Approximate Bayesian computation via the energy statistic

Hien Nguyen, Julyan Arbel, Hongliang Lu, Florence Forbes

Approximate Bayesian computation via the energy statistic

Hien D. Nguyen
Department of Mathematics and Statistics, La Trobe University
Julyan Arbel, Hongliang Lü, Florence Forbes
Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK

Abstract

Approximate Bayesian computation (ABC) has become an essential part of the Bayesian toolbox for addressing problems in which the likelihood is prohibitively expensive or entirely unknown, making it intractable. ABC defines a quasi-posterior by comparing observed data with simulated data, traditionally based on some summary statistics, the elicitation of which is regarded as a key difficulty. In recent years, a number of data discrepancy measures bypassing the construction of summary statistics have been proposed, including the Kullback–Leibler divergence, the Wasserstein distance and maximum mean discrepancies. Here we propose a novel importance-sampling (IS) ABC algorithm relying on the so-called *two-sample energy statistic*. We establish a new asymptotic result for the case where both the observed sample size and the simulated data sample size increase to infinity, which highlights to what extent the data discrepancy measure impacts the asymptotic pseudo-posterior. The result holds in the broad setting of IS-ABC methodologies, thus generalizing previous results that have been established only for rejection ABC algorithms. Furthermore, we propose a consistent V-statistic estimator of the energy statistic, under which we show that the large sample result holds. Our proposed energy statistic based ABC algorithm is demonstrated on a variety of models, including a Gaussian mixture, a moving-average model of order two, a bivariate beta and a multivariate g-and-k distribution. We find that our proposed method compares well with alternative discrepancy measures.

Keywords: approximate Bayesian computation, energy statistic, Kullback–Leibler divergence, importance sampling, maximum mean discrepancy, Wasserstein distance.

1 Introduction

In recent years, Bayesian inference has become a popular paradigm for machine learning and statistical analysis. Good introductions and references to the primary methods and philosophies of Bayesian inference can be found in texts such as Press (2003), Ghosh et al. (2006), Koch (2007), Koop et al. (2007), Robert (2007), Barber (2012), and Murphy (2012).

In this article, we are concerned with the problem of parametric, or classical Bayesian inference. For details regarding nonparametric Bayesian inference, the reader is referred to the expositions of Ghosh & Ramamoorthi (2003), Hjort et al. (2010), and Ghosh & van der Vaart (2017).

When conducting parametric Bayesian inference, we observe some realizations $\boldsymbol{x}$ of the data $\boldsymbol{X} \in \mathbb{X}$ that are generated from some data generating process (DGP), which can be characterized by a parametric likelihood, given by a probability density function (PDF) $f(\boldsymbol{x}|\boldsymbol{\theta})$, determined entirely via the parameter vector $\boldsymbol{\theta}$. Using the information that the parameter vector $\boldsymbol{\theta}$ is a realization of a random variable $\boldsymbol{\Theta} \in \mathbb{T}$, which arises from a DGP that can be characterized by some known prior PDF $\pi(\boldsymbol{\theta})$, we wish to characterize the posterior distribution

$$\pi(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{f(\boldsymbol{x}|\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})}{c(\boldsymbol{x})}, \tag{1}$$

where

$$c(\boldsymbol{x}) = \int_{\mathbb{T}} f(\boldsymbol{x}|\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}\ .$$

In very simple cases, such as cases when the prior PDF is a conjugate of the likelihood (cf. Robert, 2007, Sec. 3.3), the posterior distribution (1) can be expressed explicitly. In the case of more complex but still tractable pairs of likelihood and prior PDFs, one can sample from (1) via a variety of Monte Carlo methods, such as those reported in Press (2003, Ch. 6).

In cases where the likelihood function is known but not tractable, or when the likelihood function has entirely unknown form, one cannot exactly sample from (1) in an inexpensive manner, or at all. In such situations, a sample from an approximation of (1) may suffice in order to conduct the user's desired inference. Such a sample can be drawn via the method of approximate Bayesian computation (ABC).

It is generally agreed that the ABC paradigm originated from the works of Rubin (1984), Tavaré et al. (1997), and Pritchard et al. (1999); see Tavaré (2019) for details. Stemming from the initial listed works, there are now numerous variants of ABC methods. Some good reviews of the current ABC literature can be found in the expositions of Marin et al. (2012), Voss (2014, Sec. 5.1), Lintusaari et al. (2017), and Karabatsos & Leisen (2018). The volume of Sisson et al. (2019) provides a comprehensive treatment regarding ABC methodologies.

The core philosophy of ABC is to define a quasi-posterior by comparing data with plausibly simulated replicates. The comparison is traditionally based on some summary statistics, the choice of which being regarded as a key challenge of the approach.

In recent years, data discrepancy measures bypassing the construction of summary statistics have been proposed by viewing data sets as empirical measures. Examples of such an approach is via the use of the Kullback–Leibler divergence, the Wasserstein distance, or a maximum mean discrepancy (MMD) variant.

In this article, we develop upon the discrepancy measurement approach of Jiang et al. (2018), via the importance sampling ABC (IS-ABC) approach which makes use of a weight function (see e.g., Karabatsos & Leisen, 2018). In particular, we report on a class of ABC algorithms that utilize the two-sample energy statistic (ES) of Szekely & Rizzo (2004) (see also Baringhaus & Franz, 2004, Szekely & Rizzo, 2013, and Szekely & Rizzo, 2017). Our approach is related to the maximum MMD ABC algorithms that were implemented in Park et al. (2016), Jiang et al. (2018), and Bernton et al. (2019). The MMD is a discrepancy measurement that is closely related to the ES (cf. Sejdinovic et al., 2013).

We establish new asymptotic results that have not been proved in these previous papers. In the IS-ABC setting and in the regime where both the observation sample size and the simulated data sample size increase to infinity, our theoretical result highlights how the data discrepancy measure impacts the asymptotic pseudo-posterior. More specifically, under the assumption that the data discrepancy measure converges to some asymptotic value $\mathcal{D}_{\infty}(\boldsymbol{\theta}_0, \boldsymbol{\theta})$, we show that the pseudo-posterior distribution converges almost surely to a distribution proportional to $\pi(\boldsymbol{\theta}) w(\mathcal{D}_{\infty}(\boldsymbol{\theta}_0, \boldsymbol{\theta}))$: the prior distribution times the IS weight w function evaluated at $\mathcal{D}_{\infty}(\boldsymbol{\theta}_0, \boldsymbol{\theta})$, where $\boldsymbol{\theta}_0$ stands for the 'true' parameter value associated to the DGP that generates observations $\boldsymbol{X}$. Although devised in settings where likelihoods are assumed intractible, ABC can also be cast in the setting of robustness with respect to misspecification, where the ABC posterior distribution can be viewed as a special case of a coarsened posterior distribution (cf. Miller & Dunson, 2018).

The remainder of the article proceeds as follows. In Section 2, we introduce the general IS-ABC framework. In Section 3, we introduce the two-sample ES and demonstrate how it can be incorporated into the IS-ABC framework. Theoretical results regarding the IS-ABC framework and the two-sample ES are presented in Section 4. Illustrations of the IS-ABC framework are presented in Section 5. Conclusions are drawn in Section 6.

2 Importance sampling ABC

Assume that we observe n independent and identically distributed (IID) replicates of $\boldsymbol{X}$ from some DGP, which we put into $\mathbf{X}_n = \{\boldsymbol{X}_i\}_{i=1}^n$. We suppose that the DGP that generates $\boldsymbol{X}$ is dependent on some parameter vector $\boldsymbol{\theta}$, a realization of $\boldsymbol{\Theta}$ from space $\mathbb{T}$, which is random and has prior PDF $\pi(\boldsymbol{\theta})$.

Denote $f(\boldsymbol{x}|\boldsymbol{\theta})$ to be the PDF of $\boldsymbol{X}$, given $\boldsymbol{\theta}$, and write

$$f(\mathbf{x}_n|\boldsymbol{\theta}) = \prod_{i=1}^n f(\boldsymbol{x}_i|\boldsymbol{\theta}),$$

where $\mathbf{x}_n$ is a realization of $\mathbf{X}_n$, and each $\boldsymbol{x}_i$ is a realization of $\boldsymbol{X}_i$ ($i \in [n] = \{1, \dots, n\}$).

If $f(\mathbf{x}_n|\boldsymbol{\theta})$ were known, then we could use (1) to write the posterior PDF

$$\pi(\boldsymbol{\theta}|\mathbf{x}_n) = \frac{f(\mathbf{x}_n|\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})}{c(\mathbf{x}_n)}, \tag{2}$$

where $c(\mathbf{x}_n) = \int_{\mathbb{T}} f(\mathbf{x}_n|\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}$ is a constant that makes $\int_{\mathbb{T}} \pi(\boldsymbol{\theta}|\mathbf{x}_n)\,\mathrm{d}\boldsymbol{\theta} = 1$. When evaluating $f(\boldsymbol{x}|\boldsymbol{\theta})$ is prohibitive and ABC is required, then operating with $f(\mathbf{x}_n|\boldsymbol{\theta})$ is similarly difficult. We suppose that given any $\boldsymbol{\theta}_0 \in \mathbb{T}$, we at least have the capability of sampling from the DGP with PDF $f(\boldsymbol{x}|\boldsymbol{\theta}_0)$. That is, we have a simulation method that allows us to feasibly sample the IID vector $\mathbf{Y}_m = \{\boldsymbol{Y}_i\}_{i=1}^m$, for any $m \in \mathbb{N}$, for a DGP with PDF

$$f(\mathbf{y}_n|\boldsymbol{\theta}) = \prod_{i=1}^m f(\boldsymbol{y}_i|\boldsymbol{\theta}).$$

Using the simulation mechanism that generates samples $\mathbf{Y}_m$ and the prior distribution that generates parameters $\boldsymbol{\Theta}$, we can simulate a set of $N \in \mathbb{N}$ simulations $\mathbf{Z}_N = \{\boldsymbol{Z}_{m,k}\}_{k=1}^N$, where $\boldsymbol{Z}_{m,k}^\top = \left(\mathbf{Y}_{m,k}^\top, \boldsymbol{\Theta}_k^\top\right)$ and $(\cdot)^\top$ is the transposition operator. Here, for each $k \in [N]$, $\boldsymbol{Z}_{m,k}$ is an observation from the DGP with joint PDF $f(\mathbf{y}_m|\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})$, hence each $\boldsymbol{Z}_{m,k}$ is composed of a parameter value and a datum conditional on the parameter value. We now consider how $\mathbf{X}_n$ and $\mathbf{Z}_N$ can be combined in order to construct an approximation of (2).

Following the approach of Jiang et al. (2018), we define $\mathcal{D}(\mathbf{x}_n, \mathbf{y}_m)$ to be some non-negative real-valued function that outputs a small value if $\mathbf{x}_n$ and $\mathbf{y}_m$ are similar, and outputs a large value if $\mathbf{x}_n$ and $\mathbf{y}_m$ are different, in some sense. We call $\mathcal{D}(\mathbf{x}_n, \mathbf{y}_m)$ the data discrepancy measurement between $\mathbf{x}_n$ and $\mathbf{y}_m$, and we say that $\mathcal{D}(\cdot, \cdot)$ is the data discrepancy function.

Next, we let $w(d, \epsilon)$ be a non-negative, decreasing (in d), and bounded (importance sampling) weight function (cf. Section 3 of Karabatsos & Leisen, 2018), which takes as inputs a data discrepancy measurement $d = \mathcal{D}(\mathbf{x}_n, \mathbf{y}_m) \geq 0$ and a calibration parameter $\epsilon > 0$. Using the weight and discrepancy functions, we can propose the following approximation for (2).

In the language of Jiang et al. (2018), we call

$$\pi_{m,\epsilon}(\boldsymbol{\theta}|\mathbf{x}_n) = \frac{\pi(\boldsymbol{\theta}) L_{m,\epsilon}(\mathbf{x}_n|\boldsymbol{\theta})}{c_{m,\epsilon}(\mathbf{x}_n)} \tag{3}$$

the quasi-posterior PDF, where

$$L_{m,\epsilon}(\mathbf{x}_n|\boldsymbol{\theta}) = \int_{\mathbb{X}^m} w(\mathcal{D}(\mathbf{x}_n, \mathbf{y}_m), \epsilon) f(\mathbf{y}_m|\boldsymbol{\theta}) \mathrm{d}\mathbf{y}_m$$

is the approximate likelihood function, and

$$c_{m,\epsilon}(\mathbf{x}_n) = \int_{\mathbb{T}} \pi(\boldsymbol{\theta}) L_{m,\epsilon}(\mathbf{x}_n|\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}$$

is a normalization constant. We can use (3) to approximate (2) in the following way. For any functional of the parameter vector $\boldsymbol{\Theta}$ of interest, $g(\boldsymbol{\Theta})$ say, we may approximate the posterior Bayes estimator of $g(\boldsymbol{\Theta})$ via the expression

$$\mathbb{E}[g(\boldsymbol{\Theta})|\mathbf{x}_n] \approx \frac{\int_{\mathbb{T}} g(\boldsymbol{\theta}) \pi(\boldsymbol{\theta}) L_{m,\epsilon}(\mathbf{x}_n|\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}}{c_{m,\epsilon}(\mathbf{x}_n)}, \tag{4}$$

where the right-hand side of (4) can be unbiasedly estimated using $\mathbf{Z}_N$ via

$$\mathbb{M}[g(\boldsymbol{\Theta})|\mathbf{x}_n] = \frac{\sum_{k=1}^{N} g(\boldsymbol{\Theta}_k) w(\mathcal{D}(\mathbf{X}_n, \mathbf{Y}_{m,k}), \epsilon)}{\sum_{k=1}^{N} w(\mathcal{D}(\mathbf{X}_n, \mathbf{Y}_{m,k}), \epsilon)}. \tag{5}$$

We call the process of constructing (5), to approximate (4), the IS-ABC procedure. The general form of the IS-ABC procedure is provided in Algorithm 1.

Algorithm 1. *IS-ABC procedure for approximating* $\mathbb{E}[g(\boldsymbol{\Theta})|\mathbf{x}_n]$.

Input: *a data discrepancy function* $\mathcal{D}$*, a weight function* w*, and a calibration parameter* $\epsilon > 0$.

For $k \in [N]$;

sample $\mathbf{\Theta}_k$ *from the DGP with PDF* $\pi(\boldsymbol{\theta})$*;*
generate $\mathbf{Y}_{m,k}$ *from the DGP with PDF* $f(\mathbf{y}_m|\mathbf{\Theta}_k)$*;*
put $\boldsymbol{Z}_k = (\mathbf{Y}_{m,k}, \mathbf{\Theta}_k)$ *into* $\mathbf{Z}_N$*.*
Output: $\mathbf{Z}_N$ *and construct the estimator* $\mathbb{M}[g(\mathbf{\Theta})|\mathbf{x}_n]$*.*

3 The energy statistic (ES)

Let δ define a metric and let $\boldsymbol{X} \in \mathbb{X} \subseteq \mathbb{R}^d$ and $\boldsymbol{Y} \in \mathbb{X}$ be two random variables that are in a metric space endowed with δ, where $d \in \mathbb{N}$. Furthermore, let $\boldsymbol{X}'$ and $\boldsymbol{Y}'$ be two random variables that have the same distributions as $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively. Here, $\boldsymbol{X}$, $\boldsymbol{X}'$, $\boldsymbol{Y}$, and $\boldsymbol{Y}'$ are all independent of one another.

Upon writing

$$\mathcal{E}_\delta(\boldsymbol{X}, \boldsymbol{Y}) = 2\mathbb{E}[\delta(\boldsymbol{X}, \boldsymbol{Y})] - \mathbb{E}[\delta(\boldsymbol{X}, \boldsymbol{X}')] - \mathbb{E}[\delta(\boldsymbol{Y}, \boldsymbol{Y}')],$$

we can define the original ES of Baringhaus & Franz (2004) and Szekely & Rizzo (2004), as a function of $\boldsymbol{X}$ and $\boldsymbol{Y}$, via the expression $\mathcal{E}_{\delta_2}(\boldsymbol{X}, \boldsymbol{Y})$, where $\delta_p(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|_p$ is the metric corresponding to the ℓ_p-norm ($p \in [1, \infty]$). Thus, the original ES statistic, which we shall also denote as $\mathcal{E}(\boldsymbol{X}, \boldsymbol{Y})$, is defined using the Euclidean norm δ_2.

The original ES has numerous useful mathematic properties. For instance, under the assumption that $\mathbb{E}\|\boldsymbol{X}\|_2 + \mathbb{E}\|\boldsymbol{Y}\|_2 < \infty$, it was shown that

$$\mathcal{E}(\boldsymbol{X}, \boldsymbol{Y}) = \frac{\Gamma\left(\frac{d+1}{2}\right)}{\pi^{(d+1)/2}} \int_{\mathbb{R}^d} \frac{|\varphi_X(\boldsymbol{t}) - \varphi_Y(\boldsymbol{t})|^2}{\|\boldsymbol{t}\|_2^{d+1}} \mathrm{d}\boldsymbol{t}, \tag{6}$$

in Proposition 1 of Szekely & Rizzo (2013), where $\Gamma(\cdot)$ is the gamma function and φ_X (respectively, φ_Y) is the characteristic function of $\boldsymbol{X}$ (respectively, $\boldsymbol{Y}$). Thus, we have the fact that $\mathcal{E}(\boldsymbol{X}, \boldsymbol{Y}) \geq 0$ for any $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{X}$, and $\mathcal{E}(\boldsymbol{X}, \boldsymbol{Y}) = 0$ if and only if $\boldsymbol{X}$ and $\boldsymbol{Y}$ are identically distributed.

The result above is generalized in Proposition 3 of Szekely & Rizzo (2013), where we have the following statement. If $\delta(\boldsymbol{x}, \boldsymbol{y}) = \delta(\boldsymbol{x} - \boldsymbol{y})$ is a continuous function and $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^d$ are independent random variables, then it is necessary and sufficient that $\delta(\cdot)$ is strictly negative definite (see Szekely & Rizzo, 2013 for the precise definition) for the following conclusion to hold: $\mathcal{E}_\delta(\boldsymbol{X}, \boldsymbol{Y}) \geq 0$ for any $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{X}$, and $\mathcal{E}_\delta(\boldsymbol{X}, \boldsymbol{Y}) = 0$ if and only if $\boldsymbol{X}$ and $\boldsymbol{Y}$ are identically distributed.

We observe that there is thus an infinite variety of functions δ from which we can construct energy statistics. We shall concentrate on the use of the original ES, based on δ_2, since it is the most well known and popular of the varieties.

3.1 The V-statistic estimator

Suppose that we observe $\mathbf{X}_n = \{\boldsymbol{X}_i\}_{i=1}^n$ and $\mathbf{Y}_m = \{\boldsymbol{Y}_i\}_{i=1}^m$, where the former is a sample containing n IID replicates of $\boldsymbol{X}$, and the latter is a sample containing m IID replicates of $\boldsymbol{Y}$, respectively, with $\mathbf{X}_n$ and $\mathbf{Y}_m$ being independent. In Gretton et al. (2012), it was shown that for any δ, upon assuming that $\delta(\boldsymbol{x}, \boldsymbol{y}) < \infty$, the so-called V-statistic estimator (cf. Serfling, 1980, Ch. 5 and Koroljuk & Borovskich, 1994)

$$\mathcal{V}_\delta(\mathbf{X}_n, \mathbf{Y}_m) = \frac{2}{mn}\sum_{i=1}^n\sum_{j=1}^m \delta(\boldsymbol{X}_i, \boldsymbol{Y}_j) - \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \delta(\boldsymbol{X}_i, \boldsymbol{X}_j) - \frac{1}{m^2}\sum_{i=1}^m\sum_{j=1}^m \delta(\boldsymbol{Y}_i, \boldsymbol{Y}_j), \quad (7)$$

can be proved to converge in probability to $\mathcal{E}_\delta(\boldsymbol{X}, \boldsymbol{Y})$, as $n \to \infty$ and $m \to \infty$, under the condition that $m/n \to \alpha < \infty$, for some constant α (see also Gretton et al., 2007).

We note that the assumption of this result is rather restrictive, since it either requires the bounding of the space $\mathbb{X}$ or the function δ. In the sequel, we will present a result for the almost sure convergence of the V-statistic that depends on the satisfaction of a more realistic hypothesis.

It is noteworthy that if the ES is non-negative, then the V-statistic retains the non-negativity property of its corresponding ES (cf. Gretton et al., 2012). That is, for any continuous and negative definite function $\delta(\boldsymbol{x}, \boldsymbol{y}) = \delta(\boldsymbol{x} - \boldsymbol{y})$, we have $\mathcal{V}_\delta(\mathbf{X}_n, \mathbf{Y}_m) \geq 0$.

3.2 The ES-based IS-ABC algorithm

From Algorithm 1, we observe that an IS-ABC algorithm requires three components. A data discrepancy measurement $d = \mathcal{D}(\mathbf{X}_n, \mathbf{Y}_m) \geq 0$, a weighting function $w(d, \epsilon) \geq 0$, and a tuning parameter $\epsilon > 0$. We propose the use of the ES in the place of the data discrepancy measurement d, in combination with various weight functions that have been used in the literature. That is we set

$$\mathcal{D}(\mathbf{X}_n, \mathbf{Y}_m) = \mathcal{V}_\delta(\mathbf{X}_n, \mathbf{Y}_m),$$

in Algorithm 1.

In particular, we consider original ES, where $\delta = \delta_2$. We name our framework the ES-ABC algorithm. In Section 4, we shall demonstrate that the proposed algorithm possesses desirable large sample qualities that guarantees its performance in practice, as illustrated in Section 5.

3.3 Related methods

The ES-ABC algorithm that we have presented here is closely related to ABC algorithms based on the maximum mean discrepancy (MMD) that were implemented in Park et al. (2016), Jiang et al. (2018), and Bernton et al. (2019). For each positive definite Mercer kernel function $\chi(\boldsymbol{x}, \boldsymbol{y})$ $(\boldsymbol{x}, \boldsymbol{y} \in \mathbb{X})$, the corresponding MMD is defined via the equation

$$\mathrm{MMD}_{\chi}^{2}(\boldsymbol{X}, \boldsymbol{Y}) = \mathbb{E}\left[\chi(\boldsymbol{X}, \boldsymbol{X}')\right] + \mathbb{E}\left[\chi(\boldsymbol{Y}, \boldsymbol{Y}')\right] - 2\mathbb{E}\left[\chi(\boldsymbol{X}, \boldsymbol{Y})\right],$$

where $\boldsymbol{X}, \boldsymbol{X}', \boldsymbol{Y}, \boldsymbol{Y}'$ are random variable such that $\boldsymbol{X}$ and $\boldsymbol{Y}$ are identically distributed to $\boldsymbol{X}'$ and $\boldsymbol{Y}'$, respectively.

The MMD as a statistic for testing goodness-of-fit was studied prominently in articles such as Gretton et al. (2007), Gretton et al. (2009), and Gretton et al. (2012). It is clear that if $\delta = -\chi$, the forms of the ES and the squared MMD are identical. More details regarding the relationship between the two classes of statistics can be found in Sejdinovic et al. (2013).

We note two shortcomings with respect to the applications of the MMD as a basis for an ABC algorithm in the previous literature. Firstly, no theoretical results regarding the consistency of the MMD-based methods have been proved. And secondly, in the application by Park et al. (2016) and Jiang et al. (2018), the MMD was implemented using the unbiased U-statistic estimator, rather than the biased V-statistic estimator. Although both estimators are consistent, in the sense that they can be proved to be convergent to the desired limiting MMD value, the U-statistic estimator has the unfortunate property of not being bounded from below by zero (cf. Gretton et al., 2012). As such, it does not meet the criteria for a data discrepancy measurement.

4 Theoretical results

4.1 General asymptotic analysis

We now establish a consistency result for the quasi-posterior density (3), when n and m approach infinity. Our result generalizes the main result of Jiang et al. (2018) (i.e., Theorem 1), which is the specific case when the weight function is restricted to the form

$$w(d,\epsilon) = [\![d < \epsilon]\!] , \tag{8}$$

where $[\![\cdot]\!]$ is the Iverson bracket notation, which equals 1 when the internal statement is true, and 0, otherwise (cf. Graham et al., 1994).

The weighting function of form (8), when implemented within the IS-ABC framework, produces the common rejection ABC algorithms, that were suggested by Tavaré et al. (1997), and Pritchard et al. (1999). We extended upon the result of Jiang et al. (2018) so that we may provide theoretical guarantees for more exotic ABC procedures, such as the kernel-smoothed ABC procedure of Park et al. (2016), which implements weights of the form

$$w(d,\epsilon) = \exp\left(-d^q/\epsilon\right), \tag{9}$$

for $q > 0$. See Karabatsos & Leisen (2018) for further discussion and examples.

In order to prove our consistency result, we require Hunt's lemma, which is reported in Dellacherie & Meyer (1980), as Theorem 45 of Section V.5. For convenience to the reader, we present the result, below.

Theorem 1. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space with increasing σ-fields $\{\mathcal{F}_n\}$ and let $\mathcal{F}_\infty = \cup_n \mathcal{F}_n$. Suppose that $\{U_n\}$ is a sequence of random variables that is bounded from above in absolute value by some integrable random variable V, and further suppose that U_n converges almost surely to the random variable U. Then, $\lim_{n\to\infty} \mathbb{E}\left(U_n|\mathcal{F}_n\right) = \mathbb{E}\left(U|\mathcal{F}_\infty\right)$ almost surely, and in $\mathcal{L}_1$ mean, as $n \to \infty$.*

Define the continuity set of a function $d \mapsto w(d)$ as

$$C(w) = \{d : w \text{ is continuous at } d\}.$$

Using Theorem 1, we can now prove the following result regarding the asymptotic behavior of the quasi-posterior density function (3).

Theorem 2. *Let $\mathbf{X}_n$ and $\mathbf{Y}_m$ be IID samples from DGPs that can be characterized by PDFs $f(\mathbf{x}_n|\boldsymbol{\theta}_0) = \prod_{i=1}^{n} f(\boldsymbol{x}_i|\boldsymbol{\theta}_0)$ and $f(\mathbf{y}_m|\boldsymbol{\theta}) = \prod_{i=1}^{m} f(\boldsymbol{y}_i|\boldsymbol{\theta})$, respectively, with corresponding parameter vectors $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}$. Suppose that the data discrepancy $\mathcal{D}(\mathbf{X}_n, \mathbf{Y}_m)$ converges to some $\mathcal{D}_\infty(\boldsymbol{\theta}_0, \boldsymbol{\theta})$, which is a function of $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}$, almost surely as $n \to \infty$, for some $m = m(n) \to \infty$. If $w(d, \epsilon)$ is piecewise continuous and decreasing in d and $w(d, \epsilon) \le a < \infty$ for all $d \ge 0$ and any $\epsilon > 0$, and if*

$$\mathcal{D}_\infty(\boldsymbol{\theta}_0, \boldsymbol{\theta}) \in C(w(\cdot, \epsilon)),$$

then we have

$$\pi_{m,\epsilon}(\boldsymbol{\theta}|\mathbf{x}_n) \to \frac{\pi(\boldsymbol{\theta})\, w(\mathcal{D}_\infty(\boldsymbol{\theta}_0, \boldsymbol{\theta}), \epsilon)}{\int \pi(\boldsymbol{\theta})\, w(\mathcal{D}_\infty(\boldsymbol{\theta}_0, \boldsymbol{\theta}), \epsilon)\, \mathrm{d}\boldsymbol{\theta}}, \tag{10}$$

almost surely, as $n \to \infty$.

Proof. Using the notation of Theorem 1, we set $U_n = w(d(\mathbf{X}_n, \mathbf{Y}_m), \epsilon)$. Since $w(d, \epsilon) \le a < \infty$, for any d, we have the existence of a $|U_n| \le V < \infty$ such that V is integrable, since we can take $V = a$. Since $\mathcal{D}(\mathbf{X}_n, \mathbf{Y}_m)$ converges almost surely to $\mathcal{D}_\infty(\boldsymbol{\theta}_0, \boldsymbol{\theta})$, and $w(\cdot, \epsilon)$ is continuous at $\mathcal{D}_\infty(\boldsymbol{\theta}_0, \boldsymbol{\theta})$, we have $U_n \to U = w(\mathcal{D}_\infty(\boldsymbol{\theta}_0, \boldsymbol{\theta}), \epsilon)$ with probability one by the extended continuous mapping theorem (cf. DasGupta, 2011, Thm. 7.10).

Now, let $\mathcal{F}_n$ be the σ-field generated by the sequence $\{\boldsymbol{X}_1, \dots, \boldsymbol{X}_n\}$. Thus, $\mathcal{F}_n$ is an increasing σ-field, which approaches $\mathcal{F}_\infty = \cup_n \mathcal{F}_n$. We are in a position to directly apply Theorem 1. This yields

$$\mathbb{E}[w(\mathcal{D}(\mathbf{X}_n, \mathbf{Y}_m), \epsilon)|\mathbf{X}_n] \to \mathbb{E}[w(\mathcal{D}_\infty(\boldsymbol{\theta}_0, \boldsymbol{\theta}), \epsilon)|\mathbf{X}_\infty],$$

almost surely, as $n \to \infty$, where the right-hand side equals $w(\mathcal{D}_\infty(\boldsymbol{\theta}_0, \boldsymbol{\theta}), \epsilon)$.

Notice that the left-hand side has the form

$$\mathbb{E}[w(\mathcal{D}(\mathbf{X}_n, \mathbf{Y}_m), \epsilon)|\mathbf{X}_n] = L_{m,\epsilon}(\mathbf{X}_n|\boldsymbol{\theta})$$

and therefore $L_{m,\epsilon}(\mathbf{X}_n|\boldsymbol{\theta}) \to w(\mathcal{D}_\infty(\boldsymbol{\theta}_0, \boldsymbol{\theta}), \epsilon)$, almost surely, as $n \to \infty$. Thus, the numerator of (3) converges to

$$\pi(\boldsymbol{\theta})\, w(\mathcal{D}_\infty(\boldsymbol{\theta}_0, \boldsymbol{\theta}), \epsilon), \tag{11}$$

almost surely.

To complete the proof, it suffices to show that the denominator of (3) converges almost surely to

$$\int_{\mathbb{T}} \pi(\boldsymbol{\theta}) \, w(\mathcal{D}_\infty(\boldsymbol{\theta}_0, \boldsymbol{\theta}), \epsilon) \, \mathrm{d}\boldsymbol{\theta}. \tag{12}$$

Since $L_{m,\epsilon}(\mathbf{X}_n|\boldsymbol{\theta}) \to w(\mathcal{D}_\infty(\boldsymbol{\theta}_0, \boldsymbol{\theta}), \epsilon)$ and $c_{m,\epsilon}(\mathbf{x}_n) = \int_{\mathbb{T}} \pi(\boldsymbol{\theta}) \, L_{m,\epsilon}(\mathbf{x}_n|\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta}$, we obtain our desired convergence via the dominated convergence theorem, because $w(d, \epsilon) \le a < \infty$. An application of a Slutsky-type theorem yields the almost sure convergence of the ratio between (11) and (12) to the right-hand side of (10), as $n \to \infty$. $\square$

The following result and proof guarantees the applicability of Theorem 2 to rejection ABC procedures, and to kernel-smoothed ABC procedures, as used in Jiang et al. (2018) and Park et al. (2016), respectively.

Proposition 1. *The result of Theorem 2 applies to rejection ABC and importance sampling ABC, with weight functions of respective forms (8) and (9).*

Proof. For weights of form (8), we note that $w(d, \epsilon) = [\![d < \epsilon]\!]$ is continuous in d at all points, other than when $d = \epsilon$. Furthermore, $w(d, \epsilon) \in \{0, 1\}$ and is hence non-negative and bounded. Thus, under the condition that $\mathcal{D}_\infty(\boldsymbol{\theta}_0, \boldsymbol{\theta}) \neq \epsilon$, we have the desired conclusion of Theorem 2.

For weights of form (9), we note that for fixed ϵ, $w(d, \epsilon)$ is continuous and positive in d. Since w is uniformly bounded by 1, differentiating with respect to d, we obtain $\mathrm{d}w/\mathrm{d}d = -(q/\epsilon) \, d^{q-1} \exp(-d^q/\epsilon)$, which is negative for any $d \ge 0$ and $q > 0$. Thus, (9) constitutes a weight function and satisfies the conditions of Theorem 2.

$\square$

4.2 Asymptotic of the energy statistic

Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be arbitrary elements of $\mathbf{X}_n$ and $\mathbf{Y}_m$, respectively. That is $\boldsymbol{X}$ and $\boldsymbol{Y}$ arise from DGPs that can be characterized by PDFs $f(\boldsymbol{x}; \boldsymbol{\theta}_0)$ and $f(\boldsymbol{y}; \boldsymbol{\theta})$, respectively. Under the assumption $\mathbb{E}\|\boldsymbol{X}\|_2 + \mathbb{E}\|\boldsymbol{Y}\|_2 < \infty$, Proposition 1 of Szekely & Rizzo (2013) states that we can write the ES as

$$\mathcal{E}(\boldsymbol{X}, \boldsymbol{Y}) = \frac{\Gamma\left(\frac{d+1}{2}\right)}{\pi^{(d+1)/2}} \int_{\mathbb{R}^d} \frac{|\varphi(\boldsymbol{t}; \boldsymbol{\theta}_0) - \varphi(\boldsymbol{t}; \boldsymbol{\theta})|^2}{\|\boldsymbol{t}\|_2^{d+1}} \mathrm{d}\boldsymbol{t}, \tag{13}$$

where $\varphi(\boldsymbol{t}; \boldsymbol{\theta})$ is the characteristic function corresponding to the PDF $f(\boldsymbol{y}; \boldsymbol{\theta})$.

We write $\log^+ x = \log(\max\{1, x\})$. From Szekely & Rizzo (2004) we have the fact that for arbitrary δ,

$$\mathcal{V}_\delta(\mathbf{X}_n, \mathbf{Y}_m) = \frac{1}{n^2m^2}\sum_{i_1=1}^{n}\sum_{i_2=1}^{n}\sum_{j_1=1}^{m}\sum_{j_2=1}^{m}\kappa_\delta(\boldsymbol{X}_{i_1}, \boldsymbol{X}_{i_2}; \boldsymbol{Y}_{j_1}, \boldsymbol{Y}_{j_2}),$$

where

$$\kappa_\delta(\boldsymbol{x}_{i_1}, \boldsymbol{x}_{i_2}; \boldsymbol{y}_{j_1}, \boldsymbol{y}_{j_2}) = \delta(\boldsymbol{x}_{i_1}, \boldsymbol{y}_{j_1}) + \delta(\boldsymbol{x}_{i_2}, \boldsymbol{y}_{j_2}) - \delta(\boldsymbol{x}_{i_1}, \boldsymbol{x}_{i_2}) - \delta(\boldsymbol{y}_{j_1}, \boldsymbol{y}_{j_2})$$

is the kernel of the V-statistic that is based on the function δ. The following result is a direct consequence of Theorem 1 of Sen (1977), when applied to V-statistics constructed from functionals δ that satisfy the hypothesis of Szekely & Rizzo (2013, Prop. 3).

Lemma 1. *Make the same assumptions regarding $\mathbf{X}_n$ and $\mathbf{Y}_m$ as in Theorem 2. Let $\delta(\boldsymbol{x}, \boldsymbol{y}) = \delta(\boldsymbol{x} - \boldsymbol{y})$ be a continuous and strictly negative definite function. If*

$$\mathbb{E}\left(|\kappa_\delta(\boldsymbol{X}_1, \boldsymbol{X}_2; \boldsymbol{Y}_1, \boldsymbol{Y}_2)| \log^+ |\kappa_\delta(\boldsymbol{X}_1, \boldsymbol{X}_2; \boldsymbol{Y}_1, \boldsymbol{Y}_2)|\right) < \infty, \tag{14}$$

then $\mathcal{V}_\delta(\mathbf{X}_n, \mathbf{Y}_m)$ converges almost surely to $\mathcal{E}_\delta(\boldsymbol{X}_1, \boldsymbol{Y}_1) \geq 0$, as $\min\{n, m\} \to \infty$, where $\boldsymbol{X}_1, \boldsymbol{X}_2 \in \mathbb{X}$ and $\boldsymbol{Y}_1, \boldsymbol{Y}_2 \in \mathbb{X}$ are arbitrary elements of $\mathbf{X}_n$ and $\mathbf{Y}_m$, respectively. Furthermore, $\mathcal{E}_\delta(\boldsymbol{X}_1, \boldsymbol{Y}_1) = 0$ if and only if $\boldsymbol{X}_1$ and $\boldsymbol{Y}_1$ are identically distributed.

We may apply the result of Lemma 1 directly to the case of $\delta = \delta_2$ in order to provide an almost sure convergence result regarding the V-statistic $\mathcal{V}_{\delta_2}(\mathbf{X}_n, \mathbf{Y}_m)$.

Corollary 1. *Make the same assumptions regarding $\mathbf{X}_n$ and $\mathbf{Y}_m$ as in Theorem 2. If $\boldsymbol{X}_1 \in \mathbb{X}$ and $\boldsymbol{Y}_1 \in \mathbb{X}$ are arbitrary elements of $\mathbf{X}_n$ and $\mathbf{Y}_m$, respectively, and*

$$\mathbb{E}\left(\|\boldsymbol{X}_1\|_2^2\right) + \mathbb{E}\left(\|\boldsymbol{Y}_1\|_2^2\right) < \infty, \tag{15}$$

and if $\min\{n, m\} \to \infty$, then $\mathcal{V}_{\delta_2}(\mathbf{X}_n, \mathbf{Y}_m)$ converges almost surely to $\mathcal{E}(\boldsymbol{X}_1, \boldsymbol{Y}_1)$, of form (13).

Proof. By the law of total expectation, we apply Lemma 1 by considering the two cases of (14): when $|\kappa_{\delta_2}| \leq 1$ and when $|\kappa_{\delta_2}| > 1$, separately, to write

$$\mathbb{E}\left(|\kappa_{\delta_2}| \log^+ |\kappa_{\delta_2}|\right) = p_0\mathbb{E}\left(|\kappa_{\delta_2}| \log^+ |\kappa_{\delta_2}| \,\big|\, |\kappa_{\delta_2}| \leq 1\right) + p_1\mathbb{E}\left(|\kappa_{\delta_2}| \log^+ |\kappa_{\delta_2}| \,\big|\, |\kappa_{\delta_2}| > 1\right), \tag{16}$$

where $p_0 = \mathbb{P}\left(|\kappa_{\delta_2}| \leq 1\right)$ and $p_1 = \mathbb{P}\left(|\kappa_{\delta_2}| > 1\right)$. The first term on the right-hand side of (16) is equal to zero, since $\log^+ |\kappa_{\delta_2}| = \log(1) = 0$, whenever $|\kappa_{\delta_2}| \leq 1$. Thus, we need only be concerned with bounding the second term.

For $|\kappa_{\delta_2}| > 1$, $|\kappa_{\delta_2}| \log |\kappa_{\delta_2}| \leq |\kappa_{\delta_2}|^2$, thus

$$\mathbb{E}\left(|\kappa_{\delta_2}| \log^+ |\kappa_{\delta_2}| \mid |\kappa_{\delta_2}| > 1\right) \leq \mathbb{E}\left(|\kappa_{\delta_2}|^2 \mid |\kappa_{\delta_2}| > 1\right)$$

The condition that $\mathbb{E}\left(|\kappa_{\delta_2}| \log^+ |\kappa_{\delta_2}|\right) < \infty$ is thus fulfilled if $\mathbb{E}\left(|\kappa_{\delta_2}|^2 \mid |\kappa_{\delta_2}| > 1\right) < \infty$, which is equivalent to

$$\mathbb{E}\left(|\kappa_{\delta_2}|^2\right) = p_0 \mathbb{E}\left(|\kappa_{\delta_2}|^2 \mid |\kappa_{\delta_2}| \leq 1\right) + p_1 \mathbb{E}\left(|\kappa_{\delta_2}|^2 \mid |\kappa_{\delta_2}| > 1\right) < \infty,$$

by virtue of the integrability of $\left\{|\kappa_{\delta_2}|^2 \mid |\kappa_{\delta_2}| \leq 1\right\}$ implying the existence of

$$\mathbb{E}\left(|\kappa_{\delta_2}|^2 \mid |\kappa_{\delta_2}| \leq 1\right),$$

since it is defined on a bounded support.

Next, by the triangle inequality, $|\kappa_{\delta_2}| \leq 2\left(\|\boldsymbol{X}_1\|_2 + \|\boldsymbol{X}_2\|_2 + \|\boldsymbol{Y}_1\|_2 + \|\boldsymbol{Y}_2\|_2\right)$, and hence

$$\begin{aligned} |\kappa_{\delta_2}|^2 \leq\ & 4\left(\|\boldsymbol{X}_1\|_2^2 + \|\boldsymbol{X}_2\|_2^2 + \|\boldsymbol{Y}_1\|_2^2 + \|\boldsymbol{Y}_2\|_2^2\right) \\ & + 8(\|\boldsymbol{X}_1\|_2 \|\boldsymbol{X}_2\|_2 + \|\boldsymbol{X}_1\|_2 \|\boldsymbol{Y}_1\|_2 + \|\boldsymbol{X}_1\|_2 \|\boldsymbol{Y}_2\|_2 \\ & + \|\boldsymbol{X}_2\|_2 \|\boldsymbol{Y}_1\|_2 + \|\boldsymbol{X}_2\|_2 \|\boldsymbol{Y}_2\|_2 + \|\boldsymbol{Y}_1\|_2 \|\boldsymbol{Y}_2\|_2). \end{aligned}$$

Since $\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{Y}_1, \boldsymbol{Y}_2$ are all pairwise independent, and $\boldsymbol{X}_1$ and $\boldsymbol{Y}_1$ are identically distributed to $\boldsymbol{X}_2$ and $\boldsymbol{Y}_2$, respectively, we have

$$\begin{aligned} \mathbb{E}\left(|\kappa_{\delta_2}|^2\right) \leq\ & 8\left[\mathbb{E}\left(\|\boldsymbol{X}_1\|_2^2\right) + \mathbb{E}\left(\|\boldsymbol{Y}_1\|_2^2\right)\right] + 8\left[\left(\mathbb{E}\|\boldsymbol{X}_1\|_2\right)^2 + \left(\mathbb{E}\|\boldsymbol{Y}_1\|_2\right)^2\right] \\ & + 32\left[\mathbb{E}\|\boldsymbol{X}_1\|_2 \mathbb{E}\|\boldsymbol{Y}_1\|_2\right], \end{aligned}$$

which concludes the proof since $\mathbb{E}\|\boldsymbol{X}_1\|_2^2 + \mathbb{E}\|\boldsymbol{Y}_1\|_2^2 < \infty$ is satisfied by the hypothesis and implies $\mathbb{E}\|\boldsymbol{X}_1\|_2 + \mathbb{E}\|\boldsymbol{Y}_1\|_2 < \infty$.

□

We note that condition (15) is stronger than a direct application of condition (14), which may be preferable in some situations. However, condition (15) is somewhat more intuitive and verifiable since it is concerned with the polynomial moments of norms and does not involve the piecewise function $\log^+ x$. It is also suggested in Zygmund (1951) that one may replace $\log^+ x$ by $\log(2+x)$ if it is more convenient to do so.

Combining the result of Theorem 2 with Corollary 1 and the conclusion from Proposition 1 of Szekely & Rizzo (2013) provided in Equation (13) yields the key result below. This result justifies the use of the V-statistic estimator $\mathcal{V}_{\delta_2}(\mathbf{X}_n, \mathbf{Y}_m)$ for the energy distance $\mathcal{E}(\boldsymbol{X}, \boldsymbol{Y})$ within the IS-ABC framework.

Corollary 2. *Under the assumptions of Corollary 1. If $\mathcal{D}(\mathbf{X}_n, \mathbf{Y}_m) = \mathcal{V}_{\delta_2}(\mathbf{X}_n, \mathbf{Y}_m)$, then the conclusion of Theorem 2 follows with*

$$\mathcal{D}(\mathbf{X}_n, \mathbf{Y}_m) \to \frac{\Gamma\left(\frac{d+1}{2}\right)}{\pi^{(d+1)/2}} \int_{\mathbb{R}^d} \frac{|\varphi(\boldsymbol{t}; \boldsymbol{\theta}_0) - \varphi(\boldsymbol{t}; \boldsymbol{\theta})|^2}{\|\boldsymbol{t}\|_2^{d+1}} \mathrm{d}\boldsymbol{t} = \mathcal{D}_\infty(\boldsymbol{\theta}_0, \boldsymbol{\theta}),$$

almost surely, as $n \to \infty$, where $\mathcal{D}_\infty(\boldsymbol{\theta}_0, \boldsymbol{\theta}) \geq 0$ and $\mathcal{D}_\infty(\boldsymbol{\theta}_0, \boldsymbol{\theta}) = 0$, if and only if $\boldsymbol{\theta}_0 = \boldsymbol{\theta}$.

5 Illustrations

We illustrate the use of the ES on some standard models. The standard rejection ABC algorithm is employed (that is, we use Algorithm 1 with weight function w of form (8)) for constructing estimators (5). The proposed ES is compared to the Kullback–Leibler divergence (KL), the Wasserstein distance (WA), and the maximum mean discrepancy (MMD). Here, the ES is applied using the Euclidean metric δ_2, the Wasserstein distance using the exponent $p = 2$ (cf. Bernton et al., 2019) and the MMD using a Gaussian kernel $\chi(\boldsymbol{x}, \boldsymbol{y}) = \exp[-(\boldsymbol{x} - \boldsymbol{y})^2]$. The Gaussian kernel is commonly used in the MMD literature, and was also considered for ABC in Park et al. (2016) and Jiang et al. (2018). Details regarding the use of the Kullback–Leibler divergence as a discrepancy function for ABC algorithms can be found in Jiang et al. (2018, Sec. 2).

We use $\boldsymbol{X} \sim \mathcal{L}$ to denote that the random variable $\boldsymbol{X}$ has probability law $\mathcal{L}$. Furthermore, we denote the normal law by $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ states that the DGP of $\boldsymbol{X}$ is multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. We further denote the uniform law, in the interval (a, b), for $a < b$, by $\mathrm{Unif}(a, b)$.

We consider examples explored in Jiang et al. (2018, Sec. 4.1). For each illustration below, we sample synthetic data of the same size m as the observed data size, n, whose value is specified for each model below. We consider only the rejection weight function, and the number of ABC iterations in Algorithm 1 is set to $N = 10^5$. The tuning parameter ϵ is set so that only the 0.05% smallest discrepancies are kept to form ABC posterior sample. We postpone the discussion of the results of our simulation experiments to Section 5.5

The experiments were implemented in R, using in particular the winference package (Bernton et al., 2019) and the FNN package (Beygelzimer et al., 2013). The Kullback–Leibler divergence between two PDFs is computed within the 1-nearest neighbor framework (Boltz et al., 2009). Moreover, the k-d trees is adopted for implementing the nearest neighbor search, which is the same as the method of Jiang et al. (2018). For estimating the 2-Wasserstein distance between two multivariate empirical measures, we propose to employ the swapping algorithm (Puccetti, 2017), which is simple to implement, and is more accurate and less computationally expensive than other algorithms commonly used in the literature (Bernton et al., 2019). Regarding the MMD, the same unbiased U-statistic estimator is adopted as given in Jiang et al. (2018) and Park et al. (2016). For reproduction of the the experimental results, the original source code can be accessed at `https://github.com/hiendn/Energy_Statistics_ABC`.

5.1 Bivariate Gaussian mixture model

Let $\mathbf{X}_n$ be a sequence of IID random variables, such that each $\boldsymbol{X}_i$ has mixture of Gaussian probability law

$$\boldsymbol{X}_i \sim p\mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) + (1-p)\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), \tag{17}$$

with known covariance matrices

$$\boldsymbol{\Sigma}_0 = \begin{bmatrix} 0.5 & -0.3 \\ -0.3 & 0.5 \end{bmatrix} \text{ and } \boldsymbol{\Sigma}_1 = \begin{bmatrix} 0.25 & 0 \\ 0 & 0.25 \end{bmatrix}.$$

We aim to estimate the generative parameters $\boldsymbol{\theta}^\top = (p, \boldsymbol{\mu}_0^\top, \boldsymbol{\mu}_1^\top)$ consisting of the mixing probability p and the population means $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_1$. To this end, we perform ABC using $n = 500$ observations, sampled from model (17) with $p = 0.3$, $\boldsymbol{\mu}_0^\top = (0.7, 0.7)$ and $\boldsymbol{\mu}_1^\top = (-0.7, -0.7)$. A kernel density estimate (KDE) of the ABC posterior distribution is presented in Figure 1.

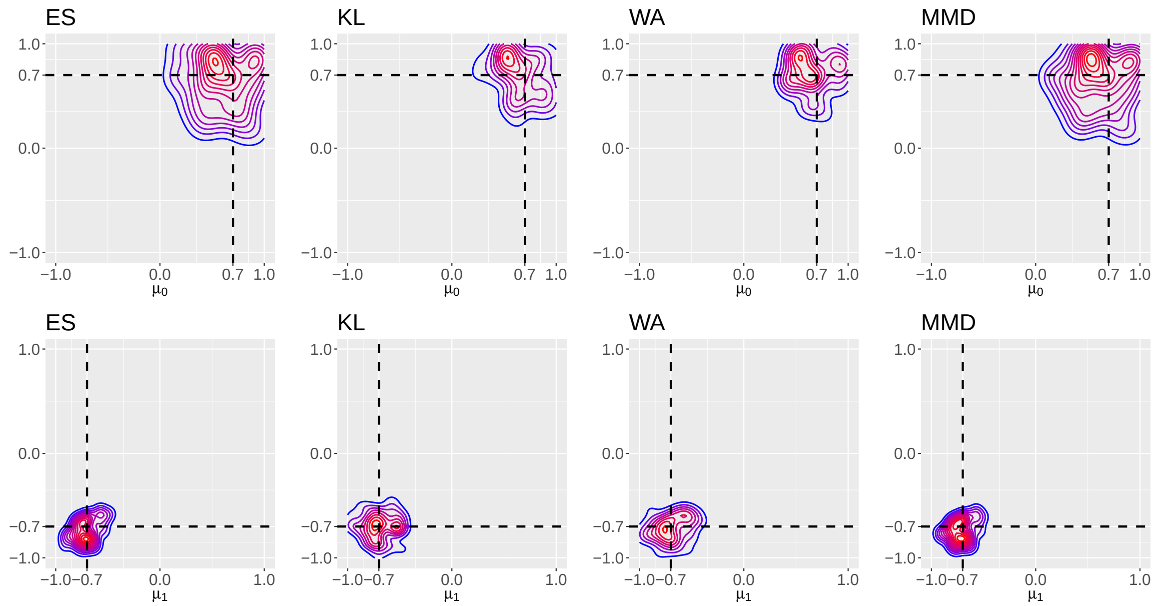

Figure 1: Marginal KDEs of the ABC posterior for the mean parameters $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_1$ of the bivariate Gaussian mixture model (17). The intersections of black dashed lines indicate the positions of the population means.

5.2 Moving-average model of order 2

The moving-average model of order q, MA(q), is a stochastic process $\{Y_t\}_{t\in\mathbb{N}^*}$ defined as

$$Y_t = Z_t + \sum_{i=1}^{q} \theta_i Z_{t-i},$$

with $\{Z_t\}_{t\in\mathbb{Z}}$ being a sequence of unobserved noise error terms. Jiang et al. (2018) used a MA(2) model for their benchmarking; namely $Y_t = Z_t + \theta_1 Z_{t-1} + \theta_2 Z_{t-2}$, $t \in [d]$. Each observation $\boldsymbol{Y}$ corresponds to a time series of length d. Here, we use the same model as that proposed in Jiang et al. (2018), where Z_t follows the Student-t distribution with 5 degrees of freedom, and $d = 10$. The priors on the model parameters θ_1 and θ_2 are taken to be uniform, that is, $\theta_1 \sim \text{Unif}(-2, 2)$ and $\theta_2 \sim \text{Unif}(-1, 1)$. We performed ABC using $n = 200$ samples generated from a model with the true parameter values $(\theta_1, \theta_2) = (0.6, 0.2)$. A KDE of the ABC posterior distribution is displayed in Figure 2.

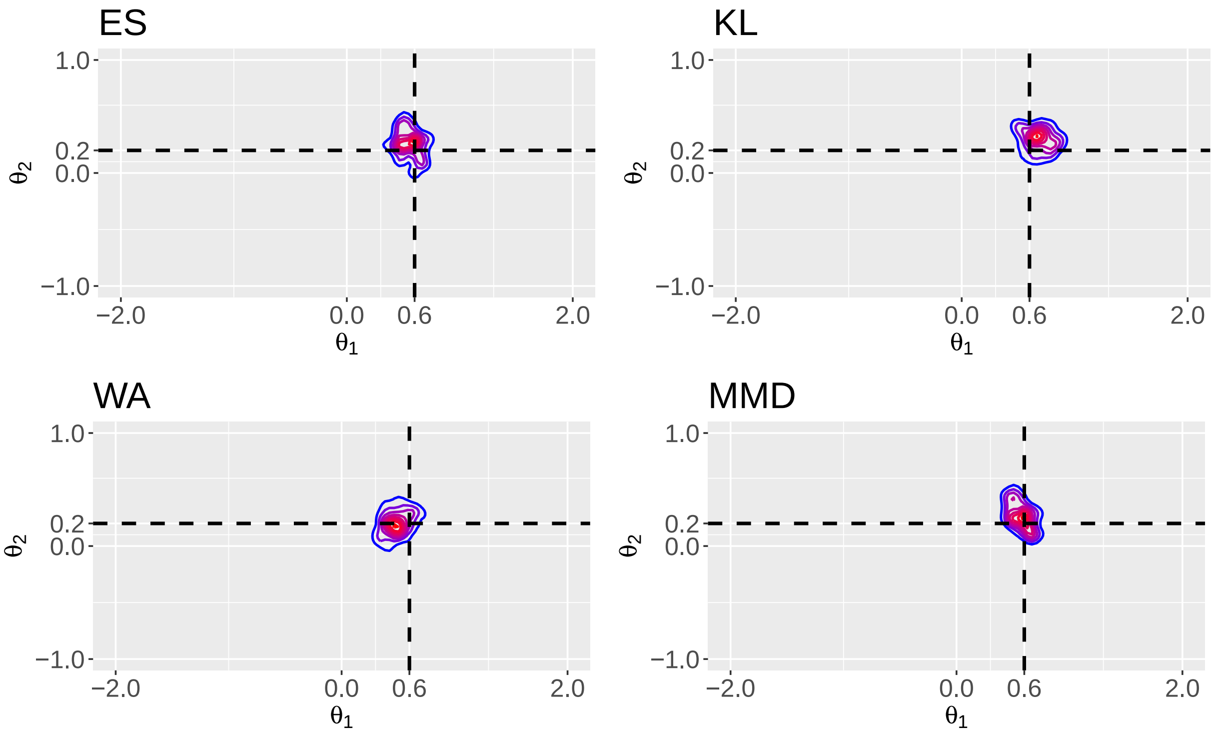

Figure 2: KDE of the ABC posterior for the parameters θ_1 and θ_2 of the MA(2) model experiment. The intersections of black dashed lines indicate the true parameter values.

5.3 Bivariate beta model

The bivariate beta model proposed by Crackel & Flegal (2017) is defined with five positive parameters $\theta_1, \ldots, \theta_5$ by letting

$$V_1 = \frac{U_1 + U_3}{U_5 + U_4}, \text{ and } V_2 = \frac{U_2 + U_4}{U_5 + U_3}, \tag{18}$$

where $U_i \sim \text{Gamma}(\theta_i, 1)$, for $i \in [5]$, and setting $Z_1 = V_1/(1 + V_1)$ and $Z_2 = V_2/(1 + V_2)$. The bivariate random variable $\boldsymbol{Z}^\top = (Z_1, Z_2)$ has marginal laws $Z_1 \sim \text{Beta}(\theta_1 + \theta_3, \theta_5 + \theta_4)$ and $Z_2 \sim \text{Beta}(\theta_2 + \theta_4, \theta_5 + \theta_3)$. We performed ABC using samples of size $n = 500$, which are generated from a DGP with true parameter values $(\theta_1, \theta_2, \theta_3, \theta_4, \theta_5) = (1, 1, 1, 1, 1)$. The prior on each of the model parameters is taken to be independent $\text{Unif}(0, 5)$. A KDE of the ABC posterior distribution is displayed in Figure 3.

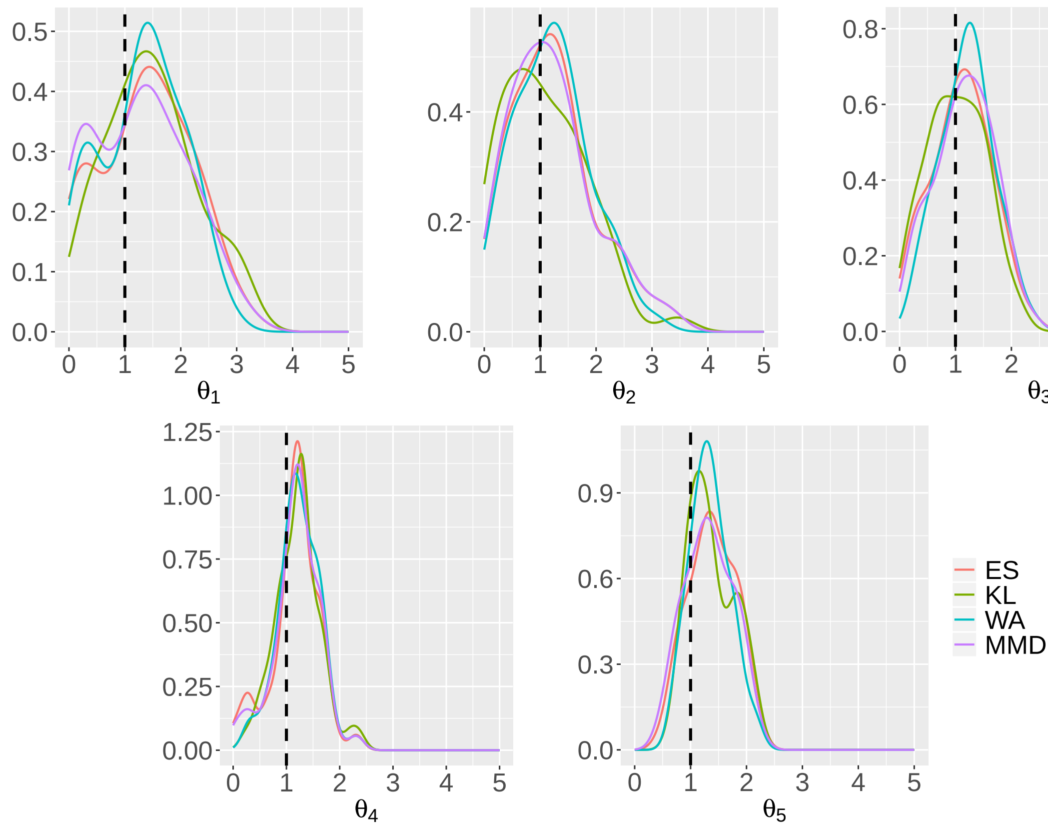

Figure 3: Marginal KDEs of the ABC posterior for the parameters $\theta_1, \ldots, \theta_5$ for the bivariate beta model. The black dashed lines indicate the true parameter values.

5.4 Multivariate g-and-k distribution

A univariate g-and-k distribution can be defined via its quantile function (Drovandi & Pettitt, 2011):

$$F^{-1}(x) = A + B\left[1 + 0.8\frac{1-\exp(-g \times z_x)}{1+\exp(-g \times z_x)}\right]\left(1+z_x^2\right)^k z_x, \tag{19}$$

where parameters (A, B, g, k) respectively relate to location, scale, skewness, and kurtosis. Here, z_x is the xth quantile of the standard normal distribution. Given a set of parameters (A, B, g, k), it is easy to simulate d observations of a DGP with quantile function (19), by generating a sequence of IID sample $\{Z_i\}_{i=1}^d$, where $Z_i \sim \mathcal{N}(0, 1)$, for $i \in [d]$.

A so-called d-dimensional g-and-k DGP can instead be defined by applying the quantile function (19) to each of the d elements of a multivariate normal vector $\boldsymbol{Z}^\top = (Z_1, ..., Z_d) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is a covariance matrix. In our experiment, we use a 5-dimensional g-and-k model with the same covariance matrix and parameter values for (A, B, g, k) as that considered by Jiang et al. (2018). That is, we generate samples of size $n = 200$ from a g-and-k DGP with the true parameter values $(A, B, g, k) = (3, 1, 2, 0.5)$ and the covariance matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & \rho & 0 & 0 & 0 \\ \rho & 1 & \rho & 0 & 0 \\ 0 & \rho & 1 & \rho & 0 \\ 0 & 0 & \rho & 1 & \rho \\ 0 & 0 & 0 & \rho & 1 \end{bmatrix},$$

where $\rho = -0.3$. Marginal KDEs of the ABC posterior distributions is presented in Figure 4.

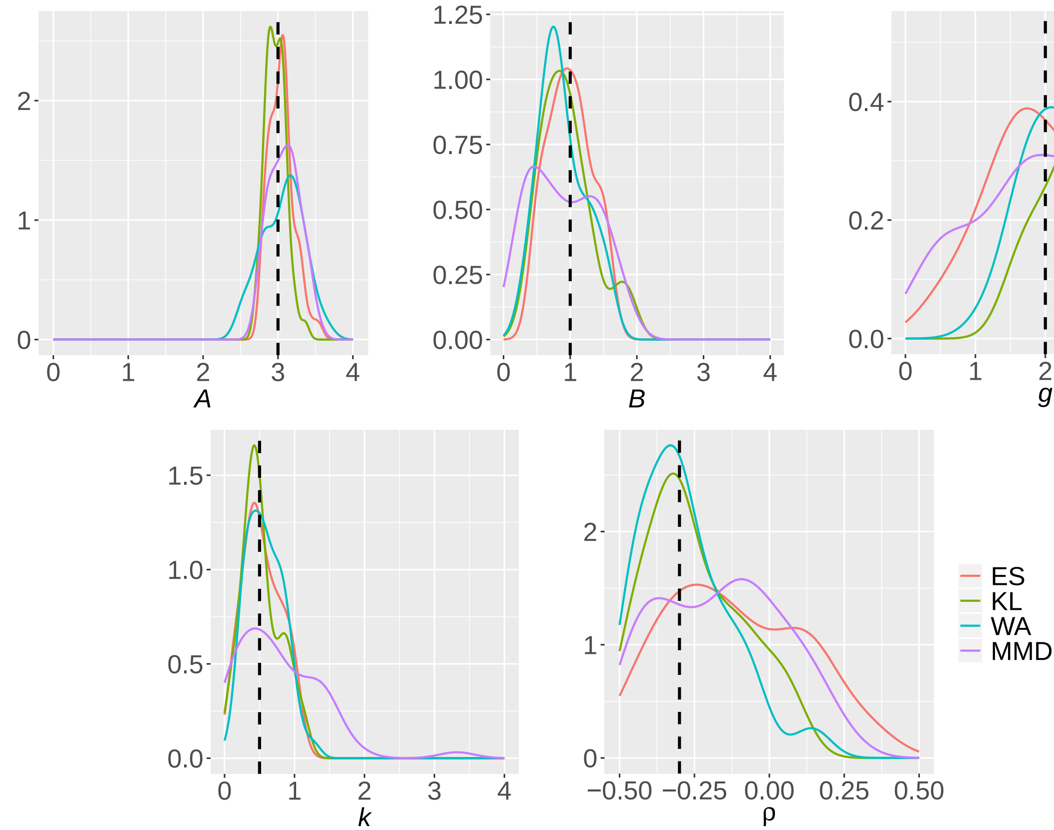

Figure 4: Marginal KDEs of the ABC posterior for the parameters A, B, g, k and ρ of the g-and-k model. The black dashed lines indicate the true parameter values.

5.5 Discussion of the results and performance

For each of the four experiments and each parameter, we computed the posterior mean $\hat{\theta}_{\mathrm{mean}}$, posterior median $\hat{\theta}_{\mathrm{med}}$, mean absolute error and mean squared error defined by

$$\mathrm{MAE} = \frac{1}{M}\sum_{k=1}^{M} |\theta_k - \theta_0|, \text{ and } \quad \mathrm{MSE} = \frac{1}{M}\sum_{k=1}^{M} |\theta_k - \theta_0|^2,$$

where $\{\theta_k\}_{k=1}^{M}$ denotes the pseudo-posterior sample and θ_0 denotes the true parameter. Here $M = 50$ since $N = 10^5$ and ϵ is chosen as to retain 0.05% of the samples. Each experiment was replicated ten times by keeping the same fixed (true) values for the parameters and by sampling new observed data each of the ten times. The estimated quantities $\hat{\theta}_{\mathrm{mean}}$, $\hat{\theta}_{\mathrm{med}}$, and errors MAE and $\mathrm{RMSE} = \mathrm{MSE}^{1/2}$ were then averaged over the ten replications, and are reported along with standard deviations $\sigma(\cdot)$ in columns associated with each estimator and true values θ_0 for each parameter in Tables 1, 2, 3 and 4.

Upon inspection, Tables 1, 2, 3 and 4 showed some advantage in performance from WA on the bivariate Gaussian mixtures, some advantage from the MMD on the bivariate beta model, and some advantage from the ES on the g-and-k model, while multiple methods are required to make the best inference in the case of the MA(2) experiment. When we further take into account the standard deviations of the estimators, we observe that all four data discrepancy measures essentially perform comparatively well across the four experimental models. Thus, we may conclude that there is no universally best performing discrepancy measure, and one must choose the right method for each problem of interest. Alternatively, one may also consider some kind of averaging over the results of the different discrepancy measures. We have not committed to an investigation of such methodologies and leave it as a future research direction.

Table 1: Estimation performance for bivariate Gaussian mixtures (Section 5.1). The best results in each column is highlighted in boldface.

		$\hat{\theta}_{\text{mean}}$	$\sigma(\hat{\theta}_{\text{mean}})$	$\hat{\theta}_{\text{med}}$	$\sigma(\hat{\theta}_{\text{med}})$	MAE	σ(MAE)	RMSE	σ(RMSE)
$\mu_{00} = 0.7$	ES	0.594	0.045	0.607	0.063	0.215	0.030	0.283	0.055
	KL	0.648	0.039	0.666	0.048	0.165	0.016	0.205	0.026
	WA	**0.675**	0.035	**0.682**	0.043	**0.152**	0.020	**0.181**	0.021
	MMD	0.564	0.079	0.582	0.076	0.234	0.054	0.311	0.101
$\mu_{01} = 0.7$	ES	0.587	0.063	0.613	0.059	0.215	0.038	0.282	0.069
	KL	0.651	0.042	0.667	0.061	0.169	0.022	0.210	0.027
	WA	**0.655**	0.050	**0.669**	0.047	**0.152**	0.015	**0.187**	0.019
	MMD	0.559	0.076	0.598	0.075	0.235	0.049	0.313	0.092
$\mu_{10} = -0.7$	**ES**	**-0.699**	0.046	-0.716	0.040	1.401	0.043	1.412	0.039
	KL	-0.709	0.029	-0.712	0.035	1.409	0.029	1.415	0.029
	WA	**-0.699**	0.030	**-0.704**	0.037	**1.399**	0.030	**1.404**	0.030
	MMD	-0.709	0.054	-0.731	0.036	1.411	0.051	1.422	0.038
$\mu_{11} = -0.7$	**ES**	**-0.696**	0.058	-0.712	0.043	1.396	0.058	1.407	0.049
	KL	-0.711	0.047	**-0.704**	0.057	1.411	0.047	1.416	0.047
	WA	-0.695	0.043	-0.695	0.053	**1.395**	0.043	**1.401**	0.043
	MMD	-0.711	0.066	-0.726	0.046	1.411	0.066	1.424	0.052

Table 2: Estimation performance for the MA(2) model (Section 5.2). The best results in each column is highlighted in boldface.

		$\hat{\theta}_{\text{mean}}$	$\sigma(\hat{\theta}_{\text{mean}})$	$\hat{\theta}_{\text{med}}$	$\sigma(\hat{\theta}_{\text{med}})$	MAE	σ(MAE)	RMSE	σ(RMSE)
$\theta_1 = 0.6$	ES	0.569	0.042	0.570	0.045	0.083	0.015	0.100	0.017
	KL	0.664	0.028	0.658	0.031	0.106	0.017	0.132	0.019
	WA	0.509	0.033	0.505	0.038	0.112	0.022	0.133	0.026
	MMD	**0.583**	0.044	**0.586**	0.048	**0.079**	0.013	**0.096**	0.015
$\theta_2 = 0.2$	ES	0.215	0.035	0.219	0.035	0.111	0.015	0.135	0.019
	KL	0.274	0.023	0.280	0.027	0.110	0.014	0.134	0.014
	WA	**0.205**	0.025	**0.207**	0.030	**0.090**	0.029	**0.112**	0.034
	MMD	0.220	0.037	0.220	0.036	0.108	0.010	0.132	0.012

Table 3: Estimation performance for the bivariate beta model (Section 5.3). The best results in each column is highlighted in boldface.

		$\hat{\theta}_{\text{mean}}$	$\sigma(\hat{\theta}_{\text{mean}})$	$\hat{\theta}_{\text{med}}$	$\sigma(\hat{\theta}_{\text{med}})$	MAE	σ(MAE)	RMSE	σ(RMSE)
$\theta_1 = 1.0$	ES	1.299	0.223	1.189	0.264	0.713	0.130	0.885	0.165
	KL	1.389	0.190	1.333	0.165	0.696	0.151	0.877	0.205
	WA	**1.286**	0.220	1.193	0.265	0.672	0.128	**0.828**	0.153
	MMD	1.229	0.188	**1.143**	0.241	**0.676**	0.092	0.836	0.121
$\theta_2 = 1.0$	ES	1.362	0.185	1.290	0.237	0.716	0.118	0.904	0.131
	KL	**1.235**	0.152	**1.153**	0.170	**0.588**	0.070	**0.745**	0.097
	WA	1.292	0.196	1.240	0.241	0.657	0.114	0.817	0.139
	MMD	1.268	0.173	1.170	0.171	0.669	0.103	0.841	0.131
$\theta_3 = 1.0$	ES	1.170	0.132	1.183	0.157	0.459	0.045	0.552	0.049
	KL	**1.083**	0.100	**1.077**	0.088	**0.394**	0.034	**0.496**	0.045
	WA	1.229	0.118	1.216	0.132	0.426	0.054	0.521	0.059
	MMD	1.181	0.116	1.182	0.143	0.456	0.051	0.548	0.061
$\theta_4 = 1.0$	**ES**	**1.128**	0.112	1.113	0.138	0.435	0.032	0.534	0.045
	KL	1.133	0.111	**1.086**	0.135	**0.390**	0.038	**0.498**	0.051
	WA	1.218	0.110	1.196	0.108	0.409	0.049	0.514	0.066
	MMD	1.150	0.098	1.133	0.130	0.423	0.041	0.518	0.049
$\theta_5 = 1.0$	ES	1.343	0.096	1.360	0.104	0.428	0.052	0.514	0.059
	KL	1.300	0.087	1.250	0.065	0.384	0.040	0.491	0.061
	WA	1.300	0.101	1.298	0.105	**0.370**	0.058	**0.446**	0.066
	MMD	**1.258**	0.115	**1.232**	0.120	0.375	0.055	0.454	0.063

Table 4: Estimation performance for the g-and-k distribution (Section 5.4). The best results in each column is highlighted in boldface.

		$\hat{\theta}_{\text{mean}}$	$\sigma(\hat{\theta}_{\text{mean}})$	$\hat{\theta}_{\text{med}}$	$\sigma(\hat{\theta}_{\text{med}})$	MAE	σ(MAE)	RMSE	σ(RMSE)
$A = 3.0$	**ES**	**3.024**	0.044	**3.009**	0.047	0.133	0.016	0.170	0.018
	KL	2.955	0.030	2.948	0.033	**0.105**	0.013	**0.128**	0.013
	WA	3.043	0.045	3.052	0.067	0.232	0.020	0.277	0.020
	MMD	3.081	0.061	3.062	0.065	0.177	0.029	0.221	0.036
$B = 1.0$	**ES**	**1.046**	0.062	**1.027**	0.079	**0.268**	0.024	**0.322**	0.029
	KL	0.918	0.071	0.885	0.068	0.313	0.026	0.375	0.029
	WA	0.894	0.127	0.869	0.136	0.277	0.044	0.334	0.045
	MMD	0.899	0.069	0.855	0.079	0.374	0.029	0.440	0.030
$g = 2.0$	ES	2.289	0.101	2.264	0.210	0.872	0.098	1.026	0.091
	KL	2.993	0.080	3.046	0.121	1.043	0.070	1.193	0.066
	WA	2.581	0.101	2.599	0.147	**0.858**	0.078	**1.025**	0.075
	MMD	**2.184**	0.128	**2.227**	0.190	0.904	0.103	1.052	0.100
$k = 0.5$	**ES**	**0.476**	0.046	0.444	0.067	0.225	0.014	0.270	0.015
	KL	0.550	0.059	**0.498**	0.064	0.252	0.029	0.317	0.045
	WA	0.544	0.095	0.526	0.094	**0.189**	0.035	**0.238**	0.046
	MMD	0.691	0.056	0.621	0.072	0.380	0.041	0.502	0.070
$\rho = -0.3$	ES	-0.163	0.047	-0.178	0.069	0.197	0.032	0.246	0.034
	KL	**-0.291**	0.034	-0.324	0.037	**0.117**	0.014	**0.144**	0.020
	WA	-0.288	0.026	**-0.314**	0.035	0.125	0.016	0.152	0.020
	MMD	-0.194	0.047	-0.210	0.063	0.174	0.030	0.218	0.035

6 Conclusion

We have introduced a novel importance-sampling ABC algorithm that is based on the so-called *two-sample energy statistic*. Along with other data discrepancy measures that view data sets as empirical measures, such as the Kullback–Leibler divergence, the Wasserstein distance and maximum mean discrepancies, our proposed approach bypasses the cumbersome use of summary statistics.

We have shown that the V-statistic estimator of the ES is consistent under mild moment conditions. Furthermore, we have established a new asymptotic result for cases when the observed sample and simulated sample sizes increasing to infinity, that shows a kind of consistency of the pseudo-posterior in the infinite data scenario. This is in concordance with previous results in such cases (see for instance Jiang et al., 2018; Bernton et al., 2019) and extends upon existing theory for the application in the general IS-ABC framework.

Illustrations of the proposed ES-ABC algorithm on four experimental models have shown that it performs comparatively well to alternative discrepancy measures. Considering computing costs, KL should be preferred over the other three discrepancy measures, with a *linearithmic* computational time of $\mathcal{O}((n+m)\log(n+m))$. This can be contrasted against the quadratic time $\mathcal{O}((n+m)^2)$ for a single computation of $\mathcal{D}\left(\mathbf{X}_n, \mathbf{Y}_m\right)$ when we consider the Wasserstein distance, instead. Both the ES and MMD estimators require quadratic computational time, like the Wasserstein distance. We note that linear time estimators are also available for the MMD and the ES, although these are unbiased and cannot be guaranteed to be positive (see Lemma 14 in Gretton et al., 2012).

In the rejection ABC setting, Proposition 2 of Bernton et al. (2019) shows that under some regularity assumptions on the DGP and if the data discrepancy measure satisfies the condition:

$$\mathcal{D}\left(\mathbf{X}_n, \mathbf{Y}_n\right) = 0 \text{ if and only if } \mathbf{X}_n = \mathbf{Y}_n, \tag{20}$$

then the ABC pseudo-posterior contracts to the posterior distribution as the rejection threshold ϵ converges to zero. It can be shown that the V-statistic estimator of the ES only satisfies the *only if* direction of (20) and thus does not necessarily enjoy the conclusions of Proposition 2 of Bernton et al. (2019). The condition is not known to be necessary and thus we do not know if the conclusion can be satisfied in another way. We observe, from our simulation experiments, that ES did not perform differently to the Wasserstein distance, which can be shown to satisfy Proposition 2 of Bernton et al. (2019).

Acknowledgments

Hien Nguyen is funded by Australian Research Council (ARC) grants DE170101134 and DP180101192, and a La Trobe University startup grant.

References

Barber, D. (2012). *Bayesian Reasoning and Machine Learning*. Cambridge: Cambridge University Press.

Baringhaus, L. & Franz, C. (2004). On a new multivariate two-sample test. *Journal of Multivariate Analysis*, 88, 190–206.

Bernton, E., Jacob, P. E., Gerber, M., & Robert, C. P. (2019). Approximate Bayesian computation with the Wasserstein distance. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81, 235–269.

Beygelzimer, A., Kakadet, S., Langford, J., Arya, S., Mount, D., & Li, S. (2013). *FNN: Fast Nearest Neighbor Search miller and Applications*. R package version 1.1.3.

Boltz, S., Debreuve, É., & Barlaud, M. (2009). High-dimensional statistical measure for region-of-interest tracking. *IEEE Transactions on Image Processing*, 18, 1266–1283.

Crackel, R. & Flegal, J. (2017). Bayesian inference for a flexible class of bivariate beta distributions. *Journal of Statistical Computation and Simulation*, 87, 295–312.

DasGupta, A. (2011). *Probability for Statistics and Machine Learning: Fundamentals and Advance*. New York: Springer.

Dellacherie, C. & Meyer, P. (1980). *Probability and Potential B: Theory of Martingales*. Amsterdam: North-Holland.

Drovandi, C. C. & Pettitt, A. N. (2011). Likelihood-free Bayesian estimation of multivariate quantile distributions. *Computational Statistics and Data Analysis*, 55, 2541–2556.

Ghosh, J. K., Delampady, M., & Samanta, T. (2006). *An Introduction to Bayesian Analysis: Theory and Methods*. New York: Springer.

Ghosh, J. K. & Ramamoorthi, R. V. (2003). *Bayesian Nonparametrics*. New York: Springer.

Ghosh, S. & van der Vaart, A. (2017). *Fundamentals of Nonparametric Bayesian Inference*. Cambridge: Cambridge University Press.

Graham, R. L., Knuth, D. E., & Patashnik, O. (1994). *Concrete Mathematics*. Reading: Addison-Wesley.

Gretton, A., Bogwardt, K. M., Rasch, M., Scholkopf, B., & Smola, A. J. (2007). A kernel method for the two-sample-problem. In *Advances in Neural Information Processing Systems*.

Gretton, A., Bogwardt, K. M., Rasch, M. J., Scholkopf, B., & Smola, A. (2012). A kernel two-sample test. *Journal of Machine Learning Research*, 13, 723–773.

Gretton, A., Fukumizu, K., Harchaoui, Z., & Sriperumbudur, B. K. (2009). A fast, consistent kernel two-sample test. In *Advances in Neural Information Processing Systems*.

Hjort, N. L., Holmes, C., Muller, P., & Walker, S. G. (2010). *Bayesian Nonparametrics*. Cambridge: Cambridge University Press.

Jiang, B., Wu, T.-Y., & Wong, W. H. (2018). Approximate Bayesian computation with Kullback-Leibler divergence as data discrepancy. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Karabatsos, G. & Leisen, F. (2018). An approximate likelihood perspective on ABC methods. *Statistical Surveys*, 12, 66–104.

Koch, K.-R. (2007). *Introduction to Bayesian Statistics*. Heidelberg: Springer.

Koop, G., Poirier, D. J., & Tobias, J. L. (2007). *Bayesian Econometric Methods*. Cambridge: Cambridge University Press.

Koroljuk, V. S. & Borovskich, Y. V. (1994). *Theory of U-Statistics*. Dordrecht: Springer.

Lintusaari, J., Gutmann, M. U., Dutta, R., Kaski, S., & Corander, J. (2017). Fundamentals and recent developments in approximate Bayesian computation. *Systems Biology*, 60, e60–e82.

Marin, J.-M., Pudlo, P., Robert, C. P., & Ryder, R. J. (2012). Approximate Bayesian computation methods. *Statistics and Computing*, 22, 1167–1180.

Miller, J. W. & Dunson, D. B. (2018). Robust Bayesian inference via coarsening. *Journal of the American Statistical Association*, 113, 340–356.

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. Cambridge: MIT Press.

Park, M., Jitkrittum, W., & Sejdinovic, D. (2016). K2-ABC: approximate Bayesian computation with kernel embeddings. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Press, S. J. (2003). *Subjective and Objective Bayesian Statistics: Principles, Models, and Applications*. Hoboken: Wiley.

Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A., & Feldman, M. W. (1999). Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular Biology and Evolution*, 16, 1791–1798.

Puccetti, G. (2017). An algorithm to approximate the optimal expected inner product of two vectors with given marginals. *Journal of Mathematical Analysis and Applications*, 451(1), 132 – 145.

Robert, C. P. (2007). *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. New York: Springer.

Rubin, D. B. (1984). Bayesian justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics*, 12, 1151–1172.

Sejdinovic, D., Sriperumbudur, B., Gretton, A., & Fukumizu, K. (2013). Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *Annals of Statistics*, 41, 2263–2291.

Sen, P. K. (1977). Almost sure convergence of generalized U-statistics. *Annals of Probability*, 5, 287–290.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. New York: Wiley.

Sisson, S. A., Fan, Y., & Beaumont, M. A., Eds. (2019). *Handbook of Approximate Bayesian Computation*. Boca Raton: CRC Press.

Szekely, G. J. & Rizzo, M. L. (2004). Testing for equal distributions in high dimension. *InterStat*, 5, 1–16.

Szekely, G. J. & Rizzo, M. L. (2013). Energy statistics: a class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143, 1249–1272.

Szekely, G. J. & Rizzo, M. L. (2017). The energy of data. *Annual Review of Statistics and Its Application*, 4(447-479).

Tavaré, S. (2019). On the history of ABC. In S. A. Sisson, Y. Fan, & M. A. Beaumont (Eds.), *Handbook of Approximate Bayesian Computation*. Boca Raton: CRC Press.

Tavaré, S., Balding, D. J., Griffiths, R. C., & Donnelly, P. (1997). Inferring coalescence times from DNA sequence data. *Genetics*, 145, 505–518.

Voss, J. (2014). *An Introduction to Statistical Computing: A Simulation-based Approach*. Chichester: Wiley.

Zygmund, A. (1951). An individual ergodic theorem for non-comutative transformations. *Acta Scientiarum Mathematicarum (Szeged)*, 14, 103–110.