

---

# K2-ABC: Approximate Bayesian Computation with Kernel Embeddings

---

Mijung Park\*<sup>†</sup>  
mijungi.p@gmail.com  
Gatsby Unit  
University College London

Wittawat Jitkrittum\*  
wittawat@gatsby.ucl.ac.uk  
Gatsby Unit  
University College London

Dino Sejdinovic  
dino.sejdinovic@stats.ox.ac.uk  
Department of Statistics  
University of Oxford

## Abstract

Complicated generative models often result in a situation where computing the likelihood of observed data is intractable, while simulating from the conditional density given a parameter value is relatively easy. Approximate Bayesian Computation (ABC) is a paradigm that enables simulation-based posterior inference in such cases by measuring the similarity between simulated and observed data in terms of a chosen set of summary statistics. However, there is no general rule to construct sufficient summary statistics for complex models. Insufficient summary statistics will “leak” information, which leads to ABC algorithms yielding samples from an incorrect posterior. In this paper, we propose a fully nonparametric ABC paradigm which circumvents the need for manually selecting summary statistics. Our approach, *K2-ABC*, uses *maximum mean discrepancy* (MMD) to construct a dissimilarity measure between the observed and simulated data. The embedding of an empirical distribution of the data into a reproducing kernel Hilbert space plays a role of the summary statistic and is sufficient whenever the corresponding kernels are characteristic. Experiments on a simulated scenario and a real-world biological problem illustrate the effectiveness of the proposed algorithm.

## 1 INTRODUCTION

ABC is an approximate Bayesian inference framework for models with intractable likelihoods. Originated in population genetics [1], ABC has been widely used in a broad range of scientific applications including ecology, cosmology, and bioinformatics [2, 3, 4]. The goal of ABC is to obtain an approximate posterior distribution over model parameters, which usually correspond to interpretable inherent mechanisms in natural phenomena. However, in many complex models of interest, exact posterior inference is intractable since the likelihood function, the probability of observations  $y^*$  for given model parameters  $\theta$ , is either expensive or impossible to evaluate. This leads to a situation where the posterior density cannot be evaluated even up to a normalisation constant.

ABC resorts to an approximation of the likelihood function using simulated observations that are *similar* to the actual observations. Most ABC algorithms evaluate the similarity between the simulated and actual observations in terms of a *pre-chosen* set of summary statistics. Since the full dataset is represented in a lower-dimensional space of summary statistics, unless the selected summary statistic  $s^*$  is sufficient, ABC results in inference on the approximate posterior  $p(\theta|s^*)$ , rather than the desired full posterior  $p(\theta|y^*)$ . Therefore, a poor choice of a summary statistic can lead to an additional bias that can be difficult to quantify and the selection of summary statistics is a crucial step in ABC [5, 6]. A number of methods have been proposed for constructing informative summary statistics by appropriate transformations of a set of candidate summary statistics: a minimum-entropy approach [7], regression from a set of candidate summary statistics to parameters [8], or a partial least-squares transformation with boosting [9, 10]. A review of these methods is given in [9]. However, they still heavily depend on the initial choice of candidate summary statistics and may not suffice for full posterior inference.

---

\*M Park and W Jitkrittum contributed equally.

<sup>†</sup>Currently at QUvA lab, University of Amsterdam

Another line of research, inspired by indirect inference framework, constructs the summary statistics using auxiliary models [11, 12, 13]. Here, the estimated parameters of an auxiliary model become the summary statistics for ABC. A thorough review of these method is given in [10]. The biggest advantage of this framework is that the dimensionality of the summary statistic can be determined by controlling the complexity of the auxiliary model using standard model selection techniques such as Bayesian information criterion (BIC) and Akaike information criterion (AIC). However, even if the complexity can be set in a principled way, one still needs to select the right parametric form of the auxiliary model.

In this light, we introduce a method that circumvents an explicit selection of summary statistics. Our method proceeds by applying a similarity measure to *data themselves*, via embedding [14, 15] of the empirical data distributions into an infinite-dimensional reproducing kernel Hilbert space (RKHS), corresponding to a positive definite kernel function defined on the data space. For suitably chosen kernels called characteristic [16], these embeddings capture all possible differences between distributions, e.g. all the high-order moment information that may be missed with a set of simple summary statistics. Thus, no information loss is incurred when going from the posterior given data  $p(\theta|y^*)$  to that given the embedding  $\mu(y^*)$  of data,  $p(\theta|\mu(y^*))$ . A flexible representation of probability measures given by kernel embeddings has already been applied to nonparametric hypothesis testing [17], inference in graphical models [18] and to proposal construction in adaptive MCMC [19]. The key quantity arising from this framework is an easily computable notion of a distance between probability measures, termed Maximum Mean Discrepancy (MMD). When the kernel used is characteristic, a property satisfied by most commonly used kernels including Gaussian, Laplacian and inverse multiquadrics, embeddings are injective, meaning that the MMD gives a metric on probability measures. Here, we adopt MMD in the context of ABC as a nonparametric distance between empirical distributions of simulated and observed data. As such, there is no need to select a summary statistic first and the kernel embedding itself plays a role of a summary statistic. In addition to the positive definite kernel used for obtaining the estimates of MMD, we apply an additional Gaussian smoothing kernel which operates on the corresponding RKHS, i.e., on embeddings themselves, to obtain a measure of similarity between simulated and observed data. For this reason, we refer to our method as *double-kernel* ABC, or *K2-ABC*. Our experimental results in section 4 demonstrate that this approach results in an effective and robust ABC method which requires no hand-crafted

summary statistics.

The rest of the paper is organised as follows. In section 2, we overview classical approaches (rejection and soft ABC) as well as several relevant recent techniques (synthetic likelihood ABC, semi-automatic ABC, indirect score ABC, and kernel ABC) to which we will compare our method in section 4. In section 3, we introduce our proposed algorithm. Experimental results including comparisons with methods discussed in section 2, are presented in section 4. We explore computational tractability of K2-ABC in section 5.

## 2 BACKGROUND

We start by introducing ABC and reviewing existing algorithms. Consider a situation where it is possible to simulate a generative model and thus sample from the conditional density  $p(y|\theta)$ , given a value  $\theta \in \Theta$  of parameters, while the computation of the likelihood  $p(y^*|\theta)$  for the observed data  $y^*$  is intractable. Neither exact posterior inference nor posterior sampling are possible in this case, as the posterior  $p(\theta|y^*) \propto \pi(\theta)p(y^*|\theta)$ , for a prior  $\pi$ , cannot be computed up to a normalizing constant. ABC uses an approximation of the likelihood obtained from simulation.

The simplest form of ABC is rejection ABC. Let  $\epsilon > 0$  be a *similarity threshold*, and  $\rho$  be a notion of distance, e.g., a premetric on domain  $\mathcal{Y}$  of observations. The rejection ABC proceeds by sampling multiple model parameters  $\theta \sim \pi$ . For each  $\theta$ , a pseudo dataset  $y$  is generated from  $p(y|\theta)$ . The parameters  $\theta$  for which the generated  $y$  are similar to the observed  $y^*$ , as decided by  $\rho(y, y^*) < \epsilon$ , are accepted. The result is an exact sample  $\{\theta_i\}_{i=1}^M$  from the approximated posterior  $\tilde{p}_\epsilon(\theta|y^*) \propto \pi(\theta)\tilde{p}_\epsilon(y^*|\theta)$ , where  $\tilde{p}_\epsilon(y^*|\theta) = \int_{B_\epsilon(y^*)} p(y|\theta)dy$  and  $B_\epsilon(y^*) = \{y : \rho(y, y^*) < \epsilon\}$ . Choice of  $\rho$  is crucial for the design of an accurate ABC algorithm. Applying a distance directly on dataset  $y$  is often challenging, when the dataset consists of a large number of (possibly multivariate) observations.

Thus, one resorts to first choosing a summary statistic  $s(y)$  and comparing them between the datasets, i.e.  $\rho(y, y') = \|s(y) - s(y')\|$ . Since it is generally difficult to construct sufficient statistics for complex models, this will often “leak” information, e.g., if  $s(y)$  represents first few empirical moments of dataset  $y$ . It is only when the summary statistic  $s$  is sufficient, that this approximation is consistent as  $\epsilon \rightarrow 0$ , i.e. that the ABC posterior  $\tilde{p}_\epsilon(\theta|y^*)$  will converge to the full posterior. Otherwise, ABC operates on the approximate posterior  $p(\theta|s(y^*))$ , rather than the full posterior  $p(\theta|y^*)$ .

Another interpretation of the approximate likelihood  $\tilde{p}_\epsilon(y^*|\theta)$  is as the convolution of the true likelihood  $p(y|\theta)$  and the ‘‘similarity’’ kernel  $\kappa_\epsilon(y, y^*) = \mathbf{1}(y \in B_\epsilon(y^*))$ . Being the indicator of the  $\epsilon$ -ball  $B_\epsilon(y^*)$  computed w.r.t.  $\rho$ , this kernel imposes a hard-constraint leading to the rejection sampling. In fact, one can use *any* similarity kernel parametrised by  $\epsilon$  which approaches delta function  $\delta_{y^*}$  as  $\epsilon \rightarrow 0$ . A frequently used similarity kernel takes the form

$$\kappa_\epsilon(y, y') = \exp\left(-\frac{\rho^q(y, y')}{\epsilon}\right), \text{ for } q > 0. \quad (1)$$

Such construction would result in a weighted sample  $\{(\theta_j, w_j)\}_{j=1}^M$ , where  $w_j = \frac{\kappa_\epsilon(y_j, y^*)}{\sum_{l=1}^M \kappa_\epsilon(y_l, y^*)}$ , which can be directly utilised in estimating posterior expectations. That is, for a test function  $f$ , the expectation  $\int_{\Theta} f(\theta)p(\theta|y^*)d\theta$  is estimated using  $\hat{\mathbb{E}}[f(\theta)] = \sum_{i=1}^M w_i f(\theta_i)$ . This is an instance of what we call *soft ABC*, in which parameter samples from the prior are weighted, rather than accepted or rejected.

**Synthetic likelihood ABC (SL-ABC)** Introduced in [20], the synthetic likelihood ABC models simulated data in terms of their summary statistics and further assumes the summary statistics have multivariate normal distribution,  $s \sim \mathcal{N}(\hat{\mu}_\theta, \hat{\Sigma}_\theta)$ , with the empirical mean and covariance defined by

$$\hat{\mu}_\theta = \frac{1}{M} \sum_{i=1}^M s_i, \quad \hat{\Sigma}_\theta = \frac{1}{M-1} \sum_{i=1}^M (s_i - \hat{\mu}_\theta)(s_i - \hat{\mu}_\theta)^\top,$$

where  $s_i$  denotes the vector of summary statistics of the  $i$ th simulated dataset. Using the following similarity kernel to measure the distance from the summary statistics of actual observations  $s^*$ ,  $\kappa_\epsilon(s^*, s) = |2\pi\epsilon I|^{-\frac{1}{2}} \exp\left(-\frac{\|s-s^*\|_2^2}{2\epsilon^2}\right)$ , the resulting synthetic likelihood is given by

$$\begin{aligned} p(s^*|\theta) &= \int \kappa_\epsilon(s^*, s) \mathcal{N}(s|\hat{\mu}_\theta, \hat{\Sigma}_\theta) ds \\ &= \mathcal{N}(s^*|\hat{\mu}_\theta, \hat{\Sigma}_\theta + \epsilon^2 I). \end{aligned}$$

Relying on the synthetic likelihood, SL-ABC algorithm performs MCMC sampling based on Metropolis-Hastings accept/reject steps with the acceptance probability given by  $\alpha(\theta'|\theta) = \min\left[1, \frac{\pi(\theta')p(s^*|\theta')q(\theta|\theta')}{\pi(\theta)p(s^*|\theta)q(\theta'|\theta)}\right]$ , where  $q(\theta|\theta')$  is a proposal distribution.

**Semi-Automatic ABC (SA-ABC)** Under a quadratic loss, [8] shows that the optimal choice of the summary statistics is the true posterior means of the parameters  $\mathbb{E}[\theta|y]$  – however, these cannot be calculated analytically. Thus, [8] proposed to use the simulated data in order to construct new summary statistics – estimates of the posterior means of the parameters – by fitting a vector-valued regression model from

a set of candidate summary statistics to parameters. Namely, a linear model  $\theta = \mathbb{E}[\theta|y] + \varepsilon = Bg(y) + \varepsilon$  with the vector of candidate summary statistics  $g(y)$  used as explanatory variables (these can be simply  $g(y) = y$  or also include nonlinear functions of  $y$ ) is fitted based on the simulated data  $\{(y_i, \theta_i)\}_{i=1}^M$ . Here, assuming  $\theta \in \Theta \subset \mathbb{R}^d$  and  $g(y) \in \mathbb{R}^r$ ,  $B$  is a  $d \times r$  matrix of coefficients. The resulting estimates  $s(y) = \hat{B}g(y)$  of  $\mathbb{E}[\theta|y]$  can then be used as summary statistics in standard ABC by setting  $\rho(y, y^*) = \|s(y) - s(y^*)\|_2$ .

**Indirect score ABC (IS-ABC)** Based on an auxiliary model  $p_A(y|\phi)$  with a vector of parameters  $\phi = [\phi_1, \dots, \phi_{\dim(\phi)}]^\top$ , the indirect score ABC uses a score vector  $S_A$  as the summary statistic [12]:  $S_A(y, \phi) = \left[\frac{\partial \log p_A(y|\phi)}{\partial \phi_1}, \dots, \frac{\partial \log p_A(y|\phi)}{\partial \phi_{\dim(\phi)}}\right]^\top$ . When the auxiliary model parameters  $\phi$  are set by the maximum-likelihood estimate (MLE) fitted with observed data  $y^*$ , the score for the observed data  $S_A(y^*, \phi_{\text{MLE}}(y^*))$  becomes zero. Based on this fact, IS-ABC searches for the parameter values whose corresponding simulated data produce a score close to zero. The discrepancy between observed and simulated data distributions in IS-ABC is given by

$$\begin{aligned} \rho(s(y), s(y^*)) \\ = \sqrt{S_A(y, \phi_{\text{MLE}}(y^*))^\top J(\phi_{\text{MLE}}(y^*)) S_A(y, \phi_{\text{MLE}}(y^*))}, \end{aligned}$$

where  $J(\phi_{\text{MLE}}(y^*))$  is the approximate covariance matrix of the observed score.

**Kernel ABC (K-ABC)** The use of a positive definite kernel in ABC has been explored recently in [21] (K-ABC) in the context of population genetics. In K-ABC, ABC is cast as a problem of estimating a conditional mean embedding operator mapping from summary statistics  $s(y)$  to corresponding parameters  $\theta$ . The problem is equivalent to learning a regression function in the RKHSs of  $s(y)$  and  $\theta$  induced by their respective kernels [22]. The training set  $\mathcal{T}$  needed for learning the regression function is generated by firstly sampling  $\{(y_i, \theta_i)\}_{i=1}^M \sim p(y|\theta)\pi(\theta)$  from which  $\mathcal{T} := \{(s_i, \theta_i)\}_{i=1}^M$  by summarising each pseudo dataset  $y_i$  into a summary statistic  $s_i$ .

In effect, given a summary statistic  $s^*$  corresponding to the observations  $y^*$ , the learned regression function allows one to represent the embedding of the posterior distribution in the form of a weighted sum of the canonical feature maps  $\{k(\cdot, \theta_i)\}_{i=1}^M$  where  $k$  is a kernel associated with an RKHS  $\mathcal{H}_\theta$ . In particular, if we assume that  $k$  is a linear kernel (as in [21]), the posterior expectation of a function  $f \in \mathcal{H}_\theta$  is given by  $\mathbb{E}[f(\theta)|s^*] \approx \sum_{i=1}^M w_i(s^*)f(\theta_i)$  where  $w_i(s^*) = \sum_{j=1}^M ((G + M\lambda I)^{-1})_{ij} k(s_j, s^*)$ ,  $G_{ij} = g(s_i, s_j)$ ,  $g$  is

a kernel on  $s$ , and  $\lambda$  is a regularization parameter. The use of a kernel  $g$  on summary statistics  $s$  implicitly transforms  $s$  non-linearly, thereby increasing the representativeness of  $s$ . Nevertheless, the need for summary statistics is not eliminated.

### 3 PROPOSED METHOD

We first overview kernel MMD, a notion of distance between probability measures that is used in the proposed K2-ABC algorithm.

**Kernel MMD** For a probability distribution  $F_x$  on a domain  $\mathcal{X}$ , its kernel embedding is defined as  $\mu_{F_x} = \mathbb{E}_{X \sim F_x} k(\cdot, X)$  [15], an element of an RKHS  $\mathcal{H}$  associated with a positive definite kernel  $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ . An embedding exists for any  $F_x$  whenever the kernel  $k$  is bounded, or if  $F_x$  satisfies a suitable moment condition w.r.t. an unbounded kernel  $k$  [23]. For any two given probability measures  $F_x$  and  $F_y$ , their maximum mean discrepancy (MMD) is simply the Hilbert space distance between their embeddings:

$$\begin{aligned} \text{MMD}^2(F_x, F_y) &= \|\mu_{F_x} - \mu_{F_y}\|_{\mathcal{H}}^2 \\ &= \mathbb{E}_X \mathbb{E}_{X'} k(X, X') + \mathbb{E}_Y \mathbb{E}_{Y'} k(Y, Y') - 2\mathbb{E}_X \mathbb{E}_Y k(X, Y), \end{aligned}$$

where  $X, X' \stackrel{i.i.d.}{\sim} F_x$  and  $Y, Y' \stackrel{i.i.d.}{\sim} F_y$ . While simple kernels like polynomial of order  $r$  capture differences in first  $r$  moments of distributions, particularly interesting are kernels with a *characteristic* property<sup>1</sup> [16], for which the kernel embedding is injective and thus MMD defined by such kernels gives a metric on the space of probability distributions. Examples of such kernels include widely used kernels such as Gaussian RBF and Laplacian. Being written in terms of expectations of kernel functions allows straightforward estimation of MMD on the basis of samples: given  $\{x^{(i)}\}_{i=1}^{n_x} \stackrel{i.i.d.}{\sim} F_x$ ,  $\{y^{(j)}\}_{j=1}^{n_y} \stackrel{i.i.d.}{\sim} F_y$ , an unbiased estimator is given by

$$\begin{aligned} \widehat{\text{MMD}}^2(F_x, F_y) &= \frac{1}{n_x(n_x - 1)} \sum_{i=1}^{n_x} \sum_{j \neq i}^{n_x} k(x^{(i)}, x^{(j)}) \\ &\quad + \frac{1}{n_y(n_y - 1)} \sum_{i=1}^{n_y} \sum_{j \neq i}^{n_y} k(y^{(i)}, y^{(j)}) \\ &\quad - \frac{2}{n_x n_y} \sum_{i=1}^{n_x} \sum_{j=1}^{n_y} k(x^{(i)}, y^{(j)}). \end{aligned}$$

Further operations are possible on kernel embeddings - one can define a positive definite kernel on probability measures themselves using their representation

in a Hilbert space. An example of a kernel on probability measures is  $\kappa_\epsilon(F_x, F_y) = \exp\left(-\frac{\widehat{\text{MMD}}^2(F_x, F_y)}{\epsilon}\right)$ , [24] with  $\epsilon > 0$ . This has recently led to a thread of research tackling the problem of learning on distributions, e.g., [25]. These insights are essential to our contribution, as we employ such kernels on probability measures in the design of the K2-ABC algorithm which we describe next.

**K2-ABC** The first component of K2-ABC is a non-parametric distance  $\rho$  between empirical data distributions. Given two datasets  $y = (y^{(1)}, \dots, y^{(n)})$  and  $y' = (y'^{(1)}, \dots, y'^{(n)})$  consisting of  $n$  i.i.d. observations<sup>2</sup>, we use MMD to measure the distance between  $y, y'$ :  $\rho^2(y, y') = \widehat{\text{MMD}}^2(F_y, F_{y'})$ , i.e.  $\rho^2$  is an unbiased estimate of  $\text{MMD}^2$  between probability distributions  $F_y$  and  $F_{y'}$  used to generate  $y$  and  $y'$ . This is almost the same as setting empirical kernel embedding  $s(y) = \mu_{\hat{F}_y} = \sum_{j=1}^n k(\cdot, y^{(j)})$  to be the summary statistic. However, in that case  $\|s(y) - s(y')\|_{\mathcal{H}}^2 = \text{MMD}^2(\hat{F}_x, \hat{F}_y)$  would have been a biased estimate of the population  $\text{MMD}^2$  [17]. Our choice of  $\rho$  is guaranteed to capture all possible differences (i.e. all moments) between  $F_y$  and  $F_{y'}$  whenever a characteristic kernel  $k$  is employed [16], i.e. we are operating on a full posterior and there is no loss of information due to the use of insufficient statistics.

---

**Algorithm 1** K2-ABC Algorithm

---

**Input:** observed data  $y^*$ , prior  $\pi$ , soft threshold  $\epsilon$

**Output:** Empirical posterior  $\sum_{i=1}^M w_i \delta_{\theta_i}$

**for**  $i = 1, \dots, M$  **do**

Sample  $\theta_i \sim \pi$

Sample pseudo dataset  $y_i \sim p(\cdot|\theta_i)$

$\tilde{w}_i = \exp\left(-\frac{\widehat{\text{MMD}}^2(F_{y_i}, F_{y^*})}{\epsilon}\right)$

**end for**

$w_i = \tilde{w}_i / \sum_{j=1}^M \tilde{w}_j$  for  $i = 1, \dots, M$

---

Further, we introduce a second kernel into the ABC algorithm (summarised in Algorithm 1), the one that operates directly on probability measures, and compute the ABC posterior sample weights,

$$\kappa_\epsilon(F_y, F_{y'}) = \exp\left(-\frac{\widehat{\text{MMD}}^2(F_y, F_{y'})}{\epsilon}\right), \quad (2)$$

with a suitably chosen parameter  $\epsilon > 0$ . Now, the datasets are compared using the estimated similarity  $\kappa_\epsilon$  between their generating distributions. There are

---

<sup>1</sup>A related notion of *universality* is often employed.

---

<sup>2</sup>The i.i.d. assumption can be relaxed in practice, as we demonstrate in section 4 on time series data.

two sets of parameters in K2-ABC, parameters of kernel  $k$  (on original domain) and  $\epsilon$  in the kernel  $\kappa_\epsilon$  (on probability measures).

K2-ABC is readily applicable to non-Euclidean input objects if a kernel  $k$  can be defined on them. Arguably the most common application of ABC is to genetic data. Over the past years there have been a number of works addressing the use of kernel methods for studying genetic data including [26] which considered genetic association analysis, and [27] which studied gene-gene interaction. Generic kernels for strings, graphs and other structured data have also been explored [28].

## 4 EXPERIMENTS

**Toy problem** We start by illustrating how the choice of summary statistics can significantly affect the inference result, especially when the summary statistics are not sufficient. We consider a symmetric Dirichlet prior  $\pi$  and a likelihood  $p(y|\theta)$  given by a mixture of uniform distributions as

$$\begin{aligned} \pi(\theta) &= \text{Dirichlet}(\theta; \mathbf{1}), \\ p(y|\theta) &= \sum_{i=1}^5 \theta_i \text{Uniform}(y; [i-1, i]). \end{aligned} \quad (3)$$

The model parameters  $\theta$  are a vector of mixing proportions. The goal is to estimate  $\mathbb{E}[\theta|y^*]$  where  $y^*$  is generated with true parameter  $\theta^* = [0.25, 0.04, 0.33, 0.04, 0.34]^\top$  (see Fig. 1A). The summary statistics are chosen to be empirical mean and variance i.e.  $s(y) = (\hat{\mathbb{E}}[y], \hat{\mathbb{V}}[y])^\top$ .

We compare three ABC algorithms: K2-ABC<sup>3</sup>, rejection ABC, and soft ABC. Here, soft ABC refers to an ABC algorithm which uses a similarity kernel in Eq. 1 with  $q = 2$  and  $\rho(y, y') = \|s(y) - s(y')\|_2$ . For K2-ABC, we set  $M = 1000$  and used a Gaussian kernel defined as  $k(a, b) = \exp\left(-\frac{\|a-b\|_2^2}{2\gamma^2}\right)$  to compute  $\widehat{\text{MMD}}_2$  where  $\gamma$  is set to  $\text{median}(\{\|y^{*(i)} - y^{*(j)}\|\}_{i,j})$  [29]. We test different values of  $\epsilon$  on a coarse grid, and report the estimated  $\mathbb{E}[\theta|y^*]$  which is closest to  $\theta^*$  as measured with a Euclidean distance.

The results are shown in Fig. 1 where the top row shows the estimated  $\mathbb{E}[\theta|y^*]$  from each method, associated with the best  $\epsilon$  as reported in the third row. The second row of Fig. 1, from left to right, shows  $y^*$  and 400 realizations of  $y$  drawn from  $p(y|\mathbb{E}[\theta|y^*])$  obtained from the three algorithms. In all cases, the mean and variance of the drawn realizations match that of  $y^*$ . However, since the first two moments are insufficient

<sup>3</sup>Code for K2-ABC is available at <https://github.com/wittawatj/k2abc>.

to characterise  $p(y|\theta^*)$ , there exists other  $\theta'$  that can give rise to the same  $s(y')$ , which yields inaccurate posterior means shown in the top row. In contrast, K2-ABC taking into account infinite-dimensional sufficient statistic correctly estimates the posterior mean.

**Ecological dynamic systems** As an example of statistical inference for ecological dynamic systems, we use observations on adult blowfly populations over time introduced in [20]. The population dynamics are modelled by a discretised differential equation:

$$N_{t+1} = PN_{t-\tau} \exp\left(-\frac{N_{t-\tau}}{N_0}\right) e_t + N_t \exp(-\delta\epsilon_t),$$

where an observation at time  $t+1$  is denoted by  $N_{t+1}$  which is determined by time-lagged observations  $N_t$  and  $N_{t-\tau}$  as well as Gamma distributed noise realisations  $e_t \sim \text{Gam}(\frac{1}{\sigma_p^2}, \sigma_p^2)$  and  $\epsilon_t \sim \text{Gam}(\frac{1}{\sigma_d^2}, \sigma_d^2)$ . Here, the parameters are  $\theta = \{P, N_0, \sigma_d, \sigma_p, \tau, \delta\}$ . We put broad Gaussian priors on log of parameters as shown in Fig. 2A. Note that the time series data given the parameters drawn from the priors vary drastically (see Fig. 2B), and therefore inference with those data is very challenging as noted in [30].

The observation (black trace in Fig. 2B) is a time-series of length  $T = 180$ , where each point in time indicates how many flies survive at each time under food limitation. For SL-ABC and K-ABC, we adopted the custom 10 summary statistics used in [30]: the log of the mean of all 25% quantiles of  $\{N_t/1000\}_{t=1}^T$  (four statistics), the mean of 25% quantiles of the first-order differences of  $\{N_t/1000\}_{t=1}^T$  (four statistics), and the maximal peaks of smoothed  $\{N_t\}_{t=1}^T$ , with two different thresholds (two statistics). For IS-ABC, following [10], we use a Gaussian mixture model with three components as an auxiliary model. In addition, we ran two versions of SA-ABC algorithm on this example: SAQ regresses to simulated parameters from the corresponding simulated instances of time-series appended with the quadratic terms, i.e.,  $g(y) = (y, y^2) \in \mathbb{R}^{2T}$ , whereas SA-custom uses the above custom 10 summary statistics from [30] appended with their squares as the candidate summary-statistics vector  $g(y)$  (which is thus 20-dimensional in this instance).<sup>4</sup>

For setting  $\epsilon$  and kernel parameters in K2-ABC, we split the data into two sets: training (75% of 180 data points) and test (the rest) sets. Using the training data, we ran each ABC algorithm given each value of  $\epsilon$  and kernel parameters defined on a coarse grid, then,

<sup>4</sup>For SL-ABC, we used the Python implementation by the author of [30]. For IS-ABC, we used the MATLAB implementation by the author of [10]. For SA-ABC, we used the R package `abctools` written by the author of [8].

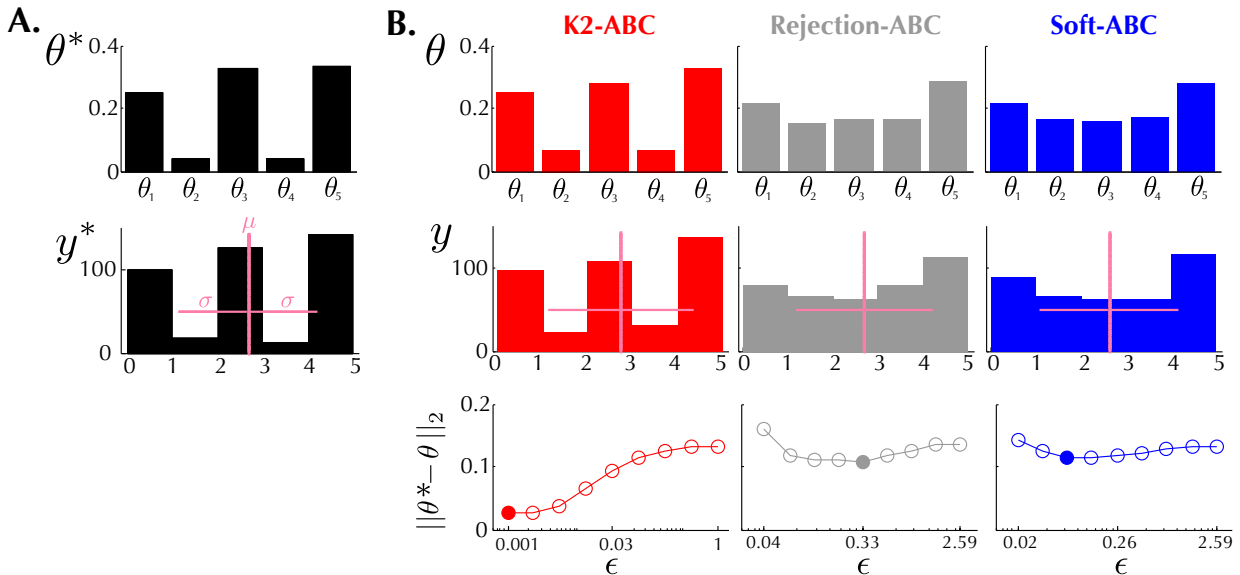


Figure 1: A possible scenario for ABC algorithms to fail due to insufficient summary statistics. **A**: Using the 5-dimensional true parameters (top), 400 observations (bottom) are sampled from the mixture of uniform distributions in Eq. 3. **B (top)**: Estimated posterior mean of parameters from each method. We drew 1000 parameters from the Dirichlet prior and 400 simulated data points given each parameter. In rejection and soft ABC algorithms, we used empirical mean and variance of observations as summary statistics to determine similarity between simulated and observed data. **B (middle)**: Histograms of 400 simulated data points given estimated posterior means by each method. Though the mean and variance of simulated data from rejection and soft ABC match that of the observed data, the shapes of the empirical distributions notably differ. **B (bottom)**: Euclidean distance between true and estimated posterior mean of parameters as a function of  $\epsilon$ . We varied the  $\epsilon$  values to find the optimal range in terms of the Euclidean distance. The magnitude of  $\epsilon$  is algorithm-specific and not comparable across methods.

computed test error<sup>5</sup> to choose the optimal values of  $\epsilon$  and kernel parameters in terms of the minimum prediction error. Finally, with the chosen  $\epsilon$  and kernel parameters, we ran each ABC algorithm using the entire data ( $M = 1000$ ).

We show the concentrated posterior mass after performing K2-ABC in Fig. 2A, as well as an example trajectory drawn from the inferred posterior mean in Fig. 2B<sup>6</sup>. To quantify the accuracy of each method, we compute the Euclidean distance between the vector of chosen 10 summary statistics  $s^* = s(y^*)$  for the observed data and  $s(y)$  for the simulated data  $y$  given the estimated posterior mean  $\hat{\theta}$  of the parameters. As shown in Fig. 3, K2-ABC outperforms other methods, although SL-ABC, SA-custom, and K-ABC all

<sup>5</sup>We used Euclidean distance between the histogram (with 10 bins) of test data and that of predictions made by each method. We chose the difference in histogram rather than in the realisation of  $y$  itself, to avoid the error due to the time shift in  $y$ .

<sup>6</sup>Note that reproducing the trajectory exactly *is not* the main goal of this experiment. We show the example trajectory to give the readers a sense of what the trajectory looks like.

explicitly operate on this vector of summary statistics  $s$  while K2-ABC does not. In other words, while those methods attempt to explicitly pin down the parts of the parameter space which produce summary statistics  $s$  similar to  $s^*$ , insufficiency of these summary statistics affects the posterior mean estimates undesirably even with respect to that very metric.

## 5 COMPUTATIONAL TRACTABILITY

In K2-ABC, given a dataset and pseudo dataset with  $n$  observations each, the cost for computing  $\widehat{\text{MMD}}^2(F_{y_i}, F_{y^*})$  is  $O(n^2)$ . For  $M$  pseudo datasets, the total cost then becomes  $O(Mn^2)$ , which can be prohibitive for a large number of observations. Since computational tractability is among the core considerations for ABC, in this section, we examine the performance of K2-ABC with different MMD approximations which reduce the computational cost.

**Linear-time MMD** The unbiased linear-time MMD estimator presented in [17, section 6] reduces the total cost to  $O(Mn)$  at the price of a higher vari-

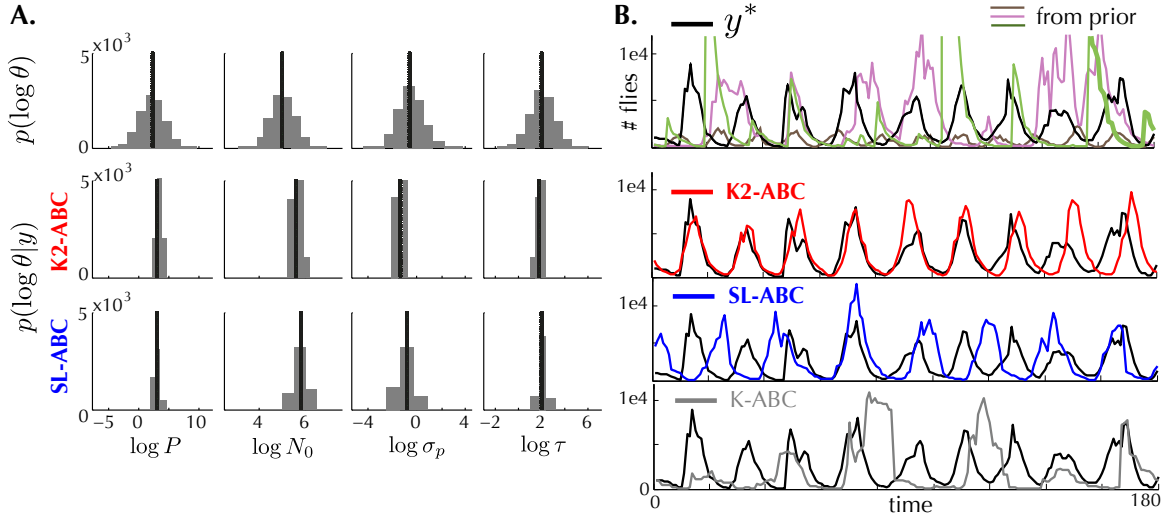


Figure 2: Blowfly data. **A (top)**: Histograms of 10,000 samples for four parameters  $\{\log P, \log N_0, \log \sigma_p, \log \tau\}$  drawn from the prior. **A (middle/bottom)**: Histogram of samples from the posteriors obtained by K2-ABC / SL-ABC (acceptance rate: 0.2, burn-in: 5000 iterations), respectively. In both cases, the posteriors over parameters are concentrated around their means (black bar). The posterior means of  $P$  and  $\tau$  obtained from K2-ABC are close to those obtained from SL-ABC, while there is noticeable difference in the means of  $N_0$  and  $\sigma_p$ . Note that we were not able to show the same histogram for K-ABC since the posterior obtained by K-ABC is improper. **B (top)**: Three realisations of  $y$  given three different parameters drawn from the prior. Small changes in  $\theta$  drastically change  $y$ . **B (middle to bottom)**: Simulated data using inferred parameters (posterior means) shown in A. Our method (in red) produces the most similar dynamic trajectory to the actual observation (in black) among all the methods.

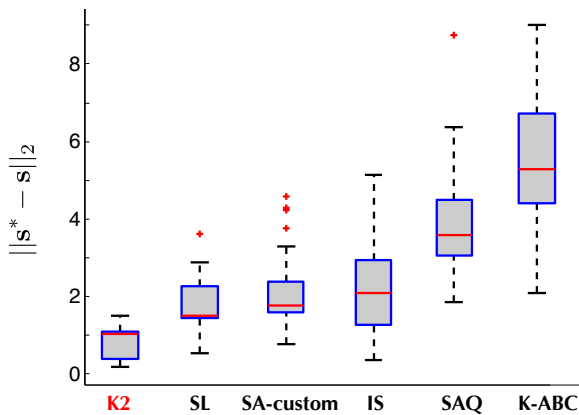


Figure 3: Euclidean distance between the chosen 10 summary statistics of  $y^*$  and  $y$  given the posterior mean of parameters estimated by various methods. Due to the fluctuations in  $y$  from the noise realisations  $(\epsilon_t, e_t)$ , we made 100 independent draws of  $y$  and computed the Euclidean distance. Note that K2-ABC achieved nearly 50% lower errors than the next best method, SL-ABC, although SL-ABC, SA-custom, and K-ABC all explicitly operate on the summary statistics in the comparison while K2-ABC does not.

ance. Due to its computational advantage, the linear-time MMD has been successfully applied in large-scale two-sample testing [31] as a test statistic. The original linear-time MMD is given by

$$\begin{aligned} \widehat{\text{MMD}}_l^2(F_x, F_y) &= \frac{2}{n} \sum_{i=1}^{n/2} \left[ k(x^{(2i-1)}, x^{(2i)}) + k(y^{(2i-1)}, y^{(2i)}) \right. \\ &\quad \left. - k(x^{(2i-1)}, y^{(2i)}) - k(x^{(2i)}, y^{(2i-1)}) \right]. \end{aligned}$$

Note that we have assumed the same number of observations  $n_x = n_y = n$  from  $F_x$  and  $F_y$ . The estimator  $\widehat{\text{MMD}}_l^2$  is constructed so that the independence of the summands allows derivation of its asymptotic distribution and the corresponding quantile computation needed for two-sample testing. However, since we do not require such independence, we employ a linear-time estimator with a larger number of summands, which also does not require  $n_x = n_y$ . Without loss of generality, we assume  $n_x \leq n_y$ . Denote  $x^{(j)} := x^{(1+\text{mod}(j-1, n_x))}$  for  $j > n_x$ , i.e., we allow a cyclic shift through the smaller dataset  $\{x^{(i)}\}_{i=1}^{n_x}$ . The

linear-time MMD estimator that we propose is

$$\widehat{\text{MMD}}_L^2(F_x, F_y) = \frac{1}{n_x - 1} \sum_{i=1}^{n_x-1} k(x^{(i)}, x^{(i+1)}) + \frac{1}{n_y - 1} \sum_{i=1}^{n_y-1} k(y^{(i)}, y^{(i+1)}) - \frac{2}{n_y} \sum_{i=1}^{n_y} k(x^{(i)}, y^{(i)}),$$

which is an unbiased estimator of  $\text{MMD}^2(F_x, F_y)$ . The total cost of the resulting K2-ABC algorithm is  $O(M(n_x + n_y))$ .

**MMD with random Fourier features** Another fast linear MMD estimator can be achieved by considering an approximation to the kernel function  $k(x, y)$  with an inner product of finite dimensional feature vectors  $\hat{\phi}(x)^\top \hat{\phi}(y)$  where  $\hat{\phi}(x) \in \mathbb{R}^D$  and  $D$  is the number of features. Given the feature map  $\hat{\phi}(\cdot)$  such that,  $k(x, y) \approx \hat{\phi}(x)^\top \hat{\phi}(y)$ ,  $\text{MMD}^2$  can be approximated as

$$\begin{aligned} \text{MMD}_{rf}^2(F_x, F_y) &\approx \mathbb{E}_X \hat{\phi}(X)^\top \mathbb{E}_{X'} \hat{\phi}(X') + \mathbb{E}_Y \hat{\phi}(Y)^\top \mathbb{E}_{Y'} \hat{\phi}(Y') \\ &\quad - 2\mathbb{E}_X \hat{\phi}(X)^\top \mathbb{E}_Y \hat{\phi}(Y) := \|\mathbb{E}_X \hat{\phi}(X) - \mathbb{E}_Y \hat{\phi}(Y)\|_2^2. \end{aligned}$$

A straightforward (biased) estimator is

$$\widehat{\text{MMD}}_{rf}^2(F_x, F_y) = \left\| \frac{1}{n_x} \sum_{i=1}^{n_x} \hat{\phi}(x^{(i)}) - \frac{1}{n_y} \sum_{i=1}^{n_y} \hat{\phi}(y^{(i)}) \right\|_2^2,$$

which can be computed in  $O(D(n_x + n_y))$ , i.e., linear in the sample size, leading to the overall cost of  $O(MD(n_x + n_y))$ .

Given a kernel  $k$ , there are a number of ways to obtain  $\hat{\phi}(\cdot)$  such that  $k(x, y) \approx \hat{\phi}(x)^\top \hat{\phi}(y)$ . One approach which became popular in recent years is based on random Fourier features [32] which can be applied to any translation invariant kernel. Assume that  $k$  is translation invariant i.e.,  $k(x, y) = \tilde{k}(x - y)$  for some function  $\tilde{k}$ . According to Bochner’s theorem [33],  $\tilde{k}$  can be written as

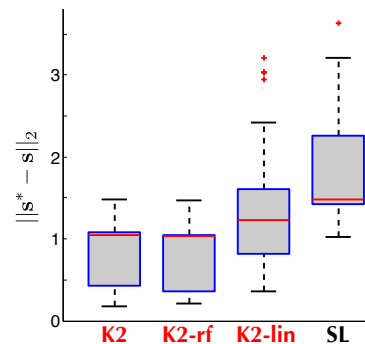
$$\begin{aligned} \tilde{k}(x - y) &= \int e^{i\omega^\top(x-y)} d\Lambda(\omega) = \mathbb{E}_{\omega \sim \Lambda} \cos(\omega^\top(x - y)) \\ &= 2\mathbb{E}_{b \sim U[0, 2\pi]} \mathbb{E}_{\omega \sim \Lambda} \cos(\omega^\top x + b) \cos(\omega^\top y + b), \end{aligned}$$

where  $i = \sqrt{-1}$  and due to positive-definiteness of  $\tilde{k}$ , its Fourier transform  $\Lambda$  is nonnegative and can be treated as a probability measure. By drawing random frequencies  $\{\omega_i\}_{i=1}^D \sim \Lambda$  and  $\{b_i\}_{i=1}^D \sim U[0, 2\pi]$ ,  $\tilde{k}(x - y)$  can be approximated with a Monte Carlo average. It follows that  $\hat{\phi}_j(x) = \sqrt{2/D} \cos(\omega_j^\top x + b_j)$  and  $\hat{\phi}(x) = (\hat{\phi}_1(x), \dots, \hat{\phi}_D(x))^\top$ . Note that a Gaussian kernel  $k$  corresponds to normal distribution  $\Lambda$ .

**Empirical results** We employ the linear-time and the random Fourier feature MMD estimators in our

K2-ABC algorithm, which we call *K2-lin* and *K2-rf*, respectively, and test these variants on the blowfly data. For K2-rf, we used 50 random features.

Figure 4: K2-ABC with different MMD estimators outperform the best existing method, SL-ABC, on the blowfly data.



## 6 CONCLUSIONS

We investigated the feasibility of using MMD as a discrepancy measure of samples from two distributions in the context of ABC. Via embeddings of empirical data distributions into an RKHS, we effectively take into account infinitely many implicit features of these distributions as summary statistics. When tested on both simulated and real-world datasets, our approach obtained more accurate posteriors, compared to other methods that rely on hand-crafted summary statistics.

While any choice of a characteristic kernel will guarantee infinitely many features and no information loss due to the use of insufficient statistics, we note that the kernel choice is nonetheless important for MMD estimation and therefore also for the efficiency of the proposed K2-ABC algorithm. As widely studied in the RKHS literature, the choice should be made to best capture characteristics of given data, i.e., by utilising domain-specific knowledge. For instance, when some data components are believed *a priori* to be on different scales, one can adopt the automatic relevance determination (ARD) kernel instead of the Gaussian kernel. Formulating explicit efficiency criteria in the context of ABC and optimizing over kernel choice, similarly as in the context of two-sample testing [31], would be an essential extension.

## Acknowledgments

We thank Arthur Gretton and Ted Meeds for their helpful comments. M Park and W Jitkrittum were funded by the Gatsby Charitable Foundation. M Park is grateful for the travel support by QUvA lab at University of Amsterdam.



## References

- [1] S. Tavaré, D. J. Balding, R. C. Griffiths, and P. Donnelly. Inferring coalescence times from DNA sequence data. *Genetics*, 145(2):505–518, 1997.
- [2] O. Ratmann, O. Jørgensen, T. Hinkley, M. Stumpf, S. Richardson, and C. Wiuf. Using likelihood-free inference to compare evolutionary dynamics of the protein networks of *H. pylori* and *P. falciparum*. *PLoS Computational Biology*, 3(11):e230, 11 2007.
- [3] E. Bazin, K. J. Dawson, and M. A. Beaumont. Likelihood-free inference of population structure and local adaptation in a bayesian hierarchical model. *Genetics*, 185(2):587–602, 06 2010.
- [4] C. M. Schafer and P. E. Freeman. Likelihood-free inference in cosmology: Potential for the estimation of luminosity functions. In *Statistical Challenges in Modern Astronomy V*, pages 3–19. Springer, 2012.
- [5] P. Joyce and P. Marjoram. Approximately sufficient statistics and bayesian computation. *Stat. Appl. Genet. Molec. Biol.*, 7(1):1544–6115, 2008.
- [6] C. P. Robert, J. Cornuet, J. Marin, and N. S. Pillai. Lack of confidence in approximate Bayesian computation model choice. *Proceedings of the National Academy of Sciences*, 108(37):15112–15117, 2011.
- [7] M. Nunes and D. Balding. On optimal selection of summary statistics for approximate bayesian computation. *Stat. Appl. Genet. Molec. Biol.*, 9(1):doi:10.2202/1544–6115.1576, 2010.
- [8] P. Fearnhead and D. Prangle. Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation. *J. R. Stat. Soc. Series B*, 74(3):419–474, 2012.
- [9] S. Aeschbacher, M. A. Beaumont, and A. Futschik. A Novel Approach for Choosing Summary Statistics in Approximate Bayesian Computation. *Genetics*, 192(3):1027–1047, 2012.
- [10] C.C. Drovandi, A.N. Pettitt, and A. Lee. Bayesian indirect inference using a parametric auxiliary model. *Statist. Sci.*, 30(1):72–95, 02 2015.
- [11] C.C. Drovandi, A.N. Pettitt, and M.J. Faddy. Approximate Bayesian computation using indirect inference. *J. R. Stat. Soc. Series C*, 60(3):317–337, 2011.
- [12] A. Gleim and C. Pigorsch. Approximate Bayesian computation with indirect summary statistics. Draft paper: <http://ect-pigorsch.mee.uni-bonn.de/data/research/papers>, 2013.
- [13] Gael M. Martin, Brendan P. M. McCabe, Worapree Maneesoonthorn, and Christian P. Robert. Approximate Bayesian Computation in State Space Models. *arXiv:1409.8363 [math, stat]*, September 2014. arXiv: 1409.8363.
- [14] A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer, 2004.
- [15] A. Smola, A. Gretton, L. Song, and B. Schölkopf. A Hilbert space embedding for distributions. In *ALT*, pages 13–31, 2007.
- [16] B. Sriperumbudur, K. Fukumizu, and G. Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. *J. Mach. Learn. Res.*, 12:2389–2410, 2011.
- [17] A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *J. Mach. Learn. Res.*, 13:723–773, 2012.
- [18] L. Song, K. Fukumizu, and A. Gretton. Kernel embeddings of conditional distributions. *IEEE Signal Process. Mag.*, 30(4):98–111, 2013.
- [19] D. Sejdinovic, H. Strathmann, M.G. Lomeli, C. Andrieu, and A. Gretton. Kernel Adaptive Metropolis-Hastings. In *ICML*, 2014.
- [20] S. N. Wood. Statistical inference for noisy nonlinear ecological dynamic systems. *Nature*, 466(7310):1102–1104, 08 2010.
- [21] S. Nakagome, K. Fukumizu, and S. Mano. Kernel approximate bayesian computation in population genetic inferences. *Stat. Appl. Genet. Molec. Biol.*, 12(6):667–678, 2013.
- [22] S. Grünewälder, G. Lever, A. Gretton, L. Baldassarre, S. Patterson, and M. Pontil. Conditional mean embeddings as regressors. In *ICML*, 2012.
- [23] D. Sejdinovic, B. Sriperumbudur, A. Gretton, and K. Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *Ann. Statist.*, 41(5):2263–2291, 2013.
- [24] A. Christmann and I. Steinwart. Universal kernels on non-standard input spaces. In *NIPS*, pages 406–414, 2010.
- [25] Z. Szabó, A. Gretton, B. Póczos, and B. Sriperumbudur. Two-stage Sampled Learning Theory on Distributions. *ArXiv e-prints:1402.1754*, February 2014.
- [26] M.C. Wu, P. Kraft, M.P. Epstein, D.M. Taylor, S.J. Chanock, D.J. Hunter, and X. Lin. Powerful SNP-Set Analysis for Case-Control Genome-wide

- Association Studies. *The American Journal of Human Genetics*, 86(6):929–942, June 2010.
- [27] S. Li and Y. Cui. Gene-centric genegene interaction: A model-based kernel machine method. *Ann. Appl. Stat.*, 6(3):1134–1161, September 2012.
- [28] T. Gärtner. A survey of kernels for structured data. *SIGKDD Explor. Newsl.*, 5(1):49–58, July 2003.
- [29] B. Schölkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA, 2001.
- [30] E. Meeds and M. Welling. GPS-ABC: Gaussian Process Surrogate Approximate Bayesian Computation. In *UAI*, volume 30, pages 593–601, 2014.
- [31] A. Gretton, , B.K. Sriperumbudur, D. Sejdinovic, H. Strathmann, S. Balakrishnan, M. Pontil, and K. Fukumizu. Optimal kernel choice for large-scale two-sample tests. In *NIPS*, pages 1205–1213. 2012.
- [32] A. Rahimi and B. Recht. Random features for large-scale kernel machines. In *NIPS*, pages 1177–1184, 2007.
- [33] Walter Rudin. *Fourier Analysis on Groups: Interscience Tracts in Pure and Applied Mathematics, No. 12*. Literary Licensing, LLC, 2013.