Statistical Analysis and Data Mining: The ASA Data Science Journal

WILEY

RESEARCH ARTICLE

On an Empirical Likelihood Based Solution to the Approximate Bayesian Computation Problem

Sanjay Chaudhuri[1] | Subhroshekhar Ghosh[2] | Kim Cuc Pham[3,4]

[1]Department of Statistics, University of Nebraska-Lincoln, Lincoln, Nebraska, USA | [2]Department of Mathematics, National University of Singapore, Singapore | [3]Department of Statistics and Applied Probability, National University of Singapore, Singapore | [4]Department of Statistics and Data Science, National University of Singapore, Singapore

Correspondence: Sanjay Chaudhuri (schaudhuri2@unl.edu)

Received: 22 February 2024 | **Revised:** 13 August 2024 | **Accepted:** 9 September 2024

Funding: Sanjay Chaudhuri was partially funded by the National Science Foundation grant DMS-2413491. Subhroshekhar Ghosh was supported in part by the MOE grants R-146-000-250-133, R-146-000-312-114, A-8002014-00-00, and MOE-T2EP20121-0013.

Keywords: approximate Bayesian computation | Bayesian inference | differential entropy | empirical likelihood | estimating equation | information projection

ABSTRACT

Approximate Bayesian computation (ABC) methods are applicable to statistical models specified by generative processes with analytically intractable likelihoods. These methods try to approximate the posterior density of a model parameter by comparing the observed data with additional process-generated simulated data sets. For computational benefit, only the values of certain well-chosen summary statistics are usually compared, instead of the whole data set. Most ABC procedures are computationally expensive, justified only heuristically, and have poor asymptotic properties. In this article, we introduce a new empirical likelihood-based approach to the ABC paradigm called ABCel. The proposed procedure is computationally tractable and approximates the target log posterior of the parameter as a sum of two functions of the data—namely, the mean of the optimal log-empirical likelihood weights and the estimated differential entropy of the summary functions. We rigorously justify the procedure via direct and reverse information projections onto appropriate classes of probability densities. Past applications of empirical likelihood in ABC demanded constraints based on analytically tractable estimating functions that involve both the data and the parameter; although by the nature of the ABC problem such functions may not be available in general. In contrast, we use constraints that are functions of the summary statistics only. Equally importantly, we show that our construction directly connects to the reverse information projection and estimate the relevant differential entropy by a k-NN estimator. We show that ABCel is posterior consistent and has highly favorable asymptotic properties. Its construction justifies the use of simple summary statistics like moments, quantiles, and so forth, which in practice produce accurate approximation of the posterior density. We illustrate the performance of the proposed procedure in a range of applications.

1 | Introduction

The concept of likelihood is central to parametric statistical inference. However, for many models encountered in natural, engineering, and environmental sciences, tractable analytic forms of their likelihoods are not available. These models are often specified by a generative process, in the sense that independent samples can be generated from them for any input value of the model parameters. Approximate Bayesian computation (ABC) methods [1–8] are useful for Bayesian inference for models like these. Given the observed data, their objective is to estimate the posterior density of the parameters associated

with the data generating process without specifying a functional relationship between those parameters and the data.

In this article, we introduce a new modified empirical likelihood-based approach to the ABC problem that we call *ABCel*. Along the line of the traditional ABC procedures, it assumes the availability of the observed data and the ability to generate independent and identically distributed data sets of the same size from the generating process for any given value of the parameter of interest. In particular compared to the empirical likelihood-based $BC_{el} - AMIS$ algorithm [5] our method does not require specifications of estimating equations depending both on the observed data and the parameters, which are typically unavailable. The estimating equations are specified by the differences in the values of the appropriate summary statistics of the observed and the replicated data sets. These equations form natural constraints for our proposed modified empirical likelihood without directly involving the parameters. ABCel can be rigorously justified using various information projections and basic principles of Bayesian statistics. Furthermore, ABCel exhibits many favorable asymptotic properties and is computationally tractable.

Because of their potential application to complex models, ABC methods have generated immense interest in statistics. Suppose $X_o = (X_{o1}, X_{o2}, \dots, X_{on})^T$ is the vector of observations of length n, simulated from the "black box" (aka the data generating process) with an unknown input θ_o. Let the parameter θ take value in the set Θ and we assign a prior $\pi(\theta)$ on this set. We want to approximate the posterior $\Pi(\theta|X_o)$ and estimate θ_o from that posterior. The inference is based on additional data sets replicated from the black box for various input parameters $\theta \in \Theta$.

The ABC procedures proposed in the literature can be classified into two broad groups. The first one tries to estimate the posterior directly, and the second one attempts to estimate the density of X_o given θ from the replicates.

1.1 | Direct Estimation of the True Posterior

The *Rejection ABC* procedures try to sample from the parameter posterior directly. The basic ABC algorithm goes through the following steps:

1: Generate θ from $\pi(\theta)$.
2: Simulate $X_1 = (X_{11}, \dots, X_{1n})$ from the black box with parameter θ.
3: **if** $X_o = X_1$ **then**
4: Accept θ.
5: **end if**
6: Return to Step 1.

It is undeniable that at least hypothetically, the above algorithm provides a sample from the target posterior, from which inference about θ_o can be drawn. However, for continuous random variables the probability that $X_o = X_1$ is zero. So the above algorithm cannot be used as it is in most applications. Furthermore, due to the curse of dimensionality, an exact match of the data, even in discrete cases might be difficult to achieve. That is direct approximation of $\Pi(\theta|X_o)$ may be computationally cumbersome [9]. The acceptance rate of the test values of θ might be minuscule, reducing the computational efficiency of the procedure significantly.

In order to avoid the above pitfalls an approximate posterior is sampled from. The steps of this *simple rejection algorithm* are as follows:

1: Choose a small tolerance $\epsilon > 0$, a distance function d, and a vector of summary statistics $s(\cdot)$.
2: Generate θ from $\pi(\theta)$.
3: Simulate $X_1 = (X_{11}, \dots, X_{1n})$ from the black box with parameter θ.
4: **if** $d(s(X_o), s(X_1)) < \epsilon$ **then**
5: Accept θ
6: **end if**
7: Return to Step 2.

Even though attractive at first glance, and in spite of the availability of sophisticated and efficient sampling algorithms [10–12] with improved efficiency, the simple rejection algorithm described above has several shortcomings. First of all, instead of the full data, a vector of summary statistics are compared. That is, samples are actually drawn from $\Pi(\theta|s(X_o))$. If $s(\cdot)$ is sufficient for θ, which of course cannot be determined, the posterior given X_o is the same as the posterior given $s(X_o)$. However, for non-sufficient $s(\cdot)$ they may not be the same.

More crucially, the accuracy of the posterior approximation depends heavily on the value of the pre-specified tolerance. Clearly, small tolerances are preferred, but they are computationally prohibitive. The same curse of dimensionality prevents the use of high-dimensional summary statistics. Available results (e.g., Frazier et al. [13], Li and Fearnhead [14], Li and Fearnhead [15], Miller and Dunson [16], Bernton et al. [17]) show that, unless the pre-specified tolerance satisfies certain conditions which depend both on the summary statistics as well as the specified distance function (Miller and Dunson [16]; Bernton et al. [17]), the resulting rejection ABC posteriors may not have desirable asymptotic properties (e.g., Bayesian consistency, correct asymptotic frequentist coverage of credible intervals). Even though the posterior obtained from the simple rejection ABC is often considered to be the "gold standard" in the literature, Frazier et al. [13] argue that the connection between the exact posterior and the rejection ABC approximate could be quite remote. We refer to Robert [18] for a more detailed and succinct discussion of the possible pitfalls of the rejection ABC procedures.

1.2 | Methods Based on Density Estimation

Alternatives to the rejection-based ABC are provided by the so-called *pseudo-likelihood methods*. For each value of the parameter, these methods attempt to estimate the likelihood of the observed summaries, from observations simulated from the data generating process. One of the most popular pseudo-likelihood method is the *synthetic likelihood* introduced by Wood [8]. Here, in order to compute the likelihood, the summary statistics are assumed to be approximately jointly distributed as a multivariate normal random vector. Their mean and the covariance matrix

vary with the parameter and are estimated using the summaries simulated from the data generating process (see Price et al. [19]). Synthetic likelihood does not perform well when the normal approximations of the summary statistics are inaccurate. This happens, for example, when extreme values of the observations are used as summaries, or often when the process generates data vectors with dependent components, for example, from a time series, and so forth. In such cases, even well-chosen marginal transformations [8] usually cannot ensure the validity of the normal approximation over the whole parameter space. Extensions that relax the requirement of normality have been a continuous topic of interest for many researchers in this area. Fasiolo et al. [20] consider an extended saddle-point approximation, whereas Dutta et al. [21] proposes a method based on logistic regression. By making use of various transformations An, Nott, and Drovandi [22] and Priddle and Drovandi [23] consider semi-parametric extensions of synthetic likelihood. Drovandi, Pettitt, and Lee [24] describe an encompassing framework for many of the above suggestions, which they call parametric Bayesian indirect inference. Frazier and Drovandi [25] have recently proposed a robustified version of synthetic likelihood that is able to detect and provide some degree of robustness to misspecification.

1.3 | An Empirical Likelihood Based Method

The $BC_{el} - AMIS$ procedure introduced by Mengersen, Pudlo, and Robert [5] is pseudo-likelihood based, where the intractable data likelihood is replaced by a non-parametric empirical likelihood [26]. This procedure follows the traditional Bayesian empirical likelihood (BayesEL) procedures [27, 28] and specifies the likelihood from the jumps of the joint empirical distribution function of the data computed under appropriate constraints.

In particular they assumed that $X_{o1}, \ldots, X_{on}$ are i.i.d and a set of constraints of the form $E[h(X_{oi}, \theta)] = 0, \ \forall i = 1, \ldots, n$ are available. Here the expectation is taken w.r.t. the unknown true distribution. An empirical likelihood can then be calculated by re-weighting the data by weights given by:

$$\hat{w} = argmax_{w \in \mathcal{W}_\theta} \prod_{i=1}^{n} w_i, \ \text{where}$$

$$\mathcal{W}_\theta = \left\{ w : \sum_{i=1}^{n} w_i h(X_{oi}, \theta) = 0 \right\} \cap \Delta_{n-1}$$

Empirical likelihood does not require the summaries to be approximately normal. However, the $BC_{el} - AMIS$ procedure typically requires constraints based on analytically tractable estimating functions of both the data and the parameters. By the nature of the ABC problem, such functions are not readily available, and thus, the proposed $BC_{el} - AMIS$ algorithm is not always easy to implement in practice. The exponentially tilted empirical likelihood [29] based ABC proposed by Grazian and Liseo [30] suffers from similar problems.

The proposed paradigm of ABCel, which is essentially a modified empirical likelihood-based method, neither uses any tolerance parameter nor assume any specific form of a pseudo-likelihood. It first finds an analytic expression of an approximation of the target posterior. This expression is then used to approximate the data density and obtain an optimal approximate of the target.

1.4 | The Proposed ABCel

The proposed ABCel procedure operates directly on the set of densities and estimates the posterior by using only the simple known properties of the data generating process. In particular, it only assumes that given the parameter θ, the generated summaries are independent and identically distributed.

For an input value θ, let s_1 be one replicated summary generated from the data generating process. Furthermore suppose that $f_0(\theta, s_1, s_o)$ is the true joint density of (θ, s_1, s_o). Clearly, our target is the posterior density $f_0(\theta|s_o)$. The conditional independence of s_1 and s_o given θ implies that the true conditional density of (θ, s_1) given s_o can be factored as $f_0(\theta, s_1|s_o) = f_0(s_1|\theta)\Pi(\theta|s_o)$. Thus, in order to approximate the posterior, we first consider a joint density $f(\theta, s_1, s_o)$ and project the corresponding conditional density $f(\theta, s_1|s_o)$ on a set of densities of the form $f_0(s_1|\theta)q(\theta)$, where $q(\theta)$ varies over all densities defined on Θ. Once the projection $f_0(s_1|\theta)q^\star(\theta)$ is determined, $q^\star(\theta)$ is the required approximation of $\Pi(\theta|s_o)$. The main result of Section 3.1 is that $q^\star(\theta)$ obtained from a direct information projection of $f(\theta, s_1|s_o)$ has a closed form (Theorem 1). In particular, the result shows that the log numerator of the optimal $q^\star(\theta)$ is the sum of the *true* expectation of the log-joint density $\log f(\theta, s_1, s_o)$ w.r.t. $f_0(s_1|\theta)$ and the true differential entropy of the density $f_0(s_1|\theta)$.

It is not difficult to argue that the optimal $q^\star(\theta)$ obtained from the true joint density $f_0(\theta, s_1, s_o)$ is indeed $\Pi(\theta|s_o)$. The approximate true posterior is by applying Theorem 1 on an estimate of $f_0(\theta, s_1, s_o)$ obtained from a reverse-information projection (i.e., m-projection) over a subset of the joint densities defined on (θ, s_1, s_o). This subset is primarily determined by the second known property of the data-generating process, namely, the conditional marginal densities of s_o give θ is identical to the conditional marginal density of s_1 given θ.

Finally we show that the proposed empirical likelihood estimator follows the above recipe in the sample. For the summary statistics which can identify the underlying density, for example, quantiles, moments, and so forth, constraints on the empirical likelihood approximately ensure the required match of the conditional marginal likelihoods. Moreover, the empirical likelihood is computed by maximizing over a vast class of non-parametrically specified likelihoods. That is, the proposed ABCel approximates the true joint likelihood well by satisfying the basic assumptions on the data-generating process.

The proposed posterior has many favorable properties. Asymptotically, under mild conditions, the proposed posterior is Bayesian and posterior consistent for the true value of the parameter when both the sample size and the number of replications grow unbounded (Section 4). We invoke the results from Ghosh, Chaudhuri, Gangopadhyay [31], to further explore its properties when the number of replications increases, but the sample size is held fixed (Section 4.2).

Finally, perhaps the biggest advantage of our procedure is its easy implementation. In order to compute the likelihood, a user only needs to specify an appropriate set of summary statistics and the number of replications to be simulated for each value of the parameter. For the estimation of the differential entropy, the order of the percentile (see Section 2.3) has to be chosen. Unlike many ABC procedures, no other parameters tuning or otherwise are either need to be specified or estimated. Moreover, both the empirical likelihood and the proposed differential entropy estimator can be computed using a fast algorithm implemented in multiple software. An easy adaptive Markov chain Monte Carlo procedure due to Haario, Saksman, and Tamminen [32] can be adapted to efficiently sample from the resulting posterior (see Section E of the Supporting Information).

2 | ABC Empirical Likelihood Posterior

In this section, we introduce the basic ideas of the proposed ABC empirical likelihood (ABCel) posterior. A modified empirical likelihood-based method which only depends on the observed data and replicated data simulated from the generating process is described. One part of the ABCel posterior is constructed using this likelihood. The other part is an estimate of a differential entropy, which is computed using a non-parametric Euclidean likelihood. We only describe the motivation and computation of ABCel posterior in this section. The justification of the procedure is presented in the subsequent sections.

2.1 | Setup

Let θ be the input parameter of the data generating process. We assume θ takes values in a set Θ and assign a prior $\pi(\theta)$ to it. For any given value $\theta \in \Theta$, the process generates i.i.d. n-dimensional random vectors from an unknown density depending on parameter θ. Since the density of the same random variable would change with the value of the parameter, we make their connection explicit in the notation. As for example, the observed data x_o is a realization of the random variable $X_o(\theta_o)$, that is, the random variable $X_o(\theta)$ generated from the process with $\theta = \theta_o$. Additionally, for each $\theta \in \Theta$, realizations from m i.i.d. replicated random variables $X_i(\theta)$, $i = 1, 2, \ldots, m$ are drawn from the process with input parameter value θ. That is in total, we consider a set of n-dimensional random vectors $\{X_i(\theta), i \in \mathbb{M}_o, \theta \in \Theta\}$, where $\mathbb{M}_o = \{o\} \cup \mathbb{N}$, that is, the set of positive integers appended with the symbol o. By construction, conditional on θ, $\{X_i(\theta), i \in \mathbb{M}_o\}$ are independent and identically distributed.

The true density $f_0(X_o(\theta)|\theta)$ is unknown, which prevents computation of the *exact* posterior $\Pi(\theta|X_o)\Pi(\theta|X_o(\theta) = x_o) \propto f_0(x_o|\theta)\pi(\theta)$. The problem is to approximate the posterior using the observation x_o and the replications $X_i(\theta)$, $i = 1, 2, \ldots, m$.

As we have noted before due to the *curse of dimensionality* direct approximation of $\Pi(\theta|X_o)$ may be computationally cumbersome [9]. Thus in most ABC applications inference on θ is drawn using a posterior conditional on $r \times 1$ vector $s(\cdot)$ of summary statistics of the observations.

Suppose that for a given $\theta \in \Theta$, $s(X(\theta))$ inherits a density $f_0(s(X(\theta))|\theta)$ from $X(\theta)$. Using the summaries $s(X_i(\theta))$, $i \in \mathbb{M}_o$, most ABC procedures estimate the *target* posterior

$$\Pi(\theta|s(X_o))\Pi(\theta|s(X_o(\theta)) = s(x_o)) = \frac{f_0(s(x_o)|\theta)\pi(\theta)}{\int f_0(s(x_o)|t)\pi(t)dt} \tag{1}$$

2.2 | Construction of ABCel Posterior

The ABCel posterior is based on the following observation. Suppose $\theta = \theta_o$. Then by construction, the random variables $s(X_o(\theta_o))$, $s(X_1(\theta))$, $\ldots$, $s(X_m(\theta))$ are identically distributed. Now if $E^0_{s|\theta_o}$ denotes the expectation w.r.t $f_0(s(X_i(\theta_o))|\theta_o)$, then for any $i = 1, \ldots, m$,

$$E^0_{s|\theta_o}[s(X_i(\theta_o)) - s(X_o(\theta_o))] = 0 \tag{2}$$

The proposed empirical likelihood estimator of the posterior consists of two parts. The first is an empirical likelihood which is constructed using constraints based on the expectation in (2). For any $\theta \in \Theta$ and for each $i = 1, 2, \ldots, m$, define

$$h_i(\theta) = s(X_i(\theta)) - s(X_o(\theta_o)) \tag{3}$$

and the random set:

$$\mathcal{W}_\theta = \left\{ w : \sum_{i=1}^{m} w_i[s(X_i(\theta)) - s(X_o(\theta_o))] = 0 \right\} \cap \Delta_{m-1} \tag{4}$$

where Δ_{m-1} is the $m-1$ dimensional simplex.

We define the optimal weights $\hat{w}$ as:

$$\hat{w} := \hat{w}(\theta) := \underset{w \in \mathcal{W}_\theta}{\arg\max} \left(\prod_{i=1}^{m} m w_i \right) \tag{5}$$

If the problem in (5) is infeasible, $\hat{w}$ is defined to be zero. These optimal weights are used in the first part of the posterior estimate.

The second part requires an estimate of the differential entropy $H^0_{s|\theta}(\theta)$ of $f_0(\cdot|\theta)$ at the input $\theta \in \Theta$, which is defined by, $H^0_{s|\theta}(\theta) = -\int f_0(s|\theta)\log f_0(s|\theta)ds$. Let, $\hat{H}^0_{s|\theta}(\theta)$ is an estimate of $H^0_{s|\theta}(\theta)$ (see Section 2.3 below for details).

By using this estimate and the optimal $\hat{w}$ we define *ABC empirical likelihood* (*ABCel*) estimate of the required posterior as,

$$\hat{\Pi}(\theta|s(X_o)) = \frac{\left[e^{\left(\frac{1}{m}\sum_{i=1}^{m} \log(\hat{w}_i(\theta)) + \hat{H}^0_{s|\theta}(\theta) \right)} \right] \pi(\theta)}{\int_{t \in \Theta} \left[e^{\left(\frac{1}{m}\sum_{i=1}^{m} \log(\hat{w}_i(t)) + \hat{H}^0_{s|t}(t) \right)} \right] \pi(t)dt} \tag{6}$$

When $\prod_{i=1}^{m} \hat{w}_i = 0$, we define $\hat{\Pi}(\theta|s(X_o)) = 0$.

The empirical likelihood used in (6) is different from the original Bayesian empirical likelihood (BayesEL) posterior [27, 28] and the previous use of Bayesian empirical likelihood in an ABC setting [5] in two ways. First, instead of the sum, it uses the mean of the log-weights. This is significant in several ways (see below) and

can be justified by an information projection argument described in Section 3.1.

The second aspect is our choice of the constraints, which is probably more significant. Usual BayesEL formulations (as in Mengersen, Pudlo, and Robert [5]) would have used constraints which are functions of $s(X_o)$ and θ. Such estimating equations are not necessarily known in an ABC problem. In our formulation, we avoid such specifications using constraints based on $s(X_o)$ and the replicated summaries $s(X_i)$, $i = 1, 2, \ldots, m$. The summaries in (3) are routinely used in *Exponential Random Graph Models* (*ERGM*) literature [33]; however, the weights are obtained by maximizing the entropy [34, 35] instead of a likelihood as in (5) above. This is equivalent to maximizing a cross-entropy term (see (12)). Unlike the rejection ABC, we do not need to specify any distance function or any tolerance parameter.

From the formulation of the constraints, the optimal weights in (5) define a constrained joint-conditional empirical distribution function supported on m observations $(s(X_i(\theta)), s(x_o))$ given θ. This is somewhat similar to the data-replication methods, discussed in Lele, Dennis, and Lutscher [36] and Doucet, Godsill, and Robert [37] (see also Gouriéroux and Monfort [38]). More importantly, as we argue in Section 3.3 below, for simple choices of summary our constraints ensure that the above joint-conditional $f(s(X_i(\theta)), s(X_o(\theta))|\theta)$ is estimated by approximately equating the underlying marginal conditional densities $f(X_i(\theta)|\theta)$ and $f(X_o(\theta)|\theta)$ of $X_i(\theta)$ and $X_o(\theta)$ respectively, which provides an argument in favor of the optimality of our procedure.

No analytic expression for the proposed ABCel posterior exists in general. By construction, each $\hat{w}_i$ is bounded for all values θ. All components of $\hat{w}$ in (5) and the ABCel posterior are strictly positive iff the origin of $\mathbb{R}^r$ is in the interior of the convex hull defined by the vectors $h_1, h_2, \ldots, h_m$. Otherwise the ABCel posterior would be zero (even though in the boundary of the above convex hull, the constrained optimization in (5) is still feasible). It is well-known (see, e.g., Chaudhuri, Mondal, and Yin [39]) that the supports of the Bayesian empirical likelihood (BayesEL) posteriors are in general non-convex. It is expected that the proposed ABCel posterior will suffer from the same deficiency as well. However, as we discuss below (see Section 5) the non-convexity of the support does not make the proposed ABCel posterior computationally expensive. One can devise easy Markov chain Monte Carlo (MCMC) techniques to draw samples from this posterior at a reasonable computational cost. Such samples are enough for making posterior inference.

Finally, the proposed method is more general than the synthetic likelihood [8]. The latter assumes normality of the joint distribution of the summary statistics. Even though many summary statistics are asymptotically normally distributed, this is not always the case. This is especially true if the process generates dependent data sets, for example, a time series, spatial data, and so forth. In such cases, the synthetic likelihood can perform quite poorly (see, e.g., Section 6.2 below). Some relaxation of normality has been proposed by various authors, but many of these procedures require specification or estimation of additional tuning parameters. In our empirical likelihood approximation, we only require the observed data and simulated data from the generating process for a given θ.

2.3 | Differential Entropy Estimation

Several estimators of differential entropy have been studied in the literature. The oracle estimator is given by $-\sum_{i=1}^{m} \log \hat{f}_0(s(X_i(\theta)))/m$. In this article we implement a weighted k-nearest neighbor based estimator due to Kozachenko and Leonenko [40] described in Berrett, Samworth, and Yuan [41]. This estimator is easy to compute and has better asymptotic properties than histogram or kernel-based estimators [42].

The nearest-neighbor estimator requires us to specify k, the order of the nearest neighbor. Ideally, k should depend on m. Our experiments suggest any value of k as long as it is not very small or not very large, makes little difference. Note that, other than the summary statistics and the number of replications m, this k is the only parameter a user needs to specify in order to compute the proposed posterior. No other parameters tuning or otherwise are required.

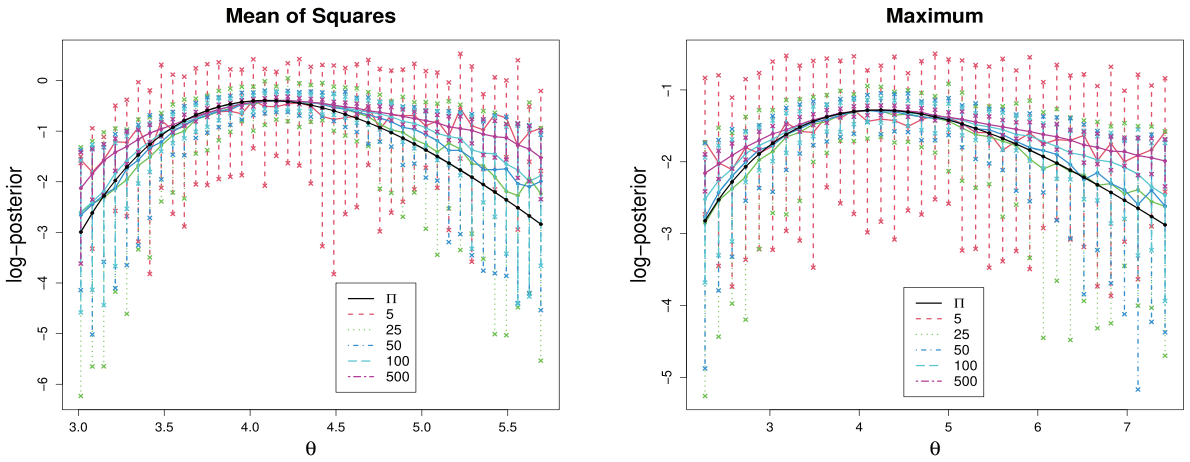
2.4 | Example

In Figure 1, we compare the shape of the ABCel log-posteriors with the true log-posteriors Π for the variance of a Normal distribution with zero mean conditional on (a) $s^{(1)}(X_i) = \sum_j X_{ij}^2/n$ (Figure 1A) and (b) $s^{(2)}(X_i) = \max_j (X_{ij})$ (Figure 2B). Here, for each $i = 1, 2, \ldots, m$, and $j = 1, 2, \ldots, 100$, the observation X_{ij} is drawn from a $N(0, \theta)$, with $\theta_o = 4$. We assume that the parameter θ follows a $U(0, 10)$ prior.

The log-posteriors were compared on a grid of parameters whose true posterior values were larger than the 0.05. Based on 100 repetitions, At each value of θ and m, the mean and the endpoints of the symmetric 95% confidence intervals are shown in the figure. To make the comparison of the shapes easier, for each m, the maximum of the mean of ABCel log-posterior was matched with the maximum value of the true log-posterior.

From Figure 1, it follows that for $m = 25$ and $m = 50$, for each value of θ the means of the estimated log-posteriors (solid colored lines) are very close to the true log-posterior (solid black line) for both $s^{(1)}(X_o)$ and $s^{(2)}(X_o)$. Furthermore, the 95% confidence bands always cover the corresponding true value of the log-posterior. It is evident that the proposed ABCel posterior is a good approximation of the true posterior up to a scaling constant. This is even true for the summary function $s^{(2)}(X_o)$, which unlike $s^{(1)}(X_o)$, asymptotically does not converge to a normal random variable under any centering or scaling.

As the number of replicates, that is, m increases (see $m = 500$), in Figure 1, the log-posterior, tends to get more flat in shape. However, the confidence bands get narrower. This is somewhat expected. We have kept the number of summaries fixed here. However, statistical intuition mandates that the number of summaries used should increase with the number of replications. Using results from [31], we discuss such phenomena in more detail in Section 4.2 below. Furthermore, an example illustrating the inter-relationship between the number and the nature of the summary statistics with the number of replications can be found in Section 6.1.



(A) $\hat{\Pi}(\theta \mid s^{(1)}(X_o))$ (B) $\hat{\Pi}(\theta \mid s^{(2)}(X_o))$

FIGURE 1 | Comparison of the true log-posterior with the logarithm of the proposed estimator for different values of m. The samples of size $n = 100$ were drawn from $N(0, \theta)$ distribution with $\theta_o = 4$. We chose (A) $s^{(1)}(X_i) = \sum_j X_{ij}^2/n$ and (B) $s^{(2)}(X_i) = \max_j (X_{ij})$ and a $U(0, 10)$ prior on θ. The true log-posterior is in black. For each value of θ and m the means and the 95% confidence intervals of the estimated log-posterior based on 100 repetitions are shown.

3 | Justification for the ABC Empirical Likelihood Posterior

In this section, we provide a rigorous justification for the proposed modified empirical likelihood-based posterior estimate $\hat{\Pi}(\theta|s(X_o))$ introduced in Section 2. Our arguments use direct and reverse information projections of appropriate conditional densities on judiciously chosen density sets. We first discuss a general functional form of a posterior approximation, then use this functional form to find an accurate approximation of the true posterior. Both of these approximations are obtained in the population. Finally, it is argued that our posterior estimate $\hat{\Pi}(\theta|s(X_o))$ approximates the above recipe in the sample.

Let $s_1 s(X_1(\theta))$ denote a replicated summary random variable corresponding to $X_1(\theta)$ obtained from the data generating process with an input $\theta \in \Theta$. As before, the observed summary random variable $s_o s(X_o(\theta))$. In order to justify the proposed empirical likelihood-based estimator, it is more convenient to work with the joint densities (denoted by $f(\theta, s_1, s_o)$) defined on (θ, s_1, s_o). Let $\mathcal{F}$ be the set of all such joint densities.

Let $f_0(\theta, s_1, s_o)$ be the *true* joint density of (θ, s_1, s_o) of the data generating process. We first explore its properties. By assumption given on θ, the s_o and s_1 are identically distributed and mutually independent. If we assume that $f_0(s_1|\theta)$ is the unknown conditional density s_1 inherits from the data-generating process, the conditional density of s_o given θ will also be the same density. We will denote the latter by $f_0(s_o|\theta)$.

Clearly, the corresponding *true* joint density of (θ, s_1, s_o), denoted by $f_0(\theta, s_1, s_o)$ is in the set $\mathcal{F}$. Furthermore, using the conditional independence of s_1 and s_o given θ we get:

$$f_0(\theta, s_1, s_o) = f_0(s_1|\theta) f_0(s_o|\theta)\pi(\theta)$$

where by construction the true marginal of θ, that is, $f_0(\theta)$ equals the prior $\pi(\theta)$. From construction, it also follows that, $\Pi(\theta|s(X_o)) = f_0(\theta|s_o)$, where $f_0(\theta|s_o)$ is the conditional density of θ given s_o corresponding to the joint $f_0(\theta, s_1, s_o)$.

We now focus on the conditional density of (θ, s_1) given s_o. Using the conditional independence of s_o and s_1 given θ, for all $\theta \in \Theta$, the *true* conditional density of (θ, s_1) given s_o can be written as:

$$f_0(\theta, s_1|s_o) = f_0(s_1|s_o, \theta) f_0(\theta|s_o) = f_0(s_1|\theta)\Pi(\theta|s(X_o)) \quad (7)$$

Suppose $\mathcal{Q}'$ is the subset of densities on (θ, s_1) defined as:

$$\mathcal{Q}' = \left\{ q'(\theta) f_0(s_1|\theta) : q'(\theta) \in \mathcal{Q}_\Theta \right\} \quad (8)$$

where $\mathcal{Q}_\Theta$ be the set of all densities defined on the set Θ. Since the posterior $\Pi(\theta|s(X_o))$ are in $\mathcal{Q}_\Theta$, the true $f_0(\theta, s_1|s_o) \in \mathcal{Q}'$. Furthermore, any density in $\mathcal{Q}'$ is a product of a density of s_1 and a density of θ. That is, $\Pi(\theta|s(X_o))$ can be approximated by integrating any density in $\mathcal{Q}'$ with respect to s_1.

Our strategy of approximating $\Pi(\theta|s(X_o))$ is as follows:

1. Take a joint density $f(\theta, s_1, s_o) \in \mathcal{F}$, and find the corresponding conditional density of (θ, s_1) given s_o denoted by $f(\theta, s_1|s_o)$.
2. Find the density $q^\star(\theta, s_1) \in \mathcal{Q}_\Theta$ closest to $f(\theta, s_1|s_o)$ under some pre-specified criterion.
3. Integrate $q^\star(\theta, s_1)$ over s_1.

The ABCel posterior described in Section 2 uses *direct information projection* in Step (2), which leads to a closed form solution of the approximate $\Pi(\theta|s_o)$ for any joint $f(\theta, s_1, s_o) \in \mathcal{F}$. Some heuristic justification of information projection can be found in Cuc [43].

3.1 | Functional Form of the Posterior Approximation

Let $f(\theta, s_1, s_o) \in \mathcal{F}$, and $f(\theta, s_1|s_o)$ be the corresponding conditional density of (θ, s_1) given s_o. We compute the information projection of $f(\theta, s_1|s_o)$ on $\mathcal{Q}'$ by minimizing Kullback–Leibler divergence between the above conditional density and each density $q(\theta, s_1) \in \mathcal{Q}'$. For any s_o, the Kullback–Leibler divergence [44] between $q(\theta, s_1)$ and $f(\theta, s_1|s_o)$ is defined as $D_{KL}\Big[q(\theta s_1)\Big\|f(\theta, s_1|s_o)\Big] = \int q(\theta, s_1)\log\Big(\frac{q(\theta,s_1)}{f(\theta,s_1|s_o)}\Big)ds_1 d\theta$. The *information projection* of $f(\theta, s_1|s_o)$ onto $\mathcal{Q}'$ is given by:

$$q^\star(\theta, s_1) := \underset{q(\theta,s_1)\in\mathcal{Q}'}{\arg\min}\, D_{KL}\Big[q(\theta s_1)\Big\|f(\theta, s_1|s_o)\Big]$$

Since the set $\mathcal{Q}'$ is convex [45], for any density $f(\theta, s_1|s_o)$ its projection is unique. Next, we find an analytic expression of $q^\star(\theta, s_1)$.

Theorem 1. *For any density $f \in \mathcal{F}$, let $E^0_{s_1|\theta}[\log f(\theta, s_1, s_o)] = \int f_0(s_1|\theta)\log f(\theta, s_1, s_o)ds_1$ and $H^0_{s_1|\theta}(\theta) = -\int f_0(s_1|\theta)\log f_0(s_1|\theta)ds_1$ be the differential entropy of the density $f_0(s_1|\theta)$. Furthermore, let us define*:

$$f'(\theta|s_o) := \frac{e^{E^0_{s_1|\theta}[\log f(\theta,s_1,s_o)]+H^0_{s_1|\theta}(\theta)}}{\int_{t\in\Theta} e^{E^0_{s_1|t}[\log f(t,s_1,s_o)]+H^0_{s_1|t}(t)}dt} \tag{9}$$

Then $q^\star(\theta, s_1) = f'(\theta|s_o)f_0(s_1|\theta)$.

The proof of above theorem is presented in the Appendix A. We show that, for any $q(\theta, s_1) = q'(\theta)f_0(s_1|\theta) \in \mathcal{Q}'$, such that $q' \in \mathcal{Q}_\Theta$, the relationship $D_{KL}[q(\theta s_1)\|f(\theta, s_1|s_o)] = D_{KL}[q'(\theta)\|f'(\theta|s_o)] + C$ holds, where C is a non-negative function of s_o and some hyper-parameters of the prior, and does not depend on q or q'. Now the L.H.S. is minimum when $q'(\theta) = f'(\theta|s_o)$, from which the result follows.

Theorem 1 shows that for any joint density $f(\theta, s_1, s_o) \in \mathcal{F}$, the density $f_0(s_1|\theta)f'(\theta|s_o)$ is the best approximation of $f_0(s_1|\theta)\Pi(\theta|s(X_o))$ over $\mathcal{Q}'$, for all θ, s_1 and s_o. The posterior $\Pi(\theta|s(X_o))$ can naturally be approximated by integrating this best approximation over s_1. Since $f'(\theta|s_o)$ is independent of s_1, the corresponding approximation of $\Pi(\theta|s(X_o))$ is trivially given by $\int f_0(s_1|\theta)f'(\theta|s_o)ds_1 = f'(\theta|s_o)$.

If $f(\theta, s_1, s_o) = f_0(\theta, s_1, s_o)$, clearly $f_0(\theta, s_1|s_o) \in \mathcal{Q}'$, and by definition it is its own information projection. That is the approximation of $\Pi(\theta|s(X_o))$ is exact. That is $f'_0(\theta|s_o) = \Pi(\theta|s(X_o))$. Furthermore, when $f_0(s_1|\theta)$ belongs to a location family $H^0_{s_1|\theta}(\theta)$ is not a function of θ. In that case $f'_0(\theta|s_o) \propto exp\Big\{E^0_{s_1|\theta}[\log f_0(\theta, s_1, s_o)]\Big\}$.

Note that, like it should in a Bayesian procedure, in the expression of $f'(\theta|s_o)$, the effect of the replicate summary s_1 gets integrated out. In the proposed empirical likelihood-based estimator, the expectation of the log-joint density is approximated by the mean of the log-optimal weights, which approximately averages out the effect of the replicated summaries from the posterior estimate. Furthermore, the proposed empirical likelihood estimates an optimal approximate of the true posterior, as we argue below.

3.2 | Optimal Posterior Approximation

Theorem 1 shows that for any joint density $f(\theta, s_1, s_o) \in \mathcal{F}$, the density $f'(\theta|s_o)$ provides an approximation of $\Pi(\theta|s(X_o))$ via information projection, with no other assumption required. Furthermore, the approximation is exact when the chosen joint density $f(\theta, s_1, s_o)$ is the true joint density $f_0(\theta, s_1, s_o)$. It, however, does not provide a way to choose the joint $f(\theta, s_1, s_o) \in \mathcal{F}$ such that $f'(\theta|s_o)$ is an optimal approximation of the true posterior in any sense. We discuss the criterion of such optimality in this section and then discuss its relationship with the proposed empirical likelihood-based procedure.

To that goal, suppose for a joint density $f(\theta, s_1, s_o) \in \mathcal{F}$, $f(\theta, s_o)$, and $f(s_1|s_o, \theta)$ respectively denote the corresponding marginal density of (θ, s_o) and the conditional density of s_1 given θ and s_o. Furthermore, suppose $f(s_1|\theta)$ and $f(s_o|\theta)$ respectively denote the conditional densities of s_1 and s_o given θ. Recall that, unless $f(\theta, s_1, s_o)$ is the true joint density f_0, the two conditional densities of s_1 and s_o given θ may not be equal. The optimality criterion is based on the following result.

Theorem 2.

a. *Let $f(\theta, s_1, s_o) \in \mathcal{F}$. Then for all θ, s_1 and s_o,*

$$\log f(\theta, s_o) - \Big\{E^0_{s_1|\theta}[\log f(\theta, s_1, s_o)] + H^0_{s_1|\theta}(\theta)\Big\} = D_{KL}\Big[f_0(s_1|\theta)\Big\|f(s_1|s_o, \theta)\Big] \geq 0 \tag{10}$$

b. *If under the joint $f(\theta, s_1, s_o)$, s_o is conditionally independent of s_1 given θ, it follows that*:

$$\log f(\theta, s_o) - \Big\{E^0_{s_1|\theta}[\log f(\theta, s_1, s_o)] + H^0_{s_1|\theta}(\theta)\Big\} = D_{KL}\Big[f_0(s_1|\theta)\Big\|f(s_1|\theta)\Big]$$

c. *If $f(\theta, s_1, s_o) = f_0(\theta, s_1, s_o)$, $E^0_{s_1|\theta}[\log f(\theta, s_1, s_o)] + H^0_{s_1|\theta}(\theta) = \log f(\theta, s_o) = \log f_0(\theta, s_o)$. Furthermore, $f'(\theta|s_o) = f(\theta|s_o) = f_0(\theta|s_o) = \Pi(\theta|s(X_o))$.*

Theorem 2a can be proved by a direct expansion of the left-hand side of the expression. The other two statements follow from the first. In particular, we get $E^0_{s_1|\theta}[\log f_0(\theta, s_1, s_o)] + H^0_{s_1|\theta}(\theta) = \log f_0(\theta, s_o)$.

This theorem shows that for any joint density $f(\theta, s_1, s_o) \in \mathcal{F}$, $f'(\theta|s_o)$ is not same as the corresponding conditional density $f(\theta|s_o)$. The log-numerator in the expression of $f'(\theta|s_o)$ is a lower bound of $\log f(\theta, s_o)$. Furthermore, their difference equals the Kullback–Leibler divergence between the true data-generating density $f_0(s_1|\theta)$ and the user-specified conditional density of s_1 given s_o and θ, that is, $f(s_1|s_o, \theta)$. Clearly, f_0 is a minima of this divergence over $\mathcal{F}$, for which, by Theorem 2c, the approximation of $\Pi(\theta|s_o)$ by $f'_0(\theta|s_o)$ is exact. Thus, we minimize the above Kullback–Leibler divergence to find the optimal approximation.

For any density $f(\theta, s_1, s_o) \in \mathcal{F}$,

$$D_{KL}\Big[f_0(s_1|\theta)\Big\|f(s_1|s_o, \theta)\Big] = 0$$

$$\Leftrightarrow f_0(s_1|\theta) = f(s_1|s_o, \theta) \text{ for all } s_1, s_o \text{ and } \theta$$

$\Leftrightarrow s_1$ is conditionally independent of s_o given θ, and

$f(s_1|\theta) = f_0(s_1|\theta)$ for all s_1, s_o and θ

Thus by minimizing the above Kullback–Leibler divergence we can only identify the density $f_0(s_1|\theta)$. The choice of $f(s_o|\theta)$ and the marginal $f(\theta)$ remains arbitrary. That is minimum is not unique and $f_0(\theta, s_1, s_o)$ is not the unique density in $\mathcal{F}$ where the minimum of the above Kullback–Leibler divergence is attained.

In order to make the minimal argument unique, define $\mathcal{F}' \subseteq \mathcal{F}$ be the collection of all joint-densities $f(\theta, s_1, s_o) \in \mathcal{F}$, such that for all values of $\theta \in \Theta$,

a. the corresponding conditional density of s_1 given θ is the same as the corresponding conditional density of s_o given θ, and
b. the corresponding marginal density of θ is the prior π.

The constraints that specify $\mathcal{F}'$ comply with our assumption about the data generating process. In particular, the true joint density $f_0 \in \mathcal{F}'$ (see Section 3). That is, it minimizes the divergence in (10) over $\mathcal{F}'$.

However, if $f \in \mathcal{F}'$ such that the above divergence is zero, then for all θ, s_1 and s_o,

$$f(\theta, s_1, s_o) = f(s_1|s_o, \theta) f(s_o|\theta) f(\theta) = f_0(s_1|\theta) f(s_o|\theta) f(\theta)$$

Furthermore, by the construction of $\mathcal{F}'$, it follows that $f(s_o|\theta) = f_0(s_o|\theta)$ and $f(\theta) = \pi(\theta)$ for all s_o and θ. So it follows that, for all θ, s_1 and s_o,

$$f(\theta, s_1, s_o) = f_0(s_1|\theta) f_0(s_o|\theta) \pi(\theta) = f_0(\theta, s_1, s_o)$$

From the arguments above, the following result is now evident.

Theorem 3. *Suppose $\mathcal{F}'$ is the subset of densities over (θ, s_1, s_o) as defined above. Then $f_0 \in \mathcal{F}'$ uniquely minimizes $D_{KL}\left[f_0(s_1|\theta) \middle\| f(s_1|s_o, \theta)\right]$ over $\mathcal{F}'$.*

An estimate of $f_o(\theta, s_1, s_o)$ can therefore be obtained as:

$$\hat{f}_0(\theta, s_1, s_o) = \arg\min_{f \in \mathcal{F}'} D_{KL}\left[f_0(s_1|\theta) \middle\| f(s_1|s_o, \theta)\right] \tag{11}$$

The estimate $\hat{f}_0(\theta, s_1, s_o)$ in (11) is actually a reverse information projection of $f_0(s_1|\theta)$ on the set of densities $f(s_1|s_o, \theta)$ such that $f(\theta, s_1, s_o) \in \mathcal{F}'$. Furthermore, since $f_0(s_1|\theta)$ is fixed, we get

$$\begin{aligned}
&\arg\min_{f \in \mathcal{F}'} D_{KL}\left[f_0(s_1|\theta) \middle\| f(s_1|s_o, \theta)\right] \\
&= \arg\max_{f \in \mathcal{F}'} \int f_0(s_1|\theta) \log f(s_1|s_o, \theta) ds_1 \\
&= \arg\max_{f \in \mathcal{F}'} \left\{ \int f_0(s_1|\theta) \log f(s_1, s_o|\theta) ds_1 \right. \\
&\quad \left. - \log \int f(s_1, s_o|\theta) ds_1 \right\}
\end{aligned} \tag{12}$$

That is, in order to minimize our loss function, we only need to maximize the cross-entropy term over the specified $\mathcal{F}'$.

3.3 | Connection to the Proposed ABCel Posterior

From the justifications presented above, for appropriate summary statistics, the task is to specify the set of joint densities $\mathcal{F}'$, at least approximately, and minimize the divergence in (10) over this specified set. Once $\hat{f}_0(\theta, s_1, s_o)$ is computed, the corresponding approximation of $\Pi(\theta|s_o)$ is given by the corresponding $\hat{f}'_0(\theta|s_o)$. This can be obtained by substituting $f(\theta, s_1, s_o)$ by $\hat{f}'_0(\theta, s_1, s_o)$ in (9).

We now argue that with simple choices of summary statistics, the proposed modified empirical likelihood-based procedure follows the same recipe. In the notations of Section 2, for $i = 1, 2, \ldots, m$, let $s_i = s(X_i(\theta))$ be the values of summary of $X_i(\theta)$ generated with input $\theta \in \Theta$. Note that, the optimal weights from (5) defines an empirical estimate of the conditional distribution of (s_1, s_o) given θ, supported over the points (s_i, s_o), $i = 1, 2, \ldots, m$. This estimate is obtained by minimizing the required divergence, over an approximated $\mathcal{F}'$. The argument takes several steps:

3.3.1 | Marginal Matching

In this section, purely for simplicity, suppose the vector of summary statistics s consists of r quantiles of the data vectors. Assuming that the problem in (5) is feasible the optimal weights $\hat{w}(\theta)$ satisfy the constraints:

$$\hat{w}(\theta) \in \Delta_{m-1} \text{ and } \sum_{i=1}^{m} \hat{w}_i(\theta)(s_i - s_o) = 0$$

By our construction, the empirical estimate of the conditional joint distribution of the random vector (s_1, s_o) given θ can be obtained as:

$$\hat{F}_m(t_1, t_o|\theta) = \sum_{i=1}^{m} \hat{w}_i(\theta) 1_{\{(s_i, s_o) \le (t_1, t_o)\}}$$

We first verify that the condition (a) in the definition of $\mathcal{F}'$ is approximately satisfied. Note that the constraints imply that:

$$\int s_1 d\hat{F}_m(t_1, t_o|\theta) = s_o$$

That is the conditional joint distribution is estimated by matching s_o with the marginal conditional expectation of s_1 given θ.

The concept of matching the expected quantiles with the observed is the key behind the goodness-of-fit plots like the Q–Q plots, probability plots, and so forth. If the match is close, the densities of the corresponding *random variables* are approximately equal. Following the same argument, the proposed empirical likelihood-based procedure computes the estimate $\hat{F}_m$ by approximately equating the conditional marginal densities of the observed data X_o and the replicated data X_1 given the input parameter value θ. Now since the summary statistics, s (in this case r quantiles) are deterministic functions of the data, consequently, the conditional marginal densities of s_1 and s_o given θ would be approximately equal. That is, the condition (a) in the definition of the set $\mathcal{F}'$ is approximately satisfied.

3.3.2 | Cross-Entropy

The proposed empirical likelihood based method maximizes the sample version of the Kullback–Leibler divergence in (12). Note that, the empirical estimate of the marginal distribution of s_1 given θ can be obtained as:

$$\hat{F}_m(t_1|\theta) = \frac{1}{m}\sum_{i=1}^{m} 1_{\{t_i<t_1\}}$$

Thus the sample version of the cross-entropy term in (12) is given by

$$\frac{1}{m}\sum_{i=1}^{m}\log(w_i(\theta)) - \log\left(\sum_{i=1}^{m} w_i(\theta)\right)$$

In view of the constraint that $\hat{w}(\theta) \in \Delta_{m-1}$, this justifies the objective function $m^{-1}\sum_{i=1}^{m}\log w_i$, that is maximized in (5). That is, the proposed empirical likelihood maximizes the sample version of the required cross-entropy in (12).

3.3.3 | Approximation of $\mathcal{F}'$

Other than the constraints which defines $\mathcal{F}'$, in the proposed method $\hat{F}_m$ is computed with minimal restrictions. For any θ, the maximum of $m^{-1}\sum_{i=1}^{m}\log w_i(\theta)$ is finite when there exist a $w \in \mathcal{W}_\theta$, such that $w_i(\theta) > 0$, for all $i = 1, 2, \ldots, m$. This is equivalent to maximizing the divergence in (12) over all joint densities $f(\theta, s_1, s_o) \in \mathcal{F}'$ such that, for all $i = 1, 2, \ldots, m$, each observation (s_i, s_o) is in the support of the conditional density $f(s_1, s_o|\theta)$. The proposed empirical likelihood-based method thus maximizes the sample version of (12) over a large flexible set of non-parametrically specified distributions approximately satisfying the constraints which define $\mathcal{F}'$. No parameter, tuning or otherwise need to be specified or estimated (as in e.g., An, Nott, and Drovandi [22]).

The above argument can be generalized for summary statistics which approximately specify the density of the underlying random variable. Such summaries have been rigorously studied in statistics. Other than the quantiles, moments, up crossing proportions, and so forth can be used. We discuss various choices for the summary statistics in Section 5 below.

4 | Properties of the ABC Empirical Likelihood Posterior

The asymptotic properties of conventional ABC methods have been a topic of much recent research [13–15]. Here we investigate some basic asymptotic properties of our proposed empirical likelihood method. The proofs of the results are deferred to the Supporting Information.

Following Owen [26] the weights in (5) can be expressed as $\hat{w}_i = \left\{m\left(1+\hat{\lambda}^T h_i\right)\right\}^{-1}$, where $\hat{\lambda}$ is obtained by solving the equation $\sum_{i=1}^{m} h_i/\left(1+\hat{\lambda}^T h_i\right) = 0$.

4.1 | Posterior Consistency

In what follows below, we consider limits as n and $m = m(n)$ grow unbounded. Furthermore, for convenience, we make the dependencies of X_o and $X_1, X_2, \ldots, X_m \in \mathbb{R}^n$ on sample size n as well as parameter θ explicit. In what follows, a sequence of events $\{E_n, n \geq 1\}$ is said to occur with high probability, if $P(E_n) \to 1$ as $n \to \infty$.

Suppose that we define $h_i^{(n)}(\theta) = \left\{s\left(X_i^{(n)}(\theta)\right) - s\left(X_o^{(n)}(\theta_o)\right)\right\}$, and assume that $E^0_{s\left(X_i^{(n)}(\theta)\right)|\theta}\left[s\left(X_i^{(n)}(\theta)\right)\right]$ is finite so that we can write $s\left(X_i^{(n)}(\theta)\right) = E^0_{s\left(X_i^{(n)}(\theta)\right)|\theta}\left[s\left(X_i^{(n)}(\theta)\right)\right] + \xi_i^{(n)}(\theta) = \mathfrak{s}^{(n)}(\theta) + \xi_i^{(n)}(\theta)$, where $E^0_{s\left(X_i^{(n)}(\theta)\right)|\theta}\left[\xi_i^{(n)}(\theta)\right] = 0$ for all i, n and θ.

We make the following assumptions:

A1. (Identifiability and convergence). There is a sequence of positive increasing real numbers $b_n \to \infty$, such that, $\mathfrak{s}^{(n)}(\theta) = b_n\{\mathfrak{s}(\theta) + o(1)\}$, where $\mathfrak{s}(\theta)$ is a one-to-one function of θ that does not depend on n. Furthermore, $\mathfrak{s}(\theta)$ is continuous at θ_o and for each $\epsilon > 0$, and for all $\theta \in \Theta$, there exists $\delta > 0$, such that whenever $\|\theta - \theta_o\| > \epsilon$, $\|\mathfrak{s}(\theta) - \mathfrak{s}(\theta_o)\| > \delta$.

A2. (Feasibility). For each θ, n and $i = o, 1, \ldots, m(n)$, the vectors $\xi_i^{(n)}(\theta)$ are identically distributed, supported over the whole space, and their distribution puts positive mass on every orthant, $\mathcal{O}_u$ of $\mathbb{R}^r$, $u = 1, 2, \ldots, 2^r$. Furthermore, for every orthant $\mathcal{O}_u$, as $n \to \infty$, $\sup_{\left\{i:\xi_i^{(n)}(\theta)\in\mathcal{O}_u\right\}}\|\xi_i^{(n)}(\theta)\| \to \infty$ in probability, uniformly in θ.

A3. (Growth of extrema of errors). As $n \to \infty$, $\sup_{i\in\{o,1,2,\ldots,m(n)\}}\|\xi_i^{(n)}(\theta)\|b_n^{-1} \to 0$ in probability, uniformly in $\theta \in \Theta$.

Assumption (A1) ensures identifiability and additionally implies that $\mathfrak{s}^{(n)}(\theta)/b_n - \mathfrak{s}(\theta)$ converges to zero uniformly in θ. Assumption (A2) is important for ensuring that with high probability the empirical likelihood ABC posterior is a valid probability measure for n large enough. Assumptions (A2) and (A3) also link the number of simulations m to n and ensure concentration of the posterior with increasing n. The proofs of the results below are given in the Appendix B. The main result, Theorem 1, shows posterior consistency for the proposed empirical likelihood method.

Let $l_n(\theta) := \exp\left(\sum_{i=1}^{m(n)}\log(\hat{w}_i(\theta))/m(n)\right)$ and for each n, we define.

$\Theta_n\left\{\theta : \|\mathfrak{s}(\theta) - \mathfrak{s}(\theta_o)\| \leq b_n^{-1}\right\}$. By continuity of $\mathfrak{s}$ at θ_0, Θ_n is nonempty for each n. Furthermore, since b_n is increasing in n, Θ_n is a decreasing sequence of sets in n.

Lemma 1. *Under Assumptions (A1)–(A3), with high probability, the likelihood $l_n(\theta) > 0$ for all $\theta \in \Theta_n$.*

Lemma 1 shows that for large n the estimated likelihood is strictly positive in a neighborhood of θ_0. Next, we show that the empirical likelihood is zero outside certain neighborhood of θ_0.

Lemma 2. *Under Assumptions (A1)–(A3), for every $\epsilon > 0$, let $B(\theta_0, \epsilon)$ be the open ball of radius ϵ centered at θ_o. The empirical likelihood is zero outside $B(\theta_0, \epsilon)$, with high probability.*

Now suppose we choose $\epsilon = b_1^{-1}$ and $n > n(b_1^{-1})$ such that $l_n(\theta)$ is positive on Θ_n with high probability. Furthermore, for all n and for all $\theta \in \Theta_n$, $\min_{i\neq j} \|s(X_j(\theta)) - s(X_i(\theta))\| > 0$ with probability 1, which implies for an appropriate choice of k, (see the Supporting Information) the estimate of the differential entropy $|\hat{H}^{0(n)}_{s|\theta}(\theta)| < \infty$ with probability 1 as well. This proves that for large values of n, with high probability: $\int_{\theta\in\Theta} l_n(\theta)e^{\hat{H}^{0(n)}_{s|\theta}(\theta)}\pi(\theta)d\theta \geq \int_{\theta\in\Theta_n} l_n(\theta)e^{\hat{H}^{0(n)}_{s|\theta}(\theta)}\pi(\theta)d\theta > 0$, and $\hat{\Pi}_n(\theta|s(X_o(\theta_o))) = \left(l_n(\theta)e^{\hat{H}^{0(n)}_{s|\theta}(\theta)}\pi(\theta)\right) / \int_{t\in\Theta} l_n(t)e^{\hat{H}^{0(n)}_{s|t}(t)}\pi(t)dt$ is a valid probability measure (with high probability). The main result, Theorem 1 below, establishes posterior consistency.

Theorem 4. *As $n \to \infty$, $\hat{\Pi}_n(\theta|s(X_o(\theta_o)))$ converges in probability to δ_{θ_o}, where δ_{θ_0} is the degenerate probability measure supported at θ_0.*

4.2 | Behavior of the Proposed Posterior With Growing Number of Replications

We now discuss how the proposed ABCel posterior behaves with fixed sample size n and observed summary and growing m. Our primary goal is to find appropriate number of replicates, that is, m for a fixed sample size n. We also discuss the bias-variance trade-off as observed in Figure 3 in more details.

Under the setup of fixed n and the observed summary, it is more appropriate to consider expectation of $h_i^{(n)}(\theta)$ conditional on $\left(\theta, s\left(X_o^{(n)}(\theta_o)\right)\right)$. Since each $X_i^{(n)}(\theta)$ is conditionally independent of $X_o^{(n)}(\theta_o)$ given θ, for each $i = 1, 2, \ldots, m$, and $\theta \in \Theta$ we get:

$$E^0_{s\left(X_i^{(n)}(\theta)\right)|\left(\theta, s\left(X_o^{(n)}(\theta_o)\right)\right)}\left[h_i^{(n)}(\theta)\right] = E^0_{s\left(X_i^{(n)}(\theta)\right)|\theta}\left[s\left(X_i^{(n)}(\theta)\right)\right] - s\left(X_o^{(n)}(\theta_o)\right) \neq 0 \text{ a.e}$$

That is, for fixed n, after conditioning on $s\left(X_o^{(n)}(\theta_o)\right)$, the constraints in the problem (5) $h_i^{(n)}(\theta), i = 1, 2, \ldots, m$ are misspecified for all $\theta \in \Theta$ almost everywhere (even when $\theta = \theta_o$). The constrained optimization problem in (5), however, could still be feasible and the resulting estimated posterior could be positive. The properties of empirical likelihood under misspecified but feasible constraint have been studied by Ghosh, Chaudhuri, Gangopadhyay [31]. We now evoke their results.

Using the notations introduced above, when $r = 1$, that is, there is only one constraint present, under conditions similar to those described above, it can be shown that, [31, Theorem 3.4] for any $\theta \in \Theta$:

$$\begin{aligned} l_m(\theta) &:= \frac{1}{m}\sum_{i=1}^{m} \log(\hat{w}(\theta)) \\ &= -\frac{1}{\mathcal{M}_m(\theta)}\left|E_{s\left(X_1^{(n)}(\theta)\right)|\theta}\left[s\left(X_1^{(n)}(\theta)\right)\right] - \left[s\left(X_o^{(n)}(\theta_o)\right)\right]\right| \\ &\quad \times \left(1 + o_p(1)\right), \\ &= -\frac{b_n}{\mathcal{M}_m(\theta)}\left|\left(\mathfrak{s}(\theta) - \mathfrak{s}(\theta_o) + o(1)\right) - \frac{\xi_o^{(n)}(\theta_o)}{b_n}\right|\left(1 + o_p(1)\right) \end{aligned} \tag{13}$$

where $\mathcal{M}_m(\theta)$ is a non-random $o(m)$ sequence such that, as $m \to \infty$, $\mathcal{M}_m \to \infty$ and both $\mathcal{M}_m^{-1}(\theta)\max_{1\leq i\leq m}\left|\xi_i^{(n)}(\theta)\right|1_{\left\{\xi_i^{(n)}(\theta)>0\right\}} = 1 + o_p(1)$, and $\mathcal{M}_m^{-1}(\theta)\max_{1\leq i\leq m}\left|\xi_i^{(n)}(\theta)\right|1_{\left\{\xi_i^{(n)}(\theta)<0\right\}} = 1 + o_p(1)$ are satisfied.

The sequence $\mathcal{M}_m(\theta)$ is the rate at which the maximum of the $s(X_i(\theta))$ grows away from its mean. The above conditions are easily satisfied. As for example, when $\xi_o^{(n)}(\theta_0)$ is a $N(0, \sigma_0^2)$ random variable, $\mathcal{M}_m \sim \sigma_0\sqrt{2\log m}$.

In the rest of this section, we assume that $r = 1$. Using the results from Ghosh, Chaudhuri, Gangopadhyay [31] it is possible to specify bounds on the rate of growth of the number of replicates with the sample size. Since the differential entropy plays a relatively minor role in determining the posterior, in what follows we assume that for each θ, the estimate of the differential entropy remains bounded, and focus on $l_m(\theta)$. Furthermore, for brevity, we present the results as *Remarks* below. More details are available in the Supporting Information.

4.2.1 | Bounds on the Growth of the Number of Replications in Terms of Sample Size

We first consider the bounds of the replication size m in terms of sample size n. Our results follow from various advantageous properties of the posterior. For the purposes of easier description and illustration, we would sometime assume that the errors $\xi_o^{(n)}$ follow a $N(0, \sigma_o^2)$ distribution.

The posterior is Bayesian consistent [13] if with high probability two things happen:

1. $\exp(l_m(\theta))$ would converge to zero for all $\theta \neq \theta_o$, and
2. $\exp(l_m(\theta_o))$ would not collapse to zero.

Remark 1. In order to ensure the first condition it is enough to choose m and n such that $b_n/\mathcal{M}_m(\theta)$ diverges. An upper bound of the rate of growth of m can thus be obtained by inverting the relation $b_n > \mathcal{M}_m(\theta)$. Depending on the distribution of $\xi_o^{(n)}$, m can be much larger than n. For example, if $\xi_o^{(n)}$ follows a normal distribution with mean zero and variance σ_o^2, $b_n = \sqrt{n}$ and $\mathcal{M}_m(\theta)$ is of the order $\sigma_o\sqrt{2\log(m)}$, which allows an upper bound of m as large as $\exp\left(n/\left(2\sigma_o^2\right)\right)$.

Remark 2. A more accurate relationship can be obtained from the second condition. The condition implies that there exists a constant $C_1 > 0$ such that, $l_m(\theta_o) > -C_1$ with a high probability. Assuming that $\xi_o^{(n)}(\theta_o)$ is a $N(0, \sigma_o^2)$ random variable, from (13), it follows that $Pr[l_m(\theta_o) \leq -C_1] \leq \exp\left(-C_1^2\log m\right) = m^{-C_1^2}$. Now if we fix the rate of reduction of the above probability to $n^{-\alpha}$ for some α, we get $m = n^{\alpha/C_1^2}$.

Other bounds can be found by controlling the rate at which the probability of a Type I error for testing the null hypothesis $\theta = \theta_o$ against the unrestricted alternative decrease to zero. By construction $l_m(\theta)$ is different from the traditional empirical likelihood, so this problem is of broad interest.

Remark 3. The log-likelihood ratio $\log LR(\theta_o)$ turns out to be $l_m(\theta_o) + \log m$. The test rejects H_0 if $\log LR(\theta_o)$ is smaller than $\log C_0$, for some pre-specified $C_0 \in (0,1)$. Assuming that, $\xi^{(n)}(\theta)$ is a $N(0, \sigma_0^2)$ random variable, it follows that, the probability of rejecting the null hypothesis is given by (see the Supporting Information):

$$Pr[\log m + l_m(\theta_o) \leq \log C_0] \leq exp\{-(\log m)^3 + 2(\log m)^2 \log C_0 - (\log m)(\log C_0)^2\}$$

Now ensuring that the probability of rejecting the null hypothesis reduces at the rate of p_n, we get $p_n = exp\{-(\log m)^3\}$, which implies the number of replications $m = exp\left\{(-\log p_n)^{1/3}\right\}$.

4.2.2 | Behavior of the Log-Likelihood When $\mathcal{M}_m(\theta)/b_n$ Diverges

This scenario includes the situation when the sample size n is fixed and the number of replication m grows. We discuss the bias-variance trade-off or the flattening of the approximate likelihood as observed in Figure 1.

Remark 4. Let us fix $\theta \neq \theta_o$ and suppose $\xi_o^{(n)}(\theta_o)$ follows a $N(0, \sigma_o^2)$ distribution. For a fixed $C_2 > 0$, it can be shown that (see the Supporting Information):

$$Pr[l_m(\theta) \leq -C_2] \leq \left(\frac{1}{m}\right)^{\left\{C_2 - \frac{b_n}{\sigma_o\sqrt{2\log m}}\left|\tilde{s}(\theta) - \tilde{s}(\theta_o)\right|\right\}^2} \tag{14}$$

Now, if $\mathcal{M}_m(\theta)/b_n = \sigma_o\sqrt{2\log m}/b_n$ diverges with m and n, clearly, for large values of m and n, $Pr[l_m(\theta) \leq -C_2] \approx m^{-C_2^2}$. That is, for any fixed $C_2 > 0$ and $\theta \neq \theta_o$, $l_m(\theta) \geq -C_2$ with a high probability. Furthermore, for a fixed n, R.H.S. of (14) is a decreasing function in m. That is if the sample size is kept fixed, increasing the number of replications will increase the probability of $l_m(\theta) \geq -C_2$. As a result, the log likelihood will become increasingly flat in shape. This is exactly the phenomenon that was observed in Figure 1. Remark 4 provides actual justification to our observation.

Statistical intuition mandates with an increase in m, we should increase the number of summary statistics. Remark 4 does not apply to such situations. We present evidence in favor of our intuition in Example 6.1 below.

5 | Choice of Summary Statistics

A judicious choice of summary statistics is crucial for a good performance of any ABC procedure [3]. The proposed method does not necessarily require summaries that are sufficient for the parameter, which according to many authors (e.g., Frazier et al. [13]; Robert [18]) are usually not available. Rather from the arguments in Section 3, it mandates an use of summaries which approximately define the density of $X_i(\theta)$, for $i \in \mathbb{M}_o$.

Sample quantiles, extreme values, or proportion of samples exceeding the certain pre-specified thresholds that directly put constraints on the data density (see D'Agostino and Stephens [46]) can be used. Moreover, moments, if they exist, may under certain conditions (e.g., Carleman's condition) specify a density (see Gut [47]). Thus, moments can be used as summaries in many cases as well.

For complex data models, with dependent components, marginal summaries may not be adequate. In such cases, constraints can be based on joint moments, joint quantiles or joint exceedances, and so forth can be used. Other than these generic choices, one can base the constraints on the functionals of transformed variables. Since a density is a one-to-one function of its characteristic function, for dependent data sets, constraints based on the smoothed spectral density of the data can be used. For example, in the case of stochastic processes, summaries based on the exceedance proportions of log-amplitudes, which actually put constraints on the auto-covariance function of the process, are often beneficial (see Section 6.3 below).

In our experience, often moments work the best. A judicious mix and match of various forms of summaries decided after some inspection of the summaries of observed data are required. It should also be recognized that summaries with widely different scales or an ill-conditioned covariance matrix may lead to a poor estimate of differential entropy and subsequently to slow mixing of the Markov Chain Monte Carlo procedures.

Finally, the number of summaries required would depend partly on their nature, partly on the number of replications m, and partly on the sample size n (see (14)). Even though some judgments is required, evidence shows (see Section 6) for any given problem, appropriate summaries can be found without much effort.

6 | Illustrative Examples and Applications

We illustrate the utility of the ABCel method with four examples involving data simulated from a standard Gaussian model, an ARCH(1) model (also considered in Mengersen, Pudlo, and Robert [5]), The simple recruitment, boom and bust model [22], and a real life example modeled as an elliptical inclusion model respectively. Here in order to address dependence we use non-Gaussian summaries based on auto-covariance function and the periodogram of the data. We also present a real application based on stereological extremes [48]. More examples on the traditional g-and-k model and an application to Erdös–Renyi random graphs can be found in the Supporting Information.

6.1 | Normal Distribution

Our first example considers inference about a mean μ for a random sample of size $n = 100$ from a normal density, $N(\mu, 1)$. The prior for μ is $N(0,1)$. The observed data X_o is generated with $\mu = 0$. The exact posterior for μ is normal, $N\left(\sum_{j=1}^n X_{oj}/(n+1), (n+1)^{-1}\right)$. The proposed empirical likelihood based method was implemented with $m = 25$. We considered several choices of constraint functions $s^{(1)}$, ..., $s^{(r)}$. Specifically, for $i = o, 1, \ldots, m$, we take (a) $s^{(1)}(X_i(\theta)) =$

$\overline{X}_{i\cdot} = \sum_{j=1}^{n} X_{ij}(\theta)/n$, (b) $s^{(2)}(X_i(\theta)) = \sum_{j=1}^{n}\left(X_{ij}(\theta) - \overline{X}_{i\cdot}\right)^2/n$, (c) $s^{(3)}(X_i(\theta)) = \sum_{j=1}^{n}\left(X_{ij}(\theta) - \overline{X}_{i\cdot}\right)^3/n$, (d) $s^{(4)}(X_i(\theta)) =$ median of $X_i(\theta)$, (e) $s^{(5)}(X_i(\theta)) =$ first quartile of $X_i(\theta)$, and (f) $s^{(6)}(X_i(\theta)) =$ third quartile of $X_i(\theta)$. Here the constrains considered use the first three central moments (a–c) and the three quartiles (d–f). Combinations of these constraints are considered within the empirical likelihood procedure.

The posteriors obtained from our proposed empirical likelihood-based ABC method with the above summaries are close to the true posterior. An illustrative example, with sample mean as a summary, is presented in Figure 2. Here, the true posterior density, that is, the dashed line, is quite close to the histogram of the samples drawn from the posterior obtained from the proposed method.

In order to compare the performance of the proposed procedure for different choices of the summary statistics, we consider frequentist coverages and the average lengths of the 95% credible intervals. The results are presented in Table 1. The coverages are based on 100 repeats of the procedure. For each repetition, MCMC approximations to the posterior are based on 50,000 samples with 50,000 iterations discarded as burn-in.

As we have shown before (see Figure 1 and Remark 4) the approximate posterior gets flatter if we keep the number of summary statistics fixed and increase the number of replications m. That is, with increasing m, one should increase the number of summary statistics used. The same argument mandates that when we increase the number of summary statistics we should also increase the number of replications. In Table 1 we report the value of m for which the Monte Carlo frequentist coverages were close to the nominal value of 95%.

From Table 1, it is clear that the proposed method performs quite well for various sets of summary statistics. For mean and the median the frequentist coverage is matches exactly the nominal value. Note that the sample mean is minimal sufficient for μ, and would be an ideal choice of summary statistic in conventional likelihood-free procedures such as ABC. However, median is not sufficient for the mean, but still produces the exact coverage. Table 1 also shows that when multiple summary statistics are

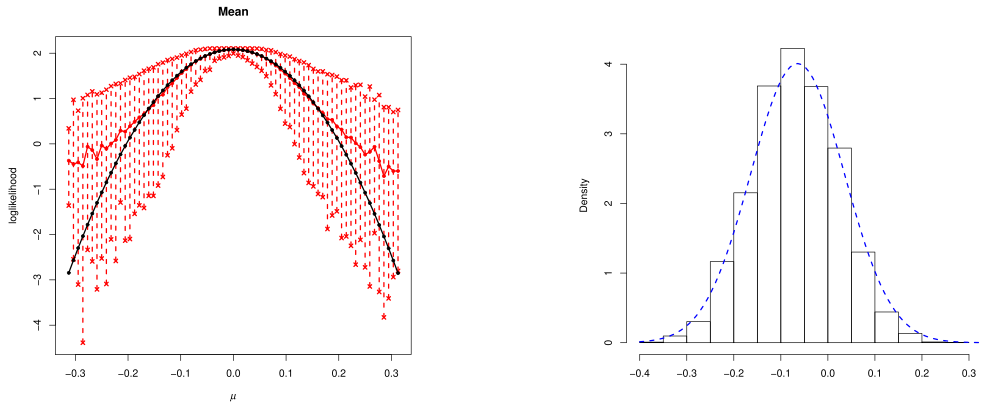


FIGURE 2 | Comparison of the true posterior of the mean of a Normal distribution with unit variance conditional on the sample mean with our proposed empirical likelihood based ABC posterior. Here $n = 100$ and $m = 25$. (A) The figure directly compares the true log-posterior (black curve) with the means and 95% credible intervals of the proposed approximate posterior based on 1000 replications for each parameter value (in red). (B) The figure compares the true posterior (dashed line) with the histogram of the samples drawn from the proposed empirical likelihood based ABC posterior (underlying histogram).

TABLE 1 | The coverage and the average length of 95% credible intervals for μ for various choices of constraint functions when $\mu = 0$ and $n = 100$.

Constraint functions	m	Coverage	Average length
Mean, (a).	25	0.95	0.360
Median, (e).	25	0.95	0.446
First two central moments, (a), (b).	40	0.94	0.331
Mean and Median, (a), (e).	40	0.94	0.330
First three central moments, (a), (b), (c).	70	0.91	0.307
Three quartiles, (e), (f), (g).	75	0.93	0.329

Note: The coverage for the true posterior is 0.95 and average length is 0.39 (2 d.p.).

used, by increasing the number of replicates it is possible to obtain an approximate posterior with frequentist coverage close to the nominal value of 95%.

6.2 | An ARCH(1) Model

We now present examples where summary statistics are not close to normal, so that the assumptions behind the synthetic likelihood are not satisfied. We consider an autoregressive conditional heteroskedastic or ARCH(1) model, where for each $i = o, 1, 2, \ldots, m$, the components $X_{i1}(\theta), X_{i2}(\theta), \ldots, X_{in}(\theta)$ are dependent for all $\theta = (\alpha_0, \alpha_1) \in \Theta$. This model was also considered in Mengersen, Pudlo, and Robert [5]. For each i, the time series $X_{ij1 \le j \le n}$ is generated by $X_{ij}(\theta) = \sigma_{ij}(\theta)\varepsilon_{ij}$, $\sigma^2_{ij}(\theta) = \alpha_0 + \alpha_1 X^2_{i(j-1)}(\theta)$, where ε_{ij} are i.i.d. $N(0,1)$ random variables, $\alpha_0 > 0$, and $0 < \alpha_1 < 1$. We assume a uniform prior over $(0,5) \times (0,1)$ for (α_0, α_1).

Our summary statistics include the three quartiles of the absolute values of the data. Since the data is dependent we also use the following summary statistic. Let, for a fixed i and for each j, $Y_{ij}(\theta) = X^2_{ij}(\theta) - \sum_{j=1}^{n} X^2_{ij}(\theta)/n$. Then for each $i = 1, 2, \ldots, m$, we define,

$$s^{(4)}(X_i(\theta)) = \frac{1}{n}\sum_{j=2}^{n}\left(1_{\{(Y_{ij}(\theta)\cdot Y_{i(j-1)}(\theta))\ge 0\}} - 1_{\{(Y_{ij}(\theta)\cdot Y_{i(j-1)}(\theta))<0\}}\right)$$

That is, $s^{(4)}$ is the difference between the proportion of the concordant and that of the discordant pairs between series Y_i with its lag-1 version. Empirical evidence suggests that s_4 performs better than the usual lag-1 auto-covariance of the series X_i^2. The quartiles of the absolute values of the data provide some information about the marginal distribution.

Our observed data were of size $n = 1000$, with $(\alpha_0, \alpha_1) = (3, 0.75)$ and we used $m = 50$ replicates for each likelihood approximation for both empirical and synthetic likelihoods in Bayesian computations. Marginal posterior densities were estimated for the parameters based on 50,000 sampling iterations with 50,000 iterations burn-in for both the synthetic likelihood and proposed empirical likelihood. We compare these methods with the posterior obtained using rejection ABC with 1,000,000 samples, a tolerance of 0.0025, and linear regression adjustment. The estimated marginal posterior densities in Figure 3 for the proposed method are quite close to those obtained from the rejection ABC. The synthetic likelihood produces quite different marginal posterior densities, especially for α_1. In this example the s_4 statistic is highly non-Gaussian, so the assumptions of the synthetic likelihood are not satisfied.

6.3 | The Simple Recruitment, Boom and Bust Model

The simple recruitment, boom and bust model is a discrete stochastic temporal model, primarily used to explain fluctuations in species population over time. The dynamics is controlled by the parameter vector $\theta = (r, \kappa, \alpha, \beta)$. For $i = o, 1, 2, \ldots, m$,

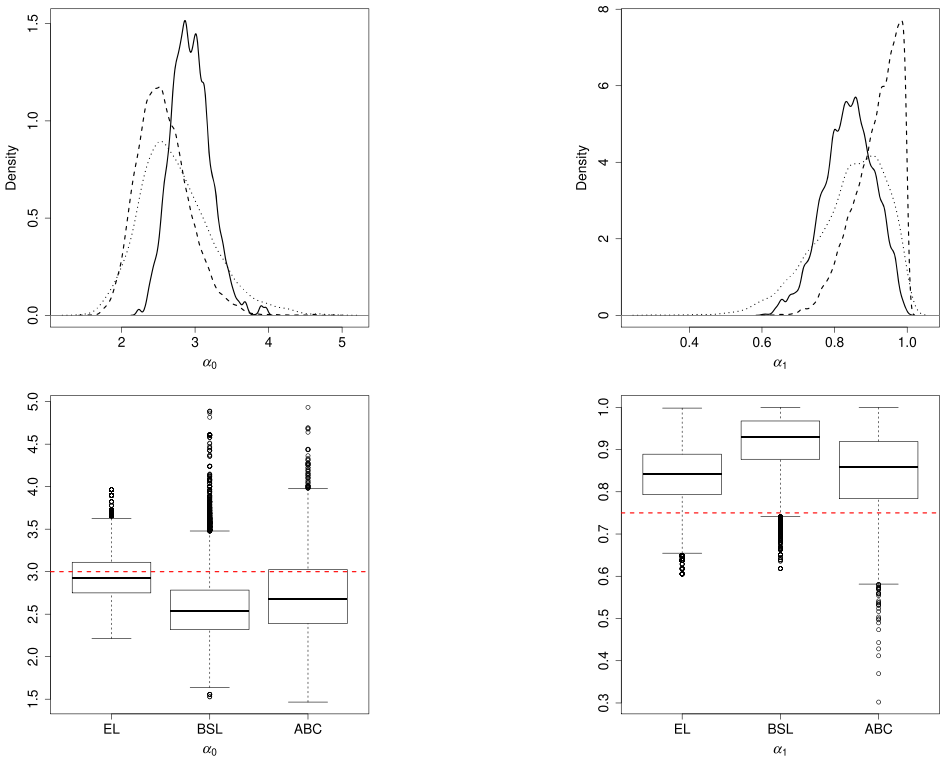


FIGURE 3 | Estimated marginal posterior densities of parameters α_0 and α_1 in the ARCH(1) model. The top row shows kernel density estimates (empirical likelihood ABC (solid), synthetic likelihood (dashed), rejection ABC (dotted)), while the bottom row shows boxplots of posterior samples. In the boxplots, the horizontal dotted lines show the true parameter values.

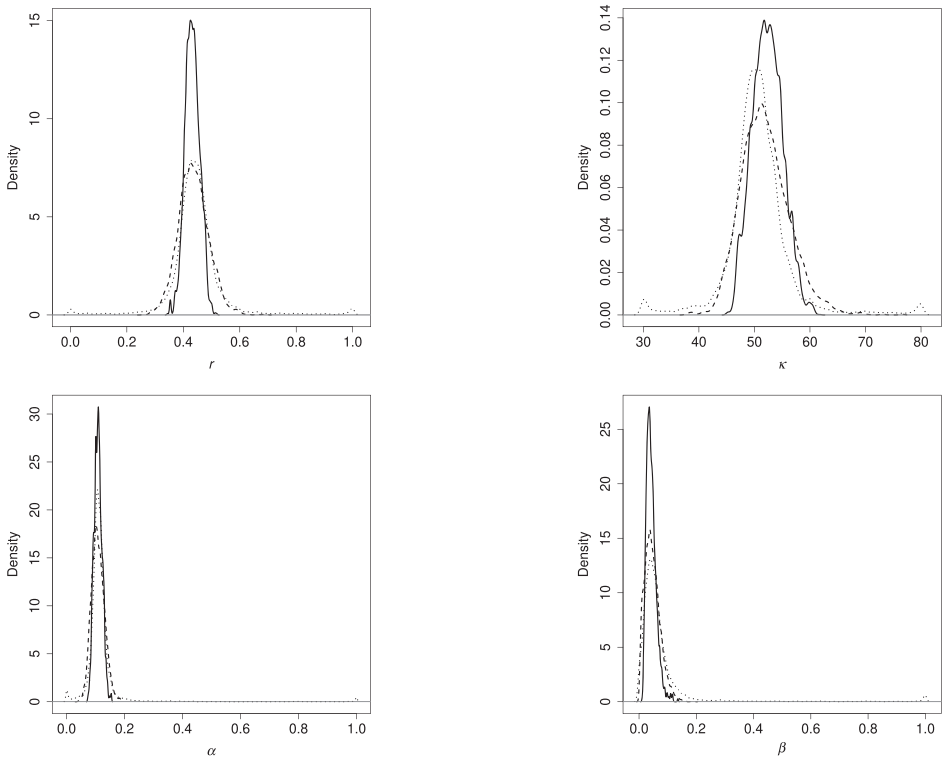


FIGURE 4 | Estimated marginal posterior densities of parameters r, κ, α, β in the boom and bust model. The figures show kernel density estimates (empirical likelihood ABC (solid), synthetic likelihood (dashed), rejection ABC (dotted)).

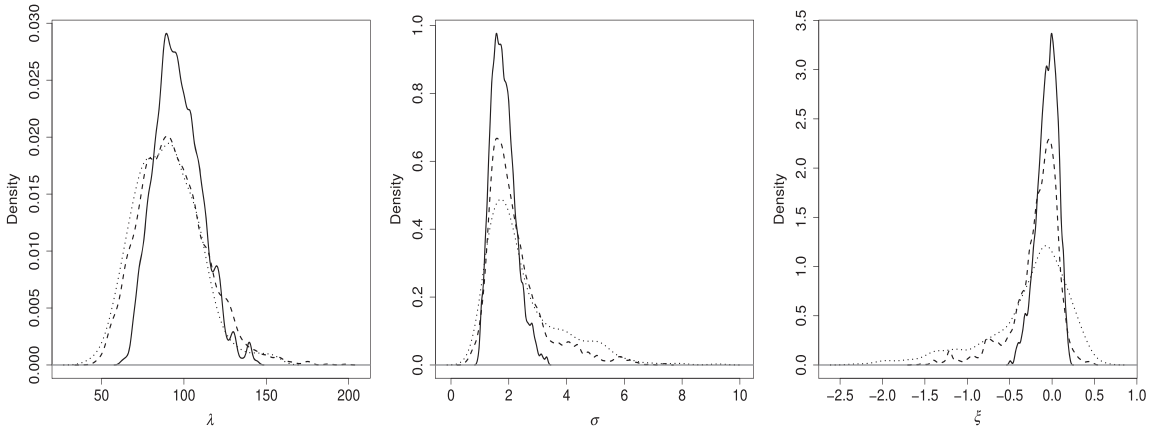


FIGURE 5 | Estimated marginal posterior densities of λ, σ and ξ using empirical likelihood ABC (solid), rejection ABC (dotted), and synthetic likelihood (dashed).

given $X_{ij}(\theta) = x_{ij}$, the distribution of $X_{i(j+1)}(\theta)$ is given by:

$$X_{i(j+1)}(\theta) \sim \begin{cases} Poisson\big(x_{ij}(1+r)\big) + \epsilon_j & \text{if } x_{ij} \le \kappa \\ Binomial\big(x_{ij}, \alpha\big) + \epsilon_j & \text{if } x_{ij} > \kappa \end{cases}$$

Here $\epsilon_j \sim Poisson(\beta)$ distribution. The sample paths rapidly cycle between the large and small non-negative integers.

For our simulation study we follow An, Nott, and Drovandi [22] and set $\theta_o = (0.4, 50, 0.09, 0.05)$, and assume a prior of $U(0,1) \times U(30,80) \times U(0,1) \times U(0,1)$ on θ. We generated observations of length $n = 200$, after discarding the first 50 values to remove the transient phase of the process.

For each θ we generated $m = 40$ replications from the model. The summary statistics used were, (a) the proportion of observations in the interval $(0, 15)$, (b) the proportion of differences $X_{ij}(\theta) - X_{i(j-1)}(\theta)$ strictly larger than 2, and (c) the proportion of log-amplitudes of the smoothed periodogram of the data lying in the interval (5.120,6.278). The intervals were chosen rather judiciously partly based on the observed data X_o.

Our choice of summary statistics is targeted toward specifying the underlying data density rather than any particular parameter. Clearly, a process can be specified by its probabilities of exceedance of certain thresholds. The use of lagged differences is also natural for the same reason. The periodogram of the process is connected to its auto-covariance function. Thus, the exceedance probabilities of its log amplitude should put constraints on the auto-covariances between the successive observations. Note that, none of the summary statistics are normally distributed in the case.

The density plots of observations sampled from the proposed ABCel (solid), synthetic likelihood (dashed), and the rejection ABC with a ridge regression adjustment with tolerance 0.001 (dotted) are presented in Figure 4. From the plot, it is clear that the rejection ABC has very heavy tails, which essentially cover the whole of the support of the priors and do not change if different tolerances or rejection methods are used. The synthetic ABC is not expected to work well in this case. However, they seem to show a lighter tail than the Rejection ABC. Among the three, the proposed ABCel posterior seems to be the most concentrated around true parameter values and seems to approximate the true posterior well.

6.4 | Stereological Data

Next, we consider an example concerning the modeling of diameters of inclusions (microscopic particles introduced in the steel production process) measured from planar cross-sections in a block of steel. The size of the largest inclusion in a block is thought to be important for steel strength. We focus on an elliptical inclusion model due to Bortot, Coles, and Sisson [49] here, which is an extension of the spherical model studied by Anderson and Coles [48]. Unlike the latter, the elliptical model does not have a tractable likelihood.

It is assumed that the inclusion centers follow a homogeneous Poisson process with a rate λ. For each inclusion, the three principal diameters of the ellipse are assumed independent of each other and of the process of inclusion centers. Given V, the largest diameter for a given inclusion, the two other principal diameters are determined by multiplying V with two independent uniform $U[0,1]$ random variables. The diameter V, conditional on exceeding a threshold value v_0 (5 μm in Bortot, Coles, and Sisson [49]) is assumed to follow a generalized Pareto distribution:

$$pr(V \leq v | V > v_0) = 1 - \left\{ 1 + \frac{\xi(v - v_0)}{\sigma} \right\}_+^{-\frac{1}{\xi}}$$

The parameters of the model are given by $\theta = (\lambda, \sigma, \xi)$. We assume independent uniform priors with ranges $(1, 200)$, $(0, 10)$ and $(-5, 5)$ respectively. A detailed implementation of ABC for this example is discussed in Erhardt and Sisson [50].

The observed data has 112 entries, measuring the largest principal diameters of elliptical cross-sections of inclusions for a planar slice. The number of inclusions L in each dataset generated from the model is random. The summary statistics used are (a) $(L - 112)/100$, (b) the mean and (c) the median of the observed planar measurements, and (d) the proportion of planar measurements less than or equal to six (approximately the median for the observed data).

Using the summary statistics described above, we compare the proposed empirical likelihood-based method with the synthetic likelihood ($m = 25$ for both) and a rejection ABC algorithm with small tolerance (0.00005) and linear regression adjustment. The resulting estimated marginal posterior densities for λ, σ, ξ are shown in Figure 5. The results for the proposed empirical likelihood-based method are more concentrated than the rejection ABC or the synthetic likelihood both of whom exhibit quite long tails. The chosen summaries mixed faster than those used in Pham, Nott, and Chaudhuri [51] and were comparable in speed to the synthetic likelihood.

7 | Discussion

This article develops a new empirical likelihood-based easy-to-use approach to the ABC paradigm called ABCel. For its implementation, the method only requires a set of summary statistics, their observed values, and the ability to simulate these summary statistics from a given black box or a suitable auxiliary model. We first use a direct information projection to derive an analytic form for an approximation of the target posterior. Using this analytic expression, the best approximation to the target posterior is then obtained from a reverse information projection. The procedure is implemented using a modified empirical likelihood. By construction, the proposed empirical likelihood estimates the joint distribution of the observed and replicated summaries by minimizing a cross-entropy over a large set of distributions. Furthermore, for appropriate summaries, at each value of the parameter, the above joint distribution is estimated by approximately equating the marginal densities of the observed and the replicated data. The construction does not require any specification of a distance function, a tolerance or a bandwidth. Neither does it assume any asymptotic distribution of the summary statistics. No constraints that are functions of the parameter and the data are required either. We explore the properties of the proposed posterior both analytically and empirically. The method is posterior consistent under reasonable conditions and shows good performance in simulated and real examples.

The modified empirical likelihood works with user-specified simple summaries like quantiles, moments or proportion of exceedance, that specify the underlying data density. Summaries based on the spectral density of the data can also be conveniently used. Even though no specific algorithm is so far available, our experience suggests appropriate simple summary statistics could easily be postulated from basic statistical considerations for almost all problems.

The number of replications depends in principle both on the number and the nature of the summary statistics used. We make recommendations on the relative magnitudes of the number of replications and the sample size.

Finally, empirical evidence as seen in the Q–Q plots in Sections F.2 and G of the Supporting Information, suggest

that under suitable conditions, the proposed posteriors would asymptotically converge to a normal density. The conditions under which such convergences would hold is a natural question for further investigation.

Author Contributions

Sanjay Chaudhuri: Conception, formulation, structural and theoretical derivations, implementation, simulation, writing. **Subhroshekhar Ghosh:** Theoretical derivation, wrting. **Kim Cuc Pham:** Some simulation.

Acknowledgements

Sanjay Chaudhuri was partially funded by the National Science Foundation grant DMS-2413491. Subhroshekhar Ghosh was supported in part by the MOE grants R-146-000-250-133, R-146-000-312-114, A-8002014-00-00, and MOE-T2EP20121-0013.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

References

1. M. A. Beaumont, W. Zhang, and D. J. Balding, "Approximate Bayesian Computation in Population Genetics," *Genetics* 162 (2002): 2025–2035.

2. M. G. B. Blum, M. A. Nunes, D. Prangle, and S. A. Sisson, "A Comparative Review of Dimension Reduction Methods in Approximate Bayesian Computation," *Statistical Science* 28 (2013): 189–208.

3. P. Fearnhead and D. Prangle, "Constructing Summary Statistics for Approximate Bayesian Computation: Semi-Automatic Approximate Bayesian Computation (With Discussion)," *Journal of the Royal Statistical Society, Series B* 74 (2012): 419–474.

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

5. K. L. Mengersen, P. Pudlo, and C. P. Robert, "Bayesian Computation via Empirical Likelihood," *Proceedings of the National Academy of Sciences* 110, no. 4 (2013): 1321–1326.

6. D. B. Rubin, "Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician," *Annals of Statistics* 12, no. 4 (1984): 1151–1172.

7. S. Tavaré, D. J. Balding, R. C. Griffiths, and P. Donnelly, "Inferring Coalescence Times From DNA Sequence Data," *Genetics* 145 (1997): 505–518.

8. S. N. Wood, "Statistical Inference for Noisy Nonlinear Ecological Dynamic Systems," *Nature* 466, no. 7310 (2010): 1102–1104.

9. C. Drovandi and D. T. Frazier, "A Comparison of Likelihood-Free Methods With and Without Summary Statistics," *Statistics and Computing* 32, no. 3 (2022): 42.

10. M. A. Beaumont, C. P. Robert, J.-M. Marin, and J. M. Corunet, "Adaptivity for ABC Algorithms: The ABC-PMC Scheme," *Biometrika* 96 (2009): 983–990.

11. P. Marjoram, J. Molitor, V. Plagnol, and S. Tavaré, "Markov Chain Monte Carlo Without Likelihoods," *Proceedings of the National Academy of Sciences of the USA* 100 (2003): 15324–15328.

12. S. A. Sisson, Y. Fan, and M. M. Tanaka, "Sequential Monte Carlo Without Likelihoods," *Proceedings of the National Academy of Sciences of the USA* 104 (2007): 1760–1765. Errata (2009), 106, 16889.

13. D. T. Frazier, G. M. Martin, C. P. Robert, and J. Rousseau, "Asymptotic Properties of Approximate Bayesian Computation," *Biometrika* 105, no. 3 (2018): 593–607.

14. W. Li and P. Fearnhead, "On the Asymptotic Efficiency of Approximate Bayesian Computation Estimators," *Biometrika* 105, no. 2 (2018b): 285–299.

15. W. Li and P. Fearnhead, "Convergence of Regression-Adjusted Approximate Bayesian Computation," *Biometrika* 105, no. 2 (2018a): 301–318.

16. J. W. Miller and D. B. Dunson, "Robust Bayesian Inference via Coarsening," *Journal of the American Statistical Association* 114, no. 527 (2019): 1113–1125.

17. E. Bernton, P. E. Jacob, M. Gerber, and C. P. Robert, "Approximate Bayesian Computation With the Wasserstein Distance," *Journal of the Royal Statistical Society, Series B: Statistical Methodology* 81, no. 2 (2019): 235–269.

18. C. P. Robert, "Approximate Bayesian Computation: A Survey on Recent Results," in *Monte Carlo and Quasi-Monte Carlo Methods*, eds. R. Cools and D. Nuyens (Cham, Switzerland: Springer International Publishing, 2016), 185–205.

19. L. F. Price, C. C. Drovandi, A. C. Lee, and D. J. Nott, "Bayesian Synthetic Likelihood," *Journal of Computational and Graphical Statistics* 27, no. 1 (2018): 1–11.

20. M. Fasiolo, S. N. Wood, F. Hartig, and M. V. Bravington, "*An Extended Empirical Saddlepoint Approximation for Intractable Likelihoods*" (2016), arXiv:1601.01849.

21. R. Dutta, J. Corander, S. Kaski, and M. U. Gutmann, "*Likelihood-Free Inference by Penalised Logistic Regression*" (2016), arXiv:1611.10242.

22. Z. An, D. Nott, and C. Drovandi, "Robust Bayesian Synthetic Likelihood via a Semi-Parametric Approach," *Statistics and Computing* 30 (2020): 543–557.

23. J. W. Priddle and C. Drovandi, "*Transformations in Semi-Parametric Bayesian Synthetic Likelihood*" (2020), arxiv:2007.01485.

24. C. C. Drovandi, A. N. Pettitt, and A. Lee, "Bayesian Indirect Inference Using a Parametric Auxiliary Model," *Statistical Science* 30, no. 1 (2015): 72–95.

25. D. T. Frazier and C. Drovandi, "*Robust Approximate Bayesian Inference With Synthetic Likelihood*" (2020), arXiv:1904.04551.

26. A. B. Owen, *Empirical Likelihood* (London, UK: Chapman and Hall, 2001).

27. S. Chaudhuri and M. Ghosh, "Empirical Likelihood for Small Area Estimation," *Biometrika* 98 (2011): 473–480.

28. N. A. Lazar, "Bayesian Empirical Likelihood," *Biometrika* 90 (2003): 319–326.

29. S. M. Schennach, "Bayesian Exponentially Tilted Empirical Likelihood," *Biometrika* 92, no. 1 (2005): 31–46.

30. C. Grazian and B. Liseo, "Approximate Bayesian Inference in Semiparametric Copula Models," *Bayesian Analysis* 12, no. 4 (2017): 991–1016.

31. S. Ghosh, S. Chaudhuri, and U. Gangopadhyay, "Maximum Likelihood Estimation Under Constraints: Singularities and Random Critical Points," *IEEE Transactions on Information Theory* 69, no. 12 (2023): 7976–7997.

32. H. Haario, E. Saksman, and J. Tamminen, "An Adaptive Metropolis Algorithm," *Bernoulli* 7, no. 2 (2001): 223–242.

33. S. Horvát, E. Czabarka, and Z. Toroczkai, "Reducing Degeneracy in Maximum Entropy Models of Networks," *Physical Review Letters* 114 (2015): 158701.

34. E. T. Jaynes, "Information Theory and Statistical Mechanics," *Physics Review* 106 (1957a): 620–630.

35. E. T. Jaynes, "Information Theory and Statistical Mechanics. II," *Physics Review* 108 (1957b): 171–190.

36. S. R. Lele, B. Dennis, and F. Lutscher, "Data Cloning: Easy Maximum Likelihood Estimation for Complex Ecological Models Using Bayesian Markov Chain Monte Carlo Methods," *Ecology Letters* 10 (2007): 551–563.

37. A. Doucet, S. Godsill, and C. Robert, "Marginal Maximum a Posteriori Estimation Using Markov Chain Monte Carlo," *Statistics and Computing* 12 (2002): 77–84.

38. C. Gouriéroux and A. Monfort, *Simulation-Based Econometric Methods* (Oxford, UK: Oxford University Press, 1996).

39. S. Chaudhuri, D. Mondal, and T. Yin, "Hamiltonian Monte Carlo Sampling in Bayesian Empirical Likelihood," *Journal of the Royal Statistical Society, Series B* 79 (2017): 293–320.

40. L. F. Kozachenko and N. N. Leonenko, "Sample Estimate of the Entropy of a Random Vector," *Problemy Peredachi Informatsii* 23, no. 2 (1987): 9–16.

41. T. B. Berrett, R. J. Samworth, and M. Yuan, "Efficient Multivariate Entropy Estimation via k-Nearest Neighbour Distances," *Annals of Statistics* 47, no. 1 (2019): 288–318.

42. P. Hall and S. Morton, "On the Estimation of Entropy," *Annals of the Institute of Statistical Mathematics* 45 (1993): 69–88.

43. P. T. K. Cuc, "Empirical Likelihood, Classifiaction and Approximate Bayesian Computation," (PhD thesis, National University of Singapore, 2016).

44. T. Cover and J. Thomas, *Elements of Information Theory* (Hoboken, New Jersey: Wiley, 2012).

45. J. Whittaker, *Graphical Models in Applied Multivariate Statistics* (Chichester, UK: Wiley, 1990).

46. R. D'Agostino and M. A. Stephens, eds., *Goodness of Fit Techniques* (New York: Marcel Dekker, 1986).

47. A. Gut, "On the Moment Problem," *Bernoulli* 8, no. 3 (2002): 407–421.

48. C. W. Anderson and S. G. Coles, "The Largest Inclusions in a Piece of Steel," *Extremes* 5, no. 3 (2002): 237–252.

49. P. Bortot, S. Coles, and S. Sisson, "Inference for Stereological Extremes," *Journal of the American Statistical Association* 102, no. 477 (2007): 84–92.

50. R. Erhardt and S. A. Sisson, "Modelling Extremes Using Approximate Bayesian Computation," in *Extreme Value Modelling and Risk Analysis: Methods and Applications*, eds. D. K. Dey and J. Yan (Boca Raton, Florida: Chapman and Hall/CRC Press, 2015), 281–306.

51. K. C. Pham, D. J. Nott, and S. Chaudhuri, "A Note on Approximating ABC-MCMC Using Flexible Classifiers," *Stat* 3, no. 1 (2014): 218–227.

52. C. Faes, J. T. Ormerod, and M. P. Wand, "Variational Bayesian Inference for Parametric and Nonparametric Regression With Missing Data," *Journal of the American Statistical Association* 106, no. 495 (2011): 959–971.

53. J. T. Ormerod and M. P. Wand, "Explaining Variational Approximation," *American Statistics* 64, no. 2 (2010): 140–153.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.

Appendix A

Proofs of Results in Section

Proof of Theorem 1. The proof proceeds by expanding the Kullback–Leibler divergence

$$D_{KL}[q(\theta, s_1) \| f(\theta, s_1|s_o)]$$

when $q(\theta, s_1) = q'(\theta) f_0(s_1|\theta)$.

For a $f \in \mathcal{F}$, suppose $f(s_o)$ is the marginal distribution of s_o. It is well known that [52, 53] the so called log evidence, that is, $\log f(s_o)$ can be expressed as:

$$\begin{aligned}\log f(s_o) &= D_{KL}[q(\theta, s_1) \| f(\theta, s_1|s_o))] \\ &\quad + \int q(\theta, s_1) \log\left(\frac{f(\theta, s_1, s_o)}{q(\theta, s_1)}\right) ds_1\, d\theta \end{aligned} \tag{A.1}$$

For the convenience of notation, for an $f \in \mathcal{F}$ we define:

$$\begin{aligned} f''(\theta, s_o) &= \frac{\exp\left(E^0_{s_1|\theta}[\log f(\theta, s_1, s_o)]\right)}{\int \exp\left(E^0_{s_1|t}[\log f(t, s_1, t')]\right) dt dt'}, \\ f''(s_o) &= \int f''(\theta, s_o) d\theta \text{ and} \\ f''(\theta|s_o) &= f''(\theta, s_o)/f''(s_o) \end{aligned}$$

By substituting the expression of $q(\theta, s_1) \in \mathcal{Q}'$ in (A.1) we get:

$$\begin{aligned} & D_{KL}[q(\theta, s_1) \| f(\theta, s_1|s_o)] \\ &= \log f(s_o) + \int q'(\theta) f_0(s_1|\theta) \log f_0(s_1|\theta) ds_1 d\theta \\ &\quad - \int q'(\theta) \left\{ \int \log f(\theta, s_1, s_o) f_0(s_1|\theta) ds_1 - \log q'(\theta) \right\} d\theta \\ &= \log f''(s_o) - \int q'(\theta) \log\left(\frac{\exp\left(E^0_{s_1|\theta}[\log f(\theta, s_1, s_o)]\right)}{q'(\theta)} \right) d\theta \\ &\quad - \int H^0_{s_1|\theta}(\theta) q'(\theta) d\theta + \log\left(\frac{f(s_o)}{f''(s_o)}\right) \\ &= \log f''(s_o) - \int q'(\theta) \left\{ \log\left(\frac{f''(\theta, s_o)}{q'(\theta)}\right) \right. \\ &\quad \left. - \log \int \exp\left(E^0_{s_1|t}[\log f(t, s_1, t')]\right) dt dt' \right\} d\theta \\ &\quad - \int H^0_{s_1|\theta}(\theta) q'(\theta) d\theta + \log\left(\frac{f(s_o)}{f''(s_o)}\right) \end{aligned} \tag{A.2}$$

Similar to (A.1) one can show that:

$$\log f''(s_o) = \int q'(\theta) \log\left(\frac{f''(\theta, s_o)}{q'(\theta)}\right) d\theta + D_{KL}[q'(\theta) \| f''(\theta|s_o)]$$

where second addendum is the Kullback–Leibler divergence between the densities $q'(\theta)$ and $f''(\theta|s_o)$. Moreover, the third addendum in (A.2) depends on the hyper-parameters of $\pi(\theta)$ and thus independent of θ. Suppose we denote $C' = \log \int \exp\left(E^0_{s_1|t}[\log f(t, s_1, t')]\right) dt dt'$.

By substituting the above result in (A.2) and from (A.1) we get:

$$\begin{aligned} & D_{KL}[q(\theta, s_1) \| f(\theta, s_1|s_o)] \\ &= \log f(s_o) - \int q'(\theta) f_0(s_1|\theta) \log\left(\frac{f(\theta, s_1, s_o)}{q'(\theta) f_0(s_1|\theta)}\right) ds_1\, d\theta \\ &= D_{KL}[q'(\theta) \| f''(\theta|s_o)] - \int H^0_{s_1|\theta}(\theta) q'(\theta) d\theta - C' + \log\left(\frac{f(s_o)}{f''(s_o)}\right) \end{aligned} \tag{A.3}$$

Now by expanding the first two addenda in (A.3) we get:

$$\begin{aligned} & D_{KL}\left[q'(\theta)\|f''(\theta|s_o)\right] - \int H^0_{s_1|\theta}(\theta)q'(\theta)d\theta \\ &= \int q'(\theta)\left\{\log\left(\frac{q'(\theta)}{f''(\theta|s_o)}\right) - H^0_{s_1|\theta}(\theta)\right\}d\theta \\ &= \int q'(\theta)\left\{\log\left(\frac{q'(\theta)}{f''(\theta|s_o)exp\left(H^0_{s_1|\theta}(\theta)\right)}\right)\right\}d\theta \\ &= \int q'(\theta)\left\{\log\left(\frac{q'(\theta)}{f'(\theta|s_o)}\right) - \left(\log\int f''(t|s_o)\exp\left(H^0_{s_1|t}(t)\right)dt\right)\right\}d\theta \end{aligned} \tag{A.4}$$

The first addendum in (A.4) is the Kullback–Leibler divergence between q' and $f'(\theta|s_o)$. The second addendum is a function of s_o and is independent of θ. By denoting it by $C(s_o)$ and collecting the terms from (A.3) and (A.4) we get:

$$\begin{aligned} & D_{KL}[q(\theta,s_1)\|f(\theta,s_1|s_o)] \\ &= D_{KL}\left[q'(\theta)\|f'(\theta|s_o)\right] - C(s_o) - C' + \log\left(\frac{f(s_o)}{f''(s_o)}\right) \end{aligned} \tag{A.5}$$

Note that, the R.H.S. of Equation (A.5) is non-negative for all $q' \in \mathcal{Q}_\Theta$. Furthermore, only the first addendum depends on q', which is also non-negative, with equality holding iff $q'(\theta) = f'(\theta|s_o)$. This implies the R.H.S. of (A.5) attains its minimum at $q'(\theta) = f'(\theta|s_o)$. So, it clearly follows that the information projection of $f(\theta,s_1|s_o)$ is given by $f'(\theta|s_o)f_0(s_1|\theta)$. □

Proof of Theorem 2. From the LHS of 2 we get:

$$\begin{aligned} & \log f(\theta,s_o) - \left\{E^0_{s_1|\theta}[\log f(\theta,s_1,s_o)] + H^0_{s_1|\theta}(\theta)\right\} \\ &= \log f(\theta,s_o) - \int f_0(s_1|\theta)\log f(s_1|s_o,\theta)ds_1 \\ &\quad - \int f_0(s_1|\theta)\log f(\theta,s_o)ds_1 + H^0_{s_1|\theta}(\theta) \\ &= \int f_0(s_1|\theta)\log\left(\frac{f_0(s_1|\theta)}{f(s_1|s_o,\theta)}\right)ds_1 \\ &= D_{KL}[f_0(s_1|\theta)\|f(s_1|s_o,\theta)] \end{aligned}$$

Rest of the theorem follows from above. □

Appendix B

Proofs of Results in Section 4.1

Proof of Lemma 1. We show that for every $\epsilon > 0$, there exists $n_0 = n_0(\epsilon)$ such that for any $n \geq n_0$ for all $\theta \in \Theta_n$ the maximization problem in (5) is feasible with probability larger than $1-\epsilon$.

By assumption, for each θ, random vectors $\xi^{(n)}_i(\theta)$ are i.i.d., put positive mass on each orthant and supremum of their lengths in each orthant diverge to infinity with n. The random vectors $\left\{\xi^{(n)}_i(\theta) - \xi^{(n)}_o(\theta_o)\right\}$ will inherit the same properties. That is, there exists integer n_0, such that for each $n \geq n_0$, the convex hull of the vectors $\left\{\xi^{(n)}_i(\theta) - \xi^{(n)}_o(\theta_o)\right\}$, $i = 1, \ldots, m(n)$, would contain the unit sphere with probability larger than $1-\epsilon/2$.

We choose an $n \geq n_0$ and a $\theta \in \Theta_n$. For this choice of θ:

$$\begin{aligned} h^{(n)}_i(\theta,\theta_o) &= b_n\{\mathfrak{s}(\theta) - \mathfrak{s}(\theta_o)\} + \xi^{(n)}_i(\theta) - \xi^{(n)}_o(\theta_o) \\ &= c_n(\theta) + \xi^{(n)}_i(\theta) - \xi^{(n)}_o(\theta_o) \end{aligned}$$

where, $\|\mathfrak{s}(\theta) - \mathfrak{s}(\theta_o)\| \leq b_n^{-1}$. That is, $\|c_n(\theta)\| \leq 1$. Now, since $-c_n(\theta)$ is in the convex hull of the vectors $\left\{\xi^{(n)}_i(\theta) - \xi^{(n)}_o(\theta_o)\right\}$, $i = 1, \ldots, m(n)$, with probability larger than $1-\epsilon/2$, there exists weights $w \in \Delta_{m(n)-1}$ such that,

$$-c_n(\theta) = \sum_{i=1}^{m(n)} w_i\left\{\xi^{(n)}_i(\theta) - \xi^{(n)}_o(\theta_o)\right\}$$

Now it follows that for the above choice of w that

$$\sum_{i=1}^{m(n)} w_i h^{(n)}_i(\theta,\theta_o) = c_n(\theta) + \sum_{i=1}^{m(n)} w_i\left\{\xi^{(n)}_i(\theta) - \xi^{(n)}_o(\theta_o)\right\} = 0$$

which shows that the problem in (5) is feasible. □

Proof of Lemma 2. Let ϵ be as in the statement. By Assumption (A1), for some $\delta > 0$, $\|\mathfrak{s}(\theta) - \mathfrak{s}(\theta_o)\| > \delta$ for all θ with $\|\theta - \theta_o\| > \epsilon$.

Consider $\eta > 0$. We show that there exists $n_0 = n_0(\eta)$ such that for any $n \geq n_0$, the constrained maximization problem in (5) is not feasible for all $\|\theta - \theta_o\| > \epsilon$, with probability larger than $1-\eta$.

Let if possible $w \in \Delta_{m(n)-1}$ be a feasible solution. Hence we get:

$$\begin{aligned} 0 = \sum_{i=1}^{m(n)} w_i h^{(n)}_i(\theta,\theta_o) &= \sum_{i=1}^{m(n)} w_i\left\{s\left(X^{(n)}_i(\theta)\right) - s\left(X^{(n)}_o(\theta_o)\right)\right\} \\ &= \left\{\mathfrak{s}^{(n)}(\theta) - \mathfrak{s}^{(n)}(\theta_o)\right\} + \left\{\sum_{i=1}^{m(n)} w_i\xi^{(n)}_i(\theta)\right\} - \xi^{(n)}_o(\theta_o) \end{aligned}$$

so that

$$-b_n\{\mathfrak{s}(\theta) - \mathfrak{s}(\theta_o) + o(1)\} = \sum_{i=1}^{m(n)} w_i\xi^{(n)}_i(\theta) - \xi^{(n)}_o(\theta_o)$$

By dividing both sides by b_n we get:

$$-\{\mathfrak{s}(\theta) - \mathfrak{s}(\theta_o)\} = \sum_{i=1}^{m(n)} w_i\left\{\frac{\xi^{(n)}_i(\theta)}{b_n} - \frac{\xi^{(n)}_o(\theta_o)}{b_n}\right\} - o(1) \tag{B.1}$$

Now, $\|\xi^{(n)}_o(\theta_o)\|/b_n \leq \sup_{i\in\{o,1,2\ldots,m(n)\}}\|\xi^{(n)}_o(\theta_o)\|/b_n$ and

$$\left\|\sum_{i=1}^{m(n)} w_i\frac{\xi^{(n)}_i(\theta)}{b_n}\right\| \leq \sum_{i=1}^{m(n)} w_i\frac{\|\xi^{(n)}_i(\theta)\|}{b_n} \leq \sup_{i\in\{o,1,2,\ldots,m(n)\}}\frac{\|\xi^{(n)}_i(\theta)\|}{b_n}$$

That is, by Assumption (A3), there exists $n_0(\eta)$ such that for any $n \geq n_0$, the RHS of (B.1) is less than δ for all $\theta \in B(\theta_o,\epsilon)$, with probability larger than $1-\eta$. However, $\|\mathfrak{s}(\theta) - \mathfrak{s}(\theta_o)\| > \delta$. We arrive at a contradiction. Thus the problem is infeasible for every $\theta \in B(\theta_o,\epsilon)^C$ with probability larger than $1-\eta$. □

Proof of Theorem 4. Let $g(\theta)$ be a continuous, bounded function. We choose an $\epsilon > 0$. Then by Lemma 2, there exists $n(\epsilon)$, such that for any $n > n(\epsilon)$ and $\theta \in B(\theta_o,\epsilon)^C$, $l_n(\theta) = 0$ and by definition (6) the posterior $\hat{\Pi}_n(\theta|s(X_o(\theta_o))) = 0$. That is for any $n > n(\epsilon)$,

$$\begin{aligned} \int_\Theta g(\theta)\hat{\Pi}_n(\theta|s(X_o(\theta_o)))d\theta &= \int_{B(\theta_o,\epsilon)} g(\theta)\hat{\Pi}_n(\theta|s(X_o(\theta_o)))d\theta \\ &= \int_{B(\theta_o,\epsilon)}\{g(\theta) - g(\theta_o)\}\hat{\Pi}_n(\theta|s(X_o(\theta_o)))d\theta \\ &\quad + g(\theta_o)\int_{B(\theta_o,\epsilon)}\hat{\Pi}_n(\theta|s(X_o(\theta_o)))d\theta \end{aligned}$$

Since the function $g(\theta)$ is bounded and continuous at θ_o, the first term is negligible. Furthermore, $\int_{B(\theta_o,\epsilon)}\hat{\Pi}_n(\theta|s(X_o(\theta_o)))d\theta = 1$. This implies the

integral converges to $g(\theta_o)$. This shows, the posterior converges weakly to δ_{θ_o}. □

Appendix C

Details of the Remarks in Section 4.2

Using the notations introduced above, when $r = 1$, that is, there is only one constraint present, under conditions similar to those described above, it can be shown that, [31, Theorem 3.4] for any $\theta \in \Theta$:

$$\begin{aligned} l_m(\theta) &:= \frac{1}{m}\sum_{i=1}^{m}\log(\hat{w}(\theta)) \\ &= -\frac{1}{\mathcal{M}_m(\theta)}\left|E_{s\left(X_1^{(n)}(\theta)\right)|\theta}\left[s\left(X_1^{(n)}(\theta)\right)\right] - \left[s(X_o^{(n)}(\theta_o))\right]\right|\left(1+o_p(1)\right) \\ &= -\frac{b_n}{\mathcal{M}_m(\theta)}\left|(\mathfrak{s}(\theta)-\mathfrak{s}(\theta_o)+o(1)) - \frac{\xi_o^{(n)}(\theta_o)}{b_n}\right|\left(1+o_p(1)\right) \end{aligned} \quad \text{(C.1)}$$

Details of Remark 1. In order to ensure the first condition, suppose $\theta \neq \theta_o$, and as $m, n \to \infty$, and in (C.1), $b_n/\mathcal{M}_m(\theta)$ diverges. Since by assumption (A3), as $m, n \to \infty$, $\sup_{i\in\{o,1,2,\ldots,m\}} |\xi_o^{(n)}(\theta_o)|/b_n \to 0$, in probability, uniformly over θ, and by assumption (A1), $\|\mathfrak{s}(\theta)-\mathfrak{s}(\theta_o)\| > 0$, for each $\theta \neq \theta_o$, the R.H.S. of (C.1) diverges to $-\infty$. So $\exp(l_m(\theta))$ converges to zero. That is, an upper bound of the rate of growth of m can thus be obtained by inverting the relation $b_n > \mathcal{M}_m(\theta)$.

Depending on the distribution of $\xi_o^{(n)}$, m can be much larger than n. For example, if $\xi_o^{(n)}$ follows a normal distribution with mean zero and variance σ_o^2, $b_n = \sqrt{n}$ and $\mathcal{M}_m(\theta)$ is of the order $\sigma_o\sqrt{2\log(m)}$, which allows an upper bound of m as large as $\exp\left(n/\left(2\sigma_o^2\right)\right)$.

Details of Remark 2. Similar to the argument for the upper bound, for Bayesian consistency $l_m(\theta_o)$ cannot diverge to $-\infty$. There exists a constant $C_1 > 0$ such that, $l_m(\theta) > -C_1$ with a high probability.

For (13), it follows that when $\theta = \theta_o$:

$$l_m(\theta_o) = -\frac{|\xi_o^{(n)}(\theta_o)|}{\mathcal{M}_m(\theta_o)}\left(1+o_p(1)\right) \quad \text{(C.2)}$$

For simplicity of presentation, we also suppose $\xi_o^{(n)}(\theta_o)$ is a $N\left(0,\sigma_o^2\right)$ variable.

For a fixed $C_1 > 0$, we first compute $Pr[l_m(\theta_o) \le -C_1]$. Using the tail bound for a $N\left(0,\sigma_o^2\right)$ random variables we get,

$$\begin{aligned} Pr[l_m(\theta_o) \le -C_1] &= Pr\left[-\frac{\left|\xi_o^{(n)}(\theta_o)\right|}{\mathcal{M}_m(\theta_o)}\left(1+o_p(1)\right) \le -C_1\right] \\ &= Pr\left[\left|\xi_o^{(n)}(\theta_o)\right| \ge C_1\frac{\mathcal{M}_m(\theta_o)}{1+o_p(1)}\right] \\ &\le \exp\left(-\frac{1}{2}\left(\frac{C_1\mathcal{M}_m(\theta_o)}{\sigma_o}\right)^2\right) \end{aligned} \quad \text{(C.3)}$$

Since $\xi_o^{(n)}(\theta_o)$ is normally distributed, $\mathcal{M}_m(\theta_o) = \sigma_o\sqrt{2\log m}$, diverges as $m \to \infty$. So the R.H.S. of (C.3) converges to zero. That is, for any $C_1 > 0$, $Pr[l_m(\theta_o) \le -C_1]$ converges to zero. Furthermore, by substituting the expression for $\mathcal{M}_m(\theta_o)$ in (C.3) we get:

$$Pr[l_m(\theta_o) \le -C_1] \le \exp\left(-C_1^2\log m\right) = \frac{1}{m^{C_1^2}} \quad \text{(C.4)}$$

Now as before by setting $p_n = m^{-C_1^2}$, we get $m = p_n^{-1/C_1^2}$. In particular, if $p_n = n^{-\alpha}$, $m = n^{\alpha/C_1^2}$.

Details of Remark 3. The likelihood ratio statistic for testing the null hypothesis of $\theta = \theta_o$ against the unrestricted alternative is given by:

$$LR(\theta_o) = \frac{\exp(l_m(\theta_o))}{\max_{w\in\Delta_{m-1}}\exp\left(\sum_{i=1}^{m}\log(w_i)/m\right)}$$

Clearly, the maximum value the denominator attains is, $1/m$. So the log-likelihood ratio $\log LR(\theta_o)$ turns out to be $l_m(\theta_o) + \log m$.

The test rejects H_0 if $\log LR(\theta_o)$ is smaller than $\log C_0$, for some pre-specified $C_0 \in (0,1)$. Ideally, C_0 should be a function of m. However, at this point we assume C_0 to be fixed.

Using (13), the probability of rejecting the null hypothesis is given by:

$$\begin{aligned} Pr[\log m + l_m(\theta_o) \le \log C_0] &= Pr[l_m(\theta_o) \le \log C_0 - \log m] \\ &= Pr\left[-\frac{1}{\mathcal{M}_m(\theta_o)}\left|\xi_o^{(n)}(\theta_o)+o(1)\right|(1+o(1)) \le \log\left(\frac{C_0}{m}\right)\right] \\ &= Pr\left[\left|\xi_o^{(n)}(\theta_o)+o(1)\right|(1+o(1)) \ge -\mathcal{M}_m(\theta_o)\log\left(\frac{C_0}{m}\right)\right] \end{aligned}$$

Now Suppose that $\xi^{(n)}(\theta)$ is a $N\left(0,\sigma_0^2\right)$ random variable. Using the tail bounds for a normal distribution, we get:

$$\begin{aligned} &Pr\left[\left|\xi_o^{(n)}(\theta_o)+o(1)\right|(1+o(1)) \ge -\mathcal{M}_m(\theta_o)\log\left(\frac{C_0}{m}\right)\right] \\ &\le exp\left(-\frac{1}{2\sigma_o^2}\left\{\mathcal{M}_m(\theta_o)\log\left(\frac{C_0}{m}\right)\right\}^2\right) \end{aligned} \quad \text{(C.5)}$$

By substituting $\mathcal{M}_m(\theta_o) = \sigma_o\sqrt{2\log m}$ in the exponent of the above expression we get:

$$\begin{aligned} \frac{1}{2\sigma_o^2}\left\{\mathcal{M}_m(\theta_o)\log\left(\frac{C_0}{m}\right)\right\}^2 &= (\log m)(\log C_0 - \log m)^2 \\ &= (\log m)^3 - 2(\log m)^2\log C_0 \\ &\quad + (\log m)(\log C_0)^2 \end{aligned}$$

Clearly, the $(\log m)^3$ term dominates and the probability of rejecting the null hypothesis decreases at the rate of $\exp\left(-(\log m)^3\right)$. This is true even if C_0 increases to one with increasing m at a suitable rate.

Finally, in order to describe some relationship between m and n, suppose we would like to ensure, that the probability of rejecting the null hypothesis reduces at the rate of p_n. Then it follows that the number of replications required to ensure such a rate is of the order $m = \exp\left((-\log p_n)^{1/3}\right)$.

Details of Remark 4. Let us fix $\theta \neq \theta_o$ and suppose $\xi_o^{(n)}(\theta_o)$ follows a $N\left(0,\sigma_o^2\right)$ distribution. Then for a fixed $C_2 > 0$, it can be shown that:

$$\begin{aligned} &Pr[l_m(\theta) \le -C_2] \\ &\le Pr\left[\left|\xi_o^{(n)}(\theta_o)\right| \ge \mathcal{M}_m(\theta)\left\{C_2 - \frac{b_n}{\mathcal{M}_m(\theta)}\left|\mathfrak{s}(\theta)-\mathfrak{s}(\theta_o)\right|\right\}\right] \\ &\le \exp\left[-\frac{(\mathcal{M}_m(\theta))^2}{2\sigma_o^2}\left\{C_2 - \frac{b_n}{\mathcal{M}_m(\theta)}\left|\mathfrak{s}(\theta)-\mathfrak{s}(\theta_o)\right|\right\}^2\right] \end{aligned} \quad \text{(C.6)}$$

Now by substituting $\mathcal{M}_m(\theta) = \sigma_o\sqrt{2\log m}$ we get:

$$Pr[l_m(\theta) \le -C_2] \le \left(\frac{1}{m}\right)^{\left\{C_2 - \frac{b_n}{\sigma_o\sqrt{2\log m}}\left|\mathfrak{s}(\theta)-\mathfrak{s}(\theta_o)\right|\right\}^2} \quad \text{(C.7)}$$

Now, if $\mathcal{M}_m(\theta)/b_n = \sigma_o\sqrt{2\log m}/b_n$ diverges with m and n, clearly, for large values of m and n, $Pr[l_m(\theta) \le -C_2] \approx m^{-C_2^2}$. That is, for any fixed

$C_2 > 0$ and $\theta \neq \theta_o$, $l_m(\theta) \geq -C_2$ with a high probability, and $\exp(l_m(\theta))$ does not collapse to zero with a high probability.

Furthermore, for a fixed n, R.H.S. of (C.7) is a decreasing function in m. That is if the sample size is kept fixed, increasing the number of replications will increase the probability of $l_m(\theta) \geq -C_2$. As a result, the log-likelihood will be flatter in shape. Note that, from (C.1), it is clear that the variance of the expected log likelihood gets reduced as m increases. This explains a bias-variance trade-off in the choice of m. Such phenomenon is evident from Figure 3, where the curve joining the means of the proposed estimated log posterior progressively flattens with the number of replications. The argument above provides a formal explanation of the phenomenon.