Approximating Predictive Probabilities of Gibbs-Type Priors

Published version:

DOI:10.1007/s13171-019-00187-y

Approximate Bayesian computation via the energy statistic

HIEN D. NGUYEN[1], JULYAN ARBEL[2],
HONGLIANG LÜ[2], FLORENCE FORBES[2]

[1]Department of Mathematics and Statistics, La Trobe University, Bundoora Melbourne 3066, Victoria Australia. (e-mail: h.nguyen5@latrobe.edu.au) [2]Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France

July 1, 2020

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 pseudo-posterior by comparing observed data with simulated data, traditionally based on some summary statistics, the elicitation of which is regarded as a key difficulty. Recently, using data discrepancy measures has been proposed in order to bypass the construction of summary statistics. Here we propose to use the importance-sampling ABC (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, and prove that the rejection ABC algorithm, based on the energy statistic, generates pseudo-posterior distributions that achieves convergence to the correct limits, when implemented with rejection thresholds that converge to zero, in the finite sample setting. 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.

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), Murphy (2012). When conducting parametric Bayesian inference, we observe some realization $\boldsymbol{x}$ 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 $\boldsymbol{\theta} \in \mathbb{T}$. Using expert knowledge, or based on computational considerations such as conjugacy, we endow the parameter $\boldsymbol{\theta}$ with some prior PDF $\pi(\boldsymbol{\theta})$. The goal of Bayesian inference is then 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 the prior predictive distribution $c(\boldsymbol{x})$ is defined by

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

In very simple cases, such as cases when the prior PDF is a conjugate of the likelihood (cf. Sec. 3.3 of Robert (2007)), the posterior PDF (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 Ch. 6 of Press (2003).

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), 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), Lintusaari et al. (2017), Karabatsos and Leisen (2018). The volume Sisson et al. (2019) provides a comprehensive treatment regarding ABC methodologies.

The core philosophy of ABC is to define a pseudo-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. Recent examples of such an approach include the use of the maximum mean discrepancy (MMD) (Park et al., 2016), Kullback–Leibler divergence (Jiang et al., 2018), and the Wasserstein distance (Bernton et al., 2019). Furthermore, Jiang et al. (2018) also considered the use of the classification accuracy method of Gutmann et al. (2018), and the indirect inference method of Drovandi et al. (2015), in the data discrepancy context.

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 and Leisen (2018). In particular, we report on a class of ABC algorithms that utilize the two-sample energy statistic (ES) of Szekely and Rizzo (2004) (see also Baringhaus and Franz (2004), Szekely and Rizzo (2013, 2017), Mak and Joseph (2018)). Our approach is related to the MMD ABC algorithms that were implemented in Park et al. (2016), Jiang et al. (2018), 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, we make the assumption that the data discrepancy measure converges to some asymptotic value $\mathcal{D}_\infty(\boldsymbol{\theta}_0, \boldsymbol{\theta})$, where $\boldsymbol{\theta}_0$ stands for the 'true' parameter value associated to the DGP that generates observations $\boldsymbol{X}$. We then show that the pseudo-posterior distribution converges almost surely to a distribution depending on the prior π and on the limiting value $\mathcal{D}_\infty(\boldsymbol{\theta}_0, \boldsymbol{\theta})$. In addition to our asymptotic results regarding large sample scenarios, we also provide corollaries regarding the performance of our ES-based ABC method, due to the general finite sample theoretical results of Bernton et al. (2019). Our asymptotic results provide useful approximations and guarantees for the practical application of our method.

The last decade has seen an active development in research on asymptotic properties of ABC. Early works revolved around the impact of the acceptance threshold on the ABC bias and the Monte Carlo error (Blum, 2010a, Barber et al., 2015, Biau et al., 2015), and on the choice of summary statistics (Blum, 2010a, Fearnhead and Prangle, 2012, Prangle et al., 2014). Further works focused on large sample size properties such as consistency for model choice (Marin et al., 2014), asymptotic efficiency (Li and Fearnhead, 2018), posterior consistency, and contraction rates (Frazier et al., 2018). It is with these results, where our article fits. Although devised in settings where likelihoods are assumed intractable, ABC can also be cast in the setting of robustness with respect to misspecification (Frazier et al., 2020). In particular, the ABC posterior distribution can be viewed as a special case of a coarsened posterior distribution (Miller and 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 $\boldsymbol{\theta}$ 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 $\boldsymbol{\theta} \in \mathbb{T}$, we at least have the capability of sampling from the DGP with PDF $f(\boldsymbol{x}|\boldsymbol{\theta})$. 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

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

Typically, one should choose $m = n$, as it fulfils the hypotheses of all of our proved theoretical results. This choice is made throughout all of our numerical demonstrations. However, we anticipate that there may be practical or computational scenarios, where it may be advantageous to be able to choose $m \neq n$, which is permissible in our methodological framework.

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

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 space endowed with a semi-metric δ, where $d \in \mathbb{N}$ (cf. Sejdinovic et al. (2013)). 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 and Franz (2004) and Szekely and Rizzo (2004), as a function of $\boldsymbol{X}$ and $\boldsymbol{Y}$, via the expression $\mathcal{E}_{\delta_1}(\boldsymbol{X}, \boldsymbol{Y})$, where $\delta_\beta(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|_2^\beta$ is the β power of the metric corresponding to the L_2-norm ($\beta \in (0, 2]$; cf. (Szekely and Rizzo, 2013, Prop. 2)). Thus, the original ES statistic, which we shall also denote as $\mathcal{E}(\boldsymbol{X}, \boldsymbol{Y})$, is defined using the Euclidean metric δ_1.

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

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

in Proposition 1 of Szekely and Rizzo (2013), where $\Gamma\left(\cdot\right)$ 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 $\mathcal{E}\left(\boldsymbol{X},\boldsymbol{Y}\right) \geq 0$ for any $\boldsymbol{X},\boldsymbol{Y} \in \mathbb{X}$, and $\mathcal{E}\left(\boldsymbol{X},\boldsymbol{Y}\right) = 0$ if and only if $\boldsymbol{X}$ and $\boldsymbol{Y}$ are identically distributed.

The result above is generalized in Proposition 3 of Szekely and Rizzo (2013), where we have the following statement. If $\delta\left(\boldsymbol{x},\boldsymbol{y}\right) = \delta\left(\boldsymbol{x}-\boldsymbol{y}\right)$ is a continuous function and $\boldsymbol{X},\boldsymbol{Y} \in \mathbb{R}^d$ are independent random variables, then it is necessary and sufficient that $\delta\left(\cdot\right)$ is strictly negative definite (see Szekely and Rizzo (2013) for the precise definition) for the following conclusion to hold: $\mathcal{E}_\delta\left(\boldsymbol{X},\boldsymbol{Y}\right) \geq 0$ for any $\boldsymbol{X},\boldsymbol{Y} \in \mathbb{X}$, and $\mathcal{E}_\delta\left(\boldsymbol{X},\boldsymbol{Y}\right) = 0$ 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_1, since it is the most well known and popular of the varieties.

3.1 The V-statistic estimator

Suppose that we observe $\mathbf{X}_n = \left\{\boldsymbol{X}_i\right\}_{i=1}^n$ and $\mathbf{Y}_m = \left\{\boldsymbol{Y}_i\right\}_{i=1}^m$, 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 $\delta\left(\boldsymbol{x},\boldsymbol{y}\right) < \infty$, the so-called V-statistic estimator (cf. (Serfling, 1980, Ch. 5) and Koroljuk and Borovskich (1994))

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

can be proved to converge in probability to $\mathcal{E}_\delta\left(\boldsymbol{X},\boldsymbol{Y}\right)$, 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)). Here, the proof was provided in the context of MMDs (see definition in Section 3.3) but is easily portable to the ES setting.

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 $\delta\left(\boldsymbol{x},\boldsymbol{y}\right) = \delta\left(\boldsymbol{x}-\boldsymbol{y}\right)$, we have $\mathcal{V}_\delta\left(\mathbf{X}_n,\mathbf{Y}_m\right) \geq 0$.

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 $d = \mathcal{D}\left(\mathbf{X}_n,\mathbf{Y}_m\right) \geq 0$, a weighting function $w\left(d,\epsilon\right) \geq 0$, 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

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

in Algorithm 1.

In particular, we consider original ES, where $\delta = \delta_1$. 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 Mercer kernel function $\chi\left(\boldsymbol{x}, \boldsymbol{y}\right)$ $\left(\boldsymbol{x}, \boldsymbol{y} \in \mathbb{X}\right)$, the corresponding MMD is defined via the equation

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

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). 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 property of not being bounded from below by zero (cf. Gretton et al. (2012)). As such, it does not meet the strict definition of a data discrepancy measurement.

For a sufficiently large sample size, the U-statistic will have low probability of having a value less than zero, and thus the difference between the U-statistic and V-statistic becomes immaterial for large n. One may also consider a truncation of the U-statistic, which causes no issues, asymptotically, as the U-statistic and V-statistic have the same limit, which is guaranteed to be non-negative.

4 Theoretical results

4.1 Behavior as $n \to \infty$ and $m \to \infty$

4.1.1 Analysis with a generic discrepancy

We now establish a consistency result for the pseudo-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

$$w\left(d, \epsilon\right) = \llbracket d < \epsilon \rrbracket, \tag{8}$$

where $\llbracket \cdot \rrbracket$ 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 and Leisen (2018) for further discussion and examples.

In order to prove our asymptotic result, we require Hunt's lemma, which is reported in Dellacherie and 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, $\lim_{n\to\infty} \mathbb{E}(U_n|\mathcal{F}_n) = \mathbb{E}(U|\mathcal{F}_\infty)$ 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 pseudo-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 $\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$. 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

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

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

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

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 continuous mapping theorem (cf. (DasGupta, 2011, Thm. 7.8)) yields the almost sure convergence of the ratio between (11) and (12) to the right-hand side of (10), as $n \to \infty$. □

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 $\mathcal{D}_{\infty}(\boldsymbol{\theta}_0, \boldsymbol{\theta}) \ne \epsilon$, 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. □

4.1.2 Analysis with the energy statistic

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

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

where

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

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 and 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*

$$\mathbb{E}\left(|\kappa_{\delta}(\boldsymbol{X}_1, \boldsymbol{X}_2; \boldsymbol{Y}_1, \boldsymbol{Y}_2)| \log^+ |\kappa_{\delta}(\boldsymbol{X}_1, \boldsymbol{X}_2; \boldsymbol{Y}_1, \boldsymbol{Y}_2)|\right) < \infty, \tag{13}$$

then $\mathcal{V}_{\delta}(\mathbf{X}_n, \mathbf{Y}_m)$ converges almost surely to $\mathcal{E}_{\delta}(\boldsymbol{X}_1, \mathbf{Y}_1) \ge 0$, 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.

We may apply the result of Lemma 1 directly to the case of $\delta = \delta_1$ in order to provide an almost sure convergence result regarding the V-statistic $\mathcal{V}_{\delta_1}(\mathbf{X}_n, \mathbf{Y}_m)$.

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

$$\mathbb{E}\left(\|\boldsymbol{X}\|_2^2\right) + \mathbb{E}\left(\|\boldsymbol{Y}\|_2^2\right) < \infty, \tag{14}$$

and if $\min\{n,m\} \to \infty$*, then* $\mathcal{V}_{\delta_1}(\mathbf{X}_n, \mathbf{Y}_m)$ *converges almost surely to*

$$\mathcal{E}(\boldsymbol{X},\boldsymbol{Y}) = \frac{\Gamma\left(\frac{d+1}{2}\right)}{\pi^{(d+1)/2}} \int_{\mathbb{R}^d} \frac{\left|\varphi(\boldsymbol{t};\boldsymbol{\theta}_0) - \varphi(\boldsymbol{t};\boldsymbol{\theta})\right|^2}{\|\boldsymbol{t}\|_2^{d+1}} \mathrm{d}\boldsymbol{t}, \tag{15}$$

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

Proof. By the law of total expectation, we apply Lemma 1 by considering the two cases of (13): when $|\kappa_{\delta_1}| \le 1$ and when $|\kappa_{\delta_1}| > 1$, separately, to write

$$\begin{aligned}\mathbb{E}\left(|\kappa_{\delta_1}|\log^+|\kappa_{\delta_1}|\right) = {} & p_0\mathbb{E}\left(|\kappa_{\delta_1}|\log^+|\kappa_{\delta_1}| \,|\, |\kappa_{\delta_1}| \le 1\right) \\ & + p_1\mathbb{E}\left(|\kappa_{\delta_1}|\log^+|\kappa_{\delta_1}| \,|\, |\kappa_{\delta_1}| > 1\right),\end{aligned} \tag{16}$$

where $p_0 = \mathbb{P}(|\kappa_{\delta_1}| \le 1)$ and $p_1 = \mathbb{P}(|\kappa_{\delta_1}| > 1)$. The first term on the right-hand side of (16) is equal to zero, since $\log^+|\kappa_{\delta_1}| = \log(1) = 0$, whenever $|\kappa_{\delta_1}| \le 1$. Thus, we need only be concerned with bounding the second term.

For $|\kappa_{\delta_1}| > 1$, $|\kappa_{\delta_1}|\log|\kappa_{\delta_1}| \le |\kappa_{\delta_1}|^2$, thus

$$\mathbb{E}\left(|\kappa_{\delta_1}|\log^+|\kappa_{\delta_1}| \,|\, |\kappa_{\delta_1}| > 1\right) \le \mathbb{E}\left(|\kappa_{\delta_1}|^2 \,|\, |\kappa_{\delta_1}| > 1\right)$$

The condition that $\mathbb{E}\left(|\kappa_{\delta_1}|\log^+|\kappa_{\delta_1}|\right) < \infty$ is thus fulfilled if $\mathbb{E}\left(|\kappa_{\delta_1}|^2 \,|\, |\kappa_{\delta_1}| > 1\right) < \infty$, which is equivalent to

$$\begin{aligned}\mathbb{E}\left(|\kappa_{\delta_1}|^2\right) = {} & p_0\mathbb{E}\left(|\kappa_{\delta_1}|^2 \,|\, |\kappa_{\delta_1}| \le 1\right) \\ & + p_1\mathbb{E}\left(|\kappa_{\delta_1}|^2 \,|\, |\kappa_{\delta_1}| > 1\right) < \infty,\end{aligned}$$

by virtue of the integrability of $\left\{|\kappa_{\delta_1}|^2 \,|\, |\kappa_{\delta_1}| \le 1\right\}$ implying the existence of

$$\mathbb{E}\left(|\kappa_{\delta_1}|^2 \,|\, |\kappa_{\delta_1}| \le 1\right),$$

since it is defined on a bounded support.

Next, by the triangle inequality,

$$|\kappa_{\delta_1}| \le 2\left(\|\boldsymbol{X}_1\|_2 + \|\boldsymbol{X}_2\|_2 + \|\boldsymbol{Y}_1\|_2 + \|\boldsymbol{Y}_2\|_2\right),$$

and hence

$$\begin{aligned}|\kappa_{\delta_1}|^2 \le {} & 4\left(\|\boldsymbol{X}_1\|_2^2 + \|\boldsymbol{X}_2\|_2^2 + \|\boldsymbol{Y}_1\|_2^2 + \|\boldsymbol{Y}_2\|_2^2\right) \\ & + 8(\|\boldsymbol{X}_1\|_2\|\boldsymbol{X}_2\|_2 + \|\boldsymbol{X}_1\|_2\|\boldsymbol{Y}_1\|_2 + \|\boldsymbol{X}_1\|_2\|\boldsymbol{Y}_2\|_2 \\ & + \|\boldsymbol{X}_2\|_2\|\boldsymbol{Y}_1\|_2 + \|\boldsymbol{X}_2\|_2\|\boldsymbol{Y}_2\|_2 + \|\boldsymbol{Y}_1\|_2\|\boldsymbol{Y}_2\|_2).\end{aligned}$$

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

$$\begin{aligned}\mathbb{E}\left(|\kappa_{\delta_1}|^2\right) \le {} & 8\left[\mathbb{E}\left(\|\boldsymbol{X}_1\|_2^2\right) + \mathbb{E}\left(\|\boldsymbol{Y}_1\|_2^2\right)\right] \\ & + 8\left[\left(\mathbb{E}\|\boldsymbol{X}_1\|_2\right)^2 + \left(\mathbb{E}\|\boldsymbol{Y}_1\|_2\right)^2\right] \\ & + 32\left[\mathbb{E}\|\boldsymbol{X}_1\|_2\,\mathbb{E}\|\boldsymbol{Y}_1\|_2\right],\end{aligned}$$

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 $\mathbb{E}\|\boldsymbol{X}_1\|_2 + \mathbb{E}\|\boldsymbol{Y}_1\|_2 < \infty$. $\square$

We note that condition (14) is stronger than a direct application of condition (13), which may be preferable in some situations. However, condition (14) 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. We further note that (14) is required for establishing almost sure convergence, and is stronger than what is needed to ensure convergence in probability, as is established in Szekely and Rizzo (2004) and Gretton et al. (2012).

Combining the result of Theorem 2 with Corollary 1 and the conclusion from Proposition 1 of Szekely and Rizzo (2013) provided in Equation (15) yields the key result below. This result justifies the use of the V-statistic estimator $\mathcal{V}_{\delta_1}(\mathbf{X}_n, \mathbf{Y}_m)$ for the energy distance $\mathcal{E}(\boldsymbol{X}, \boldsymbol{Y})$ within the IS-ABC framework, and is comparable to Corollaries 1–3 of Jiang et al. (2018) regarding the large sample asymptotics of other discrepancy measurements.

Corollary 2. *Under the assumptions of Corollary 1. If $\mathcal{D}(\mathbf{X}_n, \mathbf{Y}_m) = \mathcal{V}_{\delta_1}(\mathbf{X}_n, \mathbf{Y}_m)$, then the conclusion of Theorem 2 follows with*

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

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 $\boldsymbol{\theta}_0 = \boldsymbol{\theta}$.

4.2 Behavior as $\epsilon \to 0$

Let $\mathbb{F}$ be the set of probability distributions on $\mathbb{X}$. From (Sejdinovic et al., 2013, Thm. 22), we have the fact that $\mathcal{E}^{1/2}(\boldsymbol{X}, \boldsymbol{Y}) = \mathcal{E}^{1/2}(F_X, F_Y)$ is a metric on $\mathbb{F}$, where $\boldsymbol{X}$ and $\boldsymbol{Y}$ have data generating process that are characterized by F_X and F_Y, respectively. As such, when we take $\boldsymbol{X}$ and $\boldsymbol{Y}$ arising from two empirical distributions with an equal number of masses (defined on $\mathbf{x}_n$ and $\mathbf{y}_n$, for instance), then we obtain the fact that $\mathcal{E}^{1/2}(\boldsymbol{X}, \boldsymbol{Y}) = 0$ if and only if the two empirical distributions are the same. In other words, $\mathbf{x}_n$ and $\mathbf{y}_n$ are equal, in the sense that the elements of $\mathbf{x}_n$ and $\mathbf{y}_n$ are equal up to a permutation. Proposition 2 of Bernton et al. (2019) then provides the following result in the case when n is fixed.

Proposition 2. *Assume that $w(d, \epsilon)$ has form (8), where $d = \mathcal{V}_{\delta_1}$, and that $f(\mathbf{x}_n|\boldsymbol{\theta})$ is a continuous and exchangeable PDF. Furthermore, assume that*

$$\sup_{\boldsymbol{\theta} \in \mathbb{T} \setminus \left\{\Theta \subset \mathbb{T} : \int_\Theta \pi(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} = 0\right\}} f(\mathbf{x}_n|\boldsymbol{\theta}) < \infty$$

and suppose that there exists some $\bar{\epsilon} > 0$, where

$$\sup_{\boldsymbol{\theta} \in \mathbb{T} \setminus \left\{\Theta \subset \mathbb{T} : \int_\Theta \pi(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} = 0\right\}} \sup_{\{\mathbf{y}_n : d(\mathbf{x}_n, \mathbf{y}_n) \leq \bar{\epsilon}\}} f(\mathbf{y}_n|\boldsymbol{\theta}) < \infty.$$

Then, for fixed $\mathbf{x}_n$, the pseudo-posterior PDF (3) converges strongly to the posterior PDF (2), as $\epsilon \to 0$.

Let us suppose that the empirical distribution of $\boldsymbol{X}_n$ is denoted $\hat{F}_n$ and that each observation of $\boldsymbol{X}_n$ is generated from a process that can be characterized by the distribution F_0 (we write the joint distribution of $\boldsymbol{X}_n$ as F_n). We shall also write $F_n^{\boldsymbol{\theta}}$ as the probability distribution corresponding to the PDF $f(\mathbf{x}_n|\boldsymbol{\theta})$, and $\hat{F}_n^{\boldsymbol{\theta}}$ as the empirical distribution obtained from a sample $\boldsymbol{Y}_n$ with data generating process that is characterized by $F_n^{\boldsymbol{\theta}}$.

Next, we let the probability distribution corresponding to the prior and pseudo-posterior PDFs of the ES-based ABC process with rejection weights (i.e. $\pi(\boldsymbol{\theta})$ and (3)) as Π and Π_n^ϵ, respectively. And finally, let us denote the probability of the set $\mathbb{A}$ with respect to the probability distribution F as $F(\mathbb{A})$. In order to state our next result, we require the following assumptions.

A1 The data generating process of $\boldsymbol{X}_n$ is such that, for every $\varepsilon > 0$,

$$\lim_{n\to\infty} F_n \left(\mathcal{E}^{1/2} \left(\hat{F}_n, F_0 \right) > \varepsilon \right) = 0.$$

A2 For every $\epsilon > 0$,

$$F_n^{\theta} \left(\mathcal{E}^{1/2} \left(\hat{F}_n^{\boldsymbol{\theta}}, F_1^{\boldsymbol{\theta}} \right) > \epsilon \right) \le c\left(\boldsymbol{\theta}\right) s_n\left(\epsilon\right)$$

where $s_n(\epsilon)$ is a sequence of functions that is strictly decreasing in ϵ for all n, and $s_n(\epsilon) \to 0$ as $n \to \infty$, for fixed ϵ. Here: $c(\boldsymbol{\theta})$ is a positive function that is integrable with respect to Π and satisfies $c(\boldsymbol{\theta}) \le c_0$ for some c_0, for all $\boldsymbol{\theta}$ such that, for some $\delta_0 > 0$, $\mathcal{E}^{1/2}\left(F_0, F_1^{\boldsymbol{\theta}}\right) \le \delta_0 + \epsilon_0$, where $\epsilon_0 = \min_{\boldsymbol{\theta}\in\mathbb{T}} \mathcal{E}^{1/2}\left(F_0, F_1^{\boldsymbol{\theta}}\right)$.

A3 There exists an $L > 0$ and a $c_\pi > 0$ such that, for each sufficiently small $\epsilon > 0$,

$$\Pi\left(\boldsymbol{\theta} \in \mathbb{T} : \mathcal{E}^{1/2}\left(F_0, F_1^{\boldsymbol{\theta}}\right) \le \epsilon + \epsilon_0\right) \ge c_\pi \epsilon^L.$$

Upon making Assumptions A1–A3, we may apply the proof process of (Bernton et al., 2019, Prop. 3) directly, replacing the Wasserstein metric with the energy metric $\mathcal{E}^{1/2}$, where appropriate. Such a process yields the following result.

Proposition 3. *Along with A1–A3, assume that there exists a sequence* $\{\epsilon_n\}_{n=1}^{\infty}$*, such that* $\epsilon_n \to 0$, $s_n(\epsilon_n) \to 0$ *and* $F_n\left(\mathcal{E}^{1/2}\left(\hat{F}_n, F_0\right) \le \epsilon_n\right) \to 1$*, as* $n \to \infty$*. Then, the ES-based ABC algorithm with* $m = n$*, discrepancy* $d = \mathcal{V}_{\delta_1}^{1/2}$*, and rejection weights using* $\epsilon = \epsilon_n + \epsilon_0$ *satisfies the inequality*

$$\Pi_n^{\epsilon_n+\epsilon_0}\left[\mathcal{E}^{1/2}\left(F_0, F_1^{\boldsymbol{\theta}}\right) > \epsilon_0 + \frac{4\epsilon_n}{3} + s_n^{-1}\left(\frac{\epsilon_n^L}{R}\right)\right] \le \frac{C}{R},$$

for some $C, R \in (0, \infty)$*, with probability going to 1 as* $n \to \infty$ *(with respect to* F_0*).*

The hypotheses of Proposition 2 are straight forward and the conclusion implies that pseudo-posterior PDF of the ES-based ABC procedure can approximate the posterior PDF, based on the likelihood of the data generating process of $\mathbf{x}_n$, to an arbitrary level of accuracy, when ϵ is made sufficiently small. This however does not mean that one should make ϵ too small in practice, as the effort required to simulate data will become more difficult and the process becomes more computationally intensive in such cases. Note that the value of ϵ is often chosen in a pragmatic way as a quantile (of a small order, usually less than 5%) of all the distances that are obtained in the ABC sample, thus deciding how many samples are kept as a fraction of the entire ABC replications. This procedure was used in the ABC algorithms of Beaumont et al. (2002), Blum (2010b), and Jabot et al. (2013).

Next, we note that the assumptions (other than A1, which is generally true for stationary and ergodic data; cf. (Szekely and Rizzo, 2013, Secs. 7 and 8)) and the conclusion of Proposition 3 are more complex. Due to the lack of ease by which A2 and A3 may be validated, the proposition is more useful as an existence result regarding what can be expected in theory, with respect to how quickly the ES-based ABC algorithm converges in n, rather than providing any practical guidance. A suggestion by Bernton et al. (2019) is that one may potentially apply the theory of Fournier and Guillin (2015) and Weed et al. (2019) in order to validate assumption A2.

Under further assumptions, the concentration with respect to the discrepancy in distributions can be transferred to a concentration result, with respect to parameter vector in the space $\mathbb{T}$ (cf. (Bernton et al., 2019, Cor. 1)).

4.3 Illustration on a simple example

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 $\boldsymbol{\Sigma}$.

For illustrating the theoretical results, we investigate the pseudo-posterior limit on a simple univariate Gaussian location model $\mathcal{N}(\boldsymbol{\theta}, \sigma^2)$ (with known variance σ^2) with conjugate Gaussian prior $\boldsymbol{\theta} \sim \mathcal{N}(0, \tau^2)$ (with variance τ^2 fixed). We have IID observations $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \mid \boldsymbol{\theta}_0 \sim \mathcal{N}(\boldsymbol{\theta}_0, \sigma^2)$, and IID replicates $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_m \mid \boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\theta}, \sigma^2)$. The posterior is $\boldsymbol{\theta} \mid \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \sim \mathcal{N}(\hat{\boldsymbol{\theta}}, \hat{\sigma}^2)$, where

$$\hat{\boldsymbol{\theta}} = \frac{n\bar{\boldsymbol{X}}_n}{n + \sigma^2/\tau^2}, \quad \hat{\sigma}^{-2} = \tau^{-2} + n\sigma^{-2}.$$

In this simple model, the limiting data discrepancy takes the form (up to a proportionality constant) of $\mathcal{D}_\infty(\boldsymbol{\theta}_0, \boldsymbol{\theta}) = (\boldsymbol{\theta}_0 - \boldsymbol{\theta})^2$ for the energy distance and Kullback–Leibler divergence, and $\mathcal{D}_\infty(\boldsymbol{\theta}_0, \boldsymbol{\theta}) = |\boldsymbol{\theta}_0 - \boldsymbol{\theta}|$ for the MMD and the (second order) Wasserstein distance.

Theorem 2 establishes that the large n and m limit of the pseudo-posterior $\pi_{m,\epsilon}$ is the distribution that we denote here by $\pi_{\infty,\epsilon}(\boldsymbol{\theta}) \propto \pi(\boldsymbol{\theta})\, w(\mathcal{D}_\infty(\boldsymbol{\theta}_0, \boldsymbol{\theta}), \epsilon)$. For illustrative purposes, let us focus on the case when $\mathcal{D}_\infty(\boldsymbol{\theta}_0, \boldsymbol{\theta}) = |\boldsymbol{\theta}_0 - \boldsymbol{\theta}\ |$, and consider rejection ABC with $w(d, \epsilon) = [\![d < \epsilon]\!]$ and IS-ABC with $w(d, \epsilon) = \exp\left(-d^2/2\epsilon^2\right)$. The limiting pseudo-posterior can then be obtained in closed-form as

$$\begin{aligned} \pi_{\infty,\epsilon}(\boldsymbol{\theta}) &\propto \mathcal{N}(\boldsymbol{\theta} \mid 0, \tau^2)\, [\![|\boldsymbol{\theta} - \boldsymbol{\theta}_0| < \epsilon]\!]\,, \\ &= \mathcal{N}_{[\boldsymbol{\theta}_0 - \epsilon, \boldsymbol{\theta}_0 + \epsilon]}(\boldsymbol{\theta} \mid 0, \tau^2), \end{aligned} \tag{17}$$

a truncated Gaussian for rejection ABC and

$$\begin{aligned} \pi_{\infty,\epsilon}(\boldsymbol{\theta}) &\propto \mathcal{N}(\boldsymbol{\theta} \mid 0, \tau^2) \exp\left(\frac{(\boldsymbol{\theta}_0 - \boldsymbol{\theta})^2}{2\epsilon^2}\right), \\ &= \mathcal{N}(\boldsymbol{\theta} \mid \bar{\boldsymbol{\theta}}(\epsilon), \bar{\sigma}^2(\epsilon)), \end{aligned} \tag{18}$$

for IS-ABC, where $\bar{\boldsymbol{\theta}}(\epsilon) = \frac{\boldsymbol{\theta}_0}{1+\epsilon^2/\tau^2}$ and $\bar{\sigma}^{-2}(\epsilon) = \tau^{-2} + \epsilon^{-2}$. See Figure 1 for an illustration, for various values of ϵ.

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_1, the Wasserstein distance using the exponent $p = 2$ and the approximation by the swapping distance (Bernton et al., 2019) and the MMD using a Gaussian kernel $\chi(\boldsymbol{x}, \boldsymbol{y}) = \exp\left[-(\boldsymbol{x} - \boldsymbol{y})^2\right]$. 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 Sec. 2 of Jiang et al. (2018). With respect to the theoretical results of Section 4, the chosen examples can be shown to be sufficiently regular as to validate the hypotheses of Corollary 2 and Proposition 2. However, we believe that it would be difficult to validate Assumptions A2 and A3 of Proposition 3, without further theoretical development.

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. The ABC procedure is sensitive to the choice of the prior; we follow the benchmark examples of Jiang et al.

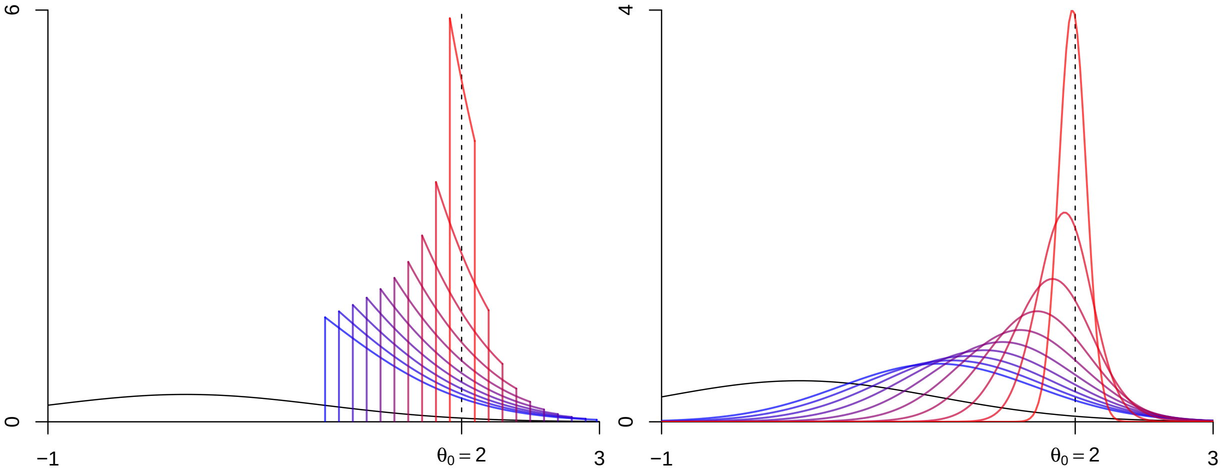


Figure 1: Illustration of theoretical results under rejection ABC (left) and IS-ABC (right) for the univariate Gaussian location model (see details in Section 4.3. The prior density is displayed in black, and the true parameter $\boldsymbol{\theta}_0 = 2$ is indicated by a vertical dashed line. Limiting pseudo-posterior densities $\pi_{\infty,\epsilon}$ provided in (17) (left) and (18) (right), are computed with ϵ varying in $\{1, 0.9, \ldots, 0.2, 0.1\}$ with colors from blue ($\epsilon = 1$) to red ($\epsilon = 0.1$).

(2018) by employing the same uniform priors, as specified in each example. 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 a mixture of bivariate Gaussian probability law

$$\boldsymbol{X}_i \sim p\mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) + (1-p)\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), \tag{19}$$

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 $\boldsymbol{\mu}_1$. We denote the uniform law, in the interval (a, b), for $a < b$, by $\mathrm{Unif}(a, b)$. The

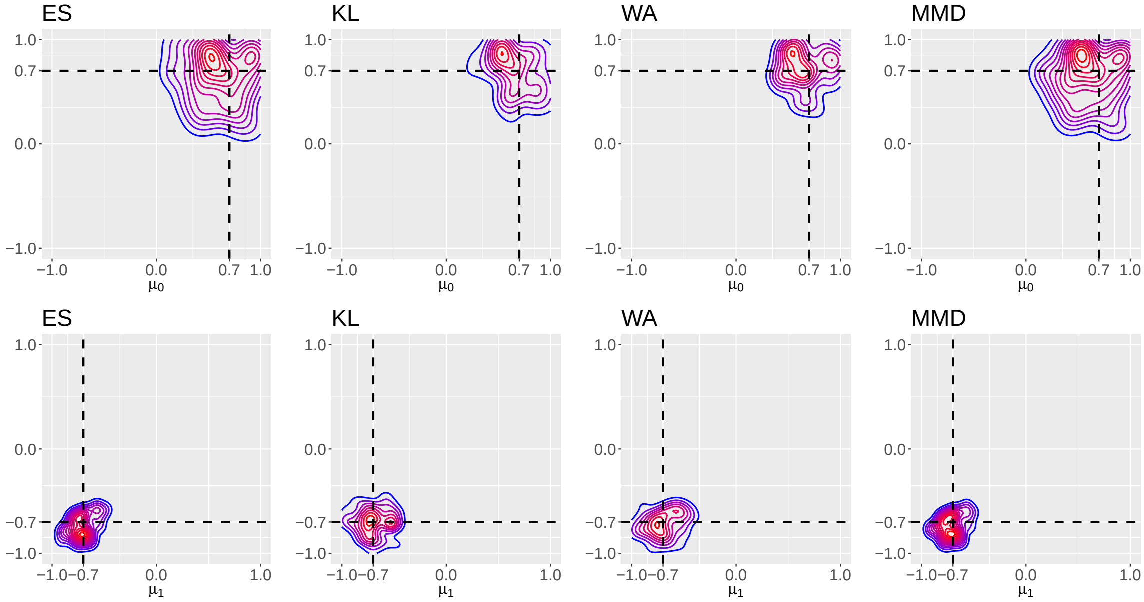


Figure 2: Marginal KDEs of the ABC posterior for the mean parameters $\boldsymbol{\mu}_0$ (top row) and $\boldsymbol{\mu}_1$ (bottom row) of the bivariate Gaussian mixture model (19). The intersections of black dashed lines indicate the positions of the population means.

priors on the model parameters are uniform; that is, $\boldsymbol{\mu}_1 \sim \text{Unif}(-1,1)^2$, $\boldsymbol{\mu}_2 \sim \text{Unif}(-1,1)^2$ and $p \sim \text{Unif}(0,1)$. We perform ABC using $n = 500$ observations, sampled from model (19) 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 (bivariate marginals of $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_1$) is presented in Figure 2.

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 joint posterior distribution of (θ_1, θ_2) is displayed in Figure 3.

5.3 Bivariate beta model

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

$$V_1 = \frac{U_1 + U_3}{U_5 + U_4}, \text{ and } V_2 = \frac{U_2 + U_4}{U_5 + U_3}, \tag{20}$$

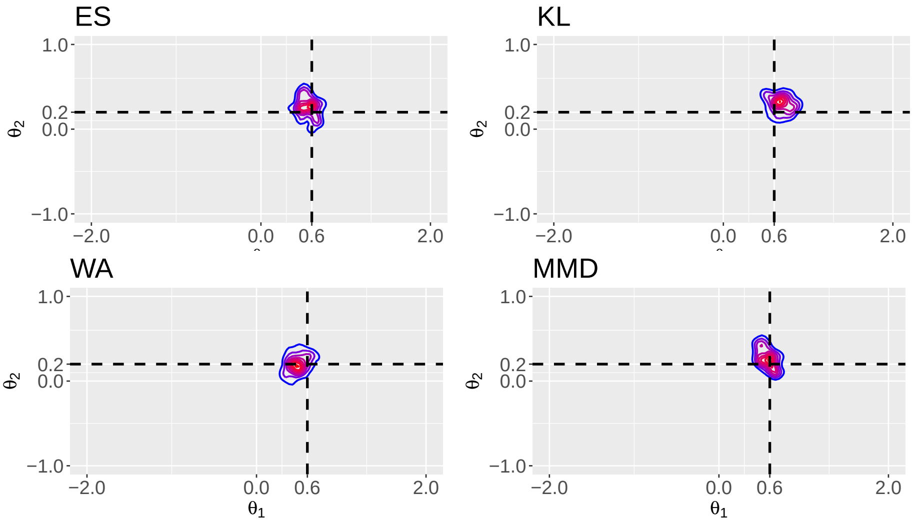


Figure 3: KDE of the ABC posterior for the parameters (θ_1, θ_2) of the MA(2) model experiment. The intersections of black dashed lines indicate the true parameter values.

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 $\text{Unif}(0, 5)$. KDEs of the marginal ABC posterior distributions of parameters $\theta_1, \theta_2, \theta_3, \theta_4$ and θ_5 are displayed in Figure 4.

5.4 Multivariate g-and-k distribution

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

$$F^{-1}(x) = A + B\left[1 + 0.8\frac{1-\exp(-g \times z_x)}{1+\exp(-g \times z_x)}\right]\left(1+z_x^2\right)^k z_x, \tag{21}$$

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 (21), 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 (21) 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

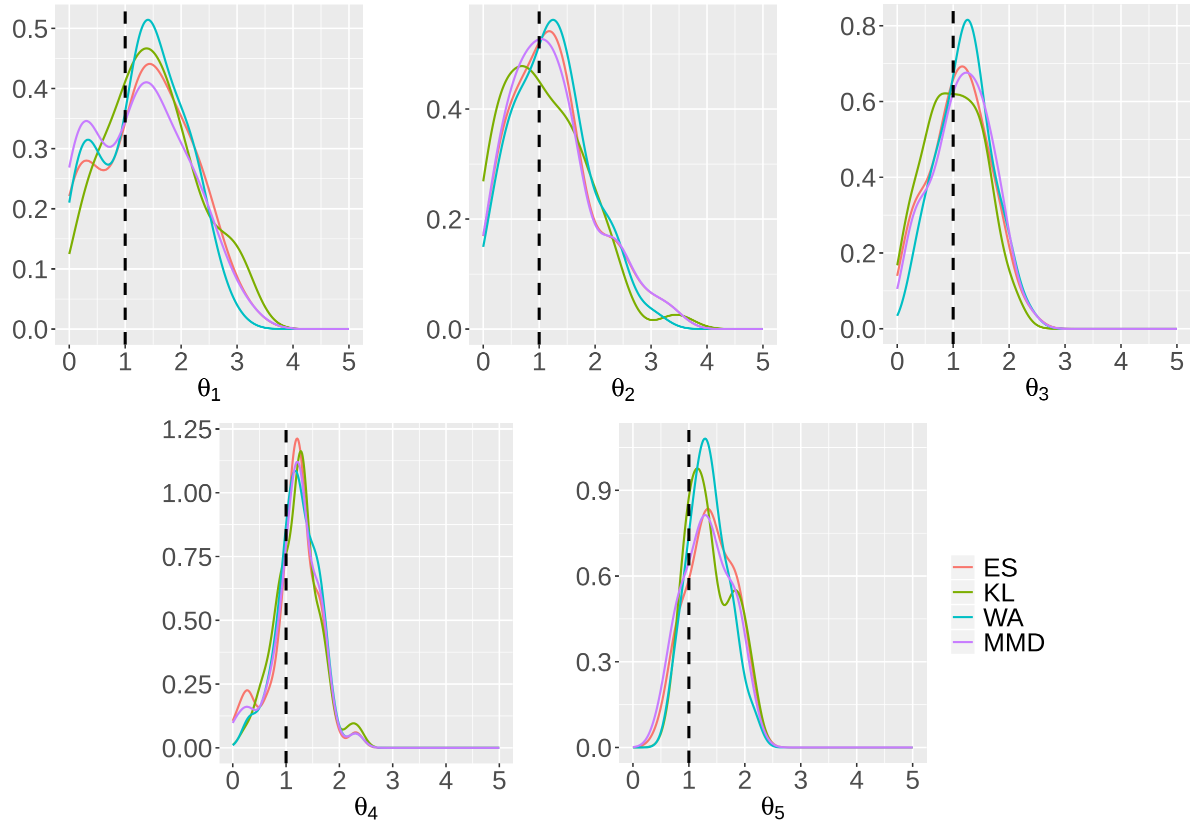


Figure 4: Marginal KDEs of the ABC posterior for the parameters $\theta_1, \dots, \theta_5$ for the bivariate beta model. The black dashed lines indicate the true parameter values.

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 $\rho = -0.3$. The prior on the model parameters A, B, g, k is taken to be independent $\mathrm{Unif}(0, 4)$, while ρ is independently assigned a $\mathrm{Unif}(-0.5, 0.5)$ prior. KDEs of the marginal ABC posterior distributions of parameters A, B, g, k and ρ are displayed in Figure 5.

5.5 Discussion of the results and performance

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

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

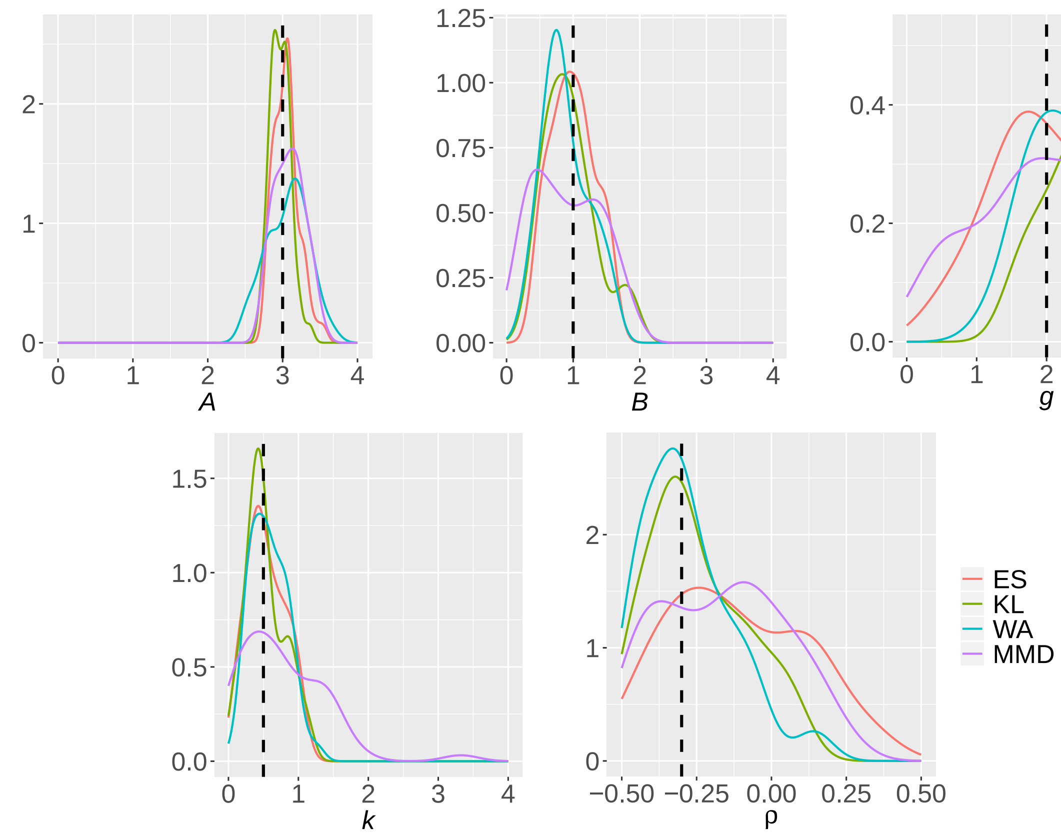


Figure 5: Marginal KDEs of the ABC posterior for the parameters A, B, g, k and ρ of the g-and-k model. The black dashed lines indicate the true parameter values.

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 $\hat{\theta}_{\text{mean}}$, $\hat{\theta}_{\text{med}}$, and errors MAE and RMSE = 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 5.

Upon inspection, Tables 1, 2, 3 and 5 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. Some considerations are therefore necessary when choosing between discrepancies. The first point of consideration is whether the data $\mathbf{X}_n$ are random variables arising from continuous or discrete measures. In the case that the data $\mathbf{X}_n$ arises from a discrete measure, the KL discrepancy measure is not applicable, since it is not defined on a set of measure greater than zero. Another consideration regarding the choice of discrepancy measures is the computational complexity of each discrepancy measure, as is summarized in Table 4.

From Table 4, we firstly note that in the case of univariate data, all methods have the same computational

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

complexity, as all of the discrepancy measures amount to comparisons between the order statistics of the observed and simulated data. Computational complexity becomes a greater separating criterion when considering the multivariate setting. In the multivariate case, the KL divergence is clearly faster than the other methods, but as mentioned before, is not applicable for discrete data. The ES and MMD methods share the same order of complexity, $\mathcal{O}((n+m)^2)$, due to their theoretical equivalence (cf. Sejdinovic et al. (2013)). It is notable that, in general, the computational complexity of the WA discrepancy is of order $\mathcal{O}((n+m)^{5/2}\log(n+m))$, which greater than that of the ES and MMD discrepancies, and is thus a significantly slower method when n and m get large. However, in our numerical results, we have used the $\mathcal{O}((n+m)^2)$ swapping distance approximation of the WA method, as was considered in Bernton et al. (2019). Although this approximation is faster than the exact WA discrepancy, it does not converge to the same value, in general, and thus theoretical results regarding the WA discrepancy cannot be directly applied to the approximation (although some theoretical statements are still available). Thus, there is a trade-off regarding theoretical outcomes when using the swap distance approximation.

We note that in the case when the MMD discrepancy measure is estimated by the V-statistic estimator, much of our theoretical results from Section 4, are applicable with minor modifications, due to the results of Sejdinovic et al. (2013). Thus, the choice between the ES and the MMD method comes down to a preference for the use of kernels or metrics. A consideration regarding the choice of the MMD discrepancy versus the ES discrepancy is that, to the best of our knowledge, a comparable result to (6) does not exist for any common kernel choice.

As an alternative to choosing one of the assessed discrepancy measures, 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.

Running times (on a MacBook Pro 3,1 GHz) for the ES, KL, MMD and WA distance computations for 10^5 ABC replications, in the four models considered in the simulations, for varying sample sizes n, and with $m = n$, are reported in Figure 6. ES is uniformly much faster than the other approaches for small samples sizes, up to the value of $n = m = 50$, where it is performing as fast as KL. For sample sizes larger than $n = m = 50$, KL is fastest. Overall, MMD and WA are slower than ES and KL.

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: Computational complexities. See discussion in Section 6.

	Complexity	References
Univariate (all methods)	$\mathcal{O}((n+m)\log(n+m))$	Jiang et al. (2018), Bernton et al. (2019), Huo and S
KL	$\mathcal{O}((n+m)\log(n+m))$	Jiang et al. (20
Multivariate ES/MMD, WA (approx.)	$\mathcal{O}((n+m)^2)$	Jiang et al. (2018), Bernto
Multivariate WA	$\mathcal{O}((n+m)^{5/2}\log(n+m))$	Bernton et al. (2

Table 5: 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. That is, we largely extend the main result of Jiang et al. (2018), regarding the large sample properties of the pseudo-posterior PDF, to the IS-ABC cases that are considered in Karabatsos and Leisen (2018) and Park et al. (2016). Thus, we provide further theoretical justification for the usage of such algorithms.

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, the ES, KL, MMD, and WA estimators in *univariate settings* are all equal in terms of order of complexity, with a *linearithmic* computational time of $\mathcal{O}((n+m)\log(n+m))$ (see Huo and Székely (2016), Chaudhuri and Hu (2019), regarding the complexity of the ES and MMD estimators). In *multivariate settings*, KL complexity is unchanged; ES and MMD have quadratic time $\mathcal{O}((n+m)^2)$, while the Wasserstein distance has complexity $\mathcal{O}((n+m)^{5/2}\log(n+m))$. The latter can be reduced to quadratic complexity if one is targetting the swapping distance, an approximation of the actual Wasserstein distance (Bernton et al., 2019). We note that linear time estimators are also available for the MMD and the ES, if one is willing to forgo precision in the estimates (see Gretton et al. (2012)). See Table 4 for a summary.

References

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

Barber, S., Voss, J., and Webster, M. (2015). The rate of convergence for approximate Bayesian computation. *Electronic Journal of Statistics*, 9(1):80–105.

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

Beaumont, M. A., Zhang, W., and Balding, D. J. (2002). Approximate bayesian computation in population genetics. *Genetics*, 162:2025–2035.

Bernton, E., Jacob, P. E., Gerber, M., and 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., and Li, S. (2013). *FNN: Fast Nearest Neighbor Search miller and Applications.* R package version 1.1.3.

Biau, G., Cérou, F., and Guyader, A. (2015). New insights into approximate Bayesian computation. *Annales de l'IHP Probabilités et statistiques*, 51(1):376–403.

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

Blum, M. G. (2010b). Choosing the summary statistics and the acceptance rate in approximate bayesian computation. In *Proceedings of COMPSTAT'2010*, pages 47–56.

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

Chaudhuri, A. and Hu, W. (2019). A fast algorithm for computing distance correlation. *Computational statistics & data analysis*, 135:15–24.

Crackel, R. and 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.* Springer, New York.

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

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

Drovandi, C. C., Pettitt, A. N., and Lee, A. (2015). Bayesian indirect inference using a parametric auxiliary model. *Statistical Science*, pages 72–95.

Fearnhead, P. and Prangle, D. (2012). Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(3):419–474.

Fournier, N. and Guillin, A. (2015). On the rate of convergence in wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162:707–738.

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

Frazier, D. T., Robert, C. P., and Rousseau, J. (2020). Model misspecification in approximate Bayesian computation: consequences and diagnostics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

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

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

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

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

Gutmann, M. U., Dutta, R., Kaski, S., and Corander, J. (2018). Likelihood-free inference via classification. *Statistics and Computing*, 28(2):411–425.

Huo, X. and Székely, G. J. (2016). Fast computing for distance covariance. *Technometrics*, 58:435–447.

Jabot, F., Faure, T., and Dumoulin, N. (2013). Easy ABC: performing efficient approximate bayesian computation sampling schemes using R. *Methods in Ecology and Evolution*, 4:684–687.

Jiang, B., Wu, T.-Y., and Wong, W. H. (2018). Approximate Bayesian computation with Kullback-Leibler divergence as data discrepancy. In Storkey, A. and Perez-Cruz, F., editors, *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 84 of *Proceedings of Machine Learning Research*, pages 1711–1721, Playa Blanca, Lanzarote, Canary Islands. PMLR.

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

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

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

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

Li, W. and Fearnhead, P. (2018). On the asymptotic efficiency of approximate Bayesian computation estimators. *Biometrika*, 105(2):285–299.

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

Mak, S. and Joseph, V. R. (2018). Support points. *The Annals of Statistics*, 46:2562–2592.

Marin, J.-M., Pillai, N. S., Robert, C. P., and Rousseau, J. (2014). Relevant statistics for Bayesian model choice. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(5):833–859.

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

Miller, J. W. and 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.* MIT Press, Cambridge.

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

Prangle, D., Fearnhead, P., Cox, M. P., Biggs, P. J., and French, N. P. (2014). Semi-automatic selection of summary statistics for ABC model choice. *Statistical applications in genetics and molecular biology*, 13(1):67–82.

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

Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A., and 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.* Springer, New York.

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., and 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.* Wiley, New York.

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

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

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

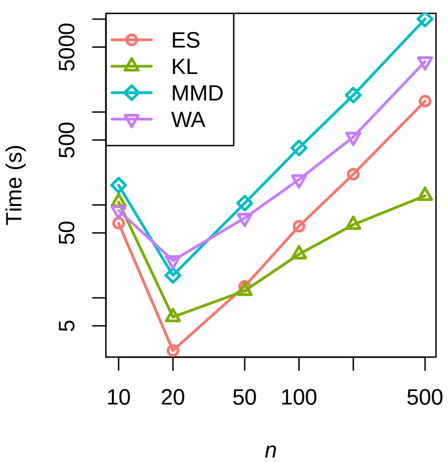
Tavaré, S. (2019). On the history of ABC. In Sisson, S. A., Fan, Y., and Beaumont, M. A., editors, *Handbook of Approximate Bayesian Computation.* CRC Press, Boca Raton.

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

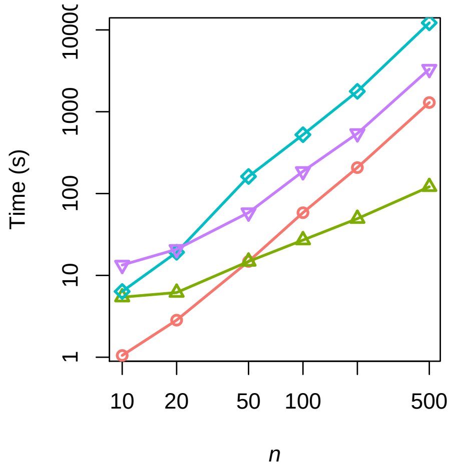
Voss, J. (2014). *An Introduction to Statistical Computing: A Simulation-based Approach.* Wiley, Chichester.

Weed, J., Bach, F., et al. (2019). Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance. *Bernoulli*, 25:2620–2648.

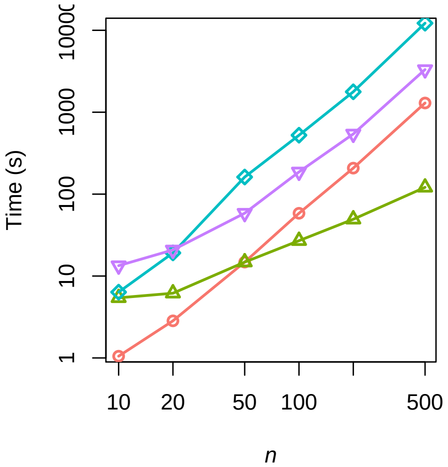
Zygmund, A. (1951). An individual ergodic theorem for non-comutative transformations. *Acta Scientiarum Mathematicarum (Szeged)*, 14:103–110.



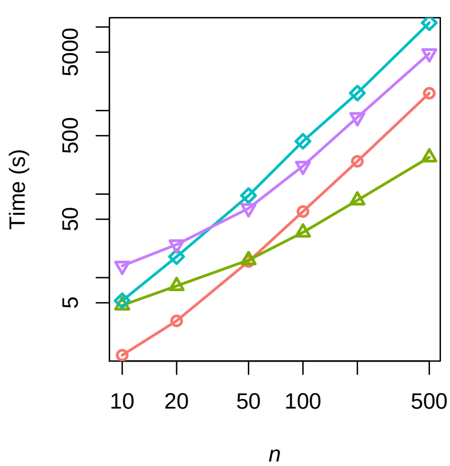
(a) Bivariate Gaussian mixtures (Section 5.1)



(b) MA(2) model (Section 5.2)



(c) Bivariate beta model (Section 5.3)



(d) g-and-k distribution (Section 5.4)

Figure 6: Running times for the ES, KL, MMD and WA distance computations in the four models considered in the simulations. Time in seconds for 10^5 ABC replications, for varying sample sizes n, and with $m = n$. Log scales on both axes.