Biometrika (2017), **xx**, x, *pp*. 1–15

Printed in Great Britain

On the Asymptotic Efficiency of Approximate Bayesian Computation Estimators

BY WENTAO LI

School of Mathematics, Statistics and Physics, Newcastle University, Newcastle upon Tyne NE1 7RU, U.K.

wentao.li@newcastle.ac.uk

AND PAUL FEARNHEAD

Department of Mathematics and Statistics, Lancaster University, Lancaster LA1 4YF, U.K.

p.fearnhead@lancaster.ac.uk

SUMMARY

Many statistical applications involve models for which it is difficult to evaluate the likelihood, but from which it is relatively easy to sample. Approximate Bayesian computation is a likelihood-free method for implementing Bayesian inference in such cases. We present results on the asymptotic variance of estimators obtained using approximate Bayesian computation in a large-data limit. Our key assumption is that the data is summarized by a fixed-dimensional summary statistic that obeys a central limit theorem. We prove asymptotic normality of the mean of the approximate Bayesian computation posterior. This result also shows that, in terms of asymptotic variance, we should use a summary statistic that is the same dimension as the parameter vector, p; and that any summary statistic of higher dimension can be reduced, through a linear transformation, to dimension p in a way that can only reduce the asymptotic variance of the posterior mean. We look at how the Monte Carlo error of an importance sampling algorithm that samples from the approximate Bayesian computation posterior affects the accuracy of estimators. We give conditions on the importance sampling proposal distribution such that the variance of the estimator will be the same order as that of the maximum likelihood estimator based on the summary statistics used. This suggests an iterative importance sampling algorithm, which we evaluate empirically on a stochastic volatility model.

Some key words: Approximate Bayesian computation; Asymptotics; Dimension Reduction; Importance Sampling; Partial Information; Proposal Distribution.

1. INTRODUCTION

Many statistical applications involve inference about models that are easy to simulate from, but for which it is difficult, or impossible, to calculate likelihoods. In such situations it is possible to use the fact we can simulate from the model to enable us to perform inference. There is a wide class of such likelihood-free methods of inference including indirect inference (Gouriéroux & Ronchetti, 1993; Heggland & Frigessi, 2004), the bootstrap filter (Gordon et al., 1993), simulated methods of moments (Duffie & Singleton, 1993), and synthetic likelihood (Wood, 2010).

We consider a Bayesian version of these methods, termed approximate Bayesian computation. This involves defining an approximation to the posterior distribution in such a way that it is possible to sample from this approximate posterior using only the ability to sample from the

model. Arguably the first approximate Bayesian computation method was that of Pritchard et al. (1999), and these methods have been popular within population genetics (Beaumont et al., 2002), ecology (Beaumont, 2010) and systems biology (Toni et al., 2009). More recently, there have been applications to areas including stereology (Bortot et al., 2007), finance (Peters et al., 2011) and cosmology (Ishida et al., 2015).

Let $K(x)$ be a density kernel, scaled, without loss of generality, so that $\max_x K(x) = 1$. Further, let $\varepsilon > 0$ be a bandwidth. Denote the data by $Y_{\text{obs}} = (y_{\text{obs},1}, \ldots, y_{\text{obs,n}})$. Assume we have chosen a finite-dimensional summary statistic $s_n(Y)$, and denote $s_{\text{obs}} = s_n(Y_{\text{obs}})$. If we model the data as a draw from a parametric density, $f_n(y \mid \theta)$, and assume prior, $\pi(\theta)$, then we define the approximate Bayesian computation posterior as

$$\pi_{\text{ABC}}(\theta \mid s_{\text{obs}}, \varepsilon) \propto \pi(\theta) \int f_n(s_{\text{obs}} + \varepsilon v \mid \theta) K(v)\, dv, \tag{1}$$

where $f_n(s \mid \theta)$ is the density for the summary statistic implied by $f_n(y \mid \theta)$. Let $f_{\text{ABC}}(s_{\text{obs}} \mid \theta, \varepsilon) = \int f_n(s_{\text{obs}} + \varepsilon v \mid \theta) K(v)\, dv$. This framework encompasses most implementations of approximate Bayesian computation. In particular, the use of the uniform kernel corresponds to the popular rejection-based rule (Beaumont et al., 2002).

The idea is that $f_{\text{ABC}}(s_{\text{obs}} \mid \theta, \varepsilon)$ is an approximation of the likelihood. The approximate Bayesian computation posterior, which is proportional to the prior multiplied by this likelihood approximation, is an approximation of the true posterior. The likelihood approximation can be interpreted as a measure of, on average, how close the summary, s_n, simulated from the model is to the summary for the observed data, s_{obs}. The choices of kernel and bandwidth determine the definition of closeness.

By defining the approximate posterior in this way, we can simulate samples from it using standard Monte Carlo methods. One approach, that we will focus on later, uses importance sampling. Let $K_\varepsilon(x) = K(x/\varepsilon)$. Given a proposal density, $q_n(\theta)$, a bandwidth, ε, and a Monte Carlo sample size, N, an importance sampler would proceed as in Algorithm 1. The set of accepted parameters and their associated weights provides a Monte Carlo approximation to π_{ABC}. If we set $q_n(\theta) = \pi(\theta)$ then this is just a rejection sampler. In practice sequential importance sampling methods are often used to learn a good proposal distribution (Beaumont et al., 2009).

Algorithm 1. Importance and rejection sampling approximate Bayesian computation

1. Simulate $\theta_1, \ldots, \theta_N \sim q_n(\theta)$;
2. For each $i = 1, \ldots, N$, simulate $Y^{(i)} = \{y_1^{(i)}, \ldots, y_n^{(i)}\} \sim f_n(y \mid \theta_i)$;
3. For each $i = 1, \ldots, N$, accept θ_i with probability $K_\varepsilon\{s_n^{(i)} - s_{\text{obs}}\}$, where $s_n^{(i)} = s_n\{Y^{(i)}\}$; and define the associated weight as $w_i = \pi(\theta_i)/q_n(\theta_i)$.

There are three choices in implementing approximate Bayesian computation: the choice of summary statistic, the choice of bandwidth, and the Monte Carlo algorithm. For importance sampling, the last of these involves specifying the Monte Carlo sample size, N, and the proposal density, $q_n(\theta)$. These, roughly, relate to three sources of approximation. To see this, note that as $\varepsilon \to 0$ we would expect (1) to converge to the posterior given s_{obs} (Fearnhead & Prangle, 2012). Thus the choice of summary statistic governs the approximation, or loss of information, between using the full posterior distribution and using the posterior given the summary. The value ε then affects how close the approximate Bayesian computation posterior is to the posterior given the summary. Finally there is Monte Carlo error from approximating the approximate Bayesian computation posterior with a Monte Carlo sample. The Monte Carlo error is not only affected by the Monte Carlo algorithm, but also by the choices of summary statistic and bandwidth,

which together affect, say, the probability of acceptance in step 3 of Algorithm 1. Having a higher-dimensional summary statistic, or a smaller value of ε, will tend to reduce this acceptance probability and hence increase the Monte Carlo error.

This work aims to study the interaction between the three sources of error, in the case where the summary statistics obey a central limit theorem for large n. We are interested in the efficiency of approximate Baysian computation, where by efficiency we mean that an estimator obtained from running Algorithm 1 has the same rate of convergence as the maximum likelihood estimator for the parameter given the summary statistic. In particular, this work is motivated by the question of whether approximate Bayesian computation can be efficient as $n \to \infty$ if we have a fixed Monte Carlo sample size. Intuitively this appears unlikely. For efficiency we will need $\varepsilon \to 0$ as $n \to \infty$, and this corresponds to an increasingly strict condition for acceptance. Thus we may imagine that the acceptance probability will necessarily tend to zero as n increases, and thus we will need an increasing Monte Carlo sample size to compensate for this.

However our results show that Algorithm 1 can be efficient if we choose an appropriate proposal distribution. The proposal distribution needs to have a suitable scale and location and have appropriately heavy tails. If we use an appropriate proposal distribution and have a summary statistic of the same dimension as the parameter vector then the posterior mean of approximate Bayesian computation is asymptotically unbiased with a variance that is $1 + O(1/N)$ times that of the estimator maximising the likelihood of the summary statistic. This is similar to asymptotic results for indirect inference (Gouriéroux & Ronchetti, 1993; Heggland & Frigessi, 2004). Our results also lend theoretical support to methods that choose the bandwidth indirectly by specifying the proportion of samples that are accepted, as this leads to a bandwidth which is of the optimal order in n.

We first prove a Bernstein-von Mises type theorem for the posterior mean of approximate Bayesian computation. This is a non-standard convergence result, as it is based on the partial information contained in the summary statistics. For related convergence results see Clarke & Ghosh (1995) and Yuan & Clarke (2004), though these do not consider the case when the dimension of the summary statistic is larger than that of the parameter. Dealing with this case introduces extra challenges.

Our convergence result for the posterior mean of approximate Bayesian computation has practically important consequences. It shows that any d-dimensional summary with $d > p$ can be projected to a p-dimensional summary statistic without any loss of information. Furthermore it shows that using a summary statistic of dimension $d > p$ can lead to an increased bias, so the asymptotic variance can be reduced if the optimal p-dimensional projected summary is used instead. If a d-dimensional summary is used, with $d > p$, it suggests choosing the variance of the kernel to match the variance of the summary statistics.

This paper adds to a growing literature on the theoretical properties of approximate Bayesian computation. Initial results focussed on comparing the bias of approximate Bayesian computation to the Monte Carlo error, and how these depend on the choice of ε. The convergence rate of the bias is shown to be $O(\varepsilon^2)$ in various settings (e.g. Barber et al., 2015). This can then be used to consider how the choice of ε should depend on the Monte Carlo sample size so as to balance bias and Monte Carlo variability (Blum, 2010; Barber et al., 2015; Biau et al., 2015). There has also been work on consistency of approximate Bayesian computation estimators. Marin et al. (2014) considers consistency when performing model choice and Frazier et al. (2016) considers consistency for parameter estimation. The latter work, which appeared after the first version of this paper, includes a similar result on the asymptotic normality of the posterior mean to our Theorem 1, albeit under different conditions. More interestingly, Frazier et al. (2016) also give results on the asymptotic form of the posterior obtained using approximate Bayesian computation. This

shows that for many implementations of approximate Bayesian computation, the posterior will over-estimate the uncertainty in the parameter estimate that it gives.

Finally, a number of papers have looked at the choice of summary statistics (e.g. Wegmann et al., 2009; Blum, 2010; Prangle et al., 2014). Our Theorem 1 gives insight into this. As mentioned above, this result shows that, in terms of minimising the asymptotic variance, we should use a summary statistic that is of the same dimension as the number of parameters. In particular it supports the suggestion in Fearnhead & Prangle (2012) of having one summary per parameter, with that summary approximating the maximum likelihood estimator for that parameter.

2. NOTATION AND SET-UP

Denote the data by $Y_{\text{obs}} = (y_{\text{obs},1}, \ldots, y_{\text{obs},n})$, where n is the sample size, and each observation, $y_{\text{obs},i}$, can be of arbitrary dimension. We make no assumption directly on the data, but make assumptions on the distribution of the summary statistics. We consider the asymptotics as $n \to \infty$, and denote the density of Y_{obs} by $f_n(y \mid \theta)$, where $\theta \in \mathcal{P} \subset \mathbb{R}^p$. We let θ_0 denote the true parameter value, and $\pi(\theta)$ its prior distribution. For a set A, let A^c be its complement with respect to the whole space.

We assume that θ_0 is in the interior of the parameter space, and that the prior is differentiable in a neighbourhood of the true parameter:

CONDITION 1. *There exists some $\delta_0 > 0$, such that $\mathcal{P}_0 = \{\theta : |\theta - \theta_0| < \delta_0\} \subset \mathcal{P}$, $\pi(\theta) \in C^1(\mathcal{P}_0)$ and $\pi(\theta_0) > 0$.*

To implement approximate Bayesian computation we will use a d-dimensional summary statistic, $s_n(Y) \in \mathbb{R}^d$; for example a vector of sample means of appropriately chosen functions. We assume that $s_n(Y)$ has a density function, which depends on n, and we denote this by $f_n(s \mid \theta)$. We will use the shorthand S_n to denote the random variable with density $f_n(s \mid \theta)$. In approximate Bayesian computation we use a kernel, $K(x)$, with $\max_x K(x) = 1$, and a bandwidth $\varepsilon > 0$. As we vary n we will often wish to vary ε, and in these situations denote the bandwidth by ε_n. For Algorithm 1 we require a proposal distribution, $q_n(\theta)$, and allow for this to depend on n. We assume the following conditions on the kernel, which are satisfied by all commonly used kernels,

CONDITION 2. *The kernel satisfies (i) $\int vK(v)\,dv = 0$; (ii) $\int \prod_{k=1}^{l} v_{i_k} K(v)\,dv < \infty$ for any coordinates $(v_{i_1}, \ldots, v_{i_l})$ of v and $l \leq p + 6$; (iii) $K(v) \propto \overline{K}(\|v\|_\Lambda^2)$ where $\|v\|_\Lambda^2 = v^T \Lambda v$ and Λ is a positive-definite matrix, and $\overline{K}(v)$ is a decreasing function of $\|v\|_\Lambda$; (iv) $K(v) = O(e^{-c_1 \|v\|_\Lambda^{\alpha_1}})$ for some $\alpha_1 > 0$ and $c_1 > 0$ as $\|v\|_\Lambda \to \infty$.*

For a real function $g(x)$ denote its kth partial derivative at $x = x_0$ by $D_{x_k} g(x_0)$, the gradient function by $D_x g(x_0)$ and the Hessian matrix by $H_x g(x_0)$. To simplify the notations, D_{θ_k}, D_θ and H_θ are written as D_k, D and H respectively. For a series x_n we use the notation that for large enough n, $x_n = \Theta(a_n)$ if there exists constants m and M such that $0 < m < |x_n/a_n| < M < \infty$, and $x_n = \Omega(a_n)$ if $|x_n/a_n| \to \infty$. For two square matrices A and B, we say $A \leq B$ if $B - A$ is semi-positive definite, and $A < B$ if $B - A$ is positive definite.

Our theory will focus on estimates of some function, $h(\theta)$, of θ, which satisfies differentiability and moment conditions that will control the remainder terms in a Taylor-expansions.

CONDITION 3. *The kth coordinate of $h(\theta)$, $h_k(\theta)$, satisfies (i) $h_k(\theta) \in C^1(\mathcal{P}_0)$; (ii) $D_k h(\theta_0) \neq 0$; and (iii) $\int h_k(\theta)^2 \pi(\theta)\,d\theta < \infty$.*

The asymptotic results presuppose a central limit theorem for the summary statistic.

CONDITION 4. *There exists a sequence a_n, with $a_n \to \infty$ as $n \to \infty$, a d-dimensional vector $s(\theta)$ and a $d \times d$ matrix $A(\theta)$, such that for all $\theta \in \mathcal{P}_0$,*

$$a_n\{S_n - s(\theta)\} \to N\{0, A(\theta)\}, \quad n \to \infty,$$

with convergence in distribution. We also assume that $s_{\text{obs}} \to s(\theta_0)$ in probability. Furthermore, (i) $s(\theta) \in C^1(\mathcal{P}_0)$ and $A(\theta) \in C^1(\mathcal{P}_0)$, and $A(\theta)$ is positive definite for $\theta \in \mathcal{P}_0$; (ii) for any $\delta > 0$ there exists a $\delta' > 0$ such that $\|s(\theta) - s(\theta_0)\| > \delta'$ for all θ satisfying $\|\theta - \theta_0\| > \delta$; (iii) $I(\theta) = Ds(\theta)^T A^{-1}(\theta) Ds(\theta)$ has full rank at $\theta = \theta_0$.

Under Condition 4, a_n is the rate of convergence in the central limit theorem. If the data are independent and identically distributed, and the summaries are sample means of functions of the data or of quantiles, then $a_n = n^{1/2}$. In most applications the data will be dependent, but if summaries are sample means (Wood, 2010), quantiles (Peters et al., 2011; Allingham et al., 2009; Blum & François, 2010) or linear combinations thereof (Fearnhead & Prangle, 2012) then a central limit theorem will often still hold, though a_n may increase more slowly than $n^{1/2}$.

Part (ii) of Condition 4 is required for the true parameter to be identifiable given only the summary of data. The asymptotic variance of the summary-based maximum likelihood estimator for θ is $I^{-1}(\theta_0)/a_n^2$. Condition (iii) ensures that this variance is valid at the true parameter.

We next require a condition that controls the difference between $f_n(s \mid \theta)$ and its limiting distribution for $\theta \in \mathcal{P}_0$. Let $N(x; \mu, \Sigma)$ be the normal density at x with mean μ and variance Σ. Define $\widetilde{f}_n(s \mid \theta) = N\{s; s(\theta), A(\theta)/a_n^2\}$ and the standardized random variable $W_n(s) = a_n A(\theta)^{-1/2}\{s - s(\theta)\}$. Let $\widetilde{f}_{W_n}(w \mid \theta)$ and $f_{W_n}(w \mid \theta)$ be the density of $W_n(s)$ when $s \sim \widetilde{f}_n(s \mid \theta)$ and $f_n(s \mid \theta)$ respectively. The condition below requires that the difference between $f_{W_n}(w \mid \theta)$ and its Edgeworth expansion $\widetilde{f}_{W_n}(w \mid \theta)$ is $o(a_n^{-2/5})$ and can be bounded by a density with exponentially decreasing tails. This is weaker than the standard requirement, $o(a_n^{-1})$, for the remainder in the Edgeworth expansion.

CONDITION 5. *There exists α_n satisfying $\alpha_n / a_n^{2/5} \to \infty$ and a density $r_{\max}(w)$ satisfying Condition 2 (ii)-(iii) where $K(v)$ is replaced with $r_{\max}(w)$, such that $\sup_{\theta \in \mathcal{P}_0} \alpha_n |f_{W_n}(w \mid \theta) - \widetilde{f}_{W_n}(w \mid \theta)| \leq c_3 r_{\max}(w)$ for some positive constant c_3.*

The following condition further assumes that $f_n(s \mid \theta)$ has exponentially decreasing tails with rate uniform in the support of $\pi(\theta)$.

CONDITION 6. *The following statements hold: (i) $r_{\max}(w)$ satisfies Condition 2 (iv); and (ii) $\sup_{\theta \in \mathcal{P}_0^c} f_{W_n}(w \mid \theta) = O(e^{-c_2 \|w\|^{\alpha_2}})$ as $\|w\| \to \infty$ for some positive constants c_2 and α_2, and $A(\theta)$ is bounded in $\mathcal{P}$.*

3. POSTERIOR MEAN ASYMPTOTICS

We first ignore any Monte Carlo error, and focus on the ideal estimator from approximate Bayesian computation. This is the posterior mean, h_{ABC}, where

$$h_{\text{ABC}} = E_{\pi_{\text{ABC}}}\{h(\theta) \mid s_{\text{obs}}\} = \int h(\theta) \pi_{\text{ABC}}(\theta \mid s_{\text{obs}}, \varepsilon_n).$$

This estimator depends on ε_n, but we suppress this from the notation. As an approximation to the true posterior mean, $E\{h(\theta) \mid Y_{\text{obs}}\}$, h_{ABC} contains errors from the choice of the bandwidth ε_n and summary statistic s_{obs}.

To understand the effect of these two sources of error, we derive results for the asymptotic distributions of h_{ABC} and the likelihood-based estimators, including the summary-based maximum likelihood estimator and the summary-based posterior mean, where we consider randomness solely due to the randomness of the data. Let $T_{\mathrm{obs}} = a_n A(\theta_0)^{-1/2}\{s_{\mathrm{obs}} - s(\theta_0)\}$.

THEOREM 1. *Assume Conditions 1–6.*

(i) Let $\hat{\theta}_{\mathrm{MLES}} = \mathit{argmax}_{\theta\in\mathcal{P}} \log f_n(s_{\mathrm{obs}} \mid \theta)$. *For* $h_s = h(\hat{\theta}_{\mathrm{MLES}})$ *or* $E\{h(\theta) \mid s_{\mathrm{obs}}\}$,

$$a_n\{h_s - h(\theta_0)\} \to N\{0, Dh(\theta_0)^T I^{-1}(\theta_0) Dh(\theta_0)\}, \quad n \to \infty,$$

with convergence in distribution.

(ii) Define $c_\infty = \lim_{n\to\infty} a_n\varepsilon_n$. *Let* Z *be the weak limit of* T_{obs}, *which has a standard normal distribution, and* $R(c_\infty, Z)$ *be a random vector with mean zero that is defined in the Supplementary Material. If* $\varepsilon_n = o(a_n^{-3/5})$, *then*

$$a_n\{h_{\mathrm{ABC}} - h(\theta_0)\} \to Dh(\theta_0)^T\{I(\theta_0)^{-1/2}Z + R(c_\infty, Z)\}, \quad n \to \infty,$$

with convergence in distribution. If either (i) $\varepsilon_n = o(a_n^{-1})$; *(ii)* $d = p$; *or (iii) the covariance matrix of* $K(v)$ *is proportional to* $A(\theta_0)$; *then* $R(c_\infty, Z) = 0$. *For other cases, the variance of* $I(\theta_0)^{-1/2}Z + R(c_\infty, Z)$ *is no less than* $I^{-1}(\theta_0)$.

Theorem 1 (i) shows the validity of posterior inference based on the summary statistics. Regardless of the sufficiency and dimension of s_{obs}, the posterior mean based on the summary statistics is consistent and asymptotically normal with the same variance as the summary-based maximum likelihood estimator.

Denote the bias of approximate Bayesian computation, $h_{\mathrm{ABC}} - E\{h(\theta) \mid s_{\mathrm{obs}}\}$, by $\mathrm{bias}_{\mathrm{ABC}}$. The choice of bandwidth impacts the size of the bias. Theorem 1 (ii) indicate two regimes for the bandwidth for which the posterior mean of approximate Bayesian computation has good properties.

The first case is when ε_n is $o(1/a_n)$. For this regime the posterior mean of approximate Bayesian computation always has the same asymptotic distribution as that of the true posterior given the summaries. The other case is when ε_n is $o(a_n^{-3/5})$ but not $o(n^{-1})$. We obtain the same asymptotic distribution if either $d = p$ or we choose the kernel variance to be proportional to the variance of the summary statistics. In general for this regime of ε_n, h_{ABC} will be less efficient than the summary-based maximum likelihood estimator.

When $d > p$, Theorem 1 (ii) shows that $\mathrm{bias}_{\mathrm{ABC}}$ is non-negligible and can increase the asymptotic variance. This is because the leading term of $\mathrm{bias}_{\mathrm{ABC}}$ is proportional to the average of $v = s - s_{\mathrm{obs}}$, the difference between the simulated and observed summary statistics. If $d > p$, the marginal density of v is generally asymmetric, and thus is no longer guaranteed to have a mean of zero. One way to ensure that there is no increase in the asymptotic variance is to choose the variance of the kernel to be proportional to the variance of the summary statistics.

The loss of efficiency we observe in Theorem 1 (ii) for $d > p$ gives an advantage for choosing a summary statistic with $d = p$. The following proposition shows that for any summary statistic of dimension $d > p$ we can find a new p-dimensional summary statistic without any loss of information. The proof of the proposition is trivial and hence omitted.

PROPOSITION 1. *Assume the conditions of Theorem 1. If* $d > p$, *define* $C = Ds(\theta_0)^T A(\theta_0)^{-1}$. *The* p*-dimensional summary statistic* CS_n *has the same information matrix,* $I(\theta)$, *as* Sn. *Therefore the asymptotic variance of* h_{ABC} *based on* Cs_{obs} *is smaller than or equal to that based on* s_{obs}.

Theorem 1 leads to following natural definition.

DEFINITION 1. *Assume that the conditions of Theorem 1 hold. Then the* asymptotic variance *of* h_{ABC} *is*

$$\mathrm{AV}_{h_{\mathrm{ABC}}} = \frac{1}{a_n^2} Dh(\theta_0)^T I_{\mathrm{ABC}}^{-1}(\theta_0) Dh(\theta_0).$$

4. Asymptotic Properties of Rejection and Importance Sampling Algorithm

4·1. *Asymptotic Monte Carlo Error*

We now consider the Monte Carlo error involved in estimating h_{ABC}. Here we fix the data and consider solely the stochasticity of the Monte Carlo algorithm. We focus Algorithm 1. Remember that N is the Monte Carlo sample size. For $i = 1, \ldots, N$, θ_i is the proposed parameter value and w_i is its importance sampling weight. Let ϕ_i be the indicator that is 1 if and only if θ_i is accepted in step 3 of Algorithm 1 and let $N_{\mathrm{acc}} = \sum_{i=1}^{N} \phi_i$ be the number of accepted parameter.

Provided $N_{\mathrm{acc}} \geq 1$ we can estimate h_{ABC} from the output of Algorithm 1 with

$$\widehat{h} = \sum_{i=1}^{N} h(\theta_i) w_i \phi_i \Big/ \sum_{i=1}^{N} w_i \phi_i.$$

Define the acceptance probability:

$$p_{\mathrm{acc},q} = \int q(\theta) \int f_n(s \mid \theta) K_\varepsilon(s - s_{\mathrm{obs}}) \mathrm{d}s \mathrm{d}\theta,$$

and the density of the accepted parameter:

$$q_{\mathrm{ABC}}(\theta \mid s_{\mathrm{obs}}, \varepsilon) = \frac{q_n(\theta) f_{\mathrm{ABC}}(s_{\mathrm{obs}} \mid \theta, \varepsilon)}{\int q_n(\theta) f_{\mathrm{ABC}}(s_{\mathrm{obs}} \mid \theta, \varepsilon)\, d\theta}.$$

Finally, define

$$\begin{aligned} \Sigma_{\mathrm{IS},n} &= E_{\pi_{\mathrm{ABC}}} \left\{ (h(\theta) - h_{\mathrm{ABC}})^2 \frac{\pi_{\mathrm{ABC}}(\theta \mid s_{\mathrm{obs}}, \varepsilon_n)}{q_{\mathrm{ABC}}(\theta \mid s_{\mathrm{obs}}, \varepsilon_n)} \right\}, \\ \Sigma_{\mathrm{ABC},n} &= p_{\mathrm{acc},q_n}^{-1} \Sigma_{\mathrm{IS},n}, \end{aligned} \tag{2}$$

where $\Sigma_{\mathrm{IS},n}$ is the importance sampling variance with π_{ABC} as the target density and q_{ABC} as the proposal density. Note that p_{acc,q_n} and $\Sigma_{\mathrm{IS},n}$, and hence $\Sigma_{\mathrm{ABC},n}$, depend on s_{obs}.

Standard results give the following asymptotic distribution of $\widehat{h}$.

PROPOSITION 2. *For a given* n *and* s_{obs}*, if* h_{ABC} *and* $\Sigma_{\mathrm{ABC},n}$ *are finite, then*

$$N^{1/2}(\widehat{h} - h_{\mathrm{ABC}}) \to N(0, \Sigma_{\mathrm{ABC},n}),$$

in distribution as $N \to \infty$.

This proposition motivates the following definition.

DEFINITION 2. *For a given* n *and* s_{obs}*, assume that the conditions of Proposition 2 hold. Then the asymptotic Monte Carlo variance of* $\widehat{h}$ *is*

$$\mathrm{MCV}_{\widehat{h}} = \frac{1}{N} \Sigma_{\mathrm{ABC},n}.$$

4·2. *Asymptotic efficiency*

We have defined the asymptotic variance as $n \to \infty$ of h_{ABC}, and the asymptotic Monte Carlo variance, as $N \to \infty$ of $\widehat{h}$. The error of h_{ABC} when estimating $h(\theta_0)$ and the Monte Carlo error of $\widehat{h}$ when estimating h_{ABC} are independent, which suggests the following definition.

DEFINITION 3. *Assume the conditions of Theorem 1, and that h_{ABC} and $\Sigma_{\mathrm{ABC},n}$ are bounded in probability for any n. Then the* asymptotic variance *of $\widehat{h}$ is*

$$\mathrm{AV}_{\widehat{h}} = \frac{1}{a_n^2} h(\theta_0)^T I_{\mathrm{ABC}}^{-1}(\theta_0) D h(\theta_0) + \frac{1}{N} \Sigma_{\mathrm{ABC},n}.$$

We can interpret the asymptotic variance of $\widehat{h}$ as a first-order approximation to the variance of our Monte Carlo estimator for both large n and N. We wish to investigate the properties of this asymptotic variance, for large but fixed N, as $n \to \infty$. The asymptotic variance itself depends on n, and we would hope it would tend to zero as n increases. Thus we will study the ratio of $\mathrm{AV}_{\widehat{h}}$ to $\mathrm{AV}_{\mathrm{MLES}}$, where, by Theorem 1, the latter is $a_n^{-2} h(\theta_0)^T I^{-1}(\theta_0) D h(\theta_0)$. This ratio measures the efficiency of our Monte Carlo estimator relative to the maximum likelihood estimator based on the summaries; it quantifies the loss of efficiency from using a non-zero bandwidth and a finite Monte Carlo sample size.

We will consider how this ratio depends on the choice of ε_n and $q_n(\theta)$. Thus we introduce the following definition:

DEFINITION 4. *For a choice of ε_n and $q_n(\theta)$, we define the asymptotic efficiency of $\widehat{h}$ as*

$$\mathrm{AE}_{\widehat{h}} = \lim_{n \to \infty} \frac{\mathrm{AV}_{\mathrm{MLES}}}{\mathrm{AV}_{\widehat{h}}}.$$

If this limiting value is zero, we say that $\widehat{h}$ is asymptotically inefficient.

We will investigate the asymptotic efficiency of $\widehat{h}$ under the assumption of Theorem 1 that $\varepsilon_n = o(a_n^{-3/5})$. We shall see that the convergence rate of the importance sampling variance $\Sigma_{\mathrm{IS},n}$ depends on how large ε_n is relative to a_n, and so we further define $a_{n,\varepsilon} = a_n$ if $\lim_{n\to\infty} a_n \varepsilon_n < \infty$ and $a_{n,\varepsilon} = \varepsilon_n^{-1}$ otherwise.

If our proposal distribution in Algorithm 1 is either the prior or the posterior, then the estimator is asymptotically inefficient:

THEOREM 2. *Assume the conditions of Theorem 1.*

(i) If $q_n(\theta) = \pi(\theta)$, $p_{\mathrm{acc},q_n} = \Theta_p(\varepsilon_n^d a_{n,\varepsilon}^{d-p})$ and $\Sigma_{\mathrm{IS},n} = \Theta_p(a_{n,\varepsilon}^{-2})$.
(ii) If $q_n(\theta) = \pi_{\mathrm{ABC}}(\theta \mid s_{\mathrm{obs}}, \varepsilon_n)$, $p_{\mathrm{acc},q_n} = \Theta_p(\varepsilon_n^d a_{n,\varepsilon}^{d})$ and $\Sigma_{\mathrm{IS},n} = \Theta_p(a_{n,\varepsilon}^{p})$.

In both cases $\widehat{h}$ is asymptotically inefficient.

The result in part (ii) shows a difference from standard importance sampling settings, where using the target distribution as the proposal leads to an estimator with no Monte Carlo error.

The estimator $\widehat{h}$ is asymptotically inefficient because the Monte Carlo variance decays more slowly than $1/a_n^2$ as $n \to \infty$. However this is caused by different factors in each case.

To see this, consider the acceptance probability of a value of θ and corresponding summary s_n simulated in one iteration of Algorithm 1. This acceptance probability depends on

$$\frac{s_n - s_{\mathrm{obs}}}{\varepsilon_n} = \frac{1}{\varepsilon_n} \left[\{s_n - s(\theta)\} + \{s(\theta) - s(\theta_0)\} + \{s(\theta_0) - s_{\mathrm{obs}}\} \right], \tag{3}$$

where $s(\theta)$, defined in Condition 4, is the limiting value of s_n as $n \to \infty$ if data is sampled from the model for parameter value θ. By Condition 4, the first and third bracketed terms within the square brackets on the right-hand side are $O_p(a_n^{-1})$. If we sample θ from the prior the middle term is $O_p(1)$, and thus (3) will blow up as ε_n goes to zero. Hence $p_{\text{acc},\pi}$ goes to zero as ε_n goes to zero, which causes the estimate to be inefficient. If we sample from the posterior, then by Theorem 1 we expect the middle term to also be $O_p(a_n^{-1})$. Hence (3) is well behaved as $n \to \infty$, and $p_{\text{acc},\pi}$ is bounded away from zero, provided either $\varepsilon_n = \Theta(a_n^{-1})$ or $\varepsilon_n = \Omega(a_n^{-1})$.

However, if we use $\pi_{\text{ABC}}(\theta \mid s_{\text{obs}}, \varepsilon_n)$ as a proposal distribution, the estimates are still inefficient due to an increasing variance of the importance weights: as n increases the proposal distribution is more and more concentrated around θ_0, while π does not change.

4·3. *Efficient Proposal Distributions*

Consider proposing the parameter value from a location-scale family. That is our proposal is of the form $\sigma_n \Sigma^{1/2} X + \mu_n$, where $X \sim q(\cdot)$, $E(X) = 0$ and $\text{var}(X) = I_p$. This defines a general form of proposal density, where the center, μ_n, the scale rate, σ_n, the scale matrix, Σ and the base density, $q(\cdot)$, all need to be specified. We will give conditions under which such a proposal density results in estimators that are efficient.

Our results are based on an expansion of $\pi_{\text{ABC}}(\theta \mid s_{\text{obs}}, \varepsilon_n)$. Consider the rescaled random variables $t = a_{n,\varepsilon}(\theta - \theta_0)$ and $v = \varepsilon_n^{-1}(s - s_{\text{obs}})$. Recall that $T_{\text{obs}} = a_n A(\theta_0)^{-1/2}\{s_{\text{obs}} - s(\theta_0)\}$. Define an unnormalised joint density of t and v as

$$g_n(t, v; \tau) = \begin{cases} N\Big[\{Ds(\theta_0) + \tau\}t; a_n\varepsilon_n v + A(\theta_0)^{1/2}T_{\text{obs}}, A(\theta_0)\Big]K(v), & a_n\varepsilon_n \to c < \infty, \\ N\Big[\{Ds(\theta_0) + \tau\}t; v + \frac{1}{a_n\varepsilon_n}A(\theta_0)^{1/2}T_{\text{obs}}, \frac{1}{a_n^2\varepsilon_n^2}A(\theta_0)\Big]K(v), & a_n\varepsilon_n \to \infty, \end{cases}$$

and further define $g_n(t; \tau) = \int g_n(t, v; \tau)\, dv$. For large n, and for the rescaled variable t, the leading term of π_{ABC} is then proportional to $g_n(t; 0)$. For both limits of $a_n\varepsilon_n$, $g_n(t; \tau)$ is a continuous mixture of normal densities with the kernel density determining the mixture weights.

Our main theorem requires conditions on the proposal density. First, that $\sigma_n = a_{n,\varepsilon}^{-1}$ and $c_\mu = \sigma_n^{-1}(\mu_n - \theta_0)$ is $O_p(1)$. This ensures that under the scaling of t, as $n \to \infty$, the proposal is not increasingly over-dispersed compared to the target density, and the acceptance probability can be bounded away from zero. Second, that the proposal distribution is sufficiently heavy-tailed:

CONDITION 7. *There exist positive constants m_1 and m_2 satisfying $m_1^2 I_p < Ds(\theta_0)^T Ds(\theta_0)$ and $m_2 I_d < A(\theta_0)$, $\alpha \in (0, 1)$, $\gamma \in (0, 1)$ and $c \in (0, \infty)$, such that for any $\lambda > 0$,*

$$\sup_{t \in \mathbb{R}^p} \frac{N(t; 0, m_1^{-2} m_2^{-2} \gamma^{-1})}{q\{\Sigma^{-1/2}(t - c)\}} < \infty, \ \sup_{t \in \mathbb{R}^p} \frac{\overline{K}^\alpha(\|\lambda t\|^2)}{q\{\Sigma^{-1/2}(t - c)\}} < \infty, \ \sup_{t \in \mathbb{R}^p} \frac{\overline{r}_{\max}(\|m_1 m_2 \gamma^{1/2} t\|^2)}{q\{\Sigma^{-1/2}(t - c)\}} < \infty,$$

where $\overline{r}_{\max}(\cdot)$ satisfies $r_{\max}(v) = \overline{r}_{\max}(\|v\|_\Lambda^2)$, and for any random series c_n in $\mathbb{R}^p$ satisfying $c_n = O_p(1)$,

$$\sup_{t \in \mathbb{R}^p} \frac{q(t)}{q(t + c_n)} = O_p(1).$$

If we choose $\varepsilon_n = \Theta(a_n^{-1})$, the Monte Carlo importance sampling variance for the accepted parameter values is $\Theta(a_n^{-2})$, and has the same order as the variance of summary-based maximum likelihood estimator.

THEOREM 3. *Assume the conditions of Theorem 1. If the proposal density* $q_n(\theta)$ *is*

$$\beta\pi(\theta) + (1-\beta)\frac{1}{\sigma_n^p|\Sigma|^{1/2}}q\{\sigma_n^{-1}\Sigma^{-1/2}(\theta-\mu_n)\},$$

where $\beta \in (0,1)$, $q(\cdot)$ *and* Σ *satisfy Condition 7,* $\sigma_n = a_{n,\varepsilon}^{-1}$ *and* c_μ *is* $O_p(1)$, *then* $p_{\mathrm{acc},q_n} = \Theta_p(\varepsilon_n^d a_{n,\varepsilon}^d)$ *and* $\Sigma_{\mathrm{IS},n} = O_p(a_{n,\varepsilon}^{-2})$. *Then if* $\varepsilon_n = \Theta(a_n^{-1})$, $\mathrm{AE}_{\widehat{h}} = \Theta_p(1)$.

Furthermore, if $d = p$, $\mathrm{AE}_{\widehat{h}} = 1 - K/(N+K)$ *for some constant* K.

The mixture with $\pi(\theta)$ here is to control the importance weight in the tail area (Hesterberg, 1995). It is not clear whether this is needed in practice, or is just a consequence of the approach taken in the proof.

Theorem 3 shows that with a good proposal distribution, if the acceptance probability is bounded away from zero as n increases, the threshold ε_n will have the preferred rate $\Theta(a_n^{-1})$. This supports using the acceptance rate to choose the threshold based on aiming for an appropriate proportion of acceptances (Del Moral et al., 2012; Biau et al., 2015).

In practice, σ_n and μ_n need to be adaptive to the observations since they depend on n. For $q(\cdot)$ and Σ, the following proposition gives a practical suggestion that satisfies Condition 7. Let $T(\cdot;\gamma)$ be the multivariate t density with degree of freedom γ. The following result says that it is theoretically valid to choose any Σ if a t distribution is chosen as the base density.

PROPOSITION 3. *Condition 7 is satisfied for* $q(\theta) = T(\theta;\gamma)$ *with any* $\gamma > 0$ *and any* Σ.

Proof. The first part of Condition 7 follows as the t-density is heavy tailed relative to the normal density, $\overline{K}(\cdot)$ and $\overline{r}_{\max}(\cdot)$. The second part can be verified easily. □

4·4. *Iterative Importance Sampling*

Taken together, Theorem 3 and Proposition 3 suggest proposing from the mixture of $\pi(\theta)$ and a t distribution with the scale matrix and center approximating those of $\pi_{\mathrm{ABC}}(\theta)$. We suggest the following iterative procedure, similar in spirit to that of Beaumont et al. (2009).

Algorithm 2. Iterative importance sampling approximate Bayesian computation

Input a mixture weight β, a sequence of acceptance rates $\{p_k\}$, and a location-scale family. Set $q_1(\theta) = \pi(\theta)$.

For $k = 1, \ldots, K$:

1. Run Algorithm 1 with simulation size N_0, proposal density $\beta\pi(\theta) + (1-\beta)q_k(\theta)$ and acceptance rate p_k, and record the bandwidth ε_k.
2. If $\varepsilon_{k-1} - \varepsilon_k$ is smaller than some positive threshold, stop. Otherwise, let μ_{k+1} and Σ_{k+1} be the empirical mean and variance matrix of the weighted sample from step 1, and let $q_{k+1}(\theta)$ be the density with centre μ_{k+1} and variance matrix $2\Sigma_{k+1}$.
3. If $q_k(\theta)$ is close to $q_{k+1}(\theta)$ or $K = K_{\max}$, stop. Otherwise, return to step 1.

After the iteration stops at the K_{th} step, run Algorithm 1 with the proposal density $\beta\pi(\theta) + (1-\beta)q_{K+1}(\theta)$, $N - KN_0$ simulations and p_{K+1}.

In this algorithm, N is the number of simulations allowed by the computing budget, $N_0 < N$ and $\{p_k\}$ is a sequence of acceptance rates, which we use to choose the bandwidth. The maximum value $K_{\max}$ of K is set such that $K_{\max}N_0 = N/2$. The rule for choosing the new proposal distribution is based on approximating the mean and variance of the density proportional to $\pi(\theta)f_{\mathrm{ABC}}(s_{\mathrm{obs}} \mid \theta, \varepsilon)^{1/2}$, which is optimal (Fearnhead & Prangle, 2012). It can be shown that these two moments are approximately equal to the mean and twice the variance of $\pi_{\mathrm{ABC}}(\theta)$ re-

spectively. For the mixture weight, β, we suggest a small value, and use 0.05 in the simulation study below.

5. Numerical Examples

5·1. *Gaussian Likelihood with Sample Quantiles*

This examples illustrates the results in Section 3 with an analytically tractable problem. Assume the observations $Y_{\rm obs} = (y_1, \ldots, y_n)$ follow the univariate normal distribution $N(\mu, \sigma)$ with true parameter values $(1, 2^{1/2})$. Consider estimating the unknown parameter (μ, σ) with the uniform prior in the region $[-10, 10] \times [0, 10]$ using Algorithm 1. The summary statistic is $(e^{\widehat{q}_{\alpha_1}/2}, \ldots, e^{\widehat{q}_{\alpha_d}/2})$ where $\widehat{q}_\alpha$ is the sample quantile of $Y_{\rm obs}$ for probability α.

The results for data size $n = 10^5$ are presented. Smaller sizes from 10^2 to 10^4 show similar patterns. The probabilities $\alpha_1, \ldots, \alpha_d$ for calculating quantiles are selected with equal intervals in $(0, 1)$, and $d = 2, 9$ and 19 were tested. In order to investigate the Monte Carlo error-free performance, N is chosen to be large enough that the Monte Carlo errors were negligible. We compare the performances of $\theta_{\rm ABC}$, the maximum likelihood estimator based on the summary statistics and the maximum likelihood estimator based on the full dataset. Since the dimension reduction matrix C in Proposition 1 can be obtained analytically, the performance of $\theta_{\rm ABC}$ using the original d-dimension summary is compared with that using the 2-dimension summary. The results of mean square error are presented in Figure 1.

The phenomena implied by Theorem 1 and Proposition 1 can be seen in this example, together with the limitations of these results. First, $E\{h(\theta) \mid s_{\rm obs}\}$, equivalent to $\theta_{\rm ABC}$ with small enough ε, and the maximum likelihood estimator based on the same summaries, have similar accuracy. Second, when ε is small, the mean square error of $\theta_{\rm ABC}$ is the same as that of the maximum likelihood based on the summary. When ε becomes larger, for $d > 2$ the mean square error increases more quickly than for $d = 2$. This corresponds to the impact of the additional bias when $d > p$.

For all cases, the two-dimensional summary obtained by projecting the original d summaries is, for small ε, as accurate as the maximum likelihood estimator given the original d summaries. This indicates that the lower-dimensional summary contains the same information as the original one. For larger ε, the performance of the reduced-dimension summaries is not stable, and is in fact worse than the original summaries for estimating μ. This deterioration is caused by the bias of $\theta_{\rm ABC}$, which for larger ε, is dominated by higher order terms in ε which could be ignored in our asymptotic results.

5·2. *Stochastic Volatility with AR(1) Dynamics*

We consider a stochastic volatility model from Sandmann & Koopman (1998) for the demeaned returns of a portfolio. Denote this return for the tth time-period as y_t. Then

$$x_t = \phi x_{t-1} + \eta_t,\ \eta_t \sim N(0, \sigma_\eta^2);\ \ y_t = \overline{\sigma} e^{x_t/2} \xi_n,\ \xi_t \sim N(0, 1),$$

where η_t and ξ_t are independent, and x_t is a latent state that quantifies the level of volatility for time-period t. By the transformation $y_t^* = \log y_t^2$ and $\xi_t^* = \log \xi_t^2$, the observation equation in the state-space model can be transformed to

$$y_n^* = 2 \log \overline{\sigma} + x_n + \xi_n^*,\ \exp(\xi_n^*) \sim \chi_1^2, \tag{4}$$

which is linear and non-Gaussian.

Approximate Bayesian computation can be used to obtain an off-line estimator for the unknown parameters of this model. Here we illustrate the effectiveness of iteratively choosing the

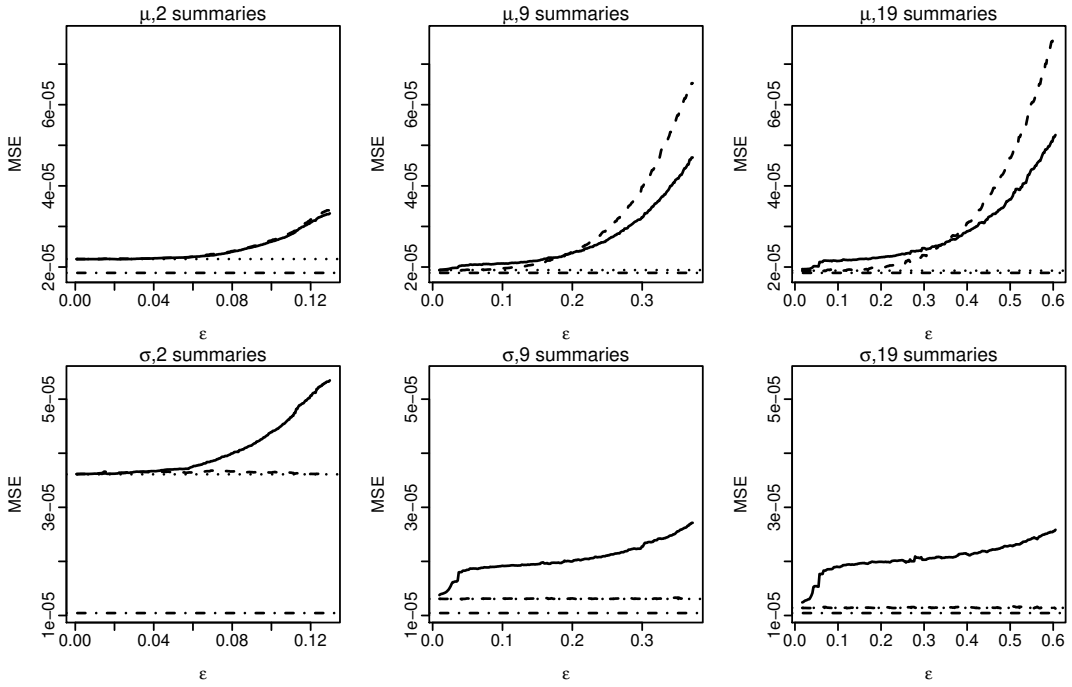


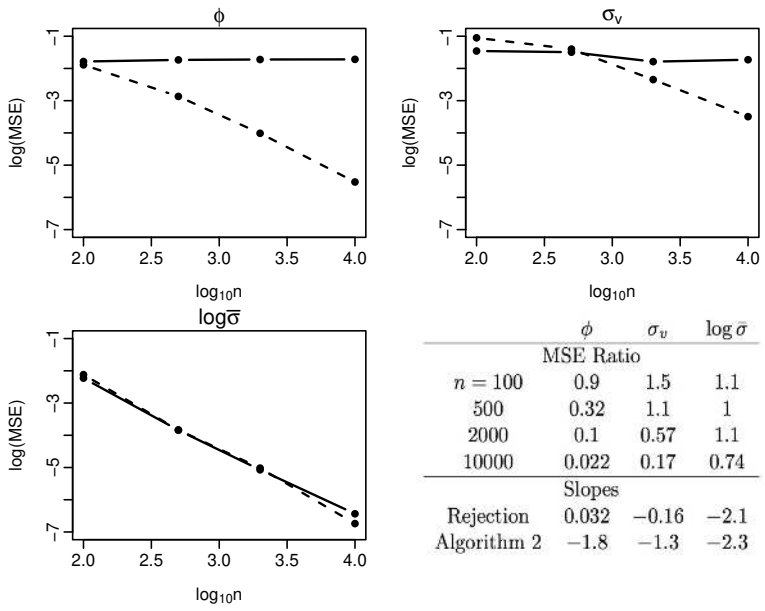
Fig. 1: Illustration of results in Section 3. Mean square errors of point estimates for 200 data sets are reported. Point estimates compared include θ_{ABC} using the original summary statistic (solid) and the transformed summary statistic (dashed), the dimension of which is reduced to 2 according to Proposition 1, the maximum likelihood estimates based on the original summary statistic (dotted) and the full data set (dash-dotted).

importance proposal for large n by comparing with rejection sampling. In the iterative algorithm, a t distribution with 5 degrees of freedom is used to construct q_k.

Consider estimating the parameter $(\phi, \sigma_\eta, \log \overline{\sigma})$ under a uniform prior in the region $[0,1) \times [0.1,3] \times [-10,-1]$. The setting with the true parameter $(\phi, \sigma_\eta, \log \overline{\sigma}) = (0.9, 0.675, -4.1)$ is studied. We use a 3-dimensional summary statistic that stores the mean, variance and lag-1 autocovariance of the transformed data. If there were no noise in the state equation for ξ_n^*, then this would be a sufficient statistic of Y^*, and hence is a natural choice for the summary statistic. The uniform kernel is used in the accept-reject step.

We evaluate rejection sampling and iterative importance sampling methods on data of length $n = 100, 500, 2000$ and 10000; and use $N = 40000$ Monte Carlo simulations. For iterative importance sampling, the sequence $\{p_k\}$ has the first five values decreasing linearly from 5% to 1%, and later values being 1%. We further set $N_0 = 2000$, and $K_{\max} = 10$. For the rejection sampler acceptance probabilities of both 5% and 1% were tried and 5% was chosen as it gave better performance. The simulation results are shown in Figure 2.

For all parameters, iterative importance sampling shows increasing advantage over rejection sampling as n increases. For larger n, the iterative procedure obtains a center for proposals closer to the true parameter and a bandwidth that is smaller than those used for rejection sampling. These contribute to the more accurate estimators. It is easy to estimate $\log \bar{\sigma}$, since the expected summary statistic $\widetilde{E}(Y^*)$ is roughly linear in $\log \bar{\sigma}$. Thus iterative importance sampling less of an advantage over rejection sampling when estimating this parameter.



	ϕ	σ_v	$\log \bar{\sigma}$
	MSE Ratio		
$n = 100$	0.9	1.5	1.1
500	0.32	1.1	1
2000	0.1	0.57	1.1
10000	0.022	0.17	0.74
	Slopes		
Rejection	0.032	−0.16	−2.1
Algorithm 2	−1.8	−1.3	−2.3

Fig. 2: Comparisons of rejection (solid) and iterative importance sampling (dashed) versions of approximate Bayesian computation. For each n, the logarithm of the average mean square error across 100 datasets is reported. For each dataset, the Monte Carlo sample size is 40000. Ratios of mean square errors of the two methods are given in the table, and smaller values indicate better performance of iterative importance sampling. For each polyline in the plots, a line is fitted and the slope is reported in the table. Smaller values indicate faster decrease of the mean square error.

6. Discussion

Our results suggest you can obtain efficient estimates using Approximate Bayesian Computation with a fixed Monte Carlo sample size as n increases. Thus the computational complexity of approximate Bayesian computation will just be the complexity of simulating a sample of size n from the underlying model.

Our results on the Monte Carlo accuracy of approximate Bayesian computation considered the importance sampling implementation given in Algorithm 1. If we do not use the uniform kernel, then there is a simple improvement on this algorithm, that absorbs the accept-reject probability within the importance sampling weight. A simple Rao–Blackwellisation argument then shows that this leads to a reduction in Monte Carlo variance. As such, our positive results about the scaling of approximate Bayesian computation with n will immediately apply to this implementation as well.

Similar positive Monte Carlo results are likely to apply to Markov chain Monte Carlo implementations of approximate Bayesian computation. A Markov chain Monte Carlo version will be efficient provided the acceptance probability does not degenerate to zero as n increases. However at stationarity, it will propose parameter values from a distribution close to the approximate Bayesian computation posterior density, and Theorems 2 and 3 suggest that for such a proposal distribution the acceptance probability will be bounded away from zero.

Whilst our theoretical results suggest that point estimates based on approximate Bayesian computation have good properties, they do not suggest that the approximate Bayesian computation posterior is a good approximation to the true posterior. In fact, results by Frazier et al. (2016) show it will over-estimate uncertainty if $\varepsilon_n = O(a_n^{-1})$. However, Li & Fearnhead (2016) show that using regression methods (Beaumont et al., 2002) to post-process approximate Bayesian

computation output can lead to both efficient point estimation and accurate quantification of uncertainty.

ACKNOWLEDGMENT

This work was support by the Engineering and Physical Sciences Research Council.

SUPPLEMENTARY MATERIAL

Proofs of the main results are included in the online supplementary material.

REFERENCES

ALLINGHAM, D., KING, R. A. R. & MENGERSEN, K. L. (2009). Bayesian estimation of quantile distributions. *Statistics and Computing* **19**, 189–201.

BARBER, S., VOSS, J., WEBSTER, M. et al. (2015). The rate of convergence for approximate Bayesian computation. *Electronic Journal of Statistics* **9**, 80–105.

BEAUMONT, M. A. (2010). Approximate Bayesian computation in evolution and ecology. *Annual Review of Ecology, Evolution, and Systematics* **41**, 379–406.

BEAUMONT, M. A., CORNUET, J.-M., MARIN, J.-M. & ROBERT, C. P. (2009). Adaptive approximate Bayesian computation. *Biometrika* **96**, 983–990.

BEAUMONT, M. A., ZHANG, W. & BALDING, D. J. (2002). Approximate Bayesian computation in population genetics. *Genetics* **162**, 2025–2035.

BIAU, G., CÉROU, F. & GUYADER, A. (2015). New insights into approximate Bayesian computation. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques* **51**, 376–403.

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

BLUM, M. G. & FRANÇOIS, O. (2010). Non-linear regression models for approximate Bayesian computation. *Statistics and Computing* **20**, 63–73.

BORTOT, P., COLES, S. G. & SISSON, S. A. (2007). Inference for stereological extremes. *Journal of the American Statistical Association* **102**, 84–92.

CLARKE, B. & GHOSH, J. (1995). Posterior convergence given the mean. *Annals of Statistics* **23**, 2116–2144.

DEL MORAL, P., DOUCET, A. & JASRA, A. (2012). An adaptive sequential Monte Carlo method for approximate Bayesian computation. *Statistics and Computing* **22**, 1009–1020.

DUFFIE, D. & SINGLETON, K. J. (1993). Simulated moments estimation of Markov models of asset prices. *Econometrica* **61**, 929–952.

FEARNHEAD, P. & PRANGLE, D. (2012). Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **74**, 419–474.

FRAZIER, D. T., MARTIN, G. M., ROBERT, C. P. & ROUSSEAU, J. (2016). Asymptotic properties of approximate Bayesian computation. *arXiv:1607.06903* .

GORDON, N., SALMOND, D. & SMITH, A. F. M. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEEE proceedings F - Radar and Signal Processing* **140**, 107–113.

GOURIÉROUX, C. & RONCHETTI, E. (1993). Indirect inference. *Journal of Applied Econometrics* **8**, s85–s118.

HEGGLAND, K. & FRIGESSI, A. (2004). Estimating functions in indirect inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **66**, 447–462.

HESTERBERG, T. (1995). Weighted average importance sampling and defensive mixture distributions. *Technometrics* **37**, 185–194.

ISHIDA, E., VITENTI, S., PENNA-LIMA, M., CISEWSKI, J., DE SOUZA, R., TRINDADE, A., CAMERON, E. & BUSTI, V. (2015). cosmoabc: Likelihood-free inference via population Monte Carlo approximate Bayesian computation. *Astronomy and Computing* **13**, 1–11.

LI, W. & FEARNHEAD, P. (2016). Improved convergence of regression adjusted approximate Bayesian computation. *arXiv:1609.07135* .

MARIN, J.-M., PILLAI, N. S., ROBERT, C. P. & ROUSSEAU, J. (2014). Relevant statistics for Bayesian model choice. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76**, 833–859.

PETERS, G. W., KANNAN, B., LASSCOCK, B., MELLEN, C., GODSILL, S. et al. (2011). Bayesian cointegrated vector autoregression models incorporating alpha-stable noise for inter-day price movements via approximate Bayesian computation. *Bayesian Analysis* **6**, 755–792.

PRANGLE, D., FEARNHEAD, P., COX, M. P., BIGGS, P. J. & FRENCH, N. P. (2014). Semi-automatic selection of summary statistics for ABC model choice. *Statistical Applications in Genetics and Molecular Biology* **13**, 67–82.

PRITCHARD, J. K., SEIELSTAD, M. T., PEREZ-LEZAUN, A. & FELDMAN, M. W. (1999). Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular Biology and Evolution* **16**, 1791–1798.

SANDMANN, G. & KOOPMAN, S. (1998). Estimation of stochastic volatility models via Monte Carlo maximum likelihood. *Journal of Econometrics* **87**, 271–301.

TONI, T., WELCH, D., STRELKOWA, N., IPSEN, A. & STUMPF, M. P. (2009). Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface* **6**, 187–202.

WEGMANN, D., LEUENBERGER, C. & EXCOFFIER, L. (2009). Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood. *Genetics* **182**, 1207–1218.

WOOD, S. N. (2010). Statistical inference for noisy nonlinear ecological dynamic systems. *Nature* **466**, 1102–1104.

YUAN, A. & CLARKE, B. (2004). Asymptotic normality of the posterior given a statistic. *Canadian Journal of Statistics* **32**, 119–137.