Approximate Bayesian Computation with Kullback-Leibler Divergence as Data Discrepancy

Bai Jiang
Princeton University

Tung-Yu Wu
Stanford University

Wing Hung Wong
Stanford University

Abstract

Complex simulator-based models usually have intractable likelihood functions, rendering the likelihood-based inference methods inapplicable. Approximate Bayesian Computation (ABC) emerges as an alternative framework of likelihood-free inference methods. It identifies a quasi-posterior distribution by finding values of parameter that simulate the synthetic data resembling the observed data. A major ingredient of ABC is the discrepancy measure between the observed and the simulated data, which conventionally involves a fundamental difficulty of constructing effective summary statistics. To bypass this difficulty, we adopt a Kullback-Leibler divergence estimator to assess the data discrepancy. Our method enjoys the asymptotic consistency and linearithmic time complexity as the data size increases. In experiments on five benchmark models, this method achieves a comparable or higher quasi-posterior quality, compared to the existing methods using other discrepancy measures.

1 Introduction

The likelihood function is of central importance in statistical inference by characterizing the connection between the observed data and the value of parameter in models. Many simulator-based models, which are stochastic data generating mechanisms taking parameter values as input and returning data as output, have arisen in evolutionary biology [1, 2], dynamic systems [3, 4], economics [5, 6], epidemiology [7–9], aeronautics [10] and other disciplines. In these models, the observed data are seen as outcomes of the data generating mechanisms given some underlying true parameter. The simulator-based models are said to be implicit [11] because their likelihood functions involve integrals over latent variables or solutions to differential equations and have no explicit form. These models are also said to be generative [12] because they specify how to generate synthetic data sets.

As the unavailability of the likelihood function renders the conventional likelihood-based inference methods inapplicable, Approximate Bayesian Computation (ABC) emerges as an alternative framework of likelihood-free inference. [13–15] provide general overviews of ABC. Rejection ABC [1, 16–18], the first and simplest ABC algorithm, repeatedly draws values of parameter θ independently from some prior π, simulates synthetic data $\boldsymbol{Y}$ for each value of θ, and rejects the parameter θ if the discrepancy $\mathfrak{D}(\boldsymbol{X}, \boldsymbol{Y})$ between the observed data $\boldsymbol{X}$ and the simulated data $\boldsymbol{Y}$ exceeds a tolerance threshold ϵ. This algorithm obtains an independent and identically distributed (i.i.d.) random sample of parameter from a quasi-posterior distribution. Later [3, 19–23] enhance the efficiency over rejection ABC by incorporating Markov Chain Monte Carlo and sequential techniques. On the other aspect, [24] interprets the acceptance-rejection rule with the tolerance threshold ϵ in ABC as convoluting the target posterior distribution with a small ϵ-noise term. This interpretation encompasses the spirit of assigning continuous weights to proposed parameter draws rather than binary weights (either accepting or rejecting) in many ABC implementations [25].

A major ingredient of ABC is the data discrepancy measure $\mathfrak{D}(\boldsymbol{X}, \boldsymbol{Y})$, which crucially influences the quality of the quasi-posterior distribution. An ABC algorithm typically reduces data $\boldsymbol{X}, \boldsymbol{Y}$ to their summary statistics $S(\boldsymbol{X}), S(\boldsymbol{Y})$ and measures the distance between $S(\boldsymbol{X})$ and $S(\boldsymbol{Y})$ instead as the data discrepancy. Naturally one would like to use a low-dimensional and quasi-sufficient summary statistic $S(\cdot)$, which offers a satisfactory tradeoff between the acceptance rate of proposed parameters and the quality of the quasi-posterior [26]. Constructing effective summary statistics presents a fundamental difficulty and is actively

Proceedings of the 21^{st} International Conference on Artificial Intelligence and Statistics (AISTATS) 2018, Lanzarote, Spain. PMLR: Volume 84. Copyright 2018 by the author(s).

pursued in the literature (see reviews [26, 27] and references therein). Most existing methods can be grouped into three main categories: *best subset selection*, *regression* and *Bayesian indirect inference*. The first category of methods select the optimal subset from a set of pre-chosen candidate summary statistics according to various information criteria (e.g. measure of sufficiency [28], entropy [29], AIC/BIC [26], impurity in random forests [30]) and then use the selected subset as the summary statistics. The candidate summary statistics are usually provided by experts in the specific scientific domain. The second category assembles methods that fit regression on candidate summary statistics [31, 25]. The last category, inspired by indirect inference [32], dispenses with candidate summary statistics and directly constructs summary statistics from an auxiliary model [33, 34, 27].

This paper mainly focuses on the setting in which no informative candidate summary statistic is available but the observed data contains a moderately large number n of i.i.d. samples $\boldsymbol{X} = \{X_i\}_{i=1}^n$, allowing direct data discrepancy measures. Most methods for constructing summary statistic fails in this setting. An exception is Fearnhead and Prangle's semi-automatic method [25], which works well for univariate distributions. It performs a regression on quantiles (and their 2nd, 3rd and 4th powers) of the empirical distributions. Still unclear is how to extend this method to multivariate distributions. Another exception is the category of Bayesian indirect inference methods [33, 34, 27]. But their performance relies on the choice of auxiliary models.

We propose using a Kullback-Leibler (KL) divergence estimator [35], which is based on the observed data $\boldsymbol{X} = \{X_i\}_{i=1}^n$ and the simulated data $\boldsymbol{Y} = \{Y_i\}_{i=1}^m$, as the data discrepancy for ABC. We denote this estimator or data discrepancy by $\mathfrak{D}_{\mathrm{KL}}(\boldsymbol{X}, \boldsymbol{Y})$. As such, we bypass the construction of summary statistics. The KL divergence $\mathrm{KL}(g_0||g_1)$, also known as information divergence or relative entropy, measures the distance between two distributions $g_0(x)$ and $g_1(x)$ [36]. Denote by $\{p_\theta :\ \theta \in \Theta\}$ the model under study, and by θ^* the true parameter that generates $\boldsymbol{X}$. Interpreting $\mathrm{KL}(p_{\theta^*}||p_\theta)$ as the expectation of the log-likelihood ratio nicely connects it to the maximum likelihood estimation. Our method leverages the KL divergence to Bayesian inference. Using a consistent KL divergence estimator developed by [35], our method is asymptotically consistent in the sense that the quasi-posterior converges to $\pi(\theta|\mathrm{KL}(p_{\theta^*}||p_\theta) < \epsilon) \propto \pi(\theta)\mathbb{I}(\mathrm{KL}(p_{\theta^*}||p_\theta) < \epsilon)$ as the sample sizes n, m increase. Here $\mathbb{I}(\cdot)$ denotes the indicator function. The KL divergence estimator used in our method might be replaced with other estimator [37–42].

The KL divergence method achieves a comparable or higher quality of the quasi-posterior distribution in the experiments on five benchmark models, compared to its three cousins: the classification accuracy method [43], the maximum mean discrepancy method [44] and the Wasserstein distance method [45]. It also enjoys a linearithmic time complexity: the cost for a single call of $\mathfrak{D}_{\mathrm{KL}}(\boldsymbol{X}, \boldsymbol{Y})$, given observed and simulated data $\boldsymbol{X}, \boldsymbol{Y}$ with n samples each, is $\mathcal{O}(n \ln n)$. This cost is smaller than $\mathcal{O}(n^2)$-cost of computing the maximum mean discrepancy in [44] and the Wasserstein distance in [45]. Computing the classification accuracy in [43] costs $\mathcal{O}(n)$ in general, but it generates much worse quais-posteriors than other methods in the experiments.

The remaining parts of the paper are structured as follows: Section 2 describes the ABC algorithm and five data discrepancy measures including our KL divergence estimator. Section 3 establishes the asymptotic consistency of our method and compares its limiting quasi-posterior distribution to those of other methods. In Section 4, we apply the methodology to five benchmark simulator-based models. Section 5 discusses our method and concludes the paper.

2 ABC and Data Discrepancy

Denote by $\mathcal{X} \subset \mathbb{R}^d$ the data space, and by Θ the parameter space of interest. The model $\mathcal{M} = \{p_\theta : \theta \in \Theta\}$ is a collection of distributions on $\mathcal{X}$. It has no explicit form of $p_\theta(x)$ but can simulate i.i.d. random samples given a value of parameter. To identify the true parameter θ^* that generates i.i.d. observed data samples $\boldsymbol{X} = \{X_i\}_{i=1}^n$, Approximate Bayesian Computation (ABC) algorithm finds the values of parameter that generate the synthetic data $\boldsymbol{Y} = \{Y_i\}_{i=1}^m \sim p_\theta$ i.i.d. resembling the observed data $\boldsymbol{X}$. The extent to which the observed and simulated data are resembling is quantified by a data discrepancy measure $\mathfrak{D}(\boldsymbol{X}, \boldsymbol{Y})$.

Since our goal is to compare difference data discrepancy measures rather than present a complete methodology for ABC, we only use rejection ABC, the simplest ABC algorithm (Algorithm 1), throughout this paper. For convenience of notations, we write $p_\theta(\boldsymbol{y}) = \prod_{i=1}^m p_\theta(y_i)$. Algorithm 1 outputs a random sample $\{\theta^{(t)}\}_{t=1}^T$ of the quasi-posterior distribution

$$\pi(\theta|\boldsymbol{X}; \mathfrak{D}, \epsilon) \propto \int \pi(\theta)\mathbb{I}\left(\mathfrak{D}(\boldsymbol{X}, \boldsymbol{y}) < \epsilon\right) p_\theta(\boldsymbol{y}) d\boldsymbol{y}. \quad (1)$$

Next, we introduce our KL divergence method and describe other direct data discrepancies. The semi-automatic method and the Bayesian indirect inference method are also included as they do not require candidate summary statistics. Let us collect more notations. g_0 and g_1 denote densities of two d-dimensional distributions on the sample space $\mathcal{X} \subseteq \mathbb{R}^d$. For a vector

Algorithm 1 Rejection ABC Algorithm

```
Input: - the observation data $\boldsymbol{X} = \{X_i\}_{i=1}^n$;
    - a prior $\pi(\theta)$ over the parameter space $\Theta$;
    - a tolerance threshold $\epsilon > 0$;
    - a data discrepancy measure $\mathfrak{D} : \mathcal{X}^n \times \mathcal{X}^m \to \mathbb{R}^+$.
for $t = 1, ..., T$ do
    repeat:  propose $\theta \sim \pi(\theta)$ and
             draw $\boldsymbol{Y} = \{Y_i\}_{i=1}^m \sim p_\theta$ i.i.d.
    until : $\mathfrak{D}(\boldsymbol{X}, \boldsymbol{Y}) < \epsilon$
    Let $\theta^{(t)} = \theta$
end for
Output: $\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(T)}$
```

x, $\|x\|$ denotes its ℓ_2 norm, and $x_{(i)}$ denotes its i-th ordered element. For two scalars a and b, we write $\max\{a, b\}$ as $a \vee b$. We write $\mathfrak{D}$ in short for a data discrepancy $\mathfrak{D}(\boldsymbol{X}, \boldsymbol{Y})$, if the observed and simulated data $\boldsymbol{X}$ and $\boldsymbol{Y}$ are clear in the context.

2.1 Kullback-Leibler (KL) Divergence as Data Discrepancy

The KL divergence between g_0 and g_1 is defined as

$$\mathrm{KL}(g_0 || g_1) = \int g_0(x) \ln \frac{g_0(x)}{g_1(x)} dx \geq 0,$$

which is zero if and only if $g_0(x) = g_1(x)$ for almost everywhere. Given the observed and simulated data $\boldsymbol{X}$ and $\boldsymbol{Y}$, we use the following estimator for $\mathrm{KL}(p_{\theta^*} || p_\theta)$:

$$\mathfrak{D}_{\mathrm{KL}} = \frac{d}{n} \sum_{i=1}^n \ln \frac{\min_j \|X_i - Y_j\|}{\min_{j \neq i}^n \|X_i - X_j\|} + \ln \frac{m}{n-1}. \quad (2)$$

This estimator is the special case of Equation (14) in [35] using 1-nearest neighbor density estimate. Theorem 2 in [35] establishes the almost-sure convergence of (2). We use (2) as the data discrepancy in Algorithm 1. As this method involves $2n$ operations of nearest neighbor search, we use k-d trees [46, 47] to implement it. The time cost per call of $\mathfrak{D}_{\mathrm{KL}}$ is $\mathcal{O}((n \vee m) \ln(n \vee m))$ on average.

2.2 Classification Accuracy (CA) as Data Discrepancy

The classification accuracy method in [43] originates from the phenomenon that distinguishing $\boldsymbol{Y}$ from $\boldsymbol{X}$, when θ is very different to θ^*, is usually easier than doing so, when θ is similar to θ^*. This method first labels X_i as class 0 and Y_i as class 1, yielding an augmented data set $\mathcal{D} = \{(X_1, 0), \ldots, (X_n, 0), (Y_1, 1), \ldots, (Y_m, 1)\}$, and then trains a prediction rule (classifier) $h : x \mapsto \{0, 1\}$ to distinguish two classes. The authors define *discriminability* or *classifiability* between two samples $\boldsymbol{X}$ and $\boldsymbol{Y}$ as the K-fold cross-validation classification accuracy, and use it as the data discrepancy for ABC. Formally, denoting by $\mathcal{D}_k$ the k-fold subset of $\mathcal{D}$, and by $|\mathcal{D}_k|$ its size,

$$\mathfrak{D}_{\mathrm{CA}} = \frac{1}{K} \sum_{k=1}^K \frac{1}{|\mathcal{D}_k|} \left[\sum_{i:\ (X_i, 0) \in \mathcal{D}_k} (1 - \hat{h}_k(X_i)) \right. \\ \left. + \sum_{i:\ (Y_i, 1) \in \mathcal{D}_k} \hat{h}_k(Y_i) \right] \quad (3)$$

where $\hat{h}_k$ is the trained prediction rule on the data set of $\mathcal{D} \setminus \mathcal{D}_k$. Our experiments set $K = 5$ and h to be the Linear Discriminant Analysis (LDA) classifier. We choose LDA because [43] reports that the quasi-posterior quality seems insensitive to the choice of classifiers, and LDA is computationally cheaper than other classifiers. The time cost per call of $\mathfrak{D}_{\mathrm{CA}}$ is $\mathcal{O}(n + m)$.

2.3 Maximum Mean (MM) Discrepancy

The kernel embedding of a probability distribution $g(x)$, defined as $\mu_g = \int k(\cdot, x) g(x) dx$, is an element in the RKHS $\mathcal{H}$ associated with a positive definite kernel kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ [48, 49]. The maximum mean (MM) discrepancy [50] between g_0 and g_1 is the distance of μ_{g_0} and μ_{g_1} in the Hilbert space $\mathcal{H}$, i.e. $\mathrm{MM}^2(g_0, g_1) = \|\mu_{g_0} - \mu_{g_1}\|_{\mathcal{H}}^2$. [44] adopts an unbiased estimator of $\mathrm{MM}^2(p_{\theta^*}, p_\theta)$ as the data discrepancy in ABC. The square of the estimator is as follows.

$$\mathfrak{D}_{\mathrm{MM}}^2 = \frac{\sum_{1 \leq i \neq j \leq n} k(X_i, X_j)}{n(n-1)} + \frac{\sum_{1 \leq i \neq j \leq m} k(Y_i, Y_j)}{m(m-1)} \\ - \frac{2 \sum_{i=1}^n \sum_{j=1}^m k(X_i, Y_j)}{nm} \quad (4)$$

In the same fashion of [44], we choose a Gaussian kernel with the bandwidth being the median of $\{\|X_i - X_j\| :\ 1 \leq i \neq j \leq n\}$. The time cost per call of $\mathfrak{D}_{\mathrm{MM}}$ is $\mathcal{O}((n + m)^2)$, as it requires computing the $(n + m) \times (n + m)$ pairwise distance matrix.

2.4 Wasserstein Distance

Let ρ be a distance on $\mathcal{X} \subseteq \mathbb{R}^d$. The q-Wasserstein distance between g_0 and g_1 is defined as

$$\mathfrak{W}_q(g_0, g_1) = \left[\inf_{\gamma \in \Gamma(g_0, g_1)} \int_{\mathcal{X} \times \mathcal{X}} \rho(x, y)^q d\gamma(x, y) \right]^{1/q},$$

where $\Gamma(g_0, g_1)$ is the set of all joint distribution $\gamma(x, y)$ on $\mathcal{X} \times \mathcal{X}$ such that γ has marginals g_0 and g_1. An estimator of $\mathfrak{W}_q(p_{\theta^*}, p_\theta)$ based on the samples $\boldsymbol{X}$ and $\boldsymbol{Y}$ can serve as the data discrepancy for ABC. In particular, with $q = 2$ and ρ being the Euclidean distance, an instance is given by

$$\mathfrak{D}_{\mathrm{W2}} = \min_{\boldsymbol{\gamma}} \left[\sum_{i=1}^{n} \sum_{j=1}^{m} \gamma_{ij} \|X_i - Y_j\|^2 \right]^{1/2} \tag{5}$$
$$\text{s.t. } \boldsymbol{\gamma} \mathbf{1}_m = \mathbf{1}_n,\ \boldsymbol{\gamma}^T \mathbf{1}_n = \mathbf{1}_m,\ 0 \leq \gamma_{ij} \leq 1$$

where $\boldsymbol{\gamma} = \{\gamma_{ij} : 1 \leq i \leq n, 1 \leq j \leq m\}$ is a $n \times m$ matrix, and $\mathbf{1}_n, \mathbf{1}_m$ are vectors of n and m ones, respectively.

For multivariate distributions ($d > 1$), exactly solving this optimization problem costs $\mathcal{O}((n+m)^3 \ln(n+m))$ [51] and approximate optimization algorithms [52, 53] reduce the cost to $\mathcal{O}((n+m)^2)$. For univariate distributions ($d = 1$), if $n = m$ and $\rho(x, y) = |x - y|$, q-Wasserstein distance has an explicit form $\left(\frac{1}{n}\sum_{i=1}^{n} |X_{(i)} - Y_{(i)}|^q\right)^{1/q}$. In this special case, the computation cost is $\mathcal{O}(n \ln n)$.

2.5 Distance between Summary Statistics

An ABC algorithm typically uses the distance between the data summaries $S(\boldsymbol{X}), S(\boldsymbol{Y})$ as the discrepancy measure $\mathfrak{D}$. In particular, if using the Euclidean distance,

$$\mathfrak{D}_S = \|S(\boldsymbol{X}) - S(\boldsymbol{Y})\|. \tag{6}$$

The subscript S specifies the choice of summary statistic S. Most methods for constructing S requires expertise in the specific scientific domain to provide candidate summary statistics. Two exceptions are the semi-automatic method [25] and the Bayesian indirect inference method [33, 34, 27].

If no candidate summary statistic is given, the **semi-automatic method** can work for univariate distributions. The method first generates a large artificial data set of pairs $(\theta, \boldsymbol{Y})$ from the prior π and the simulated model. When the distribution is one-dimensional, each $\boldsymbol{Y}$ in the artificial data set is a vector of length m. The semi-automatic method performs a linear regression on the artificial data set with θ as target and candidate summary statistics as regressors. In absence of candidate summary statistics, Fearnhead and Prangle [25] suggested using evenly-spaced quantiles (and their 2nd, 3rd and 4th powers) of $\boldsymbol{Y}$ as regressors. Formally, $\mathbb{E}[\theta|\boldsymbol{Y}] \approx \beta_0 + \sum_{k=1}^{K-1}\sum_{l=1}^{4} \beta_{kl} Y_{(km/K)}^{l}$. The summary statistic $S(\boldsymbol{y})$ is merely the prediction of θ given $\boldsymbol{y}$, which is understood as an approximation of the posterior mean $\mathbb{E}[\theta|\boldsymbol{y}]$. As the original paper [25] does not clearly show how to extend this method to multivariate distributions in which case each simulated data set $\boldsymbol{Y}$ is an $n \times d$ matrix, we simply put quantiles of each marginal together as a total of $4d(K-1)$ regressors. Our experiments set $K = 8$.

The **Bayesian indirect inference methods** construct the summary statistic from an auxiliary model $\{p_A(x|\phi) : \phi \in \Phi\}$. See a general review of the literature in [27]. An instance of these methods is given by [33]. The authors suggest the maximum likelihood estimate (MLE) of the auxiliary model as the summary statistic. Formally,

$$S(\boldsymbol{y}) = \hat{\phi}(\boldsymbol{y}) = \arg\max_{\phi \in \Phi} \prod_{i=1}^{m} p_A(y_i|\phi). \tag{7}$$

In particular, if $p_A(x|\phi)$ is d-dimensional Gaussian with parameter ϕ (which aggregates mean and covariance parameters together), the summary statistics $S(\boldsymbol{y})$ are merely the sample mean and covariance of $\boldsymbol{y} = \{y_i\}_{i=1}^{m}$. [34] also describes an approach that uses the auxiliary likelihood (AL) to set up a data discrepancy

$$\mathfrak{D}_{\mathrm{AL}} = \frac{1}{m} \ln p_A(\boldsymbol{Y}|\hat{\phi}(\boldsymbol{Y})) - \frac{1}{m} \ln p_A(\boldsymbol{Y}|\hat{\phi}(\boldsymbol{X})) \tag{8}$$

3 Asymptotic Analysis

This section analyzes the asymptotic quasi-posterior distributions as the size n of observation data $\boldsymbol{X}$ goes to infinity (and the size m of the simulated data $\boldsymbol{Y}$ increases at the same rate like $m/n \to \alpha > 0$) with different data discrepancy measures. The tolerance threshold ϵ assumed to be fixed. This analysis explains how different data discrepancy measures weight the values of parameter in their quasi-posterior distributions. The theoretical results show that the quasi-posterior distribution obtained by our KL divergence method (2) identifies θ with $\mathrm{KL}(p_{\theta^*}\|p_\theta) < \epsilon$ asymptotically. In addition, our method is related to the classification accuracy method (3) by the concept of f-divergence [54]. The KL divergence method is also asymptotically equivalent to (8), a Bayesian indirect inference method, in case that the auxiliary model is bijective to the true model. Proof of theorems and corollaries are deferred to the Appendix.

We start this section from Theorem 1, which is an application of Lévy's upward theorem.

Theorem 1 (Asymptotic Quasi-Posterior) *If the data discrepancy measure $\mathfrak{D}(\boldsymbol{X}, \boldsymbol{Y})$ in Algorithm 1 converges to some real number $\mathfrak{D}(p_{\theta^*}, p_\theta)$ almost surely as the data size $n \to \infty, m/n \to \alpha > 0$ then the quasi-posterior distribution $\pi(\theta|\boldsymbol{X}; \mathfrak{D}, \epsilon)$ defined by (1) converges to $\pi(\theta|\mathfrak{D}(p_{\theta^*}, p_\theta) < \epsilon)$ for any θ. That is,*

$$\lim_{n\to\infty} \pi(\theta|\boldsymbol{X}; \mathfrak{D}, \epsilon) = \pi(\theta|\mathfrak{D}(p_{\theta^*}, p_\theta) < \epsilon)$$
$$\propto \pi(\theta)\mathbb{I}(\mathfrak{D}(p_{\theta^*}, p_\theta) < \epsilon)$$

This theorem asserts that the asymptotic quasi-posterior is a restriction of the prior π on the region $\{\theta \in \Theta : \mathfrak{D}(p_{\theta^*}, p_\theta) < \epsilon\}$. Putting it together with the established almost sure convergence of $\mathfrak{D}_{\mathrm{KL}}$ in [35], we have a corollary for $\mathfrak{D}_{\mathrm{KL}}$ as follows.

Corollary 1 (Asymptotic Quasi-Posterior of $\mathfrak{D}_{\mathbf{KL}}$) *Let $n \to \infty$ and $m/n \to \alpha > 0$. If Algorithm 1 uses $\mathfrak{D}_{KL}$ defined by (2) as the data discrepancy measure then the quasi-posterior distribution*

$$\lim_{n\to\infty} \pi(\theta|\boldsymbol{X}; \mathfrak{D}_{KL}, \epsilon) = \pi(\theta|KL(p_{\theta^*}||p_\theta) < \epsilon) \\ \propto \pi(\theta)\mathbb{I}(KL(p_{\theta^*}||p_\theta) < \epsilon).$$

It is known that the maximum likelihood estimator minimizes the KL divergence between the empirical distribution of p_{θ^*} and p_θ. ABC with $\mathfrak{D}_{\mathrm{KL}}$ shares the same idea to find θ with small KL divergence.

Next, we find $\mathfrak{D}_{\mathrm{KL}}$ coincides with $\mathfrak{D}_{\mathrm{AL}}$ in the framework of f-divergence [54]. For a convex function $f(t)$ with $f(1) = 0$, f-divergence between two distributions g_0 and g_1 is defined as $\mathfrak{D}_f(g_0||g_1) = \int f(g_0(x)/g_1(x))\, g_0(x)dx$. The KL divergence belongs to this class of divergences with $f(t) = t\ln t$. $\mathfrak{D}_{\mathrm{CA}}$ (induced by the optimal classifier) is also related to f-divergence.

Corollary 2 (Asymptotic Quasi-Posterior of $\mathfrak{D}_{\mathbf{CA}}$) *Let $n \to \infty$ and $m/n \to \alpha > 0$. If the optimal (Bayes) classifier $h(x) = \mathbb{I}(\alpha p_\theta(x) \geq p_{\theta^*}(x))$ induces $\mathfrak{D}_{CA}$ in (3) then Algorithm 1 with $\mathfrak{D}_{CA}$ yields the quasi-posterior distribution*

$$\lim_{n\to\infty} \pi(\theta|\boldsymbol{X}; \mathfrak{D}_{CA}, \epsilon) = \pi(\theta|\mathfrak{D}_f(p_{\theta^*}||p_\theta) + c(\alpha) < \epsilon) \\ \propto \pi(\theta)\mathbb{I}(\mathfrak{D}_f(p_{\theta^*}||p_\theta) + c(\alpha) < \epsilon)$$

where the constant $c(\alpha) = (\alpha \vee 1)/(1+\alpha)$ is the classification accuracy of the naive prediction rule $h'(x) = \mathbb{I}(\alpha \leq 1)$ and the f-divergence $\mathfrak{D}_f$ corresponds to $f(t) = (\alpha/t) \vee 1 - \alpha \vee 1$.

This result suggests that $\mathfrak{D}_{\mathrm{CA}}$ with a general classifier is a suboptimal estimator for f-divergence $\mathfrak{D}_f(p_{\theta^*}||p_\theta)$ since the classifier in use is an approximation of the optimal (Bayes) classifier.

Thirdly, our KL divergence method also resembles one of the Bayesian indirect inference method (8).

Corollary 3 (Asymptotic Quasi-Posterior of $\mathfrak{D}_{\mathbf{AL}}$) *Let $n \to \infty$ and $m/n \to \alpha > 0$. If an auxiliary model $\{p_A(\cdot|\phi) : \phi \in \Phi\}$ is bijective to the model $\{p_\theta : \theta \in \Theta\}$ and induces $\mathfrak{D}_{AL}$ in (8) then Algorithm 1 with $\mathfrak{D}_{AL}$ yields the quasi-posterior distribution*

$$\lim_{n\to\infty} \pi(\theta|\boldsymbol{X}; \mathfrak{D}_{AL}, \epsilon) = \pi(\theta|KL(p_\theta||p_{\theta^*}) < \epsilon) \\ \propto \pi(\theta)\mathbb{I}(KL(p_\theta||p_{\theta^*}) < \epsilon).$$

It is worth noting that similar results may not hold for $\mathfrak{D}_{\mathrm{MM}}$ defined by (4) and $\mathfrak{D}_{\mathrm{W2}}$ defined by (5), as their convergences have not been established in the literature except that [50] shows that $\mathfrak{D}_{\mathrm{MM}}$ converges to MM in some special cases. For summary-based data discrepancy (6), in general

$$\pi(\theta|\boldsymbol{X}; \mathfrak{D}_S, \epsilon) \to \pi(\theta|\|s(\theta^*) - s(\theta)\| < \epsilon)$$

if $s(\theta^*)$ and $s(\theta)$ are the limits of $S(\boldsymbol{X})$ and $S(\boldsymbol{Y})$ as $n, m \to \infty$. The semi-automatic method advocates an approximation of the posterior mean as the summary statistics. As $S(\boldsymbol{X}) = \mathbb{E}[\theta|\boldsymbol{X}] \to \theta^*$ and $S(\boldsymbol{Y}) = \mathbb{E}[\theta|\boldsymbol{Y}] \to \theta$, we have $\pi(\theta|\boldsymbol{X}; \mathfrak{D}_S, \epsilon) \to \pi(\theta|\|\theta^* - \theta\| < \epsilon)$. However, this appealing asymptotic quasi-posterior is a mirage because the semi-automatic construction provides only a projection of $\mathbb{E}[\theta|\boldsymbol{y}]$ onto the function class spanned by the regressors. In most cases, finding the posterior mean $\mathbb{E}[\theta|\boldsymbol{y}]$, the optimal estimator in the Bayesian sense, is even harder than the task of constructing summary statistics.

4 Experiments

We run experiments on five benchmark models: a bivariate Gaussian mixture model ($d = 2$), a $M/G/1$-queuing model ($d = 5$), a bivariate beta model ($d = 2$), a moving-average model of order 2 and length $d = 10$, and a multivariate g-and-k distribution ($d = 5$). In each experiment, we set $n = m$ and the tolerance threshold ϵ adaptively such that 50 of 10^5 proposed θ are accepted.

4.1 Toy Example: Gaussian Mixture Model

The univariate Gaussian mixture model services as a benchmark model in ABC literatures [20, 24]. Here we use a more challenging bivariate Gaussian mixture model $Z \sim \mathrm{Bernoulli}(p)$, $X|Z = 0 \sim \mathcal{N}(\mu_0, [0.5, -0.3; -0.3, 0.5])$, $X|Z = 1 \sim \mathcal{N}(\mu_1, [0.25, 0; 0, 0.25])$. Unknown parameter $\theta = (p, \mu_0, \mu_1)$ consists of the mixture ratio p and subpopulation means μ_0, μ_1. We perform ABC on $n = 500$ observed samples which are generated from the true parameters $p^* = 0.3$, $\mu_0^* = (+0.7, +0.7)$, $\mu_1^* = (-0.7, -0.7)$. The prior we used is $p \sim \mathrm{Uniform}[0, 1]$, $\mu_0, \mu_1 \sim \mathrm{Uniform}[-1, +1]^2$.

Figure 1 shows that the KL divergence $\mathfrak{D}_{\mathrm{KL}}$ visibly outperforms other methods. $\mathfrak{D}_{\mathrm{W2}}, \mathfrak{D}_{\mathrm{MM}}$ and $\mathfrak{D}_{\mathrm{S}}$ with the auxiliary MLE as summary roughly figure out the true parameters but mix up the two subpopulation means. The other two methods does not find the true parameters.

For estimating the mixture ratio p, the KL divergence method achieves a mean square error of 0.001, more accurate than other methods ($\mathfrak{D}_{\mathrm{W2}}$: 0.002, $\mathfrak{D}_{\mathrm{W2}}$: 0.015; $\mathfrak{D}_{\mathrm{CA}}$: 0.053; $\mathfrak{D}_{\mathrm{S}}$ with the auxiliary MLE as summary: 0.020; $\mathfrak{D}_{\mathrm{S}}$ with the semi-automatic construction as summary: 0.025).

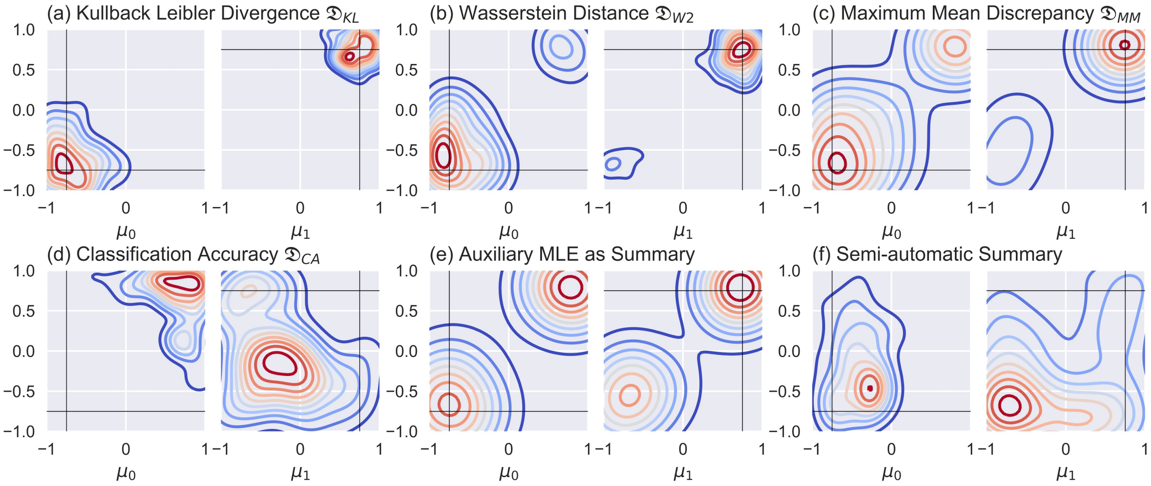

Figure 1: Quasi-posteriors for the Gaussian mixture model. Black lines cross at μ_0^* and μ_1^*.

4.2 $M/G/1$-queuing Model

Queuing models are usually easy to simulate from but have no intractable likelihoods. We adopts a specific $M/G/1$-queue that has been studied by ABC methods [55, 25]. In this model, the service times follows Uniform$[\theta_1, \theta_2]$ and the inter-arrival times are exponentially distributed with rate θ_3. Each datum is a 5-dimensional vector consisting of the first five inter-departure times $x = (x_1, x_2, x_3, x_4, x_5)$ after the queue starts from empty. Unknown parameter $\theta = (\theta_1, \theta_2, \theta_3)$. We perform ABC on $n = 500$ observed samples which are generated from the true parameters $\theta^* = (1, 5, 0.2)$. The prior we used is $\theta_1 \sim$ Uniform$[0, 10]$, $\theta_2 - \theta_1 \sim$ Uniform$[0, 10]$, $\theta_3 \sim$ Uniform$[0, 0.5]$.

Figure 2 shows that $\mathfrak{D}_{\mathrm{KL}}$ produces better inference for θ_1^* and θ_2^* then $\mathfrak{D}_{\mathrm{W2}}$ and $\mathfrak{D}_{\mathrm{MM}}$. It also captures θ_3^*, although the marginal quasi-posterior does not concentrate as much as those of $\mathfrak{D}_{\mathrm{W2}}$ and $\mathfrak{D}_{\mathrm{MM}}$. The summary statistics constructed by the semi-automatic method using marginal octiles identifies θ_1^* and give meaningless inference about θ_2^* and θ_3^*.

4.3 Bivariate Beta Model

Arnold and Ng [56] defines an 8-parameter model as follows. $U_i \sim \mathrm{Gamma}(\theta_i, 1)$, $i = 1, \ldots, 8$. Let

$$V_1 = (U_1 + U_5 + U_7)/(U_3 + U_6 + U_8),$$
$$V_2 = (U_2 + U_5 + U_8)/(U_4 + U_6 + U_7)$$

then $Z_1 = V_1/(1+V_1), Z_2 = V_2/(1+V_2))$ are marginally beta distributed, and $Z = (Z_1, Z_2)$ jointly follows a bivariate beta distribution. Crackel and Flegal [57] considers 5-parameter sub-model by restricting $\theta_3 = \theta_4 = \theta_5 = 0$. Another variant of this model was studied by [58]. We apply the methodology to Crackel and Flegal [57]'s sub-model.

We perform ABC on $n = 500$ observed samples which are generated from $\theta^* = (1, 1, 1, 1, 1)$. The parameter setting was used in [56]. The prior we used is $(\theta_1, \theta_2, \theta_6, \theta_7, \theta_8) \sim$ Uniform$[0, 5]^5$. Figure 3 shows an overall satisfactory performance of $\mathfrak{D}_{\mathrm{KL}}$ among others.

4.4 Moving-average Model of Order 2

Marin et al. [14] uses the moving-average model of order 2 as a benchmark model in their review. This model generates $Y_j = Y_j + \theta_1 Y_{j-1} + \theta_2 Y_{j-2}$, $j = 1, ..., d$ with Z_j being unobserved noise error terms. Each datum Y is a time series of length d. We take Z_j to follow Student's t distribution with 5 degrees of freedom and set $d = 10$. With the prior $(\theta_1, \theta_2) \sim$ Uniform$([-2, +2] \times [-1, +1])$, ABC was performed on $n = 200$ observed samples generated from $\theta^* = (0.6, 0.2)$. $\mathfrak{D}_{\mathrm{KL}}, \mathfrak{D}_{\mathrm{W2}}, \mathfrak{D}_{\mathrm{MM}}, \mathfrak{D}_{\mathrm{S}}$ with auxiliary MLE as summary obtain comparably high quality quasi-posteriors. See Figure 4.

4.5 Multivariate g-and-k Distribution

The univariate g-and-k distribution is defined by its inverse distribution function (9). It has no analytical form of the density function, and the numerical evaluation of the likelihood function is costly [59].

$$F^{-1}(x) = A + B\left[1 + c\frac{1 - e^{-gz_x}}{1 + e^{-gz_x}}\right](1 + z_x^2)^k z_x \quad (9)$$

where z_x is the x-th quantile of the standard normal distribution, and parameters A, B, g, k are related to location, scale, skewness and kurtosis, respectively, and $c = 0.8$ is the conventional choice [25]. As the inversion transform method can conveniently sample from this distribution by drawing $Z \sim \mathcal{N}(0, 1)$ i.i.d. and then transforming them to be g-and-k distributed random variables. [60, 5, 61, 25] have performed ABC on it. Multivariate g-and-k distribution has also been considered [62, 63]. Here we study a 5-dimensional g-and-k distribution: first draw $(Z_1, \ldots, Z_5)^T \sim \mathcal{N}(0, \Sigma)$

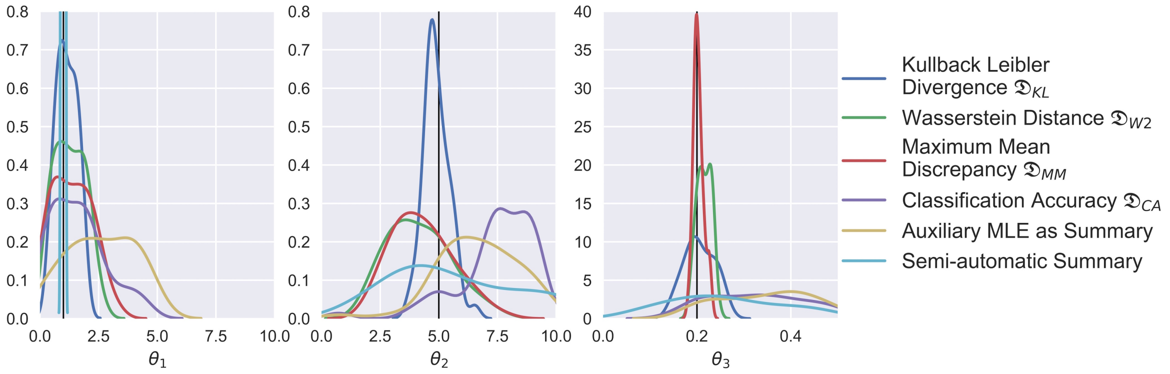

Figure 2: Quasi-posteriors for $M/G/1$-queue model. Black lines intercepts x-axis at true parameters.

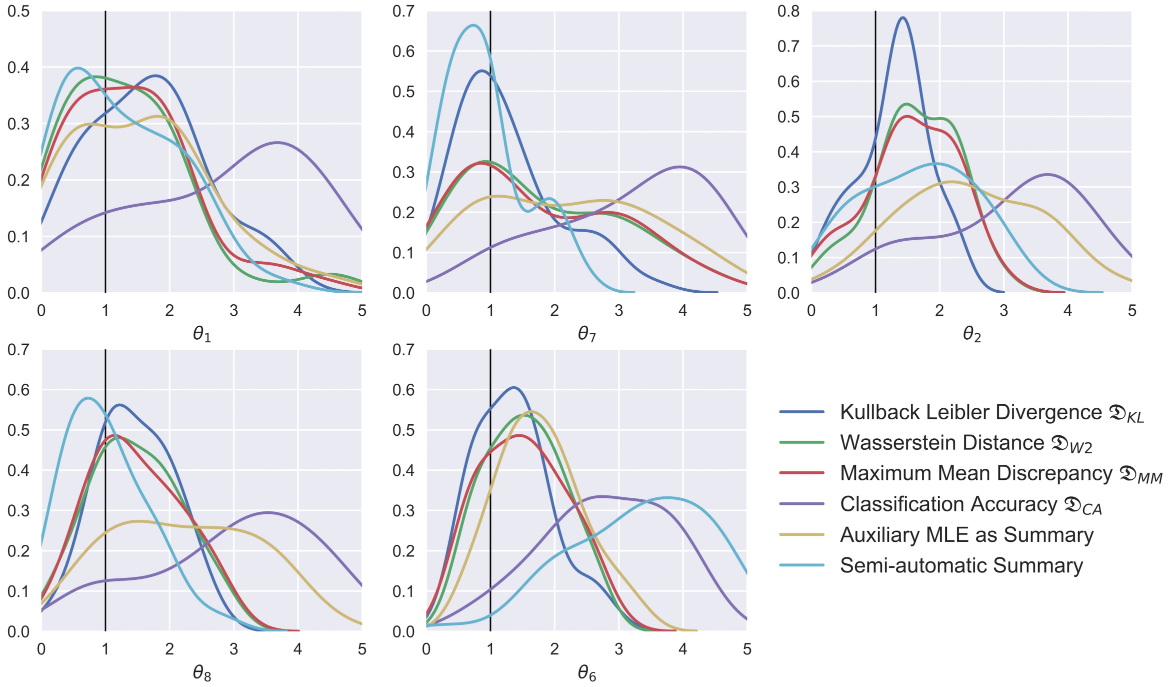

Figure 3: Quasi-posteriors for bivariate beta model. Black lines intercepts x-axis at true parameters.

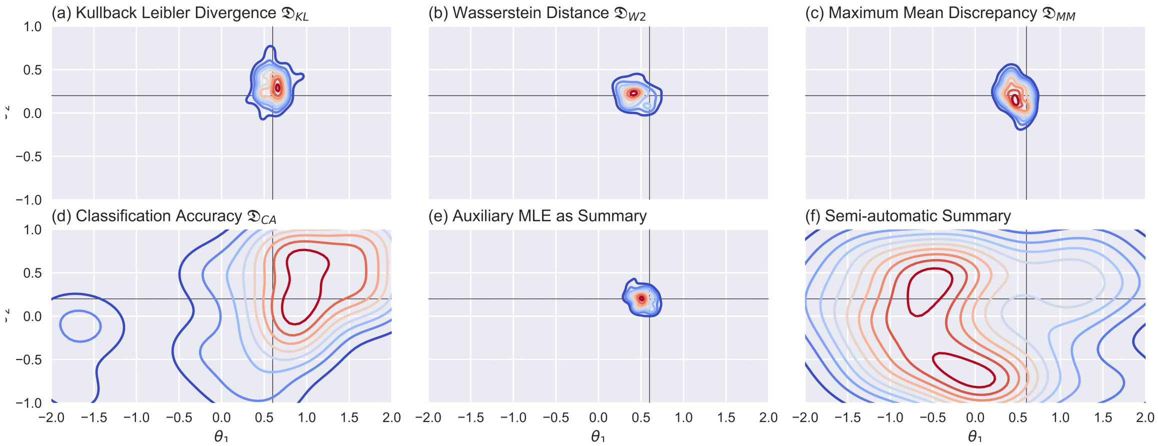

Figure 4: Quasi-posteriors for the moving-average model. Black lines cross at true parameters.

with Σ having sparse structures $\Sigma_{ii} = 1$ and $\Sigma_{ij} = \rho$ if $|i - j| = 1$ or 0 otherwise, and transform them marginally as the univariate g-and-k distribution does. Again $\mathfrak{D}_{\mathrm{KL}}$ obtain comparably high quality quasi-

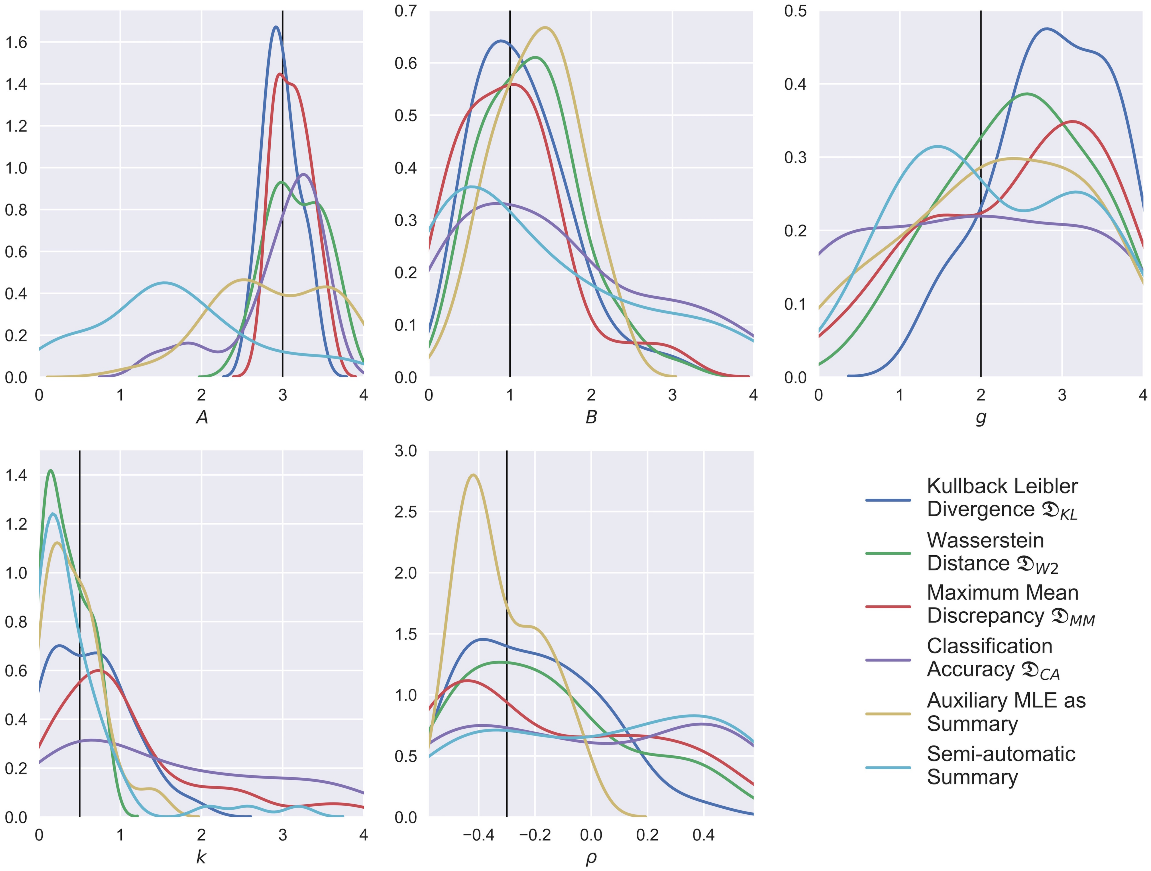

Figure 5: Quasi-posteriors for the multivariate g-and-k model. Solid lines cross at μ_0^* and μ_1^*.

posteriors. See Figure 5.

5 Discussion

We add the KL divergence estimators to the arsenal of data discrepancy measures for ABC. This estimator converges to the exact KL divergence. Thus $\mathfrak{D}_{\mathrm{KL}}$, as the data discrepancy for ABC, yields a quasi-posterior which eventually concentrates on $\{\theta : \mathrm{KL}(p_{\theta^*}||p_\theta) < \epsilon\}$ as the sample size increases. Our method and the maximum likelihood estimation share the same spirit of finding minimizers of the KL divergence. We also connect the KL divergence method to the classification accuracy method $\mathfrak{D}_{\mathrm{CA}}$ and a Bayesian indirect inference method $\mathfrak{D}_{\mathrm{AL}}$. Both two methods are somehow suboptimal to $\mathfrak{D}_{\mathrm{KL}}$, as they are equivalent to $\mathfrak{D}_{\mathrm{KL}}$ asymptotically only if their key ingredients (the prediction rule for $\mathfrak{D}_{\mathrm{CA}}$ or the auxiliary model for $\mathfrak{D}_{\mathrm{AL}}$) are "oracle" ones. We run experiments on five benchmark models and compare different data discrepancy measures. In terms of the quasi-posterior quality, $\mathfrak{D}_{\mathrm{KL}}$ performs comparably good or even better than $\mathfrak{D}_{\mathrm{MM}}, \mathfrak{D}_{\mathrm{W2}}$, and much better than other methods.

Apart from the quasi-posterior quality, another core consideration of ABC is the computational tractability of the data discrepancy function. A whole run of ABC analysis usually calls the data discrepancy function millions of times. In case of $n = m$, the $\mathcal{O}(n \ln n)$ cost per call of the KL divergence method is thus attractive, whereas $\mathfrak{D}_{\mathrm{MM}}$ and $\mathfrak{D}_{\mathrm{W2}}$ have $\mathcal{O}(n^2)$ cost per call.

Our theoretical results assume $m = \alpha n$ asymptotically, and our experiments use the typical setting $m = n$. In the experiments, our method performs well when $\alpha \in [0.5, 1.0]$. If α is close to 0 then the commonly-seen imbalance issue in two sample problems arises.

Our theoretical results also assume n goes to infinity. For finite data samples, we need the convergence rate or non-asymptotic error bounds of the KL estimator in use to quantify the quality of the quasi-posterior distribution. Unfortunately, apart from the almost sure convergence result for the KL estimator, there are few results to more precisely justify the KL estimators in the literature.

Acknowledgements

This work is supported by NSF Grants DMS-1407557 and DMS-1721550. It was completed when Bai Jiang was a Ph.D. student at Stanford University. The authors thank anonymous referees for their suggestions.

References

[1] Simon Tavaré, David J Balding, Robert C Griffiths, and Peter Donnelly. Inferring coalescence times from DNA sequence data. *Genetics*, 145(2):505–518, 1997.

[2] Mark A Beaumont, Wenyang Zhang, and David J Balding. Approximate Bayesian Computation in population genetics. *Genetics*, 162(4):2025–2035, 2002.

[3] Tina Toni, David Welch, Natalja Strelkowa, Andreas Ipsen, and Michael PH Stumpf. Approximate Bayesian Computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface*, 6 (31):187–202, 2009.

[4] Simon N Wood. Statistical inference for noisy nonlinear ecological dynamic systems. *Nature*, 466 (7310):1102–1104, 2010.

[5] Glynn W Peters and Scott A Sisson. Bayesian inference, Monte Carlo sampling and operational risk. *Journal of Operational Risk*, 1(3):27–50, 2006.

[6] Gareth W Peters, Scott A Sisson, and Yanan Fan. Likelihood-free Bayesian inference for α-stable models. *Computational Statistics & Data Analysis*, 56(11):3743–3756, 2012.

[7] Mark M Tanaka, Andrew R Francis, Fabio Luciani, and SA Sisson. Using Approximate Bayesian Computation to estimate tuberculosis transmission parameters from genotype data. *Genetics*, 173(3): 1511–1520, 2006.

[8] Michael GB Blum and Viet Chi Tran. HIV with contact tracing: a case study in Approximate Bayesian Computation. *Biostatistics*, 11(4):644–660, 2010.

[9] Trevelyan J McKinley, Joshua V Ross, Rob Deardon, and Alex R Cook. Simulation-based Bayesian inference for epidemic models. *Computational Statistics & Data Analysis*, 71:434–447, 2014.

[10] Jason Christopher, Caelan Lapointe, Nicholas Wimer, Torrey Hayden, Ian Grooms, Gregory B Rieker, and Peter E Hamlington. Parameter estimation for a turbulent buoyant jet using Approximate Bayesian Computation. In *55th AIAA Aerospace Sciences Meeting*, page 0531, 2017.

[11] Peter J Diggle and Richard J Gratton. Monte Carlo methods of inference for implicit statistical models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 193–227, 1984.

[12] Christopher M Bishop and Julia Lasserre. Generative or discriminative? getting the best of both worlds. *Bayesian statistics*, 8:3–24, 2007.

[13] Katalin Csilléry, Michael GB Blum, Oscar E Gaggiotti, and Olivier François. Approximate Bayesian Computation in practice. *Trends in ecology & evolution*, 25(7):410–418, 2010.

[14] Jean-Michel Marin, Pierre Pudlo, Christian P Robert, and Robin J Ryder. Approximate Bayesian Computational methods. *Statistics and Computing*, pages 1–14, 2012.

[15] Mikael Sunnåker, Alberto Giovanni Busetto, Elina Numminen, Jukka Corander, Matthieu Foll, and Christophe Dessimoz. Approximate Bayesian Computation. *PLoS Computational Biology*, 9(1): e1002803, 2013.

[16] Yun-Xin Fu and Wen-Hsiung Li. Estimating the age of the common ancestor of a sample of DNA sequences. *Molecular biology and evolution*, 14(2): 195–199, 1997.

[17] Gunter Weiss and Arndt von Haeseler. Inference of population history using a likelihood approach. *Genetics*, 149(3):1539–1546, 1998.

[18] Jonathan K Pritchard, Mark T Seielstad, Anna Perez-Lezaun, and Marcus W Feldman. Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular biology and evolution*, 16(12):1791–1798, 1999.

[19] Paul Marjoram, John Molitor, Vincent Plagnol, and Simon Tavaré. Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 100(26):15324–15328, 2003.

[20] Scott A Sisson, Yanan Fan, and Mark M Tanaka. Sequential Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 104 (6):1760–1765, 2007.

[21] Scott A Sisson, Yanan Fan, and Mark M Tanaka. Correction for "Sequential Monte Carlo without likelihoods". *Proceedings of the National Academy of Sciences*, 104(6):1760–1765, 2009.

[22] Christopher C Drovandi and Anthony N Pettitt. Estimation of parameters for macroparasite population evolution using Approximate Bayesian Computation. *Biometrics*, 67(1):225–233, 2011.

[23] Pierre Del Moral, Arnaud Doucet, and Ajay Jasra. An adaptive Sequential Monte Carlo method for Approximate Bayesian Computation. *Statistics and Computing*, 22(5):1009–1020, 2012.

[24] Richard D Wilkinson. Approximate Bayesian Computation (ABC) gives exact results under the assumption of model error. *Statistical applications in genetics and molecular biology*, 12(2):129–141, 2013.

[25] Paul Fearnhead and Dennis Prangle. 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, 2012.

[26] Michael GB Blum, Maria A Nunes, Dennis Prangle, and Scott A Sisson. A comparative review of dimension reduction methods in Approximate Bayesian Computation. *Statistical Science*, 28(2): 189–208, 2013.

[27] Christopher C Drovandi, Anthony N Pettitt, and Anthony Lee. Bayesian indirect inference using a parametric auxiliary model. *Statistical Science*, 30 (1):72–95, 2015.

[28] Paul Joyce and Paul Marjoram. Approximately sufficient statistics and Bayesian computation. *Statistical applications in genetics and molecular biology*, 7(1):26, 2008.

[29] Matthew A Nunes and David J Balding. On optimal selection of summary statistics for Approximate Bayesian Computation. *Statistical applications in genetics and molecular biology*, 9(1):34, 2010.

[30] Pierre Pudlo, Jean-Michel Marin, Arnaud Estoup, Jean-Marie Cornuet, Mathieu Gautier, and Christian P Robert. Reliable abc model choice via random forests. *Bioinformatics*, 32(6):859–866, 2015.

[31] Daniel Wegmann, Christoph Leuenberger, and Laurent Excoffier. Efficient Approximate Bayesian Computation coupled with Markov chain Monte Carlo without likelihood. *Genetics*, 182(4):1207–1218, 2009.

[32] Christian Gourieroux, Alain Monfort, and Eric Renault. Indirect inference. *Journal of applied econometrics*, 8(S1):S85–S118, 1993.

[33] Christopher C Drovandi, Anthony N Pettitt, and Malcolm J Faddy. Approximate bayesian computation using indirect inference. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 60(3):317–337, 2011.

[34] Alexander Gleim and Christian Pigorsch. Approximate Bayesian Computation with indirect summary statistics. *Draft paper: http://ect-pigorsch. mee. uni-bonn. de/data/research/papers*, 2013.

[35] Fernando Pérez-Cruz. Kullback-Leibler divergence estimation of continuous distributions. In *Information Theory, 2008. ISIT 2008. IEEE International Symposium on*, pages 1666–1670. IEEE, 2008.

[36] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

[37] Qing Wang, Sanjeev R Kulkarni, and Sergio Verdú. Divergence estimation of continuous distributions based on data-dependent partitions. *IEEE Transactions on Information Theory*, 51(9):3064–3074, 2005.

[38] XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. Estimating divergence functionals and the likelihood ratio by penalized convex risk minimization. In *NIPS*, pages 1089–1096, 2007.

[39] Qing Wang, Sanjeev R Kulkarni, and Sergio Verdú. Divergence estimation for multidimensional densities via k-nearest-neighbor distances. *IEEE Transactions on Information Theory*, 55(5):2392–2405, 2009.

[40] XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010.

[41] Jorge Silva and Shrikanth S Narayanan. Information divergence estimation based on data-dependent partitions. *Journal of Statistical Planning and Inference*, 140(11):3180–3198, 2010.

[42] Kirthevasan Kandasamy, Akshay Krishnamurthy, Barnabas Poczos, and Larry Wasserman. Nonparametric von mises estimators for entropies, divergences and mutual informations. In *Advances in Neural Information Processing Systems*, pages 397–405, 2015.

[43] Ritabrata Gutmann, Michael U Dutta, Samuel Kaski, and Jukka Corander. Likelihood-free inference via classification. *Statistics and Computing*, 28(2):411–425, 2018.

[44] Mijung Park, Wittawat Jitkrittum, and Dino Sejdinovic. K2-ABC: Approximate Bayesian Computation with kernel embeddings. In *Artificial Intelligence and Statistics*, pages 398–407, 2016.

[45] Espen Bernton, Pierre E Jacob, Mathieu Gerber, and Christian P Robert. Inference in generative models using the Wasserstein distance. *arXiv preprint arXiv:1701.05146*, 2017.

[46] Jon Louis Bentley. Multidimensional binary search trees used for associative searching. *Communications of the ACM*, 18(9):509–517, 1975.

[47] Songrit Maneewongvatana and David Mount. On the efficiency of nearest neighbor searching with data clustered in lower dimensions. *Computational Science—ICCS 2001*, pages 842–851, 2001.

[48] Alex Smola, Arthur Gretton, Le Song, and Bernhard Schölkopf. A Hilbert space embedding for

distributions. In *International Conference on Algorithmic Learning Theory*, pages 13–31. Springer, 2007.

[49] Alain Berlinet and Christine Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics.* Springer Science & Business Media, 2011.

[50] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

[51] Rainer Burkard, Mauro Dell'Amico, and Silvano Martello. Assignment problems: Society for industrial and applied mathematics, 2009.

[52] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, pages 2292–2300, 2013.

[53] Marco Cuturi and Arnaud Doucet. Fast computation of Wasserstein barycenters. In *International Conference on Machine Learning*, pages 685–693, 2014.

[54] Imre Csiszár and Paul C Shields. Information theory and statistics: a tutorial. *Foundations and Trends® in Communications and Information Theory*, 1(4):417–528, 2004.

[55] Michael GB Blum and Olivier François. Nonlinear regression models for Approximate Bayesian Computation. *Statistics and computing*, 20(1):63–73, 2010.

[56] Barry C Arnold and Hon Keung Tony Ng. Flexible bivariate beta distributions. *Journal of Multivariate Analysis*, 102(8):1194–1202, 2011.

[57] Roberto Crackel and James Flegal. Bayesian inference for a flexible class of bivariate beta distributions. *Journal of Statistical Computation and Simulation*, 87(2):295–312, 2017.

[58] Ingram Olkin and Ruixue Liu. A bivariate beta distribution. *Statistics & Probability Letters*, 62 (4):407–412, 2003.

[59] Glen D Rayner and Helen L MacGillivray. Numerical maximum likelihood estimation for the g-and-k and generalized g-and-h distributions. *Statistics and Computing*, 12(1):57–75, 2002.

[60] Michele Haynes and Kerrie Mengersen. Bayesian estimation of g-and-k distributions using MCMC. *Computational Statistics*, 20(1):7–30, 2005.

[61] David Allingham, R AR King, and Kerrie L Mengersen. Bayesian estimation of quantile distributions. *Statistics and Computing*, 19(2):189–201, 2009.

[62] Christopher C Drovandi and Anthony N Pettitt. Likelihood-free bayesian estimation of multivariate quantile distributions. *Computational Statistics & Data Analysis*, 55(9):2541–2556, 2011.

[63] Jingjing Li, David J Nott, Yanan Fan, and Scott A Sisson. Extending Approximate Bayesian Computation methods to high dimensions via a Gaussian copula model. *Computational Statistics & Data Analysis*, 106:77–89, 2017.