Approximate confidence distribution computing

SUZANNE THORNTON*, WENTAO LI[0], AND MINGE XIE[0]

ABSTRACT

Approximate confidence distribution computing (ACDC) offers a new take on the rapidly developing field of likelihood-free inference from within a frequentist framework. The appeal of this computational method for statistical inference hinges upon the concept of a *confidence distribution*, a special type of estimator which is defined with respect to the repeated sampling principle. An ACDC method provides frequentist validation for computational inference in problems with unknown or intractable likelihoods. The main theoretical contribution of this work is the identification of a matching condition necessary for frequentist validity of inference from this method. In addition to providing an example of how a modern understanding of confidence distribution theory can be used to connect Bayesian and frequentist inferential paradigms, we present a case to expand the current scope of so-called approximate Bayesian inference to include non-Bayesian inference by targeting a confidence distribution rather than a posterior. The main practical contribution of this work is the development of a data-driven approach to drive ACDC in both Bayesian or frequentist contexts. The ACDC algorithm is data-driven by the selection of a data-dependent proposal function, the structure of which is quite general and adaptable to many settings. We explore two numerical examples that both verify the theoretical arguments in the development of ACDC and suggest instances in which ACDC outperform approximate Bayesian computing methods computationally.

KEYWORDS AND PHRASES: Approximate Bayesian inference, Confidence distribution, Computational inference.

1. INTRODUCTION

1.1 Approximate confidence distribution computing

Approximate confidence distribution computing (ACDC) is a new take on likelihood-free inference within a frequentist setting. The development of this computational method for statistical inference hinges upon the modern notion of a *confidence distribution*, a special type of estimator which will be defined shortly. Through targeting this special distribution estimator rather than a specific likelihood or posterior distribution as in variational inference and approximate Bayesian inference, respectively, ACDC provides frequentist validation for inference in complicated settings with an unknown or intractable likelihood where dimension-reducing sufficient summary statistics may not even exist. This work demonstrates another example where confidence distribution estimators connect Bayesian and frequent inference, in the surprising context of computational methods for likelihood-free inference [22, 20].

Let $x_{\text{obs}} = \{x_1, \ldots, x_n\}$ be an observed sample originating from a data-generating model that belongs to some complex parametric family M_θ. Suppose the likelihood function is intractable (either analytically or computationally), but that this model is generative, i.e. given any $\theta \in \mathcal{P}$, we can simulate artificial data from M_θ. Let $S_n(\cdot)$ be a summary statistic that maps the sample space into a smaller dimensional space and $r_n(\theta)$ be a data-dependent function on the parameter space. The simplest version of ACDC is the rejection algorithm labeled Algorithm 1 below,

*Corresponding author.

where $K_\varepsilon(u) = \varepsilon^{-1}K(u/\varepsilon)$ for a kernel function $K(\cdot)$ and ε is a small positive value, referred to as the *tolerance level*.

---

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

Algorithm 1: Accept-reject approximate confidence distribution computing (ACDC)

The output of many iterations of Algorithm 1 are potential parameter values, and these potential parameter values are draws from the probability density

$$q_\varepsilon(\theta \mid s_{\text{obs}}) = \frac{\int_{\mathcal{S}} r_n(\theta)f_n(s \mid \theta)K_\varepsilon(s - s_{\text{obs}})\,ds}{\int_{\mathcal{P}\times\mathcal{S}} r_n(\theta)f_n(s \mid \theta)K_\varepsilon(s - s_{\text{obs}})\,dsd\theta}. \quad (1.1)$$

Here, $f_n(s \mid \theta)$ denotes the likelihood of the summary statistic, implied by the intractable likelihood of the data. Therefore $f_n(s \mid \theta)$ is typically also intractable. We will refer to $f_n(s \mid \theta)$ as an *s-likelihood* to emphasize that this is distinct from a traditional likelihood function $f_n(x_{obs} \mid \theta)$. We denote the cumulative distribution function corresponding to $q_\varepsilon(\theta \mid s_{\text{obs}})$ by $Q_\varepsilon(\theta \mid s_{\text{obs}})$.

The main contribution of this paper is the establishment of a matching condition under which $Q_\varepsilon(\theta \mid s_{\text{obs}})$ is an *approximate confidence distribution* for θ and can be used to derive various types of frequentist inferences. These conditions depend on the

choice of $r_n(\theta)$ but are rather general and we present a strategy for choosing an appropriate data-dependent function in Section 3. Practically, this new perspective allows the data to drive the algorithm in a way that can make it more computationally effective than other existing likelihood-free approaches. Theoretically, this perspective establishes frequentist validation for inference from ACDC based upon general conditions that do not depend on the sufficiency of s_{obs}.

Our justification for these practical and theoretical advantages of ACDC relies on the frequentist notion of a confidence distribution. Some background information on confidence distributions is presented next. To motivate the concept, first consider parameter estimation within a frequentist paradigm. We often desire that our estimators, whether point estimators or interval estimators, have certain properties such as unbiasedness or preform similarly under repeated, randomly sampling. A confidence distribution is an extension of this tradition in that it is a distribution estimate (i.e., it is a sample-dependent distribution function) that satisfies certain desirable properties. Following [22] and [17], we define a confidence distribution as follows:

A sample-dependent function on the parameter space is a CONFIDENCE DISTRIBUTION (CD) *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 valid confidence sets of all levels for θ.*

A confidence distribution 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, unknown quantity and the sampling of data as the random event. A confidence distribution is a sample-dependent function that can be used to estimate the parameter of interest to quantify the uncertainty of the estimation. If $H_n(\cdot)$ is a CD for some parameter θ, then one can simulate $\xi_{CD} \sim H_n(\cdot)$, conditional upon the observed data. We will refer to the random estimator $\xi_{CD} \sim H_n(\cdot)$ as a CD-RANDOM VARIABLE.

The function $r_n(\theta)$ can be viewed as if it is a data-dependent prior. From a frequentist perspective, the data-dependent function $r_n(\theta)$ acts as an initial distribution estimate for θ and Algorithm 1 is a way to update this estimate in search of a better-preforming distribution estimate. This is analogous to any updating algorithm in point estimation requiring an initial estimate that is updated in search for a better-performing one (e.g., say, a Newton-Raphson algorithm or an expectation-maximization algorithm). Of critical concern in this perspective is ensuring that the data is not 'doubly used' for inference. An appropriate choice of the initial distribution estimate, $r_n(\theta)$, addresses this concern and a general strategy for choosing $r_n(\theta)$ is proposed later in Section 3. Therefore, this perspective asserts that $Q_\varepsilon(\theta \mid s_{obs})$ can be used for valid frequentist inference on θ (e.g., deriving confidence sets, p-values, etc.) even if it may sometimes not produce the most efficient estimator (i.e., may not produce the tightest confidence sets for all $\alpha \in (0,1)$ levels).

1.2 Related work on Approximate Bayesian computing (ABC)

Approximate Bayesian computation (ABC) refers to a family of computing algorithms to approximate posterior densities of θ by bypassing direct likelihood evaluations [cf. 6, 4, 16]. The target of an ABC algorithm is the posterior distribution rather than a confidence distribution. A simple rejection sampling ABC proceeds in the same manner as Algorithm 1, but it replaces $\theta_1, \ldots, \theta_N \sim r_n(\theta)$ with $\theta_1, \ldots, \theta_N \sim \pi(\theta)$, a pre-specified prior distribution for θ, in Step 1. In this article, we view ABC as a special case of ACDC where $r_n(\theta) = \pi(\theta)$. The simple rejection sampling ABC is computationally inefficient. Some advanced computing techniques have been used to improve the simple ABC approach. One of such advanced algorithms that is comparable to Algorithm 1 is the an importance sampling version of ABC described in Algorithm 2, where the $r_n(\theta)$ from Algorithm 1 is treated as a reference distribution to facilitate and improve ABC computing efficiency.

---

1. Simulate $\theta_1, \ldots, \theta_N \sim r_n(\theta)$.
For each $i = 1, \ldots, N$,
    2. Simulate $x^{(i)} = \{x_1^{(i)}, \ldots, x_n^{(i)}\}$ from M_θ.
    3. Accept θ_i with probability $K_\varepsilon(s^{(i)} - s_{obs})$,
    where $s_{obs} = S_n(x_{obs})$ and $s^{(i)} = S_n(x^{(i)})$, and assign
    importance weights $w(\theta_i) = \pi(\theta_i)/r_n(\theta_i)$.

Algorithm 2: Importance sampling ABC (IS-ABC)

The theoretical argument behind an approximate Bayesian inference (either using the simple rejection sampling ABC or IS-ABC) depends upon $q_\varepsilon(\theta \mid s_{obs})$ converging to the posterior, $p(\theta \mid x_{obs}) = \pi(\theta)f_n(x_{obs} \mid \theta)/\int \pi(\theta)f_n(x_{obs} \mid \theta)d\theta$, as the tolerance level approaches zero; c.f., e.g. [12] and [2]. However, it is well-known that the quality of this approximation depends not only on the size of ε (and choice of prior) but, also importantly, upon the choice of summary statistic. If s_{obs} is not sufficient (as is generally the case in applications of ABC), then the s-likelihood $f_n(s_{obs} \mid \theta)$ can be very different from the likelihood of the data $f_n(x_{obs} \mid \theta)$ and thus $q_\varepsilon(\theta \mid s_{obs})$ can be a very poor approximation to $p(\theta \mid x_{obs})$, even as ε approaches zero and $n \to \infty$.

For example, consider using Algorithm 1 with two different choices of summary statistic, the sample mean or median, for estimating the location parameter of a random sample ($n = 100$) from $Cauchy(10, 0.55)$. If we suppose $r_n(\theta) \propto 1$, then our algorithm is not data-driven and $r_n(\theta)$ instead acts as an uninformative prior. Hence this example corresponds to an accept-reject version of ABC algorithm. Figure 1 shows $q_\varepsilon(\theta \mid s_{obs})$ for each choice of summary statistic (black lines) where $\varepsilon = 0.005$. The posterior distribution (solid gray lines) does not match well with either approximate ABC posterior distribution $q_\varepsilon(\theta \mid s_{obs})$ because only the entire data vector itself is sufficient in this example. This example demonstrates how a strictly ABC approach that targets $p(\theta \mid x_{obs})$ can produce inconsistent results and even misleading inferential conclusions.

Figure 1: *The gray curves below represent the target posterior distribution (gray lines), $p(\theta \mid x)$, for an $n = 100$ IID sample from $Cauchy(\theta = 10, 0.55)$. The curves in black represent $q_\varepsilon(\theta \mid s_{obs})$ for two different summary statistics, $S_{n_1} = Median(x)$ (left) and $S_{n_2} = \bar{x}$ (right). In each case $\varepsilon = 0.005$.*
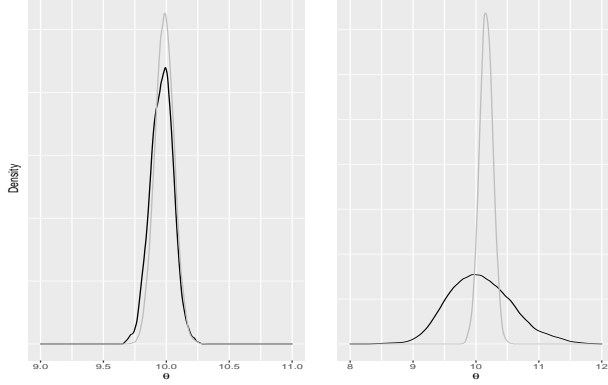
To quote [14]: "the choice of the summary statistic is essential to ensure ABC produces a reliable approximation to the true posterior distribution." Much of the current literature on ABC methods is appropriately oriented towards the selection and evaluation of the summary statistic. The theoretical justification for inference from ACDC on the other hand, does not require an optimal selection of S_n. Although less informative summary statistics may lead to less efficient CDs, the validity of the inferential conclusions can remain intact even for less informative (and non-sufficient) summary statistics. (See Sections 2.1 and 4.)

The large sample theoretical results presented in Section 3 specify conditions under which Algorithm 1 produces an asymptotically normal confidence distribution. These results are similar to those in [9] but our work is distinct because we do not target an approximation to a posterior distribution. Instead, the theoretical results in this section of our paper focus on the properties and performance of ACDC inherited through its connection to CDs. Additionally, in Section 3 we propose a regression-adjustment technique based on that of [9] and [3]. This post-processing step for ACDC is applied to Algorithms 1 and 2 in Section 4 to improve the accuracy of the CDs.

Computationally, the numerical studies in Section 4 also suggest that accept-reject ACDC is more stable than the IS-ABC even when both approaches utilize the same data-driven function $r_n(\theta)$. This difference in performance is due to the fact that the importance weights $w(\theta) = \pi(\theta)/r_n(\theta)$ in in IS-ABC can fluctuate greatly causing numerical instability in the generated parameter values. The steep computing cost associated with the generative model is an expansive area of current research on likelihood-free methods including adaptations that decrease the computing cost of approximate Bayesian methods such as MCMC methods [13] and sequential Monte Carlo techniques[19]. Although an exploration of these adaptations is beyond the scope of this paper, we expect that many of these approaches can be readily applied to improve the computational performance of ACDC as well. The

numerical examples in Section 4 demonstrate how accept-reject ACDC accepts more simulations than IS-ABC suggesting that merely incorporating $r_n(\theta)$ as a data-dependent proposal function is not computationally preferable.

1.3 Notation and outline of topics

In addition to the notation from the introduction, throughout the remainder of paper we will use the following notation. The observed data is $x_{\mathrm{obs}} \in \mathcal{X} \subset \mathbb{R}^n$, the summary statistic is a mapping $S_n : \mathcal{X} \to \mathcal{S} \subset \mathbb{R}^d$ and the observed summary statistic is $s_{\mathrm{obs}} = S_n(x_{\mathrm{obs}})$. The parameter of interest is $\theta \in \mathcal{P} \subset \mathbb{R}^p$ with $p \leq d < n$; i.e. the number of unknown parameters is no greater than the number of summary statistics and the dimension of the summary statistic is smaller than the dimension of the data. If some function of S_n is an estimator for θ, we will denote this function by $\hat{\theta}_S$.

The next section presents the core theoretical result of this paper which establishes a necessary condition for ACDC methods to produce a valid CD, thereby establishing ACDC as a likelihood-free method that provides valid frequentist inference. Section 3 presents general large sample conditions for ACDC that produce asymptotic CDs and establishes precise conditions for an appropriate choice of the data-dependent function $r_n(\theta)$. Section 4 contains two numerical examples that verify the inferential conclusions of ACDC and illustrate the computational advantages of this data-driven algorithm. Section 5 concludes with a brief discussion. All proofs for Sections 2 and 3 are contained in the Appendix (and supplementary material).

2. ESTABLISHING FREQUENTIST GUARANTEES

2.1 General conditions

In this section, we formally establish conditions under which ACDC can be used to produce confidence regions with guaranteed frequentist coverages for any significance level. To motivate our main theoretical result, first consider the simple case of a scalar parameter and a function $\hat{\theta}_{\mathrm{S}} = \hat{\theta}(S_n)$ which maps the summary statistic, $S_n \in \mathcal{S}$, into the parameter space $\mathcal{P}$.
Claim: If

$$\mathrm{pr}^*(\theta - \hat{\theta}_S \leq t \mid S_n = s_{\mathrm{obs}}) = \mathrm{pr}(\hat{\theta}_S - \theta \leq t \mid \theta = \theta_0), \quad (2.1)$$

then $H_n(t) \stackrel{def}{=} 1 - Q_\varepsilon(2\hat{\theta}_S - t \mid s_{obs})$ is a CD for θ.

In the claim, $\mathrm{pr}^*(\cdot \mid S_n = s_{\mathrm{obs}})$ refers to the probability measure on the simulation, conditional on the observed summary statistic, and $\mathrm{pr}(\cdot \mid \theta = \theta_0)$ is the probability measure on the data before it is observed. The proof of this claim (provided in the appendix) involves showing that $H_n(\theta_0)$ follows a uniform distribution. Once this is established, any $(1-\alpha)100\%$ level confidence interval for θ can be found by inverting the confidence distribution, $H_n(t)$.

This claim is conceptually similar to the bootstrap central limit theorem which states (conditions under which) the variability of the bootstrap estimator matches the variability induced by

the random sampling procedure. Equation (2.1) instead matches the variability induced by the Monte-Carlo sampling to the random sampling variability. On the left hand side, $\hat{\theta}_S$ is fixed given s_{obs} and the (conditional) probability measure is defined with respect to the Monte-Carlo copies of θ. Thus θ on the left hand side of this equation plays the role of a CD random variable. On the right hand side, the probability measure is defined with respect to the sampling variability where θ is the true parameter value. See also [20] for more discussions of similar matchings to link Monte-Carlo randomness with sample randomness across Bayesian, fiducial and frequentist paradigms.

The main condition necessary for valid frequentist inference from ACDC methods is a generalization of the claim above for vector θ.

Condition 1. *For $\mathfrak{B}$ a Borel set on $\mathbb{R}^k$,*

$$\sup_{A \in \mathfrak{B}} \big\| \text{pr}^* \{ V(\theta, S_n) \in A \mid S_n = s_{\text{obs}} \}$$
$$-\text{pr} \{ W(\theta, S_n) \in A \mid \theta = \theta_0 \} \big\| = o_p(\delta_{n,\varepsilon}),$$

where $\text{pr}^*(\cdot \mid Sn = s_{\text{obs}})$ refers to the probability measure on the simulation, conditional on the observed summary statistic, and $\text{pr}(\cdot \mid \theta = \theta_0)$ is the probability measure on the data before it is observed and finally $\delta_{n,\varepsilon}$ is a positive rate of convergence that depends on n and ε.

Rather than consider only the linear functions $(\theta - \hat{\theta}_S)$ and $(\hat{\theta}_S - \theta)$, Condition 1 considers any functions $V(\theta, S_n)$ and $W(\theta, S_n)$. (For example, the claim above is a special case of Condition 1 where $V(t_1, t_2) = -W(t_1, t_2) = t_1 - \hat{\theta}_S(t_2)$.) We use the notation $pr^*\{\cdot \mid S_n = s_{obs}\}$ because this probability measure is defined over a transformation of the $\theta \sim Q_\varepsilon(\cdot \mid s_{obs})$.

Furthermore, Condition 1 permits the parameter space and the sample space of the summary statistic to be different from each other. In short, a matching condition on the relationship between two general, multi-dimensional mappings, $V, W : \mathcal{P} \times \mathcal{S} \to \mathbb{R}^k$ is the key to establishing when ACDC can be used to produce a confidence distribution for θ.

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

$$\text{pr}^* \{ V(\theta, S_n) \in A_{1-\alpha} \mid S_n = s_{\text{obs}} \} = (1 - \alpha) + o(\delta'), \quad (2.2)$$

where $\delta' > 0$ is a pre-selected small positive precision number, designed to control Monte-Carlo approximation error. If Condition 1 holds, then

$$\Gamma_{1-\alpha}(s_{\text{obs}}) \stackrel{\text{def}}{=} \{ \theta : W(\theta, s_{\text{obs}}) \in A_{1-\alpha} \} \subset \mathcal{P} \quad (2.3)$$

is a level $(1 - \alpha)100\%$ confidence set for θ_0. We summarize this in the following lemma which is proved in the appendix.

Lemma 1. *Suppose there exist mappings V and $W : \mathcal{P} \times \mathcal{S} \to \mathbb{R}^k$ such that Condition 1 holds. Then, $\text{pr}\{\theta_0 \in \Gamma_{1-\alpha}(S_n) \mid \theta = \theta_0\} = (1 - \alpha) + o_p(\delta)$, where $\delta = \max\{\delta_{n,\varepsilon}, \delta'\}$. If Condition 1 holds almost surely, then $\text{pr}\{\theta_0 \in \Gamma_{1-\alpha}(S_n) \mid \theta = \theta_0\} \stackrel{\text{a.s.}}{=} (1 - \alpha) + o(\delta)$.*

Nowhere in Lemma 1 is the sufficiency (or near sufficiency) of S_n required. Of course, if the selected summary statistic happens to be sufficient, then inference from the CD with be equivalent to maximum likelihood inference. Furthermore, Lemma 1 may hold for finite n provided Condition 1 does not require $n \to \infty$, i.e. provided $\delta_{n,\varepsilon}$ only depends on ε. Later in this section we will consider a special case of Lemma 1 that may be independent of sample-size.

In the next sections we explore some specific situations in which Condition 1 holds. First however, we relate equation (2.2) to a random sample from $Q_\varepsilon(\cdot \mid s_{\text{obs}})$ for vector θ. Suppose θ_i', $i = 1, \ldots, m$, are m draws from $Q_\varepsilon(\cdot \mid s_{\text{obs}})$ and let $v_i = V(\theta_i', s_{\text{obs}})$. The set $A_{1-\alpha}$ may be a $(1 - \alpha)100\%$ contour set of $\{v_1, \ldots, v_m\}$ such that $o(\delta') = o(m^{-1/2})$. For example, we can directly use $\{v_1, \ldots, v_m\}$ to construct a $100(1 - \alpha)\%$ depth contour as $A_{1-\alpha} = \{\theta : (1/m) \sum_{i=1}^m \mathbb{I}\{\hat{D}(v_i) < \hat{D}(\theta)\} \geq \alpha\}$, where $\hat{D}(\cdot)$ is an empirical depth function on $\mathcal{P}$ computed from the empirical distribution of $\{v_1, \ldots, v_m\}$. See, e.g., [18] and [11] for more on the development of data depth and depth contours in nonparametric multivariate analysis.

2.2 Finite sample size case

We now explore a special case of Lemma 1 where the mappings V and W correspond an approximate pivot statistic. We call a mapping $T = T(\theta, S_n)$ from $\mathcal{P} \times \mathcal{S} \to \mathbb{R}^d$ an *approximate pivot statistic*, if

$$\text{pr}\{T(\theta, S_n) \in A \mid \theta = \theta_0\} = \int_{t \in A} g(t)dt \{1 + o(\delta'')\}, \quad (2.4)$$

where $g(t)$ is a density free of θ, $A \subset \mathbb{R}^d$ is any Borel set, and δ'' is either zero or a small number (tending to zero) that may or may not depend on the sample size n. For example, suppose $S_n|\theta = \lambda \sim \text{Poisson}(\lambda)$. Then, $T(\lambda, S_n) = (S_n - \lambda)/\sqrt{\lambda}$ is an approximate pivot when λ is large, and the density function is $\phi(t)\{1 + o(\lambda^{-1})\}$, where $\phi(t)$ the density function of the standard normal distribution [5]. The usual pivotal cases are special examples of approximate pivots that may not rely on large sample theory. Examples of approximate pivots where δ'' is a function of n are discussed later in Section 3.

Theorem 1. *Suppose $T = T(\theta, S_n)$ is an approximate pivot statistic that is differentiable with respect to the summary statistic and, for given t and θ, let $s_{t,\theta}$ denote a solution to the equation $t = T(\theta, s)$. If*

$$\int_{\mathcal{P}} r_n(\theta) K_\varepsilon \left(s_{t,\theta} - s_{\text{obs}} \right) d\theta = C, \quad (2.5)$$

where C is a constant free of t, then, Condition 1 holds almost surely, for $V(\theta, S_n) = W(\theta, S_n) = T(\theta, S_n)$.

A direct implication of Theorem 1 is that $\Gamma_{1-\alpha}(s_{\text{obs}})$, as defined in (2.3), is a level $(1 - \alpha)100\%$ confidence region where $\text{pr}\{\theta \in \Gamma_{1-\alpha}(S_n) \mid \theta_0\} \stackrel{\text{a.s.}}{=} (1 - \alpha) + o(\delta)$.

The assumption in equation (2.5) needs to be verified on a case-by-case basis. Location and scale families contain natural pivot statistics and satisfy these conditions. This is formally stated in the following corollary.

Corollary 1. *(a) Suppose S_n is a point estimator for μ such that $S_n \sim g_1(S_n - \mu)$ and suppose $r_n(\mu) \propto 1$. Then, for any u,*

$$|pr^*(\mu - S_n \le u \mid S_n = s_{obs}) - pr(S_n - \mu \le u \mid \mu = \mu_0)| \overset{a.s.}{=} o(1).$$

(b) Suppose S_n is a point estimator for σ such that $S_n \sim g_2(S_n/\sigma)/\sigma$ and suppose $r_n(\sigma) \propto 1/\sigma$. then, for any $v > 0$,

$$\left| pr^* \left(\frac{\sigma}{S_n} \le v \big| S_n = s_{obs} \right) - pr \left(\frac{S_n}{\sigma} \le v \big| \sigma = \sigma_0 \right) \right| \overset{a.s.}{=} o(1).$$

(c) If $S_{n,1}$ and $S_{n,2}$ are point estimators for μ and σ, respectively, where $S_{n,1} \sim g_1\{(S_{n,1} - \mu)/\sigma\}/\sigma$ and $S_{n,2} \sim g_2(S_{n,2}/\sigma)/\sigma$ are independent and if $r_n(\mu, \sigma) \propto 1/\sigma$, then, for any u and any $v > 0$,

$$\left| pr^* \left[\begin{pmatrix} \mu - S_{n,1} \le u \\ \frac{\sigma}{S_{n,2}} \le v \end{pmatrix} \Big| \begin{pmatrix} S_{n,1} \\ S_{n,2} \end{pmatrix} = \begin{pmatrix} s_{1,obs} \\ s_{2,obs} \end{pmatrix} \right] \right.$$
$$\left. - pr \left[\begin{pmatrix} \mu - S_{n,1} \le u \\ \frac{S_{n,2}}{\sigma} \le v \end{pmatrix} \Big| \begin{pmatrix} \mu \\ \sigma \end{pmatrix} = \begin{pmatrix} \mu_0 \\ \sigma_0 \end{pmatrix} \right] \right| \overset{a.s.}{=} o(1).$$

Consequently, $H_1(S_{n,1}, x) = \int_{-\infty}^{x} g_1(S_{n,1} - u) du$ is a CD for μ and $H_2(S_{n,2}^2, x) = 1 - \int_0^x g_2(\hat{\sigma}_S/u) du$, is a CD for σ.

Note that Theorem 1 and Corollary 1 cover some finite sample size scenarios, including the Cauchy example discussed in Section 1. For this example, Corollary 1 (part (a)) asserts that the different posterior approximations obtained by approximate Bayesian computing with either $S_n = Median(x)$ or $S_n = \bar{x}$ are both CDs. That is, both densities in black in Figure 2 are densities for confidence distributions of θ. These distribution estimators lead to valid frequentist inference even though neither summary statistic is sufficient. This development represents a departure from the typical asymptotic arguments for likelihood-free computational inference.

This section has considered the case in which the tolerance level, ε, does not necessarily depend on the sample size n. In the next section however, the tolerance may depend on the sample size and so we adopt the notation ε_n to reflect this.

3. LARGE SAMPLE THEORY

3.1 A Bernstein-von Mises theorem for ACDC

In the Bayesian ABC framework, Condition 1 holds as $n \to \infty$ by selecting a ε_n that decreases to zero at a certain rate. [10]. We now verify Condition 1 holds more generally for ACDC methods that use a data-dependent $r_n(\theta)$, in a large sample setting. The results presented here are generalizations of results in [9] and [10]. Roughly speaking, the next theorem establishes that the distribution of a centered random draw from $Q_\varepsilon(\theta \mid s_{obs})$ and the distribution of its centered expectation (before the data is observed), i.e. $\int \theta \, dQ_\varepsilon(\theta \mid S_n)$, are asymptotically the same.

The next condition concerns the asymptotic behavior of the summary statistic is crucial for the proofs of the theorems in this section (see Appendix F).

Condition 2. *There exists a sequence $\{a_n\}$, satisfying $a_n \to \infty$ as $n \to \infty$, a d-dimensional vector $s(\theta)$, a $d \times d$ matrix $A(\theta)$, and some $\delta_0 > 0$ such that for $S_n \sim f_n(\cdot \mid \theta)$ and all $\theta \in \mathcal{P}_0 \overset{def}{=} \{\theta : \|\theta - \theta_0\| < \delta_0\} \subset \mathcal{P}$,*

$$a_n\{S_n - s(\theta)\} \overset{d}{\to} N\{0, A(\theta)\}, \text{ as } n \to \infty,$$

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

This is a standard condition but, notably, does not depend on the sufficiency of this statistic. Because of this, we refrain from discussing this condition further so we may instead focus on our main contribution, the development of the following regulatory conditions on $r_n(\theta)$.

Condition 3. *For all $\theta \in \mathcal{P}_0$, $r_n(\theta) \in C^2(\mathcal{P}_0)$ and $r_n(\theta_0) > 0$.*

Condition 4. *There exists a sequence $\{\tau_n\}$ such that $\tau_n = o(a_n)$ and $\sup_{\theta \in \mathcal{P}_0} \tau_n^{-p} r_n(\theta) = O_p(1)$.*

Condition 5. *There exists constants m, M such that $0 < m < | \tau_n^{-p} r_n(\theta_0) | < M < \infty$.*

Condition 6. *It holds that $\sup_{\theta \in \mathbb{R}^p} \tau_n^{-1} D\{\tau_n^{-p} r_n(\theta)\} = O_p(1)$.*

Condition 3 is a general assumption regarding the differentiability of $r_n(\theta)$ within an open neighborhood of the true parameter value. Condition 4 and 5 essentially require $r_n(\theta)$ to be more dispersed than the s-likelihood within a compact set containing θ_0. They require $r_n(\theta)$ converge to a point mass more slowly than $f_n(\theta \mid s_{obs})$. Condition 6 requires the gradient of the standardized version of $r_n(\theta)$ to converge with rate τ_n. These are relatively weak conditions and can be satisfied with locally asymptotic $r_n(\theta)$, for example. Of course, a flat prior used in approximate Bayesian inference also satisfies Condition 3–6. The proofs of the theorems in this section also require additional conditions (Conditions 7-10 of Appendix F) that are typical of BvM-type theorems. These additional conditions are not presented here for readability reasons and because they do not directly relate to $r_n(\theta)$ which is our emphasis.

Theorem 2. *Let $\hat{\theta}_S = \hat{\theta}(S_n) = \int \theta \, dQ_\varepsilon(\theta \mid S_n)$. Assume $r_n(\theta)$ satisfies Condition 3–6 above and also Conditions 7–10 in the supplementary material. If $\varepsilon_n = o(a_n^{-1})$ as $n \to \infty$, then Condition 1 is satisfied with $V(\theta, S_n) = a_n \left(\theta - \hat{\theta}_{s_{obs}} \right)$ and $W(\theta, S_n) = a_n \left(\hat{\theta}_S - \theta \right)$.*

Theorem 2 says when $\varepsilon_n = o(a_n^{-1})$, the coverage of $\Gamma_{1-\alpha}(s_{obs})$ is asymptotically correct as n and the number of accepted parameter values increase to infinity. In practice, $\hat{\theta}_S$ typically will not have a closed form. To construct $\Gamma_{1-\alpha}(s_{obs})$, the value of $\hat{\theta}$ at $S_n = s_{obs}$ can be estimated using the accepted parameter values from ACDC. Here Condition 1 is satisfied by generalizing the limit distributions of the approximate posterior

in [9] so they hold also for $Q_\varepsilon(\theta \mid s_{obs})$, when $\varepsilon_n = o(a_n^{-1})$. Specifically, for A defined as in equation (2.2),

$$\sup_{A \in \mathfrak{B}^p} \left| \int_{\{\theta:\, a_n(\theta - \hat\theta) \in A\}} dQ_\varepsilon(\theta \mid s_{obs}) - \right.$$
$$\left. \int_A N\{t; 0, I(\theta_0)^{-1}\} \, dt \right| \xrightarrow{\text{P}} 0 \qquad (3.1)$$

and

$$a_n(\hat\theta - \theta_0) \xrightarrow{\text{d}} N\{0, I(\theta_0)^{-1}\}, \qquad (3.2)$$

as $n \to \infty$, where $I(\theta_0)$ is a non-singular matrix defined in Condition 2. Thus inference based on $Q_\varepsilon(\theta \mid s_{obs})$ is valid for $n \to \infty$ regardless of whether or not $r_n(\theta)$ depends on the data. For the same tolerance level, Theorem 2 asserts that the limiting distribution of $Q_\varepsilon(\theta \mid S_n)$ matches the limiting distribution of the approximate posterior from [9] which is the output distribution of the accept-reject version of ABC. In comparison however, Algorithm 1 has a better acceptance rate since the data-dependent $r_n(\theta)$ will concentrate more probability mass around θ_0 than a typical prior.

Although inference from ACDC is validated with $\varepsilon_n = o(a_n^{-1})$, a well-known issue in approximate Bayesian literature is that this tolerance level is too small in practice, causing the acceptance rate to degenerate as $n \to \infty$ for any proposal distribution [10]. Obviously ACDC methods will suffer from this same issue. (For an example with Normal data, see Appendix E.)

One remedy that relaxes the restriction on ε_n is to post-process the sample from $Q_\varepsilon(\theta \mid s_{obs})$ with a regression adjustment[1]. When the data-generating model is correctly specified, the regression adjusted sample correctly quantifies the CD uncertainty and yields an accurate point estimate with ε_n decaying at a rate of $o(a_n^{-3/5})$[9].

Let $\theta^* = \theta - \beta_\varepsilon(s - s_{obs})$ be the post-processed sample from $Q_\varepsilon(\theta \mid s_{obs})$, where β_ε is the minimizer from

$$\begin{pmatrix} \alpha_\varepsilon \\ \beta_\varepsilon \end{pmatrix} = \operatorname*{arg\,min}_{\alpha \in \mathbb{R}^p, \beta \in \mathbb{R}^{d \times p}} E_\varepsilon \left\{ \|\theta - \alpha - \beta(s - s_{obs})\|^2 \mid s_{obs} \right\}$$

for expectation under the joint distribution of accepted θ values and corresponding summary statistics.

Theorem 3. Under the conditions of Theorem 2, if $\varepsilon_n = o\left(a_n^{-3/5}\right)$ as $n \to \infty$, Condition 1 holds with $V(\theta, S_n) = a_n(\theta^* - \hat\theta^*_{s_{obs}})$ and $W(\theta, S_n) = a_n(\hat\theta^*_S - \theta)$, where $\hat\theta^*_S$ is the expectation of the post-processed observations of the CD random variable.

Here, Condition 1 is implied by the following convergence results (where A defined as in equation (2.2)),

$$\sup_{A \in \mathfrak{B}^p} \left| \int_{\{\theta:\, a_n(\theta - \hat\theta^*) \in A\}} dQ^*_\varepsilon(\theta \mid s_{obs}) - \right.$$
$$\left. \int_A N\{t; 0, I(\theta_0)^{-1}\} \, dt \right| \xrightarrow{\text{P}} 0,$$

and

$$a_n(\hat\theta^*_S - \theta_0) \xrightarrow{\text{d}} N\{0, I(\theta_0)^{-1}\},$$

as $n \to \infty$. The limiting distributions above are the same as those in (3.1) and (3.2), therefore $\Gamma_{1-\alpha}(s_{obs})$ constructed using the post-processed sample achieve the same efficiency as those using original ACDC sample of θ values. The benefit of permitting larger tolerance levels is a huge improvement in the computing costs associated with ACDC.

3.2 Designing r_n

Condition 4 implies that in practice, one must take care to choose $r_n(\theta)$ so that its growth with respect to the sample size is slower than the growth of the s-likelihood. In this section we propose a generic algorithm to construct such an $r_n(\theta)$ based on sub-setting the observed data.

Notably, there is a trade-off in ACDC inference between faster computations and guaranteed coverage of the approximate CD based confidence intervals (or regions). When $r_n(\theta)$ grows at a similar rate as the s-likelihood for $n \to \infty$, the computing time for ACDC methods may be reduced but this risks violating Conditions 4–6. If these assumptions are violated, the resulting simulations do not necessarily form a CD and consequently, inference may not be valid in terms of producing confidence sets with guaranteed coverage. Therefore, $r_n(\theta)$ should be designed such that its convergence rate is bounded away from that of the s-likelihood. The minibatch scheme presented below is one way to ensure $r_n(\theta)$ is approriately bounded.

Assume that a point estimator $\hat\theta_S(z)$ of θ can be computed for a dataset, z, of any size.

Minibatch scheme

1. Choose k subsets of the observations, each with size n^ν for some $0 < \nu < 1$.
2. For each subset z_i of x_{obs}, compute the point estimate $\hat\theta_{S,i} = \hat\theta_S(z_i)$, for $i = 1, \ldots, k$.
3. Let $r_n(\theta) = (1/kh) \sum_{i=1}^k K\left\{ h^{-1}\|\theta - \hat\theta_{S,i}\| \right\}$, where $h > 0$ is the bandwidth of the kernel density estimate using $\{\hat\theta_{S,1}, \ldots, \hat\theta_{S,k}\}$ and kernel function K.

If $\hat\theta_S$ is consistent, then for $\nu < 3/5$, $r_n(\theta)$ as obtained by this minibatch procedure will satisfy Conditions 4–6. Based on our experience, if n is large one may simply choose $\nu = 1/2$ to partition the data. For small n, say $n < 100$, it is better to select $\nu > 1/2$ and to overlap the subsets (or "mini" batches of the observed data) so that each subset contains a reasonable number of observations. For a given summary statistic, there are many methods to construct this type of point estimator including: a minimum distance-based optimizer [8, 15], the synthetic likelihood method and its variants [21, 7], or accept-reject ACDC with $\hat\theta_S = E\{\theta \mid S_n(z_i)\}$, the s-likelihood-based expectation over a subset of the observed data. The choice of $\hat\theta_S$ does not need to be an accurate estimator since it is only used to construct the initial rough estimate of a CD for θ. But a heavily biased $\hat\theta_S$ causes biases in confidence sets derived from the CD, since $r_n(\theta)$ does not cover parameter values resulting in high values of $f_n(s \mid \theta)$ very well. In practice, the computing cost will depend on which

particular optimization scheme is followed. However, a full study on the selection of $\hat{\theta}_S$ is beyond the scope of this paper.

The computational cost associated with implementing the minibatch scheme is comparable to the cost of constructing a proposal distribution for IS-ABC methods. Multiple runs to compute $\hat{\theta}_{S,i}$ values can be parallelized easily and any procedure to obtain a proposal distribution for IS-ABC can be applied on the mini batches of data to yield a point estimate for θ. For example, for each subset z_i, the conditional mean $E\{\theta \mid S_n(z_i)\}$ can be estimated by population Monte Carlo ABC on $S_n(z_i)$. This is not any more computationally expensive than computing the same estimate on the full data. This, together with the fact that accept-reject ACDC accepts more simulations than IS-ABC, make ACDC the favorable choice in terms of overall computational performance. The numerical examples in Section 4 support this conclusion.

At this point, our reader may wonder if $\hat{\theta}_S$, can be computed, why not simply use a non-parametric bootstrap method to construct confidence sets? Although it requires no likelihood evaluation, this method has two significant drawbacks. First, the non-parametric bootstrap method is heavily affected by the quality of $\hat{\theta}_S$. For example, a bootstrapped confidence interval for θ is based on quantiles of $\hat{\theta}_S$ from simulated data. A poor estimator typically leads to poor performing confidence sets. In contrast, in ACDC methods, $\hat{\theta}_S$ is only used to construct the initial distribution estimate which is then updated by the data. Second, when it is more computationally expensive to obtain $\hat{\theta}_S$ than the summary statistic, the non-parametric bootstrap will be much more costly than ACDC methods since $\hat{\theta}_S$ must be calculated for each pseudo data set. Example 4.2 in the next section illustrates such an example.

4. NUMERICAL EXAMPLES

4.1 Location and scale parameters for Cauchy data

In the Cauchy example presented in Figure 1 we saw how the lack of a sufficient summary statistic can change the validity of inferential conclusions from an approximate Bayesian inference approach. Through a CD perspective however, the inferential conclusions from ACDC are valid under the frequentist criterion even if the summary statistic is not sufficient. Provided Condition 1 is satisfied, different summary statistics produce different CDs. Here we present a continuation of this Cauchy example where random data ($n = 400$) is drawn from a $Cauchy(\theta, \tau)$ distribution with data-generating parameter values $(\theta_0, \tau_0) = (10, 0.55)$. We investigate the performance of 500 independent 95% confidence intervals for θ alone (settings one and two) and τ alone (setting three) and 500 independent 95% confidence regions when both parameters are unknown (settings four and five).

In each setting, the confidence intervals (regions) are generated for both accept-reject ACDC and IS-ABC utilizing the same minibatch scheme to construct r_n with the median and/or the median absolute deviation (MAD) as point estimators and $v = 1/2$.

Table 1. *Coverage proportions of confidence sets from ACDC applied to Cauchy data under five different settings. Coverage is calculated over* 500 *independent runs that draw a* $n = 400$ *IID sample from a* $Cauchy(\theta = 10, \tau = 0.55)$ *distribution. The Monte Carlo sample size for both algorithms is* $50,000$ *and the nominal coverage level in every setting is* 95%. *The last column displays the median ratio of the sizes of confidence sets from accept-reject ACDC divided by those from IS-ABC.*

	Acceptance proportion	ACDC Coverage	IS-ABC Coverage	Ratio of Widths/ Volumes
Setting 1: θ unknown				
$S_n = Median(x)$	0.005	0.93	0.94	0.94
	0.05	0.94	0.94	0.94
	0.10	0.93	0.94	0.94
Setting 2: θ unknown				
$S_n = \bar{x}$	0.005	0.97	0.98	0.65
	0.05	0.97	0.98	0.60
	0.10	0.97	0.98	0.56
Setting 3: τ unknown				
$S_n = MAD(x)$	0.005	0.93	0.94	1.00
	0.05	0.92	0.94	1.00
	0.10	0.93	0.94	1.00
Setting 4: $(\theta, \tau)'$ both unknown				
$S_n = \begin{pmatrix} \bar{x} \\ SD(x) \end{pmatrix}$	0.005	0.96	0.96	0.58
	0.05	0.99	0.98	0.48
	0.10	0.99	0.98	0.47
Setting 5: $(\theta, \tau)'$ both unknown				
$S_n = \begin{pmatrix} Median(x) \\ MAD(x) \end{pmatrix}$	0.005	0.91	0.93	0.98
	0.05	0.94	0.96	1.00
	0.10	0.94	0.97	1.00

The main difference in the two algorithms is the use of r_n. In former r_n is a data-driven initial CD estimate whereas the latter represents a Bayesian approach that assumes an uninformative prior on the parameter space and employs $r_n(\theta)$ as the proposal distribution for the importance sampling updates. Both algorithms are improved by adapting the regression adjustments mentioned in Section 3.1, so the output for every run of each algorithm is post-processed in this manner.

Table 1 compares frequentist coverage proportions of confidence regions from both algorithms. The acceptance proportion determines how many simulated parameter values are kept and thus is directly related to the tolerance level. Most coverage rates are close to the nominal levels when the acceptance proportion is small, which is expected from the asymptotic theory in Section 3. Overall the coverage performance is similar for both algorithms. For settings one, three, and five with informative summary statistics, both algorithms give similar confidence regions which undercover a bit in the finite-sample regime. For settings two and four with less informative summary statistics, accept-reject ACDC is preferable because it produces tighter confidence bounds.

The main reason for the favorable performance of accept-reject ACDC is related to the skewed importance weights for IS-ABC. This can be seen in the sizes of the confidence sets of the two algorithms in Table 1 but is even more clear when comparing the CD densities of each method as in Figure 2. Figure 2 shows the impact of importance weights in IS-ABC on the variances of

Figure 2: *These are densities of point estimators from accept-reject ACDC (red) and IS-ABC (black) for the 500 independent data sets for each of the five settings in Table 1. Additionally, this figure shows a box plot of the relative sizes of the 500 confidence sets, that is, the length (or volume) of regions produced by accept-reject ACDC divided by those of IS-ABC.*
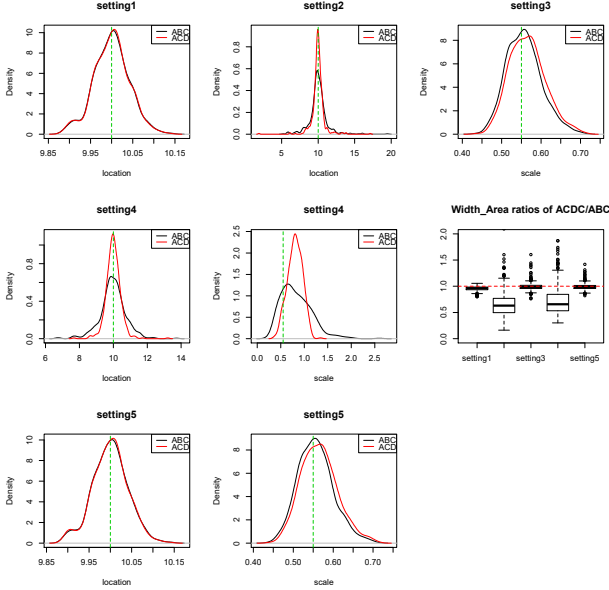
point estimators and CDs. For settings 1,3 and 5 where an informative summary statistic is used, the importance weights do not much affect either the point estimator or resulting CDs. In these cases, $r_n(\theta)$ is a good proposal distribution according to the criteria in [10]. For settings 2 and 4 however, where the summary statistic is less informative, Figure 2 shows how the importance weights inflate both point estimate and CD variances with Monte Carlo variation in IS-ABC. One reason for the severe skewedness in the importance weights $\pi(\theta)/r_n(\theta)$ is that the high variance of S_n means more parameter simulations are accepted in the tails of $r_n(\theta)$. This results in broader confidence regions for IS-ABC than accept-reject ACDC.

This numerical study validates inference for both accept-reject ACDC and IS-ABC even in the case where typical asymptotic arguments do not apply (settings 2 and 4). Furthermore, this example demonstrates two valid but distinct uses of the minibatch scheme for constructing a data driven distribution estimator. In Algorithm 1, r_n drives the search for a distribution estimator or acts as a data-dependent prior within a Bayesian context. In Algorithm 2, r_n acts as a proposal distribution for ABC with a flat prior. The computational differences in the performance of confidence regions in this example suggest that the former application of r_n is preferable to the latter if the summary statistic is not very informative. Interestingly, even though the Bayesian IS-ABC algorithm fails to give us Bayesian posterior distributions, it can still provide us valid frequentist inference.

4.2 Mulit-parameter inference for a Ricker model

A Ricker map is a non-linear dynamical system, often used in Ecology, that describes how a population changes over time. The population, N_t, is noisily observed and is described by the following model,

$$y_t \sim \text{Pois}(\phi N_t),$$
$$N_t = r N_{t-1} e^{-N_{t-1}+e_t}, e_t \sim N(0, \sigma^2),$$

where $t = 1, \ldots, T$ and parameters r, ϕ and σ are positive constants, interpreted as the intrinsic growth rate of the population, a scale parameter, and the environmental noise, respectively. This model is computationally challenging since its likelihood function is intractable for $\sigma > 0$ and is highly irregular in certain regions of the parameter space.

We investigate the performance of confidence bounds for each parameter marginally and two pairs of parameters jointly. We follow the setting and the choice of summary statistics in [21]. The output of both algorithms are post-processed using the regression adjustment.

In the minibatch scheme, for the point estimator we use $E\{\theta \mid S_n(z_i)\}$ estimated by the population Monte Carlo version of IS-ABC. The maximum synthetic likelihood estimator proposed in [21] was also tried, but the estimates obtained this way over-concentrated in a certain area of the parameter space. The corresponding $r_n(\theta)$ did not cover the target mass very well, causing biases in the coverage levels. Instead, $r_n(\theta)$ is used to initialise the population Monte Carlo iterations. Since the sample size is small in this example, overlapping minibatches are chosen with a total number of 40 where each minibatch contains a series of length 10. In this example, the parametric bootstrap method is not computationally feasible because it is computationally expensive to use the simulation-based methods in obtaining the point estimates.

In Table 2, when the acceptance proportion is small, most coverage rates of accept-reject ACDC are close to the nominal level. In contrast, the confidence bounds from IS-ABC display more over-coverage, indicating an even smaller ε is needed to reduce the variance inflation. Furthermore, all ACDC confidence regions are tighter with a size reduction up to 51%. The box plots in Figure 3 show that the CD variances from IS-ABC are inflated substantially by the importance weights, resulting in broader confidence regions as observed in the last column of Table 2.

As in Section 4.1, this numerical example validates the inferential conclusions from both algorithms but here we see accept-reject ACDC consistently producing tighter confidence regions than IS-ABC. The summary statistics in this example were carefully selected to be informative based on domain knowledge. Nevertheless, accept-reject ACDC can still avoid the excessive Monte Carlo variation that impedes IS-ABC.

5. DISCUSSION

In this article, we propose ACDC as a new inference-based approach to likelihood-free methods. It can provide valid frequestist

Table 2. Coverage proportions of marginal confidence intervals (or joint confidence regions) for accept-reject ACDC and IS-ABC applied to Ricker data. Coverage is calculated over 150 *independent runs that produce observations from* $t = 51$ *to* 100 *for data generated by a Ricker model with* $(r, \sigma, \phi) = (e^{3.8}, 0.3, 10)$. *The Monte Carlo sample size for both algorithms is* 50,000 *and the nominal coverage level in every setting is* 95%. *The last column displays the median ratio of the sizes of confidence sets from accept-reject ACDC divided by those from IS-ABC.*

Acceptance proportion	ACDC Coverage	IS-ABC Coverage	Ratio of Widths/ Volumes
Setting 1: $log(r)$ **unknown**			
0.005	0.953	0.980	0.793
0.05	0.960	0.987	0.734
0.10	0.967	0.987	0.707
Setting 2: $log(\sigma)$ **unknown**			
0.005	0.967	0.987	0.782
0.05	0.987	0.993	0.732
0.10	0.987	0.993	0.717
Setting 3: $log(\phi)$ **unknown**			
0.005	0.953	0.967	0.828
0.05	0.947	0.993	0.762
0.10	0.960	0.987	0.734
Setting 4: $(log(r), log(\sigma))'$ unknown			
0.005	0.960	0.947	0.611
0.05	0.960	0.973	0.519
0.10	0.960	0.947	0.484
Setting 5: $(log(r), log(\phi))'$ **unknown**			
0.005	0.973	0.987	0.749
0.05	1.0	1.0	0.619
0.10	1.0	1.0	0.557

Figure 3: *These are densities of point estimators from accept-reject ACDC (red) and IS-ABC (black) for the* 150 *independent data sets produced by the Ricker model. Additionally, this figure shows a box plot of the ratio of the sizes of the* 150 *confidence sets, that is, the length (or volume) of regions produced by accept-reject ACDC divided by those of IS-ABC.*
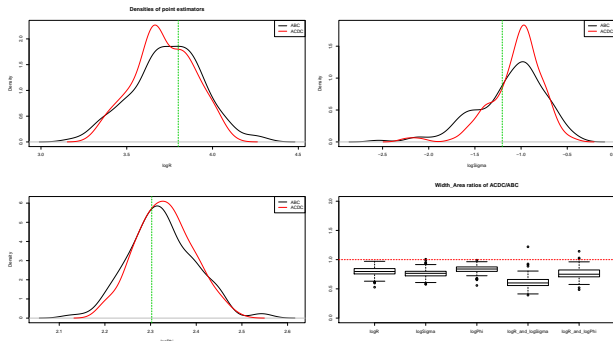
inference for target parameters from data without a tractable likelihood. ACDC can be viewed as an extension of ABC but, crucially, ACDC does not require any Bayesian assumptions nor does the validity of inferential conclusions depend upon the near-sufficiency of the summary statistic. Computationally, an ACDC approach is preferable when compared to the corresponding IS-ABC method which suffers from skewed importance weights.

The main theoretical contribution of this work is the identification of a matching condition (Condition 1) necessary for valid frequentist inference from ACDC methods. This condition is similar to the theoretical support for Bootstrap estimation and is met in cases that rely on typical asymptotic arguments (e.g. reference citations in Section 3) but also applies to certain small-sample cases. Additionally, a key practical contribution of this work is the general minibatch method for initializing ACDC estimators. This approach guides the search for a well-behaved distribution estimator using a data-dependent distribution $r_n(\theta)$. This can result in improved computational performance even compared to an IS-ABC method that is similarly data-driven. In cases where $r_n(\theta)$ does not yield reasonable acceptance probabilities we expect that many of the established techniques used in ABC can be readily adapted to ACDC to further improve its computational performance without sacrificing the frequentist inferential guarantees.

An ACDC approach quantifies the uncertainty in estimation by drawing upon a direct connection to confidence distribution estimators. Different choices of summary statistic yield different approximate CDs, some producing tighter confidence sets than others. However, inference from ACDC is validated, regardless of the sufficiency of S_n, provided Condition 1 can be established. Within a Bayesian framework, there is no clear way to choose among different posterior approximations associated with different summary statistics. By pivoting to a frequentist perspective, different summary statistics produce different (CD) estimators but all of these estimators are well-behaved in the long run, yielding valid inferential statements about θ. Supported by the theoretical developments and examples in this paper, it appears as though ACDC provides a more parsimonious solution to validating likelihood-free inference than attempts to reconcile differences among posteriors and their various approximations.

ACKNOWLEDGMENT

APPENDIX A. CLAIM IN SECTION 2

Proof. First note that $H_n(t) = 1 - Q_\varepsilon(2\hat{\theta} - t \mid S_n = s_{obs})$ is a sample-dependent cumulative distribution function on the parameter space. By equation (2.1), we denote both sides as $G(t)$, i.e. $G(t) = \text{pr}^*\{\theta - \hat{\theta}_S \leq t \mid S_n = s_{obs}\} = \text{pr}\{\hat{\theta}_S - \theta \leq t \mid \theta = \theta_0\}$. Now we can write $H_n(\theta_0) = \text{pr}^*(2\hat{\theta}_S - \theta \leq \theta_0 \mid S_n = s_{\text{obs}}) =$

$\mathrm{pr}^*(\theta - \hat{\theta}_S \geq \hat{\theta}_S - \theta_0 \mid S_n = s_{\mathrm{obs}}) = 1 - G(\hat{\theta}_S - \theta_0)$, for $G(t) = \mathrm{pr}(\hat{\theta}_S - \theta \leq t \mid \theta = \theta_0)$. The last equality holds by equation (2.1). Since $G(\hat{\theta}_S - \theta_0) \mid \theta_0 \sim Unif(0,1)$ with respect to the sampling variability of $\hat{\theta}_S$, $H_n(\theta_0) = H_n(\theta_0, s_{\mathrm{obs}}) \sim Unif(0,1)$. By definition, $H_n(\cdot)$ is a confidence distribution for θ. $\square$

APPENDIX B. LEMMA 1

Proof. Following the notation established in the claim of Section 2, first note that

$$\left| \mathrm{pr}\{\theta \in \Gamma_{1-\alpha}(S_n) | \theta = \theta_0\} - (1-\alpha) \right|$$
$$= \left| \mathrm{pr}\{W(\theta, S_n) \in A_{1-\alpha} | \theta = \theta_0\} - (1-\alpha) \right|$$
$$\leq \left| \mathrm{pr}^*\{V(\theta, S_n) \in A_{1-\alpha} | S_n = s_{\mathrm{obs}}\} - (1-\alpha) \right|$$
$$+ \left| \mathrm{pr}\{W(\theta, S_n) \in A_{1-\alpha} | \theta = \theta_0\} \right.$$
$$\left. - \mathrm{pr}^*\{V(\theta, S_n) \in A_{1-\alpha} | S_n = s_{\mathrm{obs}}\} \right|$$

and by the definition of $A_{1-\alpha}$ in (4), $| \mathrm{pr}^*\{V(\theta, S_n) \in A_{1-\alpha} \mid S_n = s_{\mathrm{obs}}\} - (1-\alpha) | = o(\delta')$, almost surely for a pre-selected precision number, $\delta' > 0$. Therefore, by Condition 1, we have $| \mathrm{pr}\{\theta \in \Gamma_{1-\alpha}(S_n) \mid \theta = \theta_0\} - (1-\alpha) | = \delta$ where $\delta = \max\{\delta_\varepsilon, \delta'\}$. Furthermore, if Condition 1 holds almost surely, then $| \mathrm{pr}\{\theta \in \Gamma_{1-\alpha}(S_n) \mid \theta = \theta_0\} - (1-\alpha) | = o(\delta)$, almost surely. $\square$

APPENDIX C. THEOREM 1

Proof. If $T = T(\theta, S_n)$ is an approximate pivot for S_n then

$$\mathrm{pr}\{T(\theta, S_n) \in A \mid \theta = \theta_0\} = \int_{t \in A} g(t)dt \{1 + o(\delta'')\}, \quad \text{(C.1)}$$

for any Borel set $A \subset \mathcal{S}$. Given θ and t, denote the solution of $t = T(\theta, s)$ by $s_{t,\theta}$. The density functions $g(t)$ and $f_n(s_{t,\theta}|\theta)$ are connected by a Jacobian matrix:

$$f_n(s_{t,\theta}|\theta)|T^{(1)}(\theta, s_{t,\theta})|^{-1} = g(t)\{1 + o(\delta'')\} \quad \text{(C.2)}$$

where $T^{(1)}(\theta, S_n) = (\partial/\partial S_n)T(\theta, S_n)$.

For $\theta' \sim Q_\varepsilon(\cdot \mid s_{obs})$ and corresponding summary S_n', the joint density of (θ', S_n') conditional on the observed data, is

$$(\theta', S_n')|S_n = s_{\mathrm{obs}} \propto r_n(\theta)f_n(S_n \mid \theta)K_\varepsilon(S_n - s_{\mathrm{obs}}).$$

Let $T' = T(\theta', S_n')$. With a variable transformation from (θ', S_n') to (θ', T'), the joint density of (θ', T'), conditional on the observed data, is

$$(\theta', T')|s_{\mathrm{obs}} \propto r_n(\theta) \left[f_n(s_{t,\theta} \mid \theta)|T^{(1)}(\theta, s_{t,\theta})|^{-1} \right] \times$$
$$K_\varepsilon(s_{t,\theta} - s_{\mathrm{obs}})$$
$$= r_n(\theta) \left[g(t)\{1 + o(\delta'')\} \right] K_\varepsilon(s_{t,\theta} - s_{\mathrm{obs}}),$$

where $s_{t,\theta}$ is the solution of $t = T(\theta, S_n)$ and the equivalence holds by (C.2). Integrating over the parameter space yields

$$T'|S_n = s_{\mathrm{obs}} \propto \left[g(t)\{1 + o(\delta'')\} \right] \int_{\mathcal{P}} r_n(\theta)K_\varepsilon(s_{t,\theta} - s_{\mathrm{obs}})d\theta$$

$$\propto g(t)\{1 + o(\delta'')\},$$

provided (2.5) holds.

Now, consider $W(\theta, S_n) = T(\theta, S_n)$ as a function of the random sample given some fixed, unknown value of θ, by (C.1)

$$\mathrm{pr}\{W(\theta, S_n) \in A \mid \theta = \theta_0\} = \int_{t \in A} g(t)dt\{1 + o(\delta)\}.$$

If we consider $V(\theta, S_n) = T(\theta, S_n)$ and the joint density of (θ', S_n') pairs then

$$\mathrm{pr}^*\{V(\theta, S_n) \in A \mid S_n = s_{\mathrm{obs}}\} = \int_{t \in A} g(t)dt\{1 + o(\delta'')\}$$

thus satisfying Condition 1. Furthermore, by Lemma 1, $\Gamma_{1-\alpha}(s_{\mathrm{obs}})$ in equation (2.3) is a $(1 - \alpha)100\%$ confidence region for θ. $\square$

APPENDIX D. COROLLARY 1

Proof. By Theorem 1, it suffices to show that equation 2.4 is free of t in each case.

(a) Suppose $S_n \sim g_1(S_n - \mu)$. Then $T_1 = T_1(\mu, S_n) = S_n - \mu \sim g_1(t)$ is a pivot for S_n. For any (t, μ) pair $s_{t,\mu} = t + \mu$. With a change of variables $u = t + \mu - s_{obs}$ and with $r_n(\mu) \propto 1$ we have

$$\int_{\mathcal{P}} r_n(\mu)K_\varepsilon(s_{t,\mu} - s_{\mathrm{obs}})d\mu = \int_{-\infty}^{\infty} K_\varepsilon(u)du,$$

which is free of t.

(b) Suppose $S_n \sim (1/\sigma)g_2(S_n/\sigma)$. Then $T = T(\sigma, S_n) = S_n/\sigma \sim g_2(t)$ is a pivot. For any (t, σ) pair $s_{t,\sigma} = t\sigma$. With $r_n(\sigma) \propto 1/\sigma$ and with a change of variables $u = t\sigma - s_{obs}$, we have

$$\int_{\mathcal{P}} r_n(\sigma)K_\varepsilon(s_{t,\sigma} - s_{\mathrm{obs}})d\sigma = \int_0^{\infty} \frac{1}{\sigma}K_\varepsilon(t\sigma - s_{\mathrm{obs}})d\sigma$$
$$= \int_0^{\infty} \frac{1}{(u + s_{obs})/t}K_\varepsilon(u)\frac{1}{t}du$$

which is free of t.

(c) Since we have already proven parts (a) and (b), part (c) follows provided we select $r_n(\theta) \propto 1/\sigma$ for $\theta = (\mu, \sigma)$.

Finally, to prove the last statement of the corollary first note that the function $H_1(S_n, x) = \int_{-\infty}^x g_1(S_n - u)du$ is a CD for μ when $S_n \sim g_1(S_n - \mu)$ because, for a given S, $H_1(S, x)$ is a distribution function on the parameter space $(-\infty, \infty)$ and given $x = \mu_0$, $H_1(S, x) \sim U(0,1)$. Similarly, the function $H_2(S^2, x) = 1 - \int_0^x g_2(S/u)du$ is a CD for σ^2 when $S_n \sim (1/\sigma)g_2(S_n/\sigma)$. $\square$

APPENDIX E. REMARK ON DEGENERACY OF ACCEPTANCE RATE

A natural question is whether Theorem 2 holds for a larger ε_n. We claim that the answer is negative, using the following basic normal mean model as a counterexample.

Consider a univariate Gaussian model with mean θ and unit variance, and observations that are IID from the model with $\theta = \theta_0$. Let $r_n(\theta)$ be a normal density with mean μ_n and variance b_n^{-2}, where μ_n and b_n are constant sequences satisfying $b_n(\mu_n - \theta_0) = O(1)$ and $b_n = o(\sqrt{n})$ as $n \to \infty$, and let S_n be the sample mean. One can verify that r_n and S_n satisfy the conditions of Theorem 2. The Gaussian kernel with variance ε_n^2 is used for the acceptance/rejection. Then the density of a linear transformation of $\theta \sim Q_\varepsilon(\theta \mid s_{\text{obs}})$ is Gaussian with a closed form

$$\sqrt{n}(\theta - \hat{\theta}_S) \mid s_{obs} \sim N(0, n\sigma_\varepsilon^2)$$

where $\sigma_\varepsilon^2 = \frac{b_n^{-2}\Delta_n}{1+\Delta_n}$ and $\Delta_n = b_n^2(n^{-1} + \varepsilon^2)$. Also,

$$\sqrt{n}(\hat{\theta}_S - \theta) \mid \theta_0 = \frac{1}{1+\Delta_n}\sqrt{n}(s_{obs} - \theta) + \frac{\sqrt{n}b_n^{-1}\Delta_n}{1+\Delta_n}b_n(\mu_n - \theta).$$

By algebra, the expectation of $\sqrt{n}(\hat{\theta}_S - \theta) \mid \theta_0$ is $o(1)$ only when $\varepsilon_n = o(b_n^{-1/2}n^{-1/4})$, and the variance is $n\sigma_\varepsilon^2 + o(1)$ only when $\varepsilon_n = o(n^{-1/2})$ or $\varepsilon_n^{-1} = o(b_n^2 n^{-1/2})$. Since $b_n = o(\sqrt{n})$, both $\varepsilon_n = o(b_n^{-1/2}n^{-1/4})$ and $\varepsilon_n^{-1} = o(b_n^2 n^{-1/2})$ can not hold simultaneously. Therefore Condition 1 is satisfied only if $\varepsilon_n = o(n^{-1/2})$.

APPENDIX F. THEOREMS 2 AND 3

The proof for Theorem 2 requires establishing Lemmas 1–5 which are given in the Supplementary Material. Theorem 3 is proved after establishing Lemmas 6–7 given in the Supplementary Material. Below, we present the additional conditions from [9], necessary for these lemmas. The proofs of these technical lemmas are contained in the Supplementary Material. It is helpful to have a copy of both [10] and [9] (and their supplementary material) on hand as these proofs are rely on results from these two publications.

F.1 Notation

For the sequence a_n in Condition 2, let $a_{n,\varepsilon} = a_n$ if $\lim_{n\to\infty} a_n\varepsilon_n < \infty$ and $a_{n,\varepsilon} = \varepsilon_n^{-1}$ otherwise. Additionally, let $c_\varepsilon = \lim_{n\to\infty} a_n\varepsilon_n$. Both $\{a_n\}$ and c_ε characterize how ε_n decreases relative to the convergence rate, a_n, of S_n.

Let $\widetilde{f}_n(s \mid \theta) = N\{s; s(\theta), A(\theta)/a_n^2\}$ be the asymptotic distribution of the summary statistic from Condition 2. Define the standardized random variables $W_n(S_n) = a_n A(\theta)^{-1/2}\{S_n - s(\theta)\}$ and $W_{\text{obs}} = a_n A(\theta)^{-1/2}\{s_{\text{obs}} - s(\theta)\}$. Finally, let $f_{W_n}(w \mid \theta)$ and $\widetilde{f}_{W_n}(w \mid \theta)$ be the density for $W_n(S_n)$ when $S_n \sim f_n(\cdot \mid \theta)$ and $\widetilde{f}_n(\cdot \mid \theta)$, respectively.

F.2 Conditions

Condition 7. The kernel in Algorithm 1 satisfies
(i) $\int v K_\varepsilon(v)dv = 0$;
(ii) $\prod_{k=1}^l v_{i_k} K_\varepsilon(v)dv < \infty$ for any coordinates $(v_{i_1}, \ldots, v_{i_l})$ of v and $l \le p + 6$;

(iii) $K_\varepsilon(v) \propto K_\varepsilon(\|v\|_\Lambda^2)$ 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) $K_\varepsilon(v) = O(\exp\{-c_1\|v\|^{\alpha_1}\})$ for some $\alpha_1 > 0$ and $c_1 > 0$ as $\|v\| \to \infty$.

Condition 8. There exists α_n satisfying $\alpha_n/a_n^{2/5} \to \infty$ and a density $r_{max}(w)$ satisfying Condition 7, where $K_\varepsilon(v)$ is replaced with $r_{max}(w)$, such that

$$\sup_{\theta \in \mathcal{P}_\delta} \alpha_n \mid f_{W_n}(w \mid \theta) - \widetilde{f}_{W_n}(w \mid \theta) \mid \le c_3 r_{max}(w)$$

for some positive constant c_3.

Condition 9. For some positive constants c_2 and α_2,
$\sup_{\theta \in \mathcal{P}_\delta^C} \widetilde{f}_{W_n}(w \mid \theta) = O(e^{-c_2\|w\|^{\alpha_2}})$ as $\|w\| \to \infty$.

Condition 10. The first two moments, $\int_{\mathcal{S}} s\widetilde{f}_n(s \mid \theta)ds$ and $\int_{\mathbb{R}^d} s^T s\widetilde{f}_n(s \mid \theta)ds$, exist.

REFERENCES

[1] BEAUMONT, M. A., ZHANG, W. and BALDING, D. J. (2002). Approximate Bayesian Computation in Population Genetics. *Genetics* **162**(4) 2025–2035.
[2] BEAUMONT, M. A., CORNUET, J. Q. M., MARIN, J. Q. M. and ROBERT, C. P. (2009). Adaptive approximate Bayesian computation. *Biometrika* **20**(10) 1–9.
[3] BLUM, M. G. (2010). Approximate Bayesian computation: A nonparametric perspective. *Journal of the American Statistical Association* **105**(491) 1178–1187.
[4] CAMERON, E. and 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**(1) 44–65.
[5] CHENG, T. T. (1949). The normal approximation to the Poisson distribution and a proof of a conjecture of Ramanujan. *Bulletin of the American Mathematical Society* **55**(4) 396–401.
[6] CSILLÉRY, K., BLUM, M. G. B., GAGGIOTTI, O. E. and FRANÇOIS, O. (2010). Approximate Bayesian Computation (ABC) in practice. *Trends in Ecology and Evolution* **25**(7) 410–418.
[7] FASIOLO, M., WOOD, S. N., HARTIG, F. and BRAVINGTON, M. V. (2018). An extended empirical saddlepoint approximation for intractable likelihoods. *Electronic Journal of Statistics* **12**(1) 1544–1578.
[8] GOURIEROUX, C., MONFORT, A. and RENAULT, E. (1993). Indirect Inference. *Journal of Applied Econometrics* **8**(S1) 85–118.
[9] LI, W. and FEARNHEAD, P. (2018). Convergence of regression-adjusted approximate Bayesian computation. *Biometrika* **105**(2) 301–318.
[10] LI, W. and FEARNHEAD, P. (2018). On the Asymptotic Efficiency of approximate Bayesian computation estimators. *Biometrika* **105**(2) 286–299.
[11] LIU, R. Y., PARELIUS, J. M. and SINGH, K. (1999). Multivariate analysis by data depth: descriptive statistics, graphics and inference (with discussion). *Annals of Statistics* **27** 783–858.
[12] MARIN, J. Q. M., PUDLO, P., ROBERT, C. P. and RYDER, R. J. (2011). Approximate Bayesian Computational methods. *Statistics and Computing* **22**(6) 1167–1180.
[13] MARJORAM, P., MOLITOR, J., PLAGNOL, V. and TAVARÉ, S. (2003). Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences* **100**(26) 15324–15328.
[14] MARQUETTE, J. Q. B. (2018) *Statistics for Astrophysics: Bayesian Methodology*.
[15] MCFADDEN, D. (1989). A method of simulated moments for estimation of discrete response models without numerical integration. *Econometrica: Journal of the Econometric Society* 995–1026.

[16] PETERS, G. W., FAN, Y. and SISSON, S. A. (2012). On sequential Monte Carlo partial rejection control approximate Bayesian computation. *Statistical Computing* **22** 1209–1222.

[17] SCHWEDER, T. and HJORT, N. L. (2016) *Confidence, Likelihood, Prabability*. Cambridge University Press.

[18] SERFLING, R. (2002). Quantile functions for multivariate analysis: approaches and applications. *Statistica Neerlandica* **56**(2) 214–232.

[19] SISSON, S., FAN, Y. and TANAKA, M. (2007). Sequential Monte Carlo without likelihoods. *Proceedings of the National Academie of Science USA* **104** 1760–1765.

[20] THORNTON, S. and XIE, M. (2022). Bridging Bayesian, frequentist and fiducial (BFF) inferences using confidence distribution. In *Handbook on Bayesian, Fiducial and Frequentist (BFF) Inferences* (J. O. Berger, X. q. L. Meng, N. Reid and M. Xie, eds.) Chapman & Hall. To appear.

[21] WOOD, S. N. (2010). Statistical inference for noisy nonlinear ecological dynamic systems. *Nature* **466**(7310) 1102.

[22] XIE, M. and SINGH, K. (2013). Confidence distribution, the frequentist distribution estimator of a parameter: A review. *International Statistical Review* **81**(1) 3–39.

Suzanne Thornton. Address of the First author, Swarthmore College, U.S.A.
E-mail address: sthornt1@swarthmore.edu

Wentao Li. Address of the Second author, The University of Manchester, U.K..
E-mail address: wentao.li@manchester.ac.uk

Minge Xie. Address of the Third author, Rutgers, The State University of New Jersey, U.S.A.
E-mail address: mxie@stat.rutgers.edu