DR-ABC: Approximate Bayesian Computation with Kernel-Based Distribution Regression

Jovana Mitrovic MITROVIC@STATS.OX.AC.UK
Dino Sejdinovic DINO.SEJDINOVIC@STATS.OX.AC.UK
Yee Whye Teh Y.W.TEH@STATS.OX.AC.UK
Department of Statistics, University of Oxford

Abstract

Performing exact posterior inference in complex generative models is often difficult or impossible due to an expensive to evaluate or intractable likelihood function. Approximate Bayesian computation (ABC) is an inference framework that constructs an approximation to the true likelihood based on the similarity between the observed and simulated data as measured by a predefined set of summary statistics. Although the choice of informative problem-specific summary statistics crucially influences the quality of the likelihood approximation and hence also the quality of the posterior sample in ABC, there are only few principled general-purpose approaches to the selection or construction of such summary statistics. In this paper, we develop a novel framework for solving this problem. We model the functional relationship between the data and the optimal choice (with respect to a loss function) of summary statistics using kernel-based distribution regression. Furthermore, we extend our approach to incorporate kernel-based regression from conditional distributions, thus appropriately taking into account the specific structure of the posited generative model. We show that our approach can be implemented in a computationally and statistically efficient way using the random Fourier features framework for large-scale kernel learning. In addition to that, our framework outperforms related methods by a large margin on toy and real-world data, including hierarchical and time series models.

Proceedings of the 33^{rd} International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48.

1. Introduction

Complex generative models arise in many application domains, e.g. when we are interested in modeling population dynamics in ecology (Wood, 2010; Lopes & Boessenkool, 2010), performing phylogenetic inference and disease dynamics modeling in epidemiology (**?**Tanaka et al., 2006) or modeling galaxy demographics in cosmology (Weyant et al., 2013; Cameron & Pettitt, 2012). In these models, it is often difficult or impossible to perform exact posterior inference due to an expensive to evaluate or intractable likelihood function. Approximate Bayesian computation (ABC) (Beaumont et al., 2002) is an inference framework that approximates the true likelihood based on the similarity between the observed and simulated data as measured by a predefined set of summary statistics. Unless the chosen summary statistics are sufficient, there is an information loss associated with the projection of the data onto the lower-dimensional subspace of the summary statistics. This results in an approximation bias in the likelihood and subsequently in the posterior sample that is difficult to estimate. More precisely, this information loss implies that ABC performs inference on the partial posterior of the model parameters given the summary statistics of the observed data $p(\theta|s(y^*))$ in lieu of doing it on the full posterior $p(\theta|y^*)$. Thus, the choice of informative problem-specific summary statistics is of crucial importance for the quality of posterior inference in ABC.

Several methods exist in the literature for the selection or construction of summary statistics. A number of these methods can be assembled around the idea of constructing summary statistics by linear or non-linear regression from the full dataset or a set of candidate statistics. In addition to considerations about the sufficiency of the derived summary statistics, all of these methods require either expert knowledge for the selection of the set of candidate statistics, e.g. Nakagome et al. (2013), or perform complex and high-dimensional regression by using the full dataset, e.g. Fearnhead & Prangle (2012).

In this paper, we develop a novel framework for the construction of informative problem-specific summary statistics. Following the approach of Fearnhead & Prangle (2012), we want to derive summary statistics that will allow inference about certain parameters of interest to be as accurate as possible. Thus, we study loss functions and reason about the optimality of summary statistics in terms of minimizing specific instances of these functions. In particular, we model the functional relationship between the data and the optimal choice of summary statistics using kernel-based distribution regression (Szabó et al., 2015). In order to properly account for the nature of the data, we take a two-step approach to distribution regression. Furthermore, we extend our method to incorporate kernel-based regression from conditional distributions. This allows us to efficiently encode the structure of the generative model as part of our method. Thus, we are able to derive more informative optimal summary statistics for problems exhibiting non-trivial dependence structure, e.g. hierarchical or spatio-temporal dependence. In summary, we are able to automatically take into account the diverse structural properties of real-world data without requiring domain-specific knowledge.

In the first variant of our framework, we assume that all aspects of the data are important for estimating the parameters of interest in ABC, i.e. we model the full marginal distribution of the data and regress from it into the space of optimal summary statistics. First, we embed the empirical distributions of newly simulated datasets via the mean embedding (Smola et al., 2007) into a reproducing kernel Hilbert space (RKHS) (Aronszajn, 1950). For this embedding, we choose a characteristic kernel (Sriperumbudur et al., 2011) to ensure that no information from the data is lost. We then regress from these embeddings to the optimal choice of summary statistics with kernel ridge regression (Friedman et al., 2001). The space of candidate regressors is thus another RKHS of functions whose domain is the space of mean-embedded data distributions. For the construction of this RKHS, one can use a simple linear kernel or more flexible kernels defined on distributions such as those described in Christmann & Steinwart (2010). The learned regression function can then be used as the summary statistics within ABC.

For the second variant of our framework, we assume that only certain aspects of the data have a direct influence on the parameter of interest in ABC and thus we restrict our attention to modeling the functional relationship between these aspects of the data and the optimal summary statistics. In particular, we assume that the observed data can be decomposed into *important* and *auxiliary* components such that the parameter of interest depends on the *auxiliary* components of the data only through the family of induced conditional distributions of the *important* data components given the *auxiliary* ones. In order to model the functional relationship between conditional distributions and the optimal summary statistics, we embed these distributions with a conditional embedding operator (Song et al., 2013) into an RKHS and use kernel ridge regression to regress from the space of conditional embedding operators into the space of optimal summary statistics. The space of candidate regressors is thus another RKHS defined on the space of bounded linear operators between RKHS defined on the auxiliary and important data components, respectively. For the construction of this RKHS, one can use any positive definite kernel defined on the space of Hilbert-Schmidt operators, e.g. a simple linear kernel or any kernel given in terms of the Hilbert-Schmidt operator norm. As before, the learned regression function can be used as the summary statistics within ABC.

In this paper, we specifically study the choice of summary statistics and use a simple rejection sampling mechanism. While more complex sampling mechanisms are possible, we take this particular approach in order to decouple the influence of these two important components of ABC on the quality of posterior inference. The rest of the paper is organized as follows. Section 2 gives an overview of related work, while section 3 introduces and discusses our framework. Experimental results on toy and real-world problems, and a comparison with related methods are given in section 4. Section 5 concludes.

2. Related Work

Most existing methods for the selection or construction of informative summary statistics can be grouped into three categories. A first category assembles methods that first perform *best subset selection* in a set of candidate statistics according to various information-theoretic criteria and then use this subset as the summary statistics within ABC. In particular, optimal subsets are selected according to, e.g. a measure of sufficiency (Joyce & Marjoram, 2008), entropy (Nunes & Balding, 2010) and AIC/BIC (Blum et al., 2013).

A second category consists of methods that construct summary statistics from auxiliary models. An example of this approach is indirect score ABC (Gleim & Pigorsch, 2013). Here, a score vector that is calculated based on the partial derivatives of the auxiliary likelihood plays the role of the summary statistics. Motivated by the fact that the score of the observed data is zero when the auxiliary model parameters are set by maximum-likelihood estimation (MLE), the method searches the parameter space for values whose simulated data produce a score close to zero under the same auxiliary model parameters. Thus, the discrepancy measure between the observed and simulated data is defined in terms of the score of the simulated data at the parameter values estimated with MLE from the observed data. A de-

tailed review of this class of approaches can be found in Drovandi et al. (2015).

A third, and last, category is comprised of methods that construct summary statistics using regression from either the full dataset or a set of candidate statistics. Examples of this approach include Blum & François (2010), Boulesteix & Strimmer (2007) and Wegmann et al. (2009) with Aeschbacher et al. (2012) providing a general overview of such methods. Here, we discuss semi-automatic ABC (SA-ABC) (Fearnhead & Prangle, 2012) in more detail. SA-ABC focuses on the construction of summary statistics that will allow inference about certain parameters of interest to be as accurate as possible with respect to specific loss functions. They show that the true posterior mean of the model parameters is the optimal choice of summary statistics under the quadratic loss function. As this quantity cannot be analytically calculated, they estimate it by fitting a regression model from simulated data. In particular, given simulated data $\{(\theta_i, y_i)\}_i$, a linear model $\theta = \beta g(y) + \epsilon$ is fitted; here, $g(\cdot)$ is taken to be either the identity function or a power function. The resulting estimates $s(y) = \hat{\beta} g(y)$ are used as the summary statistics in ABC.

In the literature, there are a few other methods that are not strictly aligned with the above categorization. Here, we review three such methods – synthetic likelihood ABC (Wood, 2010), K-ABC (Nakagome et al., 2013) and K2-ABC (Park et al., 2015). Synthetic likelihood ABC (Wood, 2010) assumes that the summary statistics follow a multivariate normal distribution and uses plug-in estimates for the mean and covariance parameters. In order to generate posterior samples, the method utilizes MCMC with a synthetic likelihood that is derived by convolving the fitted distribution of the summary statistics with a Gaussian kernel that measures the similarity between the observed and simulated data via the fitted summary statistics. On the other hand, K-ABC (Nakagome et al., 2013) and K2-ABC (Park et al., 2015) use the RKHS framework in connection with ABC, albeit in a different fashion than our method. K-ABC regresses from already chosen summary statistics $s(y)$ to posterior expectations of interest, i.e. it estimates a conditional mean embedding operator mapping from $s(y)$ to the corresponding model parameters θ. While the use of a kernel on the summary statistics increases their representative power, the method does not mitigate the challenge of selecting summary statistics. A potential solution to this shortcoming could be to choose the whole dataset to regress with, i.e. use $s(y) = y$. This differs from our approach in two ways. First, the choice of an appropriate kernel that can be defined directly on the data is not straightforward. Our approach does not suffer from this shortcoming since we treat datasets as bags of samples. Second, instead of performing regression to estimate posterior expectations, we utilize it to calculate summary statistics that can be used within ABC. This decouples the regression model from the actual ABC method and thus, does not limit the number of samples that can be used within ABC, i.e. it allows for an arbitrary large number of samples to be drawn after performing the regression step. The K-ABC method has recently been used in HIV epidemiology (**?**). On the other hand, K2-ABC embeds the empirical data distributions into an RKHS via the mean embedding and uses a dissimilarity measure on the space of these embeddings to assess the similarity between the observed and simulated data. In particular, the maximum mean discrepancy (Gretton et al., 2012) is used as the dissimilarity measure on the space of the mean-embedded data distributions, and an exponential smoothing kernel is utilized to compute the ABC posterior sample weights. In contrast to the methods discussed above, there is no explicit construction or selection of summary statistics, but rather the summary statistics are given implicitly as the mean embeddings of the empirical data distributions into an possibly infinite-dimensional RKHS. Our framework is different from this method in that it performs an additional step and regresses from the mean embeddings to the space of summary statistics that are optimal with respect to a user-specified loss function.

3. DR-ABC Method

In this section, we introduce and discuss the novel framework of ABC with kernel-based distribution regression (DR-ABC) and review some of its important building blocks.

MMD. Given a probability distribution F_A defined on a non-empty set $\mathcal{A}$, the mean embedding of F_A, $\mu_{F_A} = \mathbb{E}_{A\sim F_A} k(\cdot, A)$, is an element of the RKHS $\mathcal{H}_k$ defined by the kernel $k : \mathcal{A} \times \mathcal{A} \to \mathbb{R}$. For two probability distributions F_A and F_B, the maximum mean discrepancy (MMD) between F_A and F_B is defined as

$$\begin{aligned}\mathrm{MMD}^2(F_A, F_B) &= ||\mu_{F_A} - \mu_{F_B}||^2_{\mathcal{H}_k} \\ &= \mathbb{E}_A \mathbb{E}_{A'} k(A, A') + \mathbb{E}_B \mathbb{E}_{B'} k(B, B') \\ &\quad - 2\mathbb{E}_A \mathbb{E}_B k(A, B)\end{aligned}$$

with $A, A' \overset{i.i.d.}{\sim} F_A$ and $B, B' \overset{i.i.d.}{\sim} F_B$. Given samples $\{a_i\}_{i=1}^{n_A} \overset{i.i.d.}{\sim} F_A$ and $\{b_j\}_{j=1}^{n_B} \overset{i.i.d.}{\sim} F_B$, an unbiased estimate of the MMD can be computed as

$$\widehat{\mathrm{MMD}}^2(F_A, F_B) = \frac{1}{n_A(n_A-1)} \sum_{i=1}^{n_A} \sum_{i' \neq i} k(a_i, a_{i'}) +$$

$$\frac{1}{n_B(n_B-1)} \sum_{j=1}^{n_B} \sum_{j' \neq j} k(b_j, b_{j'}) - \frac{2}{n_A n_B} \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} k(a_i, b_j).$$

Distribution Regression. The goal of distribution regression is to establish a functional relationship between prob-

ability distributions over a given set and real-valued (possibly multidimensional) responses. In particular, given data $\{(\theta_l, P_l)\}_{l=1}^L$ drawn i.i.d. from an unknown meta distribution $\mathcal{M}$ defined on the product space of responses and probability distributions over the space of observations, we are interested in capturing this data-generating mechanism with a regression model and predicting new responses θ_{L+1} given new distributions P_{L+1}.

In this setting, one major challenge arises due to the fact that the probability distributions $\{P_l\}_{l=1}^L$ are not observed directly, but are available only in terms of their i.i.d. samples. In particular, the data is given as $\{(\theta_l, \{y_l^{(n)}\}_{n=1}^{N_l})\}_{l=1}^L$ with $y_l^{(1)}, \dots, y_l^{(N_l)} \overset{i.i.d.}{\sim} P_l$ and $\mathcal{Y}$ the underlying sample space. Thus, one is interested in predicting new responses θ_{L+1} given a new bag of samples $\{y_{L+1}^{(n)}\}_{n=1}^{N_{L+1}} \overset{i.i.d.}{\sim} P_{L+1}$. This particularity makes regressing directly from the space of probability distributions $\mathcal{M}_1^+(\mathcal{Y})$ to the response space $\Theta \subset \mathbb{R}^D$ difficult as one has to capture the two-stage sampled nature of the data in one functional relationship. If we take the kernel ridge regression approach, then the functional relationship between P and θ is modeled as an element g from the RKHS $\mathcal{G} = \mathcal{G}(k_\mathcal{G})$ of functions mapping from $\mathcal{M}_1^+(\mathcal{Y})$ to Θ with the kernel $k_\mathcal{G}$ defined on $\mathcal{M}_1^+(\mathcal{Y})$. In order to properly account for the two-stage sampled nature of the data, we take a two-stage approach to distribution regression. First, a distribution $P \in \mathcal{M}_1^+(\mathcal{Y})$ is mapped via the mean embedding μ into the RKHS $\mathcal{H}_k$ defined by the kernel $k : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$. Second, this result is composed with an element h from the RKHS $\mathcal{H}_K$ defined by the kernel $K : Y \times Y \to \mathbb{R}$, where Y is the image of $\mathcal{M}_1^+(\mathcal{Y})$ under the mean embedding. This yields $k_\mathcal{G}(P, P') = K(\mu_P, \mu_{P'})$ and $g = h \circ \mu$ with $h : Y \to \mathbb{R}^D$ such that $h(\cdot) = (h_1(\cdot), \dots, h_D(\cdot))$ and $h_d \in \mathcal{H}_K$ for every $d \in \{1, \dots, D\}$, i.e. we treat every dimension of θ separately. Taking the classical regularization approach, the solution of kernel ridge regression can be calculated as

$$h_d^\lambda = \operatorname*{arg\,min}_{h_d \in \mathcal{H}_K} \frac{1}{L} \sum_{l=1}^{L} \left| h_d\left(\mu_{\hat{P}_l}\right) - \theta_{ld} \right|^2 + \lambda ||h_d||_{\mathcal{H}_K}^2$$

with $\hat{P}_l = \frac{1}{N_l} \sum_{n=1}^{N_l} \delta_{y_l^{(n)}}$, $\theta_l = (\theta_{l1}, \dots, \theta_{lD})$, λ the regularization parameter and $h^\lambda = (h_1^\lambda, \dots, h_D^\lambda)$. Given a new $P_{L+1} \in \mathcal{M}_1^+(\mathcal{Y})$ in terms of samples $\{y_{L+1}^{(n)}\}_{n=1}^{N_{L+1}} \overset{i.i.d.}{\sim} P_{L+1}$, a prediction for θ_{L+1} can be calculated in the following way

$$\hat{\theta}_{L+1} = \boldsymbol{\Theta}(\boldsymbol{K} + L\lambda \boldsymbol{I}\boldsymbol{d})^{-1}\boldsymbol{k}, \tag{1}$$

where $\boldsymbol{K}_{ll'} = K(\mu_{\hat{P}_l}, \mu_{\hat{P}_{l'}})$, $\boldsymbol{k}_l = K(\mu_{\hat{P}_l}, \mu_{\hat{P}_{L+1}})$, $\boldsymbol{\Theta} = (\theta_1, \dots, \theta_L)$ and $l, l' \in \{1, \dots, L\}$.

Conditional Distribution Regression. Often only certain aspects of the data are assumed to have a direct influence on the response, e.g. in hierarchical or spatio-temporal settings, and thus one might be interested in modeling only the functional relationship between these aspects of the data and the response. This motivates a decomposition of the data as $y_l^{(n)} = (z_l^{(n)}, x_l^{(n)})$ with $x_l^{(n)}$ encoding the *important* data aspects (for the inference task at hand) and $z_l^{(n)}$ describing the rest of the information that we are not interested in modeling explicitly (e.g. this could correspond to locations on a grid on which the observations are recorded).[1] In other words, we assume that θ_l depends on P_l only through the induced conditionals $\{P_l(\cdot|z_l^{(n)})\}_{n=1}^{N_l}$. Thus, the problem of distribution regression reduces to the question of modeling the functional relationship between the induced family of conditional distributions $\{P(\cdot|z)\}_{z \in \mathcal{Z}} \subset \mathcal{M}_1^+(\mathcal{X})$ and the response θ, i.e. learning a map from the set of functions $\mathcal{T} = \{t : \mathcal{Z} \to \mathcal{M}_1^+(\mathcal{X}),\ t(z) = P(\cdot|z)\}$ into the response space Θ. In order to ensure that all the necessary mathematical objects exist and are well-defined, we make the following assumptions:

1. $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ is a Polish space with $\mathcal{B}(\mathcal{X})$ the associated Borel σ-algebra,
2. $(\mathcal{Z}, \mathscr{Z})$ is a measurable space with $\mathscr{Z}$ the associated σ-algebra,
3. kernel k is bounded and can be factorized as $k((z, x), (z', x')) = k_\mathcal{Z}(z, z')k_\mathcal{X}(x, x')$, and
4. $\mathbb{E}_{X|z}\left[g(X)|z\right] \in \mathcal{H}_{k_\mathcal{X}}$ for all $g \in \mathcal{H}_{k_\mathcal{X}}, z \in \mathcal{Z}$.

In contrast to distribution regression from joint distributions, here we are interested in simultaneously embedding whole families of conditional distributions into an RKHS. To achieve this, we model this embedding as a function

$$\mu_{X|\cdot} : \mathcal{Z} \to \mathcal{H}_{k_\mathcal{X}} \quad \text{with} \quad \mu_{X|Z=z} = \mathbb{E}_{X|Z=z} k_\mathcal{X}(\cdot, X),$$

i.e. for every $z \in \mathcal{Z}$, the embedding of the conditional distribution of X given $Z = z$ is a function in the RKHS $\mathcal{H}_{k_\mathcal{X}}$. Following the approach of Song et al. (2013), we encode the embedding of $\{P(\cdot|z)\}_{z \in \mathcal{Z}}$ with the conditional embedding operator $C_{X|Z}$, where

$$\mu_{X|Z=z} = C_{X|Z} k_\mathcal{Z}(\cdot, z)$$

and

$$C_{X|Z} = C_{XZ} C_{ZZ}^{-1} \in \mathcal{L}(\mathcal{H}_{k_\mathcal{Z}}, \mathcal{H}_{k_\mathcal{X}}).$$

[1] In particular, $\left\{\left(z_l^{(n)}, x_l^{(n)}\right)\right\}_{n=1}^{N_l} \overset{i.i.d.}{\sim} P_l$, $\mathcal{Y} = \mathcal{Z} \times \mathcal{X}$ and $x_l^{(n)} \sim P_l(\cdot|z_l^{(n)})$.

Thus, the embedding of a family of conditional distributions is modeled as an operator between RKHS defined on $\mathcal{Z}$ and $\mathcal{X}$, respectively. Next, we regress from the space of conditional embedding operators, i.e. from the space of bounded linear operators $\mathcal{L}(\mathcal{H}_{k_\mathcal{Z}}, \mathcal{H}_{k_\mathcal{X}})$, into Θ using kernel ridge regression. For this purpose, we define a kernel $K : \mathcal{L}(\mathcal{H}_{k_\mathcal{Z}}, \mathcal{H}_{k_\mathcal{X}}) \times \mathcal{L}(\mathcal{H}_{k_\mathcal{Z}}, \mathcal{H}_{k_\mathcal{X}}) \to \mathbb{R}$ to measure the similarity between different conditional embedding operators. Typical choices for this kernel include the linear kernel $K(C, C') = Tr(CC')$ or any other positive definite kernel given in terms of the Hilbert-Schmidt operator norm. Finally, putting all the building blocks together, the solution of kernel ridge regression can be computed as

$$h_d^\lambda = \underset{h_d \in \mathcal{H}_K}{\arg\min} \frac{1}{L} \sum_{l=1}^{L} \left| h_d\left(\hat{C}_{X|Z}^{(l)}\right) - \theta_{ld} \right|^2 + \lambda_2 ||h_d||_{\mathcal{H}_K}^2,$$

with $\hat{C}_{X|Z}^{(l)} = \boldsymbol{k}_X^{(l)} \left(\boldsymbol{k}_{ZZ}^{(l)} + \lambda_1 Id\right)^{-1} \boldsymbol{k}_Z^{(l)}$, $\boldsymbol{k}_X^{(l)} = \left[k_\mathcal{X}(\cdot, x_l^{(1)}), \ldots, k_\mathcal{X}(\cdot, x_l^{(N_l)})\right]$, $\left[\boldsymbol{k}_{ZZ}^{(l)}\right]_{ij} = k_\mathcal{Z}(z_l^{(i)}, z_l^{(j)})$, $\boldsymbol{k}_Z^{(l)} = \left[k_\mathcal{Z}(\cdot, z_l^{(1)}), \ldots, k_\mathcal{Z}(\cdot, z_l^{(N_l)})\right]^T$, $\lambda = (\lambda_1, \lambda_2)$ the regularization parameter and $\theta_l = (\theta_{l1}, \ldots, \theta_{lD})$. Given a new distribution $P_{L+1} \in \mathcal{M}_1^+(\mathcal{Z} \times \mathcal{X})$ in terms of samples $\{z_{L+1}^{(n)}, x_{L+1}^{(n)}\}_{n=1}^{N_{L+1}} \overset{i.i.d.}{\sim} P_{L+1}$ with $x_{L+1}^{(n)} \sim P_{L+1}(\cdot | z_{L+1}^{(n)})$, a prediction for θ_{L+1} can be calculated as

$$\hat{\theta}_{L+1} = \boldsymbol{\Theta}(\boldsymbol{K} + L\lambda_2 \boldsymbol{Id})^{-1} \boldsymbol{k}, \tag{2}$$

where $\boldsymbol{K}_{ll'} = K(\hat{C}_{X|Z}^{(l)}, \hat{C}_{X|Z}^{(l')})$, $\boldsymbol{k}_l = K(\hat{C}_{X|Z}^{(l)}, \hat{C}_{X|Z}^{(L+1)})$, $\boldsymbol{\Theta} = (\theta_1, \ldots, \theta_L)$ and $l, l' \in \{1, \ldots, L\}$.

DR-ABC.

Our framework, ABC with kernel-based distribution regression, provides a novel approach to the construction of informative problem-specific summary statistics. Motivated by Fearnhead & Prangle (2012), we study loss functions and use simulated data to construct approximations to summary statistics that are optimal with respect to these loss functions. While any loss function can be used within our framework,[2] we focus on the quadratic loss function $L(\theta, \theta^*) = (\theta^* - \theta)^2$ with θ^* the true value of the parameter of interest. Given simulated data $\{(\theta_l, \{y_l^{(n)}\}_{n=1}^{N_l})\}_{l=1}^{L}$, we regress into the space of optimal summary statistics, i.e. into the parameter space in the case of quadratic loss, with kernel-based distribution regression. As discussed in the previous section, we study two different variants of our framework – *full* and *conditional* DR-ABC – to account for the diverse structural properties present in real-world data.

[2]For loss functions not admitting closed-form solutions for the argument of their minimum, numerical optimization techniques might need to be used.

Full DR-ABC: In this variant of the DR-ABC framework, we assume that all aspects of the data are important for estimating the parameter of interest, i.e. we model the complete marginal distribution of the data and regress from it into the parameter space. In particular, we first embed the empirical distributions of the simulated data via the mean embedding into the RKHS $\mathcal{H}_k$ defined by the kernel $k : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$. Next, we define a second RKHS $\mathcal{H}_K$ via the kernel $K : Y \times Y \to \mathbb{R}$,

$$K(\mu_{\hat{P}_l}, \mu_{\hat{P}_{l'}}) = \exp\left(-\frac{\widehat{\mathrm{MMD}}^2(P_l, P_{l'})}{2\sigma_K^2} \right)$$

with σ_K the kernel bandwidth. This kernel provides a dissimilarity measure on the space of mean embeddings. Third, we perform kernel ridge regression from Y into the parameter space with $\mathcal{H}_K$ as the space of candidate regressors. Finally, for a particular $P_{L+1} \in \mathcal{M}_1^+(\mathcal{Y})$ given in terms of a sample $\{y_{L+1}^{(n)}\}_{n=1}^{N_{L+1}} \overset{i.i.d.}{\sim} P_{L+1}$, the approximated optimal summary statistics is equal to Equation 1, i.e. the value of the fitted distribution regression function evaluated at the empirical distribution of that sample.

Conditional DR-ABC: Unlike *full DR-ABC*, here we assume that only certain aspects of the data have a direct influence on the parameter of interest. Thus, we restrict our attention to modeling the functional relationship between these aspects of the data and the parameter of interest. First, we identify the *important* and *auxiliary* data components, i.e. we decompose the simulated data as $\{y_l^{(n)}\}_{n,l} = \{(z_l^{(n)}, x_l^{(n)})\}_{n,l}$. Second, for every l, we encode the embedding of the induced family of conditional distributions $\{P_l(\cdot | z_l^{(n)})\}_n$ with the conditional embedding operator $C_{X|Z}^{(l)} : \mathcal{H}_{k_\mathcal{Z}} \to \mathcal{H}_{k_\mathcal{X}}$, where $\mathcal{H}_{k_\mathcal{Z}}$ and $\mathcal{H}_{k_\mathcal{X}}$ are RKHS defined on $\mathcal{Z}$ and $\mathcal{X}$, respectively, and $k_\mathcal{Z}$ and $k_\mathcal{X}$ are the corresponding kernels. Third, we define a new RKHS $\mathcal{H}_K$ via the kernel $K : \mathcal{L}(\mathcal{H}_{k_\mathcal{Z}}, \mathcal{H}_{k_\mathcal{X}}) \times \mathcal{L}(\mathcal{H}_{k_\mathcal{Z}}, \mathcal{H}_{k_\mathcal{X}}) \to \mathbb{R}$,

$$K(C, C') = Tr(CC').$$

This kernel defines a dissimilarity measure on the space of conditional embedding operators. Next, we perform kernel ridge regression from this space into the parameter space and use the newly constructed RKHS $\mathcal{H}_K$ as the space of candidate regressors. Finally, the fitted distribution regression function can be used as a summary statistics within ABC; the approximated optimal summary statistics of a new dataset is given by Equation 2.

Application to ABC: Having constructed the optimal summary statistics, we can now perform ABC. First, we sample M points from the prior and generate the corresponding datasets $\{(\theta_m, \{y_m^{(j)}\}_{j=1}^{J_m})\}_{m=1}^{M}$. Depending on the inference task at hand and the structural properties of the data, a splitting of the data might be suitable, i.e. $\{y_m^{(j)}\}_{m,j} =$

$\{(z_m^{(j)}, x_m^{(j)})\}_{m,j}$. In order to assess the similarity between the observed and simulated data, we estimate the optimal summary statistics for each dataset and compare these approximations via a smoothing kernel that defines a dissimilarity measure on the parameter space. In particular, we calculate one of the following

$$\kappa(\hat{P}_m, \hat{P}^*) = \exp\left(-\frac{\left|\left|h^\lambda \circ \mu_{\hat{P}_m} - h^\lambda \circ \mu_{\hat{P}^*}\right|\right|_2^2}{\epsilon}\right)$$

$$\kappa(\hat{P}_m, \hat{P}^*) = \exp\left(-\frac{\left|\left|h^\lambda \circ \hat{C}^{(m)}_{X|Z} - h^\lambda \circ \hat{C}^*_{X|Z}\right|\right|_2^2}{\epsilon}\right)$$

depending on whether we are in the setting of full or conditional DR-ABC, respectively. Here, $\hat{P}^*$ and $\hat{P}_m$, and $\hat{C}^*_{X|Z}$ and $\hat{C}^{(m)}_{X|Z}$ are the empirical data distributions and conditional embedding operators of the observed and newly simulated data, respectively.

Putting together kernel-based distribution regression and ABC as discussed above, the following algorithms summarize the two different variants of the DR-ABC framework.

Algorithm 1 Distribution Regression

Input: prior π and data-generating process P
Output: fitted regression function $h^\lambda \circ \mu$
for $l = 1, \ldots, L$ **do**
  Sample $\theta_l \sim \pi$
  Sample dataset $\{y_l^{(n)}\}_n \sim P(\cdot|\theta_l)$
end for
Fit distribution regression with $\{(\theta_l, \{y_l^{(n)}\}_n)\}_l$

Algorithm 2 Conditional Distribution Regression

Input: prior π and data-generating process P
Output: fitted regression function $h^\lambda \circ C_{X|Z}$
for $l = 1, \ldots, L$ **do**
  Sample $\theta_l \sim \pi$
  Sample dataset $\{y_l^{(n)}\}_n \sim P(\cdot|\theta_l)$
  Split dataset $\{y_l^{(n)}\}_n = \{(z_l^{(n)}, x_l^{(n)})\}_n$
end for
Fit distribution regression from conditionals with $\{(\theta_l, \{(z_l^{(n)}, x_l^{(n)})\}_n)\}_l$

Computational complexity. Assuming that both the observed and simulated datasets are of size N, the cost of computing $\widehat{\mathrm{MMD}}^2$ between two bags of samples or computing $K(\hat{C}^{(l)}_{X|Z}, \hat{C}^{(l')}_{X|Z})$ for any l, l' is $O(N^2)$. Taking L and M as the number of simulated datasets for (conditional) distribution regression and ABC, respectively, the total computational cost of fitting the regression and running ABC is $O(N^2(ML + L^2) + L^3)$ in both *full* and *conditional DR-ABC*. In order to mitigate this large computational cost, we apply the popular large-scale kernel learning framework of random Fourier features (RFF) (Rahimi & Recht, 2007). This framework has successfully been applied in several contexts (Chitta et al., 2012; Huang et al., 2013), extended (Le et al., 2013; Yang et al., 2014) and thoroughly analyzed (Bach, 2015; Sutherland & Schneider, 2015; Sriperumbudur & Szabo, 2015). The context most similar to ours is that of Jitkrittum et al. (2015) where two layers of random Fourier features are applied in connection with distribution regression, albeit in the context of emulating Expectation Propagation messages. Using random Fourier features, we approximate the potentially infinite-dimensional feature maps that figure in the computation of kernel functions with finite-dimensional vectors. This implies that kernel evaluations can be approximated by inner products of these finite-dimensional features. Using f random Fourier features in each layer of approximation, we get a significantly reduced computational cost of $O(Nf(ML + L^2) + f^3)$ for full DR-ABC. For conditional DR-ABC, the computational cost can be reduced to $O(f^2(ML + L^2) + f^3)$. In our experiments, we use $f = 100$ and the following RFF expansion

$$\phi(x) \in \mathbb{R}^f, \quad \phi(x)_{i:(i+1)} = \sqrt{\frac{2}{f}}[\cos(w_i^T x), \sin(w_i^T x)].$$

Due to a result from Bach (2015), a comparatively small number of random Fourier features can be used even for large datasets since the number of random Fourier features needed for good approximations of kernel ridge regression solutions often scales sublinearly with the number of observations. Nevertheless, the dependence of the required num-

Algorithm 3 DR-ABC Algorithm

Input: prior π, data-generating process P, observed data $\{y_i^*\}_i$, soft threshold ϵ
Output: weighted posterior sample $\sum_m w_m \delta_{\theta_m}$
Step 1: Perform Distribution Regression ***or*** Conditional Distribution Regression depending on the nature of the data
Step 2: ABC
for $m = 1, \ldots, M$ **do**
  Sample $\theta_m \sim \pi$
  Sample dataset $\{y_m^{(j)}\}_j \sim P(\cdot|\theta_m)$
  Compute $\tilde{w}_m = \exp\left(-\frac{\left|\left|h^\lambda \circ \mu_{\hat{P}_m} - h^\lambda \circ \mu_{\hat{P}^*}\right|\right|_2^2}{\epsilon}\right)$ ***or***
  $\tilde{w}_m = \exp\left(-\frac{\left|\left|h^\lambda \circ \hat{C}^{(m)}_{X|Z} - h^\lambda \circ \hat{C}^*_{X|Z}\right|\right|_2^2}{\epsilon}\right)$
  depending on the nature of the data
end for
$w_m = \tilde{w}_m / \sum_{k=1}^M \tilde{w}_k$ for $m = 1, \ldots, M$

ber of random Fourier features on the number of datasets and the number of observations within each dataset, particularly in settings such as ours where there are two layers of random Fourier features, is not yet fully understood.

4. Experimental Results

Toy example. The first problem we study is the following Gaussian hierarchical model[3]

$$\theta \sim \mathcal{N}(2,1), \quad z \sim \mathcal{N}(0,2), \quad x|z,\theta \sim \mathcal{N}(\theta z^2, 1).$$

This simple example serves as a proof of concept for our framework. In this model, the parameter of interest is θ, and our goal is to estimate $\mathbb{E}[\theta|y^*]$ with y^* the observed dataset. In our experiments, we compare full and conditional DR-ABC against SA-ABC and K2-ABC. We specifically compare our framework against these methods as they are examples illustrating the regression and RKHS approach to the construction of summary statistics. For the performance metric, we calculate the mean square error (MSE) of the parameter of interest on synthetic data. In particular, we set $\theta^* = 2$ and generate 200 observations given this parameter value as y^*; for every newly simulated dataset, we also draw 200 datapoints. For full DR-ABC, we take the kernels k and K as Gaussian kernels, while for conditional DR-ABC, k is a Gaussian kernel and K is a linear kernel. The hyperparameters in the two DR-ABC methods are set via five-fold cross-validation on appropriately defined grids. For the grids of the different kernel bandwidth parameters, we multiply the respective median heuristics (Reddi et al., 2014) with a set of ten equally spaced points between 10^{-4} and 1000. For λ and ϵ, we choose the grids by exponentiating 10 to the powers given by ten equally spaced points between -4 and 1. In order to account for the randomness in the generative process, we run each of the methods 20 times and display the mean of MSE across the different runs. Figure 1 describes the performance of our chosen methods across different numbers of particles used in the (conditional) distribution regression phase (for full and conditional DR-ABC) and in the ABC phase (for all four methods). In order to achieve comparable results, we use the same number of particles in the regression phase as in the ABC phase for SA-ABC. K2-ABC exhibits a fairly stable reconstruction error for different numbers of ABC particles and outperforms SA-ABC only when relatively small numbers of ABC particles are used. Across the wide spectrum of the number of ABC particles used, we see both conditional and full DR-ABC outperforming K2-ABC by a large margin. While full DR-ABC usually also outperforms SA-ABC, conditional DR-ABC does this consistently by a large margin.

[3]The code for all presented experiments is available at https://github.com/jovana-mitrovic/dr-abc.

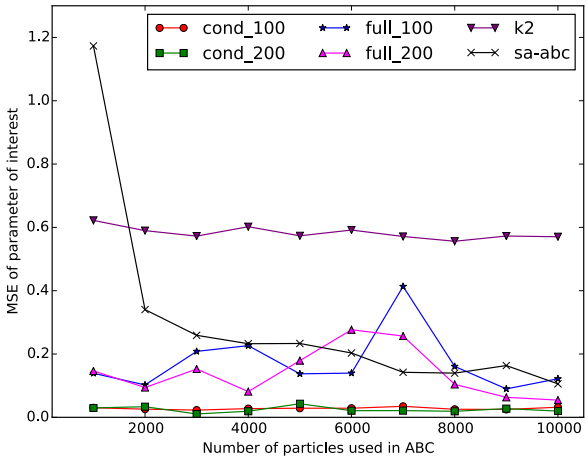

Figure 1. MSE of the parameter of interest for full and conditional DR-ABC, SA-ABC and K2-ABC averaged across 20 runs. The number of ABC particles is between 1000 – 10000. We use either 100 or 200 particles in (conditional) distribution regression.

Ecological Dynamical Systems. Many ecological problems have an intractable likelihood due to a dynamic generative process and thus rely on ABC for posterior inference. Deriving informative summary statistics in this setting is quite challenging as the dependence structure within the data needs to be appropriately taken into account. As an example of an ecological system with a dynamic generative process, we examine the problem of inferring the dynamics of the adult blowfly population as introduced in Wood (2010). In particular, the population dynamics are modeled by the following discretised differential equation

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

with N_{t+1} denoting the observation at time $t+1$ which is determined by the time-lagged observations N_t and $N_{t-\tau}$, and the Gamma distributed noise variables $e_t \sim \text{Gam}(\frac{1}{\sigma_p^2}, \sigma_p^2)$ and $\epsilon_t \sim \text{Gam}(\frac{1}{\sigma_d^2}, \sigma_d^2)$. The parameters of interest in this model are $\theta = \{P, N_0, \sigma_d, \sigma_p, \tau, \delta\}$. As before, we compare our framework with SA-ABC and K2-ABC and use the same performance metric, kernels and hyperparameter selection procedure as in the previous example. For conditional DR-ABC, we take $x^{(n)} = N_n$ and $z^{(n)} = (N_{n-1}, N_{n-1-\tau})$. From Figure 2, we see that our methods outperform SA-ABC by a large margin across the whole spectrum of test situations. Full DR-ABC displays competitive performance to K2-ABC, even outperforming it in certain instances by a large margin. On the other hand, conditional DR-ABC outperforms K2-ABC in all test situations; in some of these situations, the performance of our method is massively superior.

Lotka-Volterra Model. Another popular ecological model in which posterior inference is difficult is the Lotka-Volterra model (Lotka, 1925; Volterra, 1927). This model

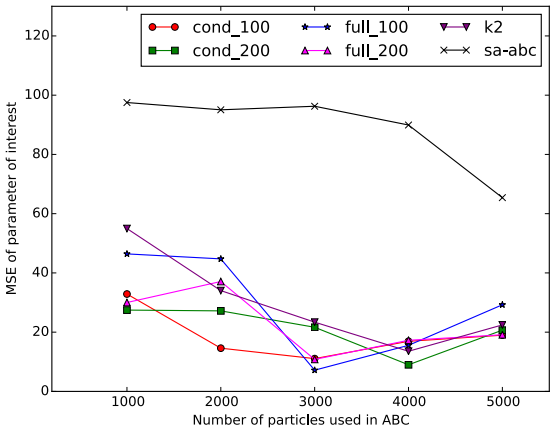

Figure 2. MSE of the parameter of interest for full and conditional DR-ABC, SA-ABC and K2-ABC averaged across 20 runs. The number of ABC particles is between 1000 - 5000. We use either 100 or 200 particles in (conditional) distribution regression.

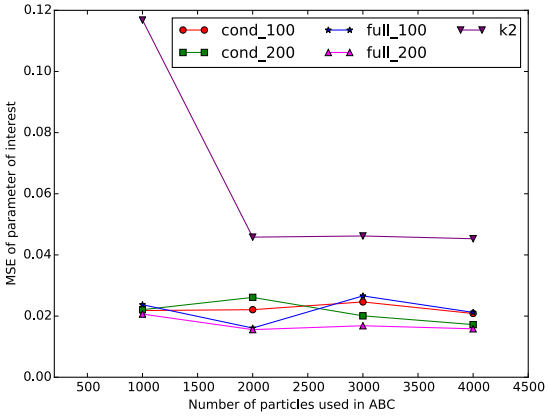

Figure 3. MSE of the parameter of interest for full and conditional DR-ABC and K2-ABC averaged across 20 runs. The number of ABC particles is between 1000 - 4000. We use either 100 or 200 particles in (conditional) distribution regression.

describes the dynamics of biological systems in which two species interact in a predator-prey relationship. The population dynamics are described by the following pair of first-order non-linear differential equations

$$\frac{dx}{dt} = \alpha x - \beta xy, \quad \frac{dy}{dt} = \gamma xy - \delta y, \tag{3}$$

where x, y are the number of prey and predators, respectively, $\alpha, \beta, \gamma, \delta$ are positive real parameters describing the interaction of the two species, t denotes time and $\frac{dx}{dt}, \frac{dy}{dt}$ are the respective growth rates. In addition to the dynamical nature of the generative process, the interaction between the two species makes deriving informative summary statistics even more challenging. The parameters of interest in this model are $\theta = \{\alpha, \beta, \gamma, \delta\}$. For conditional DR-ABC, we condition on the previous states of the model according to 3. As in the previous two experiments, we compare our framework with SA-ABC and K2-ABC and use the same performance metric, kernels and hyperparameter selection procedure.

From Figure 3, we see that our framework outperforms competing methods by a large margin. While for small numbers of ABC particles, full DR-ABC seems to perform better, for large numbers of ABC particles, conditional DR-ABC slightly outperforms full DR-ABC with a clear downward trend in the error for higher numbers of ABC particles. As for SA-ABC, the method cannot directly be applied to this problem due to the high correlation in the observations which leads to a regression problem that is ill-conditioned. Even after performing PCA and using the first 10 principal components for approximation, SA-ABC yielded 1–2 orders of magnitude larger errors than those displayed in the figure and thus, they are excluded from it for clarity.

5. Conclusion

In this paper, we developed a novel framework for the construction of informative problem-specific summary statistics using the flexible and expressive setting of reproducing kernel Hilbert spaces. We introduced two different approaches based on embeddings of probability distributions and kernel-based distribution regression. Our proposed framework has several advantages over previous general-purpose and semi-automatic summary statistics construction methods. First, by using the flexible RKHS framework, we are able to regulate the kind and amount of information that is extracted from the data and thus construct more informative problem-specific summary statistics, as opposed to mandating an ad hoc selection of a limited set of candidate statistics or postulating heuristic summary statistics which inevitably leads to a hard to evaluate approximation bias in the likelihood and subsequently in the posterior sample. Moreover, our framework compactly encodes the extracted information. Second, due to the modeling flexibility of our framework, we are able to appropriately account for different structural properties present in real-world data. Third, our methods can be implemented in a computationally and statistically efficient way using the random Fourier features framework for large-scale kernel learning. Fourth, our framework can be easily extended to any object class on which the embedding kernel(s) can be defined. Examples of such object classes include genetic data (Wu et al., 2010), phylogenetic trees (**?**), strings, graphs and other structured data (Gärtner, 2003). Fifth, although there are multiple sets of hyperparameters in each of our methods, their selection can be performed in a principled way via cross-validation. From experiments on toy and real-world problems, we see that our framework substantially reduces the bias in the posterior sample achieving superior performance when compared to related methods used for the construction of summary statistics in ABC.

Acknowledgements

We gratefully acknowledge the generous financial support of The Clarendon Fund of the University of Oxford without which the present work could not have been completed.

References

Aeschbacher, S., Beaumont, M. A., and Futschik, A. A Novel Approach for Choosing Summary Statistics in Approximate Bayesian Computation. *Genetics*, 192(3): 1027–1047, 2012.

Aronszajn, Nachman. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68 (3):337–404, 1950.

Bach, F. On the equivalence between quadrature rules and random features. Technical Report HAL-01118276, 2015.

Beaumont, M. A., Zhang, W., and Balding, D. J. Approximate Bayesian Computation in Population Genetics. *Genetics*, 162(4):2025–2035, 2002.

Blum, M. G. B. and François, Olivier. Non-linear regression models for approximate bayesian computation. *Statistics and Computing*, 20(1):63–73, 2010.

Blum, M. G. B., Nunes, M. A., Prangle, D., and Sisson, S. A. A Comparative Review of Dimension Reduction Methods in Approximate Bayesian Computation. *Statist. Sci.*, 28(2):189–208, 05 2013.

Boulesteix, A.-L. and Strimmer, K. Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Briefings in bioinformatics*, 8(1):32–44, 2007.

Cameron, E. and Pettitt, A. N. Approximate bayesian computation for astronomical model analysis: a case study in galaxy demographics and morphological transformation at high redshift. *Monthly Notices of the Royal Astronomical Society*, 425(1):44–65, 2012.

Chitta, R., Jin, R., and Jain, A. K. Efficient kernel clustering using random fourier features. In *2012 IEEE 12th International Conference on Data Mining (ICDM)*, pp. 161–170. IEEE, 2012.

Christmann, A. and Steinwart, I. Universal kernels on non-standard input spaces. In *NIPS*, pp. 406–414, 2010.

Drovandi, C. C., Pettitt, A. N., and Lee, A. Bayesian indirect inference using a parametric auxiliary model. *Statist. Sci.*, 30(1):72–95, 02 2015.

Fearnhead, P. and Prangle, D. Constructing summary statistics for approximate bayesian computation: semi-automatic approximate bayesian computation. *J. R. Stat. Soc. Series B*, 74(3):419–474, 2012.

Friedman, J., Hastie, T., and Tibshirani, R. *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin, 2001.

Gärtner, T. A survey of kernels for structured data. *SIGKDD Explor. Newsl.*, 5(1):49–58, July 2003.

Gleim, A. and Pigorsch, C. Approximate bayesian computation with indirect summary statistics. *Draft paper: http://ect-pigorsch. mee. uni-bonn. de/data/research/papers*, 2013.

Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., and Smola, A. A kernel two-sample test. *J. Mach. Learn. Res.*, 13:723–773, 2012.

Huang, P.-S., Deng, L., Hasegawa-Johnson, M., and He, X. Random features for kernel deep convex network. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3143–3147. IEEE, 2013.

Jitkrittum, W., Gretton, A., Heess, N., Eslami, S. M. A., Lakshminarayanan, B., Sejdinovic, D., and Szabó, Z. Kernel-Based Just-In-Time Learning for Passing Expectation Propagation Messages. In *UAI*, 2015.

Joyce, P. and Marjoram, P. Approximately sufficient statistics and bayesian computation. *Stat. Appl. Genet. Molec. Biol.*, 7(1):1544–6115, 2008.

Le, Q., Sarlós, T., and Smola, A. Fastfood - approximating kernel expansions in loglinear time. In *Proceedings of the International Conference on Machine Learning*, 2013.

Lopes, J. S. and Boessenkool, S. The use of approximate bayesian computation in conservation genetics and its application in a case study on yellow-eyed penguins. *Conservation Genetics*, 11(2):421–433, 2010.

Lotka, A. J. Elements of physical biology. *Williams and Wilkins*, 1925.

Nakagome, S., Fukumizu, K., and Mano, S. Kernel approximate bayesian computation in population genetic inferences. *Stat. Appl. Genet. Molec. Biol.*, 12(6):667–678, 2013.

Nunes, M. and Balding, D. On optimal selection of summary statistics for approximate bayesian computation. *Stat. Appl. Genet. Molec. Biol.*, 9(1):doi:10.2202/1544–6115.1576, 2010.

Park, M., Jitkrittum, W., and Sejdinovic, D. K2-ABC: Approximate Bayesian Computation with Infinite Dimensional Summary Statistics via Kernel Embeddings. *arXiv preprint arXiv:1502.02558*, 2015.

Rahimi, A. and Recht, B. Random features for large-scale kernel machines. In *NIPS*, pp. 1177–1184, 2007.

Reddi, S. J., Ramdas, A., Poczos, B., Singh, A., and Wasserman, L. On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions. *arXiv preprint arXiv:1406.2083*, 2014.

Smola, A., Gretton, A., Song, L., and Schölkopf, B. A Hilbert space embedding for distributions. In *ALT*, pp. 13–31, 2007.

Song, L., Fukumizu, K., and Gretton, A. Kernel embeddings of conditional distributions: A unified kernel framework for nonparametric inference in graphical models. *Signal Processing Magazine, IEEE*, 30(4):98–111, 2013.

Sriperumbudur, B., Fukumizu, K., and Lanckriet, G. Universality, characteristic kernels and RKHS embedding of measures. *J. Mach. Learn. Res.*, 12:2389–2410, 2011.

Sriperumbudur, B. K. and Szabo, Z. Optimal Rates for Random Fourier Features. *arXiv preprint arXiv:1506.02155*, 2015.

Sutherland, D. J. and Schneider, J. On the Error of Random Fourier Features. *arXiv preprint arXiv:1506.02785*, 2015.

Szabó, Z., Gretton, A., Poczos, B., and Sriperumbudur, B. Two-stage Sampled Learning Theory on Distributions. In *AISTATS*, volume 38, pp. 948–957, February 2015.

Tanaka, M. M., Francis, A. R., Luciani, F., and Sisson, S. A. Using approximate bayesian computation to estimate tuberculosis transmission parameters from genotype data. *Genetics*, 173(3):1511–1520, 2006.

Volterra, V. *Variazioni e fluttuazioni del numero d'individui in specie animali conviventi*. C. Ferrari, 1927.

Wegmann, Daniel, Leuenberger, Christoph, and Excoffier, Laurent. Efficient approximate bayesian computation coupled with markov chain monte carlo without likelihood. *Genetics*, 182(4):1207–1218, 2009.

Weyant, A., Schafer, C., and Wood-Vasey, W. M. Likelihood-free cosmological inference with type ia supernovae: approximate bayesian computation for a complete treatment of uncertainty. *The Astrophysical Journal*, 764(2):116, 2013.

Wood, S. N. Statistical inference for noisy nonlinear ecological dynamic systems. *Nature*, 466(7310):1102–1104, 08 2010.

Wu, M.C., Kraft, P., Epstein, M.P., Taylor, D.M., Chanock, S.J., Hunter, D.J., and Lin, X. Powerful SNP-Set Analysis for Case-Control Genome-wide Association Studies. *The American Journal of Human Genetics*, 86(6):929–942, June 2010.

Yang, Z., Smola, A., Song, L., and Wilson, A. A la carte-learning fast kernels. *arXiv preprint arXiv:1412.6493*, 2014.