Journal of Machine Learning Research 1 (2012) 1-48 Submitted 4/00; Published 10/00

Kernel Bayes' Rule: Bayesian Inference with Positive Definite Kernels

Kenji Fukumizu FUKUMIZU@ISM.AC.JP
The Institute of Statistical Mathematics
Tachikawa, Tokyo 190-8562 Japan

Le Song LSONG@CC.GATECH.EDU
College of Computing
Georgia Institute of Technology
266 Ferst Drive, Atlanta, GA 30332-0765 USA

Arthur Gretton ARTHUR.GRETTON@GOOGLEMAIL.COM
Gatsby Computational Neuroscience Unit
University College London
Alexandra House, 17 Queen Square, London - WC1N 3AR UK

Editor:

Abstract

A kernel method for realizing Bayes' rule is proposed, based on representations of probabilities in reproducing kernel Hilbert spaces. Probabilities are uniquely characterized by the mean of the canonical map to the RKHS. The prior and conditional probabilities are expressed in terms of RKHS functions of an empirical sample: no explicit parametric model is needed for these quantities. The posterior is likewise an RKHS mean of a weighted sample. The estimator for the expectation of a function of the posterior is derived, and rates of consistency are shown. Some representative applications of the kernel Bayes' rule are presented, including Baysian computation without likelihood and filtering with a nonparametric state-space model.

Keywords: Bayes' Rule, Positive Definite Kernel, Reproducing Kernel Hilbert Space

1. Introduction

Kernel methods have long provided powerful tools for generalizing linear statistical approaches to nonlinear settings, through an embedding of the sample to a high dimensional feature space, namely a reproducing kernel Hilbert space (RKHS) (Hofmann et al., 2008; Schölkopf and Smola, 2002). Examples include support vector machines, kernel PCA, and kernel CCA, among others. In these cases, data are mapped via a canonical feature map to a reproducing kernel Hilbert space (of high or even infinite dimension), in which the linear operations that define the algorithms are implemented. The inner product between feature mappings need never be computed explicitly, but is given by a positive definite kernel function unique to the RKHS: this permits efficient computation without the need to deal explicitly with the feature representation.

The mappings of individual points to a feature space may be generalized to mappings of probability measures (e.g. Berlinet and Thomas-Agnan, 2004, Chapter 4). We call such

mappings the *kernel means* of the underlying random variables. With an appropriate choice of positive definite kernel, the kernel mean on the RKHS uniquely determines the distribution of the variable (Fukumizu et al., 2004, 2009a; Sriperumbudur et al., 2010), and statistical inference problems on distributions can be solved via operations on the kernel means. Applications of this approach include homogeneity testing (Gretton et al., 2007, 2009a), where the empirical means on the RKHS are compared directly, and independence testing (Gretton et al., 2008, 2009b), where the mean of the joint distribution on the feature space is compared with that of the product of the marginals. Representations of conditional dependence may also be defined in RKHS, and have been used in conditional independence tests (Fukumizu et al., 2008).

In this paper, we propose a novel, nonparametric approach to Bayesian inference, making use of kernel means of probabilities. In applying Bayes' rule, we compute the posterior probability of x in $\mathcal{X}$ given observation y in $\mathcal{Y}$;

$$q(x|y) = \frac{p(y|x)\pi(x)}{q_{\mathcal{Y}}(y)}, \tag{1}$$

where $\pi(x)$ and $p(y|x)$ are the density functions of the prior and the likelihood of y given x, respectively, with respective base measures $\nu_{\mathcal{X}}$ and $\nu_{\mathcal{Y}}$, and the normalization factor $q_{\mathcal{Y}}(\mathrm{y})$ is given by

$$q_{\mathcal{Y}}(y) = \int p(y|x)\pi(x)d\nu_{\mathcal{X}}(x). \tag{2}$$

Our main result is a nonparametric estimate of the kernel mean posterior, given kernel mean representations of the prior and likelihood.

A valuable property of the kernel Bayes' rule is that the kernel posterior mean is estimated nonparametrically from data; specifically, the prior and the likelihood are represented in the form of samples from the prior and the joint probability that gives the likelihood, respectively. This confers an important benefit: we can still perform Bayesian inference by making sufficient observations on the system, even in the absence of a specific parametric model of the relation between variables. More generally, if we can sample from the model, we do not require explicit density functions for inference. Such situations are typically seen when the prior or likelihood is given by a random process: Approximate Bayesian Computation (Marjoram et al., 2003; Sisson et al., 2007; Tavaré et al., 1997) is widely applied in population genetics, where the likelihood is given by a branching process, and nonparametric Bayesian inference (Müller and Quintana, 2004) often uses a process prior with sampling methods. Alternatively, a parametric model may be known, however it might be of sufficient complexity to require Markov chain Monte Carlo or sequential Monte Carlo for inference. The present kernel approach provides an alternative strategy for Bayesian inference in these settings. We demonstrate rates of consistency for our posterior kernel mean estimate, and for the expectation of functions computed using this estimate.

An alternative to the kernel mean representation would be to use nonparametric density estimates for the posterior. Classical approaches include kernel density estimation (KDE) or distribution estimation on a finite partition of the domain. These methods are known to perform poorly on high dimensional data, however. By contrast, the proposed kernel mean representation is defined as an integral or moment of the distribution, taking the form of a function in an RKHS. Thus, it is more akin to the characteristic function approach (see e.g.

Kankainen and Ushakov, 1998) to representing probabilities. A well conditioned empirical estimate of the characteristic function can be difficult to obtain, especially for conditional probabilities. By contrast, the kernel mean has a straightforward empirical estimate, and conditioning and marginalization can be implemented easily, at a reasonable computational cost.

The proposed method of realizing Bayes' rule is an extension of the approach used in Song et al. (2009) for state-space models. In this earlier work, a heuristic approximation was used, where the kernel mean of the new hidden state was estimated by adding kernel mean estimates from the previous hidden state and the observation. Another relevant work is the belief propagation approach in Song et al. (2010a, 2011), which covers the simpler case of a uniform prior.

This paper is organized as follows. We begin in Section 2 with a review of RKHS terminology and of kernel mean embeddings. In Section 3, we derive an expression for Bayes' rule in terms of kernel means, and provide consistency guarantees. We apply the kernel Bayes' rule in Section 4 to various inference problems, with numerical results and comparisons with existing methods in Section 5. Our proofs are contained in Section 6 (including proofs of the consistency results of Section 3).

2. Preliminaries: positive definite kernel and probabilities

Throughout this paper, all Hilbert spaces are assumed to be separable. For an operator A on a Hilbert space, the range is denoted by $\mathcal{R}(A)$. The linear hull of a subset S in a vector space is denoted by Span$S$.

We begin with a review of positive definite kernels, and of statistics on the associated reproducing kernel Hilbert spaces (Aronszajn, 1950; Berlinet and Thomas-Agnan, 2004; Fukumizu et al., 2004, 2009a). Given a set Ω, a ($\mathbb{R}$-valued) positive definite kernel k on Ω is a symmetric kernel $k : \Omega \times \Omega \to \mathbb{R}$ such that $\sum_{i,j=1}^n c_i c_j k(x_i, x_j) \geq 0$ for arbitrary number of points $x_1, \ldots, x_n$ in Ω and real numbers $c_1, \ldots, c_n$. The matrix $(k(x_i, x_j))_{i,j=1}^n$ is called a Gram matrix. It is known by the Moore-Aronszajn theorem (Aronszajn, 1950) that a positive definite kernel on Ω uniquely defines a Hilbert space $\mathcal{H}$ consisting of functions on Ω such that (i) $k(\cdot, x) \in \mathcal{H}$ for any $x \in \Omega$, (ii) Span$\{k(\cdot, x) \mid x \in \Omega\}$ is dense in $\mathcal{H}$, and (iii) $\langle f, k(\cdot, x)\rangle = f(x)$ for any $x \in \Omega$ and $f \in \mathcal{H}$ (the reproducing property), where $\langle \cdot, \cdot \rangle$ is the inner product of $\mathcal{H}$. The Hilbert space $\mathcal{H}$ is called the *reproducing kernel Hilbert space* (RKHS) associated with k, since the function $k_x = k(\ , x)$ serves as the reproducing kernel $\langle f, k_x\rangle = f(x)$ for $f \in \mathcal{H}$.

A positive definite kernel on Ω is said to be *bounded* if there is $M > 0$ such that $k(x, x) \leq M$ for any $x \in \Omega$.

Let $(\mathcal{X}, \mathcal{B}_\mathcal{X})$ be a measurable space, X be a random variable taking values in $\mathcal{X}$ with distribution P_X, and k be a measurable positive definite kernel on $\mathcal{X}$ such that $E[\sqrt{k(X, X)}] < \infty$. The associated RKHS is denoted by $\mathcal{H}$. The *kernel mean* m_X^k (also written $m_{P_X}^k$) of X on the RKHS $\mathcal{H}$ is defined by the mean of the $\mathcal{H}$-valued random variable $k(\cdot, X)$. The existence of the kernel mean is guaranteed by $E[\|k(\cdot, X)\|] = E[\sqrt{k(X, X)}] < \infty$. We usually write m_X for m_X^k for simplicity, where there is no ambiguity. By the reproducing property, the kernel mean satisfies the relation

$$\langle f, m_X\rangle = E[f(X)] \tag{3}$$

for any $f \in \mathcal{H}$. Plugging $f = k(\cdot, u)$ into this relation derives

$$m_X(u) = E[k(u, X)] = \int k(u, \tilde{x}) dP_X(\tilde{x}), \tag{4}$$

which shows the explicit functional form. The kernel mean m_X is also denoted by m_{P_X}, as it depends only on the distribution P_X with k fixed.

Let $(\mathcal{X}, \mathcal{B}_\mathcal{X})$ and $(\mathcal{Y}, \mathcal{B}_\mathcal{Y})$ be measurable spaces, (X, Y) be a random variable on $\mathcal{X} \times \mathcal{Y}$ with distribution P, and $k_\mathcal{X}$ and $k_\mathcal{Y}$ be measurable positive definite kernels with respective RKHS $\mathcal{H}_\mathcal{X}$ and $\mathcal{H}_\mathcal{Y}$ such that $E[k_\mathcal{X}(X, X)] < \infty$ and $E[k_\mathcal{Y}(Y, Y)] < \infty$. The (uncentered) *covariance operator* $C_{YX} : \mathcal{H}_\mathcal{X} \to \mathcal{H}_\mathcal{Y}$ is defined as the linear operator that satisfies

$$\langle g, C_{YX} f \rangle_{\mathcal{H}_\mathcal{Y}} = E[f(X) g(Y)]$$

for all $f \in \mathcal{H}_\mathcal{X}, g \in \mathcal{H}_\mathcal{Y}$. This operator C_{YX} can be identified with $m_{(YX)}$ in the product space $\mathcal{H}_\mathcal{Y} \otimes \mathcal{H}_\mathcal{X}$, which is given by the product kernel $k_\mathcal{Y} k_\mathcal{X}$ on $\mathcal{Y} \times \mathcal{X}$ (Aronszajn, 1950), by the standard identification between the linear maps and the tensor product. We also define C_{XX} for the operator on $\mathcal{H}_\mathcal{X}$ that satisfies $\langle f_2, C_{XX} f_1 \rangle = E[f_2(X) f_1(X)]$ for any $f_1, f_2 \in \mathcal{H}_\mathcal{X}$. Similarly to Eq. (4), the explicit integral expressions for C_{YX} and C_{XX} are given by

$$(C_{YX} f)(y) = \int k_\mathcal{Y}(y, \tilde{y}) f(\tilde{x}) dP(\tilde{x}, \tilde{y}), \quad (C_{XX} f)(x) = \int k_\mathcal{X}(x, \tilde{x}) f(\tilde{x}) dP_X(\tilde{x}), \tag{5}$$

respectively.

An important notion in statistical inference with positive definite kernels is the characteristic property. A bounded measurable positive definite kernel k on a measurable space $(\Omega, \mathcal{B})$ is called *characteristic* if the mapping from a probability Q on $(\Omega, \mathcal{B})$ to the kernel mean $m_Q^k \in \mathcal{H}$ is injective (Fukumizu et al., 2009a; Sriperumbudur et al., 2010). This is equivalent to assuming that $E_{X \sim P}[k(\cdot, X)] = E_{X' \sim Q}[k(\cdot, X')]$ implies $P = Q$: probabilities are uniquely determined by their kernel means on the associated RKHS. With this property, problems of statistical inference can be cast as inference on the kernel means. A popular example of a characteristic kernel defined on Euclidean space is the Gaussian RBF kernel $k(x, y) = \exp(-\|x - y\|^2/(2\sigma^2))$. It is known that a bounded measurable positive definite kernel on a measurable space $(\Omega, \mathcal{B})$ with corresponding RKHS $\mathcal{H}$ is characteristic if and only if $\mathcal{H} + \mathbb{R}$ is dense in $L^2(P)$ for arbitrary probability P on $(\Omega, \mathcal{B})$, where $\mathcal{H} + \mathbb{R}$ is the direct sum of two RKHSs $\mathcal{H}$ and $\mathbb{R}$ Aronszajn (1950). This implies that the RKHS defined by a characteristic kernel is rich enough to be dense in L^2 space up to the constant functions. Other useful conditions for a kernel to be characteristic can be found in Sriperumbudur et al. (2010), Fukumizu et al. (2009b), and Sriperumbudur et al. (2011).

Throughout this paper, when positive definite kernels on a measurable space are discussed, the following assumption is made:

(K) Positive definite kernels are bounded and measurable.

Under this assumption, the mean and covariance always exist with arbitrary probabilities.

Given i.i.d. sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ with law P, the empirical estimator of the kernel mean and covariance operator are given straightforwardly by

$$\widehat{m}_X^{(n)} = \frac{1}{n} \sum_{i=1}^n k_\mathcal{X}(\cdot, X_i), \qquad \widehat{C}_{YX}^{(n)} = \frac{1}{n} \sum_{i=1}^n k_\mathcal{Y}(\cdot, Y_i) \otimes k_\mathcal{X}(\cdot, X_i),$$

where $\widehat{C}_{YX}^{(n)}$ is written in tensor form. It is known that these estimators are $\sqrt{n}$-consistent in appropriate norms, and $\sqrt{n}(\widehat{m}_X^{(n)} - m_X)$ converges to a Gaussian process on $\mathcal{H}_\mathcal{X}$ (Berlinet and Thomas-Agnan, 2004, Sec. 9.1). While we may use non-i.i.d. samples for numerical examples in Section 5, in our theoretical analysis we always assume i.i.d. samples for simplicity.

3. Kernel expression of Bayes' rule

3.1 Kernel Bayes' rule

Let $(\mathcal{X}, \mathcal{B}_\mathcal{X})$ and $(\mathcal{Y}, \mathcal{B}_\mathcal{Y})$ be measurable spaces, (X, Y) be a random variable on $\mathcal{X} \times \mathcal{Y}$ with distribution P, and $k_\mathcal{X}$ and $k_\mathcal{Y}$ be positive definite kernels on $\mathcal{X}$ and $\mathcal{Y}$, respectively, with respective RKHS $\mathcal{H}_\mathcal{X}$ and $\mathcal{H}_\mathcal{Y}$. Let Π be a probability on $(\mathcal{X}, \mathcal{B}_\mathcal{X})$, which serves as a *prior* distribution. For each $x \in \mathcal{X}$, define a probability $P_{Y|x}$ on $(\mathcal{Y}, \mathcal{B}_\mathcal{Y})$ by $P_{Y|x}(B) = E[I_B(Y)|X = x]$, where I_B is the index function of a measurable set $B \in \mathcal{B}_\mathcal{Y}$. The prior Π and the family $\{P_{Y|x} \mid x \in \mathcal{X}\}$ defines the joint distribution Q on $\mathcal{X} \times \mathcal{Y}$ by

$$Q(A \times B) = \int_A P_{Y|x}(B) d\Pi(x)$$

for any $A \in \mathcal{B}_\mathcal{X}$ and $B \in \mathcal{B}_\mathcal{Y}$, and its marginal distribution $Q_\mathcal{Y}$ by $Q_\mathcal{Y}(B) = Q(\mathcal{X} \times B)$. Throughout this paper, it is assumed that $P_{Y|x}$ and Q are well-defined under some regularity conditions. Let (Z, W) be a random variable on $\mathcal{X} \times \mathcal{Y}$ with distribution Q. It is also assumed that the sigma algebra generated by W includes every point $\{y\}$ $(y \in \mathcal{Y})$. For $y \in \mathcal{Y}$, the *posterior* probability given y is defined by the conditional probability

$$Q_{\mathcal{X}|y}(A) = E[I_A(Z)|W = y] \qquad (A \in \mathcal{B}_\mathcal{X}). \tag{6}$$

If the probability distributions have density functions with respect to measures $\nu_\mathcal{X}$ on $\mathcal{X}$ and $\nu_\mathcal{Y}$ on $\mathcal{Y}$, namely, if the p.d.f. of P and Π are given by $p(x, y)$ and $\pi(x)$, respectively, Eq. (6) is reduced to the well known form Eq. (1).

The goal of this subsection is to derive an estimator of the kernel mean of posterior $m_{Q_{\mathcal{X}|y}}$. The following theorem is fundamental to discuss conditional probabilities with positive definite kernels.

Theorem 3.1 (Fukumizu et al. (2004)) *If $E[g(Y)|X = \cdot] \in \mathcal{H}_\mathcal{X}$ holds for $g \in \mathcal{H}_\mathcal{Y}$, then*

$$C_{XX} E[g(Y)|X = \cdot] = C_{XY} g.$$

If C_{XX} is injective, i.e., if the function $f \in \mathcal{H}_\mathcal{X}$ with $C_{XX} f = C_{XY} g$ is unique, the above relation can be expressed as

$$E[g(Y)|X = \cdot] = {C_{XX}}^{-1} C_{XY} g. \tag{7}$$

Noting $\langle C_{XX} f, f \rangle = E[f(X)^2]$, it is easy to see that C_{XX} is injective, if $\mathcal{X}$ is a topological space, $k_\mathcal{X}$ is a continuous kernel, and $\mathrm{Supp}(P_X) = \mathcal{X}$, where $\mathrm{Supp}(P_X)$ is the support of P_X.

From Theorem 3.1, we have the following result, which expresses the kernel mean of $Q_\mathcal{Y}$.

Theorem 3.2 (Song et al. (2009), Eq. 6) *Let m_Π and $m_{Q_\mathcal{Y}}$ be the kernel means of Π in $\mathcal{H}_\mathcal{X}$ and $Q_\mathcal{Y}$ in $\mathcal{H}_\mathcal{Y}$, respectively. If C_{XX} is injective, $m_\Pi \in \mathcal{R}(C_{XX})$, and $E[g(Y)|X = \cdot] \in \mathcal{H}_\mathcal{X}$ for any $g \in \mathcal{H}_\mathcal{Y}$, then*

$$m_{Q_\mathcal{Y}} = C_{YX} {C_{XX}}^{-1} m_\Pi. \tag{8}$$

Proof Take $f \in \mathcal{H}_\mathcal{X}$ such that $f = C_{XX}^{-1} m_\Pi$. For any $g \in \mathcal{H}_\mathcal{Y}$, $\langle C_{YX} f, g \rangle = \langle f, C_{XY} g \rangle = \langle f, C_{XX} E[g(Y)|X = \cdot] \rangle = \langle C_{XX} f, E[g(Y)|X = \cdot] \rangle = \langle m_\Pi, E[g(Y)|X = \cdot] \rangle = \langle m_{Q_\mathcal{Y}}, g \rangle$, which implies $C_{YX} f = m_{Q_\mathcal{Y}}$. ■

As discussed in Song et al. (2009), the operator $C_{YX} C_{XX}^{-1}$ can be regarded as the kernel expression of the conditional probability $P_{Y|x}$ or $p(y|x)$.

Note, however, that the assumption $E[g(Y)|X = \cdot] \in \mathcal{H}_\mathcal{X}$ may not hold in general; we can easily give counterexamples in the case of Gaussian kernels[1]. In the following, we nonetheless derive a population expression of Bayes' rule under this strong assumption, use it as a prototype for defining an empirical estimator, and prove its consistency.

Eq. (8) has a simple interpretation if the probabilities have density functions and $\pi(x)/p_X(x)$ is in $\mathcal{H}_\mathcal{X}$, where p_X is the density function of the marginal P_X. From Eq. (4) we have $m_\Pi(x) = \int k_\mathcal{X}(x, \tilde{x}) \pi(\tilde{x}) d\nu_\mathcal{X}(\tilde{x}) = \int k_\mathcal{X}(x, \tilde{x}) (\pi(\tilde{x})/p_X(\tilde{x})) dP_X(\tilde{x})$, which implies $C_{XX}^{-1} m_\Pi = \pi/p_X$ from Eq. (5). Thus Eq. (8) is an operator expression of the obvious relation

$$\int \int k_\mathcal{Y}(y, \tilde{y}) p(\tilde{y}|\tilde{x}) \pi(\tilde{x}) d\nu_\mathcal{X}(\tilde{x}) d\nu_\mathcal{Y}(\tilde{y}) = \int k_\mathcal{Y}(y, \tilde{y}) (\pi(\tilde{x})/p_X(\tilde{x})) dP(\tilde{x}, \tilde{y}).$$

In deriving kernel realization of Bayes' rule, we will use the following tensor representation of the joint probability Q, based on Theorem 3.2:

$$m_Q = C_{(YX)X} C_{XX}^{-1} m_\Pi \in \mathcal{H}_\mathcal{Y} \otimes \mathcal{H}_\mathcal{X}. \tag{9}$$

In the above equation, the covariance operator $C_{(YX)X} : \mathcal{H}_\mathcal{X} \to \mathcal{H}_\mathcal{Y} \otimes \mathcal{H}_\mathcal{X}$ is defined by the random variable $((Y, X), X)$ taking values on $(\mathcal{Y} \times \mathcal{X}) \times \mathcal{X}$.

In many applications of Bayesian inference, the probability conditioned on a particular value should be computed. By plugging the point measure at x into Π in Eq. (8), we have a population expression

$$E[k_\mathcal{Y}(\cdot, Y)|X = x] = C_{YX} {C_{XX}}^{-1} k_\mathcal{X}(\cdot, x), \tag{10}$$

which has been considered in Song et al. (2009, 2010a) as the kernel mean of the conditional probability. It must be noted that for this case the assumption $m_\Pi = k(\cdot, x) \in \mathcal{R}(C_{XX})$ in Theorem 3.2 may not hold in general[2]. We will show in Theorem 6.1, however, that under some conditions a regularized empirical estimator based on Eq. (10) is a consistent estimator of $E[k_\mathcal{Y}(\cdot, Y)|X = x]$.

1. Suppose that $\mathcal{H}_\mathcal{X}$ and $\mathcal{H}_\mathcal{Y}$ are given by Gaussian kernel, and that X and Y are independent. Then, $E[g(Y)|X = x]$ is a constant function of x, which is known not to be included in a RKHS given by a Gaussian kernel (Steinwart and Christmann, 2008, Corollary 4.44).
2. Suppose $C_{XX} h_x = k_\mathcal{X}(\cdot, x)$ were to hold for some $h_x \in \mathcal{H}_\mathcal{X}$. Taking the inner product with $k_\mathcal{X}(\cdot, \tilde{x})$ would then imply $k_\mathcal{X}(x, \tilde{x}) = \int h_x(x') k_\mathcal{X}(\tilde{x}, x') dP_X(x')$, which is not possible for many popular kernels, including the Gaussian kernel.

If we replace P by Q and x by y in Eq. (10), we obtain

$$m_{Q_{\mathcal{X}|y}} = E[k_{\mathcal{X}}(\cdot, Z)|W = y] = C_{ZW}C_{WW}^{-1}k_{\mathcal{Y}}(\cdot, y). \tag{11}$$

This is exactly the kernel mean expression of the posterior, and the next step is to provide a way of deriving the covariance operators C_{ZW} and C_{WW}. Recall that the kernel mean $m_Q = m_{(ZW)} \in \mathcal{H}_{\mathcal{X}} \otimes \mathcal{H}_{\mathcal{Y}}$ can be identified with the covariance operator $C_{ZW} : \mathcal{H}_{\mathcal{Y}} \to \mathcal{H}_{\mathcal{X}}$, and $m_{(WW)}$, which is the kernel mean on the product space $\mathcal{H}_{\mathcal{Y}} \otimes \mathcal{H}_{\mathcal{Y}}$, with C_{WW}. Then from Eq. (9) and the similar expression $m_{(WW)} = C_{(YY)X}C_{XX}^{-1}m_\Pi$, we are able to obtain the operators in Eq. (11), and thus the kernel mean of the posterior.

The above argument can be rigorously implemented, if empirical estimators are considered. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be an i.i.d. sample with law P. Since the kernel method needs to express the information of variables in terms of Gram matrices given by data points, we assume that the prior is also expressed in the form of an empirical estimate, and that we have a consistent estimator of m_Π in the form

$$\widehat{m}_\Pi^{(\ell)} = \sum_{j=1}^{\ell} \gamma_j k_{\mathcal{X}}(\cdot, U_j),$$

where $U_1, \ldots, U_\ell$ are points in $\mathcal{X}$ and γ_j are the weights. The data points U_j may or may not be a sample from the prior Π, and negative values are allowed for γ_j. Negative values are observed in successive applications of the kernel Bayes rule, as in the state-space example of Section 4.3. Based on Theorem 3.2, the empirical estimators for $m_{(ZW)}$ and $m_{(WW)}$ are defined respectively by

$$\widehat{m}_{(ZW)} = \widehat{C}_{(YX)X}^{(n)}\big(\widehat{C}_{XX}^{(n)} + \varepsilon_n I\big)^{-1}\widehat{m}_\Pi^{(\ell)}, \quad \widehat{m}_{(WW)} = \widehat{C}_{(YY)X}^{(n)}\big(\widehat{C}_{XX}^{(n)} + \varepsilon_n I\big)^{-1}\widehat{m}_\Pi^{(\ell)},$$

where ε_n is the coefficient of the Tikhonov-type regularization for operator inversion, and I is the identity operator. The empirical estimators $\widehat{C}_{ZW}$ and $\widehat{C}_{WW}$ for C_{ZW} and C_{WW} are identified with $\widehat{m}_{(ZW)}$ and $\widehat{m}_{(WW)}$, respectively. In the following, G_X and G_Y denote the Gram matrices $(k_{\mathcal{X}}(X_i, X_j))$ and $(k_{\mathcal{Y}}(Y_i, Y_j))$, respectively, and I_n is the identity matrix of size n.

Proposition 3.3 *The Gram matrix expressions of $\widehat{C}_{ZW}$ and $\widehat{C}_{WW}$ are given by*

$$\widehat{C}_{ZW} = \sum_{i=1}^{n} \widehat{\mu}_i k_{\mathcal{X}}(\cdot, X_i) \otimes k_{\mathcal{Y}}(\cdot, Y_i) \quad \textit{and} \quad \widehat{C}_{WW} = \sum_{i=1}^{n} \widehat{\mu}_i k_{\mathcal{Y}}(\cdot, Y_i) \otimes k_{\mathcal{Y}}(\cdot, Y_i),$$

respectively, where the common coefficient $\widehat{\mu} \in \mathbb{R}^n$ is

$$\widehat{\mu} = \Big(\frac{1}{n}G_X + \varepsilon_n I_n\Big)^{-1}\widehat{\mathbf{m}}_\Pi, \quad \widehat{\mathbf{m}}_{\Pi,i} = \widehat{m}_\Pi(X_i) = \sum_{j=1}^{\ell} \gamma_j k_{\mathcal{X}}(X_i, U_j). \tag{12}$$

The proof is similar to that of Proposition 3.4 below, and is omitted. The expressions in Proposition 3.3 imply that the probabilities Q and $Q_{\mathcal{Y}}$ are estimated by the weighted samples $\{((X_i, Y_i), \widehat{\mu}_i)\}_{i=1}^n$ and $\{(Y_i, \widehat{\mu}_i)\}_{i=1}^n$, respectively, with common weights. Since the weight

$\widehat{\mu}_i$ may be negative, in applying Eq. (11) the operator inversion in the form $(\widehat{C}_{WW}+\delta_n I)^{-1}$ may be impossible or unstable. We thus use another type of Tikhonov regularization, thus obtaining the estimator

$$\widehat{m}_{Q_{\mathcal{X}|y}} := \widehat{C}_{ZW}\big(\widehat{C}_{WW}^2+\delta_n I\big)^{-1}\widehat{C}_{WW}k_{\mathcal{Y}}(\cdot,y). \tag{13}$$

Proposition 3.4 *For any $y\in\mathcal{Y}$, the Gram matrix expression of $\widehat{m}_{Q_{\mathcal{X}|y}}$ is given by*

$$\widehat{m}_{Q_{\mathcal{X}|y}} = \mathbf{k}_X^T R_{X|Y}\mathbf{k}_Y(y), \qquad R_{X|Y} := \Lambda G_Y((\Lambda G_Y)^2+\delta_n I_n)^{-1}\Lambda, \tag{14}$$

where $\Lambda=\mathrm{diag}(\widehat{\mu})$ is a diagonal matrix with elements $\widehat{\mu}_i$ in Eq. (12), $\mathbf{k}_X=(k_{\mathcal{X}}(\cdot,X_1),\ldots,k_{\mathcal{X}}(\cdot,X_n))^T\in{\mathcal{H}_{\mathcal{X}}}^n$, and $\mathbf{k}_Y=(k_{\mathcal{Y}}(\cdot,Y_1),\ldots,k_{\mathcal{Y}}(\cdot,Y_n))^T\in{\mathcal{H}_{\mathcal{Y}}}^n$.

Proof Let $h=(\widehat{C}_{WW}^2+\delta_n I)^{-1}\widehat{C}_{WW}k_{\mathcal{Y}}(\cdot,y)$, and decompose it as $h=\sum_{i=1}^n\alpha_i k_{\mathcal{Y}}(\cdot,Y_i)+h_\perp=\alpha^T\mathbf{k}_Y+h_\perp$, where $h_\perp$ is orthogonal to $\mathrm{Span}\{k_{\mathcal{Y}}(\cdot,Y_i)\}_{i=1}^n$. Expansion of $(\widehat{C}_{WW}^2+\delta_n I)h=\widehat{C}_{WW}k_{\mathcal{Y}}(\cdot,y)$ gives $\mathbf{k}_Y^T(\Lambda G_Y)^2\alpha+\delta_n\mathbf{k}_Y^T\alpha+\delta_n h_\perp=\mathbf{k}_Y^T\Lambda\mathbf{k}_Y(y)$. Taking the inner product with $k_{\mathcal{Y}}(\cdot,Y_j)$, we have

$$\big((G_Y\Lambda)^2+\delta_n I_n\big)G_Y\alpha = G_Y\Lambda\mathbf{k}_Y(y).$$

The coefficient ρ in $\widehat{m}_{Q_{\mathcal{X}|y}}=\widehat{C}_{ZW}h=\sum_{i=1}^n\rho_i k_{\mathcal{X}}(\cdot,X_i)$ is given by $\rho=\Lambda G_Y\alpha$, and thus

$$\rho=\Lambda\big((G_Y\Lambda)^2+\delta_n I_n\big)^{-1}G_Y\Lambda\mathbf{k}_Y(y)=\Lambda G_Y\big((\Lambda G_Y)^2+\delta_n I_n\big)^{-1}\Lambda\mathbf{k}_Y(y).$$

■

We call Eqs.(13) and (14) the *kernel Bayes' rule* (KBR). The required computations are summarized in Figure 1. The KBR uses a weighted sample to represent the posterior; it is similar in this respect to sampling methods such as importance sampling and sequential Monte Carlo (Doucet et al., 2001). The KBR method, however, does not generate samples of the posterior, but updates the weights of a sample by matrix computation. We will give some experimental comparisons between KBR and sampling methods in Section 5.1.

If our aim is to estimate the expectation of a function $f\in\mathcal{H}_{\mathcal{X}}$ with respect to the posterior, the reproducing property Eq. (3) gives an estimator

$$\langle f,\widehat{m}_{Q_{\mathcal{X}|y}}\rangle_{\mathcal{H}_{\mathcal{X}}} = \mathbf{f}_X^T R_{X|Y}\mathbf{k}_{\mathcal{Y}}(y), \tag{15}$$

where $\mathbf{f}_X=(f(X_1),\ldots,f(X_n))^T\in\mathbb{R}^n$.

3.2 Consistency of the KBR estimator

We now demonstrate the consistency of the KBR estimator in Eq. (15). For the theoretical analysis, it is assumed that the distributions have density functions for simplicity. In the following two theorems, we show only the best rates that can be derived under the assumptions, and defer more detailed discussions and proofs to Section 6. We assume here that the sample size $\ell=\ell_n$ for the prior goes to infinity as the sample size n for the likelihood goes to infinity, and that $\widehat{m}_\Pi^{(\ell_n)}$ is n^α-consistent in RKHS norm.

Input: (i) $\{(X_i, Y_i)\}_{i=1}^n$: sample to express P. (ii) $\{(U_j, \gamma_j)\}_{j=1}^\ell$: weighted sample to express the kernel mean of the prior $\widehat{m}_\Pi$. (iii) ε_n, δ_n: regularization constants.

Computation:

1. Compute Gram matrices $G_X = (k_\mathcal{X}(X_i, X_j))$, $G_Y = (k_\mathcal{Y}(Y_i, Y_j))$, and a vector $\widehat{\mathbf{m}}_\Pi = (\sum_{j=1}^\ell \gamma_j k_\mathcal{X}(X_i, U_j))_{i=1}^n$.
2. Compute $\widehat{\mu} = n(G_X + n\varepsilon_n I_n)^{-1}\widehat{\mathbf{m}}_\Pi$.
3. Compute $R_{X|Y} = \Lambda G_Y((\Lambda G_Y)^2 + \delta_n I_n)^{-1}\Lambda$, where $\Lambda = \mathrm{diag}(\widehat{\mu})$.

Output: $n \times n$ matrix $R_{X|Y}$.
Given conditioning value y, the kernel mean of the posterior $q(x|y)$ is estimated by the weighted sample $\{(X_i, \rho_i)\}_{i=1}^n$ with weight $\rho = R_{X|Y}\mathbf{k}_Y(y)$, where $\mathbf{k}_Y(y) = (k_\mathcal{Y}(Y_i, y))_{i=1}^n$.

Figure 1: Algorithm of Kernel Bayes' Rule

Theorem 3.5 *Let f be a function in $\mathcal{H}_\mathcal{X}$, (Z, W) be a random variable on $\mathcal{X} \times \mathcal{Y}$ such that the distribution is Q with p.d.f. $p(y|x)\pi(x)$, and $\widehat{m}_\Pi^{(\ell_n)}$ be an estimator of m_Π such that $\|\widehat{m}_\Pi^{(\ell_n)} - m_\Pi\|_{\mathcal{H}_\mathcal{X}} = O_p(n^{-\alpha})$ as $n \to \infty$ for some $0 < \alpha \leq 1/2$. Assume that $\pi/p_X \in \mathcal{R}(C_{XX}^{1/2})$, where p_X is the p.d.f. of P_X, and $E[f(Z)|W = \cdot] \in \mathcal{R}(C_{WW}^2)$. For the regularization constants $\varepsilon_n = n^{-\frac{2}{3}\alpha}$ and $\delta_n = n^{-\frac{8}{27}\alpha}$, we have for any $y \in \mathcal{Y}$*

$$\mathbf{f}_X^T R_{X|Y}\mathbf{k}_Y(y) - E[f(Z)|W = y] = O_p(n^{-\frac{8}{27}\alpha}), \quad (n \to \infty),$$

where $\mathbf{f}_X^T R_{X|Y}\mathbf{k}_Y(y)$ is given by Eq. (15).

It is possible to extend the covariance operator C_{WW} to one defined on $L^2(Q_\mathcal{Y})$ by

$$\tilde{C}_{WW}\phi = \int k_\mathcal{Y}(y, w)\phi(w)dQ_\mathcal{Y}(w), \qquad (\phi \in L^2(Q_\mathcal{Y})). \tag{16}$$

If we consider the convergence on average over y, we have a slightly better rate on the consistency of the KBR estimator in $L^2(Q_\mathcal{Y})$.

Theorem 3.6 *Let f be a function in $\mathcal{H}_\mathcal{X}$, (Z, W) be a random vector on $\mathcal{X} \times \mathcal{Y}$ such that the distribution is Q with p.d.f. $p(y|x)\pi(x)$, and $\widehat{m}_\Pi^{(\ell_n)}$ be an estimator of m_Π such that $\|\widehat{m}_\Pi^{(\ell_n)} - m_\Pi\|_{\mathcal{H}_\mathcal{X}} = O_p(n^{-\alpha})$ as $n \to \infty$ for some $0 < \alpha \leq 1/2$. Assume that $\pi/p_X \in \mathcal{R}(C_{XX}^{1/2})$, where p_X is the p.d.f. of P_X, and $E[f(Z)|W = \cdot] \in \mathcal{R}(\tilde{C}_{WW}^2)$. For the regularization constants $\varepsilon_n = n^{-\frac{2}{3}\alpha}$ and $\delta_n = n^{-\frac{1}{3}\alpha}$, we have*

$$\left\|\mathbf{f}_X^T R_{X|Y}\mathbf{k}_Y(W) - E[f(Z)|W]\right\|_{L^2(Q_\mathcal{Y})} = O_p(n^{-\frac{1}{3}\alpha}), \quad (n \to \infty).$$

The condition $\pi/p_X \in \mathcal{R}(C_{XX}^{1/2})$ requires the prior to be sufficiently smooth. If $\widehat{m}_\Pi^{(\ell_n)}$ is a direct empirical mean with an i.i.d. sample of size n from Π, typically $\alpha = 1/2$, with which the theorems imply $n^{4/27}$-consistency for every y, and $n^{1/6}$-consistency in the $L^2(Q_\mathcal{Y})$ sense. While these might seem to be slow rates, the rate of convergence can in practice be much faster than the above theoretical guarantees.

4. Bayesian inference with Kernel Bayes' Rule

4.1 Applications of Kernel Bayes' Rule

In Bayesian inference, we are usually interested in finding a point estimate such as the MAP solution, the expectation of a function under the posterior, or other properties of the distribution. Given that KBR provides a posterior estimate in the form of a kernel mean (which uniquely determines the distribution when a characteristic kernel is used), we now describe how our kernel approach applies to problems in Bayesian inference.

First, we have already seen that a consistent estimator for the expectation of $f \in \mathcal{H}_{\mathcal{X}}$ can be defined with respect to the posterior. On the other hand, unless $f \in \mathcal{H}_{\mathcal{X}}$ holds, there is no theoretical guarantee that it gives a good estimate. In Section 5.1, we discuss some experimental results in such situations.

To obtain a point estimate of the posterior on x, it is proposed in Song et al. (2009) to use the preimage $\widehat{x} = \arg\min_x \|k_{\mathcal{X}}(\cdot, x) - \mathbf{k}_X^T R_{X|Y} \mathbf{k}_Y(y)\|^2_{\mathcal{H}_{\mathcal{X}}}$, which represents the posterior mean most effectively by one point. We use this approach in the present paper when point estimates are considered. In the case of the Gaussian kernel $\exp(-\|x-y\|^2/(2\sigma^2))$, the fixed point method

$$x^{(t+1)} = \frac{\sum_{i=1}^n X_i \rho_i \exp(-\|X_i - x^{(t)}\|^2/(2\sigma^2))}{\sum_{i=1}^n \rho_i \exp(-\|X_i - x^{(t)}|^2/(2\sigma^2))},$$

where $\rho = R_{X|Y}\mathbf{k}_Y(y)$, can be used to optimize x sequentially (Mika et al., 1999). This method usually converges very fast, although no theoretical guarantee exists for the convergence to the globally optimal point, as is usual in non-convex optimization.

A notable property of KBR is that the prior and likelihood are represented in terms of samples. Thus, unlike many approaches to Bayesian inference, precise knowledge of the prior and likelihood distributions is not needed, once samples are obtained. The following are typical situations where the KBR approach is advantageous:

- The probabilistic relation among variables is difficult to realize with a simple parametric model, while we can obtain samples of the variables easily. We will see such an example in Section 4.3.
- The probability density function of the prior and/or likelihood is hard to obtain explicitly, but sampling is possible:
  - In the field of population genetics, Bayesian inference is used with a likelihood expressed by branching processes to model the split of species, for which the explicit density is hard to obtain. Approximate Bayesian Computation (ABC) is a popular method for approximately sampling from a posterior without knowing the functional form (Marjoram et al., 2003; Sisson et al., 2007; Tavaré et al., 1997).
  - Another interesting application along these lines is nonparametric Bayesian inference (Müller and Quintana (2004) and references therein), in which the prior is typically given in the form of a process without a density form. In this case, sampling methods are often applied (MacEachern, 1994; MacEachern et al., 1999;

West et al., 1994, among others). Alternatively, the posterior may be approximated using variational methods (Blei and Jordan, 2006).

We will present an experimental comparison of KBR and ABC in Section 5.2.

- Even if explicit forms for the likelihood and prior are available, and standard sampling methods such as MCMC or sequential MC are applicable, the computation of a posterior estimate given y might still be computationally costly, making real-time applications unfeasible. Using KBR, however, the expectation of a function of the posterior given different y is obtained simply by taking the inner product as in Eq. (15), once $\mathbf{f}_X^T R_{X|Y}$ has been computed.

4.2 Discussions concerning implementation

When implementing KBR, a number of factors should be borne in mind to ensure good performance. First, in common with many nonparametric approaches, KBR requires training data in the region of the new "test" points for results to be meaningful. In other words, if the point on which we condition appears in a region far from the sample used for the estimation, the posterior estimator will be unreliable.

Second, in computing the posterior in KBR, Gram matrix inversion is necessary, which would cost $O(n^3)$ for sample size n if attempted directly. Substantial cost reductions can be achieved if the Gram matrices are approximated by low rank matrix approximations. A popular choice is the incomplete Cholesky decomposition (Fine and Scheinberg, 2001), which approximates a Gram matrix in the form of $\Gamma\Gamma^T$ with $n \times r$ matrix Γ ($r \ll n$) at cost $O(nr^2)$. Using this and the Woodbury identity, the KBR can be approximately computed at cost $O(nr^2)$.

Third, kernel choice or model selection is key to the effectiveness of any kernel method. In the case of KBR, we have three model parameters: the kernel (or its parameter, e.g. the bandwidth), the regularization parameter ε_n, and δ_n. The strategy for parameter selection depends on how the posterior is to be used in the inference problem. If it is to be applied in regression, we can use standard cross-validation. In the filtering experiments in Section 5, we use a validation method where we divide the training sample in two.

A more general model selection approach can also be formulated, by creating a new regression problem for the purpose. Suppose the prior Π is given by the marginal P_X of P. The posterior $Q_{\mathcal{X}|y}$ averaged with respect to P_Y is then equal to the marginal P_X itself. We are thus able to compare the discrepancy of the empirical kernel mean of P_X and the average of the estimators $\widehat{m}_{Q_{\mathcal{X}|y=Y_i}}$ over Y_i. This leads to a K-fold cross validation approach: for a partition of $\{1, \ldots, n\}$ into K disjoint subsets $\{T_a\}_{a=1}^K$, let $\widehat{m}_{Q_{\mathcal{X}|y}}^{[-a]}$ be the kernel mean of posterior computed using Gram matrices on data $\{(X_i, Y_i)\}_{i \notin T_a}$, and based on the prior mean $\widehat{m}_X^{[-a]}$ with data $\{X_i\}_{i \notin T_a}$. We can then cross validate by minimizing $\sum_{a=1}^K \left\| \frac{1}{|T_a|} \sum_{j \in T_a} \widehat{m}_{Q_{\mathcal{X}|y=Y_j}}^{[-a]} - \widehat{m}_X^{[a]} \right\|_{\mathcal{H}_\mathcal{X}}^2$, where $\widehat{m}_X^{[a]} = \frac{1}{|T_a|} \sum_{j \in T_a} k_\mathcal{X}(\cdot, X_j)$.

4.3 Application to nonparametric state-space model

We next describe how KBR may be used in a particular application: namely, inference in a general time invariant state-space model,

$$p(X,Y) = \pi(X_1)\prod_{t=1}^{T} p(Y_t|X_t)\prod_{t=1}^{T-1} q(X_{t+1}|X_t),$$

where Y_t is an observable variable, and X_t is a hidden state variable. We begin with a brief review of alternative strategies for inference in state-space models with complex dynamics, for which linear models are not suitable. The extended Kalman filter (EKF) and unscented Kalman filter (UKF, Julier and Uhlmann, 1997) are nonlinear extensions of the standard linear Kalman filter, and are well established in this setting. Alternatively, nonparametric estimates of conditional density functions can be employed, including kernel density estimation or distribution estimates on a partitioning of the space (Monbet et al., 2008; Thrun et al., 1999). The latter nonparametric approaches are effective only for low-dimensional cases, however. Most relevant to this paper are Song et al. (2009) and Song et al. (2010b), in which the kernel means and covariance operators are used to implement the nonparametric HMM.

In this paper, we apply the KBR for inference in the nonparametric state-space model. We do not assume the conditional probabilities $p(Y_t|X_t)$ and $q(X_{t+1}|X_t)$ to be known explicitly, nor do we estimate them with simple parametric models. Rather, we assume a sample $(X_1,Y_1),\ldots,(X_{T+1},Y_{T+1})$ is given for both the observable and hidden variables in the training phase. The conditional probability for observation process $p(y|x)$ and the transition $q(x_{t+1}|x_t)$ are represented by the empirical covariance operators as computed on the training sample,

$$\widehat{C}_{XY} = \frac{1}{T}\sum_{i=1}^{T} k_{\mathcal{X}}(\cdot,X_i)\otimes k_{\mathcal{Y}}(\cdot,Y_i), \quad \widehat{C}_{X_{+1}X} = \frac{1}{T}\sum_{i=1}^{T} k_{\mathcal{X}}(\cdot,X_{i+1})\otimes k_{\mathcal{X}}(\cdot,X_i), \tag{17}$$
$$\widehat{C}_{YY} = \frac{1}{T}\sum_{i=1}^{T} k_{\mathcal{Y}}(\cdot,Y_i)\otimes k_{\mathcal{Y}}(\cdot,Y_i), \quad \widehat{C}_{XX} = \frac{1}{T}\sum_{i=1}^{T} k_{\mathcal{X}}(\cdot,X_i)\otimes k_{\mathcal{X}}(\cdot,X_i).$$

While the sample is not i.i.d., we can use the empirical covariances, which are consistent by the mixing property of Markov models.

Typical applications of the state-space model are filtering, prediction, and smoothing, which are defined by the estimation of $p(x_s|y_1,\ldots,y_t)$ for $s=t$, $s>t$, and $s<t$, respectively. Using the KBR, any of these can be computed. For simplicity we explain the filtering problem in this paper, but the remaining cases are similar. In filtering, given new observations $\tilde{y}_1,\ldots,\tilde{y}_t$, we wish to estimate the current hidden state x_t. The sequential estimate for the kernel mean of $p(x_t|\tilde{y}_1,\ldots,\tilde{y}_t)$ can be derived via KBR. Suppose we already have an estimator of the kernel mean of $p(x_t|\tilde{y}_1,\ldots,\tilde{y}_t)$ in the form

$$\widehat{m}_{x_t|\tilde{y}_1,\ldots,\tilde{y}_t} = \sum_{i=1}^{T} \alpha_i^{(t)} k_{\mathcal{X}}(\cdot,X_i),$$

where $\alpha_i^{(t)} = \alpha_i^{(t)}(\tilde{y}_1,\ldots,\tilde{y}_t)$ are the coefficients at time t.

From $p(x_{t+1}|\tilde{y}_1,\ldots,\tilde{y}_t) = \int p(x_{t+1}|x_t)p(x_t|\tilde{y}_1,\ldots,\tilde{y}_t)dx_t$, Theorem 3.2 tells us the kernel mean of x_{t+1} given $\tilde{y}_1,\ldots,\tilde{y}_t$ is estimated by $\widehat{m}_{x_{t+1}|\tilde{y}_1,\ldots,\tilde{y}_t} = \widehat{C}_{X_{+1}X}(\widehat{C}_{XX}+\varepsilon_T I)^{-1}\widehat{m}_{x_t|\tilde{y}_1,\ldots,\tilde{y}_t} = \mathbf{k}_{X_{+1}}^T(G_X + T\varepsilon_T I_T)^{-1}G_X\alpha^{(t)}$, where $\mathbf{k}_{X_{+1}}^T = (k_{\mathcal{X}}(\cdot,X_2),\ldots,k_{\mathcal{X}}(\cdot,X_{T+1}))$. Applying Theorem 3.2 again with $p(y_{t+1}|\tilde{y}_1,\ldots,\tilde{y}_t) = \int p(y_{t+1}|x_{t+1})p(x_{t+1}|\tilde{y}_1,\ldots,\tilde{y}_t)dx_t$, we have an estimate for the kernel mean of the prediction $p(y_{t+1}|\tilde{y}_1,\ldots,\tilde{y}_t)$,

$$\widehat{m}_{y_{t+1}|\tilde{y}_1,\ldots,\tilde{y}_t} = \widehat{C}_{YX}(\widehat{C}_{XX}+\varepsilon_T I)^{-1}\widehat{m}_{x_{t+1}|\tilde{y}_1,\ldots,\tilde{y}_t} = \sum_{i=1}^T \widehat{\mu}_i^{(t+1)} k_{\mathcal{Y}}(\cdot,Y_i),$$

where the coefficients $\widehat{\mu}^{(t+1)} = (\widehat{\mu}_i^{(t+1)})_{i=1}^T$ are given by

$$\widehat{\mu}^{(t+1)} = \big(G_X + T\varepsilon_T I_T\big)^{-1} G_{XX_{+1}} \big(G_X + T\varepsilon_T I_T\big)^{-1} G_X \alpha^{(t)}. \tag{18}$$

Here $G_{XX_{+1}}$ is the "transfer" matrix defined by $\big(G_{XX_{+1}}\big)_{ij} = k_{\mathcal{X}}(X_i, X_{j+1})$. From $p(x_{t+1}|\tilde{y}_1,\ldots,\tilde{y}_{t+1}) = \frac{p(y_{t+1}|x_{t+1})p(x_{t+1}|\tilde{y}_1,\ldots,\tilde{y}_t)}{\int p(y_{t+1}|x_{t+1})p(x_{t+1}|\tilde{y}_1,\ldots,\tilde{y}_t)dx_{t+1}}$, kernel Bayes' rule with the prior $p(x_{t+1}|\tilde{y}_1,\ldots,\tilde{y}_t)$ and the likelihood $p(y_{t+1}|x_{t+1})$ yields

$$\alpha^{(t+1)} = \Lambda^{(t+1)} G_Y \big((\Lambda^{(t+1)} G_Y)^2 + \delta_T I_T\big)^{-1} \Lambda^{(t+1)} \mathbf{k}_Y(\tilde{y}_{t+1}), \tag{19}$$

where $\Lambda^{(t+1)} = \mathrm{diag}(\widehat{\mu}_1^{(t+1)},\ldots,\widehat{\mu}_T^{(t+1)})$. Eqs. (18) and (19) describe the update rule of $\alpha^{(t)}(\tilde{y}_1,\ldots,\tilde{y}_t)$.

If the prior $\pi(x_1)$ is available, the posterior estimate at x_1 given $\tilde{y}_1$ is obtained by the kernel Bayes' rule. If not, we may use Eq. (10) to get an initial estimate $\widehat{C}_{XY}(\widehat{C}_{YY} + \varepsilon_n I)^{-1} k_{\mathcal{Y}}(\cdot,\tilde{y}_1)$, yielding $\alpha^{(1)}(\tilde{y}_1) = T(G_Y + T\varepsilon_T I_T)^{-1}\mathbf{k}_Y(\tilde{y}_1)$.

In sequential filtering, a substantial reduction in computational cost can be achieved by low rank matrix approximations, as discussed above. Given an approximation of rank r for the Gram matrices and transfer matrix, and employing the Woodbury identity, the computation costs just $O(Tr^2)$ for each time step.

4.4 Bayesian computation without likelihood

We next address the setting where the likelihood is not known in analytic form, but sampling is possible. In this case, Approximate Bayesian Computation (ABC) is a popular method for Bayesian inference. The simplest form of ABC, which is called the rejection method, generates a sample from $q(Z|W=y)$ as follows: (i) generate a sample X_t from the prior Π, (ii) generate a sample Y_t from $P(Y|X_t)$, (iii) if $D(y,Y_t) < \tau$, accept X_t; otherwise reject, (iv) go to (i). In step (iii), D is a distance measure of the space $\mathcal{X}$, and τ is tolerance to acceptance.

In the same setting as ABC, KBR gives the following sampling-based method for computing the kernel posterior mean:

1. Generate a sample $X_1,\ldots,X_n$ from the prior Π.
2. Generate a sample Y_t from $P(Y|X_t)$ $(t=1,\ldots,n)$.
3. Compute Gram matrices G_X and G_Y with $(X_1,Y_1),\ldots,(X_n,Y_n)$, and $R_{X|Y}\mathbf{k}_Y(y)$.

Alternatively, since (X_t, Y_t) is an sample from Q, it is possible to use Eq. (10) for the kernel mean of the conditional probability $q(x|y)$. As in Song et al. (2009), the estimator is given by

$$\sum_{t=1}^{n} \nu_j k_{\mathcal{X}}(\cdot, X_t), \quad \nu = (G_Y + N\varepsilon_N I_N)^{-1}\mathbf{k}_Y(y).$$

The distribution of a sample generated by ABC approaches to the true posterior if τ goes to zero, while empirical estimates via the kernel approaches converge to the true posterior mean in the limit of infinite sample size. The efficiency of ABC, however, can be arbitrarily poor for small τ, since a sample X_t is then rarely accepted in Step (iii).

The ABC method generates a sample, hence any statistics based on the posterior can be approximated. Given a posterior mean obtained by one of the kernel methods, however, we may only obtain expectations of functions in the RKHS, meaning that certain statistics (such as confidence intervals) are not straightforward to obtain. In Section 5.2, we present an experimental evaluation of the trade-off between computation time and accuracy for ABC and KBR.

5. Numerical Examples

5.1 Nonparametric inference of posterior

The first numerical example is a comparison between KBR and a kernel density estimation (KDE) approach to obtaining conditional densities. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be an i.i.d. sample from P on $\mathbb{R}^d \times \mathbb{R}^r$. With probability density functions $K^{\mathcal{X}}(x)$ on $\mathbb{R}^d$ and $K^{\mathcal{Y}}(y)$ on $\mathbb{R}^r$, the conditional probability density function $p(y|x)$ is estimated by

$$\widehat{p}(y|x) = \frac{\sum_{j=1}^{n} K^{\mathcal{X}}_{h_X}(x - X_j) K^{\mathcal{Y}}_{h_Y}(y - Y_j)}{\sum_{j=1}^{n} K^{\mathcal{X}}_{h}(x - X_j)},$$

where $K^{\mathcal{X}}_{h_X}(x) = h_X^{-d} K^{\mathcal{X}}(x/h_X)$ and $K^{\mathcal{Y}}_{h_Y}(x) = h_Y^{-r} K^{\mathcal{Y}}(y/h_Y)$ $(h_X, h_Y > 0)$. Given an i.i.d. sample $U_1, \ldots, U_\ell$ from the prior Π, the particle representation of the posterior can be obtained by importance weighting (IW). Using this scheme, the posterior $q(x|y)$ given $y \in \mathbb{R}^r$ is represented by the weighted sample (U_i, ζ_i) with $\zeta_i = \widehat{p}(y|U_i)/\sum_{j=1}^{\ell} \widehat{p}(y|U_j)$.

We compare the estimates of $\int x q(x|y) dx$ obtained by KBR and KDE + IW, using Gaussian kernels for both the methods. Note that the function $f(x) = x$ does not belong to the Gaussian kernel RKHS, and the consistency of KBR is not rigorously guaranteed for this function (c.f. Theorem 3.5). That said, Gaussian kernels are known to be able to approximate any continuous function on a compact subset of the Euclidean space with arbitrary accuracy (Steinwart, 2001). With such kernels, we can expect the posterior mean to be approximated with high accuracy on any compact set, and thus on average. In our experiments, the dimensionality was given by $r = d$ ranging from 2 to 64. The distribution P of (X, Y) was $N((0, \mathbf{1}_d^T)^T, V)$ with $V = A^T A + 2I_d$, where $\mathbf{1}_d = (1, \ldots, 1)^T \in \mathbb{R}^d$ and each component of A was randomly generated as $N(0, 1)$ for each run. The prior Π was $P_X = N(0, V_{XX}/2)$, where V_{XX} is the X-component of V. The sample sizes were $n = \ell = 200$. The bandwidth parameters h_X, h_Y in KDE were set $h_X = h_Y$, and chosen over the set $\{2 * i \mid i = 1, \ldots, 10\}$ in two ways: least square cross-validation (Bowman, 1984; Rudemo,

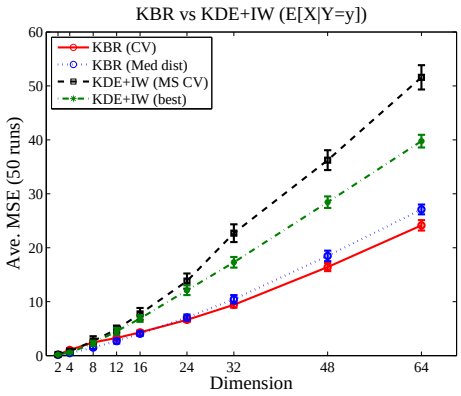


Figure 2: Comparison between KBR and KDE+IW.

1982) and the best mean performance. For the KBR, we chose σ in $e^{-\|x-x'\|^2/(2\sigma^2)}$ in two ways: the median over the pairwise distances in the data (Gretton et al., 2008), and the 10-fold cross-validation approach described in Section 4.1. Figure 2 shows the mean square errors (MSE) of the estimates over 1000 random points $y \sim N(0, V_{YY})$. KBR significantly outperforms the KDE+IW approach. Unsurprisingly, the MSE of both methods increases with dimensionality.

5.2 Bayesian computation without likelihood

We compare ABC and the kernel methods, KBR and conditional mean, in terms of estimation accuracy and computational time, since they have an obvious tradeoff. To compute the estimation accuracy rigorously, the ground truth is needed: thus we use Gaussian distributions for the true prior and likelihood, which makes the posterior easy to compute in closed form. The samples are taken from the same model used in Section 5.1, and $\int xq(x|y)dx$ is evaluated at 10 different points of y. We performed 10 random runs with different random generation of the true distributions.

For ABC, we used only the rejection method; while there are more advanced sampling schemes (Marjoram et al., 2003; Sisson et al., 2007), their implementation is dependent on the problem being solved. Various values for the acceptance region τ are used, and the accuracy and computational time are shown in Fig. 3 together with total sizes of the generated samples. For the kernel methods, the sample size n is varied. The regularization parameters are given by $\varepsilon_n = 0.01/n$ and $\delta_n = 2\varepsilon_n$ for KBR, and $\varepsilon_n = 0.01/\sqrt{n}$ for the conditional kernel mean. The kernels in the kernel methods are Gaussian kernels for which the bandwidth parameters are chosen by the median of the pairwise distances on the data (Gretton et al., 2008). The incomplete Cholesky decomposition is employed for the low-rank approximation. The results indicate that kernel methods achieve more accurate results than ABC at a given computational cost, and the conditional kernel mean shows better results.

5.3 Filtering problems

We next compare the KBR filtering method (proposed in Section 4.3) with EKF and UKF on synthetic data.

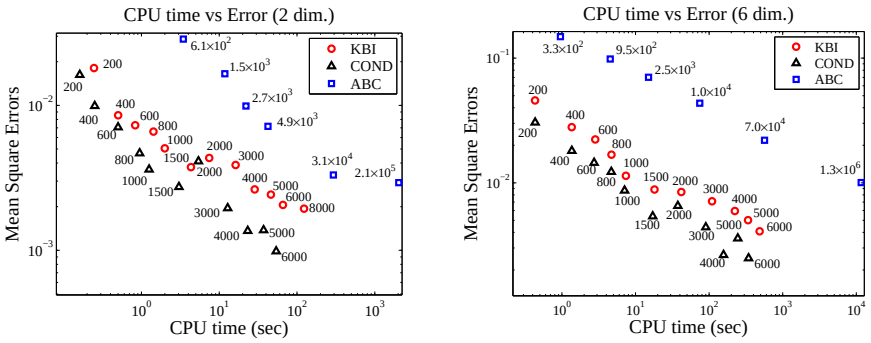


Figure 3: Comparison of estimation accuracy and computational time with KBR and ABC for Bayesian computation without likelihood. The numbers at the marks are the sample sizes generated for computation.

KBR has the regularization parameters ε_T, δ_T, and kernel parameters for $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$ (e.g., the bandwidth parameter for an RBF kernel). Under the assumption that a training sample is available, cross-validation can be performed on the training sample to select the parameters. By dividing the training sample into two, one half is used to estimate the covariance operators Eq. (17) with a candidate parameter set, and the other half to evaluate the estimation errors. To reduce the search space and attendant computational cost, we used a simpler procedure, setting $\delta_T = 2\varepsilon_T$, and using the Gaussian kernel bandwidths $\beta\sigma_{\mathcal{X}}$ and $\beta\sigma_{\mathcal{Y}}$, where $\sigma_{\mathcal{X}}$ and $\sigma_{\mathcal{Y}}$ are the median of pairwise distances in the training samples (Gretton et al., 2008). This leaves only two parameters β and ε_T to be tuned.

We applied the KBR filtering algorithm from Section 4.3 to two synthetic data sets: a simple nonlinear dynamical system, in which the degree of nonlinearity can be controlled, and the problem of camera orientation recovery from an image sequence. In the first case, the hidden state is $X_t = (u_t, v_t)^T \in \mathbb{R}^2$, and the dynamics are given by

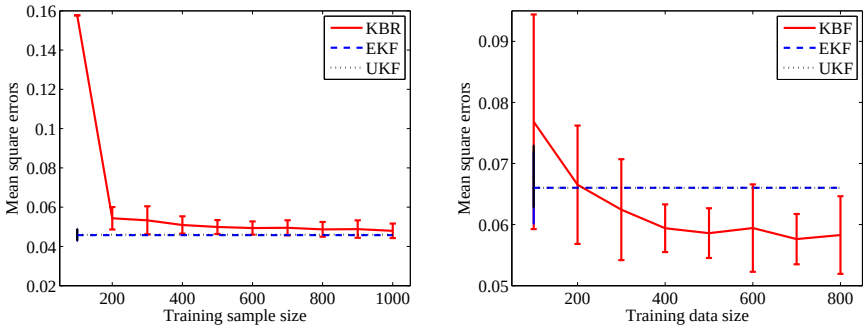
$$\begin{pmatrix} u_{t+1} \\ v_{t+1} \end{pmatrix} = (1 + b\sin(M\theta_{t+1})) \begin{pmatrix} \cos\theta_{t+1} \\ \sin\theta_{t+1} \end{pmatrix} + \zeta_t, \quad \theta_{t+1} = \theta_t + \eta \pmod{2\pi},$$

where $\eta > 0$ is an increment of the angle and $\zeta_t \sim N(0, \sigma_h^2 I_2)$ is independent process noise. Note that the dynamics of (u_t, v_t) are nonlinear even for $b = 0$. The observation Y_t follows

$$Y_t = (u_t, v_t)^T + \xi_t, \qquad \xi_t \sim N(0, \sigma_o^2 I),$$

where ξ_t is independent noise. The two dynamics are defined as follows. (a) (rotation with noisy observation) $\eta = 0.3$, $b = 0$, $\sigma_h = \sigma_o = 0.2$. (b) (oscillatory rotation with noisy observation) $\eta = 0.4$, $b = 0.4$, $M = 8$, $\sigma_h = \sigma_o = 0.2$. (See Fig.5).

We assume the correct dynamics are known to the EKF and UKF. The results are shown in Fig. 4. In all the cases, EKF and UKF show unrecognizably small difference. The dynamics in (a) are weakly nonlinear, and KBR has slightly worse MSE than EKF and UKF. For dataset (b), which has strong nonlinearity, KBR outperforms the nonlinear Kalman filter for $T \geq 200$.



Data (a) Data (b)

Figure 4: Comparisons with the KBR Filter and EKF. (Average MSEs and standard errors over 30 runs.)

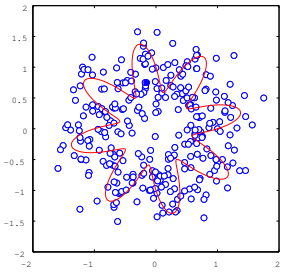


Figure 5: Example of data (b) (X_t, $N = 300$)

In our second synthetic example, we applied the KBR filter to the camera rotation problem used in Song et al. (2009). The angle of a camera, which is located at a fixed position, is a hidden variable, and movie frames recorded by the camera are observed. The data are generated virtually using a computer graphics environment. As in Song et al. (2009), we are given 3600 downsampled frames of 20×20 RGB pixels ($Y_t \in [0,1]^{1200}$), where the first 1800 frames are used for training, and the second half are used to test the filter. We make the data noisy by adding Gaussian noise $N(0, \sigma^2)$ to Y_t.

Our experiments cover two settings. In the first, we assume we do not know that the hidden state S_t is included in $SO(3)$, but only that it is a general 3×3 matrix. In this case, we use the Kalman filter by estimating the relations under a linear assumption, and the KBR filter with Gaussian kernels for S_t and X_t as Euclidean vectors. In the second setting, we exploit the fact that $S_t \in SO(3)$: for the Kalman Filter, S_t is represented by a quanternion, which is a standard vector representation of rotations; for the KBR filter the kernel $k(A, B) = \mathrm{Tr}[AB^T]$ is used for S_t, and S_t is estimated within $SO(3)$. Table 1 shows the Frobenius norms between the estimated matrix and the true one. The KBR filter significantly outperforms the EKF, since KBR has the advantage in extracting the complex nonlinear dependence between the observation and the hidden state.

	KBR (Gauss)	KBR (Tr)	Kalman (9 dim.)	Kalman (Quat.)
$\sigma^2 = 10^{-4}$	0.210 ± 0.015	0.146 ± 0.003	1.980 ± 0.083	0.557 ± 0.023
$\sigma^2 = 10^{-3}$	0.222 ± 0.009	0.210 ± 0.008	1.935 ± 0.064	0.541 ± 0.022

Table 1: Average MSE and standard errors of estimating camera angles (10 runs).

6. Proofs

The proof idea for the consistency rates of the KBR estimators is similar to Caponnetto and De Vito (2007) and Smale and Zhou (2007), in which the basic techniques are taken from the general theory of regularization (Engl et al., 2000).

The first preliminary result is a rate of convergence for the mean transition in Theorem 3.2. In the following $\mathcal{R}(C_{XX}^0)$ means $\mathcal{H}_\mathcal{X}$.

Theorem 6.1 *Assume that $\pi/p_X \in \mathcal{R}(C_{XX}^\beta)$ for some $\beta \geq 0$, where π and p_X are the p.d.f. of Π and P_X, respectively. Let $\widehat{m}_\Pi^{(n)}$ be an estimator of m_Π such that $\|\widehat{m}_\Pi^{(n)} - m_\Pi\|_{\mathcal{H}_\mathcal{X}} = O_p(n^{-\alpha})$ as $n \to \infty$ for some $0 < \alpha \leq 1/2$. Then, with $\varepsilon_n = n^{-\max\{\frac{2}{3}\alpha, \frac{\alpha}{1+\beta}\}}$, we have*

$$\left\| \widehat{C}_{YX}^{(n)} \left(\widehat{C}_{XX}^{(n)} + \varepsilon_n I \right)^{-1} \widehat{m}_\Pi^{(n)} - m_{Q\mathcal{Y}} \right\|_{\mathcal{H}_\mathcal{Y}} = O_p(n^{-\min\{\frac{2}{3}\alpha, \frac{2\beta+1}{2\beta+2}\alpha\}}), \quad (n \to \infty).$$

Proof Take $\eta \in \mathcal{H}_\mathcal{X}$ such that $\pi/p_X = C_{XX}^\beta \eta$. Then, we have

$$m_\Pi = \int k_\mathcal{X}(\cdot, x) \frac{\pi(x)}{p_X(x)} p_X(x) d\nu_\mathcal{X}(x) = C_{XX}^{\beta+1} \eta. \tag{20}$$

First we show the rate of the estimation error:

$$\left\| \widehat{C}_{YX}^{(n)} \left(\widehat{C}_{XX}^{(n)} + \varepsilon_n I \right)^{-1} \widehat{m}_\Pi^{(n)} - C_{YX} \left(C_{XX} + \varepsilon_n I \right)^{-1} m_\Pi \right\|_{\mathcal{H}_\mathcal{Y}} = O_p\left(n^{-\alpha} \varepsilon_n^{-1/2} \right), \tag{21}$$

as $n \to \infty$. By using $B^{-1} - A^{-1} = B^{-1}(A - B)A^{-1}$ for any invertible operators A and B, the left hand side of Eq. (21) is upper bounded by

$$\begin{aligned} &\left\| \widehat{C}_{YX}^{(n)} \left(\widehat{C}_{XX}^{(n)} + \varepsilon_n I \right)^{-1} \left(\widehat{m}_\Pi^{(n)} - m_\Pi \right) \right\|_{\mathcal{H}_\mathcal{Y}} + \left\| \left(\widehat{C}_{YX}^{(n)} - C_{YX} \right) \left(C_{XX} + \varepsilon_n I \right)^{-1} m_\Pi \right\|_{\mathcal{H}_\mathcal{Y}} \\ &\qquad + \left\| \widehat{C}_{YX}^{(n)} \left(\widehat{C}_{XX}^{(n)} + \varepsilon_n I \right)^{-1} \left(C_{XX} - \widehat{C}_{XX}^{(n)} \right) \left(C_{XX} + \varepsilon_n I \right)^{-1} m_\Pi \right\|_{\mathcal{H}_\mathcal{Y}}. \end{aligned}$$

By the decomposition $\widehat{C}_{YX}^{(n)} = \widehat{C}_{YY}^{(n)1/2} \widehat{W}_{YX}^{(n)} \widehat{C}_{XX}^{(n)1/2}$ with $\|\widehat{W}_{YX}^{(n)}\| \leq 1$ (Baker, 1973), we have $\|\widehat{C}_{YX}^{(n)} \left(\widehat{C}_{XX}^{(n)} + \varepsilon_n I \right)^{-1}\| = O_p(\varepsilon_n^{-1/2})$, which implies the first term is of $O_p(n^{-\alpha} \varepsilon_n^{-1/2})$. From the $\sqrt{n}$ consistency of the covariance operators and $m_\Pi = C_{XX}^{\beta+1} \eta$, a similar argument to the first term proves that the second and third terms are of the order $O_p(n^{-1/2})$ and $O_p(n^{-1/2} \varepsilon_n^{-1/2})$, respectively, which means Eq. (21).

Next, we show the rate for the approximation error

$$\left\| C_{YX} \left(C_{XX} + \varepsilon_n I \right)^{-1} m_\Pi - m_{Q\mathcal{Y}} \right\|_{\mathcal{H}_\mathcal{Y}} = O(\varepsilon_n^{\min\{(1+2\beta)/2, 1\}}) \qquad (n \to \infty). \tag{22}$$

Let $C_{YX} = C_{YY}^{1/2} W_{YX} C_{XX}^{1/2}$ be the decomposition with $\|W_{YX}\| \leq 1$. It follows from Eq. (20) and the relation

$$m_{Q_{\mathcal{Y}}} = \int\int k(\cdot, y) \frac{\pi(x)}{p_X(x)} p(x,y) d\nu_{\mathcal{X}}(x) d\nu_{\mathcal{Y}}(y) = C_{YX} C_{XX}^{\beta} \eta$$

that the left hand side of Eq. (22) is upper bounded by

$$\|C_{YY}^{1/2} W_{YX}\| \, \| \left(C_{XX} + \varepsilon_n I\right)^{-1} C_{XX}^{(2\beta+3)/2} \eta - C_{XX}^{(2\beta+1)/2} \eta \|_{\mathcal{H}_{\mathcal{X}}}.$$

By the eigendecomposition $C_{XX} = \sum_i \lambda_i \phi_i \langle \phi_i, \cdot \rangle$, where $\{\lambda_i\}$ are the positive eigenvalues and $\{\phi_i\}$ are the corresponding unit eigenvectors, the expansion

$$\left\| \left(C_{XX} + \varepsilon_n I\right)^{-1} C_{XX}^{(2\beta+3)/2} \eta - C_{XX}^{(2\beta+1)/2} \eta \right\|_{\mathcal{H}_{\mathcal{X}}}^2 = \sum_i \left(\frac{\varepsilon_n \lambda_i^{(2\beta+1)/2}}{\lambda_i + \varepsilon_n} \right)^2 \langle \eta, \phi_i \rangle^2$$

holds. If $0 \leq \beta < 1/2$, we have $\frac{\varepsilon_n \lambda_i^{(2\beta+1)/2}}{\lambda_i + \varepsilon_n} = \frac{\lambda_i^{(2\beta+1)/2}}{(\lambda_i+\varepsilon_n)^{(2\beta+1)/2}} \frac{\varepsilon_n^{(1-2\beta)/2}}{(\lambda_i+\varepsilon_n)^{(1-2\beta)/2}} \varepsilon_n^{(2\beta+1)/2} \leq \varepsilon_n^{(2\beta+1)/2}$. If $\beta \geq 1/2$, then $\frac{\varepsilon_n \lambda_i^{(2\beta+1)/2}}{\lambda_i + \varepsilon_n} \leq \|C_{XX}\| \varepsilon_n$. The dominated convergence theorem shows that the the above sum converges to zero of the order $O(\varepsilon_n^{\min\{2\beta+1, 2\}})$ as $\varepsilon_n \to 0$.

From Eqs. (21) and (22), the optimal order of ε_n and the optimal rate of consistency are given as claimed. ■

The following theorem shows the consistency rate of the estimator used in the conditioning step Eq. (11).

Theorem 6.2 *Let f be a function in $\mathcal{H}_{\mathcal{X}}$, and (Z, W) be a random variable taking values in $\mathcal{X} \times \mathcal{Y}$. Assume that $E[f(Z)|W = \cdot] \in \mathcal{R}(C_{WW}^{\nu})$ for some $\nu \geq 0$, and $\widehat{C}_{WZ}^{(n)} : \mathcal{H}_{\mathcal{X}} \to \mathcal{H}_{\mathcal{Y}}$ and $\widehat{C}_{WW}^{(n)} : \mathcal{H}_{\mathcal{Y}} \to \mathcal{H}_{\mathcal{Y}}$ be compact operators, which may not be positive definite, such that $\|\widehat{C}_{WZ}^{(n)} - C_{WZ}\| = O_p(n^{-\gamma})$ and $\|\widehat{C}_{WW}^{(n)} - C_{WW}\| = O_p(n^{-\gamma})$ for some $\gamma > 0$. Then, for a positive sequence $\delta_n = n^{-\max\{\frac{4}{9}\gamma, \frac{4}{2\nu+5}\gamma\}}$, we have as $n \to \infty$*

$$\left\| \widehat{C}_{WW}^{(n)} \left((\widehat{C}_{WW}^{(n)})^2 + \delta_n I \right)^{-1} \widehat{C}_{WZ}^{(n)} f - E[f(X)|W = \cdot] \right\|_{\mathcal{H}_{\mathcal{X}}} = O_p(n^{-\min\{\frac{4}{9}\gamma, \frac{2\nu}{2\nu+5}\gamma\}}).$$

Proof Let $\eta \in \mathcal{H}_{\mathcal{X}}$ such that $E[f(Z)|W = \cdot] = C_{WW}^{\nu} \eta$. First we show

$$\left\| \widehat{C}_{WW}^{(n)} \left((\widehat{C}_{WW}^{(n)})^2 + \delta_n I \right)^{-1} \widehat{C}_{WZ}^{(n)} f - C_{WW} (C_{WW}^2 + \delta_n I)^{-1} C_{WZ} f \right\|_{\mathcal{H}_{\mathcal{X}}} = O_p(n^{-\gamma} \delta_n^{-5/4}). \quad (23)$$

The left hand side of Eq. (23) is upper bounded by

$$\begin{aligned}
&\left\| \widehat{C}_{WW}^{(n)} \left((\widehat{C}_{WW}^{(n)})^2 + \delta_n I \right)^{-1} (\widehat{C}_{WZ}^{(n)} - C_{WZ}) f \right\|_{\mathcal{H}_{\mathcal{Y}}} \\
&+ \left\| (\widehat{C}_{WW}^{(n)} - C_{WW}) (C_{WW}^2 + \delta_n I)^{-1} C_{WZ} f \right\|_{\mathcal{H}_{\mathcal{Y}}} \\
&+ \left\| \widehat{C}_{WW}^{(n)} \left((\widehat{C}_{WW}^{(n)})^2 + \delta_n I \right)^{-1} \left((\widehat{C}_{WW}^{(n)})^2 - C_{WW}^2 \right) (C_{WW}^2 + \delta_n I)^{-1} C_{WZ} f \right\|_{\mathcal{H}_{\mathcal{Y}}}.
\end{aligned}$$

Let $\widehat{C}_{WW}^{(n)} = \sum_i \lambda_i \phi_i \langle \phi_i, \cdot \rangle$ be the eigendecomposition, where $\{\phi_i\}$ is the unit eigenvectors and $\{\lambda_i\}$ is the corresponding eigenvalues. From $\left|\lambda_i/(\lambda_i^2 + \delta_n)\right| = 1/|\lambda_i + \delta_n/\lambda_i| \leq 1/(2\sqrt{|\lambda_i|}\sqrt{\delta_n/|\lambda_i|}) = 1/(2\sqrt{\delta_n})$, we have $\|\widehat{C}_{WW}^{(n)}\big((\widehat{C}_{WW}^{(n)})^2 + \delta_n I\big)^{-1}\| \leq 1/(2\sqrt{\delta_n})$, and thus the first term of the above bound is of $O_p(n^{-\gamma}\delta_n^{-1/2})$. A similar argument by the eigendecomposition of C_{WW} combined with the decomposition $C_{WZ} = C_{WW}^{1/2} U_{WZ} C_{ZZ}^{1/2}$ with $\|U_{WZ}\| \leq 1$ shows that the second term is of $O_p(n^{-\gamma}\delta_n^{-3/4})$. From the fact $\|(\widehat{C}_{WW}^{(n)})^2 - C_{WW}^2\| \leq \|\widehat{C}_{WW}^{(n)}(\widehat{C}_{WW}^{(n)} - C_{WW})\| + \|(\widehat{C}_{WW}^{(n)} - C_{WW})C_{WW}\| = O_p(n^{-\gamma})$, the third term is of $O_p(n^{-\gamma}\delta_n^{-5/4})$. This implies Eq. (23).

From $E[f(Z)|W = \cdot] = C_{WW}^{\nu}\eta$ and $C_{WZ}f = C_{WW}E[f(Z)|W = \cdot] = C_{WW}^{\nu+1}\eta$, the convergence rate

$$\big\|C_{WW}(C_{WW}^2 + \delta_n I)^{-1} C_{WZ} f - E[f(Z)|W = \cdot]\big\|_{\mathcal{H}_\mathcal{Y}} = O(\delta_n^{\min\{1, \frac{\nu}{2}\}}). \tag{24}$$

can be proved by the same way as Eq. (22).

Combination of Eqs.(23) and (24) proves the assertion. ■

Recall that $\tilde{C}_{WW}$ is the integral operator on $L^2(Q_\mathcal{Y})$ defined by Eq. (16). The following theorem shows the consistency rate on average. Here $\mathcal{R}(\tilde{C}_{WW}^0)$ means $L^2(Q_\mathcal{Y})$.

Theorem 6.3 *Let f be a function in $\mathcal{H}_\mathcal{X}$, and (Z, W) be a random variable taking values in $\mathcal{X} \times \mathcal{Y}$ with distribution Q. Assume that $E[f(Z)|W = \cdot] \in \mathcal{R}(\tilde{C}_{WW}^{\nu}) \cap \mathcal{H}_\mathcal{Y}$ for some $\nu > 0$, and $\widehat{C}_{WZ}^{(n)} : \mathcal{H}_\mathcal{X} \to \mathcal{H}_\mathcal{Y}$ and $\widehat{C}_{WW}^{(n)} : \mathcal{H}_\mathcal{Y} \to \mathcal{H}_\mathcal{Y}$ be compact operators, which may not be positive definite, such that $\|\widehat{C}_{WZ}^{(n)} - C_{WZ}\| = O_p(n^{-\gamma})$ and $\|\widehat{C}_{WW}^{(n)} - C_{WW}\| = O_p(n^{-\gamma})$ for some $\gamma > 0$. Then, for a positive sequence $\delta_n = n^{-\max\{\frac{1}{2}\gamma, \frac{2}{\nu+2}\gamma\}}$, we have as $n \to \infty$*

$$\big\|\widehat{C}_{WW}^{(n)}\big((\widehat{C}_{WW}^{(n)})^2 + \delta_n I\big)^{-1}\widehat{C}_{WZ}^{(n)} f - E[f(X)|W = \cdot]\big\|_{L^2(Q_\mathcal{Y})} = O_p(n^{-\min\{\frac{1}{2}\gamma, \frac{\nu}{\nu+2}\gamma\}}).$$

Proof Note that for $f, g \in \mathcal{H}_\mathcal{X}$ we have $(f, g)_{L^2(Q_\mathcal{Y})} = E[f(W)g(W)] = \langle f, C_{WW} g\rangle_{\mathcal{H}_\mathcal{X}}$. It follows that the left hand side of the assertion is equal to

$$\big\|C_{WW}^{1/2}\widehat{C}_{WW}^{(n)}\big((\widehat{C}_{WW}^{(n)})^2 + \delta_n I\big)^{-1}\widehat{C}_{WZ}^{(n)} f - C_{WW}^{1/2} E[f(Z)|W = \cdot]\big\|_{\mathcal{H}_\mathcal{Y}}.$$

First, by the similar argument to the proof of Eq. (23), it is easy to show that the rate of the estimation error is given by

$$\big\|C_{WW}^{1/2}\big\{\widehat{C}_{WW}^{(n)}\big((\widehat{C}_{WW}^{(n)})^2 + \delta_n I\big)^{-1}\widehat{C}_{WZ}^{(n)} f - C_{WW}(C_{WW}^2 + \delta_n I)^{-1} C_{WZ} f\big\}\big\|_{\mathcal{H}_\mathcal{Y}} = O_p(n^{-\gamma}\delta_n^{-1}).$$

It suffices then to prove

$$\big\|C_{WW}(C_{WW}^2 + \delta_n I)^{-1} C_{WZ} f - E[f(Z)|W = \cdot]\big\|_{L^2(Q_\mathcal{Y})} = O(\delta_n^{\min\{1, \frac{\nu}{2}\}}).$$

Let $\xi \in L^2(Q_{\mathcal{Y}})$ such that $E[f(Z)|W = \cdot] = \tilde{C}_{WW}^{\nu}\xi$. In a similar way to Theorem 3.1, $\tilde{C}_{WW}E[f(Z)|W] = \tilde{C}_{WZ}f$ holds, where $\tilde{C}_{WZ}$ is the extension of C_{WZ}, and thus $C_{WZ}f = \tilde{C}_{WW}^{\nu+1}\xi$. The left hand side of the above equation is equal to

$$\left\|\tilde{C}_{WW}(\tilde{C}_{WW}^2 + \delta_n I)^{-1}\tilde{C}_{WW}^{\nu+1}\xi - \tilde{C}_{WW}^{\nu}\xi\right\|_{L^2(Q_{\mathcal{Y}})}.$$

By the eigendecomposition of $\tilde{C}_{WW}$ in $L^2(Q_{\mathcal{Y}})$, a similar argument to the proof of Eq. (24) shows the assertion. ■

The consistency of KBR follows by combining the above theorems.

Theorem 6.4 *Let f be a function in $\mathcal{H}_{\mathcal{X}}$, (Z, W) be a random variable that has the distribution Q with p.d.f. $p(y|x)\pi(x)$, and $\widehat{m}_{\Pi}^{(n)}$ be an estimator of m_{Π} such that $\|\widehat{m}_{\Pi}^{(n)} - m_{\Pi}\|_{\mathcal{H}_{\mathcal{X}}} = O_p(n^{-\alpha})$ $(n \to \infty)$ for some $0 < \alpha \leq 1/2$. Assume that $\pi/p_X \in \mathcal{R}(C_{XX}^{\beta})$ with $\beta \geq 0$, and $E[f(Z)|W = \cdot] \in \mathcal{R}(C_{WW}^{\nu})$ for some $\nu \geq 0$. For the regularization constants $\varepsilon_n = n^{-\max\{\frac{2}{3}\alpha, \frac{1}{1+\beta}\alpha\}}$ and $\delta_n = n^{-\max\{\frac{4}{9}\gamma, \frac{4}{2\nu+5}\gamma\}}$, where $\gamma = \min\{\frac{2}{3}\alpha, \frac{2\beta+1}{2\beta+2}\alpha\}$, we have for any $y \in \mathcal{Y}$*

$$\mathbf{f}_X^T R_{X|Y}\mathbf{k}_Y(y) - E[f(Z)|W = y] = O_p(n^{-\min\{\frac{4}{9}\gamma, \frac{2\nu}{2\nu+5}\gamma\}}), \quad (n \to \infty),$$

where $\mathbf{f}_X^T R_{X|Y}\mathbf{k}_Y(y)$ is given by Eq. (14).

Proof By applying Theorem 6.1 to $Y = (Y, X)$ and $Y = (Y, Y)$, we see that both of $\|\widehat{C}_{WZ} - C_{WZ}\|$ and $\|\widehat{C}_{WW} - C_{WW}\|$ are of $O_p(n^{-\gamma})$. Since

$$\begin{aligned}&\mathbf{f}_X^T R_{X|Y}\mathbf{k}_Y(y) - E[f(Z)|W = y] \\ &\qquad = \langle k_{\mathcal{Y}}(\cdot, y), \widehat{C}_{WW}\left((\widehat{C}_{YY})^2 + \delta_n I\right)^{-1}\widehat{C}_{WZ}f - E[f(Z)|W = \cdot]\rangle_{\mathcal{H}_{\mathcal{Y}}},\end{aligned}$$

combination of Theorems 6.1 and 6.2 proves the theorem. ■

The next theorem shows the rate on average w.r.t. $Q_{\mathcal{Y}}$. The proof is similar to the above theorem, and omitted.

Theorem 6.5 *Let f be a function in $\mathcal{H}_{\mathcal{X}}$, (Z, W) be a random variable that has the distribution Q with p.d.f. $p(y|x)\pi(x)$, and $\widehat{m}_{\Pi}^{(n)}$ be an estimator of m_{Π} such that $\|\widehat{m}_{\Pi}^{(n)} - m_{\Pi}\|_{\mathcal{H}_{\mathcal{X}}} = O_p(n^{-\alpha})$ $(n \to \infty)$ for some $0 < \alpha \leq 1/2$. Assume that $\pi/p_X \in \mathcal{R}(C_{XX}^{\beta})$ with $\beta \geq 0$, and $E[f(Z)|W = \cdot] \in \mathcal{R}(\tilde{C}_{WW}^{\nu}) \cap \mathcal{H}_{\mathcal{Y}}$ for some $\nu > 0$. For the regularization constants $\varepsilon_n = n^{-\max\{\frac{2}{3}\alpha, \frac{1}{1+\beta}\alpha\}}$ and $\delta_n = n^{-\max\{\frac{1}{2}\gamma, \frac{2}{\nu+2}\gamma\}}$, where $\gamma = \min\{\frac{2}{3}\alpha, \frac{2\beta+1}{2\beta+2}\alpha\}$, we have*

$$\left\|\mathbf{f}_X^T R_{X|Y}\mathbf{k}_Y(W) - E[f(Z)|W]\right\|_{L^2(Q_{\mathcal{Y}})} = O_p(n^{-\min\{\frac{1}{2}\gamma, \frac{\nu}{\nu+2}\gamma\}}), \quad (n \to \infty).$$

We also have consistency of the estimator for the kernel mean of posterior $m_{Q_{\mathcal{X}|y}}$, if we make stronger assumptions. First, we formulate the expectation with the posterior in terms of operators. Let (Z, W) be a random variable with distribution Q. Assume that for

any $f \in \mathcal{H}_\mathcal{X}$ the conditional expectation $E[f(Z)|W = \cdot]$ is included in $\mathcal{H}_\mathcal{Y}$. We then have a linear operator S defined by

$$S : \mathcal{H}_\mathcal{X} \to \mathcal{H}_\mathcal{Y}, \qquad f \mapsto E[f(Z)|W = \cdot].$$

If we further assume that S is bounded, the adjoint operator $S^* : \mathcal{H}_\mathcal{Y} \to \mathcal{H}_\mathcal{X}$ satisfies

$$\langle S^* k_\mathcal{Y}(\cdot, y), f\rangle_{\mathcal{H}_\mathcal{X}} = \langle k_\mathcal{Y}(\cdot, y), Sf\rangle_{\mathcal{H}_\mathcal{Y}} = E[f(Z)|W = y]$$

for any $y \in \mathcal{Y}$, and thus $S^* k_\mathcal{Y}(\cdot, y)$ is equal to the kernel mean of the conditional probability of Z given $W = y$.

We make the following further assumptions:

Assumption (S)

1. The covariance operator C_{WW} is injective.
2. There exists $\nu > 0$ such that for any $f \in \mathcal{H}_\mathcal{X}$ there is $\eta_f \in \mathcal{H}_\mathcal{X}$ with $Sf = C_{WW}^\nu \eta_f$, and the linear map
$$C_{WW}^{-\nu} S : \mathcal{H}_\mathcal{X} \to \mathcal{H}_\mathcal{Y}, \qquad f \mapsto \eta_f$$
is bounded.

Theorem 6.6 *Let (Z, W) be a random variable that has the distribution Q with p.d.f. $p(y|x)\pi(x)$, and $\widehat{m}_\Pi^{(n)}$ be an estimator of m_Π such that $\|\widehat{m}_\Pi^{(n)} - m_\Pi\|_{\mathcal{H}_\mathcal{X}} = O_p(n^{-\alpha})$ $(n \to \infty)$ for some $0 < \alpha \le 1/2$. Assume (S) above, and $\pi/p_X \in \mathcal{R}(C_{XX}^\beta)$ with some $\beta \ge 0$. For the regularization constants $\varepsilon_n = n^{-\max\{\frac{2}{3}\alpha, \frac{1}{1+\beta}\alpha\}}$ and $\delta_n = n^{-\max\{\frac{4}{9}\gamma, \frac{4}{2\nu+5}\gamma\}}$, where $\gamma = \min\{\frac{2}{3}\alpha, \frac{2\beta+1}{2\beta+2}\alpha\}$, we have for any $y \in \mathcal{Y}$*

$$\left\|\mathbf{k}_X^T R_{X|Y} \mathbf{k}_Y(y) - m_{Q_{\mathcal{X}|y}}\right\|_{\mathcal{H}_\mathcal{X}} = O_p(n^{-\min\{\frac{4}{9}\gamma, \frac{2\nu}{2\nu+5}\gamma\}}),$$

as $n \to \infty$, where $m_{Q_{\mathcal{X}|y}}$ is the kernel mean of the posterior given y.

Proof First, in a similar manner to the proof of Eq. (23), we have

$$\left\|\widehat{C}_{ZW}^{(n)}\left((\widehat{C}_{WW}^{(n)})^2 + \delta_n I\right)^{-1} \widehat{C}_{WW}^{(n)} k_\mathcal{Y}(\cdot, y) - C_{ZW}(C_{WW}^2 + \delta_n I)^{-1} C_{WW} k_\mathcal{Y}(\cdot, y)\right\|_{\mathcal{H}_\mathcal{X}} = O_p(n^{-\gamma}\delta_n^{-5/4}).$$

The assertion is thus obtained if

$$\left\|C_{ZW}(C_{WW}^2 + \delta_n I)^{-1} C_{WW} k_\mathcal{Y}(\cdot, y) - S^* k_\mathcal{Y}(\cdot, y)\right\|_{\mathcal{H}_\mathcal{X}} = O(\delta_n^{\min\{1, \frac{\nu}{2}\}}) \tag{25}$$

is proved. The left hand side of Eq. (25) is upper-bounded by

$$\left\|C_{ZW}(C_{WW}^2 + \delta_n I)^{-1} C_{WW} - S^*\right\| \|k_\mathcal{Y}(\cdot, y)\|_{\mathcal{H}_\mathcal{Y}} = \left\|C_{WW}(C_{WW}^2 + \delta_n I)^{-1} C_{WZ} - S\right\| \|k_\mathcal{Y}(\cdot, y)\|_{\mathcal{H}_\mathcal{Y}}.$$

It follows from Theorem 3.1 that $C_{WZ} = C_{WW} S$, and thus $\|C_{WW}(C_{WW}^2 + \delta_n I)^{-1} C_{WZ} - S\| = \|C_{WW}(C_{WW}^2 + \delta_n I)^{-1} C_{WW} S - S\| \le \delta_n \|(C_{WW}^2 + \delta_n I)^{-1} C_{WW}^\nu\| \, \|C_{WW}^{-\nu} S\|$. The eigen-decomposition of C_{WW} together with the inequality $\frac{\delta_n \lambda^\nu}{\lambda^2 + \delta_n} \le \delta_n^{\min\{1, \nu/2\}}$ $(\lambda \ge 0)$ completes the proof. ∎

Acknowledgements

We thank Arnaud Doucet, Lorenzo Rosasco, Yee Whye Teh and Shuhei Mano for their helpful comments. KF has been supported in part by JSPS KAKENHI (B) 22300098.

References

N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.

C.R. Baker. Joint measures and cross-covariance operators. *Transactions of the American Mathematical Society*, 186:273–289, 1973.

A. Berlinet and C. Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Kluwer Academic Publisher, 2004.

D. Blei and M. Jordan. Variational inference for dirichlet process mixtures. *Journal of Bayesian Analysis*, 1(1):121–144, 2006.

Aedian W. Bowman. An alternative method of cross-validation for the smoothing of density estimates. *Biometrika*, 71(2):353–360, 1984.

A. Caponnetto and E. De Vito. Optimal rates for regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

A. Doucet, N. De Freitas, and N.J. Gordon. *Sequential Monte Carlo Methods in Practice*. Springer, 2001.

H.W. Engl, M. Hanke, and A. Neubauer. *Regularization of Inverse Problems*. Kluwer Academic Publishers, 2000.

S. Fine and K. Scheinberg. Efficient SVM training using low-rank kernel representations. *Journal of Machine Learning Research*, 2:243–264, 2001.

K. Fukumizu, F.R. Bach, and M.I. Jordan. Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *Journal of Machine Learning Research*, 5:73–99, 2004.

K. Fukumizu, F.R. Bach, and M.I. Jordan. Kernel dimension reduction in regression. *Annals of Statistics*, 37(4):1871–1905, 2009a. ISSN 0090-5364. doi: 10.1214/08-AOS637.

K. Fukumizu, B. Sriperumbudur, A. Gretton, and B. Schoelkopf. Characteristic kernels on groups and semigroups. In *Advances in Neural Information Processing Systems 21*, pages 473–480, Red Hook, NY, 2009b. Curran Associates Inc.

Kenji Fukumizu, Arthur Gretton, Xiaohai Sun, and Bernhard Schölkopf. Kernel measures of conditional dependence. In *Advances in Neural Information Processing Systems 20*, pages 489–496. MIT Press, 2008.

A. Gretton, K.M. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola. A kernel method for the two-sample-problem. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 513–520. MIT Press, Cambridge, MA, 2007.

A. Gretton, K. Fukumizu, Z. Harchaoui, and B. Sriperumbudur. A fast, consistent kernel two-sample test. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 673–681. 2009a.

Arthur Gretton, Kenji Fukumizu, Choon Hui Teo, Le Song, Bernhard Schölkopf, and Alex Smola. A kernel statistical test of independence. In *Advances in Neural Information Processing Systems 20*, pages 585–592. MIT Press, 2008.

Arthur Gretton, Kenji Fukumizu, and Bharath K. Sriperumbudur. Discussion of: Brownian distance covariance. *Annals of Applied Statistics*, 3(4):1285–1294, 2009b.

Thomas Hofmann, Bernhard Schölkopf, and Alexander J. Smola. Kernel methods in machine learning. *The Annals of Statistics*, 36(3):1171–1220, 2008.

S.J. Julier and J.K. Uhlmann. A new extension of the kalman filter to nonlinear systems. In *Proceedings of AeroSense: The 11th International Symposium on Aerospace/Defence Sensing, Simulation and Controls*, 1997.

A. Kankainen and N.G. Ushakov. A consistent modification of a test for independence based on the empirical characteristic function. *Journal of Mathematical Sciencies*, 89: 1582–1589, 1998.

S MacEachern. Estimating normal means with a conjugate style dirichlet process prior. *Communications in Statistics – Simulation and Computation*, 23(3):727–741, 1994.

Steven N. MacEachern, Merlise Clyde, and Jun S. Liu. Sequential importance sampling for nonparametric bayes models: The next generation. *The Canadian Journal of Statistics*, 27(2):251–267, 1999.

Paul Marjoram, John Molitor, Vincent Plagnol, and Simon Tavare. Markov chain monte carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 100(26): 15324–15328, 2003.

Sebastian Mika, Bernhard Schölkopf, Alex Smola, Klaus-Robert Müller, Matthias Scholz, and Gunnar Rätsch. Kernel PCA and de-noising in feature spaces. In *Advances in Neural Information Pecessing Systems 11*, pages 536–542. MIT Press, 1999.

V. Monbet, P. Ailliot, and P.F. Marteau. l^1-convergence of smoothing densities in non-parametric state space models. *Statistical Inference for Stochastic Processes*, 11:311–325, 2008.

P. Müller and F.A. Quintana. Nonparametric bayesian data analysis. *Statistical Science*, 19(1):95–110, 2004.

Mats Rudemo. Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics*, 9(2):pp. 65–78, 1982.

B. Schölkopf and A.J. Smola. *Learning with Kernels*. MIT Press, 2002.

S. A. Sisson, Y. Fan, and Mark M. Tanaka. Sequential monte carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 104(6):1760–1765, 2007.

Steve Smale and Ding-Xuan Zhou. Learning theory estimates via integral operators and their approximation. *Constructive Approximation*, 26:153–172, 2007.

L. Song, J. Huang, A. Smola, and K. Fukumizu. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *Proceedings of the 26th International Conference on Machine Learning (ICML2009)*, pages 961–968. 2009.

L. Song, A. Gretton., and C. Guestrin. Nonparametric tree graphical models via kernel embeddings. In *Proceedings of AISTATS 2010*, pages 765–772, 2010a.

L. Song, S. M. Siddiqi, G. Gordon, and A. Smola. Hilbert space embeddings of hidden markov models. In *Proceedings of the 27th International Conference on Machine Learning (ICML2010)*, pages 991–998. 2010b.

L. Song, A. Gretton, D. Bickson, Y. Low, and C. Guestrin. Kernel belief propagation. In *Proceedings of AISTATS 2011*, pages 707–715, 2011.

Bharath K. Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert R.G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11:1517–1561, 2010.

Bharath K. Sriperumbudur, Kenji Fukumizu, and Gert Lanckriet. Universality, characteristic kernels and rkhs embedding of measures. *Journal of Machine Learning Research*, 12:2389–2410, 2011.

I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, 2:67–93, 2001.

Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Information Science and Statistics. Springer, 2008.

S. Tavaré, D.J. Balding, R.C. Griffithis, and P. Donnelly. Inferring coalescence times from dna sequece data. *Genetics*, 145:505–518, 1997.

S. Thrun, J. Langford, and D. Fox. Monte carlo hidden markov models: Learning nonparametric models of partially observable stochastic processes. In *Proceedings of International Conference on Machine Learning (ICML 1999)*, pages 415–424, 1999.

Mike West, Peter Müller, and Michael D. Escobar. Hierarchical priors and mixture models, with applications in regression and density estimation. In P. Freeman et al, editor, *Aspects of Uncertainty: A Tribute to D.V. Lindley*, pages 363–386. 1994.