponding AMLE of (N, α, β) by using the priors $N \sim \text{Unif}\{1, 2, 3, 4, 5\}$, $\alpha \sim \text{Unif}(5, 13)$ and $\beta \sim \text{Unif}(0.5, 1.5)$, tolerances $\varepsilon \in \{0.5, 0.4, 0.3, 0.2, 0.15\}$ and the two sets of summary statistics mentioned above. Figure 8 shows the boxplots of the AMLEs for (α, β) obtained using

TABLE 2
Replicate study with 50 data realisations. Empirical distribution of $\widehat{N}$ for different values of ε

	8 summary statistics				9 summary statistics			
ε	1	2	3	4	1	2	3	4
0.5	7	0	24	19	5	0	43	2
0.4	8	0	41	1	3	0	24	23
0.3	1	27	22	0	0	43	7	0
0.2	0	46	4	0	0	44	6	0
0.15	0	47	3	0	0	46	4	0

ABC samples of size $5,000$. We can observe that the behaviour of the estimators of (α, β) is fairly similar for both sets of summary statistics. Table 1 also suggests an improvement in the estimation of N produced by the inclusion of the sample skewness.

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 can itself be informative. 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. In principle this should allow straightforward extension of the method to computation of approximate confidence intervals and profile likelihoods, although some extension of the theoretical results might be required to formally justify doing so. Furthermore, we would strongly recommend that the approximation of the likelihood 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

We thank two reviewers and an Associate Editor for helpful comments.

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. MR1985502

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.

Biau, G., Cérou, F. and Guyader, A. (2012). New Insights into Approximate Bayesian Computation. ArXiv preprint arXiv:1207.6461.

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. MR2236862

Blum, M. G. B. (2010). Approximate Bayesian computation: a nonparametric perspective. *Journal of the American Statistical Association* 105: 1178–1187. MR2752613

Bretó, C., Daihi, H., Ionides, E. L. and King, A. A. (2009). Time series analysis via mechanistic models. *Annals of Applied Statistics* 3: 319–348. MR2668710

Cox, D. R. and Kartsonaki, C. (2012). The fitting of complex parametric models. *Biometrika* 99: 741–747. MR2966782

Cox, D. R. and Reid, N. (2004). A note on pseudolikelihood constructed from marginal densities. *Biometrika* 91: 729–737. MR2090633

Cox, D. R. and Smith, W. L. (1954). On the superposition of renewal processes. *Biometrika* 41: 91–9. MR0062995

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. MR2758237

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 Association* 99: 523–536. MR2062837

Didelot, X., Everitt, R. G., Johansen, A. M. and Lawson, D. J. (2011). Likelihood-free estimation of model evidence. *Bayesian Analysis* 6: 49–76. MR2781808

Diggle, P. J. and Gratton, R. J. (1984) Monte Carlo Methods of Inference for Implicit Statistical Models. *Journal of the Royal Statistical Society* B 46:193–227. MR0781880

Duong, T. and Hazleton, M. L. (2005). Cross-validation bandwidth matrices for multivariate kernel density estimation. *Scandinavian Journal of Statistics* 32:485–506. MR2204631

Duong, T. (2011). *ks: Kernel smoothing*. R package version 1.8.5. http://CRAN.R-project.org/package=ks

Ehrlich, E., Jasra, A. and Kantas, N. (2012). Static parameter estimation for ABC approximations of hidden Markov models. ArXiv preprint arXiv:1210.4683.

Fan, Y., Nott, D. J. and Sisson, S. A. (2012). Approximate Bayesian Computation via Regression Density Estimation. *Stat* 2: 34–48.

Fearnhead, P. and Prangle, D. (2012). Constructing Summary Statistics for Approximate Bayesian Computation: Semi-automatic ABC (with discussion). *Journal of the Royal Statistical Society* B 74: 419–474. MR2925370

Fukunaga, K. and Hostetler, L. D. (1975). The Estimation of the Gradient of a Density Function, with Applications in Pattern Recognition. *IEEE Transactions on Information Theory* 21: 32–40. MR0388638

Gaetan, C. and Yao, J. F. (2003). A multiple-imputation Metropolis version of the EM algorithm. *Biometrika* 90: 643–654. MR2006841

Gouriéroux, C., Monfort, A. and Renault, E. (1993). Indirect Inference. *Journal of Applied Econometrics* 8:S85–S118.

Ionides, E. L. (2005). Maximum Smoothed Likelihood Estimation. *Statistica Sinica* 15: 1003–1014. MR2234410

Jaki, T. and West, R. W. (2008). Maximum Kernel Likelihood Estimation. *Journal of Computational and Graphical Statistics* 17: 976. MR2649075

Jasra, A., Kantas, N. and Ehrlich, E. (2013). Approximate Inference for Observation Driven Time Series Models with Intractable Likelihoods. ArXiv Preprint arXiv:1303.7318.

Jing, J., Koch, I. and Naito, K. (2012). Polynomial Histograms for Multivariate Density and Mode Estimation. *Scandinavian Journal of Statistics* 39: 75–96. MR2896792

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. MR2416438

Konakov, V. D. (1973). On asymptotic normality of the sample mode of multivariate distributions. *Theory of Probability and its Applications* 18: 836–842. MR0336874

Lehmann, E. and Casella, G. (1998). *Theory of Point Estimation* (revised edition). Springer-Verlag, New York. MR1639875

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.

Marin, J.-M., Pudlo, P., Robert, C. P. and Ryder, R. (2011). Approximate Bayesian Computational methods. *Statistics and Computing* 21: 289–291. MR2992292

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

Marjoram, P., Molitor, J., Plagnol, V. and Tavaré, S. (2003). Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences of the United States of America*: 15324–15328.

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

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. MR1833706

Parzen, E. (1962). On estimation of a probability density function and mode. *Annals of Mathematical Statistics* 33: 1065–1076. MR0143282

Peters, G. W., Sisson, S. A. and Fan, Y. (2010). Likelihood-free Bayesian inference for $\alpha-$stable models. *Computational Statistics & Data Analysis* 56: 3743–3756. MR2943924

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., 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. MR0947566

Rubio, F. J. and Johansen, A. M. (March, 2012). On Maximum Intractable Likelihood Estimation. University of Warwick, Dept. of Statistics. CRiSM working paper 12–04.

Rudin, W. (1976). *Principles of Mathematical Analysis*. New York: McGraw-Hill. MR0385023

Sisson, S. A., Fan, Y. and Tanaka, M. M. (2007). Sequential Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences of the United States of America*, 104: 1760–1765. MR2301870

Sköld, M. and Roberts, G. O. (2003). Density estimation for the Metropolis–Hastings algorithm. *Scandinavian Journal of Statistics* 30: 699–718. MR2155478

Toni, T., Welch, D., Strelkowa, N., Ipsen, A. and 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. MR1113551

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

Wuertz, D. and R core team members (2010). *fBasics: Rmetrics - Markets and Basic Statistics*. R package version 2110.79. http://CRAN.R-project.org/package=fBasics