An automatic statistical inference approach using Hilbert space embeddings and approximate Bayesian computation

Wilson González Vanegas

This dissertation is submitted for the degree of
Master in Electrical Engineering

Maestría en Ingeniería Eléctrica
Universidad Tecnológica de Pereira
Pereira - Risaralda - Colombia

Automatics Research Group March 2019

I dedicate this thesis to my loving parents, Guillermo González and Margarita Vanegas, and my younger brother Sebastián González; your unstoppable love, encouragement, and strength have been the engine to achieve this step...

Acknowledgements

I would like to acknowledge my supervisor Álvaro A. Orozco G. for his support and guidance and my advisor Andrés M. Álvarez M. for his contributions, motivation, patience, and guidance. I am also thankful to my collegues from the Automatics Research Group who have contributed to my personal and proffesional life.

This research was funded by the Department of Science, Technology and Innovation, Colciencias, Colombia, under the project "*Desarrollo de una plataforma para el cálculo de confiabilidad en la operación interdependiente de los sistemas de gas natural y sector eléctrico de Colombia que permita evaluar alternativas de inversión y regulación para optimizar los costos de operación*" with code: 111074558696. I also was supported by Vicerrectoría de Investigaciones, Innovación y Extensión from Universidad Tecnológica de Pereira, under the project with code E6-18-2.

Abstract

In many science and engineering applications, the central aim concerns the estimation of some parameters that describe a system, using a set of samples or observations representing a particular behavior. In most cases, such an estimation is a difficult task due to the noise contained in the samples, demanding a treatment from probabilistic perspectives. In particular, Bayesian estimation is a useful tool for inferring model parameters since it includes prior knowledge to handle the uncertainty, especially in real-life applications. For straightforward systems, it is easy to find and assess an expression for the likelihood function. However, in non-trivial systems, the application of any Bayesian framework is a very challenging task, because the complexity of a model means that the associated likelihood is computationally intractable, or that it is not even possible to determine an analytical formula for the likelihood function.

In this regard, numerical simulation-based techniques have been widely used across the literature since they provide an alternative to the problem of applying Bayesian inference with intractable likelihoods. In particular, Approximate Bayesian Computation (ABC) allows statistical inference without using a likelihood function via an auxiliary model that generates simulations of the system, which are *somehow* compared to the observations. Recent progress in this field has included kernel methods and Hilbert spaces to support the Bayesian inference scheme, using probability density functions over the observed and simulated data. The estimation strategy of these densities is an essential component during the inference process, and the selection of free parameters not only in the density estimation but also in the general ABC-based scheme is an increasing challenging task.

The present project aims to develop an automatic statistical inference approach based on Hilbert space embeddings and kernel methods, taking into account both the relevance of data in the non-parametric density estimation and the matching between parameters and simulations spaces used in ABC techniques. In particular, the central aim is to develop a strategy that favors the flexibility and accuracy for both supervised and unsupervised inference scenarios, looking for an automatic scheme that encodes relevant structures in the data and allows to tuning of the parameters used in ABC automatically.

Contents

List of figures

List of tables

Chapter 1

Introduction

1.1 Towards an automatic approximate Bayesian computation approach

Estimation Theory is considered as one of the most attractive areas within the Inferential Statistics field. In particular, Estimation Theory allows making of predictions about some model parameters that describe a particular system, based on a set of samples that arise from measuring some behavior on the system [45]. In most cases, these measures are corrupted by noise that comes from various sources. Thus, the estimation of the model parameters becomes a difficult task demanding a treatment in probabilistic terms. To deal with such an issue, we have three different perspectives arising from the Estimation Theory point of view: the point estimation method, the interval estimation method, and the Bayesian estimation approach [7]. Concerning the point estimation method, a closed mathematical expression is employed to find a single value for each parameter to be estimated, using some particular criterion, for instance, that the likelihood function is maximized assuming some specific distribution for the data (like in the Maximum Likelihood Estimation method), or that the mean square error is minimized, as is the case of least squares [5]. On the other hand, in the interval estimation method, the aim concerns the computation of a range where the true value of the parameters could be found with an associated probability, that is, there are *confidence intervals* for the value of the model parameters [27]. Finally, from a Bayesian estimation perspective, a probability density function for the model parameter (known as *posterior distribution* or simply *posterior*) stands for the probability that the model parameters could take a particular value [6].

Regarding the Bayesian estimation approach, previous knowledge about the model parameters (expressed through a probability density function known as *prior distribution* or simply *prior*) is required to calculate the posterior via the Bayes' theorem, where a likelihood function states the probability of the observed data under a given statistical model in order to leverage the inclusion of prior knowledge into the posterior distribution [45]. For straightforward models, to find an expression for the likelihood function is a direct task. However, for complex systems, such as those that present high nonlinearity or a stochastic behavior, the model complexity means that there is no analytical formula for the likelihood function or that it is computationally intractable and can not be evaluated in any practical amount of time, standing for a really challenging scenario to perform statistical inference using Bayesian techniques [42].

To deal with the intractability of the likelihood function, numerical simulation-based techniques like Approximate Bayesian Computation (ABC) have been proposed, where an approximated posterior is obtained using an *auxiliary model* (which is a mathematical description of the model) in lieu of a likelihood function. The main idea behind an ABC-based method is to assess the auxiliary model with samples drawn from the prior distribution to compute *simulated data* which is compared with the observed data in the sense of a distance function, leading a set of weights that define an approximation of the posterior of the model parameters [43].

The number of different ABC methods that have been proposed across the literature is so large that it would be difficult to classify them. However, two large groups could be highlighted: the ones that work with statistics of the simulations and observations, and those that use distribution embeddings of the observed and simulated data using kernel functions. In the former, *summary statistics* (sufficient statistics) summarize the information contained in observations and simulations before computing the distance to reduce the computational burden due to the large amount of features and/or observations [22]. In the latter, distributions over the observed and simulated data are embedded into a Reproducing Kernel Hilbert Space (RKHS) via a *reproducing kernel*, where a comparison between probability measures supports the inference procedure [30].

Regarding the Distribution-based ABC scheme, since the probability density functions of the observed and simulated data are unknown in practice (because such data belong to the parameter space rather than a probability space), either parametric or non-parametric approaches can be used to estimate such densities. A parametric

estimation constrains the data to follow a particular family of distribution while a non-parametric estimation stands for more flexible modeling where the density approximation is the result of a data-driven process. The latter is preferred in this research due to the type of models we are dealing with. In this regard, the Parzen window estimator is one of the most popular techniques to approximate probability density functions from data, where a kernel function is utilized to compute the probability of a single sample based on the other ones [37]. However, it commonly assumes that all samples have the same importance, and relevant information is not taken into account; this is not appropriated especially in those applications with high dynamics and stochasticity. Besides, like the number of bins in a histogram, the selection of hyperparameters in the desired kernel function is a crucial step since they determine the shape of the estimated density. In particular, for the Gaussian kernel (widely used for its mathematical properties), to determine the covariance matrix is essential because it states the statistical relationship between features of the data set.

Concerning the Hilbert embedding-based procedure for ABC, statistical inference of the model parameters is performed using a comparison between distributions associated to the observed and simulated data and the inference is not performed directly in the parameter space but in the probability space [30], that is, any information about the parameter space is taken into account to compute the posterior distribution (we refer to this setting as *unsupervised inference*). This *blind* approach could lead to scenarios where the uncertainty modeling via probability density functions is not enough to understand the complexity of the system. In such a case, if additional information about the parameters space at the input of the auxiliary model is available, it could be included the inference procedure to obtain a more accurate estimation (we refer to this setting as *supervised inference*).

The present work aims to develop an automatic statistical inference approach based on Hilbert space embeddings and kernel methods, taking into account both the relevance of data in the non-parametric density estimation and the matching between parameters and simulations spaces used in ABC techniques. In particular, the central aim is to develop a strategy that favors the flexibility and accuracy for both supervised and unsupervised inference scenarios, looking for an automatic scheme that encodes relevant structures in the data and allows the tuning of the parameters used in ABC automatically.

1.2 Aims

1.2.1 General aim

To develop an automatic approximate Bayesian computation approach based on Hilbert space embeddings that allows the automatic selection of free parameters for both supervised and unsupervised statistical inference scenarios.

1.2.2 Specific aims

1. To propose a nonparametric density estimation methodology that highlights relevant data information in order to support the posterior estimations of ABC-based unsupervised inference scenarios.

2. To develop an unsupervised inference approach that allows an automatic free parameter selection in the nonparametric density estimation, in order to improve the accuracy, precision, and flexibility of ABC-based models.

3. To propose a statistical alignment methodology that correlates dependencies between the parameter space and the simulations space used in ABC, in order to support the posterior estimations of ABC-based supervised inference scenarios.

4. To develop a supervised inference approach that includes information about the parameter space and allows the automatic selection of all free parameters, in order to improve the accuracy, precision, and flexibility of ABC-based models.

1.3 Research contributions

Bayesian statistical inference under intractability of the likelihood function is a very challenging task. Different approaches based on Approximate Bayesian Computation have been proposed for leading the lack or expensive assessment of a likelihood function [42, 43, 30]. However, all these approaches rely on the use of free parameters that have to be tuned by the user and essentially affect the accuracy of the inference. Cross-validation or grid search can be used to find proper values for the free parameters; nevertheless, these approaches are often too expensive in computational terms and state problem dependent solutions [45, 7, 6]. On the other hand, in most of the ABC-based statistical inference models, the posterior has to be computed without any information about the relationship between the parameter space and the simulations space. In this

regard, the development of an automatic ABC-based statistical inference approach that can automatically find values for the free parameters based on data including supervised and unsupervised scenarios would be an important step towards more flexible statistical inference tasks, in systems with stochastic and dynamic properties. Furthermore, this project would state an essential contribution to the fields of statistical inference and kernel methods through the development of a methodology that combines Bayesian perspectives and Hilbert embeddings seeking for more robust statistical inference frameworks.

Chapter 2

Background

This chapter provides a brief introduction to the fundamental ideas behind Approximate Bayesian Computation (ABC). It introduces the straightforward ABC method based on rejection and summary statistics and describes the usage of kernel methods and Hilbert embeddings in the context of ABC. A short review about Hilbert spaces and metrics over distributions is also given. The chapter also presents the related work concerning the most significant ABC methods proposed in the state-of-the-art.

2.1 Bayesian inference under intractable likelihoods

From a Bayesian perspective, entire knowledge about a vector of model parameters $\theta \in \Theta$ is completely expressed through the posterior distribution $p(\theta|y)$, where $y \in \mathcal{X}$ stands for the observed data. Therefore, a likelihood function $p(y|\theta)$ updates beliefs about the model parameters, as expressed in a prior distribution $p(\theta)$, according to Bayes' theorem:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\int_{\Theta} p(y|\theta)p(\theta)d\theta}. \tag{2.1}$$

The posterior distribution contains all necessary information for analysis of the model, from predictive inference and model checking to decision making and beyond [38].

In practice, the complexity of a model, especially in those cases that involve nonlinearity or stochasticity, means that the calculation of the posterior as in equation 2.1 can not be performed due to two main reasons: 1) the mathematical expressions for

the likelihood function and/or the prior distribution lead to such complex calculations that there is no way to obtain a posterior in close form and it is just accessible through its samples using a numerical method; 2) the assessment of the likelihood function used to obtain samples from the posterior is computationally intractable and it can not be performed numerically in any amount of time. Under this situation, neither exact nor sampled posterior can be obtained using a Bayesian approach.

To face any of these challenging scenarios, a simple option is to fit a different model that is less complex and more amenable in terms of mathematical and statistical computations. However, this could be performed at the expense of wrong or less realistic conclusions about the phenomenon that is being analyzed. A more elegant alternative concerns an approximation for the desired model maintaining realism under some approximation error. In this sense, several alternatives have emerged as "free-likelihood" techniques and approximate Bayesian computation is a particular case of them.

2.2 ABC fundamentals

Approximate Bayesian computation was originally introduced as a solution for performing statistical inference in the field of molecular biology where systems with high dynamics and stochasticity are typically found. The first ABC algorithm was proposed by authors [33] who studied the demographic history of the Y chromosome. However, the use of ABC techniques has influenced several research areas like systems biology [23], climate analysis [20], ecological modeling [14], and nuclear imaging [13], just to mention some of them. The fundamental idea behind an ABC framework is to replace the calculation of the likelihood function with an idea of how likely it is that the desired model could have produced the observations, using a set of simulated data generated from an *auxiliary model* [43]. These simulations are then compared with the observed data in order to find an approximation of the posterior distribution of the model parameters, that is, instead of finding a value for the model parameters such that a particular function is minimized (as is the case of least squares), the goal of ABC is to estimate the *posterior* distribution of those parameters [11].

The most straightforward ABC approach can be summarized in the skecth shown in figure 2.1. The idea behind this framework, known as *ABC rejection*, is remarkably simple: a set of candidates $\{\theta_n \backsim \zeta(\theta)\}_{n=1}^{N}$, drawn from the desired prior distribution $\zeta(\theta)$,

is employed to assess an auxiliary model $\mathcal{M}{:}\Theta{\to}\mathcal{X}$, which stands for the mathematical description of the system under analysis, in order to generate simulated data $\{x_n{\in}\mathcal{X}\}_{n=1}^{N}$. Therefore, each simulation is compared with the observed data in the sense of a distance $d{:}\mathcal{X}{\times}\mathcal{X}{\to}\mathbb{R}^{+}$ and a small threshold ξ, that is, if the distance is less than ξ, the corresponding prior sample is selected to follow the posterior distribution; otherwise it is rejected [46]. Typically, it is difficult to apply a distance directly on $\mathcal{X}$ due to a large number of samples and features in real data. In such a case, some strategies use a mapping $s{=}\vartheta(x)$ before calculating the distance, where $s{\in}\mathcal{S}$ is a feature space and $\vartheta{:}\mathcal{X}{\to}\mathcal{S}$ [22].

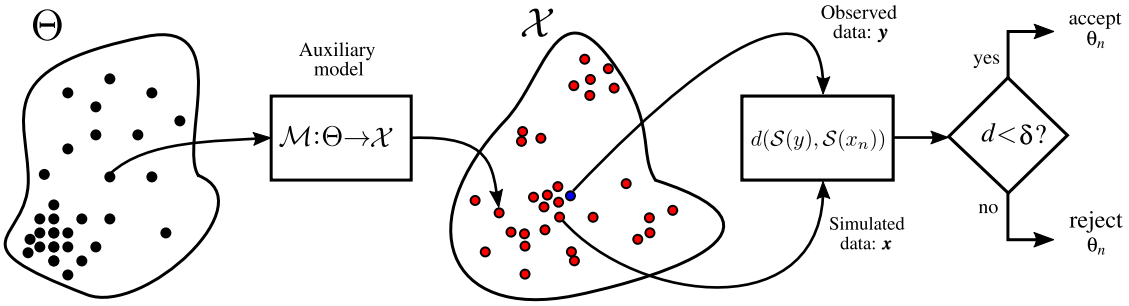


Fig. 2.1 A sketch representing the *ABC rejection* algorithm.

In practice, the simplicity of the previous approach leads to some drawbacks regarding a given inference task. For instance, the selection of a proper distance (according to the problem) to compare observations and simulations, the value of a suitable threshold to accept or reject candidates, the selection of *sufficient summary statistics* to extract information properly from data, among others [46]. As a consequence, more complex algorithms based on sampling methods have been proposed in the last years, e. g., the ABC MCMC algorithm that uses Markov chains and Monte Carlo simulation to find better approximations for the posterior of the model parameters [25]. A more detailed description about these methods can be found in [43, 38] and references therein.

Recently, Park et al. [30] proposed to use Hilbert space embeddings in the context of ABC using kernels functions as an alternative to summary statistics. Figure 2.2 summarizes this novel approach. In particular, both simulated and observed data are considered to follow probability distributions $\{x_n{\backsim}P_{X_n}\}_{n=1}^{N}$ and $y{\backsim}P_Y$, which are embedded into a Reproducing Kernel Hilbert Space (RKHS) of functions $\mathcal{H}$ generated by a characteristic kernel $k(\cdot,\cdot)$ (see section 2.3 for more details). Therefore, a distance between probability distributions $d_{\mathcal{H}}{:}\mathcal{H}{\times}\mathcal{H}{\to}\mathbb{R}^{+}$ supports the computation of a similarity kernel $\kappa(P_{X_n},P_Y)$ that assigns a weight w_n to each prior sample:

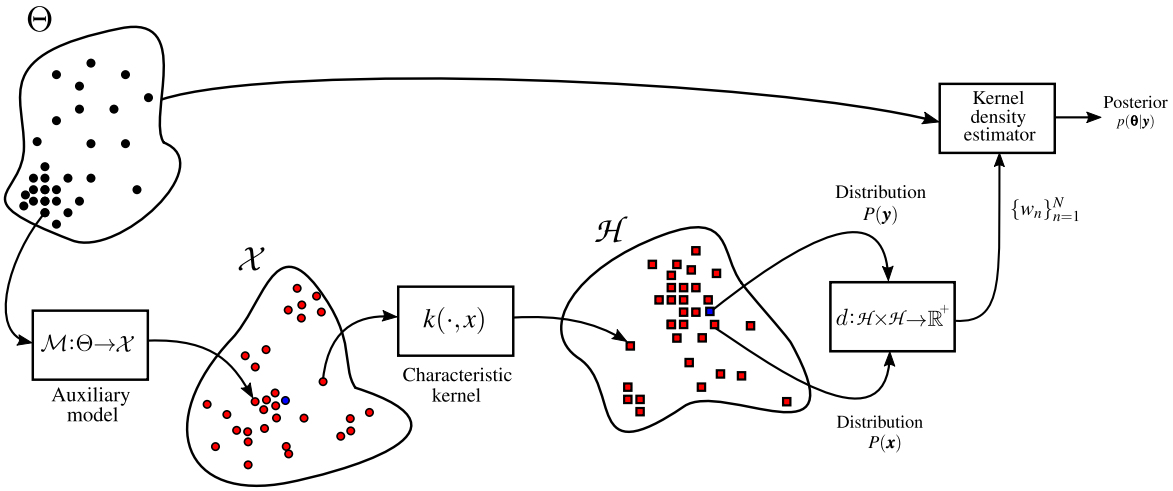


Fig. 2.2 A sketch representing Hilbert embeddings in the context of ABC.

$$w_n = \frac{\kappa(P_{X_n}, P_Y)}{\sum_{n=1}^{N} \kappa(P_{X_n}, P_Y)}. \tag{2.2}$$

A frequently used similarity kernel over probability measures takes the form [30]:

$$\kappa(d^2_{\mathcal{H}}(P_{X_n}, P_Y); \epsilon) = \exp\left(- \frac{d^2_{\mathcal{H}}(P_{X_n}, P_Y)}{\epsilon} \right), \tag{2.3}$$

where $\epsilon \in \mathbb{R}^+$ is a small threshold.

The application of equations 2.2 and 2.3 over simulations results in a weighted sample set $\Psi = \{\theta_n, w_n\}_{n=1}^N$ that can be used to approximate the posterior $p(\theta|y)$ via empirical posterior estimation or a kernel density estimator. Algorithm 1 presents the main steps to perform ABC based on Hilbert embeddings, where $\delta(a-b)$ is the Delta function.

Algorithm 1 ABC based on Hilbert embeddings

Input: Observed data: $y \backsim P_Y$, prior: $\zeta(\theta)$, threshold ϵ, distance parameters.

Output: Posterior estimation: $\hat{p}(\theta|y) = \sum_{i=1}^{N} w_n \delta(\theta - \theta_n)$.

1: **for** $n = 1, \cdots, N$ **do**
2: $\quad \theta_n \backsim \zeta(\theta)$ ▷ Draw a candidate θ_n from the prior.
3: $\quad x_n = \mathcal{M}(\theta_n);\ x_n \backsim P_{X_n}$ ▷ Draw a sample from the model.
4: $\quad \tilde{w}_n = \kappa(d_{\mathcal{H}}(P_{X_n}, P_Y); \epsilon)$ ▷ Compute the n-th weight value.
5: **end for**
6: $w_n = \tilde{w}_n / \sum_{n=1}^{N} \tilde{w}_n$ ▷ Normalize weights

2.3 Hilbert spaces and metrics over distributions

An *inner product space* $\mathbb{H}$ (the largest and most inclusive space of vectors equipped with $\langle \cdot, \cdot \rangle$ as inner product) that has an orthonormal set of basis $\{\boldsymbol{\beta}_k\}_{k=1}^{\infty}$ is known as a *Pre-Hilbert* space [1]. A vector $\boldsymbol{x}$ in $\mathbb{H}$ can be spanned by the basis as a linear combination:

$$\boldsymbol{x} = \sum_{k=1}^{\infty} a_k \boldsymbol{\beta}_k, \tag{2.4}$$

where the scalars a_k are the coefficients of the representation.

The squared norm between two vectors $\boldsymbol{x_n}{=}\sum_{k=1}^{n} a_k \boldsymbol{\beta}_k$ and $\boldsymbol{x_m}{=}\sum_{k=1}^{m} a_k \boldsymbol{\beta}_k$, with $m > n$ in $\mathbb{H}$, can be defined using the inner product as:

$$\begin{aligned} ||\boldsymbol{x_n} - \boldsymbol{x_m}||_2^2 &= \langle \boldsymbol{x_n} - \boldsymbol{x_m}, \boldsymbol{x_n} - \boldsymbol{x_m} \rangle \\ &= \left\langle \sum_{k=n+1}^{m} a_k \boldsymbol{\beta_k} \, , \sum_{k=n+1}^{m} a_k \boldsymbol{\beta_k} \right\rangle \\ &= \sum_{k=n+1}^{m} a_k^2. \end{aligned} \tag{2.5}$$

When the coefficients a_k are defined such that the following conditions are satisfied: 1) $\sum_{k=1}^{n} a_k^2 < \infty$; 2) $\sum_{k=n+1}^{m} a_k^2 \to 0$ as both $n, m \to \infty$; then, a sequence of vectors $\{\boldsymbol{x_k}\}_{k=1}^{\infty}$ so defined is a *Cauchy sequence*. Namely, a vector $\boldsymbol{x}$ can be expressed via the basis $\{\boldsymbol{\beta}_k\}_{k=1}^{\infty}$ if, an only if, $\boldsymbol{x}$ is a linear combination of such basis and the coefficients of the representation are square summable [24].

Definition. *An inner product space* $\mathbb{H}$ *is complete if every Cauchy sequence of vectors selected from the space* $\mathbb{H}$ *converges to a limit in* $\mathbb{H}$*. A complete inner product space is called a* Hilbert space.

Hilbert spaces can be either finite or infinite-dimensional vector spaces. The latter, for instance, are the foundation of continuous-time signal processing [32]. Another useful Hilbert space is the Reproducing Kernel Hilbert Space (RKHS). An RKHS is a special Hilbert space associated with a non-negative definite kernel function κ such that it reproduces (via a inner product) each function f in the space [1]. Let $\mathcal{H}$ be a Hilbert space of real-valued functions defined on a set $\mathcal{X}$, equipped with an inner

product $\langle \cdot\ ,\ \cdot \rangle$ and a real-valued bivariate function $\kappa{:}\mathcal{X}\times\mathcal{X}\to\mathbb{R}^+$. The function $\kappa(x, x')$ is said to be non-negative definite if for any finite set $\{x_n\}_{n=1}^N \subset \mathcal{X}$ and any not all zero real numbers $\{\alpha_n\}_{n=1}^N$, the following condition is satisfied:

$$\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j \kappa(x_i, x_j) \geq 0. \tag{2.6}$$

Theorem (Moore-Aronszajn). *For any non-negative definite function $\kappa(x, x')$, there exists a uniquely determined (and possible infinite-dimensional)* Hilbert *space $\mathcal{H}$ consisting of functions on $\mathcal{X}$ such that:*

(1) $\forall x{\in}\mathcal{X}, \quad \kappa(\cdot, x){\in}\mathcal{H}$

(2) $\forall x{\in}\mathcal{X}, \forall f{\in}\mathcal{H}, \quad f(x){=}\langle f\ ,\ \kappa(\cdot, x)\rangle_{\mathcal{H}}.$

From condition (1) it is easy to infer that each element in the input space is mapped onto a function in the RKHS generated by the selected *reproducing kernel* κ. On the other hand, the condition (2) is known as the reproducing property of $\kappa(x, x')$ in $\mathcal{X}$. Namely, if the non-linear mapping function $\phi{:}\mathcal{X}\to\mathcal{H}$ is defined as $\phi(x){=}\kappa(\cdot, x)$, it follows that

$$\langle \phi(x)\ ,\ \phi(x')\rangle_{\mathcal{H}} = \langle \kappa(\cdot, x)\ ,\ \kappa(\cdot, x')\rangle = \kappa(x, x'). \tag{2.7}$$

As a consequence, $\phi(x){=}\kappa(\cdot, x)$ defines completely and uniquely the RKHS associated with the kernel κ.

A valuable conclusion concerning the last decade of research in statistics states that it is possible to define distances between probability distributions via the expected value and the concept of norm in an RKHS [41]. Let P_X be a probability distribution associated to a random variable $X{\backsim}P_X$. Define the embedding $\mu[\cdot]$ of P_X into the RKHS $\mathcal{H}$ as [40]:

$$\mu[P_X] := \mathbb{E}_X[\kappa(x, \cdot)], \tag{2.8}$$

where $\mathbb{E}_X$ stands for the expected value operator over X and the reproducing kernel (commonly named in this case as *characteristic kernel*) satisfies the sufficient condition $\kappa(x, \cdot) < \infty$.

A distance between two distributions P_X and P_Y is then defined as:

$$d^2_{\mathcal{H}}(P_X, P_Y) = \left\| \mu[P_X] - \mu[P_Y] \right\|^2_{\mathcal{H}} = \Big\langle \mu[P_X] - \mu[P_Y] \, , \, \mu[P_X] - \mu[P_Y] \Big\rangle_{\mathcal{H}} . \tag{2.9}$$

Using the definition of expected value and assigning probability density functions as $f(x)$ and $g(y)$ to P_X and P_Y, respectively, the distance in equation 2.9 is rewritten as:

$$\begin{aligned} d^2_{\mathcal{H}} &= \Big\| \mu[P_X] \Big\|^2_{\mathcal{H}} - 2 \Big\langle \mu[P_X] \, , \, \mu[P_Y] \Big\rangle_{\mathcal{H}} + \Big\| \mu[P_Y] \Big\|^2_{\mathcal{H}} \\ &= \left\langle \int k(x, \cdot) f(x) dx \, , \, \int k(x', \cdot) f(x') dx' \right\rangle_{\mathcal{H}} - 2 \left\langle \int k(x, \cdot) f(x) dx \, , \, \int k(y, \cdot) g(y) dy \right\rangle_{\mathcal{H}} \\ &\qquad + \left\langle \int k(y, \cdot) g(y) dy \, , \, \int k(y', \cdot) g(y') dy' \right\rangle_{\mathcal{H}} \\ &= \int \int \langle k(x, \cdot) \, , \, k(x', \cdot) \rangle_{\mathcal{H}} \, f(x) f(x') dx dx' - 2 \int \int \langle k(x, \cdot) \, , \, k(y, \cdot) \rangle_{\mathcal{H}} \, f(x) g(y) dx dy \\ &\qquad + \int \int \langle k(y, \cdot) \, , \, k(y', \cdot) \rangle_{\mathcal{H}} \, g(x) g(y') dy dy', \end{aligned} \tag{2.10}$$

and using the reproducing property $k(a, b)) = \langle k(a, \cdot) \, , \, k(b, \cdot) \rangle$ (Moore-Aronszajn):

$$\begin{aligned} d^2_{\mathcal{H}}(P_X, P_Y) = &\int \int k(x, x') f(x) f(x') dx dx' - 2 \int \int k(x, y) f(x) g(y) dx dy \\ &+ \int \int k(y, y') g(x) g(y') dy dy'. \end{aligned} \tag{2.11}$$

2.4 Related work

Several ABC methods have been proposed in the last years in response to the increasing need to perform statistical inference in situations that involve intractability of the likelihood function. As an initial attempt, Wood [47] introduced the synthetic likelihood ABC (SL-ABC) where an analytic expression for a synthetic likelihood can be obtained by assuming that the summary statistics of observed and simulated data have multivariate Gaussian distributions that can be compared using a similarity function; the resulting artificial likelihood supports posterior sampling via a Markov chain Monte Carlo (MCMC) approach. Although this is a quite practical idea given the easy access to an MCMC-based algorithm, e.g., the Metropolis-Hastings sampler,

it restricts data to follow a normal-based parametric model that could compromise the inference quality in complex applications [38].

In lieu of using synthetic likelihoods, alternative approaches employ a known conditional probability of the simulated data given a vector of parameters as auxiliary model to perform *indirect inference* in the context of ABC [12, 10]. A clear example is the well-known Indirect Score ABC (IS-ABC) proposed by Gleim and Pigorsch [16], where a score vector concerning the partial derivatives of the log-likelihood with respect to the vector parameters is fixed as summary statistics to select the best candidates values whose corresponding simulated data produce a score close to zero, based on the fact that under a Maximum Likelihood Estimator (MLE) fitted with observed data such score vector becomes exactly zero.

With the fast spreading of kernel methods over several research areas, recent approximate inference techniques have introduced kernel in the context of ABC. The first kernel-based ABC framework, the Kernel-ABC (K-ABC) introduced by Nakagome et al. [29], concerns the estimation of a conditional mean embedding operator mapping from the summary statistics space to the parameter space. Essentially, a regression function represents the embedding of the posterior distribution in the form of a weighted sum of feature maps associated with a kernel over summary statistics that produces an RKHS. Although the use of kernel functions stands for a more general representation in a nonlinear way, the need for proper and sufficient summary statistics in the K-ABC method is still a remarkable drawback. In this regard, the K2-ABC algorithm proposed by Park et al. [30] was introduced as a free-summary statistics approach as explained in section 2.2. In particular, empirical approximations for the distributions associated to the observed and simulated data support the calculation of a distance that only depends on the characteristic kernel in an RKHS when the Maximum Mean Discrepancy (MMD) dissimilarity criterion is employed [18]. As an extension to K2-ABC, authors in [50] used traditional Parzen windows-based estimators for the associated densities using a Gaussian kernel for both, the characteristic kernel and the kernel used to estimate each density. The obtained Parzen-ABC (P-ABC) has a more robust distance in the sense of the number of samples to perform the posterior approximation but at the expense of more free parameters to tune. In practice, none of the mentioned kernel-based approaches take into account the relevance of each sample in the data-sets while computing the distance over distributions, setting an interesting challenge for unsupervised inference scenarios regarding the improvement of distances

in an RKHS for ABC.

Devoted to automatic frameworks for ABC, an initial attempt proposed by Fearnhead and Prangle [15] introduced a Semi-Automatic ABC (SA-ABC) where a linear model under a quadratic loss function leads to an optimal (and so automatic) selection of summary statistics using simulated data; the resulting estimates for such statistics can be used in a standard ABC approach using a euclidean distance as similarity measure between simulated and observed data but with the remaining drawbacks discussed in section 2.2. In contrast to the automatic construction of summary statistics, Prangle [31] focused on adapting the ABC comparison stage via a weighted Euclidean distance designed to work in the most efficient iterative ABC algorithms based on population Monte Carlo [39, 4, 42]. On the other hand, Meeds and Weling [26] developed a surrogate model as synthetic likelihood to define a suitable number of simulations for ABC using a Gaussian process-based framework. Moreover, Mitrovic et al. [28] modeled the functional relationship between simulations and the optimal choice of summary statistics to encode the structure of a generative model using a kernel ridge regression for conditional distributions. However, the techniques mentioned above require the estimation of different free parameters to approximate the posterior. As a consequence, expensive tuning procedures as grid search and cross-validation are carried out. Besides, the user requires a vast knowledge concerning the ABC algorithm and the studied data to properly tune the free parameters, demanding for improved methods concerning the automatic selection of free parameters in the overall posterior estimation process for supervised inference scenarios.

Chapter 3

An improved ABC approach for unsupervised inference scenarios

This chapter presents a general way to highlight relevant information from observations and simulations concerning the nonparametric density estimation procedure for improving posterior estimations in unsupervised inference scenarios. It focuses on the automatic selection of free parameters when approximating such densities to support the comparison between probability measures of observed and simulated data in Hilbert embedding-based ABC. The chapter introduces a general distance to compare distributions in a reproducing kernel Hilbert space and discusses how other distances can be obtained from it as particular cases. It suggests that extracting relevant information from simulations and observations improves the posterior approximation in those unsupervised inference tasks where no information about the model parameter is directly included in the overall ABC procedure.

3.1 Learning from data to compare distributions

As it was explained in section 2.3, a distance between probability distributions can be evaluated in a RKHS using their associated probability density functions. Typically, in a practical ABC inference task, there is no idea about the analytic expressions for such densities and they can be accessed only through a pair of independent and identically distributed sets $X_n=\{x_i\in\mathbb{R}^d\}_{i=1}^{Q_x}$ and $Y=\{y_i\in\mathbb{R}^d\}_{j=1}^{Q_y}$, for simulated and observed data of size Q_x and Q_y, respectively, where $\{X_n\backsim P_{X_n}\}_{n=1}^{N}$ and $Y\backsim P_Y$. In this research, the following general kernel-based density estimators are proposed:

$$\hat{f}_n(x) = \sum_{i=1}^{Q_x} \alpha_i K_{H_i^p}(x, x_i); \quad \sum_{m=1}^{Q_x} \alpha_m = 1, \tag{3.1}$$

$$\hat{g}(y) = \sum_{j=1}^{Q_y} \beta_j K_{H_j^q}(y, y_j); \quad \sum_{m=1}^{Q_y} \beta_m = 1, \tag{3.2}$$

where $\{\alpha_i \in [0,1]\}_{i=1}^{Q_x}$ and $\{\beta_j \in [0,1]\}_{j=1}^{Q_y}$ are representation weights and $\hat{f}_n$ and $\hat{g}$ are the approximated densities associated with P_{X_n} and P_Y, respectively. Moreover, K_H stands for the multivariate Gaussian kernel with covariance matrix $H \in \mathbb{R}^{d \times d}$. Although various kernel functions can be tested, the Gaussian function is preferred since it aims at finding densities with universal approximating ability, not to mention its mathematical tractability [32].

Using a multivariate Gaussian kernel with covariance matrix H_K as characteristic kernel and substituting 3.1 and 3.2 in equation 2.11 it follows that (see Appendix A):

$$d^2_{\mathcal{H}}(P_{X_n}, P_Y) = \sum_{i=1}^{Q_x}\sum_{j=1}^{Q_x} \alpha_i \alpha_j K_{H_X}(x_i, x_j) - 2\sum_{i=1}^{Q_x}\sum_{j=1}^{Q_y} \alpha_i \beta_j K_{H_{XY}}(x_i, y_j) + \sum_{i=1}^{Q_y}\sum_{j=1}^{Q_y} \beta_i \beta_j K_{H_Y}(y_i, y_j). \tag{3.3}$$

where

$$H_X = H_K + H_i^p + H_j^p, \quad H_{XY} = H_K + H_i^p + H_j^q, \quad H_Y = H_K + H_i^q + H_j^q.$$

The expression in 3.3 aims to highlight relevant information from data in two different ways when comparing distributions: 1) by weighting the sample sets using a sparse representation that find relevant data structures via a sparse Hilbert space embedding distance; 2) by adapting the covariance matrices according to relevant data structures contained in the sample sets through an adaptive Hilbert space embedding distance. Complete description about these approaches and its usage in the context of ABC is provided in the following sections.

3.1.1 Sparse Hilbert space embedding distance

An initial idea arising from equation 3.3 aims to highlight representative information contained in both distributions to be compared by means of a set of weighting

coefficients. In particular, the probability density functions are estimated using relevant samples through sparseness in the weights such that few elements in $\boldsymbol{\alpha}\in[0,1]^{Q_x}$ and $\boldsymbol{\beta}\in[0,1]^{Q_y}$ are nonzero. In this regard, the covariance matrices are fixed as $H_i^p{=}H_P, \forall i{=}1,2,\ldots,Q_x$, and $H_j^q{=}H_Q, \forall j{=}1,2,\ldots,Q_y$ leading to the following compact form of the Sparse Hilbert Space Embedding Distance (SHSED):

$$d^2_{\text{SHSED}}(P_{X_n},P_Y)=\boldsymbol{\alpha}^\top\boldsymbol{A}^{(H_X)}\boldsymbol{\alpha}+\boldsymbol{\beta}^\top\boldsymbol{B}^{(H_Y)}\boldsymbol{\beta}-2\boldsymbol{\alpha}^\top\boldsymbol{L}^{(H_{XY})}\boldsymbol{\beta} \tag{3.4}$$

where $H_X{=}H_K+2H_P$, $H_Y{=}H_K+2H_Q$, and $H_{XY}{=}H_K+H_P+H_Q$. Moreover, $\boldsymbol{A}^{(H_X)}\in\mathbb{R}^{Q_x\times Q_x}$, $\boldsymbol{B}^{(H_Y)}\in\mathbb{R}^{Q_y\times Q_y}$, and $\boldsymbol{L}^{(H_{XY})}\in\mathbb{R}^{Q_x\times Q_y}$ stand for kernel matrices holding elements $a_{ij}{=}K_{H_X}(x_i,x_j)$, $b_{ij}{=}K_{H_Y}(y_i,y_j)$, and $\ell_{ij}{=}K_{H_{XY}}(x_i,y_j)$, respectively.

To find the weighting coefficients, a constrained quadratic optimization problem based on an Integrated Squared Error (ISE), $\varepsilon(\boldsymbol{\alpha}){=}\int_X(f(x){-}\hat{f}(x))^2dx$, is proposed as follows [8]:

$$\begin{aligned}\underset{\boldsymbol{\alpha}}{\operatorname{argmin}}\,\varepsilon(\boldsymbol{\alpha})=\;&\boldsymbol{\alpha}^\top\boldsymbol{A}^{(2H_P)}\boldsymbol{\alpha}{-}\tfrac{2}{Q_x}\boldsymbol{\alpha}^\top\boldsymbol{A}^{(H_P)}\mathbf{1}\\ \text{s.t.}\quad&||\boldsymbol{\alpha}||_1=1\\ &\alpha_i\geq 0,\quad\forall i=1,2,\ldots,Q_x.\end{aligned} \tag{3.5}$$

The nonlinear problem in 3.5 can be solved using a Sequential Minimal Optimization (SMO) algorithm. In particular, a forward constrained regression approach is utilized that stands for a fast sparse kernel density estimation [21]. Finally, the $\boldsymbol{\beta}$ weights can be found by solving an analogous optimization problem based on $\varepsilon(\boldsymbol{\beta})$.

3.1.2 Adaptive Hilbert space embedding distance

Concerning the extraction of relevant information from data to compare distributions, another attractive approach emerge from the idea to adaptively select the covariance matrix of the kernel used to estimate the probability density functions. In particular, each sample in the observed and simulated data has the same contribution in the sense of the weighting coefficients by fixing $\alpha_i{=}1/Q_x, \forall i{=}1,2,\ldots,Q_x$ and $\beta_j{=}1/Q_y, \forall j{=}1,2,\ldots,Q_y$, leading to the following compact form of the Adaptive Hilbert Space Embedding Distance (AHSED):

$$d^2_{\text{AHSED}}(P_{X_n},P_Y)=\frac{1}{Q_x^2}\mathbf{1}_{Q_x}^\top\boldsymbol{A}^{(H_X)}\mathbf{1}_{Q_x}+\frac{1}{Q_y^2}\mathbf{1}_{Q_y}^\top\boldsymbol{B}^{(H_Y)}\mathbf{1}_{Q_y}-\frac{2}{Q_xQ_y}\mathbf{1}_{Q_x}^\top\boldsymbol{L}^{(H_{XY})}\mathbf{1}_{Q_y}, \tag{3.6}$$

where the kernel matrices $\boldsymbol{A}^{(H_X)}$, $\boldsymbol{B}^{(H_Y)}$, and $\boldsymbol{L}^{(H_{XY})}$ are the same sizes as in 3.4 but whose elements a_{ij}, b_{ij}, and ℓ_{ij} stand for more revealing similarity calculations by using covariance $H_X{=}H_K + H_i^p + H_j^p$, $H_Y{=}H_K + H_i^q + H_j^q$, and $H_{XY}{=}H_K + H_i^p + H_j^q$, respectively. Moreover, $\mathbf{1}_{\mathcal{D}}{\in}\mathbb{R}^{\mathcal{D}}$ is the all-ones vector of size $\mathcal{D}$.

To achieve a more robust comparison approach, a Bayesian-based inference methodology can be used where the minimization of the Bayes risk coefficient (under a particular loss function) leads to a closed form for computing each covariance matrix. In particular, a likelihood function $p(x_i|H_i^p){=}1/(Q_x-1)\sum_{t=1,t\neq i}^{Q_x} K_{H_i^p}(x_i-x_t)$ and an inverse Wishart prior $H_i^p{\backsim}\mathcal{W}^{-1}(r,C)$ can be used to obtain the Bayes estimator, $\widehat{H_i^p}$, by solving an optimization problem based on the Bayes risk $\mathcal{R}(H_i^p){=}\mathbb{E}[\mathcal{L}(\widehat{H_i^p},H_i^p)]$ [49]:

$$\underset{H_i^p}{\operatorname{argmin}}\,\mathcal{R}(H_i^p) = \int_{\mathscr{H}} \mathcal{L}(\widehat{H_i^p},H_i^p)p(H_i^p|x_i)dH_i^p \tag{3.7}$$

where $\mathscr{H}$ is the space of positive definite matrices, $\mathcal{L}(\widehat{H_i^p},H_i^p)$ represents a desired loss function, and $p(H_i^p|x_i)$ stands for the posterior distribution of H_i^p. It is easy to demonstrate that under both quadratic and entropy-based loss functions, the optimization problem in 3.7 has the following close form solutions [49]:

$$\widehat{H_i^p}\Big|_{quad} = \frac{1}{r-d}\frac{\sum\limits_{\substack{t=1\\t\neq i}}^{Q_x}\Big|(x_i{-}x_t)^\top(x_i{-}x_t)+C\Big|^{-\frac{r+1}{2}}\Big[(x_i{-}x_t)^\top(x_i{-}x_t)+C\Big]}{\sum\limits_{\substack{t=1\\t\neq i}}^{Q_x}\Big|(x_i{-}x_t)^\top(x_i{-}x_t)+C\Big|^{-\frac{r+1}{2}}}, \tag{3.8}$$

$$\widehat{H_i^p}\Big|_{ent} = \frac{1}{r+1}\left[\frac{\sum\limits_{\substack{t=1\\t\neq i}}^{Q_x}\Big|(x_i{-}x_t)^\top(x_i{-}x_t)+C\Big|^{-\frac{r+1}{2}}\Big[(x_i{-}x_t)^\top(x_i{-}x_t)+C\Big]^{-1}}{\sum\limits_{\substack{t=1\\t\neq i}}^{Q_x}\Big|(x_i{-}x_t)^\top(x_i{-}x_t)+C\Big|^{-\frac{r+1}{2}}}\right]^{-1}, \tag{3.9}$$

where $C{\in}\mathbb{R}^{d\times d}$ and $r{\in}\mathbb{R}\geq d$ stand for the scale matrix and the degrees of freedom associated to the inverse Wishart prior distribution, respectively. A suitable choice for these hyperparameters concerns $r{=}(Q_x)^{2/(d+4)}$ and the sample covariance matrix $C{=}1/Q_x\sum_{t=1}^{Q_x}(x_t-\bar{x})(x_t-\bar{x})^\top$, $\bar{x}{=}1/Q_x\sum_{t=1}^{Q_x}x_t$ [49]. In turn, the Bayes estimators

for matrices $\widehat{H_j^q}$ can be found by solving an analogous optimization problem based on $\mathcal{R}(H_j^q)$.

3.2 An enhanced ABC for unsupervised inference

Once the comparison between the probability distributions of simulated and observed data has been improved via the so called distances SHSED and AHSED, an enhanced ABC for supervised inference scenarios can be obtained by introducing such distances into the general Hilbert embedding-based ABC scheme. Algorithm 2 summarizes the proposed unsupervised ABC method.

Algorithm 2 ABC based on SHSED/AHSED

Input: Observed data: $\{y^{(j)} \backsim P_Y\}_{j=1}^{Q_y}$, prior: $\zeta(\theta)$, threshold: ϵ, distance parameters, bandwidth: $\sigma_\theta \in \mathbb{R}^+$.

Output: Posterior estimation: $\hat{p}(\theta|y)$.

1: **for** $n = 1, \cdots, N$ **do**
2: $\quad \theta_n \backsim \zeta(\theta)$ ▷ Draw a solution θ_n from the prior.
3: $\quad x_n \backsim p(x|\theta_n)$; $\{x_n^{(i)} \backsim P_{X_n}\}_{i=1}^{Q_x}$ ▷ Draw a sample from the model.
4: $\quad \tilde{w}_n = \kappa_G(\mathrm{d}^2_{\mathcal{H}}(P_{X_n}, P_Y); \epsilon)$ ▷ Compute the n-weight using SHSED or AHSED.
5: **end for**
6: $w_n = \tilde{w}_n / \sum_{n=1}^N \tilde{w}_n$ ▷ Normalize the weights
7: $\hat{p}(\theta|y) = \sum_{n=1}^N w_n \kappa_G(\mathrm{d}_e(\theta, \theta_n); \sigma_\theta)$, ▷ Parzen-based posterior approximation.

3.3 Results

To evaluate the performance of the proposed unsupervised ABC framework, two statistical inference task are considered: a toy problem that comprises a Poisson mixture model and a nonlinear ecological dynamic system termed the Ricker model [47]. As benchmark, the straightforward ABC Rejection and two state-of-the-art ABC methods based on HSE are studied. The former, Maximum Mean Discrepancy (MMD)-based ABC [30], uses empirical density approximations: $\hat{f}(x_i) = 1/Q_x \sum_{j=1}^{Q_x} \delta(x_i - x_j)$, $\hat{g}(y_i) = 1/Q_y \sum_{j=1}^{Q_y} \delta(y_i - y_j)$, being $\delta(\cdot)$ the Dirac delta function. The latter, Parzen-based ABC [50], uses traditional Parzen window estimators as follows: $\hat{f}(x_i) = 1/Q_x \sum_{j=1}^{Q_x} \kappa_G(\mathrm{d}_e(x_i, x_j); \sigma_X)$, $\hat{g}(y_i) = 1/Q_y \sum_{j=1}^{Q_y} \kappa_G(\mathrm{d}_e(y_i, y_j); \sigma_Y)$, where $\kappa_G(\mathrm{d}_e(\cdot,\cdot); \sigma)$ is a Gaussian kernel with covariance $H = \sigma I$, being $I \in \mathbb{R}^{d \times d}$ an identity matrix. Both

methods lead to different distances that can be directly incorporated into the fourth line in algorithm 2 taking into account their corresponding distance parameters. On the other hand, as quantitative assessment, the following relative error is used:

$$\mathcal{E}_{\theta^{(z)}} = 100 \times \frac{||\theta^{(z)} - \sum_{n=1}^{N} w_n \hat{\theta}_n^{(z)}||}{||\theta^{(z)}||}, \tag{3.10}$$

where $\theta^{(z)}$ is the value of the z-th target parameter and $\hat{\theta}_n^{(z)}$ is the n-th ABC-based approximation with weight w_n.

3.3.1 Inference for a Poisson mixture model

Initially, a finite mixture of Poisson distributions is considered as follows:

$$p(x|\boldsymbol{\lambda}, \boldsymbol{\pi}) = \sum_{c=1}^{C} \pi_c \frac{\exp(-\lambda_c)\lambda_c^x}{x!}, \tag{3.11}$$

where $\boldsymbol{\pi} = \{\pi_c\}_{c=1}^{C}$ are the mixing coefficients holding the condition $\sum_{c=1}^{C} \pi_c = 1$, $\boldsymbol{\lambda} = \{\lambda_c\}_{c=1}^{C}$ are the number of average events for each Poisson density ($\lambda_1 < \lambda_2 < \cdots < \lambda_C$), and C is the number of components [44].

For concrete testing, the aim is to estimate the posterior $p(\boldsymbol{\pi}|\boldsymbol{\lambda}, x)$, with $C = 2$ and $\boldsymbol{\lambda} \in \{1, 8\}$. In particular, a Dirichlet distribution is imposed over $\boldsymbol{\pi}$ as prior distribution [44], that is, $\boldsymbol{\pi} \backsim \text{Dirichlet}(1, 1)$. Initially, ten samples for $\boldsymbol{\pi}$ are generated from the prior, then, 50 observations are computed for each of them using the model. Figure 3.1 shows the obtained inference results by fixing $\sigma_X = \sigma_Y = 0.33$ and $\sigma_K = 2.154$ (selected manually). Notice that the same bandwidths hold for matrices H_X, H_Y, and H_K in SHSED-ABC due to the unidimensional properties of data. As seen, the proposed SHSED extracts relevant information from simulations when suitable values for σ_X and σ_Y are selected. In fact, Figure 3.1b shows how sample π_6 is the farthest from the true values of $\boldsymbol{\pi}$ but its associated simulation x_6 in Figure 3.1a is the closest from the observed data; nevertheless, Figure 3.1c shows that the introduced sparse approach assigns a high value for the distance, leading to a low weight. On the other hand, the proposed SHSED approach assigned high weights to the samples π_3, π_4, and π_5, which exihibit close distance values to the true parameters in $\boldsymbol{\pi}$ but distant dependencies with simulations x_3, x_4, and x_5 concerning the observed data.

Now, to test the robustness of the ABC-based estimations regarding the number of simulations, an observed dataset holding 50 (scalar) samples drawn from the Poisson

mixture model is formed, with $\boldsymbol{\pi} \in \{0.3, 0.7\}$. Afterwards, to generate the simulated data, 50 samples are drawn from the model to compute the posterior. This procedure is repeated 100 times. The number of simulations (N) is increased from 10 to 300 using a step size of 10. For the ABC rejection, the Euclidean distance with ϵ=10 was fixed to compare the observed and simulated data; this threshold was defined empirically. For the MMD, the Parzen, and the SHSED-based ABC methods, a soft threshold value of $\epsilon = 0.158$ was used.

Figures 3.1d and 3.1e show the error bar curves of the posterior distribution for the mixing coefficients. As seen, the SHSED improves the posterior estimation leading to the lowest errors and deviations for $\sigma_X = \sigma_Y = 0.406$ (fixed using grid-search). See how the sparse-based weighting procedure favors the identification of relevant structures in an RKHS before computing the dependencies between observed and simulated data within an ABC framework. Additionally, note that for low kernel bandwidth values in 3.5, the SHSED approach converges to the MMD and Parzen-based ones, since the weights $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ become closer to $1/Q_x$ and $1/Q_y$, respectively.

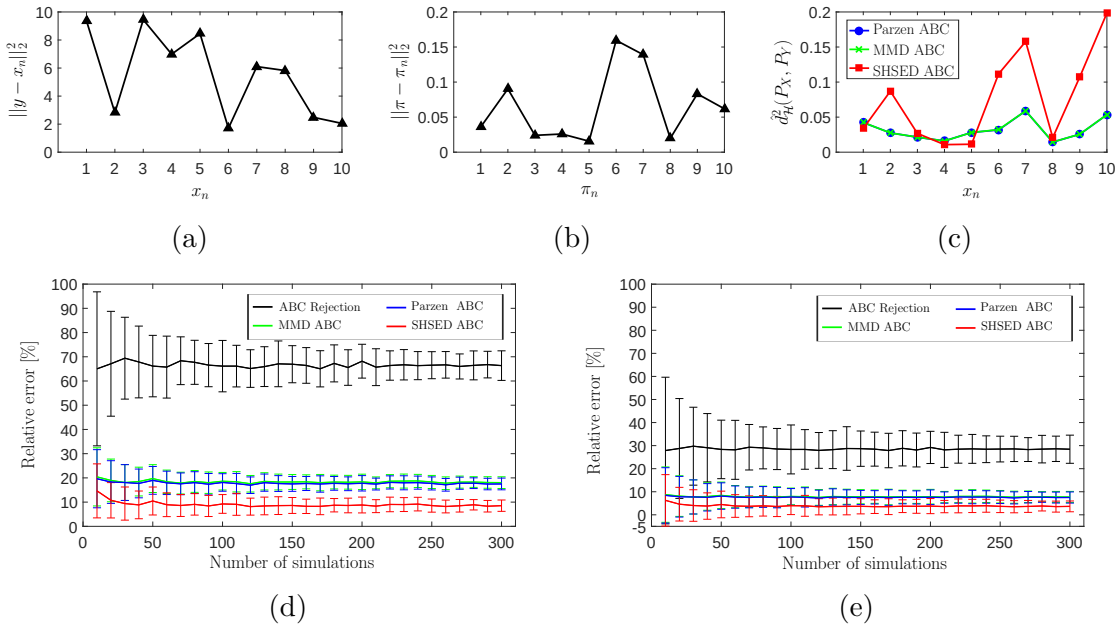


Fig. 3.1 Poisson mixture model results. (a) Euclidean distance between the observed and simulated data. (b) Euclidean distance between candidates sampled from the prior and the true values of the parameters. (c) Different distances based on HSE between distributions associated with the observed and simulated data. (d),(e) Relative error bar curves for the posterior of the mixing coefficients: π_1 and π_2 respectively.

3.3.2 Inference in an ecological non-linear Ricker map

This nonlinear ecological dynamic system can be modeled using a discrete differential equation as follows [17]:

$$\ln(M^{(t)}) = \ln(r)+\ln(M^{(t-1)})-M^{(t-1)}+e^{(t)} \tag{3.12}$$

where $M^{(t)}\in\mathbb{R}$ is the size of some animal population at time t, $e^{(t)}\backsim\mathcal{N}(0,\sigma_e^2)$, being $\sigma_e\in\mathbb{R}^+$ the standard deviation of the innovations, and $\ln(r)$ is related to the growth rate parameter of the model ($r\in\mathbb{R}^+$). Additionally, the observation y is a time series drawn from a Poisson distribution as $y\backsim\text{Poisson}(\phi M^{(t)})$, with $y\in\mathbb{N}$, where ϕ is a scale parameter [19]. Thus, the Ricker model is completely parametrized by $\boldsymbol{\theta}=[\ln(r),\phi,\sigma_e]$.

In this experiment, 50 samples from the model are drawn with $\boldsymbol{\theta} = [3.8, 10, 0.3]$ by fixing the following priors [17]: $\ln(r)\backsim\mathcal{N}(4, 0.5)$; $\phi\backsim\mathcal{X}^2(10)$; $\sigma_e\backsim\text{invgamma}(3, 1.3)$. Figure 3.2 shows the observed data. See how infering the model parameters from this data is a quite challenging task due to high variability concerning the stochastic property of the scaled Ricker map [47].

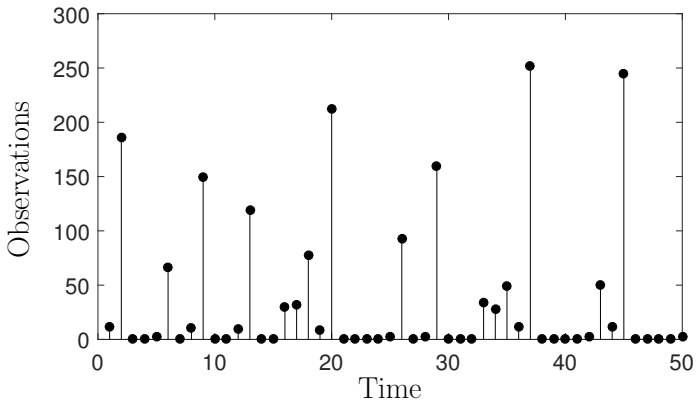


Fig. 3.2 Observed data to perform inference in the scaled Ricker map.

Regarding the free parameter selection, the threshold for the ABC rejection approach was fixed empirically as ϵ=75, while for the MMD, Parzen and, SHSED ABC, the similarity kernel parameter was selected as ϵ=0.158. Besides, the values σ_X=σ_Y=0.0158 were employed to compute the sparse weights for simulations and observations in SHSED-ABC and the characteristic kernel width was treated as σ_K=2.5.

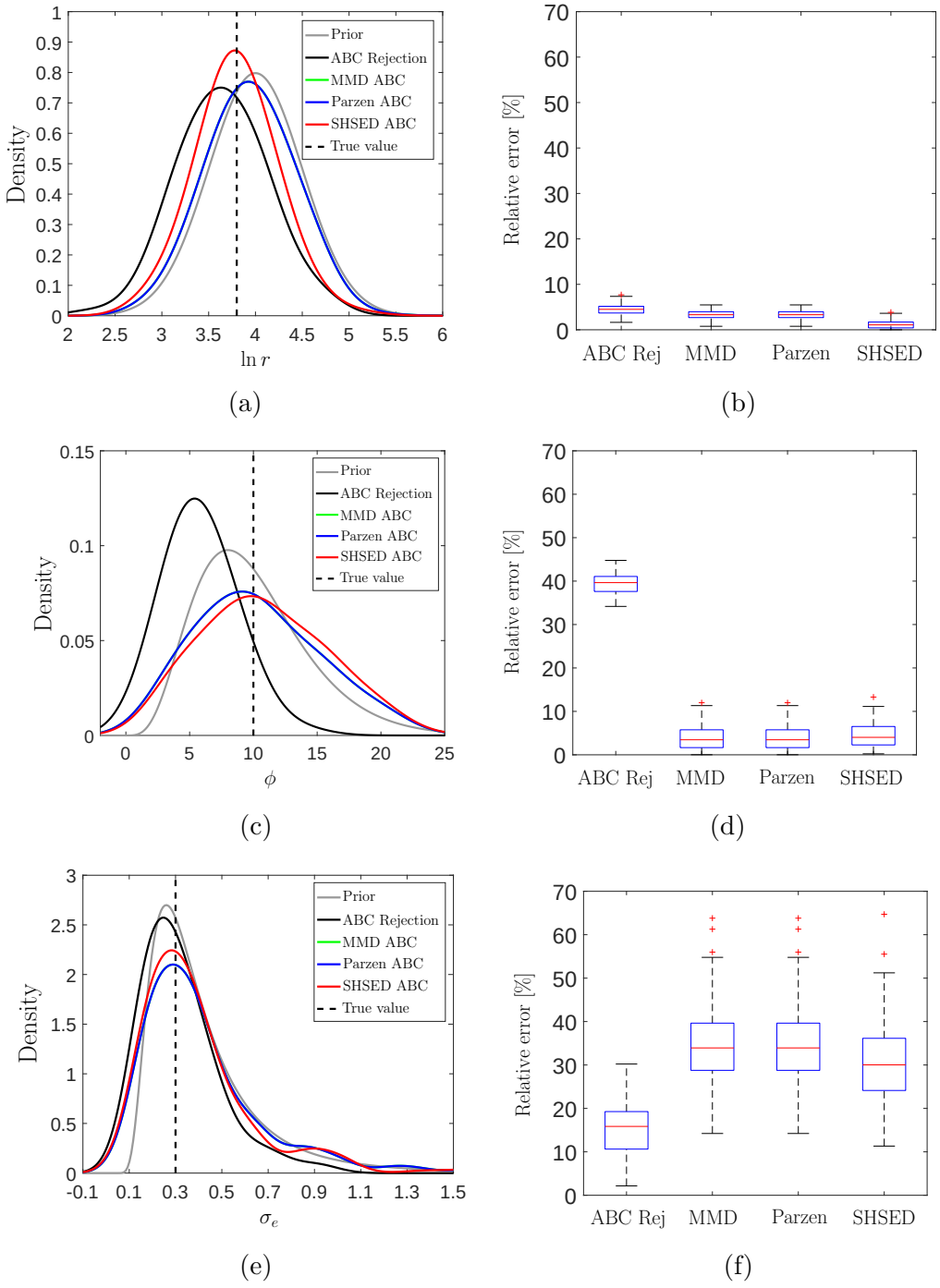


Fig. 3.3 Ricker model results. Left column: Different ABC-based estimated posteriors for $\{\ln r,\ \phi,\ \sigma_e\}$ with $\sigma_\theta = \{0.253,\ 2.190,\ 0.082\}$, respectively. Right column: Relative error boxplots for the posterior of the Ricker model parameters: $\ln r$, ϕ, and σ_e, respectively.

According to the results in Figures 3.3a, 3.3c, and 3.3e the MMD-based posteriors are almost identical to the ones obtained by the Parzen-based approach. Furthermore, notice how the posterior distribution for $\ln(r)$ and σ_e parameters obtained using the proposed SHSED has their maximum closer to the true value in comparison to the benchmarks, and the rejection-based ABC has the worst performance. In the case of ϕ, it can be seen that the results obtained by the SHSED are close enough to the other ABC techniques based on HSED.

Lastly, to test the stability of the inference, the procedure to approximate the posterior of the model parameters concerning the relative error assessment was repeated 100 times. Obtained results in Figures 3.3b, 3.3d, and 3.3f reveal how the proposed SHSED reaches the lowest uncertainty for the $\ln(r)$ and ϕ inferences. In the case of σ_e, since the true value is small ($\sigma_e = 0.3$), little changes in the posterior shape produce a substantial change in the expected value; hence, relevant changes in the index error are gathered. Remarkably, the lowest uncertainty for σ_e is related to the ABC rejection method due to the smoother form of the posterior in comparison to the other HSED methods; nonetheless, SHSED achieves the best overall performance (see Figure 3.4).

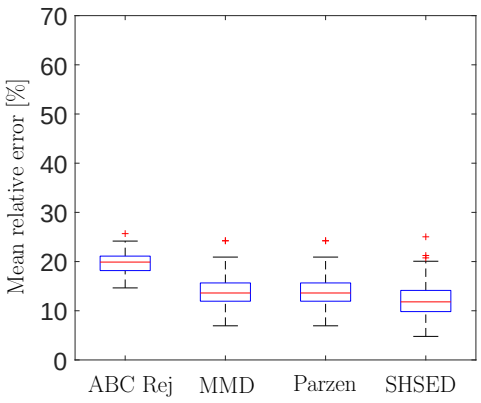


Fig. 3.4 Overall performance of the different ABC methods in terms of the mean relative error.

Chapter 4

An automatic ABC approach for supervised inference scenarios

This chapter provides mathematical foundations for achieving an automatic ABC-based statistical inference approach for supervised tasks. It shows how information about the parameter space can be included as a reference in the ABC procedure to define similarities over simulations using a metric learning-based statistical alignment. It also describes a local linear embedding-based framework to introduce the concept of neighborhood in the context of ABC for highlighting relevant samples in the posterior estimation. As a result, the chapter presents a novel ABC algorithm that does not require the tunning of any free parameter seeking for a quite competitive method compared to other non-automatic state-of-the-art ABC techniques.

4.1 Initial remarks

As it was mentioned in previous chapters, the ABC framework based on Hilbert embedding as introduced by Park et al. [30] only works for unsupervised inference in the sense that decisions about samples in the parameter space (posterior approximations) are made in terms of comparisons over the simulation space. This approach is somehow *blinded* and does not take into account information related to parameter candidates directly into the ABC procedure. As a consequence, the quality of the posterior estimations strongly relies on a proper tuning procedure of the free parameters leading to expensive and/or time-consuming routines like cross-validation or grid search [38].

Inspired by the idea of including relevant information contained in the parameter space into the ABC procedure, a two stage-based methodology is introduced in this

research according to the illustrative diagram shown in figure 4.1. Highlighted in light blue, a first stage comprises the matching between similarities defined over candidates in the parameter space and mappings of simulated data as defined by $\vartheta{:}\mathcal{X}\to\mathcal{S}$ in the feature space $\mathcal{S}$. This strategy aims to set the notion of similarity over simulated and observed data as close as possible to the idea of likeness over prior samples via a statistical alignment methodology that learns a distance $\mathrm{d}_{\mathcal{S}}{:}\mathcal{S}\times\mathcal{S}\to\mathbb{R}^{+}$. Then, in a second stage (indicated in light orange), a novel way to compute the posterior weights uses a similarity kernel that is a function of the learned distance to highlight relevant samples into the posterior distribution. In particular, the assigned posterior weights are calculated by means of a truncated representation that searches for the optimal number of nearest neighbors according to the detection of local relationships over samples in Θ. The resulting ABC algorithm has a potential advantage: the two additional stages introduce additional information into the ABC framework that can be used to select all the free parameters yielding to an automatic ABC method.

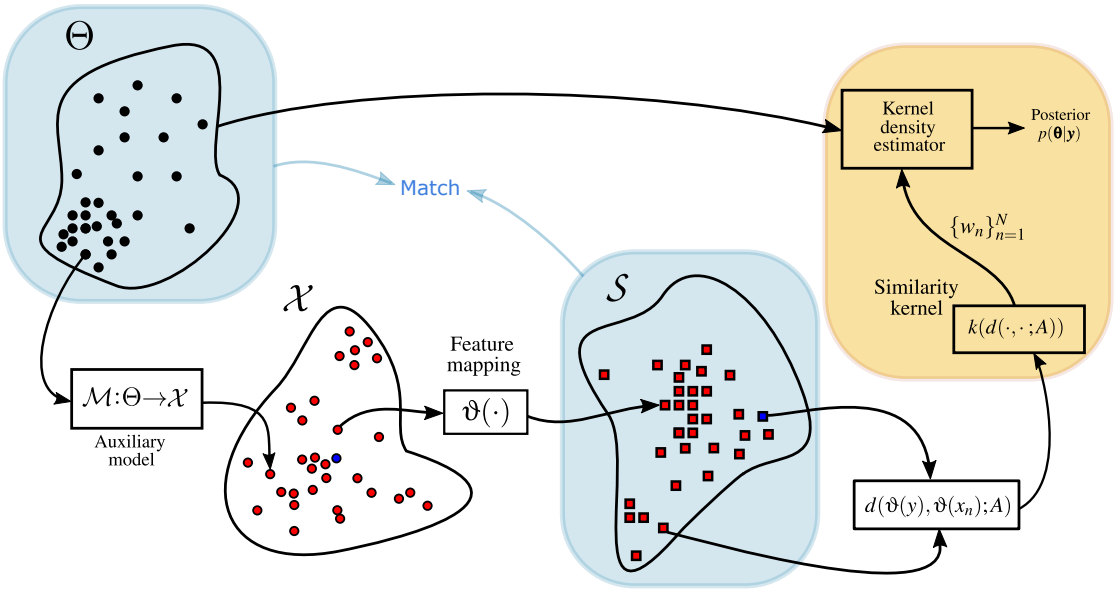


Fig. 4.1 A sketch for the proposed automatic ABC method.

4.2 Proposal fundamentals

4.2.1 Aligning the parameter and simulation spaces in ABC

To avoid the influence of the ϵ value and the kernel parameters while computing the ABC-based posterior as in equation 2.3, an automatic statistical alignment approach for enhancing and automating the inference task is introduced. The aim of such an

alignment is to include the information contained in the candidates $\{\theta_n\}_{n=1}^N$ to improve the comparison stage carried out over simulations and observations. Let $\Psi=\{\theta_n, x_n\}_{n=1}^N$ be the set of N candidates $\theta_n \in \mathbb{R}^P \backsim \zeta(\theta)$ drawn from the desired prior distribution $\zeta(\theta)$ and their corresponding simulations $x_n \in \mathbb{R}^Q \backsim p(x|\theta)$. Further, let the kernel function $\kappa_\theta : \Theta \times \Theta \to \mathbb{R}^+$ be a similarity measure between candidates in Θ, that define the kernel matrix $\boldsymbol{K_\theta} \in \mathbb{R}^{N \times N}$ holding elements:

$$\kappa_\theta(\theta_n, \theta_{n'}) = \begin{cases} \exp(-\mathrm{d}_\Theta^2(\theta_n, \theta_{n'})), & \theta_n \in \Omega_{n'} \\ 0, & \text{otherwise,} \end{cases} \tag{4.1}$$

where $\Omega_{n'}$ is a set holding the M-nearest neighbors of $\theta_{n'}$ in the sense of the distance $\mathrm{d}_\Theta : \Theta \times \Theta \to \mathbb{R}^+$ (see section 4.2.2).

In this research, to avoid large variations among components of θ_n, the Mahalanobis distance is employed as follows:

$$\mathrm{d}_\Theta^2(\theta_n, \theta_{n'}) = (\theta_n - \theta_{n'})^{\mathrm{T}} \boldsymbol{\Sigma}_\Theta^{-1} (\theta_n - \theta_{n'}), \tag{4.2}$$

where $\boldsymbol{\Sigma}_\Theta \in \mathbb{R}^{P \times P}$ is the sample covariance matrix of $\{\theta_n\}_{n=1}^N$.

Concerning the feature space $\mathcal{S}$, the similarity assessment is computed via the kernel $\kappa_s : \mathcal{S} \times \mathcal{S} \to \mathbb{R}^+$ to build the matrix $\boldsymbol{K_s} \in \mathbb{R}^{N \times N}$ holding elements:

$$\kappa_s(\vartheta(x_n), \vartheta(x_{n'})) = \exp(-\mathrm{d}_\mathcal{S}^2(\vartheta(x_n), \vartheta(x_{n'}))), \tag{4.3}$$

where $\mathrm{d}_\mathcal{S}^2 : \mathcal{S} \times \mathcal{S} \to \mathbb{R}^+$ and $\vartheta : \mathcal{Y} \to \mathcal{S}$ is a feature mapping. To perform the pairwise comparison between simulations in $\mathcal{S}$, a Mahalanobis distance is introduced as [3]:

$$\mathrm{d}_\mathcal{S}^2(\vartheta(x_n), \vartheta(x_{n'})) = (\vartheta(x_n) - \vartheta(x_{n'}))^\top \boldsymbol{A}\boldsymbol{A}^\top (\vartheta(x_n) - \vartheta(x_{n'})), \tag{4.4}$$

where $\boldsymbol{\Sigma}_\mathcal{S}^{-1} = \boldsymbol{A}\boldsymbol{A}^\top$ stands for the inverse covariance matrix of $\vartheta(x_n) \in \mathbb{R}^D$ and $\boldsymbol{A} \in \mathbb{R}^{D \times d}$. In this sense, the information concerning the similarity over candidates in Θ, represented via $\boldsymbol{K_\theta}$, is used to state the notion of similarity over simulations and observation in $\mathcal{S}$, represented via $\boldsymbol{K_s}$ (see figure 4.1). The statistical alignment between two kernel matrices has been studied from the Metric Learning perspective where the most popular approach is the Centered Kernel Alignment (CKA) [48]. In particular, the following CKA-based measure between the above kernel matrices is employed [9]:

$$\hat{\rho}(\boldsymbol{K}_{\boldsymbol{\theta}}, \boldsymbol{K}_{\boldsymbol{s}}) = \frac{\langle \bar{\boldsymbol{K}}_{\boldsymbol{\theta}}, \bar{\boldsymbol{K}}_{\boldsymbol{s}} \rangle_{\mathrm{F}}}{\sqrt{\langle \bar{\boldsymbol{K}}_{\boldsymbol{\theta}} \bar{\boldsymbol{K}}_{\boldsymbol{\theta}} \rangle_{\mathrm{F}} \langle \bar{\boldsymbol{K}}_{\boldsymbol{s}} \bar{\boldsymbol{K}}_{\boldsymbol{s}} \rangle_{\mathrm{F}}}}, \tag{4.5}$$

where $\bar{\boldsymbol{K}}$ stands for the centered kernel as $\bar{\boldsymbol{K}} = \tilde{\boldsymbol{I}}\boldsymbol{K}\tilde{\boldsymbol{I}}$, being $\tilde{\boldsymbol{I}}{=}\boldsymbol{I}{-}\mathbf{1}^{\top}\mathbf{1}/N$ the empirical centering matrix, $\boldsymbol{I}{\in}\mathbb{R}^{N\times N}$ is the identity matrix, and $\mathbf{1}{\in}\mathbb{R}^{N}$ is the all-ones vector. Moreover, the notation $\langle\cdot,\cdot\rangle_{\mathrm{F}}$ represents the matrix-based Frobenius norm. In equation 4.5, $\hat{\rho}(\cdot,\cdot)$ is a data driven estimator that aims to quantify the similarity between the parameter space and the feature space. To find the projection matrix $\boldsymbol{A}$, the following optimization problem can be solved:

$$\hat{\boldsymbol{A}} = \arg\max_{A} \log\left(\hat{\rho}(\boldsymbol{K}_{\boldsymbol{s}}(\boldsymbol{A}), \boldsymbol{K}_{\boldsymbol{\theta}})\right), \tag{4.6}$$

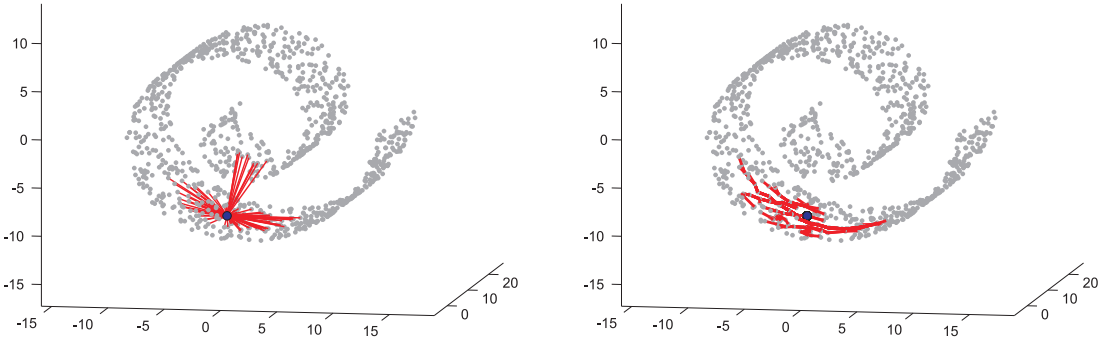
where the logarithm function is used for mathematical convenience. The optimization problem in 4.6 can be solved using a gradient descent-based approach [3].

4.2.2 Revealing local relationships over parameter samples

Setting the ϵ value to find the posterior weights in equation 2.3 is a crucial step. Depending on the distance output values, a particular choice for ϵ could produce a *peaked* posterior when just a few number of weights have larger values or lead to a posterior similar to the prior distribution in the limit condition when $w_n{\to}1/N, \forall n{=}1,2,\ldots,N$. In this regard, the truncated representation to define similarities over the parameter space introduced in equation 4.1 sets an alternative path to avoid the influence of ϵ via the concept of neighborhood. The central aim concerns the selection of a number $M{\in}\mathbb{N}$ of nearest neighbors that reveals representative prior samples into the posterior. The M-value could be fixed manually after an exhaustive search based on cross-validation. However, that would change one problem by another.

In this research, an automatic technique based on Locally Linear Embedding (LLE) and graph theory, the Local Neighborhood Selection (LNS) algorithm introduced by Álvarez et al. [2], facilitates the selection of the optimal number of nearest neighbors. The idea behind LNS is to define a suitable number of neighbors for each sample in the data set taking into account the structure of the manifold. In particular, it computes the neighborhood as the balance between a neighborhood found by the Euclidean distance and a neighborhood found by the geodesic distance based on the principle that when the region around a point is linear and dense, the Euclidean and geodesic distances obtain a similar set of nearest neighbors for each sample; otherwise,

the Euclidean distance will detect short connections while the geodesic distance will identify the right neighbors of each sample [2]. For a better illustration, figure 4.2 shows the nearest neighbors for a particular sample in the manifold (filled bullet) using both the Euclidean and the geodesic distance. Notice how the Euclidean distance selects neighbors that do not follow the structure of the manifold while the geodesic distance understands the actual structure of the samples leading to a proper selection of the nearest neighbors.



(a) (b)

Fig. 4.2 Neighborhoods of a sample according to different distances. (a) Euclidean distance. (b) Geodesic distance. Source: [2].

The LNS algorithm can be summarized as follows [2]:

1. Compute the Euclidean distance D_E for all points in Θ.
2. Construct the minimal connected neighborhood graph $\mathcal{G}$ of the given data set Θ by the k-nearest neighbors method with k_{min}, initializing $k_{min} = 1$. Check the full connectivity of the graph by using the Breadth-first search (BFS) [35]. If the graph is not full connected, update $k_{min} = k_{min} + 1$ and start again this step.
3. Compute the geodesic distance D_G over $\mathcal{G}$ by using Dijistra's algorithm [35].
4. Define $k_{max} = N^2/(k_{min}N_E)$, where N_E is the number of edges in $\mathcal{G}$ and N the number of samples in Θ.
5. Set the vector $\mathbf{k}_s = [k_{min} + 1, ..., k_{max}]$, with $\mathbf{k}_s \in \mathbb{R}^B$. The vector $\mathbf{k}_s$ contains the possible values of k for every $\boldsymbol{\theta}_i$.

6. For each $\boldsymbol{\theta}_i$ define the sets $\boldsymbol{\eta}_D^{(c)}$ and $\boldsymbol{\eta}_{D_G}^{(c)}$, with $c = 1, ..., B$. Each element in $\boldsymbol{\eta}_D^{(c)}$ and $\boldsymbol{\eta}_{D_G}^{(c)}$ corresponds to the k_{s_c} nearest neighbors $\boldsymbol{\theta}_j$ of $\boldsymbol{\theta}_i$ ($j = 1, ..., k_{s_c}$) according to D_E and D_G respectively.

7. Calculate the linearity conservation matrix $\mathbf{V}$ of size $N \times B$, which analyzes the similarity of the neighborhoods obtained by D_E and D_G, taking into account the patch size. Each element of $\mathbf{V}$ can be computed as, $\mathbf{V}_{ic} = |\{\overline{\boldsymbol{\eta}_D^{(c)} \cap \boldsymbol{\eta}_{D_G}^{(c)}}\}|/k_{s_c}$, where $|\cdot|$ calculates the cardinality of a set and $\overline{\{\cdot\}}$ the complement.

8. Initially, for each $\boldsymbol{\theta}_i$ define the set $\boldsymbol{k}_o = \emptyset$. Verify the equality $\mathbf{V}_{ic} = \min\{\mathbf{v}_i\}$, where $\mathbf{v}_i$ is a row vector of $\mathbf{V}$ of size $1 \times B$. If the equality is fulfilled update $\boldsymbol{k}_o = \boldsymbol{k}_o \cup k_{s_c}$.

9. Define k_i for each $\boldsymbol{\theta}_i$ as $k_i = \max\{\boldsymbol{k}_o\}$.

10. Smooth k_i to obtain similar properties in near neighborhoods according to $k_i = (k_i + \mathbf{k}_{\boldsymbol{\eta}}\mathbf{1}) / (k_i + 1)$, where $\mathbf{k}_{\boldsymbol{\eta}}$ is a vector of size $1 \times k_i$, with the sizes of the neighborhoods of each element in $\boldsymbol{\eta}$ (set with the $\boldsymbol{\theta}_j$ nearest neighbors of $\boldsymbol{\theta}_i$ using Euclidean distance, with $j = 1, .., k_i$), and $\mathbf{1}$ is a column vector of size $k_i \times 1$.

11. Store all the values k_i in the vector $\mathbf{k}$.

12. Remove the outliers in $\mathbf{k}$ (see [34]), and replace them by the average of the elements in $\mathbf{k}$, which were not identified as outliers.

13. Each element in $\mathbf{k}$ contains the number of nearest neighbors k_i for each $\boldsymbol{\theta}_i$.

Finally, to accomplish a global representation of the manifold, the M-value is fixed as M=median($\mathbf{k}$).

4.3 An automatic metric learning-based ABC for supervised scenarios

Once the statistical alignment between the spaces of parameters and features is completely automated by means of the CKA-based Metric Learning approach and the LNS algorithm, the distance $\mathrm{d}_{\mathcal{S}}^2$:$\mathcal{S}\times\mathcal{S}\to\mathbb{R}^+$ is properly learned (see equation 4.3). Namely, the projection matrix $\boldsymbol{A}$ defines the notion of similarity over the feature space of simulations and observations as close as possible to the idea of similarity over prior

candidates in the parameter space. Thereby, for the sake of neighborhood preservation, a weighted sample set $\Psi{=}\{(\theta_n, w_n)\}_{n=1}^{N}$ can be form by fixing:

$$w_n{=}\frac{\kappa_E(z, z_n)}{\sum_{n=1}^{N} \kappa_E(z, z_n)} \tag{4.7}$$

being $\kappa_E{:}\mathbb{R}^d{\times}\mathbb{R}^d{\to}\mathbb{R}$ a simmilarity kernel defined as:

$$\kappa_E(z, z_n) = \begin{cases} \exp(-||z - z_n||_2^2), & z_n \in \Upsilon \\ 0, & \text{otherwise,} \end{cases} \tag{4.8}$$

where Υ is a set holding the M-nearest neighbors of $z{=}\vartheta(y)^\top \hat{\boldsymbol{A}}$ in the sense of the Euclidean distance.

The previous setting results in an Automatic Metric Learning-based ABC, named AML-ABC, that can be summarizes in algorithm 3.

Algorithm 3 AML-ABC algorithm

Input: Observed data: y, prior: $\zeta(\theta)$, mapping: ϑ, M-nearest neighbors, width: σ_θ.
Output: Posterior estimation: $\hat{p}(\theta|y)$.
Metric learning stage:
$\Psi^{'}{=}\left\{(\theta_n^{'}, x_n^{'})\right\}_{n=1}^{N}$; $\theta_n^{'} \backsim \zeta(\theta)$, $x_n^{'} \backsim p(x|\theta_n^{'})$ ▷ Draw training data.
$\hat{\boldsymbol{A}} = \arg\max_{\boldsymbol{A}} \log\left(\hat{\rho}(\boldsymbol{K_s}(\boldsymbol{A}), \boldsymbol{K_\theta})\right)$ ▷ Compute CKA based on ϑ, M, $\theta_n^{'}$, and $x_n^{'}$.
Inference stage:
$\Psi{=}\{(\theta_n, x_n)\}_{n=1}^{N}$; $\theta_n \backsim \zeta(\theta)$, $x_n \backsim p(x|\theta_n)$ ▷ Draw simulated data.
$z{=}\vartheta(y)^{\mathrm{T}}\hat{\boldsymbol{A}}$ ▷ Project features of observed data
for $n = 1, \cdots, N$ **do**
    $z_n{=}\vartheta(x_n)^{\mathrm{T}}\hat{\boldsymbol{A}}$ ▷ Project features of simulated data
    $\tilde{w}_n = \kappa_E(z, z_n)$ ▷ Compute the n-th weight value.
end for
$w_n = \tilde{w}_n / \sum_{n=1}^{N} \tilde{w}_n$ ▷ Normalize the weights
$\hat{p}(\theta|y){=}\sum_{n=1}^{N} w_n \kappa_G(\mathrm{d}_e(\theta, \theta_n; \sigma_\theta))$ ▷ Compute the posterior.

4.4 Results

To test the AML-ABC performance, two experiments are considered following [30]. Firstly, a toy problem concerning synthetic data from a mixture of uniform distributions is studied, where the central aim is to approximate the posterior of the mixing coefficients. In a second experiment, inference for a real insect population is analyzed using Nicholson's classic blowfly data; the inference problem comprises the posterior approximation of six model parameters given the real observed data. For comparison purposes, the K2-ABC method is selected as the benchmark in the mixture model due to its nice performance over other methods [30]. On the other hand, the K-ABC, IS-ABC, SA-ABC, and Synthetic Likelihood ABC are selected as a benchmark for the Nicholson's experiment.

4.4.1 Inference for a uniform mixture model

In this case, a mixture of uniform distributions is studied as:

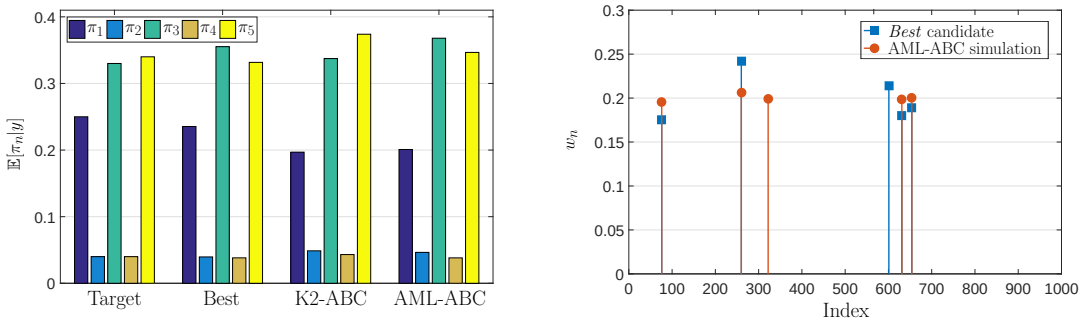
$$p(x|\boldsymbol{\pi}) = \sum_{c=1}^{C} \pi_c \mathscr{U}(c{-}1, c), \tag{4.9}$$

where $\boldsymbol{\pi}{=}\{\pi_c\}_{c=1}^{C}$ are the mixing coefficients holding $\sum_{c=1}^{C} \pi_c{=}1$, and C is the number of components. Moreover, $\mathscr{U}(a, b)$ stands for the uniform distribution with boundary parameters a and b.

In particular, the aim is to estimate the posterior $p(\boldsymbol{\pi}|y)$ for $C{=}5$, given synthetic observations y drawn from the mixture with true parameters (target): $\boldsymbol{\pi}^{*}{=}\{0.25, 0.04, 0.33, 0.04, 0.34\}$. For concrete testing, $N{=}1000$ samples from a symmetric Dirichlet prior were drawn,$\boldsymbol{\pi}{\backsim}$Dirichlet($\mathbf{1}$), and then used the mixture model to form the simulated data by drawing 400 observations for each prior candidate. Moreover, a histogram with 10 bins is utilized as feature mapping in AML-ABC, while the kernel widths in K2-ABC were fixed as $\gamma{=}0.1$, $\epsilon{=}0.001$ [30]. As quantitative assessment, the Euclidean distance $\mathcal{E}{=}||\boldsymbol{\pi}^{*}{-}\hat{\boldsymbol{\pi}}||_2$ was selected, where $\hat{\boldsymbol{\pi}}$ is the expected value of the posterior using the weights $\{w_n\}_{n=1}^{N}$ obtained by using each method.

Since this is a controlled experiment with known parameters $\boldsymbol{\pi}^{*}$, the best possible performance of the AML-ABC can be found by running the inference stage with $\tilde{w}_n{=}\kappa_E(\boldsymbol{\pi}^{*}, \boldsymbol{\pi}_n)$ in Algorithm 3. This approach is referred to as *Best*. The previous

setting is equivalent to think that the CKA between $\boldsymbol{K_\theta}$ and $\boldsymbol{K_s}$ is perfect ($\boldsymbol{K_\theta}$=$\boldsymbol{K_s}$). Figure 4.3 shows the *Best* performance along with K2-ABC and AML-ABC results over the uniform mixture problem. In Figure 4.3a, the expected value of the posterior computed for all methods is close to the target. In particular, the obtained index errors were: $\mathcal{E}_{Best}$=0.030, $\mathcal{E}_{\text{K2-ABC}}$= 0.063, and $\mathcal{E}_{\text{AML-ABC}}$=0.064. These results show that the AML-ABC is a competitive estimator to K2-ABC with a significant advantage concerning the automatic selection of free parameters. In addition, to provide a better understanding of the AML-ABC efectiveness, Figure 4.3b provides the weights for the M=5 nearest neighbors needed to compute the posteriors. As noted, the majority of the simulations found via the LNS algorithm match the selected candidates using the *Best* approach, even though the target values were never introduced in the AML-ABC inference procedure.



(a) (b)

Fig. 4.3 Uniform mixture model results. (a) Estimated mean posterior of mixing coefficients using various methods (b) Weights of the 5 nearest neighbors in AML-ABC.

4.4.2 Inference in a real blowfly data-set

In this real dataset experiment, the problem concerns inferring the dynamics of an adult blowfly population as introduced in [47]. Mathematically, the population dynamics are modeled via a discretized differential equation of the form:

$$N_{t+1}=PN_{t-\tau}\exp\left(-\frac{N_{t-\tau}}{N_0}\right)e_t + N_t\exp(-\delta\epsilon_t), \tag{4.10}$$

where N_{t+1} denotes the observation time at t+1 which is determined by the time-lagged observations N_t and $N_{t-\tau}$; e_t and ϵ_t stand for Gamma distributed noise $e_t \backsim \mathcal{G}(1/\sigma_p^2, \sigma_p^2)$ and $\epsilon_t \backsim \mathcal{G}(1/\sigma_d^2, \sigma_d^2)$.

In this problem, the aim is to estimate the posterior of the model parameters $\boldsymbol{\theta}=\{P, N_0, \sigma_d, \sigma_p, \tau, \delta\}$ given observed data concerning a time series of 180 observations[1]. The following Log-normal distributions are used for setting priors over $\boldsymbol{\theta}$ [26]: $\log P \backsim \mathcal{N}(2, 2^2)$, $\log N_0 \backsim \mathcal{N}(6, 1)$, $\log \sigma_d \backsim \mathcal{N}(-0.5, 1)$, $\log \sigma_p \backsim \mathcal{N}(-0.5, 1)$, $\log \tau \backsim \mathcal{N}(2.7, 1)$, $\log \delta \backsim \mathcal{N}(-1, 0.4^2)$. Figure 4.4 shows the observed data. Inferring the model parameters in this blowfly data-set is a very challenging task since the system dynamics can easily move from stable to chaotic regimes: a small change in any of the model parameters could produce a tremendous change in the trajectory of the system [26, 47]. This states an interesting scenario to test the performance and robustness of the AML-ABC.

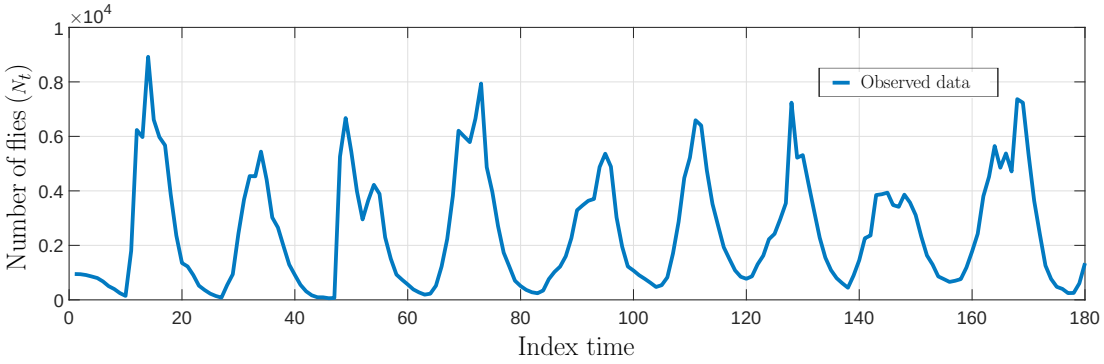


Fig. 4.4 Observations to perform inference in the real blowfly data-set.

For concrete testing, concerning the AML-ABC settings, N=5000 samples were drawn from the prior to form the simulated data by drawing 180 observations for each prior candidate using the model as in equation 4.10. Besides, as feature mapping, the custom 10 statistics in this kind of data were selected as: the log of the mean of all 25% quantiles of $\{N_t/1000\}_{t=1}^{180}$ (four statistics), the mean of 25% quantiles of the first-order derivatives of $\{N_t/1000\}_{t=1}^{180}$ (four statistics) and the maximal peaks of smoothed $\{N_t\}_{t=1}^{180}$ using two different thresholds (two statatistics) [30]. Moreover, the Euclidean distance $\mathscr{E}=||\vartheta(y)-\vartheta(x_n|\hat{\boldsymbol{\theta}})||_2$ was selected as quantitative assessment, where $x_n|\hat{\boldsymbol{\theta}}$ is a simulation from the model given the expected value of the posterior using each method. Namely, due to fluctuations produced by ϵ_t and e_t, 100 simulation for each method were computed using the expected value and then the median and standard deviation for $\mathscr{E}$ provide the performance of each method [30].

[1] Available on the supplementary materials of [47].

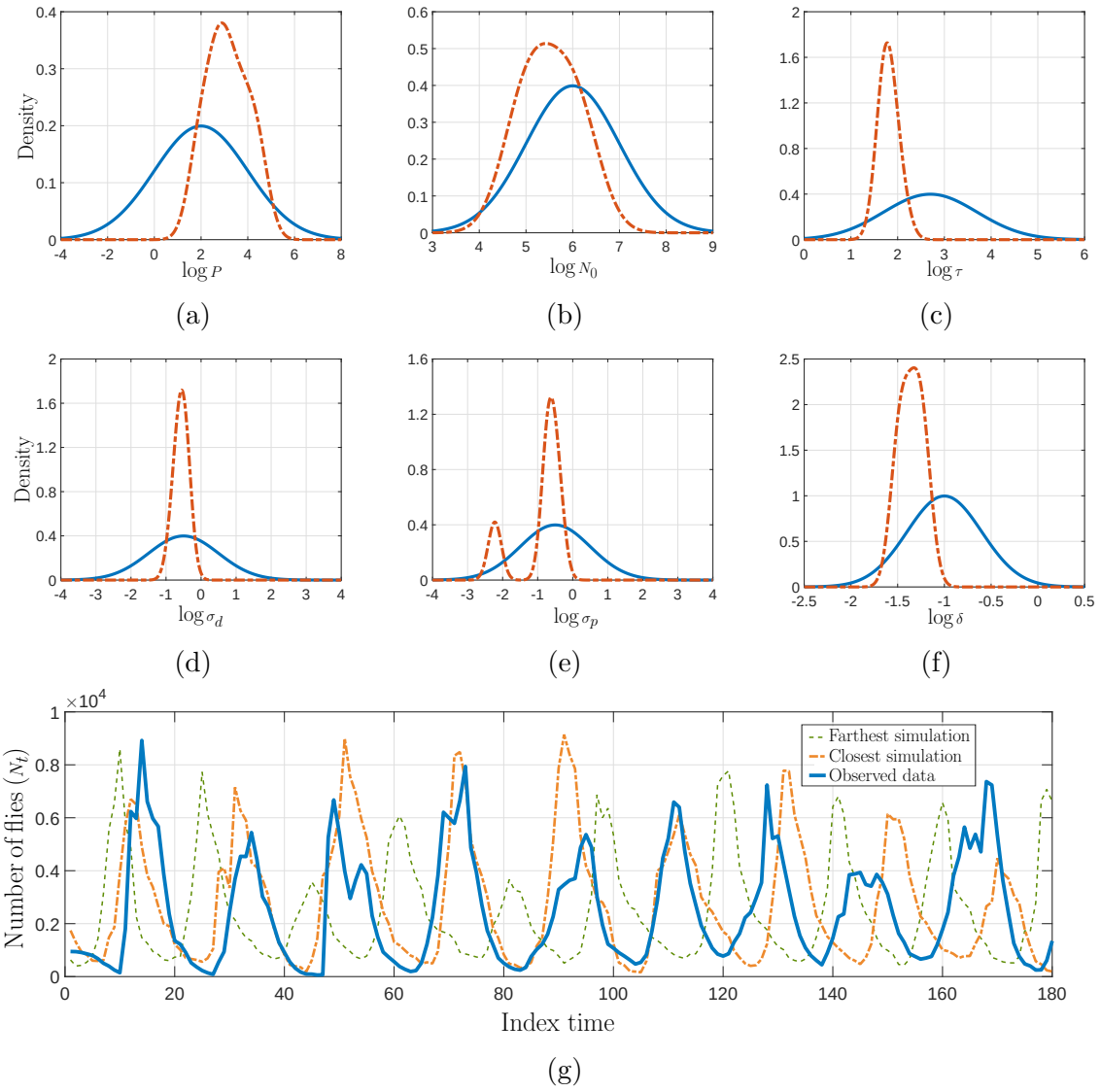


Fig. 4.5 Non-linear ecological dynamic system results. (a)-(f) Prior distribution (solid line) and AML-ABC-based posterior estimation (dashed line) of model parameters in the log-space. (g) Some realizations from the model using the expected value of the parameters found via AML-ABC.

Figures 4.5a to 4.5f show the prior and posterior approximation for each parameter fixing σ_θ according to [36]. Notice how the proposed AML-ABC updates the beliefs about the model parameters leading to more concentrated posteriors. In the case of $\log \sigma_p$, two modes reflect different intervals with probable values for driving the noise realization associated with egg production in the blowfly population. However, there is a predominant mode that states higher probabilities for this parameter. Furthermore, Figure 4.5g shows the closest and farthest simulation to the observed data from 100 real-

ization used to compute $\mathscr{E}$, showing that the obtained posteriors lead to a stable regime.

Table 4.1 Performance of different ABC schemes over the blowfly dataset.

Method	**median**($\mathscr{E}$)$\pm$**std**($\mathscr{E}$)
Kernel ABC [29]	5.2 ± 3.0
Indirect score ABC [16]	2.1 ± 1.8
Semi-Automatic ABC [15]	1.9 ± 1.2
Synthetic Likelihood ABC [47]	1.7 ± 1.3
K2-ABC [30]	$\mathbf{1.0 \pm 0.8}$
AML-ABC	$\mathbf{1.2 \pm 0.4}$

Finally, Table 4.1 shows the performance of AML-ABC compared to different ABC-based methods tested on the blowfly dataset by authors in [30], where clearly the AML-ABC is a quite competitive approach to K2-ABC.

Chapter 5

Final remarks

5.1 Conclusions

Unsupervised inference scenarios. An enhanced methodology for performing statistical inference using a Hilbert embedding-based ABC framework was proposed. In particular, two novel distances to compare distributions associated with two random variables in an RKHS were introduced: one of them highlights relevant information through sparse estimations of the densities (SHSED) while the other reveals information via an adaptive computation of similarities in an RHKS (AHSED). To test the introduced approach, two statistical inference tasks were studied: a Poisson mixture model and a nonlinear ecological dynamic system concerning a scaled Ricker map. Attained results demonstrated how the proposed SHSED-based ABC outperforms other state-of-the-art ABC-based inference approaches, including the well-known ABC rejection. In synthesis, the posterior quality estimation can be improved when enhanced distances that reveal relevant information from simulations and observations are introduced in the context of ABC.

Supervised inference scenarios. A novel automatic enhancement of the well-known ABC algorithm devoted to Bayesian inference was developed, called AML-ABC. In particular, a Metric Learning approach based on a CKA methodology to quantify the matching between parameter and simulation spaces was introduced. Particularly, a Mahalanobis distance learned through CKA and graph theory is employed to reveal local relationships among parameter and simulation samples. Notably, the AML-ABC does not require the tuning of any free parameter. Obtained results on a synthetic data-set and a real-world ecological system show how the introduced AML-ABC is a

competitive approach compared to other non-automatic state-of-the-art ABC methods. In conclusion, it is possible to obtain an automatic version of an ABC approach when additional criteria are introduced such that optimization routines lead to a suitable selection of free parameters, preventing expensive tuning procedures like grid search or cross-validation.

5.2 Future work

Concerning the unsupervised inference setting, future work includes the development of an automatic selection of the RKHS regarding the input data dynamics (optimal selection of the characteristic kernel). Moreover, applying ABC approaches using HSE for multivariate data is a research line of interest. On the other hand, regarding the proposed inference approach for supervised scenarios, some potential lines of research include the extension of AML-ABC for other multi-dimensional applications, the development of an adaptive selection of the number of nearest neighbors to calculate the distance between observed and simulated data, besides the median criterion, and the inclusion of other dissimilarity measures, besides the Mahalanobis distance, to deal with complex and/or noisy data.

References

[1] Alabiso, C. and Weiss, I. (2015). *A Primer on Hilbert Space Theory: Linear Spaces, Topological Spaces, Metric Spaces, Normed Spaces, and Topological Groups*, pages 23–74. Springer International Publishing, Cham.

[2] Álvarez-Meza, A., Valencia-Aguirre, J., Daza-Santacoloma, G., and Castellanos-Domínguez, G. (2011). Global and local choice of the number of nearest neighbors in locally linear embedding. *Pattern Recognition Letters*, 32(16):2171–2177.

[3] Alvarez-Meza, A. M., Orozco-Gutierrez, A., and Castellanos-Dominguez, G. (2017). Kernel-based relevance analysis with enhanced interpretability for detection of brain activity patterns. *Frontiers in neuroscience*, 11:550.

[4] Beaumont, M. A., Cornuet, J.-M., Marin, J.-M., and Robert, C. P. (2009). Adaptive approximate bayesian computation. *Biometrika*, 96(4):983–990.

[5] Bickel, P. J., Klaassen, C. A., Bickel, P. J., Ritov, Y., Klaassen, J., Wellner, J. A., and Ritov, Y. (1993). *Efficient and adaptive estimation for semiparametric models*. Johns Hopkins University Press Baltimore.

[6] Box, G. E. and Tiao, G. C. (2011). *Bayesian inference in statistical analysis*, volume 40. John Wiley & Sons.

[7] Casella, G. and Berger, R. L. (2002). *Statistical inference*, volume 2. Duxbury Pacific Grove, CA.

[8] Chen, S., Hong, X., and Harris, C. J. (2008). An orthogonal forward regression technique for sparse kernel density estimation. *Neurocomputing*, 71(4-6):931–943.

[9] Cortes, C., Mohri, M., and Rostamizadeh, A. (2012). Algorithms for learning kernels based on centered alignment. *JMLR*, 13(Mar):795–828.

[10] Creel, M. and Kristensen, D. (2011). Indirect likelihood inference. *Working paper, Barcelona Graduate School of Econimics*.

[11] Csilléry, K., Blum, M. G., Gaggiotti, O. E., and François, O. (2010). Approximate bayesian computation (abc) in practice. *Trends in ecology & evolution*, 25(7):410–418.

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

[13] Fan, Y., Meikle, S. R., Angelis, G., and Sitek, A. (2018). Abc in nuclear imaging. *Handbook of Approximate Bayesian Computation*.

[14] Fasiolo, M. and Wood, S. N. (2015). Approximate methods for dynamic ecological models. *arXiv preprint arXiv:1511.02644*.

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

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

[17] Golchi, S. and Campbell, D. A. (2016). Sequentially constrained monte carlo. *Computational Statistics & Data Analysis*, 97:98–113.

[18] Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773.

[19] Gutmann, M. U. and Corander, J. (2015). Bayesian optimization for likelihood-free inference of simulator-based statistical models. *Journal of Machine Learning Research*.

[20] Holden, P. B., Edwards, N. R., Hensman, J., and Wilkinson, R. D. (2018). Abc for climate: dealing with expensive simulators. *Handbook of Approximate Bayesian Computation*, pages 569–95.

[21] Hong, X. and Chen, S. (2013). A fast algorithm for sparse probability density function construction. In *Digital Signal Processing (DSP), 2013 18th International Conference on*, pages 1–6. IEEE.

[22] Joyce, P. and Marjoram, P. (2008). Approximately sufficient statistics and bayesian computation. *Statistical applications in genetics and molecular biology*, 7(1).

[23] Liepe, J., Kirk, P., Filippi, S., Toni, T., Barnes, C. P., and Stumpf, M. P. (2014). A framework for parameter estimation and model selection from experimental data in systems biology using approximate bayesian computation. *Nature protocols*, 9(2):439.

[24] Liu, W., Principe, J. C., and Haykin, S. (2011). *Kernel adaptive filtering: a comprehensive introduction*, volume 57. John Wiley & Sons.

[25] Marjoram, P., Molitor, J., Plagnol, V., and Tavaré, S. (2003). Markov chain monte carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 100(26):15324–15328.

[26] Meeds, E. and Welling, M. (2014). Gps-abc: Gaussian process surrogate approximate bayesian computation. *arXiv preprint arXiv:1401.2838*.

[27] Meeker, W. Q., Hahn, G. J., and Escobar, L. A. (2017). *Statistical Intervals: A Guide for Practitioners and Researchers*, volume 541. John Wiley & Sons.

[28] Mitrovic, J. et al. (2016). Dr-abc: Approximate bayesian computation with kernel-based distribution regression. In *International Conference on Machine Learning - Volume 48*, ICML'16, pages 1482–1491.

[29] Nakagome, S., Fukumizu, K., and Mano, S. (2013). Kernel approximate bayesian computation in population genetic inferences. *Statistical applications in genetics and molecular biology*, 12(6):667–678.

[30] Park, M., Jitkrittum, W., and Sejdinovic, D. (2016). K2-abc: Approximate bayesian computation with kernel embeddings. In *Artificial Intelligence and Statistics*, pages 398–407.

[31] Prangle, D. et al. (2017). Adapting the abc distance function. *Bayesian Analysis*, 12(1):289–309.

[32] Principe, J. C. (2010). *Information theoretic learning: Renyi's entropy and kernel perspectives*. Springer Science & Business Media.

[33] Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A., and Feldman, M. W. (1999). Population growth of human y chromosomes: a study of y chromosome microsatellites. *Molecular biology and evolution*, 16(12):1791–1798.

[34] Rencher, A. C. (2003). *Methods of multivariate analysis*, volume 492. John Wiley & Sons.

[35] Russell, S. J. and Norvig, P. (2016). *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited,.

[36] Shimazaki, H. and Shinomoto, S. (2010). Kernel bandwidth optimization in spike rate estimation. *Journal of computational neuroscience*, 29(1-2):171–182.

[37] Silverman, B. W. (1986). *Density estimation for statistics and data analysis*, volume 26. CRC press.

[38] Sisson, S. A., Fan, Y., and Beaumont, M. (2018). *Handbook of Approximate Bayesian Computation*. Chapman and Hall/CRC.

[39] Sisson, S. A., Fan, Y., and Tanaka, M. M. (2009). Correction: Sequential monte carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 106(39):16889–16889.

[40] Smola, A., Gretton, A., Song, L., and Schölkopf, B. (2007). A hilbert space embedding for distributions. In *International Conference on Algorithmic Learning Theory*, pages 13–31. Springer.

[41] Sriperumbudur, B. K., Gretton, A., Fukumizu, K., Schölkopf, B., and Lanckriet, G. R. (2010). Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11(Apr):1517–1561.

[42] Toni, T., Welch, D., Strelkowa, N., Ipsen, A., and Stumpf, M. P. (2009). Approximate bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface*, 6(31):187–202.

[43] Turner, B. M. and Van Zandt, T. (2012). A tutorial on approximate bayesian computation. *Journal of Mathematical Psychology*, 56(2):69–85.

[44] Viallefont, V., Richardson, S., and Green, P. J. (2002). Bayesian analysis of poisson mixtures. *Journal of nonparametric statistics*, 14(1-2):181–202.

[45] Wasserman, L. (2013). *All of statistics: a concise course in statistical inference*. Springer Science & Business Media.

[46] Wilkinson, R. D. (2013). Approximate bayesian computation (abc) gives exact results under the assumption of model error. *Statistical applications in genetics and molecular biology*, 12(2):129–141.

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

[48] Zhang, Q., Filippi, S., Gretton, A., and Sejdinovic, D. (2018). Large-scale kernel methods for independence testing. *Statistics and Computing*, 28(1):113–130.

[49] Zougab, N., Adjabi, S., and Kokonendji, C. C. (2014). Bayesian estimation of adaptive bandwidth matrices in multivariate kernel density estimation. *Computational Statistics & Data Analysis*, 75:28–38.

[50] Zuluaga, C. D., Valencia, E. A., Álvarez, M. A., and Orozco, Á. A. (2015). A parzen-based distance between probability measures as an alternative of summary statistics in approximate bayesian computation. In *International Conference on Image Analysis and Processing*, pages 50–61. Springer.

Appendix A

Derivation of the general distance over distributions in an RHKS

Let $\{x_i \in \mathbb{R}^d\}_{i=1}^{Q_x} \backsim P_X$ and $\{y_i \in \mathbb{R}^d\}_{j=1}^{Q_y} \backsim P_Y$ be a pair of independent and identically distributed sample sets drawn from probability distributions P_X and P_Y, respectively. Let the associated probability density function $f(x)$ and $g(y)$ have the following forms:

$$f(x) = \sum_{i=1}^{Q_x} \alpha_i K_{H_i^p}(x, x_i); \quad \sum_{m=1}^{Q_x} \alpha_m = 1, \tag{A.1}$$

$$g(y) = \sum_{j=1}^{Q_y} \beta_j K_{H_j^q}(y, y_j); \quad \sum_{m=1}^{Q_y} \beta_m = 1, \tag{A.2}$$

where $\{\alpha_i \in [0,1]\}_{i=1}^{Q_x}$ and $\{\beta_j \in [0,1]\}_{j=1}^{Q_y}$ are representation weights. Moreover, $K_H(\cdot,\cdot)$ stands for the multivariate Gaussian kernel with covariance matrix $H \in \mathbb{R}^{d\times d}$ as:

$$K_{H_k}(z, z') = \frac{1}{(2\pi)^{d/2}|H|^{1/2}} \exp\left(-\frac{(z-z')^\top H^{-1}(z-z')}{2}\right) \tag{A.3}$$

Using a multivariate Gaussian kernel with covariance matrix H_K as characteristic kernel and substituting 3.1 and 3.2 in equation 2.11 we obtain:

$$
\begin{aligned}
d^2_{\mathcal{H}}(P_X, P_Y) = &\sum_{i=1}^{Q_x}\sum_{j=1}^{Q_x}\alpha_i\alpha_j \int\int K_{H_K}(x,x')K_{H_i^p}(x,x_i)K_{H_j^p}(x',x_j)dxdx' \\
&- 2\sum_{i=1}^{Q_x}\sum_{j=1}^{Q_y}\alpha_i\beta_j \int\int K_{H_K}(x,y)K_{H_i^p}(x,x_i)K_{H_j^q}(y,y_j)dxdy \\
&+ \sum_{i=1}^{Q_y}\sum_{j=1}^{Q_y}\beta_i\beta_j \int\int K_{H_K}(y,y')K_{H_i^q}(y,y_i)K_{H_j^q}(y',y_j)dydy'. \qquad \text{(A.4)}
\end{aligned}
$$

From the fact that the integral of the product of two Gaussians is exactly evaluated as the value of the Gaussian computed at the difference of the arguments and whose covariance is the sum of the covariance of the two original Gaussian functions, because the Gaussian maintains the functional form under convolution [32] (page 75), it follows that:

$$
\begin{aligned}
d^2_{\mathcal{H}}(P_X, P_Y) = &\sum_{i=1}^{Q_x}\sum_{j=1}^{Q_x}\alpha_i\alpha_j \int K_{H_K+H_i^p}(x',x_i)K_{H_j^p}(x',x_j)dx' \\
&- 2\sum_{i=1}^{Q_x}\sum_{j=1}^{Q_y}\alpha_i\beta_j \int K_{H_K+H_i^p}(y,xi)K_{H_j^q}(y,y_j)dy \\
&+ \sum_{i=1}^{Q_y}\sum_{j=1}^{Q_y}\beta_i\beta_j \int K_{H_K+H_i^q}(y,y_i)K_{H_j^q}(y',y_j)dy'. \qquad \text{(A.5)}
\end{aligned}
$$

Finally, applying again the convolution properties of the Gaussian in equation A.5:

$$
\begin{aligned}
d^2_{\mathcal{H}}(P_{X_n}, P_Y) = &\sum_{i=1}^{Q_x}\sum_{j=1}^{Q_x}\alpha_i\alpha_j K_{H_X}(x_i,x_j) - 2\sum_{i=1}^{Q_x}\sum_{j=1}^{Q_y}\alpha_i\beta_j K_{H_{XY}}(x_i,y_j) \\
&+ \sum_{i=1}^{Q_y}\sum_{j=1}^{Q_y}\beta_i\beta_j K_{H_Y}(y_i,y_j). \qquad \text{(A.6)}
\end{aligned}
$$

where

$$
H_X{=}H_K{+}H_i^p{+}H_j^p, \quad H_{XY}{=}H_K{+}H_i^p{+}H_j^q, \quad H_Y{=}H_K{+}H_i^q{+}H_j^q. \quad \textit{Q.E.D.}
$$

Appendix B

Publications

The development of this research has lead to the following publications:

- González-Vanegas W., Alvarez-Meza A., Orozco-Gutierrez Á. (2018). *Sparse Hilbert Embedding-Based Statistical Inference of Stochastic Ecological Systems.* In: Mendoza M., Velastín S. (eds). Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. CIARP 2017. Lecture Notes in Computer Science, vol 10657. pp. 255-262. Springer, Cham

- González-Vanegas W., Alvarez-Meza A., Orozco-Gutierrez Á. (2019). *An Automatic Approximate Bayesian Computation Approach Using Metric Learning.* In: Vera-Rodriguez Ruben., Fierrez Julian., Morales Aythami. (eds). Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. CIARP 2018. Lecture Notes in Computer Science, vol 11401. Springer, Cham

- González-Vanegas W., Alvarez-Meza A., Orozco-Gutierrez Á. (2019). *AKL-ABC: An automatic approximate Bayesian computation approach based on kernel learning.* Submitted on March 2019 to the journal: Statistics, Taylor & Francis.