Biometrika (0000), **000**, 0, *pp*. 1–20
Printed in Great Britain
Advance Access publication on Day Month Year

Approximate confidence distribution computing: An effective likelihood-free method with statistical guarantees

BY S. THORNTON

Department of Statistics and Biostatistics, Rutgers, The State University of New Jersey New Brunswick, New Jersey 08901 U.S.A.

suzanne.thornton@rutgers.edu

W. LI

Department of Mathematics, Statistics, and Physics, provide current address - Newcastle University, U.K.

w.li@lancaster.ac.uk

AND M. XIE

Department of Statistics and Biostatistics, Rutgers, The State University of New Jersey New Brunswick, New Jersey 08901 U.S.A.

mxie@stat.rutgers.edu

SUMMARY

Approximate Bayesian computing is a powerful likelihood-free method that has grown increasingly popular since early applications in population genetics. However, complications arise in the theoretical justification for Bayesian inference conducted from this method with a non-sufficient summary statistic. In this paper, we seek to re-frame approximate Bayesian computing within a frequentist context and justify its performance by standards set on the frequency coverage rate. In doing so, we develop a new computational technique called *approximate confidence distribution computing*, that yields theoretical support for the use of non-sufficient summary statistics in likelihood-free methods. Furthermore, we demonstrate that approximate confidence distribution computing extends the scope of approximate Bayesian computing to include data-dependent priors without damaging the inferential integrity. This data-dependent prior can be viewed as an initial 'distribution estimate' of the target parameter which is updated with the results of the approximate confidence distribution computing method. A general strategy for constructing an appropriate data-dependent prior is also discussed and is shown to often increase the computing speed while maintaining statistical guarantees. We supplement the theory with simulation studies illustrating the benefits of the confidence distribution method, namely the potential for broader applications than the Bayesian method and the increased computing speed compared to approximate Bayesian computing.

Some key words: Approximate Bayesian computing; Bernstein-von Mises; Confidence distribution; Exact inference; Large sample theory.

1. INTRODUCTION

1·1. *Background*

Approximate Bayesian computing is a likelihood-free method that approximates a posterior distribution while avoiding direct calculation of the likelihood. This procedure originated in population genetics where complex demographic histories yield intractable likelihoods. Since then, approximate Bayesian computing has been applied to many other areas besides the biological sciences including astronomy and finance; cf., e.g., Cameron & Pettitt (2012); Marin et al. (2011); Sisson et al. (2007). Despite its practical popularity in providing a Bayesian solution for complex data problems, the theoretical justification for inference from this method is under-developed and has only recently been explored in statistical literature; cf., e.g., Robinson et al. (2014); Barber et al. (2015); Frazier et al. (2018); Li & Fearnhead (2018b). In this paper, we seek to re-frame the problem within a frequentist setting and help address two weaknesses of approximate Bayesian computing: (1) lack of theoretical justification for Bayesian inference when using a non-sufficient summary statistic and (2) slow computing speed. We propose a novel likelihood-free method as a bridge connecting Bayesian and frequentist inferences and examine it within the context of the existing literature on approximate Bayesian computing.

Rather than deriving a likelihood, approximate Bayesian computing uses the assumption that one may treat simulations as artificial experiments and compare observed and simulated summary statistics. We assume these simulations are observations of some data generating model, M_θ, where $\theta \in \mathcal{P} \subset \mathbb{R}^p$ is unknown. To apply this computing method, we need not have an analytically tractable expression for the likelihood of the data; instead, we need only assume that data may be generated by simulations of this scientific model.

The goal of approximate Bayesian computing is to produce draws from an approximation to a posterior distribution for θ, given one has observed a random sample, $x_{obs} = \{x_1, \ldots, x_n\}$ from some unknown distribution with density $f(x_i; \theta)$. The standard accept-reject version of approximate Bayesian computing proceeds as follows:

Algorithm 1. (Accept-reject approximate Bayesian computing)

1. Simulate $\theta_1, \ldots, \theta_N \sim \pi(\theta)$;
2. For each $i = 1, \ldots, N$, simulate $x^{(i)} = \{x_1^{(i)}, \ldots, x_n^{(i)}\}$ from M_θ;
3. For each $i = 1, \ldots, N$, accept θ_i with probability $K\{\varepsilon^{-1}(s^{(i)} - s_{\text{obs}})\}$, where $s_{\text{obs}} = S_n(x_{obs})$ and $s^{(i)} = S_n(x^{(i)})$.

In the above algorithm, the data is summarized by some low-dimension summary statistic, $S_n(\cdot)$ (e.g., $S_n(\cdot)$ is a mapping from the sample space in $\mathbb{R}^n$ to $\mathcal{S} \subset \mathbb{R}^d$), and some distance metric, defined by the kernel probability $K(\cdot)$ with bandwidth ε, compares it to simulated data. We refer to ε as the *tolerance level* and typically assume it goes to zero. In many cases, ε is required to go to zero at a certain rate of n (cf., e.g., Li & Fearnhead (2018b)), but there are cases in finite sample development in which ε is independent of sample size n. Sometimes the summary statistic, $s_n^{(i)}$, can be directly simulated from M_θ. The underlying distribution from which these N copies or draws of θ are generated is called the *ABC posterior*, with the probability density,

$$\pi_\varepsilon(\theta \mid s_{\text{obs}}) = \frac{\int_{\mathcal{S}} \pi(\theta) \tilde{f}_n(s; \theta) K\{\varepsilon^{-1}(s - s_{\text{obs}})\}\, ds}{\int_{\mathcal{P}\times\mathcal{S}} \pi(\theta) \tilde{f}_n(s; \theta) K\{\varepsilon^{-1}(s - s_{\text{obs}})\}\, ds d\theta}, \tag{1}$$

and corresponding cumulative distribution function denoted $\Pi_\varepsilon(\theta \in A \mid s_{\text{obs}})$. Here $\tilde{f}_n(s; \theta)$ denotes the (typically unknown) density of the summary statistic. We will refer to $\tilde{f}_n(s; \theta)$ as an *s-likelihood* to emphasize that it is not a likelihood in any traditional sense. Since this is a Bayesian

procedure, in addition to the assumption of the existence of a data-generating model, M_θ, Algorithm 1 assumes a prior distribution, $\pi(\cdot)$, on θ. In the absence of prior information, the user may select a flat prior.

A common assertion (see, e.g. Marin et al. (2011)) is that this ABC posterior is close enough to the target posterior distribution, $p(\theta \mid x) \propto \pi(\theta) \prod_{i=1}^{n} f(x_i; \theta)$; however, the quality of the approximation of an ABC posterior to its target posterior distribution depends on the closeness of the tolerance level to zero and, more crucially for our purposes, on the choice of summary statistic. Indeed, we have the following lemma:

Lemma 1. Let $K(\cdot)$ be a symmetric kernel function with $\int uK(u)du = 0$ and $\int \|u\|^2 K(u)du < \infty$. Suppose $\tilde{f}_n(s; \theta)$ has a bounded second-derivative with respective to s. Then

$$\pi_\varepsilon(\theta \mid s_{\text{obs}}) \propto \pi(\theta)\tilde{f}_n(s_{\text{obs}}; \theta) + O(\varepsilon^2). \tag{2}$$

Various versions of this result are known (cf., e.g., Barber et al. (2015) and Li & Fearnhead (2018a)); for completeness, we provide a brief proof of Lemma 1 in the appendix. If the summary statistic is not sufficient, $\tilde{f}_n(s_{\text{obs}}; \theta)$ can be very different from $\prod_{i=1}^{n} f(x_i; \theta)$, in which case the ABC posterior can be a very poor approximation to the target posterior, even if $\varepsilon \to 0$.

Figure 1 provides such an example where we consider random data from a Cauchy distribution with a known scale parameter. Only the data itself is sufficient for the location parameter, θ; therefore, any reasonable choice of summary statistic will not be sufficient. Fig. 1 illustrates that, without sufficiency, the ABC posterior approximations will never converge to the targeted posterior distribution so the approximations to the target posterior can be quite poor. Specifically, Fig. 1 shows two applications: one using the sample mean and the other using the sample median as the summary statistic, each with a flat prior on the unknown location parameter. In neither case is the ABC posterior the same as the targeted posterior regardless of sample size or the rate of $\varepsilon \to 0$, including the rate typically required in the existing literature; cf., Li & Fearnhead (2018b).

For this reason, inference from the ABC posterior can produce misleading results within a Bayesian context when the summary statistic used is not sufficient. Questions arise such as, if an ABC posterior is different from the target posterior distribution, can it still be used in Bayesian inference? Or, since different summary statistics can produce different ABC posteriors, can one or more of these distributions be used to make statistical inferences?

In this paper, we attempt to address these questions by instead re-framing Algorithm 1 within a frequentist context and consider a more general likelihood-free method based on confidence distribution theory. To this end, we introduce a new computational method called *approximate confidence-distribution computing*. Before introducing this new algorithm, we first quickly review the concept of a confidence distribution.

1·2. *Confidence distributions*

When estimating an unknown parameter, we often desire that our estimators, whether point estimators or interval estimators, have certain properties, such as unbiasedness or correct coverage of the true parameter value in the long run. A confidence distribution is an extension of this tradition in that it is a distribution estimate (i.e., it uses a sample-dependent distribution function to estimate the target parameter) that satisfies certain properties. Following Xie & Singh (2013), Schweder & Hjort (2016) and references therein, we define a confidence distribution as follows.

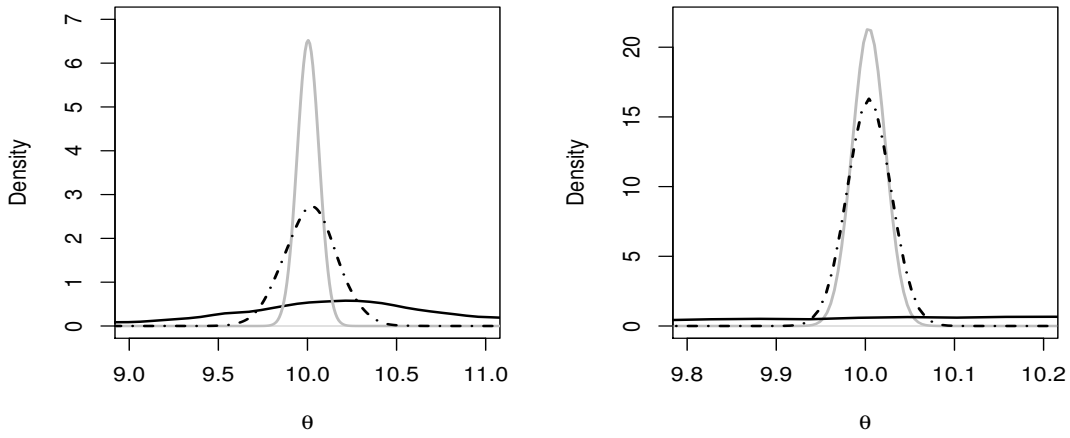


Fig. 1: *Target posterior (gray) and ABC posteriors for data from a Cauchy distribution with known scale parameter for both* $S_n = \bar{x}$ *(solid black) and* $S_n = Median(x)$ *(dashed black) for a small observed sample sizes* $(n = 50)$ *on the left and a very large sample size on the right* $(n = 5000)$.

Definition 1. *A sample-dependent function on the parameter space is a confidence distribution for a parameter* θ *if 1) For each given sample the function is a distribution function on the parameter space; 2) The function can provide confidence intervals/regions of all levels for* θ.

Consider the following example taken from Singh et al. (2007). Suppose $X_1, \ldots, X_n$ is a sample from $N(\mu, \sigma^2)$ where both μ and σ^2 are unknown. A confidence distribution for parameter μ is the function $H_n(y) = F_{t_{(n-1)}}\left\{(y - \bar{X})/(s_n/\sqrt{n})\right\}$ where $F_{t_{(n-1)}}(\cdot)$ is the cumulative distribution function of a Student's t-random variable with $(n-1)$ degrees of freedom and $\bar{X}$ and s_n^2 are the sample mean and variance, respectively. Here $H_n(y)$ is a cumulative distribution function in the parameter space of μ from which we can construct confidence intervals of μ at all levels. For example, for any $\alpha \in (0, 1)$, one sided confidence intervals for μ are $(\infty, H_n^{-1}(\alpha)]$ and $[H_n^{-1}(\alpha), \infty)$. Similarly, a confidence distribution for parameter σ^2 is the function $H_n(\sigma^2) = 1 - F_{\chi^2_{n-1}}\left[\{(n-1)s_n^2\}/(\sigma^2)\right]$, where $F_{\chi^2_{n-1}}(\cdot)$ is the distribution function of a Chi-squared random variable with $(n-1)$ degrees of freedom. Again, $H_n(\sigma^2)$ is a cumulative distribution function in the parameter space of σ^2 from which we can construct confidence intervals of σ at all levels.

We emphasize that, by definition, a confidence distribution is a sample-dependent distribution function that can represent confidence intervals/regions of all levels for a parameter of interest. Any confidence distribution approach will utilize a sample-dependent distribution function to estimate the unknown parameter; thus a confidence distribution is actually an expression of inference (i.e. an estimator of the parameter in the form of a distribution function), rather than a distribution for the parameter.

A confidence distribution estimator has a similar appeal to a Bayesian posterior in that it is a distribution function carrying much information about the parameter. A confidence distribution however, is a frequentist notion which treats the parameter as a fixed and unknown quantity, and we judge its performance by the frequency coverage such that a interval/region for θ obtained from a confidence distribution can contain the true parameter value, θ_0, at any specified frequency. We will refer to this property as the *frequentist coverage property* of confidence dis-

tributions. We hope to demonstrate that the construction of approximate confidence distribution computing as a likelihood-free method provides one of many examples in which confidence distribution theory provides a useful inferential tool for a problem where a statistical method with desirable properties was previously unavailable. For more details on confidence distributions see Xie & Singh (2013), Schweder & Hjort (2016), and references therein.

1·3. *Approximate confidence distribution computing*

We now formally introduce approximate confidence distribution computing, as an alternative to approximate Bayesian computing (Algorithm 1). The theoretical foundation for approximate confidence distribution computing relies upon the frequentist coverage property of confidence distributions. It also provides a computational method with potential applications extending beyond the scope of Algorithm 1. Additionally, as will be discussed later, approximate confidence distribution computing introduces some flexibility that can greatly decrease computing costs.

Approximate confidence distribution computing proceeds in the same manner as Algorithm 1, but no longer requires a prior assumption on θ; instead, the user is free to select a data-dependent distribution function, $r_n(\cdot)$, from which potential parameter values will be generated. Specifically, the new algorithm proceeds as following:

Algorithm 2. (Accept-reject approximate confidence distribution computing)

1. Simulate $\theta_1, \ldots, \theta_N \sim r_n(\theta)$;
2. and 3. are identical with steps 2. and 3. of Algorithm 1.

The underlying distribution from which these N copies of θ are simulated is called the *ACC confidence distribution*, and it has the probability density $r_\varepsilon(\theta \mid s_{\text{obs}})$ defined by replacing $\pi(\theta)$ in (1) with $r_n(\theta)$. Denote the corresponding distribution function by $R_\varepsilon(\theta \in A \mid s_{\text{obs}})$. We use the notation θ_{ACC} to represent a random draw from this function. When $r_n(\theta) = \pi(\theta)$, Algorithm 2 is the same as Algorithm 1; in this way, approximate Bayesian computing can be viewed as a special case of approximate confidence distribution computing.

From a Bayesian perspective, one may view Algorithm 2 as as an extension permitting the use of approximate Bayesian computing in the presence of a data-dependent prior. However, there is another natural, frequentist interpretation that views the function $r_n(\cdot)$ as an initial distribution estimate for θ and views Algorithm 2 as a method that updates this estimates in pursuit of a better-performing distribution estimate. The logic of this frequentist interpretation is analogous to any updating algorithm in point estimation (e.g., say, a Newton-Raphson algorithm or an Expectation-maximization algorithm), which requires an initial estimate and then updates a search for a better-performing estimate. There is a question whether the data are 'doubly used'. The answer depends on how the initial $r_n(\cdot)$ is chosen. Under some constraints on $r_n(\cdot)$, Algorithm 2 can guarantee a distribution estimator for θ that satisfies the frequentist coverage property, although Algorithm 2 may not ensure the efficiency of this distribution estimator unless the summary statistic is sufficient.

1·4. *Related work*

Likelihood-free methods such as approximate Bayesian computing have existed for more than 20 years, but research regarding the theoretical properties of these methods is a newly active area (e.g. Li & Fearnhead (2018b); Frazier et al. (2018)). Here we do not attempt to give a full review of all likelihood-free methods (e.g. Marin et al. (2011)) but we acknowledge the existence of alternatives such as indirect inference (e.g. Creel & Kristensen (2013); Gourieroux et al. (1993)).

One of our theoretical results specifies conditions under which approximate confidence distribution computing produces an asymptoticly normal confidence distribution. This result, pre-

sented in Section 3, mirrors the work of Li & Fearnhead (2018b) and Frazier et al. (2018) on the asymptotic normality of the ABC posterior distribution. However, in contrast to these papers, we are not concerned with viewing the result of Algorithm 2 as an approximation to some posterior distribution, rather we focus on the properties of this distribution inherited through its connection to confidence distributions. More importantly, the properties we develop here allow us to conduct statistical inference with a guaranteed performance standard. In Section 2 we discuss how Algorithm 2 can be used in exact inference by not relying on any sort of asymptotic (large n) assumptions or normally distributed populations. Aside from the errors of Monte-Carlo approximation and the choice of tolerance level, exact inference from Algorithm 2 ensures the targeted repetitive coverage rates and type-I errors.

This paper presents the novel idea that the continued study of likelihood-free methods would benefit from the incorporation of confidence distribution theory. To this end, and for the ease of presentation, we mainly focus on the basic accept-reject version of Algorithm 2. We conclude that much of the existing work in the approximate Bayesian computation literature can also be applied to Algorithm 2 to further improve upon its computational performance as discussed in Sections 2 and 5.

1·5. *Notation*

Throughout the paper we will use the following notation. The observed data is $x_{obs} \in \mathcal{X} \subset \mathbb{R}^n$, the summary statistic is a mapping $S_n : \mathcal{X} \to \mathcal{S} \subset \mathbb{R}^d$ and the observed summary statistic is $s_{obs} = S_n(x_{obs})$. The parameter of interest is $\theta \in \mathcal{P} \subset \mathbb{R}^p$ with $p \leq d \leq n$; i.e. the number of unknown parameters is no greater than the number of summary statistics and dimension of the summary statistic is no greater than the dimension of the data. If some function of S_n is an estimate for θ, we denote this function by $\hat{\theta}_S$. Let θ_0 represent the fixed, true value of the parameter θ. We will refer to the distributions resulting from Algorithms 1 and 2 by their mathematical notations π_ε and r_ε, respectively. Let θ_{ACC} represent a random draw from r_ε.

Additionally, for a series z_n, we use the notation that $z_n \approx a_n$, if there exists constants m and M such that $0 < m < |z_n/a_n| < M < 1$ as $n \to \infty$ and for a real function $g(x)$, denote its gradient function at $x = x_0$ by $D_x\{g(x_0)\}$.

2. ESTABLISHING FREQUENTIST GUARANTEES FOR APPROXIMATE CONFIDENCE DISTRIBUTION COMPUTING

Loosely speaking, if the randomness in the Monte-Carlo simulation from r_ε matches that of the sampling population, then approximate confidence distribution computing can be used to help us answer inference questions with frequentist guarantees on performance. In this section, we formally derive this statement and establish conditions under which Algorithm 2 can be used to produce confidence regions with guaranteed coverages at all levels.

To motivate our main theoretical result, we first consider the simple case where we have a scalar parameter, θ, and $\hat{\theta}_S$ is a function that maps the summary statistic into the parameter space $\mathcal{P} = (-\infty, \infty)$. Suppose further that the Monte-Carlo copy of $(\theta_{ACC} - \hat{\theta}_S) \mid S_n = s_{obs}$ and the sampling population copy of $(\hat{\theta}_S - \theta) \mid \theta = \theta_0$ have the same distribution:

$$(\theta_{ACC} - \hat{\theta}_S) \mid S_n = s_{obs} \sim (\hat{\theta}_S - \theta) \mid \theta = \theta_0, \tag{3}$$

Then, we can conduct inference for θ with a guaranteed performance. On the left hand side of (3), $\hat{\theta}_S$ is fixed given s_{obs} so the probability measure is with respect to θ_{ACC}, meaning the randomness is due to the simulation conducted in Algorithm 2. Conversely, on the right hand side, $\hat{\theta}_S$ is a random variable since randomness is due to the random data before observation.

This is very similar to the bootstrap central limit theorem that $n^{1/2}(\theta_B - \hat{\theta}_B) \mid S_n = s_{obs} \sim n^{1/2}(\hat{\theta}_B - \theta) \mid \theta = \theta_0$, as $n \to \infty$, where appropriate; cf, Singh (1981) and Freedman & Bickel (1981). There, the randomness on the left hand side is from the bootstrap estimator, θ_B given $S_n = s_{obs}$, and the randomness on the right hand side is from the random sample, S_n.

Let $G(t) = \text{pr}(\hat{\theta}_S - \theta_0 \leq t \mid \theta = \theta_0)$, and for simplicity assume that we also have $\text{pr}^*(\theta_{ACC} - \hat{\theta}_S \leq t \mid s_{obs}) = G(t)$. Here $\text{pr}^*(\cdot \mid s_{obs})$ refers to the probability measure on simulation given $S_n = s_{obs}$ corresponding to the left hand side of (3). Define $D_n(t) = D(t, s_{obs}) = \text{pr}^*(2\hat{\theta}_S - \theta_{ACC} \leq t \mid s_{obs})$, a mapping from $\mathcal{P} \times \mathcal{S} \to (0, 1)$. Given s_{obs}, $D_n(t)$ is a sample-dependent cumulative distribution function on $\mathcal{P}$.

[Claim] Under the setup above, $D_n(t)$ is a confidence distribution for θ and, for any $\alpha \in (0, 1)$, $(-\infty, D_n^{-1}(1-\alpha)] = (-\infty, 2\hat{\theta}_S - \theta_{ACC,\alpha}]$ is an $(1-\alpha)$-level confidence interval of θ.

In the claim, $D_n^{-1}(\alpha)$ is the quantile of $D_n(\cdot)$, i.e., the solution of $D_n(t) = \alpha$, and $\theta_{ACC,\alpha}$ is a quantile of θ_{ACC}, defined by $\text{pr}^*(\theta_{ACC} \leq \theta_{ACC,\alpha} \mid sob) = \alpha$. A proof of the claim is provided in the appendix.

Now we introduce a key lemma that generalizes the argument above to a multidimensional parameter and a wider range of relationships between S_n and θ_{ACC}. This lemma assumes a relationship between two mappings V and $W : \mathcal{P} \times \mathcal{S} \to \mathbb{R}^k$, where $V(\cdot, S_n)$ is a function that acts on the parameter space $\mathcal{P}$, given $S_n = sob$, and $W(\theta, \cdot)$ is a function that acts on the space of the summary statistic $\mathcal{S} \subset \mathbb{R}^d$, given $\theta = \theta_0$. For example, in the argument above, $V(t_1, t_2) = -W(t_1, t_2) = t_1 - \hat{\theta}(t_2)$, where $\hat{\theta}$ is a function of the summary statistic; however, we may also wish to consider other non-linear mappings. Corresponding to (3), we require a matching equation: $V(\theta_{ACC}, S_n) \mid S_n = sob \sim W(\theta, S_n) \mid \theta = \theta_0$. More formally, we consider the following condition:

[Condition A] For $\mathfrak{B}$ a Borel set on $\mathbb{R}^k$,

$$\sup_{A \in \mathfrak{B}} \|\text{pr}^*\{V(\theta_{ACC}, S_n) \in A \mid S_n = sob\} - \text{pr}\{W(\theta, S_n) \in A \mid \theta = \theta_0\}\| = o_p(\varepsilon),$$

For a given sob and $\alpha \in (0, 1)$, define a set $A_{1-\alpha} \subset \mathbb{R}^k$ such that,

$$\text{pr}^*\{V(\theta_{ACC}, S_n) \in A_{1-\alpha} \mid S_n = sob\} = (1-\alpha) + o(\delta), \tag{4}$$

where $\delta > 0$ is a pre-selected small positive precision number $\delta \to 0$. Condition A implies that $\Gamma_{1-\alpha}(sob) = \{\theta : W(\theta, sob) \in A_{1-\alpha}\} \subset \mathcal{P}$ is a level $(1-\alpha)100\%$ confidence region for θ_0. We summarize this in the following lemma which is proved in the appendix.

Lemma 2. Suppose that there exists mappings V and $W : \mathcal{P} \times \mathcal{S} \to \mathbb{R}^k$ such that Condition A holds. Then, $\text{pr}\{\theta_0 \in \Gamma_{1-\alpha}(S_n)\} = (1-\alpha) + o_p(\varepsilon \vee \delta)$. If further Condition A holds almost surely, then $\text{pr}\{\theta_0 \in \Gamma_{1-\alpha}(S_n)\} = (1-\alpha) + o(\varepsilon \vee \delta)$, almost surely.

In the above, there are no requirements on the sufficiency of the summary statistic. Also, ε in Condition A is the tolerance level for the matching of simulated $S^{(i)}$ and s_{obs} in Step 3 of Algorithm 2, and it may or may not depend on the sample size n. Furthermore, δ in (4) is designed to control Monte-Carlo approximation error. So whether or not Lemma 2 is a large sample result depends only on whether or not we require $\varepsilon \to 0$ at a certain rate of the sample size n. Later in this section, we will consider a special case that is sample-size independent; then Section 3 extends the large sample Bernstein-von Mises theory to Algorithm 2, using a tolerance that does depend on n.

Before we move on to verify Condition A for different cases, we first relate equation (4) to θ_{ACC} samples from r_ε. Suppose θ_i, $i = 1, \ldots, N$, are N Monte-Carlo copies of θ_{ACC}. Let

$v_i = V(\theta_i, sob)$. The set $A_{1-\alpha}$ can typically be a $(1-\alpha)100\%$ contour set of $\{v_1, \ldots, v_N\}$ with $o(\delta) = o(N^{-1/2})$. For example, we can directly use $v_1, \ldots, v_N$ to construct a $100(1-\alpha)\%$ depth contour as $A_{1-\alpha} = \{\theta : (1/N)\sum_{i=1}^{N} \mathbb{I}\{\hat{D}(v_i) < \hat{D}(\theta)\} \geq \alpha\}$, where $\hat{D}(\cdot)$ is an empirical depth function on $\mathcal{P}$ computed based on the empirical distribution of $\{v_1, \ldots, v_N\}$. See, e.g., Serfling (2002) and Liu et al. (1999) for the development of data depth and depth contours in nonparametric multivariate analysis. In the special case where $k = 1$, by defining $\hat{q}_\alpha = v_{[N\alpha]}$, the $[N\alpha]$th largest $v_1, \ldots, v_N$, a $(1-\alpha)100\%$ confidence region for θ_0 can then be constructed as $\Gamma_{1-\alpha}(sob) = \{\theta : \hat{q}_{\alpha/2} \leq W(\theta, sob) \leq \hat{q}_{1-\alpha/2}\}$ or $\Gamma_{1-\alpha}(sob) = \{\theta : W(\theta, sob) \leq \hat{q}_{1-\alpha}\}$.

We also remark that the existing literature on likelihood-free methods typically relies upon obtaining a "nearly sufficient" summary statistic to justify inferential results; see e.g., Joyce & Marjoram (2008). In this paper however, we explore guaranteed frequentist properties of Algorithm 2 that hold without regard to a "sufficient enough" summary statistic. However, if the summary statistic happens to be sufficient, then an appropriate choice of the initial ballpark estimate, r_n, means that inference based on the resulting distribution, r_ε, is also efficient.

To end this section, Theorem 1 explores a special case of Algorithm 2 in location and scale families. In this case, the integrity of the inference based on r_ε is ensured without relying on large sample theory or a Gaussian assumption. The proof is given in the appendix.

Theorem 1. Assume $\hat{\mu}_S = \hat{\mu}(S_1)$ and $\hat{\sigma}_S = \hat{\sigma}(S_2)$ are point estimators for location and scale parameters μ and τ, respectively, and $S_1, S_2 \in \mathbb{R}$ are two summery statistics.

Part 1 Suppose $\hat{\mu}_S \sim g_1(\hat{\mu}_S - \mu)$. If $r_n(\mu) \propto 1$ then, for any u,

$$| \text{ pr}^*\{\mu_{ACC} - \hat{\mu}_S \leq u \mid \hat{\mu}_{obs}\} - \text{pr}\{\hat{\mu}_S - \mu \leq u \mid \mu = \mu_0\} | = o(\varepsilon), \quad \text{almost surely.}$$

Part 2 Suppose $\hat{\sigma}_S \sim 1/\sigma g_2\left(\hat{\sigma}_S/\sigma\right)$. If $r_n(\sigma) \propto 1/\sigma$ then, for any $v > 0$,

$$| \text{ pr}^* \left\{ \frac{\sigma_{ACC}}{\hat{\sigma}_S} \leq v \mid \hat{\sigma}_{obs} \right\} - \text{pr} \left\{ \frac{\hat{\sigma}_S}{\sigma} \leq v \mid \sigma = \sigma_0 \right\} | = o(\varepsilon), \quad \text{almost surely.}$$

Part 3 Suppose $\hat{\mu}_S \sim (1/\sigma) g_1\left((\hat{\mu}_S - \mu)/\sigma\right)$ and $\hat{\sigma}_S \sim (1/\sigma) g_2\left(\hat{\sigma}_S/\sigma\right)$ are independent. If $r_n(\mu, \sigma) \propto 1/\sigma$, then, for any u and any $v > 0$,

$$\begin{aligned} | \text{ pr}^* \left\{ \mu_{ACC} - \hat{\mu}_S \leq u, \frac{\sigma_{ACC}}{\hat{\sigma}_S} \leq v \mid \hat{\mu}_{obs}, \hat{\sigma}_{obs} \right\} - \\ \text{pr} \left\{ \mu_{ACC} - \hat{\mu}_S \leq u, \frac{\hat{\sigma}_S}{\sigma} \leq v \mid \mu = \mu_0, \sigma = \sigma_0 \right\} | = o(\varepsilon), \quad \text{almost surely.} \end{aligned}$$

Furthermore, we may derive $H_1(\hat{\mu}_S, x) = 1 - \int_{-\infty}^{\hat{\mu}_S - x} g_1(w)dw$, a confidence distribution for μ induced by $(\hat{\mu}_S - \mu)$ given $\mu = \mu_0$ or $H_2(\hat{\sigma}_S^2, x) = 1 - \int_0^{\hat{\sigma}_S^2/x} g(w)dw$, a confidence distribution for σ^2 induced by $\hat{\sigma}_S^2/\sigma^2$ given $\sigma = \sigma_0$.

Theorem 1 represents a departure from the typical asymptotic arguments supporting likelihood-free methods and permits the use of Algorithm 2 in forming confidence intervals/regions with potentially exact correct frequentist coverage. If we have an exact pivot for some unknown parameter, then the only source of approximation in the inference resulting from Algorithm 2 is due to the computational requirements on ε. We remark that Theorem 1 can be generalized beyond the location-scale family with some additional technical conditions on the initial ballpark estimate.

3. FREQUENTIST COVERAGE PF APPROXIMATE CONFIDENCE DISTRIBUTION COMPUTING FOR LARGE SAMPLES

3·1. *Bernstein-von Mises result for approximate confidence distribution computing*

Theorem 2 below enables the construction of a confidence region with asymptotically correct coverage property using the output of Algorithm 2. For approximate Bayesian computing, it is known that Condition A can be satisfied by a Bernstein-von Mises type convergence when ε_n degenerates to 0 fast enough (Li & Fearnhead, 2018a). Theorem 2 extends the Bernstein-von Mises type convergence to approximate confidence distribution computing. The key condition to obtain this result is the following central limit theorem for the summary statistic.

[C1] There exists a sequence a_n, satisfying $a_n \to \infty$ as $n \to \infty$, a d-dimensional vector $\eta(\theta)$ and a $d \times d$ matrix $A(\theta)$, such that for all $\theta \in \mathcal{P}_0$,

$$a_n\{s_n - \eta(\theta)\} \to N(0, A(\theta)), \text{ as } n \to \infty,$$

in distribution. We also assume that $s_{obs} \to \eta(\theta_0)$ in probability. Furthermore, it holds that (i) $\eta(\theta)$ and $A(\theta) \in C^1(\mathcal{P}_0)$, and $A(\theta)$ is positive definite for any θ; (ii) for any $\delta > 0$ there exists a $\delta' > 0$ such that $\mid \eta(\theta) - \eta(\theta_0) \mid > \delta'$ for all θ satisfying $\mid \theta - \theta_0 \mid > \delta$; and (iii) $I(\theta) \triangleq \left\{\frac{\partial}{\partial\theta}\eta(\theta)\right\}^T A^{-1}(\theta) \left\{\frac{\partial}{\partial\theta}\eta(\theta)\right\}$ has full rank at $\theta = \theta_0$.

Assuming that we also satisfy the regulatory conditions C5–C9 in the appendix, as established in Li & Fearnhead (2018a), we now show that this convergence can be generalized to apply to a data-dependent $r_n(\cdot)$, so that Condition A can be satisfied for Algorithm 2.

Below are conditions on $r_n(\cdot)$: Suppose there exists a sequence $\{\tau_n\}$ and $\delta > 0$, such that

[C2] $\tau_n = o(a_n)$ and $\sup_{\theta \in B_\delta} \tau_n^{-p} r_n(\theta) = O_p(1)$,
[C3] $r_n(\theta) \in C^1(\mathcal{P}_0)$ and $\tau_n^{-p} r_n(\theta_0) = \Theta_p(1)$, and
[C4] $\sup_{\theta \in \mathbb{R}^p} \tau_n^{-1} \frac{\partial}{\partial\theta}[\tau_n^{-p} r_n(\theta)] = O_p(1)$.

In the conditions, a_n, the convergence rate of s_n, is dominated by τ_n, a rate that standardizes the multivariate function $r_n(\theta)$ within some δ-ball. So in essence, C2 and C3 require $r_n(\cdot)$ to be more dispersed than the s-likelihood for θ within a compact set $\mathcal{P}_0$. C4 requires the first derivative of standardized $r_n(\theta)$ to converge with rate τ_n. These are weak conditions and can be satisfied by, e.g. $r_n(\theta)$ being a local asymptotic normality model. Let θ_ε be the expectation of θ under $r_\varepsilon(\theta, s_n) \propto r_n(\theta)\tilde{f}_n(s_n \mid \theta)$.

Theorem 2. Assume C1 and that $r_n(\theta)$ satisfies C2—C4 and $\varepsilon_n = o(a_n^{-1})$ as $n \to \infty$. If we can also assume C5–C8 of the appendix and if the following statements hold:

$$\sup_{A \in \mathfrak{B}^p} \left\| R_\varepsilon\{a_n(\theta_{ACC} - \theta_\varepsilon) \in A \mid s_n = s_{\text{obs}}\} - \int_A N\left(t; 0, I(\theta_0)^{-1}\right) dt \right\| \to 0,$$

in probability, and

$$a_n(\theta_\varepsilon - \theta_0) \to N\left(0, I(\theta_0)^{-1}\right),$$

in distribution, then Condition A is satisfied with $\{V(\theta_{ACC}, s_n) \mid s_n = s_{obs}\} = a_n(\theta_{ACC} - \theta_\varepsilon)$, $W(s_n, \theta_0) = a_n(\theta_\varepsilon - \theta_0)$ and replacing ε with sample-size-dependent ε_n.

3·2. *Comparing approximate Bayesian computing and approximate confidence distribution computing*

Theorem 2 says that the sufficiently small ε_n for the ACC confidence distribution is $o(a_n^{-1})$, the same as that for the ABC posterior distribution, and with such ε_n, the confidence region for

θ_0 with asymptotically correct coverage can be constructed as outlined in Section 2. In practice, since in most cases θ_ε does not have closed form, it is estimated by the sample of θ_{ACC}. In conjunction with Proposition 1 of Li & Fearnhead (2018a) on the convergence of the ABC posterior, Theorem 2 shows that R_ε and Π_ε are the same to the first order for sufficiently small ε. This demonstrates the computational advantage of using a data-dependent $r_n(\cdot)$ because, if the same ε_n is used for Algorithm 1 and Algorithm 2, the output of both will have the same asymptotic distribution but the latter has higher Monte Carlo efficiency due to the fact that any $r_n(\cdot)$ satisfying C2–C4 is closer to the s-likelihood than $\pi(\cdot)$ thus resulting in higher acceptance probabilities.

One drawback of using the sufficiently small ε in Algorithm 1 is the degeneracy of Monte Carlo efficiency as $n \to \infty$ which occurs since the acceptance probability for any proposal distribution will degenerate to zero (Li & Fearnhead, 2018a). This means that most simulated datasets, which are often computationally expensive, will be wasted when the data are informative. It is easy to see that Algorithm 2 also suffers from the same issue with any choice of $r_n(\cdot)$. From this point of view it may be more practical to use ε_n outside this regime; however, a larger ε_n will always cause the parameter uncertainty to be overestimated by π_ε.

This highlights the tension in approximate Bayesian computation between choice of bandwidth that will lead to more accurate inferences versus choices to reduce the computational cost. For Algorithm 2, although a larger ε_n does not necessarily invalidate the existence of the desired mappings V and W, the overall performance might not be comparable to Algorithm 1 for larger ε_n and hence it is not always preferable to adapt a data-dependent $r_n(\theta)$. This is illustrated in the following Gaussian example.

Example 1. Consider a univariate normal model with mean θ and unit variance and consider observations that are independent identically distributed from the model with mean value θ_0. Assume the prior distribution of θ is standard normal, and $r_n(\theta)$ is the density $N(\theta; \mu_n, b_n^{-2})$ for some sequences μ_n and b_n. Assume μ_n and b_n satisfy that $b_n(\mu_n - \theta_0) = O(1)$ and $b_n = o(n^{1/2})$ as $n \to \infty$. This is a natural assumption for $r_n(\theta)$ to be a reasonable proposal density, because it guarantees that $r_n(\theta)$ well covers the true parameter θ_0 and is more dispersed than the s-likelihood. In Algorithms 1 and 2, the summary statistic is the sample mean and the acceptance kernel is Gaussian with variance ε_n^2. For this toy model, both π_ε and r_ε have the same closed form $N(\theta; \theta_\varepsilon, \sigma_\varepsilon^2)$ where

$$\theta_\varepsilon = \frac{sob + b_n^2(1/n + \varepsilon^2)\mu_n}{1 + b_n^2(1/n + \varepsilon^2)}, \ \sigma_\varepsilon^2 = \frac{1/n + \varepsilon^2}{1 + b_n^2(1/n + \varepsilon^2)}.$$

Consider the scaled mean $n^{1/2}(\theta_\varepsilon - \theta_0)$ as in Theorem 2. For the ABC posterior, some algebra shows that ε_n is negligible when $\varepsilon = o(n^{-1/4})$ and the scaled ABC posterior mean has the asymptotic distribution $N(0, 1)$ as $n \to \infty$. For the ACC confidence distribution, by decomposing the scaled mean of the confidence distribution as $\Delta_1 n^{1/2}(sob - \theta_0) + \Delta_2 b_n(\mu_n - \theta_0)$ where

$$\Delta_1 = \frac{1}{1 + b_n^2(1/n + \varepsilon^2)}, \ \Delta_2 = \frac{\sqrt{n} b_n(1/n + \varepsilon^2)}{1 + b_n^2(1/n + \varepsilon^2)},$$

it can be seen that ε_n is negligible only when $\varepsilon_n = o(b_n^{-1/2} n^{-1/4})$ and only then $n^{1/2}(\theta_\varepsilon - \theta_0)$ has the same asymptotic distribution, $N(0, 1)$. As $n \to \infty$, if $b_n^{1/2} n^{1/4} \varepsilon_n$ does not degenerate to zero, then neither does Δ_2 so the asymptotic variance of $n^{1/2}(\theta_\varepsilon - \theta_0)$ is overinflated by a constant or diverging factor. Therefore with such sufficiently small ε_n, the asymptotic performance of the ACC confidence distribution mean is inferior to that of the ABC posterior mean. However,

since Algorithm 2 is still superior to Algorithm 1 computationally by having a higher acceptance probability, their overall performance is incomparable.

On the other hand, when ε_n is larger than the sufficiently small ε, i.e. not in the order of $o(n^{-1/2})$, the ABC posterior variance σ_ε over-inflates the targeted posterior variance by a constant or a rate that goes to infinity. Hence it is more difficult to estimate the posterior variance accurately than it is to obtain an accurate point estimate.

3·3. *Comparison with regression adjustment*

One remedy to reduce the overinflated uncertainty in the distribution estimate of Algorithm 1 is to apply the regression adjustment on its output (Beaumont et al., 2002). It has been suggested to routinely apply the regression adjustment on Algorithm 1 in order to correctly quantify the estimated uncertainty, yielding inference that is accurate in terms of both point estimates and the variance of distributional estimates with ε_n decaying more slowly than $o(n^{-1/2})$ (Li & Fearnhead, 2018a).

We suggest applying the regression adjustment to Algorithm 2 for a similar reason. Denote a sample from the ACC conditional distribution by $\{(\theta_i, s^{(i)})\}_{i=1,\ldots,N}$. A new sample can be obtained by using $\{\theta_i - \hat{\beta}_\varepsilon(s^{(i)} - s_{\text{obs}})\}_{i=1,\ldots,N}$ where $\hat{\beta}_\varepsilon$ is the least square estimate of the coefficient matrix in the linear model

$$\theta_i = \alpha + \beta(s^{(i)} - s_{\text{obs}}) + e_i,\ i = 1, \ldots, N,$$

where e_i are independent identically distributed error. The new sample can be seen as a draw from the distribution $\theta^* = \theta - \beta_\varepsilon(s - s_{\text{obs}})$, where $(\theta, s) \sim r_\varepsilon(\theta, s)$ and β_ε is from the minimizer

$$(\alpha_\varepsilon, \beta_\varepsilon) = \text{argmin}_{\alpha,\beta} E_\varepsilon \left\{ || \ \theta - \alpha - \beta(s - s_{\text{obs}}) \ ||^2 \ s_{\text{obs}} \right\}$$

for expectation under the joint distribution of (θ, s), but with β_ε replaced by its estimator.

The following theorem states that regression adjusted approximate confidence distribution computing has the same favored property as regression adjusted approximate Bayesian computing. Let θ^*_ε be the expectation of the regression adjusted θ_{ACC} values, θ^*_{ACC}, under $r_\varepsilon(\theta, s)$.

Theorem 3. Assume the conditions of Theorem 2 and additionally assume C9 of the appendix. If $\varepsilon_n = o(a_n^{-3/5})$ as $n \to \infty$ and if the following statements hold:

$$\sup_{A \in \mathfrak{B}^p} \ | \ R_\varepsilon\{a_n(\theta^*_{ACC} - \theta^*_\varepsilon) \in A \mid s_n = s_{\text{obs}}\} - \int_A N\left(t; 0, I(\theta_0)^{-1}\right) dt \ | \xrightarrow{P} 0, \quad \text{and}$$

$$a_n(\theta^*_\varepsilon - \theta_0) \xrightarrow{d} N\left(0, I(\theta_0)^{-1}\right),$$

then Condition A is satisfied with $[V(\theta_{ACC}, s_n) \mid s_n = sob] = a_n(\theta^*_{ACC} - \theta^*_\varepsilon)$, $W(\theta_0, s_n) = a_n(\theta^*_\varepsilon - \theta_0)$ and replacing ε with ε_n.

Theorem 3 indicates the adjusted sufficiently small ε for the ACC confidence distribution is $o(a_n^{-3/5})$, the same as for the adjusted ABC posterior. With such ε_n, the desired confidence region for θ_0 can be constructed as in Section 2 using the new sample, again illustrating the computational advantage of a data-dependent $r_n(\cdot)$.

3·4. *Guidelines for selecting the initial ballpark distribution estimate*

The generality of approximate confidence distribution computing is that it can produce justifiable inferential results with weak conditions on a possibly data-dependent initial ballpark distribution estimate. In general, one should be careful in choosing this estimate, r_n, to ensure its growth with respect to the sample size is slower than the growth of the s-likelihood, according

to C2. We now propose a generic algorithm to construct this initial ballpark estimate. Suppose the observed dataset is x of size n. Given the summary statistic $s = S(x)$, assume that a point estimate $\widehat{\theta}(s)$ of θ can be computed.

[Step1] Choose k subsets of the observations, each with size n^δ for some $0 < \delta < 1$.

[Step2] For each subset x_i of x, compute the point estimate $\widehat{\theta}_i = \widehat{\theta}(s_i)$ where $s_i = s(x_i)$, for $i = 1, \ldots, k$.

[Step3] Let $r_n(\theta) = 1/(kh) \sum_{i=1}^{k} K\left\{(\theta - \widehat{\theta}_i)/h\right\}$, where K is any kernel function satisfying $C6$ and $h > 0$ is the bandwidth of the kernel density estimate using $\{\widehat{\theta}_1, \ldots, \widehat{\theta}_k\}$.

By choosing $\delta < 3/5$, we ensure C2–C4 are met and that the ε selected by accepting a reasonable proportion of simulations in Algorithm 2 is sufficiently small, provided the rate of s is a power function of n. Based on our experience, if n is large one may simply choose $\delta = 1/2$; however, for small n, say $n < 100$, it is better to select $\delta > 1/2$. For problems with an intractable likelihood, possible choices of $\widehat{\theta}(s)$ include the estimator maximizing an approximate likelihood that is a function of s or the minimizer of the average distance as a function of θ between simulated s and s_{obs} (Meeds & Welling, 2015). A full study of the choice of $\widehat{\theta}(s)$ is beyond the scope of this paper.

It is important to recognize the trade-off in approximate confidence-distribution computing between faster computations and guaranteed frequentist inference. Whilst one may choose to iteratively update $r_n(\cdot)$ for the sake of computing time, this may risk violating C2–C4. If these assumptions are violated, then the resulting simulations do not necessarily form a confidence distribution and consequently inference based on a sample from r_ε may not be valid in the frequentist sense. However, provided C1–C4 hold and the observed data is large enough, Theorem 2 shows that regardless of the choice of the initial ballpark estimate, Algorithm 2 always produces the same confidence distribution.

4. EMPIRICAL EXAMPLES

4·1. *Cauchy data*

The first few examples we discuss are continuations of Fig. 1 from the introduction. Suppose we observe random data, $(x_1, \ldots, x_n)$, from a $Cauchy(\theta, \tau)$ distribution. We wish to produce 95% confidence intervals or regions for each or both of the model parameters.

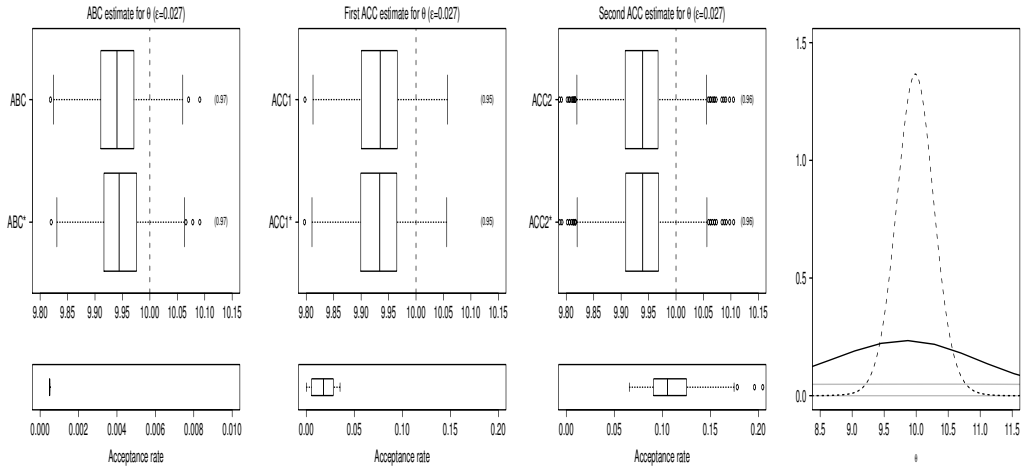
First, consider the case where θ is unknown but τ is known. Here we consider applying ACC with summary statistic $S_n = Median(x_1, \ldots, x_n)$ and we consider two different data-driven choices of $r_n(\theta)$ and note the requirement of C1 is met since

$$n^{1/2}(S_n - \theta) \xrightarrow{\mathcal{L}} N\left(0, \pi^2\tau^2/4\right).$$

Hence, provided we choose $r_n(\theta)$ to satisfy C2–C4, we can use ACC to find valid frequentist confidence intervals for θ. We first consider $r_{n1}(\theta) \propto \left[1 + \{(\bar{x} - \theta)/\tau_2\}^2\right]^{-1}$, where $\tau_2 > \tau$, which focuses the ACC draws around the mean of the observed data. Although this choice of r_{n1} is already faster than the analogous ABC method with a flat prior, $\pi(\theta) \propto 1$, (see Figure 2), we can improve upon the efficiency of the ACC algorithm with an even more informative choice, say, $r_{n2}(\theta)$. Define r_{n2} according to the scheme outlined in Section 3.4, i.e. $r_{n2}(\theta) \propto \sum_{i=1}^{k} \phi\left\{(\theta - S_{ni})/h\right\}$, where $h > 0$ is some choice of kernel bandwidth, ϕ is the standard normal density function, and $k = n^\delta$ for some $\delta \in (0, 1)$. The plots of r_{n1} and r_{n2} are contrasted with the plot of a flat prior used in ABC in Fig. 2.

In the numerical study, we observe a sample of size $n = 400$ and assume $\tau = 0.55$ is known and compare the results of ABC to ACC in Fig. 2. For ACC with r_{n1} we choose $\tau_2 = 10$ and for ACC with r_{n2} we set $h = 0.09198$ (chosen by R's default *density()* function) and $\delta = 1/2$. Each box-plot represents the kept unadjusted or regression-adjusted parameter draws from the corresponding ACC or ABC distribution.

The coverage of the 95% confidence intervals for the true parameter value ($\theta = 10$) over 100 independent runs of ACC is given in parentheses. In Fig. 2, these two applications of ACC clearly display the computational advantage in the flexible choice of $r_n(\cdot)$ which drastically improves the acceptance rates compared to ABC.



(a) (b) (c) (d)

Fig. 2: *Plots of the estimated parameter values from 100 independent samples of Cauchy location data with $S_n = Median(x_1, \ldots, x_n)$. The true parameter value is $\theta = 10$; the coverage of the 95% ACC-based confidence intervals and the regression adjusted intervals is in parentheses; regression adjusted estimates are denoted with $(*)$; acceptance rates are on the bottom row. (a) ABC estimate for θ using a flat $\pi(\theta)$. (b) ACC estimate for θ using $r_{n1}(\theta)$. (c) ACC estimate for θ using $r_{n2}(\theta)$. (d) Comparison of the distributions on the parameter space defined by $\pi(\theta)$ (gray), $r_{n1}(\theta)$ (solid black), and $r_{n2}(\theta)$ (dotted black).*

Next, consider the case where τ is unknown and θ is known. Consider $\hat{\tau}_S$ as a summary statistic for τ where

$$\hat{\tau}_S = \exp(0.5[Median\{\log \mid (x_i - \hat{\theta}_S)(x_j - \hat{\theta}_S) \mid\}]),$$

for $\hat{\theta}_S = Median(x_1, \ldots, x_n)$ and $1 \leq i, j \leq n$, $i \leq j$. This is the Hodges-Lehman scale estimator for the Cauchy distribution which follows a scale family distribution as shown in Kravchuk & Pollett (2012). Specifically,

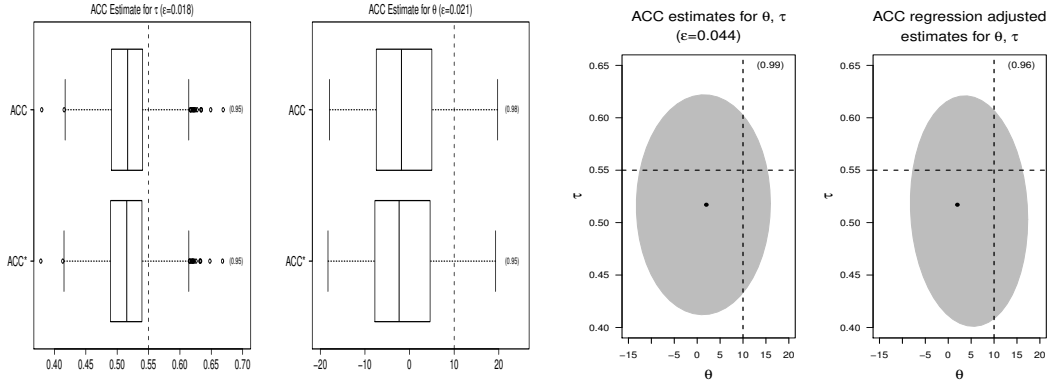
$$n^{1/2}(\log \hat{\tau}_S - \log \tau) \xrightarrow{\mathcal{L}} N(0, 2).$$

Hence, in-line with Theorem 2, we can derive valid frequentist confidence intervals for τ using an $r_n(\tau)$ that satisfies assumptions C2–C4. Here we choose $r_n(\tau) \propto 1/\tau$.

In this simulation, we observe a sample of size $n = 400$ and assume $\theta = 10$ is known. The resulting estimates for τ from ACC and ABC (both with and without the regression adjustment)

and the coverage of the 95% confidence intervals for the true parameter value ($\tau = 0.55$) based on 100 independent runs of ACC are shown in Figure 3. In this example (as well as the next), the application of ACC is the same as that of ABC since we choose $r_n(\tau) \propto 1/\tau$ (and $r_n(\theta, \tau) \propto 1/\tau$ in the next example); however, we emphasize the correct coverage of the true parameter value.

Now, consider the case where both (θ, τ) are unknown. Here, we choose the summary statistics $\hat{\theta}_S = Median(x_1, \ldots, x_n)$ and $\hat{\tau}_S$, defined above. Setting $r_n(\theta, \tau) \propto 1/\tau$, we can again derive valid frequentist confidence regions for (θ, τ). In this simulation, we observe a sample of size $n = 400$. The resulting joint estimates for (θ, τ) (both with and without the regression adjustment) and the coverage of the 95% confidence regions for the true parameter values ($(\theta, \tau) = (10, 0.55)$) based on 100 independent runs of ACC are shown in Fig. 3.



(a) (b) (c)

Fig. 3: *Plots of the estimated parameter values from 100 independent samples of Cauchy data. The true parameter values are denoted by dotted lines; the coverage of the* 95% *ACC-based confidence intervals/regions is in parentheses; regression adjusted estimates are denoted with* $(*)$*. (a) ACC estimate for* τ *supposing* θ *is known. (b) ACC estimate for* θ *for a pivotal summary statistic* $S_n = \bar{x}$ *supposing* τ *is known. (c) ACC estimates and regression adjusted estimates where both* (θ, τ) *are unknown parameters.*

Finally, to close this Cauchy example, we consider an example where the Bernstein-von Mises-type asymptotic theorems do not apply but we can still use ACC to derive valid frequentist inference. Consider again, the case where θ is unknown but τ is known but now suppose we apply ACC with the summary statistic chosen to be the sample mean, $S_n = \bar{x}$. This choice of summary statistic is not asymptotically normal and that the ABC-posterior will never come near the Bayesian target posterior (see Fig. 1). However, $(S_n - \theta) \sim Cauchy\,(0, \tau)$ is a distribution in a location family so by Theorem 1, we can derive valid frequentist confidence intervals for θ with ACC if we choose $r_n(\theta) \propto 1$.

In the numerical study, we observe a sample of size $n = 400$ and assume $\tau = 0.55$ is known. The resulting estimates for θ and the coverage of the 95% confidence intervals for the true parameter value ($\theta = 10$) based on 100 independent runs of ACC are shown in Fig. 3. In this example, the confidence intervals based on ACC are shown to have the correct coverage of the true parameter value despite the asymptotic non-normality of the summary statistic.

This result illustrates an example where the typical Bayesian justification for ABC does not hold since the asymptotic distribution of the summary statistic is non-Gaussian. ACC however,

still provides valid frequentist inference by taking advantage of the exact pivotal structure of a summary statistic without relying on a large sample size.

4·2. *A discretely observed birth-death process*

We also illustrate our approximate computing method using a more complex birth-death process. Consider the problem of estimating the parameters for a discretely observed simple linear birth-death process, known also as a Kendall's process. This model forms a continuous-time Markov chain $\{X_u : 0 \leq u \leq t\}$, where X_u counts the number of individuals present in the population at time u and t is the end time point. Let X_0 be the initial size of the population and let μ and λ be the per-particle death and birth rates, respectively; the continuous Kendall's process is described by the following differential equation:

$$\begin{aligned}\frac{dPr(X_u = b \mid X_0 = a)}{du} = (b-1)\lambda Pr(X_u = b-1 \mid X_0 = a) \\ +(b+1)\mu Pr(X_u = b+1 \mid X_0 = a) \\ -b(\lambda+\mu)Pr(X_u = b \mid X_0 = a).\end{aligned}$$

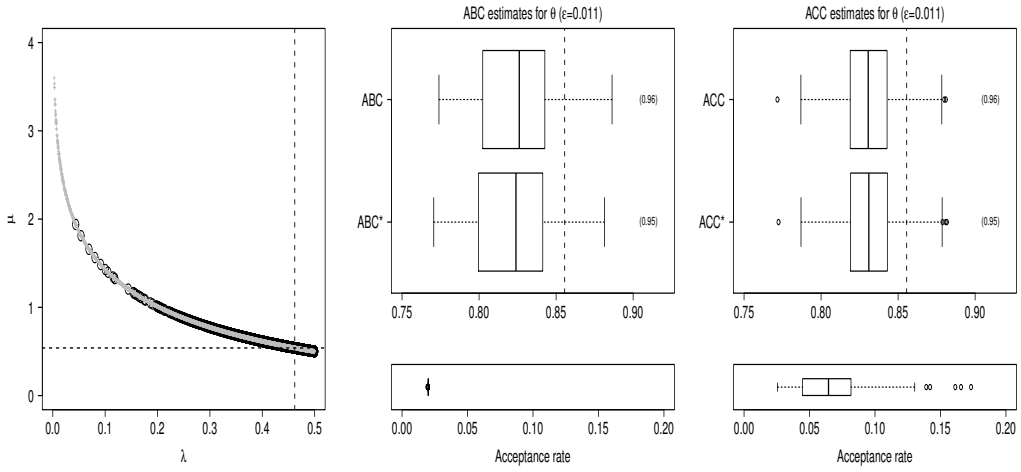
In practice, the continuous process is not observed, instead the data may consist of discretely observed points $\{X_S : s = 0, \tau, 2\tau, \ldots, k\tau = t\}$, where $\tau \in \mathbb{Z}^+$ defines the size of the time intervals over which we observe the birth-death process. Though the mathematical properties of this simple process are well known, for illustrative purposes, we consider the problem of conducting inference on the model parameter $\theta = \lambda/\mu$. As described in Immel (1951) and Keiding (1975), if $\tau = (\log\lambda - \log\mu)/(\lambda - \mu)$, then the estimator $\hat{\theta}_S = \sum_{s=1}^{k\tau} X_S \left\{\sum_{s=0}^{(k-1)\tau} X_S\right\}^{-1}$ is the maximum likelihood estimator for θ in the discretely observed birth-death process. Furthermore, if we consider $\hat{\theta}_S$ to be an estimate derived from a single (k-dimensional) vector observation from a population of size nX_0, then for fixed X_0,

$$(nX_0)^{1/2}(\hat{\theta}_S - \theta) \xrightarrow{\mathcal{L}} N\left(0, \frac{\theta(1-\theta^2)}{1-\theta^k}\right).$$

Therefore C1 holds for this choice of summary statistic and through simulations we verify that ACC-based confidence intervals have at least the nominal coverage of level of the true parameter value θ_0, even for a data-driven choice of $r_n(\theta)$. Furthermore, we compare the performance of ACC to an analogous application of ABC demonstrating the computational advantage of ACC. (See Figure 4.)

To simulate observations from a discrete linear birth-death process, we set the initial population size X_0 and generate a continuous-time Markov chain where $\text{pr}(birth) = \lambda/(\lambda+\mu)$ and $\text{pr}(death) = 1 - \text{pr}(birth)$ and the time until the next event is distributed exponentially with a mean equal to the product of the current size of the population and $(\mu + \lambda)$. The process continues until either the population dies out or until we reach some fixed time point t. For the examples summarized in Fig. (4) we set $X_0 = 300$ and $t = 100$ and chose true parameter values $\mu = 0.54$, $\lambda = 0.462$, therefore $\theta = 0.85\bar{5}$ and $\tau = 2$. We assume it is known that $\tau \approx \{\log(\lambda) - \log(\mu)\}/(\lambda - \mu)$ and that $0 < \theta < 1$, i.e. that $\mu > \lambda$.

To ensure we meet assumptions C2–C4 we define a data-driven $r_n(\theta)$ in-line with the guidelines of Section 3.4. Partition the observed ordered data into smaller subsets of size $k = m^{1/2}$ where m is the number of non-zero observations from the original data set of size n. Then, compute $\hat{\theta}_{Si}$ for $i = 1, \ldots, k$ and these values of $\hat{\theta}_{S_i}$ are independent of each other. Denote the mean



(a) (b) (c)

Fig. 4: *Application of ABC and ACC to data from a birth-death process. (a) A sample from the flat prior $\pi(\theta)$ in ABC (gray) to a sample from the data-dependent distribution $r_n(\theta)$ in ACC (black). Note the more concentrated region defined by $r_n(\theta)$. Compare the estimated θ values and acceptance rates of ABC (b) and ACC (c). The true parameter value is shown by a dotted line, the coverage of the 95% confidence intervals is show in parentheses. Note the improvement in acceptance rates for ACC.*

of these values k by $\bar{\theta}_{S_i}$ and the variance by $\hat{\sigma}^2_{\theta_{Si}}$ and define

$$r_n(\theta) \sim N\left(\bar{\theta}_{Si}, \hat{\sigma}^2_{\theta_{Si}}\right) \mathbb{I}\{0 < \theta < 1\}.$$

The region defined by the analogous flat prior in ABC, $\pi(\theta) \sim U(0,1)$, is more disperse than the region defined by r_n as shown in Fig. 4. In fact, the median acceptance rate for parameter values drawn by $r_n(\theta)$ in ACC is 0.066 while the median acceptance rate for parameter values drawn from the ABC prior is 0.020, illustrating the huge gain in computational efficiency by using a data-dependent $r_n(\cdot)$.

5. DISCUSSION

We have introduced ACC, a likelihood-free method that does not depend on any Bayesian assumptions such as prior information. Rather than compare the output to a target posterior distribution, ACC quantifies the uncertainty in estimation by drawing upon a direct connection to a confidence distribution. This connection guarantees that confidence intervals/regions based on ACC methods capture the truth about the parameters of interest at least at the nominal level and thus we provide theoretical support for ACC-based inference including, but not limited to, the special case where we do have prior information (i.e. ABC). Furthermore, in the case where the selected summary statistic is sufficient, the ACC method is equivalent to maximum likelihood inference. In addition to providing sound theoretical results for inference, the framework of ACC sets the user up for better computational performance by allowing the data to drive the algorithm through the choice of r_n. The potential computational advantage of ACC has been illustrated through several simulation examples.

There are three main sources of error in applying ACC; the error due to the choice of ε_n, the Monte Carlo error due to a finite choice of N, and possibly the asymptotic error due to a finite sample size n. The first type of error is comparable to the error in ABC methods as detailed in Section 3. The Monte Carlo error, which is not discussed in this paper is also common to ABC methods and represents a practical limitation of computational methods. Finally, the error due to a finite sample size may or may not be an issue in ACC, depending on the problem at hand.

Variants of Algorithm 1 (ABC) aim to improve the Monte Carlo sampling from $\Pi_\varepsilon(\theta \mid s_{\text{obs}})$ and are frequently used due to the inefficiency of Algorithm 1 when the prior distribution is very different from the targeted posterior distribution. Such improved sampling techniques are not necessary for Algorithm 2 (ACC) provided the chosen $r_n(\cdot)$ is good in the sense that a large part of the simulated summary statistics lie within the sufficiently small ε neighborhood. For example, $r_n(\cdot)$ chosen as in Section 3·4 works well provided the distribution of the point estimate $\widehat{\theta}$ converges to θ_0 with rate $o(a_n^{-3/5})$. The efficiency of Algorithm 2 can also be seen by comparing it to the following importance sampling variant of Algorithm 1:

[Step1] Generate $\theta_1, \ldots, \theta_N$ from $q(\theta)$ and simulate $s^{(i)}$ from M_θ
[Step2] Accept $s^{(i)}$ when $\mid s^{(i)} - s_{obs} \mid < \varepsilon$ and assign weight $w_i = \pi(\theta_i)/q(\theta_i)$.

For importance sampling ABC, even if the proposal density $q(\cdot)$ is set to $r_n(\cdot)$, Algorithm 1 is less efficient than Algorithm 2 because the importance weight w_i is unbounded while the corresponding weight in Algorithm 2 is unity.

While it may be difficult to find a satisfactory $r_n(\cdot)$ that yields reasonable acceptance probabilities for Algorithm 2, the various improved sampling techniques of ABC can be naturally adapted to improve the performance of ACC. These variants of Algorithm 2 will be more efficient than the corresponding variants of Algorithm 1 provided $r_n(\cdot)$ is closer to the proposal density $q(\cdot)$ than $\pi(\cdot)$.

We find the philosophical interpretation of the results admitted through ACC to be more natural than the Bayesian interpretation of ABC. Within a frequentist setting, it makes sense to view the many different potential confidence distributions produced by ACC using different summary statistics as various choices of estimators. However, within the Bayesian framework, there is no clear way to choose from among the different ABC posteriors due to various choices of summary statistics. In particular, there is an ambiguity in defining the probability measure on the joint space $(\mathcal{P}, \mathcal{X})$ when choosing among different ABC posteriors. Rather than engaging in a pursuit to define a moving target such as this, ACC maintains a clear frequentist interpretation thereby offering a consistently cohesive interpretation of likelihood-free methods.

ACKNOWLEDGEMENT

Acknowledgements should appear after the body of the paper but before any appendices and be as brief as possible subject to politeness. Information, such as contract numbers, of no interest to readers, must be excluded.

SUPPLEMENTARY MATERIAL

Further instructions will be given when a paper is accepted.

APPENDIX 1

Additional Conditions

[C5] There exists some $\delta_0 > 0$ such that $\mathcal{P}_\prime = \{\theta :\mid \theta - \theta_\prime \mid < \delta_\prime\} \subset \mathcal{P}$, $r_n(\theta) \in C^2(\mathcal{P}_\prime)$, and $r_n(\theta_0) > 0$.

[C6] The kernel satisfies (i) $\int vK(v)dv = 0$; (ii)$\prod_{k=1}^{l} v_{i_k} K(v)dv < \infty$ for any coordinates $(v_{i_1}, \ldots, v_{i_l})$ of v and $l \leq p+6$; (iii)$K(v) \propto K(\|v\|_\Lambda^2)$ where $\|v\|_\Lambda^2 = v^T \Lambda v$ and Λ is a positive-definite matrix, and $K(v)$ is a decreasing function of $\|v\|_\Lambda$; (iv) $K(v) = O(\exp\{-c_1\|v\|^{\alpha_1}\})$ for some $\alpha_1 > 0$ and $c_1 > 0$ as $\|v\| \to \infty$.

For C7–C8 define the random variable $W_n(s) = a_n A(\theta)^{-1/2}\{s - \eta(\theta)\}$ and let $\tilde{f}_{W_n}$ be the density for W_n under $\tilde{f}_n(w;\theta)$ and let $\tilde{\tilde{f}}_{W_n}(w;\theta)$ be the density for W_n under the normal approximation model from C1.

[C7] There exists α_n satisfying $\alpha_n/\alpha_m^{2/5} \to \infty$ and a density $r_{max}(w)$ satisfying C6 (ii)-(iii) where $K(v)$ is replaced with $r_{max}(w)$, such that $\sup_{\theta\in\mathcal{P}_0} \alpha \mid \tilde{f}_{W_n}(w;\theta) - \tilde{\tilde{f}}_{W_n}(w;\theta) \mid \leq c_3 r_{max}(w)$ for some positive constant c_3.

[C8] The following statements hold: (i) $r_{max}(w)$ satisfies C6 (iv); and (ii) $\sup_{\theta\in\mathcal{P}_0^C} \tilde{f}_{W_n}(x;\theta) = O(e^{-c_2\|w\|^{\alpha_2}})$ as $\|w\| \to \infty$ for some positive constants c_2 and α_2, and $A(\theta)$ is bounded in $\mathcal{P}$.

[C9] The first two moments, $\int_{\mathbb{R}^d} s\tilde{f}_n(s;\theta)ds$ and $\int_{\mathbb{R}} ss^T\tilde{f}_n(s;\theta)$, exist.

Proof of Lemma 1

The density of r_ε can be expressed by

$$\begin{aligned}\pi_\varepsilon(\theta|s_{\text{obs}}) &\propto \int_{\mathbb{R}^d} \pi(\theta)\tilde{f}_n(s \mid \theta)K\{\varepsilon^{-1}(s - s_{\text{obs}})\}ds \\ &= \pi(\theta)\int \left\{\tilde{f}_n(s_{\text{obs}} \mid \theta) + \tilde{f}_n'(\bar{s} \mid \theta)(\bar{s} - s) + (1/2)\tilde{f}_n''(\bar{s} \mid \theta)(\bar{s} - s)^2\right\} K\{\varepsilon^{-1}(s - s_{\text{obs}})\}ds \\ &\propto \pi(\theta)\tilde{f}_n(s_{\text{obs}}|\theta) + O(\varepsilon^2),\end{aligned}$$

where $\tilde{f}_n''(\cdot \mid \theta)$ is the second derivative of $\tilde{f}_n(\cdot \mid \theta)$ and $\bar{s}$ is a value/vector between s_{obs} and $s_{\text{obs}} + u\varepsilon$. The equality above holds due to a Taylor expansion of $\tilde{f}_n(\cdot \mid \theta)$ with respect to s_{obs} and the final proportion holds using the substitution $u = \varepsilon^{-1}(s - s_{\text{obs}})$ and that $\int_{\mathbb{R}^d} K(u)\,du = 1$ and $\int_{\mathbb{R}^d} uK(u)\,du = 0$.

Proof of the Claim in Section 2

By its definition, $D_n(\cdot) = D(\cdot, sob)$ is a sample-dependent cumulative distribution function on the parameter space. We also have $D_n(\theta_0) = D(\theta_0, sob) = \text{pr}^*\{2\hat{\theta}_\text{S} - \theta_{ACC} \leq \theta_0 \mid sob\} = \text{pr}^*\{\theta_{ACC} - \hat{\theta}_\text{S} \geq \hat{\theta}_\text{S} - \theta_0 \mid sob\} = 1 - G(\hat{\theta}_\text{S} - \theta_0)$. Since $G(t) = \text{pr}\{\hat{\theta}_\text{S} - \theta_0 \leq t\}$, we have $G(\hat{\theta}_\text{S} - \theta_0) \sim Unif(0,1)$ under the probability measure of the random sample population. Thus, as a function of the random S_n, $D_n(\theta_0) = D_n(\theta_0, S_n) \sim Unif(0,1)$. By the univariate confidence distribution definition, $D_n(\cdot)$ is a confidence distribution function.

Furthermore, $D_n(\cdot)$ can provide us confidence intervals of any level. In particular, for any $\alpha \in (0,1)$, $\text{pr}\{\theta_0 \leq D_n^{-1}(1-\alpha)\} = \text{pr}\{D_n(\theta_0) \leq 1-\alpha\} = 1-\alpha$. Thus, $(-\infty, D_n^{-1}(1-\alpha)]$ is a $(1-\alpha)$-level confidence interval. Note that, $D_n(2\hat{\theta}_\text{S} - \theta_{ACC,\alpha}) = \text{pr}^*\{2\hat{\theta}_\text{S} - \theta_{ACC} \leq 2\hat{\theta}_\text{S} - \theta_{ACC,\alpha} \mid sob\} = 1 - \text{pr}^*\{\theta_{ACC} < \theta_{ACC,\alpha} \mid sob\} = 1-\alpha$. So, $D_n^{-1}(1-\alpha) = 2\hat{\theta}_\text{S} - \theta_{ACC,\alpha}$. Therefore, $(-\infty, 2\hat{\theta}_\text{S} - \theta_{ACC,\alpha}]$ is also a $(1-\alpha)$-level confidence interval for θ.

Proof of Lemma 2

Since

$$\begin{aligned}&\mid \text{pr}\{\theta_0 \in \Gamma_{1-\alpha}(S_n)\} - (1-\alpha) \mid = \mid \text{pr}\{W(\theta, S_n) \in A_{1-\alpha} \mid \theta = \theta_0\} - (1-\alpha) \mid \\ \leq &\mid \text{pr}^*\{V(\theta_{ACC}, S_n) \in A_{1-\alpha} \mid S_n\} - (1-\alpha) \mid + \mid \text{pr}\{W(\theta, S_n) \in A_{1-\alpha} \mid \theta = \theta_0\} \\ &\quad -\text{pr}^*\{V(\theta_{ACC}, S_n) \in A_{1-\alpha} \mid S_n\} \mid,\end{aligned}$$

by the definition of $A_{1-\alpha}$ in (3), it follows that $\mid \text{pr}^*\{V(\theta_{ACC}, S_n) \in A_{1-\alpha} \mid S_n\} - (1-\alpha) \mid = o(\delta)$, almost surely. Therefore, by Condition A, we have $\mid \text{pr}\{\theta_0 \in \Gamma_{1-\alpha}(S_n)\} - (1-\alpha) \mid \leq o_p(\delta) + o_p(\varepsilon) =$

$o_p(\varepsilon \vee \delta)$. Furthermore, if Condition A holds almost surely, $\mid \text{pr}\{\theta_0 \in \Gamma_{1-\alpha}(S_n)\} - (1-\alpha) \mid \leq o(\delta) + o(\varepsilon) = o(\varepsilon \vee \delta)$, almost surely.

Proof of Theorem 1

Let's consider a more general pivotal case where $T = \Psi(S_n, \theta_0)$ is free of parameter θ_0. In Algorithm 2, we simulate $\theta^* \sim r_n(\theta)$ and $S_n^* \sim M_{\theta^*}$. Since Ψ is a pivot function, we have $\Psi(S_n^*, \theta^*) \mid \theta^* \sim g(T)$. Denote by $T^* = \Psi(S_n^*, \theta^*)$, it follows that unconditionally, $(T^*, \theta^*) \sim r_n(\theta) g(T)$. In Algorithm 2, we only keep those θ^* which generate S_n^* such that $\mid S_n^*, sob \mid \leq \varepsilon_n$. So conditional on sob, $T_{ACC} = \Psi(sob, \theta_{ACC})$ follows the distribution with density

$$r_{\varepsilon_n}(T \mid sob) \propto \int r_n(\theta) g(T) K_\varepsilon \left(\|u_{T,\theta}, sob\| \right) d\theta \propto g(T) \int r_n(\theta) K_\varepsilon \left(\|u_{T,\theta}, sob\| \right) d\theta$$

In the above formula, $u_{T,\theta}$ is the solution of $T = \Psi(u, \theta)$ for any given values of T and θ. Also, $K_\varepsilon(\cdot) = K(\cdot/\varepsilon)/\varepsilon$ is a scaled kernel function by the bandwidth ε. If $\int r_n(\theta) K_\varepsilon \left(\|u_{T,\theta}, sob\| \right) d\theta$ is free of T and finite, then $\Psi(S_n, \theta_{ACC}) \mid sob \sim G(T)$ as desired.

We now verify that $\int r_n(\theta) K_\varepsilon \left(\|u_{T,\theta}, sob\| \right) d\theta$ is free of T for S_n from a scale family and remark that the proofs for Part 1 and Part 3 are similar. In particular, suppose S_n has distribution $(1/\theta) g(S_n/\theta)$, then $T = \Psi(S_n, \theta) = S_n/\theta \sim g(t)$ is a pivot. So, for any given (t, θ) pair we have $u_{t,\theta} = t\theta$. Thus, with variable transformation $u = (t\theta - s_{obs})/\varepsilon$ and assuming that $K(\cdot)$ is symmetric, we have

$$\begin{aligned} \int r_n(\theta) K_\varepsilon \left(\|u_{T,\theta}, sob\| \right) d\theta &= \int \frac{1}{\varepsilon\theta} K \left(|t\theta - s_{obs}|/\varepsilon \right) d\theta \\ &= \int \frac{1}{\varepsilon u + s_{obs}} K(u) du = s_{obs}^{-1} + o(\varepsilon), \end{aligned}$$

which is free of t.

Proof of Theorem 2

[Wentao, please insert proof here]

Proof of Theorem 3

[Wentao, please insert proof here]

REFERENCES

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

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

CAMERON, E. & PETTITT, A. N. (2012). Approximate bayesian computation for astronomical model analysis: A case study in galaxy demographics and morphological transformation at high redshift. *Monthly Notices of the Royal Astronomical Society* **425**, 44–65.

CREEL, M. & KRISTENSEN, D. (2013). Indirect likelihood inference. *Manuscript, Department of Economics. Columbia University* .

FRAZIER, D. T., MARTIN, G. M., ROBERT, C. P. & ROUSSEAU, J. (2018). Asymptotic properties of approximate Bayesian computation. *Biometrika* .

FREEDMAN, D. A. & BICKEL, P. J. (1981). Some asymptotic theory for the bootstrap. *Annals of Statistics* **9**, 1196–1217.

GOURIEROUX, C., MONFORT, A. & RENAULT, E. (1993). Indirect inference. *Journal of Applied Econometrics* **8**, S85–S118.

IMMEL, E. (1951). Problems of estimation and of hypothesis testing connected with birth-and-death markov processes. Abstract:" Ann. Math. Statisti. 22, 485.

JOYCE, P. & MARJORAM, P. (2008). Approximately sufficient statistics and Bayesian computation. *Statistical Applications in Genetics and Molecular Biology* **7**, 26.

KEIDING, N. (1975). Maximum likelihood estimation in the birth-and-death process. *The Annals of Statistics* **3**, 363–372.

KRAVCHUK, O. Y. & POLLETT, P. K. (2012). Hodges-lehmann scale estimator for Cauchy distribution. *Communications in Statistics- Theory and Methods* **41**, 3621–3632.

LI, W. & FEARNHEAD, P. (2018a). Convergence of regression-adjusted approximate bayesian computation. *Biometrika* **105**, 301–318.

LI, W. & FEARNHEAD, P. (2018b). On the asymptotic efficiency of approximate bayesian computation estimators. *Biometrika* **105**, 286–299.

LIU, R., PARELIUS, J. & SINGH, K. (1999). Multivariate analysis by data depth: descriptive statistics, graphics and inference (with discussion). *Annals of Statistics* **27**, 783 – 858.

MARIN, J.-M., PUDLO, P., ROBERT, C. P. & RYDER, R. J. (2011). Approximate Bayesian computational methods. *Statistics and Computing* **22**, 1167–1180.

MEEDS, E. & WELLING, M. (2015). Optimization Monte Carlo: Efficient and embarrassingly parallel likelihood-free inference. In *Advances in Neural Information Processing Systems*.

ROBINSON, J. D., BUNNEFELD, L., HEARN, J., STONE, G. N. & HICKERSON, M. J. (2014). ABC inference of multi-population divergence with admixture from unphased population genomic data. *Molecular Ecology* **23**, 4458–4471.

SCHWEDER, T. & HJORT, N. L. (2016). *Confidence, Likelihood, Prabability*. Cambridge University Press.

SERFLING, R. (2002). Quantile functions for multivariate analysis: approaches and applications. *Statistica Neerlandica* **56**, 214–232.

SINGH, K. (1981). On the asymptotic accuracy of Efron's bootstrap. *The Annals of Statistics* **9**, 1187–1195.

SINGH, K., XIE, M. & STRAWDERMAN, W. E. (2007). Confidence distribution (CD) - distribution estimator of a parameter. *IMS Lecture Notes* **54**, 132–150.

SISSON, S., FAN, Y. & TANAKA, M. (2007). Seuqential Monte Carlo without likelihoods. *Proceedings of the National Academie of Science USA* **104**, 1760–1765.

XIE, M. & SINGH, K. (2013). Confidence distribution, the frequentist distribution estimator of a parameter: A review. *International Statistical Review* **81**, 3–39.

[*Received* Day $Month$ Year. *Editorial decision on* Day $Month$ Year]