ADVANCED COMPUTING METHODS FOR STATISTICAL INFERENCE

by

SUZANNE THORNTON

A dissertation submitted to the

School of Graduate Studies

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Statistics

Written under the direction of

Min-ge Xie

and approved by

_____

_____

_____

_____

New Brunswick, New Jersey

October, 2019

ABSTRACT OF THE DISSERTATION

Advanced computing methods for statistical inference

by Suzanne Thornton

Dissertation Director: Min-ge Xie

In this thesis, we provide some new and interesting solutions to problems of computational inference. In particular, the two problems we address are (1.) How to obtain valid confidence sets for parameters from models with no tractable likelihood function and (2.) How to obtain valid exact confidence sets for the odds ratio when the signal is very difficult to detect. Our approach to solving these problems is to develop algorithmic procedures that result in confidence distributions for the parameters of interest. A confidence distribution can be thought of as a frequentist analog to a Bayesian posterior. It is a distribution estimate for a parameter of interest that provides inferential results with respect to the *Repeated Sampling Principle*.

1. Most likelihood-free computational methods for statistical inference are performed under a Bayesian paradigm, even though they are driven by the need for inferential results in instances where the likelihood principle may fail. We develop a frequentist computational method to apply in situations where one has an intractable likelihood and instead rely on the Repeated Sampling Principle to justify our inferential results. Our method expands the applications of approximate Bayesian computing methods from and permits faster computational speed by eliminating the need for any prior information. Rather than attempting to work within a Bayesian

framework without a tractable likelihood function, our method creates a special type of estimate, a confidence distribution, for the parameter of interest.

2. Establishing drug safety entails detecting relationships between treatments and rare, but adverse, events. For a 2×2 contingency tables of drug treatment and adverse events, this means that we are interested in inference for an odds ratio with a weak signal. In these situations, we will encounter very few adverse events, even if the number of patients under study is large. We develop a frequentist computational method for inference on sparse contingency tables that does not rely on large sample assumptions. Our method works under the assumption that one margin is fixed, enabling us to compare the observed data to simulated data through a data generating equation and a modified statistic. We make use of a stabilization parameter which allows us to consider smaller potential parameter values even if we have a zero observation in the data. This stabilization parameter makes our method distinct from the standard tail method approach. We show that our method can out-perform the overly-conservative existing exact methods and a Bayesian method.

In both of these problems, the algorithmic approaches we propose attempt to capture the sample variability using a known random variable connected to the data through a data-generating equation. In order to validate the inferential results within a frequentist framework, the algorithmic approaches to both of the above problems work by producing a specific type of estimator, a confidence distribution, for the unknown parameter. We think these two problems illustrate the rich possibilities for incorporating confidence distribution theory into the world of statistical computing.

Acknowledgements

There are many people to thank for the achievement of this dissertation. First of all, I would like to to thank my dissertation advisor, Min-ge Xie, for his patience and wisdom as he advised me throughout my graduate career. Whenever I felt doubtful or insecure about my work, knowing that I could trust his judgment and expertise brought me peace of mind and kept me motivated. I would also like to thank my research collaborators Wentao Li and Zeshi Zheng for their thoughtful input and valuable contributions to this work. I would also like to extend my thanks to John Kolassa for being such an excellent graduate advisor and Bill Strawderman and Steve Buyske for helping me realize my interest in teaching as a strength and as a step towards a future successful career.

Thanks is also due to Rutgers University for providing me with generous graduate support in the form of the Presidential Fellowship and due to the entire Statistics department faculty, staff, students, and especially the chair, Regina Liu for their support and encouragement.

From the University of Florida, I thank Jim Hobert, my undergraduate statistics advisor, for my first experience in statistical research. He enthusiastically entertained my questions, piquing my interested in this field and boosted my confidence by encouraging me to ask more questions. From the Icahn School of Medicine at Mount Sinai, I thank Emma Benn for mentoring me after meeting at WSDS 2016 and encouraging me to pursue my passion of advocating for minorities in STEM. She has helped me realize the kind of statistician I want to be.

Finally, thank you to Abigail Militano, to whom this work is dedicated. I hardly believe I finished this without you, but I know you would be proud. You are greatly missed.

Dedication

To Abigail

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

The field of Statistics has most assuredly benefited from the advancement of computational computing. The goal of the work presented in this thesis is to quantify uncertainty about model parameters but the methods developed herein are meant to be congruent with the rapid intertwining of the computational and statistical sciences. Broadly speaking, the problems explored in this thesis are those of computational inference, by which we mean,

> "any statistical methods [that] involve direct simulation of the hypothesized data-generating process rather than formal computations of probabilities that would result under a given model of the data-generating process." [Gentle (2009)]

In the methods of computational inference proposed here, the probabilities associated with confidence intervals are estimated by the simulation of a hypothesized data-generating process rather than by resampling an observed sample. A guiding paradigm we reference in this work is the heuristic that one should reject models under which the probability of generated data matching the observed sample is small. ([Gentle (2009)]) This work evaluates the performance of statistical measures of uncertainty with respect to the Repeated Sampling Principle. As such, all novel methods proposed in this work are frequentist and assume that the model parameters of interest are fixed unknowns. The problems addressed in this dissertation are ones for which computational inference either replaces or supplements asymptotic inference; we will refer to instances of the former case as "exact" computational inferential procedures.

A key concept underlying the computational methods developed in this thesis is the frequentist notion of a confidence distribution. When estimating an unknown parameter within the frequentist paradigm, we often desire that our estimators, whether point

estimators or interval estimators, have certain properties such as unbiasedness or a certain 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 desirable properties. We define a confidence distribution as follows.

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 sets of all levels for θ.* [Xie & Singh(2013), Schweder & Hjort(2016)]

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, unknown quantity. It is not a distribution of the parameter; rather, it is a sample-dependent function used to estimate the parameter of interest, including to quantify the uncertainty of the estimation. In fact, one of the appeals of drawing inference from a confidence distribution is the similarity of these estimators to Bayesian posterior distributions and the flexibility they provide for inference because of this. For a comprehensive review on confidence distributions, see [Xie & Singh(2013)] and references therein.

In this work, we propose frequentist solutions to two different computation inference problems and in so doing, we are able to offer empirical procedures that do not require prior information. The two problems we address are (1.) How to obtain valid confidence sets for parameters from models with no tractable likelihood function and (2.) How to obtain valid exact confidence sets for the odds ratio when the signal is very difficult to detect.

1. 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 work,

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*, yielding 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 frequentist inferential guarantees. We supplement the theory with simulation studies illustrating the benefits of the proposed method, namely the potential for broader applications and the increased computing speed compared to approximate Bayesian computing methods.

2. Inference for sparse contingency tables with binomial sampling is a challenging but important area of statistical research still subject to much debate. Most commonly used methods rely on large sample, asymptotic results, but finite sample inference, especially in the case where we observe zero or a small number of positive outcomes, remains a problematic question in many applications. In the work presented here, we provide a computational inferential method for the log odds ratio of a 2×2 contingency table which we call a *repro sampling method*. Our method does not rely on large sample size assumptions but instead attempts to "reproduce" the sample variability though simulations of known random variables and a data-generating equation. Our method differs from the standard tail method in that we utilize a modified statistic that depends on a positive stabilization parameter. We demonstrate that our methods can produce better confidence intervals than the exact tail method and compare our method to several other common exact

approaches. We also examine the results of applying the repro sampling method to many real clinical trials and compare our confidence intervals to the standard approach.

The remainder of this dissertation is organized as follows. In Chapter 2 we propose a new likelihood-free computational method for inference called *approximate confidence distribution computing*. We show that this method can greatly improve upon the computational cost of existing likelihood-free methods without damaging the inferential integrity of the results, within an entirely frequentist framework. Though the main results of this section are developed around a Bernstein von Mises-type of large sample theorem, we show that there are possible extensions of this method that are independent of sample size. An appendix is included at the end of this chapter for additional proofs. In Chapter 3 we propose an exact computational inferential method for inference on the odds ratio of sparse 2×2 contingency tables, called *repro sampling*. We compare the algorithmic procedure developed here to other existing methods exact methods of interest and demonstrate the potential improvement afforded by incorporating a positive stabilization parameter. Chapter 4 concludes this dissertation with some additional remarks on the role of confidence distributions in the development of statistical computing methods and with some possible directions for future research.

Chapter 2

Approximate confidence distribution computing (ACC)

2.1 Introduction

2.1.1 Background to approximate Bayesian computing

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), Csilléry et al.(2010), Peters et al.(2012)]. Despite its practical popularity in providing a Bayesian solution for complex data problems in which there is no tractable likelihood function, the theoretical justification for inference from this method is under-developed and has only recently been explored in the statistical literature; cf., e.g., [Robinson et al.(2014), Barber et al.(2015), Frazier et al.(2018), Li & Fearnhead(2018b)]. Here, 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 computing.

Let $x_{\text{obs}} = \{x_1, \ldots, x_n\}$ be an observed sample from some intractable distribution. Assume however, that there exists some data generating model, M_θ, depending on the parameter of interest, $\theta \in \mathcal{P} \subset \mathbb{R}^p$. That is, given any θ, we can simulate artificial data from M_θ, even though we cannot work with the likelihood directly. 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_i}*;*
3. *For each* $i = 1, \ldots, N$*, accept* θ_i *with probability* $K_\varepsilon(s^{(i)} - s_{\text{obs}})$*, where* $s_{\text{obs}} = S_n(x_{\text{obs}})$ *and* $s^{(i)} = S_n(x^{(i)})$*.*

In the above algorithm, $\pi(\cdot)$ is a prior distribution function and 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 $\mathcal{S} \subset \mathbb{R}^d$ with $d \leq n$). The kernel probability $K_\varepsilon(\cdot)$ follows the notation $K_\varepsilon(u) = \varepsilon^{-1}K(u/\varepsilon)$, where $K(\cdot)$ is a kernel function which, without loss of generality, we assume satisfies $\max_x K(x) = 1$. 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 under development in which ε is independent of sample size n, see e.g. [Barber et al.(2015)].

The underlying distribution from which the accepted copies or draws of θ are generated in Algorithm 1 is called the *approximate Bayesian computing posterior* and has the density

$$\pi_\varepsilon(\theta \mid s_{\text{obs}}) = \frac{\int_{\mathcal{S}} \pi(\theta) f_n(s \mid \theta) K_\varepsilon(s - s_{\text{obs}})\, ds}{\int_{\mathcal{P}\times\mathcal{S}} \pi(\theta) f_n(s \mid \theta) K_\varepsilon(s - s_{\text{obs}})\, ds d\theta}, \tag{2.1}$$

and corresponding cumulative distribution function, $\Pi_\varepsilon(\theta \mid s_{\text{obs}})$. Here $f_n(s \mid \theta)$ denotes the probability density of the summary statistic, implied by the intractable likelihood and, as such, is typically unknown. We will refer to $f_n(s \mid \theta)$ as an *s-likelihood*. Since this is a Bayesian procedure, Algorithm 1 assumes a prior distribution, $\pi(\cdot)$, on θ. In the absence of prior information, the user may select a noninformative prior.

A common assertion is that $\pi_\varepsilon(\theta \mid s_{\text{obs}})$ is close enough to the target posterior distribution, e.g. [Marin et al.(2012)]; however, the quality of this approximation depends on the closeness of the tolerance level to zero and, more crucially for our purposes, on the choice of summary statistic $S_n(\cdot)$. Provided the prior is proper, we have the

following lemma:

Lemma 1 *Let $K(\cdot)$ be a kernel density function symmetric about zero with $\int \|u\|^2 K(u)du < \infty$ where $\|\cdot\|$ is the Euclidean norm. Suppose the matrix of second derivatives of $f_n(s \mid \theta)$ is bounded with respect to s. Then*

$$\pi_\varepsilon(\theta \mid s_{\text{obs}}) \propto \pi(\theta) f_n(s_{\text{obs}} \mid \theta) + O(\varepsilon^2). \tag{2.2}$$

Various versions of this result are known (cf., e.g., [Barber et al.(2015), Li & Fearnhead(2018a)]); for completeness, we provide a brief proof of Lemma 1 in the appendix. Note that, if the summary statistic $S_n(\cdot)$ is not sufficient, $f_n(s_{\text{obs}} \mid \theta)$ can be very different from the actual likelihood function and, in this way, $\pi_\varepsilon(\theta \mid s_{\text{obs}})$ can be a very poor approximation to the target posterior, even as $\varepsilon \to 0$.

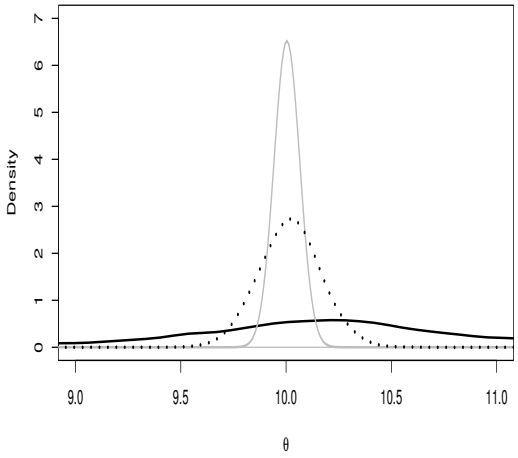
Figure 2.1 provides such an example where we consider random data from a Cauchy distribution with a known scale parameter. Only the data itself is sufficient for the location parameter, θ; therefore, any summary statistic, including the commonly used sample mean and median, will not be sufficient. Figure 2.1 illustrates that, without sufficiency, the posterior approximation resulting from Algorithm 1, using either sample mean and sample median as $S_n(\cdot)$, will never converge to the targeted posterior distribution, thus indicating that the approximations to the target posterior can be quite poor. What's more, the two different summary statistics lead to quite different approximate Bayesian computing posteriors $\pi_\varepsilon(\theta \mid s_{\text{obs}})$. The approximate Bayesian computing posteriors obtained using the sample mean are much flatter than those obtained using the sample median. Further details about Figure 2.1 can be found in Sections 2.4.1 and 2.4.2.

For this reason, inference from $\Pi_\varepsilon(\cdot \mid s_{\text{obs}})$ can produce misleading results within a Bayesian context when the summary statistic used is not sufficient. Questions arise such as, if $\Pi_\varepsilon(\cdot \mid s_{obs})$ is different from the target posterior distribution, can it still be used in Bayesian inference? Or, since different summary statistics can produce different approximate posterior distributions, can one or more of these distributions be used to make statistical inferences?

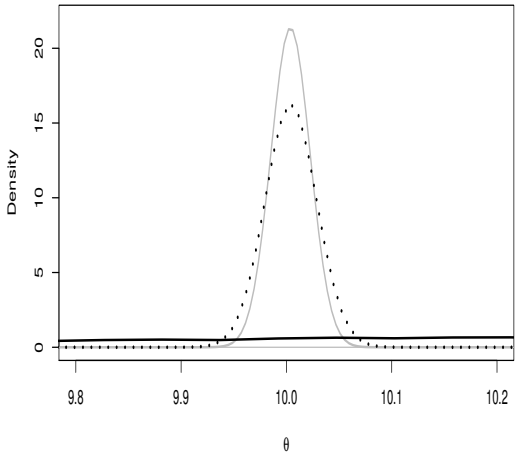
We attempt to address these questions by instead re-framing Algorithm 1 within

Figure 2.1: *The three curves in each of the two plots are the target posterior (gray) and approximate Bayesian computing posteriors for data from a Cauchy distribution with known scale parameter for summary statistic $S_n = \bar{x}$ (solid black) and $S_n = Median(x)$ (dashed black). The prior density is a constant in $\mathbb{R}$.*

(a) Cauchy data of sample size $n = 50$.



(b) Cauchy data of sample size $n = 5000$.



a frequentist context, thus creating 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*.

2.1.2 Approximate confidence distribution computing

The theoretical foundation for approximate confidence distribution computing relies upon the Repeated Sampling Principle. Confidence distributions are, by definition, estimators that follow the *frequentist coverage property* since they are able to produce confidence sets for θ that contain this true parameter value, θ_0, at any specified frequency with repeated experimental runs.

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. Furthermore, approximate confidence distribution computing provides a computational method with potential

applications extending beyond the scope of Algorithm 1 and, as will be discussed later, it 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 function, $r_n(\theta)$, from which potential parameter values will be generated. Specifically, the new algorithm proceeds as follows:

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 the accepted draws of θ are simulated is denoted by $Q_\varepsilon(\theta \mid s_{\text{obs}})$. We refer to $Q_\varepsilon(\theta \mid s_{\text{obs}})$ as an *approximate confidence distribution* and denote the corresponding density by $q_\varepsilon(\theta \mid s_{\text{obs}})$ as defined by replacing $\pi(\theta)$ in (2.1) with $r_n(\theta)$:

$$q_\varepsilon(\theta \mid s_{\text{obs}}) = \frac{\int_{\mathcal{S}} r_n(\theta) f_n(s \mid \theta) K_\varepsilon(s - s_{\text{obs}})\, ds}{\int_{\mathcal{P}\times\mathcal{S}} r_n(\theta) f_n(s \mid \theta) K_\varepsilon(s - s_{\text{obs}})\, ds d\theta}, \tag{2.3}$$

In this way, approximate Bayesian computing can be viewed as a special case of approximate confidence distribution computing with $r_n(\theta) = \pi(\theta)$.

From a Bayesian perspective, one may view Algorithm 2 as as an extension permitting the use of Algorithm 1 in the presence of a data-dependent prior. However, there is another natural, frequentist interpretation that views the function $r_n(\theta)$ as an initial distribution estimate for θ and views Algorithm 2 as a method to update this estimate 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 in search for a better-performing estimate. One may ask if the data are, thus being 'doubly used'. The answer depends on how the initial distribution estimate is chosen. Under some constraints on $r_n(\theta)$, Algorithm 2 can guarantee a distribution estimator for θ that satisfies the frequentist coverage property

thus $q_\varepsilon(\theta \mid s_{\text{obs}})$ can be used to make inferences (e.g., deriving confidence sets, p-values, etc.), although Algorithm 2 may not guarantee 'estimation efficiency' (i.e., producing the tightest confidence sets for all levels) unless the summary statistic is sufficient.

2.1.3 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, 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 Algorithm 2 produces an asymptotically normal confidence distribution. This result, presented in Section 2.3, generalizes the work of [Li & Fearnhead(2018a)] on the asymptotic normality of the approximate Bayesian computing posterior. 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 and performance of this distribution inherited through its connection to confidence distributions. More importantly, the properties we develop here allow us to conduct inference while guaranteeing the frequentist coverage property. Additionally, presented separately in Section 2.2, we specify general conditions under which Algorithm 2 can be used to conduct frequentist inference that is beyond the Bernstein-von Mises type convergence, including exact inference that does not rely 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, the exact inference from Algorithm 2 ensures the targeted repetitive coverage rates and type-I errors.

The main goal of the paper is to present the 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, although we will compare the performance of Algorithm 2 with a typical importance sampling approximate Bayesian computing method and also 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 2.2 and 2.5.

2.1.4 Notation

In addition to the notation from the introduction, throughout the remainder of paper we will use the following notation. The observed data is $x_{\text{obs}} \in \mathcal{X} \subset \mathbb{R}^n$, the summary statistic is a mapping $S_n : \mathcal{X} \to \mathcal{S} \subset \mathbb{R}^d$ and the observed summary statistic is $s_{\text{obs}} = S_n(x_{\text{obs}})$. 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 estimator for θ, we denote this function by $\hat{\theta}_S$. Denote a random draw from $Q_\varepsilon(\theta \mid s_{\text{obs}})$ by θ_{ACC}. Additionally, for a real function $g(x)$, denote its gradient function at some $x = x_0$ by $D_x\{g(x_0)\}$; for simplicity and when it is clear from context, x is omitted from D_x.

2.2 Establishing frequentist guarantees for Algorithm 2

In this section, we formally establish conditions under which Algorithm 2 can be used to produce confidence regions with guaranteed frequentist coverages at any level.

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

$$(\theta_{\text{ACC}} - \widehat{\theta}) \mid S_n = s_{\text{obs}} \sim (\widehat{\theta} - \theta) \mid \theta = \theta_0. \tag{2.4}$$

Then, we can conduct inference for θ with a guaranteed frequentist standard of performance. On the left hand side of (2.4), $\widehat{\theta}$ is fixed given s_{obs} and the (conditional) 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, $\widehat{\theta}$ is a random variable since the data is random for a given parameter θ_0. That is, equation (2.4) states that the 'randomness' in θ_{ACC} from the Monte-Carlo simulation match that in $\widehat{\theta}$ of the sampling population. This is very similar to the bootstrap central limit theorem that $n^{1/2}(\theta_B - \hat{\theta}_S) \mid S_n = s_{\text{obs}} \sim n^{1/2}(\hat{\theta}_S - \theta) \mid \theta = \theta_0$, 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 $S_n = s_{\text{obs}}$, and the randomness on the right hand side is from the random sample of the sampling population.

Given (2.4), let $G(t) = \text{pr}(\widehat{\theta} - \theta \leq t \mid \theta = \theta_0)$. Then $\text{pr}^*(\theta_{\text{ACC}} - \widehat{\theta} \leq t \mid S_n = s_{\text{obs}}) = G(t)$ where $\text{pr}^*(\cdot \mid S_n = s_{\text{obs}})$ refers to the probability measure on simulation given $S_n = s_{\text{obs}}$ corresponding to the left hand side of (2.4). Define $H(t, s_{\text{obs}}) = \text{pr}^*(2\widehat{\theta} - \theta_{\text{ACC}} \leq t \mid S_n = s_{\text{obs}})$, a mapping from $\mathcal{P} \times \mathcal{S} \to (0, 1)$. Conditional on s_{obs}, $H(t, s_{\text{obs}})$ is a sample-dependent cumulative distribution function on $\mathcal{P}$; We use the shorthand $H_n(t)$ to denote $H(t, s_{\text{obs}})$. The following statement Remark1 holds as proved in the appendix. In the remark, $H_n^{-1}(\alpha)$ is the quantile of $H_n(\cdot)$, i.e., the solution of $H_n(t) = \alpha$, and $\theta_{ACC,\alpha}$ is a quantile of θ_{ACC}, defined by $\text{pr}^*(\theta_{\text{ACC}} \leq \theta_{ACC,\alpha} \mid S_n = s_{\text{obs}}) = \alpha$.

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

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 $S_n = s_{\text{obs}}$, 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 one dimensional 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. Corresponding to (2.4), we require a

matching equation: $V(\theta_{\text{ACC}}, S_n) \mid S_n = s_{\text{obs}} \sim W(\theta, S_n) \mid \theta = \theta_0$. Formally, for general mappings V and W, we consider Condition 1 below. In the condition, $\delta_\varepsilon \to 0$, as $\varepsilon \to 0$. Here, ε 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.

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

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

where $\text{pr}^*(\cdot \mid s_{\text{obs}})$ refers to the probability measure on the simulation given $S_n = s_{\text{obs}}$ and $\text{pr}(\cdot \mid \theta_0)$ is the probability measure on the data before it is observed.

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

$$\text{pr}^*\{V(\theta_{\text{ACC}}, S_n) \in A_{1-\alpha} \mid S_n = s_{\text{obs}}\} = (1 - \alpha) + o(\delta'), \tag{2.5}$$

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

$$\Gamma_{1-\alpha}(s_{\text{obs}}) \stackrel{\text{def}}{=} \{\theta : W(\theta, s_{\text{obs}}) \in A_{1-\alpha}\} \subset \mathcal{P} \tag{2.6}$$

is a level $(1 - \alpha)100\%$ confidence region for θ_0. We summarize this in the following lemma which is proved in the appendix. In the next lemma let $\delta = \max\{\delta_\varepsilon, \delta'\}$ thus, 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.

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

Note that there are no requirements on the sufficiency of the summary statistic S_n in Lemma 2. However, if the selected summary statistic happens to be sufficient,

then inference based on the results of Algorithm 2 is equivalent to maximum likelihood inference.

Later in this section we will consider a special case of Lemma 2 that is sample-size independent. In this case, (aside from the errors of Monte-Carlo approximation and the choice of tolerance level) Algorithm 2 can result in *exact* inference procedures that do not rely on large sample asymptotics. Later, in Section 2.3, we use large sample asymptotics to extend Bernstein-von Mises theory to Algorithm 2.

Before we move on to verify Condition 1 for different cases, we first relate equation (2.5) to θ_{ACC} samples from $Q_\varepsilon(\cdot \mid s_{\text{obs}})$. Suppose $\theta_{\text{ACC},i}$, $i = 1, \ldots, m$, are m Monte-Carlo copies of θ_{ACC}. Let $v_i = V(\theta_{\text{ACC},i}, s_{\text{obs}})$. The set $A_{1-\alpha}$ can typically be a $(1-\alpha)100\%$ contour set of $\{v_1, \ldots, v_m\}$ satisfying $o(\delta') = o(m^{-1/2})$. 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 based on the empirical distribution of $\{v_1, \ldots, v_m\}$. 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 $\hat{q}_\alpha = v_{[m\alpha]}$, the $[m\alpha]$th largest $v_1, \ldots, v_m$, a $(1-\alpha)100\%$ confidence region for θ_0 can then be constructed as $\Gamma_{1-\alpha}(s_{\text{obs}}) = \{\theta : \hat{q}_{\alpha/2} \leq W(\theta, s_{\text{obs}}) \leq \hat{q}_{1-\alpha/2}\}$ or $\Gamma_{1-\alpha}(s_{\text{obs}}) = \{\theta : W(\theta, s_{\text{obs}}) \leq \hat{q}_{1-\alpha}\}$.

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 work 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 rough initial estimate, $r_n(\theta)$, means that inference based on the resulting distribution, $Q_\varepsilon(\cdot \mid s_{obs})$, is also efficient.

To end this section, we explore a special case of Algorithm 2 where the mappings V and W correspond an approximate pivotal statistic. Here, we call a mapping $T = T(\theta, S_n)$

from $\mathcal{P} \times \mathcal{S} \to \mathbb{R}^d$ an *approximate pivot statistic*, if

$$\text{pr}\{T(\theta, S_n) \in A \mid \theta = \theta_0\} = \int_{t \in A} g(t)dt \, \{1 + o(\delta^{''})\}, \tag{2.7}$$

where $g(t)$ is a density function that is free of the parameter θ and $A \subset \mathbb{R}^d$ is any Borel set. Also, $\delta^{''}$ is either zero or a small number (tending to zero) that may or may not depend on the sample size n. The usual pivotal cases are special examples of such. Other examples, including that to be discussed in Section 2.3, involve large sample asymptotics where $\delta^{''}$ is a function of n, in particular, $\delta^{''} \to 0$ as $n \to 0$. However, there are also cases where $\delta^{''}$ does not involve 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. In this case, the density function is $\phi(t)\{1 + o(\lambda^{-1})\}$, where $\phi(t)$ the density function of the standard normal distribution [Cheng(1949)].

We have the following theorem for approximate pivot statistics. A proof is given in the appendix.

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

$$\int r_n(\theta) K_\varepsilon \left(s_{t,\theta} - s_{\text{obs}}\right) d\theta = C\{1 + o(\delta_\varepsilon^{'})\}, \text{ where } C \text{ is a constant free of } t, \tag{2.8}$$

Here, $r_n(\theta)$, $K(\cdot)$, and ε are as specified in Algorithm 2, and $\delta_\varepsilon^{'} \to 0$ as $\varepsilon \to 0$. Then, Condition 1 holds almost surely, for $V(\theta, S_n) = W(\theta, S_n) = T(\theta, S_n)$ and $\delta = \max\{\delta^{''}, \delta_\varepsilon^{'}\}$. Furthermore, by Lemma 2 and for observed $S_n = s_{\text{obs}}$, $\Gamma_{1-\alpha}(s_{\text{obs}})$ defined in (2.6) is a level $(1-\alpha)100\%$ confidence region with $pr\{\theta \in \Gamma_{1-\alpha}(S_n) \mid \theta = \theta_0\} = (1-\alpha) + o(\delta)$, almost surely.

Location and scale families contain natural pivot statistics. We verify requirement (2.8) for the location and scale families, which leads to the following corollary. A proof of the corollary is also given in the appendix.

Corollary 1 *Assume $\hat{\mu}_S$ and $\hat{\sigma}_S$ are point estimators for location and scale parameters μ and σ, respectively.*

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

$$|pr^*(\mu_{\text{ACC}} - \hat{\mu}_S \leq u \mid \hat{\mu}_{\text{obs}}) - pr(\hat{\mu}_S - \mu \leq u \mid \mu = \mu_0)| = o(1), \quad \textit{almost surely.}$$

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

$$\left| pr^* \left(\frac{\sigma_{\text{ACC}}}{\hat{\sigma}_S} \leq v \mid \hat{\sigma}_{\text{obs}} \right) - pr \left(\frac{\hat{\sigma}_S}{\sigma} \leq v \mid \sigma = \sigma_0 \right) \right| = o(1), \quad \textit{almost surely.}$$

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

$$\left| pr^* \left(\mu_{\text{ACC}} - \hat{\mu}_S \leq u, \frac{\sigma_{\text{ACC}}}{\hat{\sigma}_S} \leq v \mid \hat{\mu}_{\text{obs}}, \hat{\sigma}_{\text{obs}} \right) - pr \left(\mu_{\text{ACC}} - \hat{\mu}_S \leq u, \frac{\hat{\sigma}_S}{\sigma} \leq v \mid \mu = \mu_0, \sigma = \sigma_0 \right) \right| = o(1), \quad \textit{almost surely.}$$

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 $H_2(\hat{\sigma}_S^2, x) = 1 - \int_0^{\hat{\sigma}_S^2/x} g_2(w) dw$, a confidence distribution for σ^2 induced by $\hat{\sigma}_S^2/\sigma^2$ given $\sigma = \sigma_0$.

Note that Theorem 1 and Corollary 1 cover some finite sample examples that do not require $n \to \infty$, one of which is illustrated in Figure 2.1. Specifically, Corollary 1 Part 1 suggests that the ABC posteriors obtained in the Cauchy example in Figure 2.1, using either the sample mean or sample median as the summary statistic, are both confidence distributions. Thus, they both are 'distribution estimators' that can be utilized to make inference. Both are not efficient, and the one by the sample median is more efficient than the one by the sample mean (in terms of having shorted confidence intervals or a higher power level-α test). This development represents a departure from the typical asymptotic arguments and permits the use of Algorithm 2 in forming confidence sets with guaranteed frequentist coverages even when n is finite.

The next section considers the case in which the tolerance level ε does depend on

the sample size n. We will now denote ε by ε_n and study the large sample performance of the proposed approximate confidence distribution computing method.

2.3 Frequentist Coverage of Algorithm 2 for Large Samples

2.3.1 Bernstein-von Mises theorem for Algorithm 2

For Algorithm 1, Condition **??** holds as $n \to \infty$ by the Bernstein-von Mises type convergence of $\pi_\varepsilon(\theta \mid s_{\text{obs}})$ ([Li & Fearnhead(2018b)]) and selecting ε_n decreasing to zero. Roughly speaking, the distribution of a properly scaled draw from $\Pi_\varepsilon(\theta \mid s_{\text{obs}})$ and the distribution of the corresponding expectation (before the data is observed) are asymptotically the same. Therefore, the development in Section 2.2, a confidence region with asymptotically correct coverage can be constructed using a sample from Algorithm 1.

Here we show that Condition 1 also holds for the more general Algorithm 2 where $r_n(\theta)$ may depend upon the data. The results are based on the same set of conditions as those in [Li & Fearnhead(2018b)]. The key condition is a central limit theorem of the summary statistic: for all θ in a neighborhood of θ_0,

$$a_n\{S_n - \eta(\theta)\} \to N\{0, A(\theta)\},$$

in distribution as $n \to \infty$, together with requirement on the identifiability of θ_0 through $\eta(\theta)$ and regulatory requirements of $A(\theta)$. This condition is denoted by Condition 6 in the supplementary materials. The set of conditions in [Li & Fearnhead(2018b)] is given in the appendix. Here we define some regulatory conditions for $r_n(\theta)$, which is not included in [Li & Fearnhead(2018b)] or, to our knowledge, any of the existing literature on approximate Bayesian computing.

Condition 2 *There exists some* $\delta_0 > 0$ *such that* $\mathcal{P}_0 = \{\theta : \|\theta - \theta_0\| < \delta_0\} \subset \mathcal{P}$, $r_n(\theta) \in C^2(\mathcal{P}_0)$, *and* $r_n(\theta_0) > 0$.

Condition 3 *There exists a sequence* $\{\tau_n\}$ *and* $\delta > 0$, *such that* $\tau_n = o(a_n)$ *and* $\sup_{\theta \in \mathcal{P}_0} \tau_n^{-p} r_n(\theta) = O_p(1)$.

Condition 4 *There exists constants m, M such that* $0 < m <| \tau_n^{-p} r_n(\theta_0) |< M < \infty$.

Condition 5 *It holds that* $\sup_{\theta \in \mathbb{R}^p} \tau_n^{-1} D\{\tau_n^{-p} r_n(\theta)\} = O_p(1)$.

Condition 3 and 4 above essentially requires $r_n(\theta)$ to be more dispersed than the s-likelihood within a compact set containing the true θ_0. It requires that $r_n(\theta)$ converges to a point mass more slowly than $f_n(\theta \mid s_{\text{obs}})$. Condition 5 requires the gradient of the standardized $r_n(\theta)$ to converge with rate τ_n. These are relatively weak conditions and can be satisfied by, e.g., $r_n(\theta)$ satisfying local asymptotic normality. We have the following theorem with the proof provided in the appendix. Note that, in the theorem, $\theta_\varepsilon(s_{\text{obs}})$ is an estimate for θ, whereas $\theta_\varepsilon(S_n)$ is an estimator; when clear, we shorten the notation of both to θ_ε.

Theorem 2 *Assume $r_n(\theta)$ satisfies Condition 2–5 and 6–10 in the supplementary material. If $\varepsilon_n = o(a_n^{-1})$ as $n \to \infty$, then Condition 1 is satisfied with $V(\theta_{\text{ACC}}, \boldsymbol{s}_{\text{obs}}) = a_n\{\theta_{\text{ACC}} - \theta_\varepsilon(s_{\text{obs}})\}$ and $W(\theta_0, S_n) = a_n\{\theta_\varepsilon(S_n) - \theta_0\}$, where $\theta_\varepsilon(s) = \int \theta \, dQ_\varepsilon(\theta \mid s)$.*

Theorem 2 says when $\varepsilon_n = o(a_n^{-1})$, the coverage of $\Gamma_{1-\alpha}(s_{\text{obs}})$ is asymptotically correct as $m_{\text{ACC}} \to \infty$ and $n \to \infty$, where m_{ACC} is the number of accepted particles in Algorithm 2. In practice, $\theta_\varepsilon(s_{\text{obs}})$, needed for constructing $\Gamma_{1-\alpha}$, does not have a closed form in most cases, and is estimated by the sample of θ_{ACC}.

In Theorem 2, Condition 1 is implied by the following convergence results,

$$\sup_{A \in \mathfrak{B}^p} \left| \int_{\{\theta:\, a_n(\theta - \theta_\varepsilon) \in A\}} dQ_\varepsilon(\theta \mid S_n = s_{\text{obs}}) - \int_A N\{t; 0, I(\theta_0)^{-1}\} \, dt \right| \to 0, \tag{2.9}$$

in probability, and

$$a_n(\theta_\varepsilon - \theta_0) \to N\{0, I(\theta_0)^{-1}\}, \tag{2.10}$$

in distribution, as $n \to \infty$, where $I(\theta) = D\eta(\theta)^T A^{-1}(\theta) D\eta(\theta)$. These results generalize the limit distributions of Π_ε in [Li & Fearnhead(2018a)] for the case of $\varepsilon_n = o(a_n^{-1})$, since the prior distribution $\pi(\theta)$ satisfies Condition 2–5. We show that, in the sense of large-sample behavior, inference based on Q_ε is validated whether or not information from the data is used in constructing $r_n(\theta)$.

2.3.2 Comparison between Algorithm 1 and Algorithm 2

Since Π_ε and Q_ε share the same limit distributions according to (2.9) and (2.10), when the same tolerance level is used, confidence regions $\Gamma_{1-\alpha}(\boldsymbol{s}_{\text{obs}})$ constructed using the sample from Π_ε and Q_ε have the same asymptotic efficiency. Therefore it is computationally more efficient to use Algorithm 2 with $r_n(\theta)$ depending on data, since any $r_n(\theta)$ with $\tau_n \to \infty$ is closer to the output distribution than $\pi(\theta)$ thus providing a higher acceptance probability for the same ε.

When $r_n(\theta)$ is available, an alternative to Algorithm 1 is its importance sampling variant which proposes from $r_n(\theta)$ ([Fearnhead & Prangle(2012)]), as specified in the following.

Algorithm 3 *(Importance sampling approximate Bayesian computing)*

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

Though Algorithm 3 is an improvement over Algorithm 1, Algorithm 2 still has a computational advantage over Algorithm 3, because $w(\theta)$ is unbounded as $n \to \infty$ while the sample weights in Algorithm 2 are unity. [Li & Fearnhead(2018b)] mention that certain techniques can be applied to control the skewed importance weight in Algorithm 3, but Algorithm 2 does not have the same issue and therefore does not require such controls.

[Li & Fearnhead(2018b)] point out in Algorithms 1 and 3, that although using $\varepsilon_n = o(a_n^{-1})$ gives valid inference, this leads to the degeneracy of Monte Carlo efficiency as $n \to \infty$, since the acceptance probability of any proposal distribution degenerates to zero for such a small tolerance level. This means that if the dataset is informative, most of the simulated datasets in Algorithms 1 and 3, will be wasted. If ε_n is outside this regime, [Li & Fearnhead(2018b)] show that Π_ε over-inflates the target posterior uncertainty and is not calibrated, i.e its uncertainty can not correctly quantify the uncertainty of the target posterior mean. A similar phenomena occurs in Algorithm 2

when too large ε_n is used. Instead of giving a formal statement, we illustrate this in the following basic Gaussian example.

Example 1 *Consider a univariate normal model with mean θ and unit variance, and observations that are independent identically distributed from the model with $\theta = \theta_0$. Keeping the location and scale parameters of the prior as fixed constants, assume a standard normal density for the prior density of θ. Let $r_n(\theta)$ also be a normal density with mean μ_n and variance b_n^{-2}, where μ_n and b_n are some sequences satisfying $b_n(\mu_n - \theta_0) = O(1)$ and $b_n = o(\sqrt{n})$ as $n \to \infty$. The choice of μ_n and b_n makes $r_n(\theta)$ a reasonable proposal density, since it covers the true parameter θ_0 and is more dispersed than the s-likelihood where the sample mean is the summary statistic in both Algorithm 1 and 2. The Gaussian kernel with variance ε_n^2 is used for the acceptance/rejection.*

For this model, limit distributions of $V(\theta_{\mathrm{ACC}}, s_{\mathrm{obs}})$ and $W(\theta_0, S_n)$ in Theorem 2 for different regimes of ε can be obtained analytically, since $q_\varepsilon(\theta \mid s_{\mathrm{obs}})$ has the closed form $N(\theta; \theta_\varepsilon, \sigma_\varepsilon^2)$ where

$$\theta_\varepsilon = \frac{\boldsymbol{s}_{\mathrm{obs}} + b_n^2(1/n + \varepsilon^2)\mu_n}{1 + b_n^2(1/n + \varepsilon^2)},\ \sigma_\varepsilon^2 = \frac{1/n + \varepsilon^2}{1 + b_n^2(1/n + \varepsilon^2)}.$$

In order for Condition 1 to hold, $V(\theta_{\mathrm{ACC}}, s_{\mathrm{obs}})$, which has the density $N(\cdot\,; 0, n\sigma_\varepsilon^2)$, and $W(\theta_0, S_n)$, which is equal to $\sqrt{n}(\theta_\varepsilon - \theta_0)$, should have the same asymptotic distributions.

By decomposing $W(\theta_0, S_n)$ into $\Delta_1\sqrt{n}(S_n - \theta_0) + \Delta_2 b_n(\mu_n - \theta_0)$ 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 the expectation of $W(\theta_0, S_n)$ is $o(1)$ only when $\varepsilon_n = o(b_n^{-1/2}n^{-1/4})$. On the other hand, the variance of $W(\theta_0, S_n)$ and $n\sigma_\varepsilon^2$ having the same limit requires $n\sigma_\varepsilon^2 - \Delta_1^2 = o(1)$ which holds only when $\varepsilon_n = o(n^{-1/2})$ or $\varepsilon_n^{-1} = o(b_n^2 n^{-1/2})$. Because $b_n = o(\sqrt{n})$, 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 when $\varepsilon_n = o(n^{-1/2})$.

One remedy to reduce the overinflated uncertainty in $\Pi_\varepsilon(\theta \mid \boldsymbol{s}_{\mathrm{obs}})$ from Algorithms 1 and 3 is to post-process its sample by the regression adjustment. ([Beaumont et al.(2002)])

Likewise, this adjustment can be applied to Algorithm 2. In the next subsection, we compare these regression adjusted approximate computing methods.

2.3.3 Comparison between Algorithm 1 and Algorithm 2 with regression adjustment

For Algorithms 1 and 3, it is known that the distribution of the regression adjusted sample is able to correctly quantify the posterior uncertainty and yield an accurate point estimate with ε_n decaying in the rate of $o(a_n^{-3/5})$, which is slower than $o(a_n^{-1/2})$ ([Li & Fearnhead(2018a)]). Here, we suggest applying the same regression adjustment to Algorithm 2 to produce valid inference on the sample of Algorithm 2 with a larger ε_n.

Let $q_\varepsilon(\theta, s)$ be the joint density of accepted θ and its associated summary statistic in Algorithm 2, i.e.

$$q_\varepsilon(\theta, s) = \frac{r_n(\theta) f_n(s \mid \theta) K_\varepsilon(s - s_{\text{obs}})}{\int_{\mathbb{R}^p \times \mathbb{R}^d} r_n(\theta) f_n(s \mid \theta) K_\varepsilon(s - s_{\text{obs}})\, d\theta ds}, \tag{2.11}$$

where $\theta \in \mathbb{R}^p$ and $s \in \mathbb{R}^d$. Denote a sample from $q_\varepsilon(\theta, s)$ by $\{(\theta_i, s^{(i)})\}_{i=1,\ldots,N}$. A new sample can be obtained as $\{\theta_i - \hat{\beta}_\varepsilon(s^{(i)} - s_{\text{obs}})\}_{i=1,\ldots,N}$ 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 errors, $\alpha \in \mathbb{R}^p$ and $\beta \in \mathbb{R}^{p \times d}$. Let $\theta^*_{\text{ACC}} = \theta - \beta_\varepsilon(s - s_{\text{obs}})$, where β_ε is from the minimizer

$$(\alpha_\varepsilon, \beta_\varepsilon) = \text{argmin}_{\alpha \in \mathbb{R}^p, \beta \in \mathbb{R}^{d \times p}} E_\varepsilon \left\{ \|\theta - \alpha - \beta(s - s_{\text{obs}})\|^2 \mid s_{\text{obs}} \right\}$$

for expectation under the joint distribution $q_\varepsilon(\theta, s)$. The new sample can be seen as a draw from the distribution of θ^*_{ACC} where $(\theta, s) \sim q_\varepsilon(\theta, s)$, but with β_ε replaced by its estimator. Let θ^*_ε be the expectation of θ^*_{ACC}.

The following theorem states that the regression adjusted Q_ε has the same favored

property as the adjusted Π_ε. Here, the regression adjusted Q_ε, say $Q^*_\varepsilon(\cdot|S_n = s_{\text{obs}})$, is the distribution of θ^*_{ACC} given $S_n = s_{\text{obs}}$.

Theorem 3 *Assume the conditions of Theorem 2 and Condition 10 of the supplementary materials. If $\varepsilon_n = o(a_n^{-3/5})$ as $n \to \infty$, Condition 1 is satisfied with $V(\theta^*_{\text{ACC}}, s_{\text{obs}}) = a_n(\theta^*_{\text{ACC}} - \theta^*_\varepsilon)$ and $W(\theta_0, S_n) = a_n(\theta^*_\varepsilon - \theta_0)$.*

In the above, Condition 1 is implied by the following convergence results which generalize the results in [Li & Fearnhead(2018a)],

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

in probability, and

$$a_n(\theta^*_\varepsilon - \theta_0) \to N\{0, I(\theta_0)^{-1}\},$$

in distribution, as $n \to \infty$. The limit distributions above are the same as those in (2.9) and (2.10), therefore $\Gamma_{1-\alpha}(\boldsymbol{s}_{\text{obs}})$ constructed using θ^*_{ACC} can achieve the same efficiency as those using θ_{ACC} while permitting much larger tolerance levels. Asymptotically, inference based on the regression adjusted Q^*_ε is not affected by an $r_n(\theta)$ that depends on the data, again illustrating the computational advantage of Algorithm 2.

2.3.4 Guidelines for Selecting r_n in Algorithm 2

The generality of approximate confidence distribution computing is that it can produce justifiable inferential results with weak conditions on a possibly data-dependent function $r_n(\theta)$. In general, one should be careful in choosing $r_n(\theta)$ to ensure its growth with respect to the sample size is slower than the growth of the s-likelihood, according to Condition 3. A generic algorithm to construct $r_n(\theta)$ based on sub-setting the data is proposed below. Assume that a point estimator $\widehat{\theta}(z)$ of θ can be computed for a dataset z of any size.

Algorithm 4 *(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* $\widehat{\theta}_i = \widehat{\theta}(z_i)$*, for* $i = 1, \ldots, k$.
3. *Let* $r_n(\theta) = (1/kh) \sum_{i=1}^{k} K\left\{h^{-1}\|\theta - \widehat{\theta}_i\|\right\}$*, where* $h > 0$ *is the bandwidth of the kernel density estimate using* $\{\widehat{\theta}_1, \ldots, \widehat{\theta}_k\}$ *and kernel function* K.

By choosing $\nu < 3/5$, we ensure that Conditions 3–5 are met. Furthermore, if $\widehat{\theta}(z)$ converges with a rate not faster than that of the summary statistic, then the tolerance level, ε_n, selected by accepting a reasonable proportion of simulations is sufficiently small, provided the rate of S_n is a power function of n. 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 overlap the subsets so that each subset contains a reasonable number of observations.

The choice of $\widehat{\theta}$ does not have to be very accurate, since it is only used to construct the initial estimate, $r_n(\theta)$. For problems of intractable likelihoods, possible choices of $\widehat{\theta}$ include the point maximizing an easy-to-obtain approximate likelihood or the point minimizing the average distance between the simulated s and $\boldsymbol{s}_{\text{obs}}$ ([Meeds & Welling(2015)]). However, a poor choice, for instance, a $\widehat{\theta}$ with a large bias, might cause bias in the inference if the mass of $Q_\varepsilon(\theta \mid \boldsymbol{s}_{\text{obs}})$ is not well covered by the simulated parameter values. For a subset, z_i, of the data, x_{obs}, we suggest choosing the point estimate to be the s-likelihood-based expectation over the subset, i.e. $E\{\theta \mid S_n(z_i)\} \propto \int \theta f_n\{S_n(z_i) \mid \theta\} d\theta$. This choice of $\widehat{\theta}$ has two benefits. First, when the summary statistic satisfies Condition 6, $E\{\theta \mid S_n(z_i)\}$ is asymptotically unbiased. Second, $E\{\theta \mid S_n(z_i)\}$ converges with the same rate as S_n, which is desirable as discussed above.

For each subset z_i of x_{obs}, $E\{\theta \mid S_n(z_i)\}$ can be approximated using the population Monte Carlo variant of Algorithm 1. ([Beaumont et al.(2009), Del Moral et al.(2012)]) This variant extends the importance sampling step of Algorithm 3 to a sequence of sampling importance resampling operations, in order to iteratively update the approximate posterior distribution starting from the prior distribution. For an initial choice of $\widehat{\theta}$, say $\widehat{\widehat{\theta}}$, let $\bar{r}_n(\theta)$ be the proposal distribution constructed by Algorithm 4 together with $\widehat{\widehat{\theta}}$. Here the user can now propose from $\bar{r}_n(\theta)$ in the first iteration of the algorithm rather than proposing from the prior distribution, helping to reduce the associated computational

cost. This approximation is straightforward to execute in parallel for multiple subsets and can be applied to Algorithm 2 as well. We call this scheme the *refined-minibatch scheme*, since it updates the $r_n(\theta)$ obtained from the minibatch scheme (i.e. Algorithm 4) by improving the quality of $\widehat{\theta}$. From our experience, the additional computational cost of the refined version is relatively small compared to the other parts of Algorithm 1 and 2 because a small particle size and several iterations are usually enough to achieve convergence of the population Monte Carlo algorithm with the proposed techniques. A full study on the choice of $\widehat{\theta}$ is beyond the scope of this study.

Remark 2 *There is a trade-off in Algorithm 2 between faster computations and guaranteed frequentist inference. When the growth of $r_n(\theta)$ is at a similar rate as the s-likelihood while the sample size $n \to \infty$, the computing time may be reduced but Algorithm 2 may also risk violating Conditions 3–5. If these assumptions are violated, the resulting simulations do not necessarily form a confidence distribution and consequently, inference based on Algorithm 2 may not be valid in terms of producing confidence sets with guaranteed coverage. However, if Conditions 3–6 do hold and the observed data is large enough, Theorem 2 shows that regardless of the choice of $r_n(\theta)$, Algorithm 2 always produces the same confidence distribution.*

2.4 Numerical Examples

2.4.1 Cauchy example

In Figure 2.1 we saw how the lack of a sufficient summary statistic could drastically change the inferential results of approximate Bayesian computing. In this section the following, we revisit the problem of finding confidence intervals for the parameters of IID data, $(x_1, \ldots, x_n)$, from a $Cauchy(\theta, \tau)$ distribution. In this section we take $\tau = 0.55$ as known but we consider τ unknown in the following section.

In Table 2.1, we fix the tolerance at three different levels and compare the proportion of accepted θ values out of a Monte Carlo sample size of 10^6 among Algorithms 1, 2, and 3. In Algorithms 1 and 3 we use a "non-informative" $\pi(\theta) \propto 1/(1 + \frac{\theta-\theta_0}{3})^2$. (Note that this prior has been centered around θ_0. This was done for practicality but for a truly

Table 2.1: *Comparison of r-ABC, IS-ABC, and r-ACC without the regression adjustment for inference on θ using the median as the summary statistic and assuming a flat prior on θ. We fix ε_n and compare the median acceptance proportions of each algorithm using a Monte Carlo sample size of 10^6. Coverage is computed over 300 runs. IS-ABC and r-ACC perform similarly.*

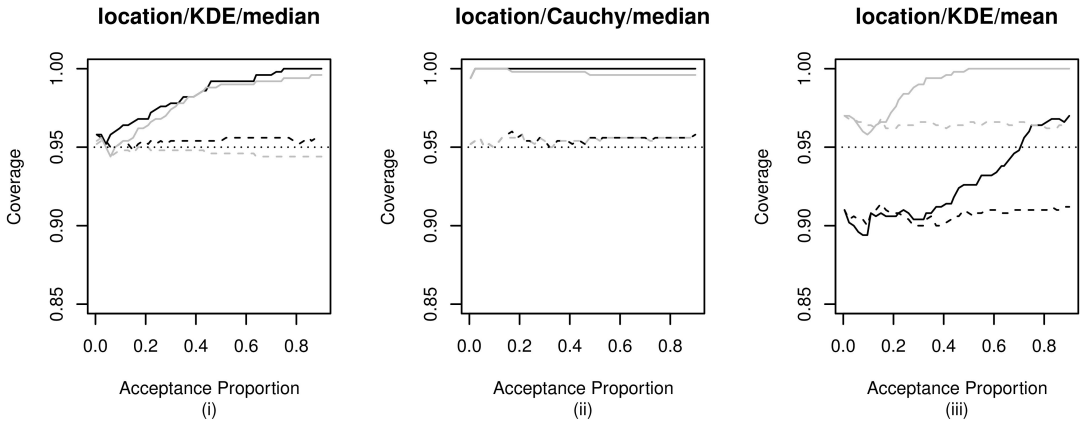
ε_n	Acceptance Proportion		
	r-ABC	IS-ABC	r-ACC
0.1	0.0001	0.001	0.001
0.01	0.001	0.008	0.008
0.001	0.008	0.079	0.079

non-informative prior, the acceptance proportion of Algorithm 1 and 3 can be much lower.) The wights for Algorithm 3 are $w(\theta) = \pi(\theta)/r_n(\theta))$ and for Algorithms 2 and 3, $r_n(\theta)$ is constructed using Algorithm 4 with $\nu = 1/2$. We see a drastic difference in the acceptance proportion between Algorithm 1 and the other two methods, demonstrating the computational advantage inherent to Algorithm 2.

The conclusion of Table 2.1 is that it is more reasonable to compare the performance of Algorithm 3 to Algorithm 2 than it is to compare Algorithm 1 and Algorithm 2. However, as the solid lines in Figure 2.2 indicate, both Algorithm 2 and 3 suffer from sever over-coverage. To address this issue, we also consider the performance of post-processed outputs of IS-ABC and rejection ACC using the regression adjustment from Section 2.3.3. Note that in Figure 2.2 we consider two different choices of summary statistic. For $S = Med(x_1, \ldots, x_n)$, we also consider $r_n(\theta) \propto 1/(\frac{\theta - \bar{x}}{\tau_0})^2$ which still satisfies Conditions 1–5 from Section 2.3.

Figure 2.2 shows that in all cases, intervals constructed by the unadjusted samples are much wider and over-cover the true parameter values for almost all acceptance proportions, in accordance with the discussion in Section 2.3.3. The performance of Algorithms 2 and 3 are similar for a more informative choice of summary statistic, however for a less informative summary statistic, there is a considerable difference in the performance of these two methods as the coverage of Algorithm 3 is much lower than the nominal level. Following the conclusions of this discussion, in the next section we will only compare the regression adjusted versions of Algorithms 2 and 3, for a fairer

Figure 2.2: Confidence interval coverage results of rejection ACC (gray) and importance sampling ABC (black) for Cauchy location parameter estimation both with (dashed) and without (solid) regression adjustment. In plots (i) and (ii), $S_n = Median(x_1, \ldots, x_n)$ and in plot (iii) $S_n = \bar{x}$. Coverage is calculated over 500 runs and the Monte Carlo size of each run is 5×10^5. Dashed black line is the nominal coverage level. KDE means that $r_n(\theta)$ is constructed using Algorithm 4 with $\nu = 1/2$ and Cauchy means that $r_n(\theta) \propto 1/(\frac{\theta - \bar{x}}{\tau_0})^2$.



comparison and a more realistic understanding of the advantages of Algorithm 2.

2.4.2 Cauchy example expanded

To expand upon the example in the previous section, we will now compare the performance of regression-adjusted Algorithm 2 and 3 in seven different experiment settings. For reference, all experiment settings explored in this section are summarized in Table 2.2. In each of these seven settings, we construct $r_n(\theta)$ using Algorithm 4 with $\nu = 1/2$.

The different prior distribution choices for each of the Bayesian methods are the Jeffrey's priors for a location-scale family. In settings (i)-(iii) of Table 2.2, we consider inference for one or both of the unknown parameters. We choose the summary statistics as the sample median and sample median absolute deviation for the location and scale parameters, respectively. These summary statistics are asymptotically normal and unbiased and satisfy Condition 6; thus Theorem 3 guarantees at least nominal coverage for the intervals/regions of Algorithm 2. In settings (iv)-(vii), we consider inference for the location parameter only using the sample mean as the summary statistic. Here, the

Table 2.2: *Experiment settings of Example 1. Improper priors are considered for (i)–(iv). In the table,* $t_4(\mu, \sigma)$ *denotes the Student's t density with degree of freedom four, location* μ *and scale* σ *and* $\mathrm{MAD}(x)$ *represents the sample median absolute deviation.*

	Unknown parameter	Prior density	Summary statistic
(i)	θ	1	$\mathrm{median}(x)$
(ii)	τ	$\tau^{-1} I_{\tau>0}$	$\mathrm{MAD}(x)$
(iii)	(θ, τ)	$\tau^{-1} I_{\tau>0}$	$\{\mathrm{median}(x), \mathrm{MAD}(x)\}$
(iv)	θ	1	$\overline{x}$
(v)	θ	$t_4(\theta_0, 1)$	$\overline{x}$
(vi)	θ	$t_4(\theta_0, 3)$	$\overline{x}$
(vii)	θ	$t_4(\theta_0 + 3, 3)$	$\overline{x}$

summary statistic does not satisfy condition 6, but it does satisfy the conditions for Corollary 1.

The results of these seven experiments are summarized in Table 2.3. For settings (i)-(iii), Table 2.3.A shows that both algorithms perform very similarly. This is not surprising, since the data size is large enough that the asymptotic behaviors of all estimates are similar. As discussed in Section 2.3.3, we see here that $Q_\varepsilon(\cdot \mid s_{obs})$ and $\Pi_\varepsilon(\cdot \mid s_{obs})$ share the same limiting normal distribution and thus the credible intervals/region from Algorithm 3 are similar to the confidence intervals/region from Algorithm 2.

For settings (iv)-(vii), Table 2.3.B shows that although the summary statistic is less informative, because it is a pivotal quantity, Corollary 1 guarantees that Algorithm 2 will produce confidence intervals with at least the nominal coverage level. Here we consider four different choices of $\pi(\cdot)$ for Algorithm 3: a non-informative prior in (iv), informative priors in (v) and (vi) and a misspecified prior in (vii). Regardless of the choice of prior, Algorithm 3 uses a summary statistic that does not meet the conditions for a Bernstein von-Mises type of theorem and the coverage of intervals from Algorithm 3 are far lower than the 95% nominal level. Not only do the intervals based on Algorithm 2 attain the nominal coverage level, as indicated in Table 2.3, the intervals from Algorithm 2 are more efficient than the credible intervals from Algorithm 3, the former having widths

about half the widths of the latter (except in the somewhat less realistic setting (v) where the prior is highly informative).

2.4.3 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,

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

where $t = 1, \ldots, T$. Parameters r, ϕ and σ are positive constants, interpreted as the intrinsic growth rate of the population, a scale parameter and the environmental noise. This model is statistically challenging since its likelihood function is intractable when σ is non-zero and highly irregular in certain regions of the parameter space. [Wood(2010)] suggests a summary statistic-based inference, instead of likelihood-based inference, to overcome the noise-driven nature of the model. [Fearnhead & Prangle(2012)] applies Algorithm 3 with the regression adjustment on the above model. In this section, we apply Algorithm 2 with the regression adjustment and compare its performance with that of regression-adjusted Algorithm 3.

We consider inference on the unknown parameter $\theta = (r, \phi, \sigma)$. A total of four different methods are compared. (i) Algorithm 2 with the regression adjustment; (ii) Algorithm 3 with the regression adjustment; both using Algorithm 4 to choose $r_n(\theta)$. (iii) Algorithm 2 with the regression adjustment; (iv) Algorithm 3 with the regression adjustment; both using Algorithm 4 with the refinement to choose $r_n(\theta)$. The main computational cost of all four algorithms is associated with the calculation of the point estimate in Algorithm 4, for which we select the maximum synthetic likelihood estimator as defined in [Wood(2010)]. Because each point estimate requires the simulation of a Markov chain Monte Carlo sample for the synthetic likelihood, each of the four algorithms spend over 50% of CPU time on obtaining $r_n(\theta)$. Relative to this cost,

Table 2.3: *Coverage proportions and the median width/volume of confidence or credible intervals/regions, calculated using* 300 *datasets under settings of Table 2.2. For credible intervals, both the frequentist coverage proportions and the Bayesian coverage probabilities are reported, the latter are given in the parenthesis. Each dataset contains* 400 *observations, and in each algorithm run, a Monte Carlo sample of size* 10^5 *is simulated. The nominal level is* 95% *and we report the median widths and volumes of the resulting intervals/regions.*

(A) Using an informative summary statistics for θ and τ.

Setting		r-ACC		IS-ABC	
	Acceptance proportion	Coverage	Width/ Volume	Coverage	Width/ Volume
(i) θ/ Median	0.005	0.947	0.162	0.950 (0.955)	0.169
	0.1	0.947	0.165	0.950 (0.957)	0.17
	0.4	0.947	0.166	0.950 (0.958)	0.17
(ii) τ/ MAD	0.005	0.950	0.163	0.947 (0.955)	0.169
	0.1	0.937	0.165	0.950 (0.958)	0.170
	0.4	0.943	0.164	0.950 (0.957)	0.171
(iii) (θ, τ) / (Median,MAD)	0.005	0.913	0.059	0.917	0.059
	0.1	0.933	0.100	0.92	0.100
	0.4	0.94	0.141	0.927	0.141

(B) Using an un-informative summary statistic for θ, i.e. $S_n = \bar{x}$.

Setting	Acceptance proportion	r-ACC		IS-ABC	
		Coverage	Width	Coverage	Width
(iv) $1_{\theta \in \mathbb{R}}$	0.005	0.970	2.56	0.983 (1)	4.65
	0.1	0.973	2.56	0.973 (1)	5.39
	0.4	0.963	2.65	0.967 (1)	5.58
(v) $t_4(\theta_0, 1)$	0.005	0.970	2.56	1 (1)	2.69
	0.1	0.973	2.56	1 (1)	2.65
	0.4	0.963	2.65	1 (1)	2.76
(vi) $t_4(\theta_0, 3)$	0.005	0.970	2.56	1 (1)	3.93
	0.1	0.973	2.56	1 (1)	4.32
	0.4	0.963	2.65	1 (1)	4.42
(vii) $t_4(\theta_0 + 3, 3)$	0.005	0.970	2.56	0.93 (1)	4.40
	0.1	0.973	2.56	0.89 (1)	5.33
	0.4	0.963	2.65	0.89 (1)	5.61

the additional cost of the population Monte Carlo algorithm in the refined-minibatch scheme is negligible when using 10^4 particles and 10 iterations run in parallel. In this example, the parametric bootstrap method is not feasible due to the large number of point estimates it would need to calculate.

Following the settings used in [Wood(2010)], our dataset contains observations from $t = 51$ to 100, generated using parameter value $\theta = (e^{3.8}, 0.3, 10)$, and using the same summary statistic therein. We assume θ follows an improper uniform prior distribution over all positive values. In Algorithm 4, each minibatch has size 10 and a total number of 40 batches are used. They are chosen with overlaps in order to ensure a reasonable number of point estimates are available in the current small data size setting. Results are given in Table 2.4. Because the regression adjustment methods are better in all cases, to save time and space we only report here results for regression adjustment methods. The simulation results without the minibatch refinement, show that IS-ABC has somewhat better coverage than r-ACC since the point estimates (and thus $r_n(\cdot)$) are biased in the small data size setting. However, with the refined-minibatch scheme, the width of the confidence intervals for r-ACC are smaller than those in IS-ABC in all cases, although both methods are over-coverage (here the target is 0.95). This result illustrates the benefit of improving $r_n(\theta)$ through the population Monte Carlo procedure on problems with poor initial choice of $r_n(\theta)$. In the Cauchy example above, using the refined-minibatch scheme would improve upon the results however the improvement would be minimal and not as strong as in the Ricker example.

2.5 Discussion

In this work, we re-frame the well-studied popular approximate Bayesian computing method 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*, 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, the new method quantifies

Table 2.4: *Coverage proportions and the median width of confidence/coverage intervals calculated using* 150 *datasets for the four different methods of the Ricker model in Example 2 with* $\delta = 3/5$ *for r-ACC and a flat prior for IS-ABC. Each dataset contains* 50 *observations, and in each algorithm run, a Monte Carlo sample of size* 10^6 *is simulated. The nominal level is* 95%.

(A) Using Algorithm 4 to construct $r_n(\theta)$.

		r-ACC		IS-ABC	
	Acceptance proportion	Coverage	Width	Coverage	Width
$\log R$	0.005	0.91	0.59	0.91	0.72
	0.1	0.91	0.59	0.99	0.89
	0.4	0.9	0.61	0.99	0.99
$\log \sigma$	0.005	0.96	2.46	0.95	2.59
	0.1	0.95	2.78	0.96	2.90
	0.4	0.94	2.9	0.97	2.89
$\log \phi$	0.005	0.89	0.21	0.92	0.24
	0.1	0.91	0.21	0.94	0.30
	0.4	0.91	0.23	0.97	0.33

(B) Using the refined version of Algorithm 4 to construct $r_n(\theta)$.

		r-ACC		IS-ABC	
	Acceptance proportion	Coverage	Width	Coverage	Width
$\log R$	0.005	0.96	0.85	0.97	0.95
	0.1	0.99	0.97	0.99	1.24
	0.4	1.00	1.17	0.99	1.96
$\log \sigma$	0.005	0.96	1.3	0.97	1.63
	0.1	0.97	1.37	0.99	1.92
	0.4	1.00	1.51	0.99	2.29
$\log \phi$	0.005	0.96	0.28	0.97	0.31
	0.1	0.99	0.35	0.99	0.43
	0.4	0.98	0.55	1.00	0.86

the uncertainty in estimation by drawing upon a direct connection to a confidence distribution. This connection guarantees that confidence sets based on approximate confidence distribution computing methods attain the frequentist coverage property even in cases where one has a finite sample size and the cases when the summary statistic used in the computing is not sufficient. Thus we provide theoretical support for inference from approximate confidence distribution methods which include, but are not limited to, the special case where we do have prior information (i.e. approximate Bayesian computing). Furthermore, in the case where the selected summary statistic is sufficient, inference based on the results of Algorithm 2 is equivalent to maximum likelihood inference. In addition to providing sound theoretical results for inference, the framework of approximate confidence distribution computing sets the user up for better computational performance by allowing the data to drive the algorithm through the choice of $r_n(\theta)$. The potential computational advantage of our method has been illustrated through numerical examples.

Different choices of summary statistics often lead to different approximate Bayesian computing posteriors $\pi_\varepsilon(\theta \mid s_{\text{obs}})$ in Algorithms 1 and 3 and different approximate confidence distribution $q_\varepsilon(\theta \mid s_{\text{obs}})$ in Algorithm 2. We find the philosophical interpretation of the results admitted through approximate confidence distribution computing to be more natural than the Bayesian interpretation of approximate Bayesian computing posteriors. Within a frequentist setting, it makes sense to view the many different potential confidence distributions produced by our method resulting from different choices of summary statistics as various choices of (distribution) estimators. However, within the Bayesian framework, there is no clear way to choose from among the different approximate 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 approximate Bayesian computing posteriors. Rather than engaging in a pursuit to define a moving target such as this, our method maintains a clear frequentist interpretation thereby offering a consonantly cohesive interpretation of likelihood-free methods.

In Section 2.3.4, one may wonder if an estimate, $\widehat{\theta}$, can be computed, then why not apply the parametric bootstrap method to construct confidence regions for θ as opposed to using Algorithm 2? Although no likelihood evaluation is needed, this bootstrap method has two drawbacks. First, the parametric bootstrap method is heavily affected by the quality of $\widehat{\theta}$. For example, a bootstrapped confidence interval is based on quantiles of $\widehat{\theta}$ from simulated datasets. A poor estimator $\widehat{\theta}$ typically leads to poor performing confidence sets. In contrast, in Section 2.3.4, $\widehat{\theta}$ is only used to construct the initial function estimate which is then updated by the data. Second, when it is more expensive to obtain $\widehat{\theta}$ than the summary statistic, the parametric bootstrap method is computationally more costly than Algorithm 2, since $\widehat{\theta}$ needs to be calculated for each pseudo dataset. Example 4.2 in Section 2.4 provided an example of this type of scenario.

The function $r_n(\theta)$ serves as the role of an initial 'distributional estimate'. Even in the instance where $r_n(\theta)$ does not yield reasonable acceptance probabilities for Algorithm 2, many of the established techniques used in approximate Bayesian computing can be adapted naturally to Algorithm 2 to improve computational performance. For example, the likelihood-free Markov chain Monte Carlo ([Marjoram et al.(2003)]) and the dimension-reduction methods on the summary statistics ([Fearnhead & Prangle(2012)]), among others, can improve Algorithm 2 without sacrificing frequentist inferential guarantees. Furthermore, these variants of Algorithm 2 will be more efficient than the corresponding variants of Algorithm 1, since $r_n(\theta)$ is less dispersed than the prior.

Appendix 1

Example of a confidence distribution

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.

Lemma 1

Proof: The density of π_ε can be expressed by

$$\begin{aligned}
\pi_\varepsilon(\theta|s_{\text{obs}}) &\propto \int_{\mathbb{R}^d} \pi(\theta) f_n(s \mid \theta) K_\varepsilon(s - s_{\text{obs}}) ds \\
&= \pi(\theta) \int \big\{ f_n(s_{\text{obs}} \mid \theta) + Df_n(\bar{s} \mid \theta)^T (\bar{s} - s) \\
&\qquad + (1/2)(\bar{s} - s)^T Hf_n(\bar{s} \mid \theta)(\bar{s} - s) \big\} K_\varepsilon(s - s_{\text{obs}}) ds \\
&\propto \pi(\theta) f_n(s_{\text{obs}}|\theta) + O(\varepsilon^2),
\end{aligned}$$

where $Df_n(\cdot \mid \theta)$ and $Hf_n(\cdot \mid \theta)$ are the vector of first derivatives and matrix of second derivatives of $f_n(\cdot \mid \theta)$, respectively, 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 $f_n(\cdot \mid \theta)$ with respect to s_{obs} and the final proportion holds using the substitution $u = (s - s_{\text{obs}})$ and that $\int_{\mathbb{R}^d} K_\varepsilon(u)\, du = 1$ and $\int_{\mathbb{R}^d} uK_\varepsilon(u)\, du = 0$. □

Remark 1 in Section 2

Proof: By its definition, $H_n(\cdot) = H(\cdot, s_{\text{obs}})$ is a sample-dependent cumulative distribution function on the parameter space. We also have $H_n(\theta_0) = H(\theta_0, s_{\text{obs}}) = \text{pr}^*(2\widehat{\theta} - \theta \le \theta_0 \mid S_n = s_{\text{obs}}) = \text{pr}^*(\theta - \widehat{\theta} \ge \widehat{\theta} - \theta_0 \mid S_n = s_{\text{obs}}) = 1 - G(\widehat{\theta} - \theta_0)$. Since $G(t) = \text{pr}(\widehat{\theta} - \theta \le t \mid \theta = \theta_0)$, we have $G(\widehat{\theta} - \theta_0) \sim Unif(0,1)$ under the probability measure of the random sample population. Thus, as a function of the random S_n, $H_n(\theta_0) = H_n(\theta_0, S_n) \sim Unif(0,1)$. By the univariate confidence distribution definition, $H_n(\cdot)$ is a confidence distribution function.

Furthermore, $H_n(\cdot)$ can provide us confidence intervals of any level. In particular, for any $\alpha \in (0,1)$, $\text{pr}\{\theta \le H_n^{-1}(1-\alpha) \mid \theta = \theta_0\} = \text{pr}\{H_n(\theta) \le 1-\alpha \mid \theta = \theta_0\} = 1-\alpha$. Thus, $(-\infty, H_n^{-1}(1-\alpha)]$ is a $(1-\alpha)$-level confidence interval. Note that, $H_n(2\widehat{\theta} - \theta_\alpha) = \text{pr}^*(2\theta_{ACC} - \theta \le 2\theta - \theta_\alpha \mid S_n = s_{\text{obs}}) = 1 - \text{pr}^*(\theta < \theta_\alpha \mid S_n = s_{\text{obs}}) = 1-\alpha$. So, $H_n^{-1}(1-\alpha) = 2\widehat{\theta} - \theta_\alpha$. Therefore, $(-\infty, 2\widehat{\theta} - \theta_\alpha]$ is also a $(1-\alpha)$-level confidence interval for θ. □

Lemma 2

Proof: First note that

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

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

Theorem 1

Proof: Setting $W(\theta, S_n) = T(\theta, S_n)$ and by (2.7), we immediately have

$$
\text{pr}\{W(\theta, S_n) \in A \mid \theta = \theta_0\} = \int_{t \in A} g(t)dt \, \{1 + o(\delta'')\}, \tag{2.12}
$$

for any Borel set $A \subset \mathbb{R}^d$.

Let $f(s|\theta)$ be the conditional density of S_n, given θ. Note that t and S_n have the same dimension. For a given θ and with the variable transformation $T = T(\theta, S_n)$, the density functions $g(t)$ and $f(s_{t,\theta}|\theta)$ are connected by a Jacobi matrix: $f(s_{t,\theta}|\theta)|T^{(1)}(\theta, s_{t,\theta})|^{-1} = g(t)\{1 + o(\delta'')\}$, where $T^{(1)}(\theta, s) = \frac{\partial}{\partial s} T(\theta, s)$ and $s_{t,\theta}$ is the solution of $t = T(\theta, s)$.

In Algorithm 2, we simulate $\theta' \sim r_n(\theta)$ and $s' = S_n(x')$ with $x'|\theta = \theta' \sim M_{\theta'}$. Furthermore, we only keep those pairs (θ', s') with the kernel probability $K_\varepsilon(s' - s_{\text{obs}})$. Thus, the joint density function of a copy of (θ', s') that are simulated and kept by Algorithm 2, conditional on observing $S_n = s_{\text{obs}}$, is

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

Now, let $T' = T(\theta', s')$. Perform a variable transformation from (θ', s') to (θ', T') with the Jacobi term $|T^{(1)}(\theta', s_{T',\theta'})|^{-1}$, where $s_{T',\theta'}$ is a solution to $T' = T(\theta', s)$. Then, the joint conditional density of (θ', T'), conditional on $S_n = s_{\text{obs}}$, is

$$\begin{aligned}(\theta', T')|S_n = s_{\text{obs}} &\propto r_n(\theta') f_n(s_{T',\theta'} \mid \theta')|T^{(1)}(\theta', s_{T',\theta'})|^{-1} K_\varepsilon(s_{T',\theta'} - s_{\text{obs}}). \\ &= r_n(\theta') g(T') K_\varepsilon(s_{T',\theta'} - s_{\text{obs}})\{1 + o(\delta^{''})\}.\end{aligned}$$

Therefore, $T' = T(\theta', s')$, the approximate pivot statistic generated from Algorithm 2, with distribution conditional on $S_n = s_{\text{obs}}$:

$$T'|S_n = s_{\text{obs}} \propto g(t')\{1 + o(\delta^{''})\} \int r_n(\theta') K_\varepsilon(s_{t',\theta'} - s_{\text{obs}}) d\theta'$$

If requirement (2.8) is satisfied, then we have

$$T'|S_n = s_{\text{obs}} \sim g(T')\{1 + o(\delta^{''})\}\{1 + o(\delta_\varepsilon^{'})\}.$$

Set $V(\theta', s') = T' = T(\theta', s')$ and denote by θ_{ACC} the θ' accepted by the ACC algorithm. We have

$$\text{pr}^*\{V(\theta_{\text{ACC}}, S_n) \in A \mid S_n = s_{\text{obs}}\} = \int_{t \in A} g(t) dt \{1 + o(\delta^{''})\}\{1 + o(\delta_\varepsilon^{'})\}$$

Thus, together with (2.12), Condition 1 is satisfied for $\delta_\varepsilon = \max\{\delta^{''}, \delta_\varepsilon^{'}\}$. Furthermore, by Lemma 2, the rest of the statements in the theorem also hold. □

Corollary 1

Proof: Here we prove requirement (2.8) for Part 2, data from a scale family. The proofs for Part 1 (location family) and Part 3 (location and scale family) are similar and thus omitted.

In particular, in a scale family suppose S_n has the density $(1/\sigma)g_2(S_n/\sigma)$. Then $T = T(\sigma, S_n) = S_n/\sigma \sim g_2(t)$ is a pivot. So, for any given (t, σ) pair we have $s_{t,\sigma} = t\sigma$. Thus, with variable transformation $u = t\sigma - s_{\text{obs}}$ we have

$$\begin{aligned} \int r_n(\sigma) K_\varepsilon \left(s_{T,\sigma} - s_{\text{obs}}\right) d\sigma &= \int \frac{1}{\sigma} K_\varepsilon \left(s_{T,\sigma} - s_{\text{obs}}\right) d\sigma \\ &= \int \frac{1}{u + s_{\text{obs}}} K_\varepsilon \left(u + s_{\text{obs}} - s_{\text{obs}}\right) du \end{aligned}$$

which is free of t. Therefore, the requirement (2.8) is satisfied in this case. Furthermore, the function $H_2(\hat{\sigma}_S^2, x) = 1 - \int_0^{\hat{\sigma}_S^2/x} g_2(w)dw$ is a confidence distribution for σ^2 since (1) given S, $H_2(\hat{\sigma}_S^2, x)$ is a distribution function on the parameter space $(0, \infty)$ and (2) given $x = \sigma_0^2$, $H_2(\hat{\sigma}_S^2, x) \sim U(0, 1)$. □

Notation and additional conditions

Let $N(x; \mu, \Sigma)$ be the normal density at x with mean μ and variance Σ, and $\widetilde{f}_n(s \mid \theta) = N\{s; s(\theta), A(\theta)/a_n^2\}$, the asymptotic distribution of the summary statistic. We define $a_{n,\varepsilon} = a_n$ if $\lim_{n\to\infty} a_n\varepsilon_n < \infty$ and $a_{n,\varepsilon} = \varepsilon_n^{-1}$ otherwise, and $c_\varepsilon = \lim_{n\to\infty} a_n\varepsilon_n$, both of which summarize how ε_n decreases relative to the converging rate, a_n, of S_n in Condition 6 below. Define the standardized random variables $W_n(S_n) = a_n A(\theta)^{-1/2}\{S_n - \eta(\theta)\}$ and $W_{\text{obs}} = a_n A(\theta)^{-1/2}\{s_{\text{obs}} - \eta(\theta)\}$ according to Condition 6 below. 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. Let $B_\delta = \{\theta \mid \|\theta - \theta_0\| \le \delta\}$ for $\delta > 0$. Define the initial density truncated in B_δ, i.e. $r_n(\theta)\mathbb{I}_{\theta\in B_\delta} / \int_{B_\delta} r_n(\theta)\, d\theta$, by $r_\delta(\theta)$. Let $t(\theta) = a_{n,\varepsilon}(\theta - \theta_0)$ and $v(s) = \varepsilon_n^{-1}(s - s_{\text{obs}})$. For any $A \in \mathcal{B}^p$ where $\mathcal{B}^p$ is the Borel sigma-field on $\mathbb{R}^p$, let $t(A)$ be the set $\{\phi : \phi = t(\theta) \text{ for some } \theta \in A\}$. For a non-negative function $h(x)$, integrable in $\mathbb{R}^l$, denote the normalized function $h(x) / \int_{\mathbb{R}^l} h(x)\, dx$ by $h(x)^{(\text{norm})}$. For a function $h(x)$, denote its

gradient by $D_x h(x)$, and for simplicity, omit θ from D_θ. For a sequence x_n, we use the notation $x_n = \Theta(a_n)$ to mean that there exist some constants m and M such that $0 < m <\mid x_n/a_n \mid < M < \infty$.

Condition 6 *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 $S_n \sim f_n(\cdot \mid \theta)$ and 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 $s_{\text{obs}} \to \eta(\theta_0)$ 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 $\|\eta(\theta) - \eta(\theta_0)\| > \delta'$ for all θ satisfying $\|\theta - \theta_0\| > \delta$; 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 $\theta = \theta_0$.

Condition 7 *The kernel satisfies (i) $\int vK_\varepsilon(v)dv = 0$; (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 $l \leq p + 6$; (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(ii)–(iii) where $K_\varepsilon(v)$ is replaced with $r_{max}(w)$, such that $\sup_{\theta\in B_\delta} \alpha_n \mid f_{W_n}(w \mid \theta) - \widetilde{f}_{W_n}(w \mid \theta) \mid \leq c_3 r_{max}(w)$ for some positive constant c_3.*

Condition 9 *The following statements hold: (i) $r_{max}(w)$ satisfies Condition 7(iv); and (ii) $\sup_{\theta\in B_\delta^C} \widetilde{f}_{W_n}(w \mid \theta) = O(e^{-c_2\|w\|^{\alpha_2}})$ as $\|w\| \to \infty$ for some positive constants c_2 and α_2, and $A(\theta)$ is bounded in $\mathcal{P}$.*

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

Proof of Theorem 2

Let $\tilde{Q}(\theta \in A \mid s) = \int_A r_\delta(\theta)\widetilde{f}_n(s \mid \theta)\, d\theta / \int_{\mathbb{R}^p} r_\delta(\theta)\widetilde{f}_n(s \mid \theta)\, d\theta$.

Lemma 3 *Assume Condition 2–8. If $\varepsilon_n = O(a_n^{-1})$, for any fixed $\nu \in \mathbb{R}^d$ and small enough δ,*

$$\sup_{A\in\mathfrak{B}^p}\left|\tilde{Q}\{a_n(\theta-\theta_0)\in A \mid s_{\text{obs}}+\varepsilon_n\nu\} - \int_A N[t;\beta_0\{A(\theta_0)^{1/2}W_{\text{obs}}+c_\varepsilon\nu\}, I(\theta_0)^{-1}]dt\right| \to 0,$$

in probability as $n\to\infty$, where $\beta_0 = I(\theta_0)^{-1}D\eta(\theta_0)^T A(\theta_0^{-1})$.

Proof of Lemma 3: With Lemma 1 from [Li & Fearnhead(2018a)], it is sufficient to show that

$$\sup_{A\in\mathfrak{B}^p} \mid \tilde{Q}\{t(\theta)\in A \mid s_{\text{obs}}+\varepsilon_n\nu\} - \widetilde{\Pi}\{t(\theta)\in A \mid s_{\text{obs}}+\varepsilon_n\nu\} \mid = o_P(1),$$

where $\widetilde{\Pi}$ denotes $\tilde{Q}$ using $r_n(\theta)$ rather than a prior $\pi(\theta)$ with a density satisfying Condition 2. With the transformation $t=t(\theta)$ and $v=v(s)$, the left hand side of the above equation can be written as

$$\sup_{A\in\mathfrak{B}^p} \mid \frac{\int_A r_\delta(\theta+a_n^{-1}t)\widetilde{f}_n(s_{\text{obs}}+\varepsilon_n\nu \mid \theta+a_n^{-1}t)dt}{\int_{\mathbb{R}^p} r_\delta(\theta+a_n^{-1}t)\widetilde{f}_n(s_{\text{obs}}+\varepsilon_n\nu \mid \theta+a_n^{-1}t)dt} - \frac{\int_A \pi(\theta+a_n^{-1}t)\widetilde{f}_n(s_{\text{obs}}+\varepsilon_n\nu \mid \theta+a_n^{-1}t)dt}{\int_{\mathbb{R}^p} \pi(\theta+a_n^{-1}t)\widetilde{f}_n(s_{\text{obs}}+\varepsilon_n\nu \mid \theta+a_n^{-1}t)dt} \mid . \tag{2.13}$$

For a function $\tau:\mathbb{R}^p\to\mathbb{R}$, define the following auxiliary functions,

$$\begin{aligned}
\phi_1\{\tau(\theta);n\} &= \frac{\int_{t(B_\delta)} |\tau(\theta+a_n^{-1}t)-\tau(\theta)|\widetilde{f}_n(s_{\text{obs}}+\varepsilon_n\nu \mid \theta+a_n^{-1}t)\,dt}{\int_{t(B_\delta)} \tau(\theta+a_n^{-1}t)\widetilde{f}_n(s_{\text{obs}}+\varepsilon_n\nu \mid \theta+a_n^{-1}t)\,dt},\\
\phi_2\{\tau(\theta);n\} &= \frac{\tau(\theta)\int_{t(B_\delta)} \widetilde{f}_n(s_{\text{obs}}+\varepsilon_n\nu \mid \theta+a_n^{-1}t)dt}{\int_{t(B_\delta)} \tau(\theta+a_n^{-1}t)\widetilde{f}_n(s_{\text{obs}}+\varepsilon_n\nu \mid \theta_0+a_n^{-1}t)dt}.
\end{aligned}$$

Then by adding and subtracting $\phi_2\{\tau_n^{-p}r_\delta(\theta);n\}\phi_2\{\pi(\theta);n\}$ in the absolute sign of (2.14), (2.14) can be bounded by

$$\phi_1\{\tau_n^{-p}r_\delta(\theta);n\} + \phi_1\{\pi(\theta);n\}\phi_2\{\tau_n^{-p}r_\delta(\theta);n\} + \phi_1\{\tau_n^{-p}r_\delta(\theta);n\}\phi_2\{\pi(\theta);n\} + \phi_1\{\pi(\theta);n\}.$$

Consider a class of function $\tau(\theta)$ satisfying the following conditions:

There exists a series $\{k_n\}$, such that $\sup_{\theta\in\mathcal{P}_0}\|k_n^{-1}D\tau(\theta)\|<\infty$ and $k_n=o(a_n)$; $\tau(\theta_0)>0$ and $\tau(\theta)\in C^1(B_\delta)$.

By Conditions 2–5, $\tau_n^{-p}r_\delta(\theta)$ and $\pi(\theta)$ belong to the above class. Then if $\phi_1\{\tau(\theta);n\}$ is $o_p(1)$ and $\phi_2\{\tau(\theta);n\}$ is $O_p(1)$, (2.14) is $o_p(1)$ and the lemma holds.

First, from (ii), there exists an open set $\omega\subset B_\delta$ such that $\inf_{\theta\in\omega}\tau(\theta)>c_1$, for a constant $c_1>0$. Then for $\phi_2\{\tau(\theta);n\}$, it is bounded by

$$\frac{\tau(\theta)}{c_1\int_{t(\omega)}\widetilde{f}_n(s_{\text{obs}}+\varepsilon_n\nu\mid\theta_0+a_n^{-1}t)^{(norm)}dt},$$

where $h(x)^{(norm)}$ represents the normalized version of $h(x)$. From equation (7) in the supplementary material of [Li & Fearnhead(2018b)], $\widetilde{f}_n(s_{\text{obs}}+\varepsilon_n\nu\mid\theta+a_n^{-1}t)$ can be written in the following form,

$$a_n^d\widetilde{f}_n(s_{\text{obs}}+\varepsilon_n\nu\mid\theta+a_n^{-1}t)=\frac{1}{\|B_n(t)\|^{1/2}}N[C_n(t)\{A_n(t)t-b_n\nu-c_2\};\theta,I_d],\quad(2.14)$$

where $A_n(t)$ is a series of $d\times p$ matrix functions, $\{B_n(t)\}$ and $\{C_n(t)\}$ are a series of $d\times d$ matrix functions, b_n converges to a non-negative constant and c_2 is a constant, and the minimum of absolute eigenvalues of $A_n(t)$ and eigenvalues of $B_n(t)$ and $C_n(t)$ are all bounded and away from 0. Then for fixed ν, by continuous mapping, (2.14) is away from zero with probability one. Therefore $\phi_2\{\tau(\theta);n\}=O_P(1)$.

Second, by Taylor expansion, $\tau(\theta+a_n^{-1}t)=\tau(\theta)+a_n^{-1}D\tau(\theta+e_tt)t$, where $\|e_t\|\le a_n^{-1}$. Then

$$\begin{aligned}\phi_1\{\tau(\theta);n\} &= \frac{k_n\phi_2\{\tau(\theta);n\}}{a_n\tau(\theta)}\frac{\int_{t(B_\delta)}|k_n^{-1}D\tau(\theta+e_tt)t|\widetilde{f}_n(s_{\text{obs}}+\varepsilon_n\nu\mid\theta+a_n^{-1}t)\,dt}{\int_{t(B_\delta)}\widetilde{f}_n(s_{\text{obs}}+\varepsilon_n\nu\mid\theta+a_n^{-1}t)\,dt}\\ &\le \frac{k_n\phi_2\{\tau(\theta);n\}}{a_n\tau(\theta)}\sup_{\theta\in B_\delta}\|k_n^{-1}D\tau(\theta)\|\frac{\int_{t(B_\delta)}\|t\|a_n^d\widetilde{f}_n(s_{\text{obs}}+\varepsilon_n\nu\mid\theta+a_n^{-1}t)dt}{\int_{t(B_\delta)}a_n^d\widetilde{f}_n(s_{\text{obs}}+\varepsilon_n\nu\mid\theta+a_n^{-1}t)\,dt},\quad(2.15)\end{aligned}$$

where the inequality holds by the triangle inequality. By the expression (2.14) and Lemma 7 in the supplementary material of [Li & Fearnhead(2018b)], the right hand side of (2.15) is $O_P(1)$. Then together with $\phi_2\{\tau(\theta);n\}=\Theta_P(1)$, $\phi_1\{\tau(\theta);n\}=o_P(1)$.

Therefore the Lemma holds. □

Define the joint density of (θ, s) in Algorithm 2 and its approximation, where the s-likelihood is replaced by its Gaussian limit and $r_n(\theta)$ by its truncation, by $q_\varepsilon(\theta, s)$ and $\tilde{q}_\varepsilon(\theta, s)$. It is easy to see that,

$$q_\varepsilon(\theta, s) = \frac{r_n(\theta) f_n(s|\theta) K_{\varepsilon_n}(s - s_{\text{obs}})}{\int_{\mathbb{R}^p \times \mathbb{R}^d} r_n(\theta) f_n(s|\theta) K_{\varepsilon_n}(s - s_{\text{obs}})\, d\theta ds},$$
$$\tilde{q}_\varepsilon(\theta, s) = \frac{r_\delta(\theta) \widetilde{f}_n(s|\theta) K_{\varepsilon_n}(s - s_{\text{obs}})}{\int_{\mathbb{R}^p \times \mathbb{R}^d} r_\delta(\theta) \widetilde{f}_n(s|\theta) K_{\varepsilon_n}(s - s_{\text{obs}})\, d\theta ds}.$$

Let $\tilde{Q}_\varepsilon(\theta \in A \mid s_{\text{obs}})$ be the approximate confidence distribution function, $\int_A \int_{\mathbb{R}^d} \tilde{q}_\varepsilon(\theta, s)\, ds d\theta$. With the transformation $t = t(\theta)$ and $v = v(s)$, let $\tilde{q}_{\varepsilon,t\nu}(t, v) = \tau_n^{-p} r_\delta(\theta + a_{n,\varepsilon}^{-1} t) \widetilde{f}_n(s_{\text{obs}} + \varepsilon_n \nu \mid \theta + a_{n,\varepsilon}^{-1} t) K_\varepsilon(\nu)$ be the transformed and unnormalized $\tilde{q}_\varepsilon(\theta, s)$, and $\tilde{q}_{A,tv}(h) = \int_A \int_{\mathbb{R}^d} h(t, v) \tilde{q}_{\varepsilon,t\nu}(t, v)\, dv dt$ for any function $h(\cdot, \cdot)$ in $\mathbb{R}^p \times \mathbb{R}^d$. Denote the factor of $\tilde{q}_{\varepsilon,t\nu}(t, v)$, $\tau_n^{-p} r_\delta(\theta + a_{n,\varepsilon}^{-1} t)$, by $\gamma_n(t)$. Let $\gamma = \lim_{n\to\infty} \tau_n^{-p} r_\delta(\theta)$ and $\gamma(t) = \lim_{n\to\infty} \tau_n^{-p} r_\delta(\theta + \tau_n^{-1} t)$, the limits of $\gamma_n(t)$ when $a_{n,\varepsilon} = a_n$ and $a_{n,\varepsilon} = \tau_n$ respectively. By Condition 3 and 4, $\gamma(t)$ exists and γ is non-zero with positive probability. Here several functions of t and v defined in [Li & Fearnhead(2018a), proofs for Section 3.1] and relate to the limit of $\tilde{q}_{\varepsilon,t\nu}(t, v)$ are used, including $g(v; A, B, c)$, $g_n(t, v)$, $G_n(v)$ and $g'_n(t, v)$. Furthermore several functions defined by integration as following are used: for any $A \in \mathfrak{B}^p$, let $g_{A,r}(h) = \int_{\mathbb{R}^d} \int_{t(A)} h(t, v) \gamma_n(t) g_n(t, v)\, dt dv$, $G_{n,r}(v) = \int_{t(B_\delta)} \gamma_n(t) g_n(t, v)\, dt$, $q_A(h) = \int_A \int_{\mathbb{R}^d} h(\theta, s) r_n(\theta) f_n(s \mid \theta) K_\varepsilon(s - s_{\text{obs}}) \varepsilon_n^{-d}\, ds d\theta$ and $\tilde{q}_A(h) = \int_A \int_{\mathbb{R}^d} h(\theta, s) r_\delta(\theta) \widetilde{f}_n(s \mid \theta) K_\varepsilon(s - s_{\text{obs}}) \varepsilon_n^{-d}\, ds d\theta$, which generalize those defined in [Li & Fearnhead(2018a), proofs for Section 3.1] for the case $r_n(\theta) = \pi(\theta)$.

Lemma 4 *Assume Condition 2–7. If $\varepsilon_n = o(a_n^{-1/2})$, then*

(i) $\int_{\mathbb{R}^d} \int_{t(B_\delta)} |\tilde{q}_{\varepsilon,t\nu}(t, \nu) - \gamma_n(t) g_n(t, \nu)|\, dt d\nu = o_p(1)$;

(ii) $g_{B_\delta,r}(1) = \Theta_P(1)$;

(iii) $\tilde{q}_{B_\delta,tv}(t^{k_1} v^{k_2}) / \tilde{q}_{B_\delta,tv}(1) = g_{B_\delta,r}(t^{k_1} v^{k_2}) / g_{B_\delta,r}(1) + O_P(a_{n,\varepsilon}^{-1}) + O_P(a_n^2 \varepsilon_n^4)$ *for k_1 and k_2*

(iv) $\tilde{q}_{B_\delta}(1) = \tau_n^p a_{n,\varepsilon}^{d-p} \left\{ \int_{t(B_\delta)} \int_{\mathbb{R}^d} \gamma_n(t) g_n(t, \nu) d\tau d\nu + O_P(a_{n,\varepsilon}^{-1}) + O_P(a_n^2 \varepsilon_n^4) \right\}$.

Proof of Lemma 4: These results generalize Lemma 2 in [Li & Fearnhead(2018a)] and Lemma 5 in [Li & Fearnhead(2018b)]. In Lemma 2 of [Li & Fearnhead(2018a)] where $\gamma_n(t) = \pi(\theta + a_{n,\varepsilon}^{-1}t)$, (i) holds by expanding $\tilde{q}_{\varepsilon,t\nu}(t,v)$ according to the proof of Lemma 5 of [Li & Fearnhead(2018b)]. Here the lines can be followed similarly by changing the terms involving $\pi(\theta)$ in equations (10) and (11) in the supplements of [Li & Fearnhead(2018b)]. Equation (10) is replaced by

$$\frac{\gamma_n(t)}{\mid A(\theta + a_{n,\varepsilon}^{-1}t) \mid^{1/2}} = \frac{\gamma_n(t)}{\mid A(\theta) \mid^{1/2}} + a_{n,\varepsilon}^{-1}\gamma_n(t)D\frac{1}{\mid A(\theta + e_t) \mid^{1/2}}t,$$

where $\|e_\tau\| \leq \delta$, and this leads to replacing $\pi(\theta)\int_{\tau(B_\delta)\times\mathbb{R}^d} g_n(t,\nu)dtd\nu$ in equation (11) by $\int_{\tau(B_\delta)\times\mathbb{R}^d} \gamma_n(t)g_n(t,\nu)dtd\nu$. These changes have no effect on the arguments therein since $\sup_{t\in t(B_\delta)} \gamma_n(t) = O_P(1)$ by Condition 3. Therefore (i) holds.

For (ii), By Condition 4 and Lemma 2 of [Li & Fearnhead(2018a)], there exists a $\delta' < \delta$ such that $\inf_{t\in t(B_{\delta'})} \gamma_n(t) = \Theta_p(1)$ and $\int_{\mathbb{R}^d}\int_{t(B_{\delta'})} g_n(t,\nu)\,dtdv = \Theta_p(1)$. Then since $g_{B_\delta,r}(1) \geq \inf_{t\in t(B_{\delta'})} \gamma_n(t)\int_{\mathbb{R}^d}\int_{t(B_{\delta'})} g_n(t,\nu)\,dtd\nu$, (ii) holds.

For (iii), $\tilde{q}_{B_\delta,tv}(t)/\tilde{q}_{B_\delta,tv}(1)$ can be expanded by following the arguments in the proof of Lemma 5 of [Li & Fearnhead(2018b)]. For $\tilde{q}_{B_\delta,tv}(t^{k_1}v^{k_2})/\tilde{q}_{B_\delta,tv}(1)$, it can be expanded similarly as in the proof of Lemma 4 of [Li & Fearnhead(2018a)].

For (iv), $\gamma_n(t)$ plays the same role as $\pi(\theta)$ in the proof of Lemma 5 in [Li & Fearnhead(2018b)], and the arguments therein can be followed exactly. The term τ_n^p is from the definition of $\gamma_n(t)$ that $r_n(\theta + a_{n,\varepsilon}^{-1}t) = \tau_n^p\gamma_n(t)$. □

Define the expectation of θ with distribution $\tilde{Q}_\varepsilon(\theta \in A \mid s_{\text{obs}})$ as $\widetilde{\theta}_\varepsilon$, and that of θ^*_{ACC} with density $\tilde{q}_\varepsilon(\theta, s)$ as $\widetilde{\theta}^*_\varepsilon$. Let $E_{G,r}(\cdot)$ be the expectation with the density $G_n(v)^{(\text{norm})}$, and $E_{G,r}\{h(v)\}$ can be written as $g_{B_\delta,r}\{h(v)\}/g_{B_\delta,r}(1)$. Let $\psi(\nu) = k_n^{-1}\beta_0\{A(\theta_0)^{1/2}W_{\text{obs}} + a_n\varepsilon_n\nu\}$, where $k_n = 1$, if $c_\varepsilon < \infty$, and $a_n\varepsilon_n$, if $c_\varepsilon = \infty$.

Lemma 5 *Assume Condition 2–5 and 7. Then if* $\varepsilon_n = o(a_n^{-1/2})$,

(i) $\widetilde{\theta}_\varepsilon = \theta_0 + a_n^{-1}\beta_0 A(\theta_0)^{1/2}W_{\text{obs}} + \varepsilon_n\beta_0 E_{G_n,r}(\nu) + r_1$, *where* $r_1 = o_P(a_n^{-1})$;

(ii) $\widetilde{\theta}^*_\varepsilon = \theta_0 + a_n^{-1}\beta_0 A(\theta_0)^{1/2}w_{\text{obs}} + \varepsilon_n(\beta_0 - \beta_\varepsilon)E_{G_n,r}(\nu) + r_2$, *where* $r_2 = o_P(a_n^{-1})$.

Proof of Lemma 5: These results generalize Lemma 3(c) and Lemma 5(c) in [Li & Fearnhead(2018a)]. With the transformation $t = t(\theta)$, by Lemma 2, if $\varepsilon_n = o(a_n^{-1/2})$,

$$\begin{cases} \widetilde{\theta}_\varepsilon = \theta_0 + a_{n,\varepsilon}^{-1}\tilde{q}_{B_\delta,t\nu}(t)/\tilde{q}_{B_\delta,t\nu}(1) = \theta_0 + a_{n,\varepsilon}^{-1} g_{B_\delta,r}(t)/g_{B_\delta,r}(1) + o_p(a_n^{-1}), \\ \widetilde{\theta}_\varepsilon^x = \theta_0 + a_{n,\varepsilon}^{-1}\tilde{q}_{B_\delta,t\nu}(t)/\tilde{q}_{B_\delta,t\nu}(1) - \varepsilon_n\beta_\varepsilon \tilde{q}_{B_\delta,t\nu}(\nu)/\tilde{q}_{B_\delta,t\nu}(1) \\ \quad\quad = \theta_0 + a_{n,\varepsilon}^{-1} g_{B_\delta,r}(t)/g_{B_\delta,r}(1) - \varepsilon_n\beta_\varepsilon E_{a_n,r}(\nu) + o_p(a_n^{-1}), \end{cases} \tag{2.16}$$

where the remainder term comes from the fact that $(a_{n,\varepsilon}^{-1} + \varepsilon_n)\left\{O_p(a_{n,\varepsilon}^{-1}) + O_p(a_n^2\varepsilon_n^4)\right\} = o_p(a_n^{-1})$.

First the leading term of $g_{B_\delta,r}(t\nu^k)$ is derived for $k = 0$ or 1. The case of $k = 1$ will be used later. Let $t' = t - \psi(\nu)$, then

$$\begin{aligned} g_{B_\delta,r}(t\nu^{k_2}) &= \int_{\mathbb{R}^d}\int_{t(B_\delta)} \{t' + \psi(\nu)\}\nu^{k_2}\gamma_n(t) g_n(t,\nu)\, dt d\nu \\ &= \int_{\mathbb{R}^d} \psi(\nu)\nu^{k_2} G_{n,r}(\nu)\, d\nu + \int_{\mathbb{R}^d}\int_{t(B_\delta)} t'\nu^{k_2}\gamma_n(t) g_n(t,\nu)\, dt d\nu. \end{aligned}$$

By matrix algebra, it is straightforward to show that $g_n(t,v) = N\{t; \psi(v), k_n^{-2} I(\theta_0)^{-1}\} G_n(v)$. Then with the transformation t', we have

$$\begin{aligned} & g_{B_\delta,r}(t\nu^{k_2}) - \int_{\mathbb{R}^d} \psi(\nu)\nu^{k_2} G_{n,r}(\nu)\, d\nu \\ = & \int_{\mathbb{R}^d}\int_{t(B_\delta)-\psi(\nu)} t'\nu^{k_2}\gamma_n\{\psi(\nu) + t'\} N\left\{t'; 0, k_n^{-2} I(\theta_0)^{-1}\right\} G_n(\nu)\, dt' d\nu. \end{aligned}$$

By applying the Taylor expansion on $\gamma_n\{\psi(\nu) + t'\}$, the right hand side of the above

equation is equal to

$$\begin{aligned}
&\int_{\mathbb{R}^d}\int_{t(B_\delta)-\psi(\nu)} t' N\{t';0,k_n^{-2}I(\theta_0)^{-1}\}\,dt'\cdot\gamma_n\{\psi(\nu)\}\nu^{k_2}G_n(\nu)\,d\nu \\
&+\int_{\mathbb{R}^d}\int_{t(B_\delta)-\psi(\nu)} t'^2 D_t\gamma_n\{\psi(\nu)+e_t\}N\{t';0,k_n^{-2}I(\theta_0)^{-1}\}\,dt'\cdot\nu^{k_2}G_n(\nu)d\nu \\
= \;& k_n^{-1}\int_{\mathbb{R}^d}\int_{Q_v} t'' N\{t'';0,I(\theta_0)^{-1}\}\,dt''\cdot\gamma_n\{\psi(\nu)\}\nu^{k_2}G_n(\nu)\,d\nu \\
&+k_n^{-2}\int_{\mathbb{R}^d}\int_{Q_v} t''^2 D_t\gamma_n\{\psi(\nu)+e_t\}N\{t'';0,I(\theta_0)^{-1}\}\,dt''\cdot\nu^{k_2}G_n(\nu)\,d\nu, \qquad (2.17)
\end{aligned}$$

where $Q_v=\{a_n(\theta-\theta_0)-k_n\psi(\nu)\mid\theta\in B_\delta\}$ and $t''=k_nt'$. Since Q_v can be written as $\left\{a_n(\theta-\theta_0-\varepsilon_n\nu)-\beta_0A(\theta_0)^{1/2}W_{\rm obs}\mid\theta\in B_\delta\right\}$, it converges to $\mathbb{R}^p$ for any fixed v with probability one. Then $\int_{Q_v}t''N\{t'';0,\tau(\theta_0)^{-1}\}\,dt''=o_P(1)$ for fixed v, and by the continuous mapping theorem and Condition 3, the first term in the right hand side of (2.17) is of the order $o_p(k_n^{-1})$. The second term is bounded by

$$k_n^{-2}\sup_{t\in\mathbb{R}}\|D_t\gamma_n(t)\|\int_{\mathbb{R}^p}\|t''\|^2N\{t'';0,I(\theta_0^{-1})\}\,dt''\int_{\mathbb{R}^d}\nu^{k_2}G_n(\nu)\,d\nu,$$

which is of the order $O_p(k^{-2}\tau_n/a_{n,\varepsilon})$ by Condition 5. Therefore

$$g_{B_\delta,r}(t\nu^{k_2})=\int_{\mathbb{R}^d}\psi(\nu)\nu^{k_2}G_n(\nu)d\nu+o_P(k_n^{-1}). \qquad (2.18)$$

By algebra, $k_n=a_{n,\tilde\varepsilon}^{-1}a_n$, and

$$\begin{aligned}
&\int_{\mathbb{R}^d}\psi(\nu)\nu^{k_2}G_n(\nu)d\nu \\
=\;& a_{n,\varepsilon}\beta_0\{a_n^{-1}A(\theta_0)^{1/2}W_{\rm obs}\int_{\mathbb{R}^d}\nu^{k_2}G_{n,r}(\nu)\,d\nu+\varepsilon_n\int_{\mathbb{R}^d}\nu^{k_2+1}G_{n,r}(\nu)\,d\nu\}. \qquad (2.19)
\end{aligned}$$

Then (i) and (ii) in the Lemma holds by plugging the expansion of $g_{B_\delta,r}(t)$ into (2.16). □

Lemma 6 *Assume Condition 2, 3, 6–9. Then as $n\to\infty$,*

(i) For any $\delta<\delta_0$, $r_{B_\delta^c}(1)$ and $\tilde q_{B_\delta^c}(1)$ are $o_p(\tau_n^p)$. More specifically, they are of the order $O_p\left(\tau_n^pe^{-a_{n,\varepsilon}^{\alpha_\delta}c_\delta}\right)$ for some positive constants c_δ and α_δ depending on δ.

(ii) $q_{B_\delta}(1) = \tilde{q}_{B_\delta}(1)\{1 + O_p(\alpha_n^{-1})\}$ *and* $\sup_{A \subset B_\delta} |q_A(1) - \tilde{q}_A(1)|/\tilde{q}_{B_\delta}(1) = O_p(\alpha_n^{-1})$*;*

(iii) if $\varepsilon_n = o(a_n^{-1/2})$*, then* $\tilde{q}_{B_\delta}(1)$ *and* $r_{B_\delta}(1)$ *are* $\Theta_P(\tau_n^p a_{n,\varepsilon}^{d-p})$*, and thus* $\tilde{q}_{\mathcal{P}_0}(1)$ *and* $q_{\mathcal{P}_0}(1)$ *are* $\Theta_P(\tau_n^p a_{n,\varepsilon}^{d-p})$*;*

(iv) if $\varepsilon_n = o(a_n^{-1/2})$*,* $\theta_\varepsilon = \widetilde{\theta}_\varepsilon + o_p(a_{n,\varepsilon}^{-1})$*. If* $\varepsilon_n = o(a_n^{-3/5})$*,* $\theta_\varepsilon = \widetilde{\theta}_\varepsilon + o_P(a_n^{-1})$*.*

Proof of Lemma 6: This generalizes Lemma 7 in [Li & Fearnhead(2018a)]. The arguments therein can be followed exactly, by Condition 3 and the fact that regarding $\pi(\theta)$, only the condition $\sup_{\theta \in \mathbb{R}^p} \pi(\theta) < \infty$ is used. □

Lemma 7 *Assume Condition 2, 3, 6–9.*

(i) For any $\delta < \delta_0$*,* $Q_\varepsilon(\theta \in B_\delta^c \mid s_{\text{obs}})$ *and* $\tilde{Q}_\varepsilon(\theta \in B_\delta^c \mid s_{\text{obs}})$ *are* $o_p(1)$*;*

(ii) There exists some $\delta < \delta_0$ *such that*

$$\sup_{A \in \mathfrak{B}^p} |Q_\varepsilon(\theta \in A \cap B_\delta \mid s_{\text{obs}}) - \tilde{Q}_\varepsilon(\theta \in A \cap B_\delta \mid s_{\text{obs}})| = o_p(1);$$

(iii) $a_{n,\varepsilon}(\theta_\varepsilon - \widetilde{\theta}_\varepsilon) = o_p(1)$.

Proof of Lemma 7: This lemma generalizes Lemma 3 of [Li & Fearnhead(2018a)]. The proof of Lemma 3 in [Li & Fearnhead(2018a)] only needs Lemma 3 and 5 from [Li & Fearnhead(2018b)] to hold. The result that $q_{B_\delta^c}\{h(\theta)\} = O_p(\tau_n^p e^{-a_{n,\varepsilon}^{\alpha_\delta} c_\delta})$ for some positive constants α_δ and c_δ, which generalizes the case of $r_n(\theta) = \pi(\theta)$ in Lemma 3 of [Li & Fearnhead(2018b)], holds by Condition 3, since the latter only uses the fact that $\sup_{\theta \in B_\delta^c} \pi(\theta) < \infty$. Then the arguments in the proof of Lemma 3 in [Li & Fearnhead(2018b)] can be followed exactly, despite the term τ_n^p that is not included in the order of $\pi_{B_\delta^c}\{h(\theta)\}$, since $Q_\varepsilon(\theta \in A \mid s_{\text{obs}})$ is the ratio $q_A(1)/q_{\mathbb{R}^p}(1)$. Since Lemma 5 in [Li & Fearnhead(2018b)] has been generalized by Lemma (4) above, the arguments of the proof of Lemma 3 in [Li & Fearnhead(2018a)] can be followed exactly. □

This result generalizes the case (i) of Proposition 1 in [Li & Fearnhead(2018a)]. With the above lemmas, lines for proving case (i) of Proposition 1 in [Li & Fearnhead(2018a)] can be followed exactly to finish the proof of Theorem 2. □

Proof of Theorem 3

Lemma 8 *Assume Condition 2–10. If* $\varepsilon_n = o_p(a_n^{-3/5})$*, then* $a_n\varepsilon_n(\beta_\varepsilon - \beta_0) = o(1)$.

Proof of Lemma 8: This generalizes Lemma 4 in [Li & Fearnhead(2018a)] by replacing $\pi(\theta_0 + a_{n,\varepsilon}^{-1}t)$ therein with $\gamma_n(t)$. By Condition 3 and the arguments in the proof of Lemma 4 in [Li & Fearnhead(2018a)], it can be shown that

$$\frac{q_{\mathbb{R}^p}\{(\theta-\theta_0)^{k_1}(s-s_{\text{obs}})^{k_2}\}}{q_{\mathbb{R}^p}(1)} = a_{n,\varepsilon}^{-k_1}\varepsilon_n^{-k_2}\left\{\frac{\tilde{q}_{B_\delta,tv}(t^{k_1}\nu^{k_2})}{\tilde{q}_{B_\delta,tv}(1)} + O_p(\alpha_n^{-1})\right\}.$$

Then by Lemma 2 (iii), the right hand side of the above is equal to

$$a_{n,\varepsilon}^{-k_1}\varepsilon_n^{-k_2}\left\{\frac{g_{B_\delta,r}(t^{k_1}\nu^{k_2})}{g_{B_\delta,r}(1)} + O_p(a_{n,\varepsilon}^{-1}) + O_p(a_n^2\varepsilon_n^4) + O_p(\alpha_n^{-1})\right\}.$$

Since $\beta_\varepsilon = \text{Cov}_\varepsilon(\theta, S_n)\text{Var}_\varepsilon(S_n)^{-1}$,

$$\begin{aligned} a_n\varepsilon_n(\beta_\varepsilon - \beta_0) =& k_n\left[\frac{g_{B_\delta,r}(t\nu)}{g_{B_\delta,r}(1)} - \frac{g_{B_\delta,r}(t)g_{B_\delta,r}(\nu)}{g_{B_\delta,r}(1)^2} + o_p(k_n^{-1})\right]\cdot \\ & \left[\frac{g_{B_\delta,r}(\nu\nu^T)}{g_{B_\delta,r}(1)} - \frac{g_{B_\delta,r}(\nu)g_{B_\delta,r}(\nu)^T}{g_{B_\delta,r}(1)^2} + o_p(k_n^{-1})\right] - a_n\varepsilon_n\beta_0, \end{aligned}$$

where the equations that $a_{n,\varepsilon}^{-1}k_n = o(1)$, $a_n^2\varepsilon_n^4 k_n = o(p)$, and $\alpha_n^{-1}k_n = o(a_n^{-2/5}k_n) = o(1)$ are used. By algebra, the right hand side of the equation above can be rewritten as

$$\begin{aligned} &\left\{\frac{g_{B_\delta,r}\{(k_nt - a_n\varepsilon_n\beta_0\nu)\nu\}}{g_{B_\delta,r}(1)} - \frac{g_{B_\delta,r}(k_nt - a_n\varepsilon_n\beta_0\nu)g_{B_\delta,r}(\nu)}{g_{B_\delta,r}(1)^2} + o_p(1)\right\}\cdot \\ &\quad\left\{E_{G,r}(\nu\nu^T) - E_{G,r}(\nu)E_{G,r}(\nu)^T + o_p(k_n^{-1})\right\}^{-1}. \end{aligned}$$

By plugging (2.18) and (2.19) in the above, $a_n\varepsilon_n(\beta_\varepsilon - \beta_0)$ is equal to

$$\begin{aligned} &\left\{E_{G,r}(\nu)\beta_0 A(\theta_0)^{1/2}W_{\text{obs}} - E_{G,r}(\nu)\beta_0 A(\theta_0)^{1/2}W_{\text{obs}} + o_p(1)\right\}\cdot\{\text{Var}_{G,r}(\nu) + o_p(k_n^{-1})\}^{-1} \\ &\quad = o_P(1)\{\text{Var}_{G,r}(\nu) + o_p(k_n^{-1})\}^{-1}. \end{aligned}$$

Since

$$\mathrm{Var}_{G,r}(\nu) \geq \frac{\inf_{t\in t(B_{\delta'})}\gamma_n(t)}{g_{B_\delta,r}(1)}\int_{\mathbb{R}^d}\int_{t(B_{\delta'})}\{\nu - E_{G,r}(\nu)\}^2 g_n(t,\nu)\,dtd\nu,$$

where δ' is defined in the proof of Lemma 4(ii), we have $\mathrm{Var}_{G,r}(\nu)^{-1} = \Theta_p(1)$. Therefore $a_n\varepsilon_n(\beta_\varepsilon - \beta_0) = o_p(1)$. □

Lemma 9 *Results generalizing Lemma 5 in [Li & Fearnhead(2018a)], i.e. replacing* Π_ε *and*
$Pitil_\varepsilon$ *therein with* Q_ε *and* $\tilde{Q}_\varepsilon$*, hold.*

Proof of Lemma 9: In [Li & Fearnhead(2018a)], the proof of Lemma 5 requires Lemma 4 and 7 in [Li & Fearnhead(2018a)] to hold. Since their generalized results have been proved, the proof of this lemma follows the same arguments. □

Lemma 10 *Results generalizing Lemma 10 in [Li & Fearnhead(2018a)] hold.*

Proof of Lemma 10: The same arguments can be followed. □

With all above lemmas, the proof of Theorem 3 holds by following the same arguments in the proof of Theorem 1 in [Li & Fearnhead(2018a)]. □

Chapter 3

Exact inference for 2×2 contingency tables of rare events

3.1 Introduction

Contingency tables are a useful way to depict categorical data as a matrix of discrete values. In particular, 2×2 tables represent binary categorical data with a broad range of applications to clinical research and more. For various sampling schemes, i.e. model assumptions, establishing exact inference for model parameters is an interesting and challenging area of active research, especially when the outcome of interest is rare (e.f. e.g. [Radavicius & Zidanaviciute (2018), Kroonenberg (2018), Li & Fu (2018)]). One major application of these methods is to help establish drug safety (as opposed to drug efficacy) in clinical trials.

The most common sampling scheme assumes only one of the marginal totals is fixed as in Table 3.1. This type of design is applicable to randomized clinical trials and cohort studies. In the work presented here, we address the problem of conducting exact inference on the odds ratio from a 2×2 table of rare events with one fixed margin. That is, we are interested in quantifying uncertainty about the odds ratio without relying on any assumptions about the sample sizes, and, we allow small or zero entries in one or more cells of Table 3.1.

Table 3.1: 2×2 *contingency table with binomial sampling*

	Non-Events	Events	
Non-exposure	X	$n_x - X$	n_x
Exposure	Y	$n_y - Y$	n_y

In other words, we are concerned with inference on the model parameters of the system

of independent random variables

$$\begin{cases} X \sim Binomial(n_x, p_x) \\ Y \sim Binomial(n_y, p_y) \end{cases}$$

where n_x and n_y are known and both p_x and p_y are small, but non-zero.

Here, we propose an inferential method to quantify our uncertainty of the log odds ratio, $\theta = \log[p_x/(1-p_x)] - \log[p_y/(1-p_y)]$ without relying on any large sample approximations, in contrast to most other existing methods. We demonstrate that using our method, we can achieve tighter confidence intervals for θ in comparison to the standard exact, conditional approach and to a Bayesian approach with noninformative priors. Our method is entirely frequentist and its performance is evaluated with respect to the Repeated Sampling Principle.

We call the computational method developed here a *repro sampling method* because it works by "reproducing" the data. Our algorithmic approach mimics the sample variability with a grid search across different possible parameter values. The key assumption for our method is that there exists a known data-generating equation. That is, we assume the observed data can be generated by some function of the parameters and of a random variable U, where the distribution of U is known. Mathematically, this means we assume $x_{obs} = T(\theta_0, u)$, where u is a particular, unobserved instance of U, θ_0 represents the true unknown value of the parameter, and x_{obs} is our observed data. The function T is a modified statistic between known random variables and the unknown sampling distribution. The repro sampling method we introduce incorporates a positive, data-driven tuning parameter, λ, that helps to stabilize our inferential results especially in the case where the signal of θ is difficult to detect because the true (p_x, p_y) values are quite small.

3.2 Repro sampling for sparse 2×2 tables

3.2.1 Sampling method setting

The repro sampling method we develop here, provides information about parameter uncertainty by mimicking the sampling mechanism that generated the observed data. Under the model specifications described above, we can rely on a parameter dependent modified statistic, $T\left[(X,Y) \mid (p_x, p_y)\right]$, and the following data-generating equations

$$\begin{cases} x_{obs} = \sum_{i=1}^{n_x} \mathbb{I}\{u_j \leq p_x\} \\ y_{obs} = \sum_{j=1}^{n_y} \mathbb{I}\{v_j \leq p_y\}. \end{cases}$$

Here each of the u_i and v_j are some (unobserved but fixed) values of independent $Unif(0,1)$ random variables. We are particularly interested in inference for θ when either one or both of the true (p_x, p_y) values are close to zero. In this setting, it is likely that we may have zero counts in one or more cell of Table 3.1 and so the log likelihood may not be well defined. One standard method to deal with this is to consider $(\max\{1/2, x_{obs}\}, \max\{1/2, y_{obs}\})$ rather than the pair (x_{obs}, y_{obs}). This adjustment is called a continuity correction.([Plackett (1964), Cox (1970)]) In the following, we adopt this continuity correction for all methods being considered, though a goal of our future work is to eliminate the need for this adjustment.

The standard continuity corrected frequentist estimator for θ is the statistics

$$T_{sd}(X,Y) = \log\left[\left(\frac{\max\{1/2, X\}}{n_x - \max\{1/2, X\}}\right)\right] - \log\left[\left(\frac{\max\{1/2, Y\}}{n_y - \max\{1/2, Y\}}\right)\right].$$

In the repro sampling method, we propose using a modified version of T_{sd} that incorporates a positive tuning parameter, λ,

$$\begin{aligned} T\left[(X,Y) \mid (p_x, p_y)\right] &= \log\left[\left(\frac{\max\{1/2, X\}}{n_x - \max\{1/2, X\}} + \lambda\frac{p_x}{1-p_x}\right)\right] \\ &\quad - \log\left[\left(\frac{\max\{1/2, Y\}}{n_y - \max\{1/2, Y\}} + \lambda\frac{p_y}{1-p_y}\right)\right]. \end{aligned}$$

Of course, this estimator is not a true statistic as it can only be calculated given particular values of (p_x, p_y). The repro sampling method however, exploits this fact by inputting points (p_x, p_y) along a predefined grid across $(0,1) \times (0,1)$.

Because we are primarily focused on inference for θ, it is helpful to re-parameterize T in terms of θ. To do this, we introduce a nuisance parameter that is orthogonal to the parameter of interest,

$$\psi = \log\left[p_x/(1-p_x)\right] + \log\left[p_y/(1-p_y)\right].$$

Now, for each (p_x, p_y) pair, there is a corresponding (θ, ψ) pair yielding

$$2\log\left(\frac{p_x}{1-p_x}\right) = \psi + \theta \quad \text{and} \quad 2\log\left(\frac{p_y}{1-p_y}\right) = \psi - \theta.$$

Thus we can write

$$\begin{aligned} T\left[(X,Y) \mid (\theta,\psi)\right] &= \log\left[\frac{\max\{1/2, X\}}{n_x - \max\{1/2, X\}} + \lambda e^{\frac{1}{2}(\psi+\theta)}\right] \\ &\quad - \log\left[\frac{\max\{1/2, Y\}}{n_y - \max\{1/2, Y\}} + \lambda e^{\frac{1}{2}(\psi-\theta)}\right]. \end{aligned} \tag{3.1}$$

Note that the value of T corresponding to the data, (x_{obs}, y_{obs}), is never actually observed because it is a function of unknown parameter values (θ_0, ψ_0). Given any value of (θ, ψ) (and choice of λ) however, we can generate T from independent $Unif(0,1)$ samples through the data generating equations which we rewrite now as

$$\begin{cases} x_{obs} = \sum_{j=1}^{n_x} \mathbb{I}\left\{u_j \leq \frac{e^{\frac{1}{2}(\psi+\theta)}}{1+e^{\frac{1}{2}(\psi+\theta)}}\right\} \\ y_{obs} = \sum_{j=1}^{n_y} \mathbb{I}\left\{v_j \leq \frac{e^{\frac{1}{2}(\psi-\theta)}}{1+e^{\frac{1}{2}(\psi-\theta)}}\right\}. \end{cases} \tag{3.2}$$

3.2.2 Repro sampling algorithm

For a set of variables $W_1, \ldots, W_m$, let $W_{(\alpha)}$ represent the α^{th} lower empirical quantile of the set. With this notation in mind, we now present the repro sampling algorithm for inference on θ.

Algorithm 5 *(Repro sampling method)*

Given some (θ, ψ)

1. *Compute* $T_{obs} = T\left[(\max\{1/2, x_{obs}\}, \max\{1/2, y_{obs}\}) \mid (\theta, \psi)\right]$*;*
2. *Simulate* N *copies of independent uniform random vectors* $\boldsymbol{u} = (u_1, \ldots, u_{n_x})$ *and* $\boldsymbol{v} = (v_1, \ldots, v_{n_y})$*;*
3. *For each* $i = 1, \ldots, N$

   3.1 Compute $x^{(i)} = \max\left\{1/2, \sum_{j=1}^{n_x} \mathbb{I}\{u_{ij} \leq \frac{e^{\frac{1}{2}(\psi+\theta)}}{1+e^{\frac{1}{2}(\psi+\theta)}}\}\right\}$,
   $y^{(i)} = \max\left\{1/2, \sum_{j=1}^{n_y} \mathbb{I}\{v_{ij} \leq \frac{e^{\frac{1}{2}(\psi-\theta)}}{1+e^{\frac{1}{2}(\psi-\theta)}}\}\right\}$,
   $T^{(i)} = T\left[(x^{(i)}, y^{(i)}) \mid (\theta, \psi)\right]$;

   3.2 Using the empirical distribution of the N *values of* T*, create the set* $S_\alpha = [T_{(\alpha/2)}, T_{(1-\alpha/2)}]$*; and retain* θ_i *if* $T_{obs} \in S_\alpha(T_i)$*;*

Repeat Steps 1–3 for different values of (θ, ψ) *along a predetermined grid.*

To understand Algorithm 5, let us first consider using the conventional statistic T_{sd}. A typical Monte Carlo approach for forming exact confidence sets for θ would be to use the tail method and the statistic

$$T_{sd}\left[(x_{obs}, y_{obs})\right] = \log\left[\frac{\max\{1/2, x_{obs}\}}{n_x - \max\{1/2, x_{obs}\}}\right] - \log\left[\frac{\max\{1/2, y_{obs}\}}{n_y - \max\{1/2, y_{obs}\}}\right],$$

to computationally solve for upper and lower confidence bounds.([Cornfield(1956)])

where (θ, ψ_{typ}) are subject to equations (3.2) and the nuisance parameter is $\psi_{typ} = \log\left[p_y/(1-p_y)\right]$. This Monte Carlo approach is exactly what Algorithm 5 does if we set $\lambda = 0$ and use ψ_{typ} as the nuisance parameter (rather than ψ). So using the tail method to find an exact confidence interval for θ is the same as using Algorithm 5 to produce confidence sets for θ.

For a positive λ however, we lend more weight to "pretend" values of the parameters along the grid. We can view λ as a stabilization parameter that allows us to consider the effect of small (p_x, p_y) values on the odds ratio, even though these parameter values likely result in zero observations. Also, note that taking $\lambda = 1$ will have no effect on our inference for the odds ratio. Later, we discuss how to choose λ based on the data.

Letting α be arbitrary, the output of Algorithm 5 are observations from a distribution estimator for θ; this special type of estimator is called a confidence distribution. We now prove this fact by showing that the output of Algorithm 5 can be used to form α-level confidence sets for θ. To see this, first define the set $S_\alpha(\theta, \psi)$ such that

$$pr\left[T\left[(X,Y) \mid (\theta,\psi)\right] \in S_\alpha(\theta,\psi)\right] \geq 1-\alpha. \tag{3.3}$$

Here, the probability measure is with respect to the empirical distribution of the N simulated $T^{(i)}$ values calculated in Step 3 of Algorithm 5 and $X = \max\{x, 1/2\}, Y = \max\{y, 1/2\}$. In Algorithm 5, we define $S_\alpha(\theta) = S_\alpha(\theta,\psi) = [T_{(\alpha/2)}, T_{(1-\alpha/2)}]$ for each of the N Monte Carlo samples of T. The output of Algorithm 5 will be (say) $m \leq N$ retained values of θ and we can define

$$\Gamma_\alpha(X,Y) = \left[\theta_{(\alpha/2)}, \theta_{(1-\alpha/2)}\right] \tag{3.4}$$

as a $100(1-\alpha)\%$ confidence interval for θ since

$$pr\left[\theta_0 \in \Gamma_\alpha(X,Y)\right] \geq pr\left[T\left[(X,Y) \mid (\theta_0,\psi_0)\right] \in S_\alpha(\theta_0,\psi_0)\right] \geq 1-\alpha,$$

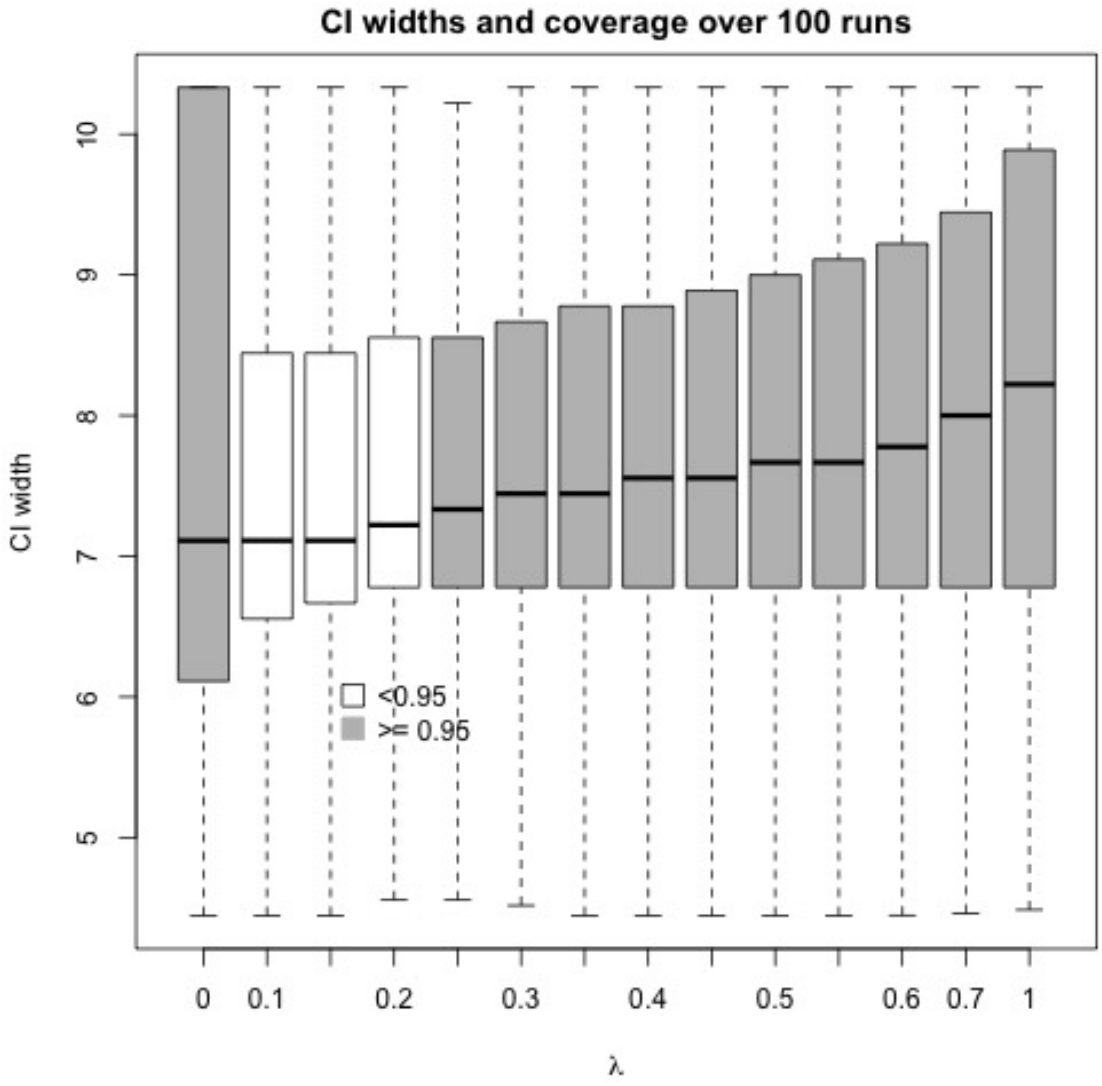
with $X = \max\{x, 1/2\}$ and $Y = \max\{y, 1/2\}$.

3.2.3 Choosing λ

To determine how Algorithm 5 performs with respect to the Repeated Sampling Principle, we explored the behavior of our modified statistic, T, for different λ choices $\lambda = 0$, $0 < \lambda < 1$, and $\lambda \geq 1$. In Figure 3.1 for example, we see that the confidence intervals for θ are potentially much smaller with $\lambda > 0$ while still achieving the nominal coverage level.

Although the coverage of confidence intervals for θ based on Algorithm 5 is at least at the nominal $(1-\alpha)100\%$ level, better coverage (closer to the nominal level) is obtained with smaller $\lambda > 0$ values. In practice, there is no way to test the actual coverage of the resulting confidence intervals and so an empirical rule for choosing a suitable λ_n,

Figure 3.1: Widths of 95% confidence interval for various λ choices in Algorithm 5 with $n_x = n_y = 100$ and the unknown truth $(p_x, p_y) = (0.01, 0.01)$. This picture illustrates the potential improvement by choosing some $\lambda > 0$.



dependent on the observed data, is desirable. Our simulations suggest Algorithm 6 as a data-driven method for choosing a small λ value.

Algorithm 6 *(Empirical choice of λ_n)*

For some $m \in \mathbb{Z}^+$, consider a sequence of λ values,

$0 \leq \lambda_1 < \lambda_2 < \cdots < \lambda_m \leq 1$.

1. *Set $i = 1$;*
2. *Run Algorithm 5 with $\lambda = \lambda_i$;*
   - *2.1 Compute the length of the interval $Q^{(i)} = \theta^{(i)}_{(1-\alpha/2)} - \theta^{(i)}_{(\alpha/2)}$;*
   - *2.2 Set $i = i + 1$;*
3. *Repeat Step 2 until $Q^{(i)} > Q^{(i-1)}$.*

Simulations suggest that for a finer sequence of initial $\lambda_1, \lambda_2, \ldots, \lambda_m$ values, the resulting confidence intervals are less likely to be overly-conservative. Also, through our simulations, we discovered that the confidence intervals for any $\lambda < 1$ will be smaller than the confidence intervals for $\lambda = 0$ when the true values of p_x and p_y are small.

3.3 Comparison to other methods

We compare the confidence intervals from the standard exact method (i.e. Algorithm 5 with $\lambda = 0$) to those with an empirically chosen $\lambda_n > 0$ (as specified in Algorithm 6). In these simulation studies, we see that by incorporating a $\lambda_n > 0$ term, the coverage of the resulting confidence intervals for θ can be closer to the nominal level and are generally more narrow.

We also compare the repro sampling method confidence intervals to Bayesian credible intervals using Jeffrey's prior and to the exact frequentist confidence intervals formed by inverting the score statistic as suggested in [Agresti(2003)] and finally to the intervals resulting from inverting Fisher's exact test. These first two methods were implemented using the built-in functions of the PropCIs R package while adjusting for the continuity correction mentioned earlier.([Scherer (2018)]) For Fisher's exact test, we did not apply a continuity correction. As indicated in Table 3.2, we consider the performance of each

Table 3.2: *Coverage and confidence interval width for four methods of estimating the log likelihood in different settings. The method with λ_n corresponds to setting λ based on Algorithm 6 and $\lambda = 0$ corresponds to the standard tail method for constructing confidence intervals based on T_{sd}. The continuity correction was used in all of the methods except for Fisher's exact test. The nominal confidence level is* 0.95*.*

Scenario	**Method**	**Coverage**	**Width**	
			Median	**Std. dev.**
1	λ_n	0.96	7.86	1.48
$n_x = n_y = 100$	$\lambda = 0$	0.92	7.13	1.96
$(p_x, p_y) = (0.01, 0.01)$	Score	1.00	9.65	17.28
	Bayes	1.00	15.58	38.03
	Fisher	1.00	38.85	22.75*
2	λ_n	0.96	6.28	1.34
$n_x = 250$, $n_y = 100$	$\lambda = 0$	0.95	7.85	1.35
$(p_x, p_y) = (0.01, 0.01)$	Score	0.96	8.33	12.94
	Bayes	0.96	13.65	28.07
	Fisher	0.97	31.23	16.31*
3	λ_n	0.99	7.93	1.58
$n_x = n_y = 100$	$\lambda = 0$	0.98	10.32	2.09
$(p_x, p_y) = (0.02, 0.01)$	Score	1.00	16.46	18.89
	Bayes	1.00	30.05	41.59
	Fisher	1.00	69.26	32.40*
4	λ_n	0.96	7.18	1.74
$n_x = 250$, $n_y = 100$	$\lambda = 0$	0.98	6.79	1.73
$(p_x, p_y) = (0.02, 0.01)$	Score	0.99	14.05	21.38
	Bayes	0.99	23.09	46.23
	Fisher	0.99	51.32	21.99*

* The standard deviation was only calculated for finite confidence intervals.

method for four different situations that consider whether or not the true parameter values are equal and whether or not the sample sizes are equal.

Although for Bayesian inference, we do not actually need to adjust our observations for discontinuities at zero counts, the performance of these credible intervals can vary wildly depending on the choice of prior. In Table 3.2, we also consider Fisher's exact test which does not use a continuity correction. If one or more observations are zero, the Fisher's exact method will produce an interval with an infinite bound. In order to more clearly compare all methods, we disregard any infinite confidence bounds when

computing the standard deviation of the widths of the Fisher confidence intervals.

We see that the width of the confidence intervals for using $\lambda_n > 0$ can be the same as (Scenarios 2 and 4) or much smaller than (Scenarios 1 and 3) the confidence intervals using $\lambda = 0$. The performance of Algorithm 5, for either choice of λ, is superior to the other methods considered. Also we note that the Bayesian intervals could perhaps be improved upon with a more informative prior.

3.4 Real data application

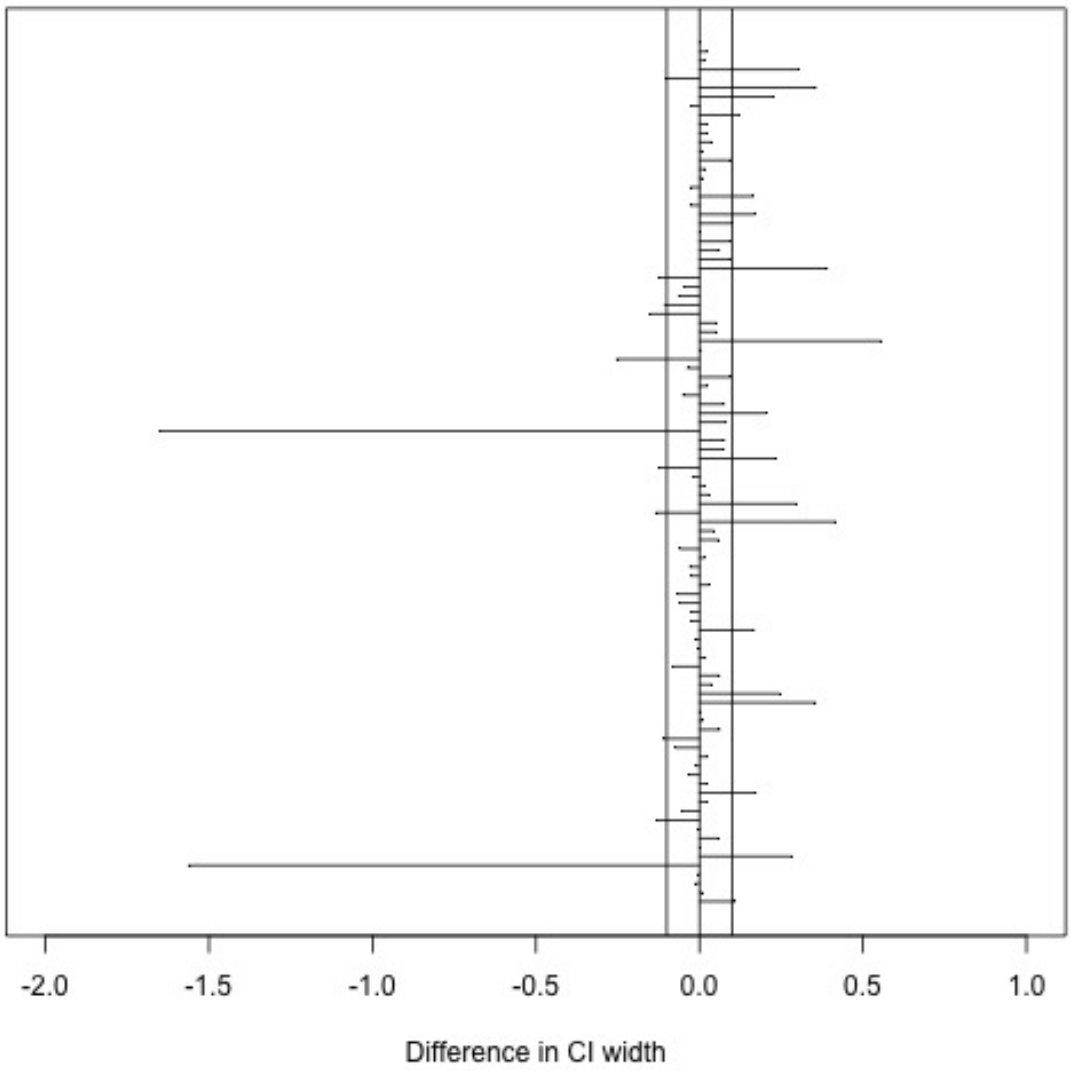
[Nissen & Wolski(2008)] collect data for a meta-analysis to examine whether the diabetes drug Avandia is associated with myocardial infarction or cardiovascular death. The datasets in [Nissen & Wolski(2008)] represent 96 individual clinical trials of moderate to large sizes with low adverse event rates. Thus many of the individual trials consist of zero counts.

To further investigate the performance of Algorithm 5 with an empirical choice of λ_n as in Algorithm 6 versus setting $\lambda = 0$ (i.e. the tail method), Figure 3.2 compares the resulting confidence intervals for θ for each study in [Nissen & Wolski(2008)]. Negative values indicate that the repro sampling method with $\lambda_n > 0$ produced a smaller confidence interval and positive values indicate that the tail method (repro with $\lambda = 0$) produced a smaller confidence interval. In Figure 3.2, we mostly see that the intervals for θ from Algorithm 6 match the confidence intervals resulting from the standard method where $\lambda = 0$ since most of the difference in widths lie within 0.1 units of the origin. However, there are several instances where we see that Algorithm 6 yields much smaller intervals and in some cases this leads to a different conclusion regarding θ.

3.5 Discussion

The repro sampling method we have introduced in Algorithm 5 can produce better performing confidence intervals than other exact methods, both in terms of matching the nominal coverage level and in terms of interval width, especially in the case where the true (p_x, p_y) values are suspected to be small. Furthermore, we present an empirical

Figure 3.2: *Plot of the difference in* 95% *CI widths from the repro sampling method using* $\lambda_n > 0$ *and* $\lambda = 0$ *for* 96 *different clinical trials studying the relationship between rare and adverse events and the drug Avandia. Larger negative values indicate smaller CIs for* $\lambda_n > 0$*. Vertical lines are plotted at* -0.1*,* 0*, and* 0.1 *for scale.*



method for choosing $\lambda_n > 0$ that can yield smaller intervals than we would achieve without considering this additional stabilization parameter. However, the case where both observed counts are zero is still problematic for Algorithm 5, as it is for all current methods. Additionally, our simulation studies indicate that if the true parameter values are such that $p_x << p_y$, then using Algorithm 6 to choose λ_n performs worse than setting $\lambda = 0$.

The method we present is a computational inference method specifically for inference on a log odds ratio that is difficult to detect. In further applications, our method could be applied to sequential online testing problems, especially when one is unable to rely on large-sample asymptotics.

Chapter 4

Concluding remarks and directions for future research

This dissertation explores frequentist solutions to two different computational inference problems. Namely, we establish ways to conduct valid statistical inference, with respect to the Repeated Sampling Procedure, in the case where (1.) we are interested in model parameters but have no tractable likelihood function, given the observed data and (2.) we are interested in a hard to detect signal from the odds ratio of a 2×2 contingency table. Though these problems may seem disparate on the surface, the work in this dissertation develops a similar computational inference solution for each situation. In both circumstances, the key to validating the resulting inference is to develop a confidence distribution as an estimator for the parameters of interest.

There are many possible directions for future research with both of the problems addressed in this thesis.

1. For the problem of inference when there is no tractable likelihood function, we establish conditions under which the approximate confidence distribution computing method will produce a confidence distribution. These conditions do not depend on the sufficiency of the summary statistic, unlike other existing work in approximate Bayesian computing. We do however require other, lighter conditions on the summary statistic such as asymptotic normality (Theorem 2 and 3) or an approximate pivotal structure (Theorem 1). For future research, it would be useful to look for other, weaker conditions on the summary statistic that will establish the results of Algorithm 2 is a confidence distribution estimator. Since an approximately pivotal summary statistic is potentially much easier to determine than approximate sufficiency in the likelihood-free setting, it would be worthwhile to consider other pivotal structures beyond the location and scale cases in Theorem 1. There is also

plenty of work left to be done regarding the computational efficiency of Algorithm 2. This is an active area of research in approximate Bayesian computing methods however this work does not generally consider the frequentist performance of the resulting approximate posterior distribution. Since the perspective of approximate confidence distribution computing is fundamentally frequentist, work that explores the frequentist coverage property of the resulting distribution estimators from these modified approximate Bayesian computing methods (e.g. IS-ABC, MCMC-ABC, ABC-PMC, ABC-SMC, etc.) is largely unexplored. ([Marin et al.(2012)])

2. For the problem of detecting a signal from a sparse 2×2 contingency table, our work explores inference for the odds ratio under the assumption that one marginal total is fixed. Naturally, future work could explore the computational inferential results under other model assumptions. Additionally, it may be useful to explore other modified statistics and/or other ways to empirically choose λ besides Algorithm 6. Since our work on this problem utilizes a continuity correction, it is not directly comparable to other methods which do not (e.g. Fisher's exact test). It would be useful to explore whether or not we can establish valid frequentist inference, without a continuity correction. Also, the scope of this work is limited to apply in instances which we believe the true (p_x, p_y) values are small but non-zero. In future work, this framework could be relaxed.

Although the work of this dissertation is limited to these two particular problems, the scope of future research questions is much larger. We suggest that confidence distributions, and the flexible frequentist perspective under which they are developed, may prove useful in solving other problems of computational inference.

References

Agresti, A. (2003). Dealing with discreteness: making exact confidence intervals for proportions, difference of proportions, and odds ratios more exact. *Statistical Methods in Medical Research* **12**, 3–21.

Barnard, G. (1963). *Journal of the Royal Statistical Society* B, **25**, 294. (In discussion.)

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

Beaumont, M., Cornuet, J-M., Marin, J-M., & Robert, C. (2009). Adaptive approximate Bayesian computation. *Biometrika* **20**, 1–9.

Beaumont, M., Zhang, W., & Balding, D. (2002). Approximate Bayesian computation in population genetics. *Genetics* **162**, 2025–2035.

Blum, Michael G.B. (2010). Approximate Bayesian computation: A nonparametric perspective. *Journal of the American Statistical Association.* **105**.491, 1178–1187.

Cameron, E. & Pettitt, A. (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.

Cheng, T. (1949). The normal approximation to the Poisson distribution and a proof of a conjecture of Ramanujan. *Bulletin of the American Mathematical Society* **55**, 396–401.

Choi, L., Blume, J., & Dupont, W. (2015). Elucidating the foundations of statistical inference with 2×2 tables. *PLoS ONE* **10** 4, 1–22.

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

Creel, M. & Kristensen, D. (2015). ABC of SV: Limited information likelihood inference in stochastic volatility jump-diffusion models. *Journal of Empirical Finance* **31** 85–108.

Csilléry, K., Blum, M., Gaggiotti, O., & François, O. (2010). Approximate Bayesian computation (ABC) in practice. *Trends in Ecology and Evolution* **25**, 410–418.

Cornfield, J. (2010). A statistical problem arising from retrospective studies. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability* **4**, 135–148.

Cox, D. (1970). The continuity correction. *Biometrika* **57**, 217–219.

Del Moral, P., Doucet, A., & Jasra, A. (2012). An adaptive sequential Monte Carlo method for approximate Bayesian computation. *Statistics and Computing* **22**, 1009–1020.

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

Frazier, D., Martin, G., Robert, C., & Rousseau, J. (2018). Asymptotic properties of approximate Bayesian computation. *Biometrika* **105** 3 593–607.

Freedman, D. & Bickel, P. (1981). Some asymptotic theory for the bootstrap. *Annals of Statistics* **9**, 1196–1217.

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

Gentle, J. (2009). Statistical Computing. *Springer Science+Business Media, LLC.* Ch 1,11–13.

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

Kroonenberg, P. & Verbeek, A. (2018). The take of Cochran's rule: My contingency table has so many expected values smaller than 5, what am I to do?. *The American Statistician* **72** 2, 175–183.

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.

Li, B. & Fu, L. (2018). Exact test of goodness of fit for binomial distribution. *Statistical papers* **59** 3, 851–860.

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., & Ryder, R. (2012). Approximate Bayesian computational methods. *Statistics and Computing* **22**, 1167–1180.

Marjoram, P., Molitor, J., Plagnol, V., & Tavaré, S. (2003). Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences* **100**, 15324–15328.

Martin, G., McCabe, B., Frazier, D., Maneesoonthorn, W., & Robert, C. (2019). Auxiliary likelihood-based approximate Bayesian computation in state space models. *Journal of Computational and Graphical Statistics.*

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

Nissen, S. & Wolski, K. (2007). Effect of Rosiglitazone on the risk of myocardial infarction and death from cardiovascular causes. *The New England Journal of Medicine* **356**, 2457-2471.

Peters, G., Fan, Y., & Sisson, S. (2012). On sequential Monte Carlo partial rejection control approximate Bayesian computation. *Statistical Computing* **22**, 1209–1222.

Plackett (1964). The continuity correction in 2×2 tables. *Biometrika* **51**, 327–227.

Radavičius, M. & Židanavičiūtė, J. (2018). Semiparametric smoothing of sparse contingency tables. *Journal of Statistical Planning and Inference* **139**, 3900–3907.

Robinson, J., Bunnefeld, L., Hearn, J., Stone, G., & Hickerson, M. (2014). ABC inference of multi-population divergence with admixture from unphased population genomic data. *Molecular Ecology* **23**, 4458–4471.

Scherer, R. (2018). *PropCIs: Various Confidence Interval Methods for Proportions*. R package version 0.3-0. https://CRAN.R-project.org/package=PropCIs

Schweder, T. & Hjort, N. (2016). *Confidence, Likelihood, Probability*. 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. (2007). Confidence distribution (CD) - distribution estimator of a parameter. *IMS Lecture Notes* **54**, 132–150.

Wood, S. (2010). Statistical inference for noisy nonlinear ecological dynamic systems. *Nature* **466**, 1102.

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