Journal of Machine Learning Research 14 (2013) 3753-3783 Submitted 12/11; Revised 6/13; Published 12/13

Kernel Bayes' Rule: Bayesian Inference with Positive Definite Kernels

Kenji Fukumizu FUKUMIZU@ISM.AC.JP
The Institute of Statistical Mathematics
10-3 Midoricho, Tachikawa
Tokyo 190-8562 Japan

Le Song LSONG@CC.GATECH.EDU
College of Computing
Georgia Institute of Technology
1340 Klaus Building, 266 Ferst Drive
Atlanta, GA 30332, USA

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

Editor: Ingo Steinwart

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 Bayesian computation without likelihood and filtering with a nonparametric state-space model.

Keywords: kernel method, Bayes' rule, 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) (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; Harchaoui et al., 2008; Gretton et al., 2009a, 2012), 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; Zhang et al., 2011).

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

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

Our main result is a nonparametric estimate of posterior kernel mean, given kernel mean representations of the prior and likelihood. We call this method *kernel Bayes' rule*.

A valuable property of the kernel Bayes' rule is that the kernel posterior mean is estimated nonparametrically from data. The prior is represented by a weighted sum over a sample, and the probabilistic relation expressed by the likelihood is represented in terms of a sample from a joint distribution having the desired conditional probability. 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 (Tavaré et al., 1997; Marjoram et al., 2003; Sisson et al., 2007) is widely applied in population genetics, where the likelihood is expressed as 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 consistency for our posterior kernel mean estimate, and derive convergence rates 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. In addition, computation of the posterior with KDE requires importance weights, which may not be accurate in low density areas. 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 by 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 Kernels 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 the following three conditions hold:

(i) $k(\cdot, x) \in \mathcal{H}$ for any $x \in \Omega$,

(ii) Span$\{k(\cdot, x) \mid x \in \Omega\}$ is dense in $\mathcal{H}$,

(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 $k_x = k(\cdot, x)$ 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 $P_\mathcal{X}$, 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 will generally write m_X in place of m_X^k for simplicity, where there is no ambiguity. By the reproducing property, the kernel mean satisfies the relation

$$\langle f, m_X \rangle = E[f(X)] \tag{2}$$

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

$$m_X(u) = E[k(u,X)] = \int k(u,\tilde{x}) dP_X(\tilde{x}), \tag{3}$$

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 Equation (3), the explicit integral expressions for C_{YX} and C_{XX} are given by

$$(C_{YX} f)(y) = \int k_\mathcal{Y}(y,\tilde{y}) f(\tilde{x}) dP(\tilde{x},\tilde{y}) \quad \text{and} \quad (C_{XX} f)(x) = \int k_\mathcal{X}(x,\tilde{x}) f(\tilde{x}) dP_\mathcal{X}(\tilde{x}), \tag{4}$$

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 $k(x,y) = \exp(-\|x-y\|^2/(2\sigma^2))$. 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 for arbitrary probabilities.

Given i.i.d. sample $(X_1,Y_1), \ldots, (X_n,Y_n)$ with law P, the empirical estimators 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. 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, Section 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

We review Bayes' rule and the notion of kernel conditional mean embeddings in Section 3.1. We demonstrate that Bayes' rule may be expressed in terms of these conditional mean embeddings. We provide consistency results for the empirical estimators of the conditional mean embedding for the posterior in Section 3.2.

3.1 Kernel Bayes' Rule

We first review Bayes' rule in a general form without using density functions, since the kernel Bayes' rule can be applied to situations where density functions are not available.

Let $(\mathcal{X}, \mathcal{B}_\mathcal{X})$ and $(\mathcal{Y}, \mathcal{B}_\mathcal{Y})$ be measurable spaces, $(\Omega, \mathcal{A}, \mathbb{P})$ a probability space, and $(X,Y) : \Omega \to \mathcal{X} \times \mathcal{Y}$ be a $(\mathcal{X} \times \mathcal{Y}$-valued) random variable with distribution P. The marginal distribution of X is denoted by $P_\mathcal{X}$. Suppose that Π is a probability measure on $(\mathcal{X}, \mathcal{B}_\mathcal{X})$, which serves as a *prior* distribution. For each $x \in \mathcal{X}$, let $P_{\mathcal{Y}|x}$ denote the conditional probability of Y given $X = x$; namely, $P_{\mathcal{Y}|x}(B) = E[I_B(Y)|X = x]$, where I_B is the indicator function of a measurable set $B \in \mathcal{B}_\mathcal{Y}$.[1] We assume that the conditional probability $P_{\mathcal{Y}|x}$ is *regular*; namely, it defines a probability measure on $\mathcal{Y}$ for each x. The prior Π and the family $\{P_{\mathcal{Y}|x} \mid x \in \mathcal{X}\}$ defines the joint distribution Q on $\mathcal{X} \times \mathcal{Y}$ by

$$Q(A \times B) = \int_A P_{\mathcal{Y}|x}(B) d\Pi(x) \tag{5}$$

for any $A \in \mathcal{B}_\mathcal{X}$ and $B \in \mathcal{B}_\mathcal{Y}$, and its marginal distribution $Q_\mathcal{Y}$ by

$$Q_\mathcal{Y}(B) = Q(\mathcal{X} \times B).$$

Let (Z,W) be a random variable on $\mathcal{X} \times \mathcal{Y}$ with distribution Q. 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 a measure $\nu_\mathcal{X}$ 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, Equations (5) and (6) are reduced to the well known form Equation (1). To make Bayesian inference meaningful, we make the following assumption:

(A) The prior Π is absolutely continuous with respect to the marginal distribution $P_\mathcal{X}$.

1. The $\mathcal{B}_\mathcal{X}$-measurable function $P_{\mathcal{Y}|x}(B)$ is always well defined. In fact, the finite measure $\mu(A) := \int_{\{\omega \in \Omega | X(\omega) \in A\}} I_B(Y) dP$ on $(\mathcal{X}, \mathcal{B}_\mathcal{X})$ is absolutely continuous with respect to $P_\mathcal{X}$. The Radon-Nikodym theorem then guarantees the existence of a $\mathcal{B}_\mathcal{X}$-measurable function $\eta(x)$ such that $\int_{\{\omega \in \Omega | X(\omega) \in A\}} I_B(Y) dP = \int_A \eta(x) dP_\mathcal{X}(x)$. We can define $P_{\mathcal{Y}|x}(B) := \eta(x)$. Note that $P_{\mathcal{Y}|x}(B)$ may not satisfy the σ-additivity in general. For details on conditional probability, see, for example, Shiryaev (1995, §9).

The conditional probability $P_{\mathcal{Y}|x}(B)$ can be uniquely determined only almost surely with respect to P_X. It is thus possible to define Q appropriately only if assumption (A) holds.

In ordinary Bayesian inference, we need only the conditional probability density (likelihood) $p(y|x)$ and prior $\pi(x)$, and not the joint distribution P. In kernel methods, however, the information on the relation between variables is expressed by covariance, which leads to finite sample estimates in terms of Gram matrices, as we see below. It is then necessary to assume the existence of the variable (X,Y) on $\mathcal{X}\times\mathcal{Y}$ with probability P, which gives the conditional probability $P_{\mathcal{Y}|x}$ by conditioning on $X=x$.

Let $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$ be positive definite kernels on $\mathcal{X}$ and $\mathcal{Y}$, respectively, with respective RKHS $\mathcal{H}_{\mathcal{X}}$ and $\mathcal{H}_{\mathcal{Y}}$. The goal of this subsection is to derive an estimator of the kernel mean of posterior $m_{Q_{X|y}}$. The following theorem is fundamental to discuss conditional probabilities with positive definite kernels.

Theorem 1 (Fukumizu et al., 2004) *If $E[g(Y)|X=\cdot]\in\mathcal{H}_{\mathcal{X}}$ holds[2] for $g\in\mathcal{H}_{\mathcal{Y}}$, then*

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

The above relation motivates to introduce a regularized approximation of the conditional expectation

$$\big(C_{XX}+\varepsilon I\big)^{-1}C_{XY}g,$$

which is shown to converge to $E[g(Y)|X=\cdot]$ in $\mathcal{H}_{\mathcal{X}}$ under appropriate assumptions, as we will see later.

Using Theorem 1, we have the following result, which expresses the kernel mean of $Q_{\mathcal{Y}}$, and implements the Sum Rule in terms of mean embeddings.

Theorem 2 (Song et al., 2009, Equation 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,[3] $m_\Pi\in\mathcal{R}(C_{XX})$, and $E[g(Y)|X=\cdot]\in\mathcal{H}_{\mathcal{X}}$ for any $g\in\mathcal{H}_{\mathcal{Y}}$, then*

$$m_{Q_{\mathcal{Y}}}=C_{YX}C_{XX}^{-1}m_\Pi, \tag{7}$$

where $C_{XX}^{-1}m_\Pi$ denotes the function mapped to m_Π by C_{XX}.

Proof Take $f\in\mathcal{H}_{\mathcal{X}}$ such that $C_{XX}f=m_\Pi$. It follows from Theorem 1 that 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 by Song et al. (2009), we can regard the operator $C_{YX}C_{XX}^{-1}$ as the kernel expression of the conditional probability $P_{\mathcal{Y}|x}$ or $p(y|x)$. Note, however, that the assumptions $m_\Pi\in\mathcal{R}(C_{XX})$ and $E[g(Y)|X=\cdot]\in\mathcal{H}_{\mathcal{X}}$ may not hold in general; we can easily give counterexamples for the latter in the case of Gaussian kernels.[4] A regularized inverse $(C_{XX}+\varepsilon I)^{-1}$ can be used to remove this

2. The assumption "$E[g(Y)|X=\cdot]\in\mathcal{H}_{\mathcal{X}}$" means that a version of the conditional expectation $E[g(Y)|X=x]$ is included in $\mathcal{H}_{\mathcal{X}}$ as a function of x.
3. 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.
4. 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).

strong assumption. An alternative way of obtaining the regularized conditional mean embedding is as the solution to a vector-valued ridge regression problem, as proposed by Grünewälder et al. (2012) and Grünewälder et al. (2013). A connection between conditional embeddings and ridge regression was noted independently by Zhang et al. (2011, Section 3.5). Following Grünewälder et al. (2013, Section 3.2), we seek a bounded linear operator $F : \mathcal{H}_\mathcal{Y} \to \mathcal{H}_\mathcal{X}$ that minimizes the loss

$$\mathcal{E}_c[F] = \sup_{\|h\|_{\mathcal{H}_\mathcal{Y}} \leq 1} E\left(E[h(Y)|X] - \langle Fh, k_\mathcal{X}(X,\cdot)\rangle_{\mathcal{H}_\mathcal{X}}\right)^2,$$

where we take the supremum over the unit ball in $\mathcal{H}_\mathcal{Y}$ to ensure worst-case robustness. Using the Jensen and Cauchy-Schwarz inequalities, we may upper bound this as

$$\mathcal{E}_c[F] \leq E_u[F] := E\|k_\mathcal{Y}(Y,\cdot) - F^*[k_\mathcal{X}(X,\cdot)]\|^2_{\mathcal{H}_\mathcal{Y}},$$

where F^* denotes the adjoint of F. If we stipulate that[5] $F^* \in \mathcal{H}_\mathcal{Y} \otimes \mathcal{H}_\mathcal{X}$, and regularize by the squared Hilbert-Schmidt norm of F^*, we obtain the ridge regression problem

$$\operatorname*{argmin}_{F^* \in \mathcal{H}_\mathcal{Y} \otimes \mathcal{H}_\mathcal{X}} E\|k_\mathcal{Y}(Y,\cdot) - F^*[k_\mathcal{X}(X,\cdot)]\|^2_{\mathcal{H}_\mathcal{Y}} + \varepsilon\|F^*\|^2_{\mathrm{HS}},$$

which has as its solution the regularized kernel conditional mean embedding. In the following, we nonetheless use the unregularized version to derive a population expression of Bayes' rule, use it as a prototype for defining an empirical estimator, and prove its consistency.

Equation (7) has a simple interpretation if the probability density function or Radon-Nikodym derivative $d\Pi/dP_\mathcal{X}$ is included in $\mathcal{H}_\mathcal{X}$ under Assumption (A). From Equation (3) we have $m_\Pi(x) = \int k_\mathcal{X}(x,\tilde{x})d\Pi(\tilde{x}) = \int k_\mathcal{X}(x,\tilde{x})(d\Pi/dP_\mathcal{X})(\tilde{x})dP_\mathcal{X}(\tilde{x})$, which implies $C_{XX}^{-1}m_\Pi = d\Pi/dP_\mathcal{X}$ from Equation (4) under the injective assumption of C_{XX}. Thus Equation (7) is an operator expression of the obvious relation

$$\int\int k_\mathcal{Y}(y,\tilde{y})dP_{\mathcal{Y}|x}(\tilde{y})d\Pi(\tilde{x}) = \int k_\mathcal{Y}(y,\tilde{y})\Big(\frac{d\Pi}{dP_\mathcal{X}}\Big)(\tilde{x})dP(\tilde{x},\tilde{y}).$$

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 Equation (7), we have a population expression[6]

$$E[k_\mathcal{Y}(\cdot,Y)|X = x] = C_{YX}{C_{XX}}^{-1}k_\mathcal{X}(\cdot,x), \tag{8}$$

or more rigorously we can consider a regularized inversion

$$E^{reg}_\varepsilon[k_\mathcal{Y}(\cdot,Y)|X = x] := C_{YX}(C_{XX} + \varepsilon I)^{-1}k_\mathcal{X}(\cdot,x) \tag{9}$$

as an approximation of the conditional expectation $E[k_\mathcal{Y}(\cdot,Y)|X = x]$. Note that in the latter expression we do not need to assume $k_\mathcal{X}(\cdot,x) \in \mathcal{R}(C_{XX})$, which is too strong in many situations. We

5. Note that more complex vector-valued RKHS are possible; see, for example, Micchelli and Pontil (2005).

6. The expression Equation (8) 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 2 may not hold in general. 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 $k_\mathcal{X}(x,\tilde{x}) = \int h_x(x')k_\mathcal{X}(\tilde{x},x')dP_\mathcal{X}(x')$, which is not possible for many popular kernels, including the Gaussian kernel.

will show in Theorem 8 that under some mild conditions a regularized empirical estimator based on Equation (9) is a consistent estimator of $E[k_{\mathcal{Y}}(\cdot,Y)|X=x]$.

To derive kernel realization of Bayes' rule, suppose that we know the covariance operators C_{ZW} and C_{WW} for the random variable $(Z,W) \sim Q$, where Q is defined by Equation (5). The conditional probability $E[k_{\mathcal{X}}(\cdot,Z)|W=y]$ is then exactly the kernel mean of the posterior distribution for observation $y \in \mathcal{Y}$. Equation (9) gives the regularized approximate of the kernel mean of posterior;

$$C_{ZW}(C_{WW}+\delta I)^{-1}k_{\mathcal{Y}}(\cdot,y), \tag{10}$$

where δ is a positive regularization constant. The remaining task is thus to derive the covariance operators C_{ZW} and C_{WW}. This can be done by recalling 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 $C_{ZW} : \mathcal{H}_{\mathcal{Y}} \to \mathcal{H}_{\mathcal{X}}$ by the standard identification of a tensor $\sum_i f_i \otimes g_i$ ($f_i \in \mathcal{H}_{\mathcal{X}}$ and $g_i \in \mathcal{H}_{\mathcal{Y}}$) and a Hilbert-Schmidt operator $h \mapsto \sum_i g_i \langle f_i, h\rangle_{\mathcal{H}_{\mathcal{X}}}$. From Theorem 2, the kernel mean m_Q is given by the following tensor representation:

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

where 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 $(\mathcal{Y}\times\mathcal{X})\times\mathcal{X}$. We can alternatively and more generally use an approximation by the regularized inversion $m_Q^{reg} = C_{(YX)X}(C_{XX}+\delta)^{-1}m_\Pi$, as in Equation (9). This expression provides the covariance operator C_{ZW}. Similarly, the kernel mean $m_{(WW)}$ on the product space $\mathcal{H}_{\mathcal{Y}} \otimes \mathcal{H}_{\mathcal{Y}}$ is identified with C_{WW}, and the expression

$$m_{(WW)} = C_{(YY)X}C_{XX}^{-1}m_\Pi$$

gives a way of estimating the operator C_{WW}.

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 we need to express the information in the variables in terms of Gram matrices given by data points, we assume 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. Such negative weights may appear in successive applications of the kernel Bayes rule, as in the state-space example of Section 4.3. Based on Equation (9), the empirical estimators for $m_{(ZW)}$ and $m_{(WW)}$ are defined respectively by

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

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 *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

$$\widehat{\mu} = \Big(\frac{1}{n} G_X + \varepsilon_n I_n\Big)^{-1} \widehat{\mathbf{m}}_{\Pi}, \quad \widehat{\mathbf{m}}_{\Pi,i} = \widehat{m}_{\Pi}(X_i) = \sum_{j=1}^{\ell} \gamma_j k_{\mathcal{X}}(X_i, U_j). \tag{11}$$

The proof is similar to that of Proposition 4 below, and is omitted. The expressions in Proposition 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, the operator inversion $(\widehat{C}_{WW} + \delta_n I)^{-1}$ in Equation (10) may be impossible or unstable. We thus use another type of Tikhonov regularization,[7] resulting in the estimator

$$\widehat{m}_{Q_{X|y}} := \widehat{C}_{ZW} \big(\widehat{C}_{WW}^2 + \delta_n I\big)^{-1} \widehat{C}_{WW} k_{\mathcal{Y}}(\cdot, y). \tag{12}$$

Proposition 4 *For any $y \in \mathcal{Y}$, the Gram matrix expression of $\widehat{m}_{Q_{X|y}}$ is given by*

$$\widehat{m}_{Q_{X|y}} = \mathbf{k}_X^T R_{X|Y} \mathbf{k}_Y(y), \qquad R_{X|Y} := \Lambda G_Y \big((\Lambda G_Y)^2 + \delta_n I_n\big)^{-1} \Lambda, \tag{13}$$

where $\Lambda = \mathrm{diag}(\widehat{\mu})$ is a diagonal matrix with elements $\widehat{\mu}_i$ in Equation (11), and $\mathbf{k}_X \in {\mathcal{H}_{\mathcal{X}}}^n := \mathcal{H}_{\mathcal{X}} \times \cdots \times \mathcal{H}_{\mathcal{X}}$ (n direct product) and $\mathbf{k}_Y \in {\mathcal{H}_{\mathcal{Y}}}^n := \mathcal{H}_{\mathcal{Y}} \times \cdots \times \mathcal{H}_{\mathcal{Y}}$ are given by

$$\mathbf{k}_X = (k_{\mathcal{X}}(\cdot, X_1), \ldots, k_{\mathcal{X}}(\cdot, X_n))^T \quad \textit{and} \quad \mathbf{k}_Y = (k_{\mathcal{Y}}(\cdot, Y_1), \ldots, k_{\mathcal{Y}}(\cdot, Y_n))^T.$$

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 $\widehat{m}_{Q_{X|y}} = \widehat{C}_{ZW} h = \sum_{i=1}^n \rho_i k_{\mathcal{X}}(\cdot, X_i)$ 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 Equations (12) and (13) 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. Note also that the weights in KBR may take negative values. The interpretation as a probability is then not straightforward, hence the mean

7. An alternative thresholding approach is proposed by Nishiyama et al. (2012), although its consistency remains to be established.

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 $\widehat{\mu} = n(G_X + n\varepsilon_n I_n)^{-1} \widehat{\mathbf{m}}_\Pi$.
   [If the inversion fails, increase ε_n by $\varepsilon_n := c\varepsilon_n$ with $c > 1$.]
3. Compute $R_{X|Y} = \Lambda G_Y ((\Lambda G_Y)^2 + \delta_n I_n)^{-1} \Lambda$, where $\Lambda = \mathrm{diag}(\widehat{\mu})$.
   [If the inversion fails, increase δ_n by $\delta_n := c\delta_n$ with $c > 1$.]

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 weights $\rho = R_{X|Y} \mathbf{k}_Y(y)$, where $\mathbf{k}_Y(y) = (k_\mathcal{Y}(Y_i, y))_{i=1}^n$.

Figure 1: Algorithm for Kernel Bayes' Rule

embedding viewpoint should take precedence (i.e., even if some weights are negative, we may still use result of KBR to estimate posterior expectations of RKHS functions). We will give 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 of Equation (2) gives an estimator

$$\langle f, \widehat{m}_{Q_{X|y}} \rangle_{\mathcal{H}_\mathcal{X}} = \mathbf{f}_X^T R_{X|Y} \mathbf{k}_\mathcal{Y}(y), \tag{14}$$

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 Equation (14). We first show consistency of the estimator, and next the rate of consistency under stronger conditions.

Theorem 5 *Let (X,Y) be a random variable on $\mathcal{X} \times \mathcal{Y}$ with distribution P, (Z,W) be a random variable on $\mathcal{X} \times \mathcal{Y}$ such that the distribution is Q defined by Equation (5), and $\widehat{m}_\Pi^{(\ell_n)}$ be a consistent estimator of m_Π in $\mathcal{H}_\mathcal{X}$ norm. Assume that C_{XX} is injective and that $E[k_\mathcal{Y}(Y,\tilde{Y})|X = x, \tilde{X} = \tilde{x}]$ and $E[k_\mathcal{X}(Z,\tilde{Z})|W = y, \tilde{W} = \tilde{y}]$ are included in the product spaces $\mathcal{H}_\mathcal{X} \otimes \mathcal{H}_\mathcal{X}$ and $\mathcal{H}_\mathcal{Y} \otimes \mathcal{H}_\mathcal{Y}$, respectively, as a function of $(x,\tilde{x})$ and $(y,\tilde{y})$, where $(\tilde{X},\tilde{Y})$ and $(\tilde{Z},\tilde{W})$ are independent copies of (X,Y) and (Z,W), respectively. Then, for any sufficiently slow decay of the regularization coefficients ε_n and δ_n, we have for any $y \in \mathcal{Y}$*

$$\left\| \mathbf{k}_X^T R_{X|Y} \mathbf{k}_Y(y) - m_{Q_{X|y}} \right\|_{\mathcal{H}_\mathcal{X}} \to 0$$

in probability as $n \to \infty$, where $\mathbf{k}_X^T R_{X|Y} \mathbf{k}_Y(y)$ is the KBR estimator given by Equation (13) and $m_{Q_{X|y}} = E[k_\mathcal{X}(\cdot, Z)|W = y]$ is the kernel mean of posterior given $W = y$.

It is obvious from the reproducing property that this theorem also guarantees the consistency of the posterior expectation in Equation (14). The rate of decrease of ε_n and δ_n depends on the convergence rate of $\widehat{m}_\Pi^{(\ell_n)}$ and other smoothness assumptions.

Next, we show convergence rates of the KBR estimator for the expectation with posterior under stronger assumptions. In the following two theorems, we show only the rates that can be derived under certain specific 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.

Theorem 6 *Let f be a function in $\mathcal{H}_\mathcal{X}$, (X,Y) be a random variable on $\mathcal{X}\times\mathcal{Y}$ with distribution P, (Z,W) be a random variable on $\mathcal{X}\times\mathcal{Y}$ with the distribution Q defined by Equation (5), 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 $0<\alpha\le 1/2$. Assume that the Radon Nikodym derivative $d\Pi/dP_\mathcal{X}$ is included in $\mathcal{R}(C_{XX}^{1/2})$, and $E[f(Z)|W=\cdot]\in\mathcal{R}(C_{WW}^2)$. With 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 Equation (14).

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

$$\tilde{C}_{WW}\phi = \int k_\mathcal{Y}(y,w)\phi(w)dQ_\mathcal{Y}(w), \qquad (\phi\in L^2(Q_\mathcal{Y})). \tag{15}$$

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 7 *Let f be a function in $\mathcal{H}_\mathcal{X}$, (Z,W) be a random vector on $\mathcal{X}\times\mathcal{Y}$ with the distribution Q defined by Equation (5), 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 $0<\alpha\le 1/2$. Assume that the Radon Nikodym derivative $d\Pi/dP_\mathcal{X}$ is included in $\mathcal{R}(C_{XX}^{1/2})$, and $E[f(Z)|W=\cdot]\in\mathcal{R}(\tilde{C}_{WW}^2)$. With 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 $d\Pi/dP_\mathcal{X}\in\mathcal{R}(C_{XX}^{1/2})$ 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.

While the convergence rates shown in the above theorems do not depend on the dimensionality of original spaces, the rates may not be optimal. In fact, in the case of kernel ridge regression, the optimal rates are known under additional information on the spectrum of covariance operators (Caponnetto and De Vito, 2007). It is also known (Eberts and Steinwart, 2011) that, given the target function is in the Sobolev space of order α, the convergence rates is arbitrary close to $O_p(n^{-2\alpha/(2\alpha+d)})$, the best rate for any linear estimator (Stone, 1982), where d is the dimensionality of the predictor. Similar convergence rates for KBR incorporating the information on eigenspectrum or smoothness will be interesting future works, in the light of the equivalence of the conditional mean embedding and operator-valued regression shown by Grünewälder et al. (2012) and Grünewälder et al. (2013).

4. Bayesian Inference with Kernel Bayes' Rule

We discuss problem settings for which KBR may be applied in Section 4.1. We then provide notes on practical implementation in Section 4.2, including a cross-validation procedure for parameter selection and suggestions for speeding computation. In Section 4.3, we apply KBR to the filtering problem in a nonparametric state-space model. Finally, in Section 4.4, we give a brief overview of Approximate Bayesian Computation (ABC), a widely used sample-based method which applies to similar problem domains.

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 under appropriate assumptions, a consistent estimator for the expectation of $f \in \mathcal{H}_X$ can be defined with respect to the posterior. On the other hand, unless $f \in \mathcal{H}_X$ holds, there is no theoretical guarantee that it gives a good estimate. In Section 5.1, we discuss experimental results observed in these situations.

To obtain a point estimate of the posterior on x, Song et al. (2009) propose to use the preimage $\widehat{x} = \arg\min_x \|k_X(\cdot,x) - \mathbf{k}_X^T R_{X|Y} \mathbf{k}_Y(y)\|^2_{\mathcal{H}_X}$, which represents the posterior mean most effectively by one point. We use this approach in the present paper when point estimates are sought. 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 (Tavaré et al., 1997; Marjoram et al., 2003; Sisson et al., 2007).

– In nonparametric Bayesian inference (Müller and Quintana, 2004 and references therein), 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; West et al., 1994; MacEachern et al., 1999, 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 infeasible. Using KBR, however, the expectation of a function of the posterior given different y is obtained simply by taking the inner product as in Equation (14), 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, when 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 replaced by low rank matrix approximations. A popular choice is the incomplete Cholesky factorization (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 $O(nr^2)$, which is linear in the sample size n. It is known that in some typical cases the eigenspectram of a Gram matrix decays fast (Widom, 1963, 1964; Bach and Jordan, 2002). We can therefore expect that the incomplete Cholesky factorization to reduce the computational cost effectively without much degrading estimation accuracy.

Third, kernel choice or model selection is key to effective performance of any kernel method. In the case of KBR, we have three model parameters: the kernel (or its parameter, e.g., the bandwidth), and the regularization parameters 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 or classification, 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 $Q_{X|y}$ 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 $\widehat{m}_{Q_{X|y=Y_i}}$ 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 $\widehat{m}_{Q_{X|y}}^{[-a]}$ 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 $\sum_{a=1}^K \left\| \frac{1}{|T_a|}\sum_{j\in T_a} \widehat{m}_{Q_{X|y=Y_j}}^{[-a]} - \widehat{m}_X^{[a]} \right\|_{\mathcal{H}_X}^2$, where $\widehat{m}_X^{[a]} = \frac{1}{|T_a|}\sum_{j\in T_a} k_X(\cdot, X_j)$.

4.3 Application to a 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,

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

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). Kernel density estimates converges more slowly in L^2 as the dimension d of the space increases, however: for an optimal bandwidth choice, this error drops as $O(n^{-4(4+d)})$ (Wasserman, 2006, Section 6.5). If the goal is to take expectations of smooth functions (i.e., RKHS functions), rather than to obtain consistent density estimates, then we can expect better performance. We show our posterior estimate of the expectation of an RKHS function converges at a rate independent of d. 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 a 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,

$$\widehat{C}_{XY} = \frac{1}{T}\sum_{i=1}^{T} k_{\mathcal{X}}(\cdot,X_i)\otimes k_{\mathcal{Y}}(\cdot,Y_i), \quad \widehat{C}_{X_{+1}X} = \frac{1}{T}\sum_{i=1}^{T} k_{\mathcal{X}}(\cdot,X_{i+1})\otimes k_{\mathcal{X}}(\cdot,X_i), \tag{16}$$

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

While the sample is not i.i.d., it is known that the empirical covariances converge to the covariances with respect to the stationary distribution as $T\to\infty$, under a mixing condition.[8] We therefore use the above estimator for the covariance operators.

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

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

8. One such condition to guarantee the central limit theorem for Markov chains in separable Hilbert spaces is *geometrical ergodicity*. See, for example, Merlevéde et al. (1997) and Stachurski (2012) for details.

where $\alpha_i^{(t)} = \alpha_i^{(t)}(\tilde{y}_1, \ldots, \tilde{y}_t)$ are the coefficients at time t. We wish to derive an update rule to obtain $\alpha^{(t+1)}(\tilde{y}_1, \ldots, \tilde{y}_{t+1})$.

For the forward propagation $p(x_{t+1}|\tilde{y}_1, \ldots, \tilde{y}_t) = \int q(x_{t+1}|x_t)p(x_t|\tilde{y}_1, \ldots, \tilde{y}_t)dx_t$, based on Equation (9) 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 $\mathbf{k}_{X_{+1}}^T = (k_{\mathcal{X}}(\cdot, X_2), \ldots, k_{\mathcal{X}}(\cdot, X_{T+1}))$. Using the similar estimator 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

$$\widehat{\mu}^{(t+1)} = \left(G_X + T\varepsilon_T I_T\right)^{-1} G_{XX_{+1}} \left(G_X + T\varepsilon_T I_T\right)^{-1} G_X \alpha^{(t)}. \tag{17}$$

Here $G_{XX_{+1}}$ is the "transfer" matrix defined by $\left(G_{XX_{+1}}\right)_{ij} = k_{\mathcal{X}}(X_i, X_{j+1})$. From

$$p(x_{t+1}|\tilde{y}_1, \ldots, \tilde{y}_{t+1}) = \frac{p(y_{t+1}|x_{t+1})p(x_{t+1}|\tilde{y}_1, \ldots, \tilde{y}_t)}{\int q(y_{t+1}|x_{t+1})p(x_{t+1}|\tilde{y}_1, \ldots, \tilde{y}_t)dx_{t+1}},$$

the kernel Bayes' rule with the prior $p(x_{t+1}|\tilde{y}_1, \ldots, \tilde{y}_t)$ and the likelihood $q(y_{t+1}|x_{t+1})$ yields

$$\alpha^{(t+1)} = \Lambda^{(t+1)} G_Y \left((\Lambda^{(t+1)} G_Y)^2 + \delta_T I_T\right)^{-1} \Lambda^{(t+1)} \mathbf{k}_Y(\tilde{y}_{t+1}), \tag{18}$$

where $\Lambda^{(t+1)} = \mathrm{diag}(\widehat{\mu}_1^{(t+1)}, \ldots, \widehat{\mu}_T^{(t+1)})$. Equations (17) and (18) 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 Equation (8) 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 in Section 4.2. 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 an approximate 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 simply Equation (8) for the kernel mean of the conditional probability $q(x|y)$. As in Song et al. (2009), the estimator is given by

$$\sum_{t=1}^{n} \nu_j k_X(\cdot, X_t), \quad \nu = (G_Y + n\varepsilon_n I_n)^{-1}\mathbf{k}_Y(y). \tag{19}$$

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 embedding 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 compute. In Section 5.2, we present an experimental evaluation of the trade-off between computation time and accuracy for ABC and KBR.

5. Experimental Results

This section demonstrates experimental results with the KBR estimator. In addition to the basic comparison with kernel density estimation, we show simple experiments for Bayesian inference without likelihood and filtering with nonparametric hidden Markov models. More practical applications are discussed in other papers; Nishiyama et al. (2012) propose a KBR approach to reinforcement learning in partially observable Markov decision processes, Boots et al. (2013) apply KBR to reinforcement learning with predictive state representations, and Nakagome et al. (2013) consider a KBR-based method for problems in population genetics.

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. We should bear in mind, however, that the function $f(x) = x$ does not

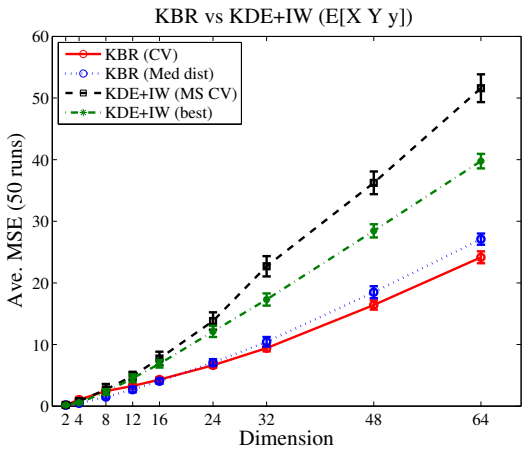


Figure 2: Comparison between KBR and KDE+IW.

belong to the Gaussian kernel RKHS, hence the consistency result of Theorem 6 does not apply to this function. 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 $N(0, V_{XX}/2)$, 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 squares cross-validation (Rudemo, 1982; Bowman, 1984) 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 4.2. In the latter, σ_x for k_X is first chosen with σ_y for k_Y set as the median distances, and then σ_y is chosen with the best σ_x. 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 values of y. We performed 10 runs with different randomly chosen parameter values A, representing 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

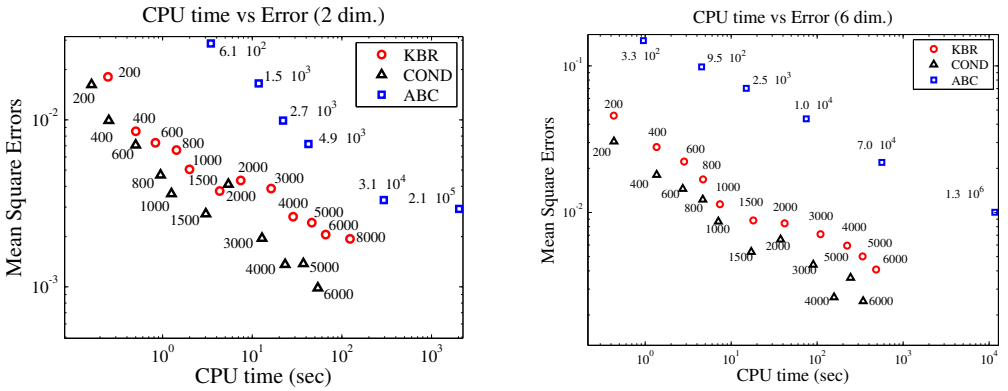


Figure 3: Comparison of estimation accuracy and computational time with KBR and ABC for Bayesian computation without likelihood. COND reprensents the method based on Equation (19). The numbers at the marks are the sample sizes generated for computation.

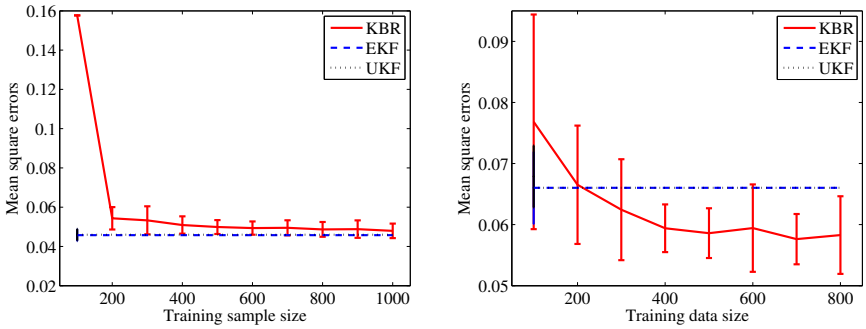
are Gaussian kernels for which the bandwidth parameters are chosen by the median of the pairwise distances on the data (Gretton et al., 2008). The incomplete Cholesky decomposition with tolerance 0.001 is employed for the low-rank approximation. The resulting ranks are, for instance, around 30 for $N = 200$ and around 50 for $N = 6000$ in the case of KBR dimension 2; around 180 for $N = 200$ and 1200 for $N = 6000$ in the case of dimension 6. This implies considerable reduction of computational time especially in the case of large sample sizes. The experimental results indicate that kernel methods achieve more accurate results than ABC at a given computational cost. The conditional kernel mean yields the best results, since in this instance, it is not necessary to correct for a difference in distribution between Π and P_X. In the next experiment, however, this simplification can no longer be made.

5.3 Filtering Problems

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

KBR has the regularization parameters ε_T, δ_T, and kernel parameters for k_X and k_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 Equation (16) 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 Gaussian kernel bandwidths $\beta\sigma_X$ and $\beta\sigma_Y$, where σ_X and σ_Y were the median of pairwise distances in the training samples (Gretton et al., 2008). This left 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 could be controlled, and the problem of camera orientation recovery from an image sequence. In the first case, the hidden state was



Data (a) Data (b)

Figure 4: Comparisons with the KBR Filter and EKF. (Average MSEs and standard errors over 30 runs.) (a): dynamics with weak nonlinearity (b): dynamics with strong nonlinearity.

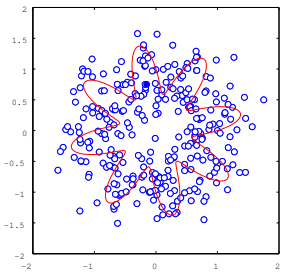


Figure 5: Example of data (b) (X_t, $N = 300$)

$X_t = (u_t, v_t)^T \in \mathbb{R}^2$, and the dynamics were 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) were nonlinear even for $b = 0$. The observation Y_t was

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

where ξ_t was independent noise. The two dynamics were defined as follows: (a) rotation with noisy observations $\eta = 0.3$, $b = 0$, $\sigma_h = \sigma_o = 0.2$; (b) oscillatory rotation with noisy observations $\eta = 0.4$, $b = 0.4$, $M = 8$, $\sigma_h = \sigma_o = 0.2$. (See Fig.5). We assumed the correct dynamics were known to the EKF and UKF.

Results are shown in Fig. 4. In all cases, the EKF and UKF show an indistinguishably small difference. The dynamics in (a) are weakly nonlinear, and KBR has slightly worse MSE than EKF and UKF. For data set (b), which has strong nonlinearity, KBR outperforms the nonlinear Kalman filter for $T \geq 200$.

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 was located at a fixed position, was a

	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 for camera angle estimation (10 runs).

hidden variable, and movie frames recorded by the camera were observed. The data were generated virtually using a computer graphics environment. As in Song et al. (2009), we were given 3600 downsampled frames of 20×20 RGB pixels ($Y_t \in [0,1]^{1200}$), where the first 1800 frames were used for training, and the second half were used to test the filter. We made the data noisy by adding Gaussian noise $N(0, \sigma^2)$ to Y_t.

Our experiments covered two settings. In the first, we did not use that the hidden state S_t was included in $SO(3)$, but only that it was a general 3×3 matrix. In this case, we formulated 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 exploited the fact that $S_t \in SO(3)$: for the Kalman filter, S_t was represented by a quanternion, which is a standard vector representation of rotations; for the KBR filter the kernel $k(A,B) = \mathrm{Tr}[AB^T]$ was used for S_t, and S_t was 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 is able to extract the complex nonlinear dependence between the observation and the hidden state.

6. Proofs

This section includes the proofs of the theoretical results in Section 3.2. The proof ideas are 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). Before stating the main proofs, we provide a proof of consistency of the empirical counterparts in Proposition 3 to the kernel Sum rule (Theorem 2). We then proceed to the proofs of Theorems 5 and 6, covering the consistency of the KBR procedure in RKHS, followed by a proof of Theorem 7 for consistency in L^2.

6.1 Consistency of the Kernel Sum Rule

We show consistency of an empirical estimate of $m_{Q_\mathcal{Y}}$ in Theorem 2. The same proof also applies in establishing consistency for the empirical estimates $\widehat{C}^{(n)}_{WW}$ and $\widehat{C}^{(n)}_{WZ}$ in Proposition 3.

Theorem 8 *Assume that C_{XX} is injective, $\widehat{m}_\Pi$ is a consistent estimator of m_Π in $\mathcal{H}_\mathcal{X}$ norm, and that $E[k_\mathcal{Y}(Y,\tilde{Y})|X = x, \tilde{X} = \tilde{x}]$ is included in $\mathcal{H}_\mathcal{X} \otimes \mathcal{H}_\mathcal{X}$ as a function of $(x,\tilde{x})$, where $(\tilde{X},\tilde{Y})$ is an independent copy of (X,Y). Then, if the regularization coefficient ε_n decays to zero sufficiently slowly,*

$$\left\| \widehat{C}^{(n)}_{YX} \left(\widehat{C}^{(n)}_{XX} + \varepsilon_n I \right)^{-1} \widehat{m}_\Pi - m_{Q_\mathcal{Y}} \right\|_{\mathcal{H}_\mathcal{Y}} \to 0$$

in probability as $n \to \infty$.

Proof The assertion is proved if, as $n \to \infty$,

$$\| \widehat{C}^{(n)}_{YX} (\widehat{C}^{(n)}_{XX} + \varepsilon_n I)^{-1} \widehat{m}_\Pi - C_{YX} (C_{XX} + \varepsilon_n I)^{-1} m_\Pi \|_{\mathcal{H}_\mathcal{Y}} \to 0 \tag{20}$$

in probability and

$$\|C_{YX}(C_{XX}+\varepsilon_n I)^{-1}m_\Pi - m_{Q_\mathcal{Y}}\|_{\mathcal{H}_\mathcal{Y}} \to 0 \tag{21}$$

with an appropriate choice of ε_n.

By using the fact that $B^{-1}-A^{-1}=B^{-1}(A-B)A^{-1}$ holds for any invertible operators A and B, the left hand side of Equation (20) is upper bounded by

$$\begin{aligned}\left\|\widehat{C}_{YX}^{(n)}\left(\widehat{C}_{XX}^{(n)}+\varepsilon_n I\right)^{-1}\left(\widehat{m}_\Pi - m_\Pi\right)\right\|_{\mathcal{H}_\mathcal{Y}} + \left\|\left(\widehat{C}_{YX}^{(n)}-C_{YX}\right)\left(C_{XX}+\varepsilon_n I\right)^{-1}m_\Pi\right\|_{\mathcal{H}_\mathcal{Y}} \\ + \left\|\widehat{C}_{YX}^{(n)}\left(\widehat{C}_{XX}^{(n)}+\varepsilon_n I\right)^{-1}\left(C_{XX}-\widehat{C}_{XX}^{(n)}\right)\left(C_{XX}+\varepsilon_n I\right)^{-1}m_\Pi\right\|_{\mathcal{H}_\mathcal{Y}}. \end{aligned}\tag{22}$$

By the decomposition

$$\widehat{C}_{YX}^{(n)} = \widehat{C}_{YY}^{(n)1/2}\widehat{W}_{YX}^{(n)}\widehat{C}_{XX}^{(n)1/2}$$

with $\|\widehat{W}_{YX}^{(n)}\| \le 1$ (Baker, 1973), we have

$$\|\widehat{C}_{YX}^{(n)}(\widehat{C}_{XX}^{(n)}+\varepsilon_n I)^{-1}\| \le \|\widehat{C}_{YY}^{(n)1/2}\widehat{W}_{YX}^{(n)}(\widehat{C}_{XX}^{(n)}+\varepsilon_n I)^{-1/2}\| = O_p(\varepsilon_n^{-1/2}),$$

which implies the first term is of $O_p(\varepsilon_n^{-1/2}\|\widehat{m}_\Pi - m_\Pi\|_{\mathcal{H}_\mathcal{X}})$. From the $\sqrt{n}$-consistency of the covariance operators, the second and third terms are of the order $O_p(n^{-1/2}\varepsilon_n^{-1})$ and $O_p(n^{-1/2}\varepsilon_n^{-3/2})$, respectively. If ε_n is taken so that $\varepsilon_n \gg n^{-1/3}$ and $\varepsilon_n \gg \|\widehat{m}_\Pi^{(n)} - m_\Pi\|^2_{\mathcal{H}_\mathcal{X}}$, Equation (20) converges to zero in probability.

For Equation (21), first note

$$\begin{aligned}&\|C_{YX}(C_{XX}+\varepsilon_n I)^{-1}m_\Pi - m_{Q_\mathcal{Y}}\|^2_{\mathcal{H}_\mathcal{Y}} \\ &\quad = \|C_{YX}(C_{XX}+\varepsilon_n I)^{-1}m_\Pi\|^2_{\mathcal{H}_\mathcal{X}} - 2\langle C_{YX}(C_{XX}+\varepsilon_n I)^{-1}m_\Pi, m_{Q_\mathcal{Y}}\rangle_{\mathcal{H}_\mathcal{Y}} + \|m_{Q_\mathcal{Y}}\|^2_{\mathcal{H}_\mathcal{Y}}.\end{aligned}$$

Let $\theta(x,\tilde{x}) := E[k_\mathcal{Y}(Y,\tilde{Y})|X=x,\tilde{X}=\tilde{x}]$. The third term in the right hand side is

$$\begin{aligned}\langle m_{Q_\mathcal{Y}}, m_{Q_\mathcal{Y}}\rangle_{\mathcal{H}_\mathcal{Y}} &= \int\int m_{Q_\mathcal{Y}}(y)dP_{\mathcal{Y}|x}(y)d\Pi(x) \\ &= \int\int \langle m_{Q_\mathcal{Y}}, k_\mathcal{Y}(\cdot,y)\rangle_{\mathcal{H}_\mathcal{Y}} dP_{\mathcal{Y}|x}(y)d\Pi(x) = \int\int \theta(x,\tilde{x})d\Pi(x)d\Pi(\tilde{x}).\end{aligned}$$

From the assumption $\theta \in \mathcal{H}_\mathcal{X}\otimes\mathcal{H}_\mathcal{X}$ and the fact $E[m_{Q_\mathcal{Y}}(Y)|X=\cdot] = \int\theta(\cdot,\tilde{x})d\Pi(\tilde{x})$, Lemma 9 below shows $E[m_{Q_\mathcal{Y}}(Y)|X=\cdot]\in\mathcal{H}_\mathcal{X}$. It follows from Theorem 1 that $C_{XY}m_{Q_\mathcal{Y}} = C_{XX}E[m_{Q_\mathcal{Y}}(Y)|X=\cdot]$ and thus

$$\begin{aligned}\langle C_{YX}(C_{XX}+\varepsilon_n I)^{-1}m_\Pi, m_{Q_\mathcal{Y}}\rangle_{\mathcal{H}_\mathcal{Y}} &= \langle m_\Pi, (C_{XX}+\varepsilon_n I)^{-1}C_{XY}m_{Q_\mathcal{Y}}\rangle_{\mathcal{H}_\mathcal{X}} \\ &= \langle m_\Pi, (C_{XX}+\varepsilon_n I)^{-1}C_{XX}E[m_{Q_\mathcal{Y}}(Y)|X=\cdot]\rangle_{\mathcal{H}_\mathcal{X}},\end{aligned}$$

which converges to

$$\langle m_\Pi, E[m_{Q_\mathcal{Y}}(Y)|X=\cdot]\rangle_{\mathcal{H}_\mathcal{X}} = \int\int\theta(x,\tilde{x})d\Pi(x)d\Pi(\tilde{x})$$

from Lemma 10 below.

Note that

$$\|C_{YX}f\|^2_{\mathcal{H}_\mathcal{Y}} = \langle C_{YX}f, C_{YX}f\rangle_{\mathcal{H}_\mathcal{X}} = E[f(X)(C_{YX}f)(Y)] \\ = E[f(X)E[k_\mathcal{Y}(Y,\tilde{Y})f(\tilde{X})]] = E[f(X)f(\tilde{X})\theta(X,\tilde{X})]$$

for any $f \in \mathcal{H}_\mathcal{X}$, where $(\tilde{X},\tilde{Y})$ is an independent copy of (X,Y). By taking $f = (C_{XX}+\varepsilon_n I)^{-1}m_\Pi$, the first term is given by

$$\begin{aligned} &E\left[\theta(X,\tilde{X})((C_{XX}+\varepsilon_n I)^{-1}m_\Pi)(X)((C_{XX}+\varepsilon_n I)^{-1}m_\Pi)(\tilde{X})\right] \\ &= E\left[\theta(X,\tilde{X})f(X)f(\tilde{X})\right] \\ &= \left\langle \theta, (C_{XX}\otimes C_{XX})f\otimes f)\right\rangle_{\mathcal{H}_\mathcal{X}\otimes\mathcal{H}_\mathcal{X}} \\ &= \left\langle \theta, (C_{XX}(C_{XX}+\varepsilon_n I)^{-1}m_\Pi)\otimes(C_{XX}(C_{XX}+\varepsilon_n I)^{-1}m_\Pi)\right\rangle_{\mathcal{H}_\mathcal{X}\otimes\mathcal{H}_\mathcal{X}}, \end{aligned}$$

which, by Lemma 10, converges to

$$\langle \theta, m_\Pi\otimes m_\Pi\rangle_{\mathcal{H}_\mathcal{X}\otimes\mathcal{H}_\mathcal{X}} = \int\int \theta(x,\tilde{x})d\Pi(x)d\Pi(\tilde{x}).$$

This completes the proof. ■

Lemma 9 *Let $\mathcal{H}_\mathcal{X}$ and $\mathcal{H}_\mathcal{Y}$ be RKHS's on $\mathcal{X}$ and $\mathcal{Y}$, respectively. If a function θ on $\mathcal{X}\times\mathcal{Y}$ is in $\mathcal{H}_\mathcal{X}\otimes\mathcal{H}_\mathcal{Y}$ (product space), then for any fixed $y\in\mathcal{Y}$ the function $\theta(\cdot,y)$ of the first argument is in $\mathcal{H}_\mathcal{X}$.*

Proof Let $\{\phi_i\}_{i=1}^I$ and $\{\psi_j\}_{j=1}^J$ be complete orthonormal bases of $\mathcal{H}_\mathcal{X}$ and $\mathcal{H}_\mathcal{Y}$, respectively, where $I,J\in\mathbb{N}\cup\{\infty\}$. Then θ is expressed as

$$\theta = \sum_{i=1}^I\sum_{j=1}^J \alpha_{ij}\phi_i\psi_j$$

with $\sum_{i,j}|\alpha_{ij}|^2<\infty$ (e.g., Aronszajn, 1950). We have $\theta(\cdot,y)=\sum_{i=1}^I\beta_i\phi_i$ with $\beta_i=\sum_{j=1}^J\alpha_{ij}\psi_j(y)$. Since

$$\begin{aligned} &\sum_i|\beta_i|^2 = \sum_i\Big|\sum_j\alpha_{ij}\psi_j(y)\Big|^2 \\ &\le \sum_i\sum_j|\alpha_{ij}|^2\sum_j|\psi_j(y)|^2 = \sum_{ij}|\alpha_{ij}|^2\sum_j\langle\psi_j,k_\mathcal{Y}(\cdot,y)\rangle^2_{\mathcal{H}_\mathcal{Y}} = \sum_{ij}|\alpha_{ij}|^2\|k_\mathcal{Y}(\cdot,y)\|^2_{\mathcal{H}_\mathcal{Y}}<\infty, \end{aligned}$$

we have $\theta(\cdot,y)\in\mathcal{H}_\mathcal{X}$. ■

Lemma 10 *Let $\mathcal{H}$ be a separable Hilbert space and C be a positive, injective, self-adjoint, compact operator on $\mathcal{H}$. then, for any $f\in\mathcal{H}$,*

$$((C+\varepsilon I)^{-1}Cf\to f,\qquad(\varepsilon\to+0).$$

Proof By the assumptions, there exist an orthonormal basis $\{\phi_i\}$ for $\mathcal{H}$ and positive eigenvalues λ_i such that

$$Cf = \sum_i \lambda_i \langle f, \phi_i \rangle_{\mathcal{H}} \phi_i.$$

Then, we have

$$\|(C+\varepsilon I)^{-1} Cf - f\|_{\mathcal{H}}^2 = \sum_i \left| \frac{\varepsilon}{\lambda_i + \varepsilon} \right|^2 |\langle f, \phi_i \rangle_{\mathcal{H}}|^2.$$

Since $\left|(\varepsilon/\lambda_i + \varepsilon)\right|^2 \leq 1$ and $\sum_i |\langle f, \phi_i \rangle_{\mathcal{H}}|^2 = \|f\|_{\mathcal{H}}^2 < \infty$, the dominated convergence theorem ensures

$$\lim_{\varepsilon \to +0} \|(C+\varepsilon I)^{-1} Cf - f\|_{\mathcal{H}}^2 = \sum_i \lim_{\varepsilon \to +0} \left| \frac{\varepsilon}{\lambda_i + \varepsilon} \right|^2 |\langle f, \phi_i \rangle_{\mathcal{H}}|^2 = 0,$$

which completes the proof. ■

6.2 Consistency Results in RKHS Norm

We first prove Theorem 5.

Proof [Proof of Theorem 5] By replacing $Y \mapsto (X,Y)$ and $Y \mapsto (Y,Y)$, it follows from Theorem 8 that $\widehat{C}_{ZW}^{(n)}$ and $\widehat{C}_{WW}^{(n)}$ are consistent estimators of C_{ZW} and C_{WW}, respectively, in operator norm. For the proof, it then suffices to show

$$\left\| \widehat{C}_{ZW}^{(n)} \left(\left(\widehat{C}_{WW}^{(n)} \right)^2 + \delta_n I \right)^{-1} \widehat{C}_{WW}^{(n)} k_{\mathcal{Y}}(\cdot, y) - C_{ZW} (C_{WW}^2 + \delta_n I)^{-1} C_{WW} k_{\mathcal{Y}}(\cdot, y) \right\|_{\mathcal{H}_{\mathcal{X}}} \to 0$$

in probability and

$$\left\| C_{ZW} (C_{WW}^2 + \delta_n I)^{-1} C_{WW} k_{\mathcal{Y}}(\cdot, y) - m_{Q_{X|y}} \right\|_{\mathcal{H}_{\mathcal{X}}} \to 0$$

with an appropriate choice of δ_n. The proof of the first convergence is similar to the proof of Equation (20) in Theorem 8, and we omit it. The proof of the second convergence is also similar to Theorem 8. The square of the left hand side is decomposed as

$$\begin{aligned} &\left\| C_{ZW} (C_{WW}^2 + \delta_n I)^{-1} C_{WW} k_{\mathcal{Y}}(\cdot, y) \right\|_{\mathcal{H}_{\mathcal{X}}}^2 \\ &\qquad - 2 \left\langle C_{ZW} (C_{WW}^2 + \delta_n I)^{-1} C_{WW} k_{\mathcal{Y}}(\cdot, y), m_{Q_{X|y}} \right\rangle_{\mathcal{H}_{\mathcal{X}}} + \left\| m_{Q_{X|y}} \right\|_{\mathcal{H}_{\mathcal{X}}}^2. \end{aligned}$$

Let $\xi \in \mathcal{H}_{\mathcal{Y}} \otimes \mathcal{H}_{\mathcal{Y}}$ be defined by $\xi(y, \tilde{y}) := E[k_{\mathcal{X}}(Z, \tilde{Z}) | W = y, \tilde{W} = \tilde{y}]$, where $(\tilde{Z}, \tilde{W})$ be an independent copy of (Z, W). The third term is then equal to

$$\xi(y, y) = E[k_{\mathcal{X}}(Z, \tilde{Z}) | W = y, \tilde{W} = y].$$

For the second term, by the same argument as the proof of Theorem 8, we have

$$\xi(y, \cdot) = E[k_{\mathcal{X}}(Z, \tilde{Z}) | W = y, \tilde{W} = \cdot] \in \mathcal{H}_{\mathcal{Y}}$$

via Lemma 9, and

$$C_{WZ} m_{Q_{X|y}} = C_{WW} E[k_{\mathcal{X}}(Z, \tilde{Z}) | W = \cdot, \tilde{W} = y] = C_{WW} \xi(\cdot, y) \in \mathcal{H}_{\mathcal{Y}}.$$

We then obtain

$$\left\langle C_{ZW}(C_{WW}^2+\delta_n I)^{-1}C_{WW}k_{\mathcal{Y}}(\cdot,y), m_{Q_{X|y}}\right\rangle_{\mathcal{H}_X} = \left\langle k_{\mathcal{Y}}(\cdot,y), (C_{WW}^2+\delta_n I)^{-1}C_{WW}^2\xi(\cdot,y)\right\rangle_{\mathcal{H}_X},$$

which converges to $\xi(y,y)$.

Finally, defining $\varphi_\delta := (C_{WW}^2+\delta I)^{-1}C_{WW}k_{\mathcal{Y}}(\cdot,y)$, the first term is equal to

$$\begin{aligned}
&E\left[\varphi_\delta(W)\varphi_\delta(\tilde{W})E[k_{\mathcal{X}}(Z,\tilde{Z})|W,\tilde{W}]\right] = \left\langle C_{ZW}\varphi_\delta, C_{ZW}\varphi_\delta\right\rangle_{\mathcal{H}_X} = E\left[\varphi_\delta(W)\left[C_{ZW}\varphi_\delta\right](Z)\right]\\
&= E\left[k(Z,Z')\varphi_\delta(W)\varphi_\delta(\tilde{W})\right] = E\left[\xi(W,\tilde{W})\varphi_\delta(W)\varphi_\delta(\tilde{W})\right]\\
&= \left\langle (C_{WW}\otimes C_{WW})\varphi_\delta\otimes\varphi_\delta, \xi\right\rangle_{\mathcal{H}_{\mathcal{Y}}\otimes\mathcal{H}_{\mathcal{Y}}} = \left\langle (C_{WW}\varphi_\delta)\otimes(C_{WW}\varphi_\delta), \xi\right\rangle_{\mathcal{H}_{\mathcal{Y}}\otimes\mathcal{H}_{\mathcal{Y}}},
\end{aligned}$$

which converges to $\langle k_{\mathcal{Y}}(\cdot,y)\otimes k_{\mathcal{Y}}(\cdot,y), \xi\rangle_{\mathcal{H}_{\mathcal{Y}}\otimes\mathcal{H}_{\mathcal{Y}}} = \xi(y,y)$. This completes the proof. ■

We next show the convergence rate of expectation (Theorem 6) under stronger assumptions. The first result is a rate of convergence for the mean transition in Theorem 2. In the following, $\mathcal{R}(C_{XX}^0)$ means $\mathcal{H}_X$.

Theorem 11 *Assume that the Radon-Nikodym derivative $d\Pi/dP_X$ is included in $\mathcal{R}(C_{XX}^\beta)$ for some $\beta \geq 0$, and let $\widehat{m}_\Pi^{(n)}$ be an estimator of m_Π such that $\|\widehat{m}_\Pi^{(n)} - m_\Pi\|_{\mathcal{H}_X} = O_p(n^{-\alpha})$ 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 $\eta\in\mathcal{H}_X$ such that $d\Pi/dP_X = C_{XX}^\beta\eta$. Then, we have

$$m_\Pi = \int k_{\mathcal{X}}(\cdot,x)\left(\frac{d\Pi}{dP_{\mathcal{X}}}\right)(x)dP_{\mathcal{X}}(x) = C_{XX}^{\beta+1}\eta. \tag{23}$$

First we show the rate of the estimation error:

$$\left\|\widehat{C}_{YX}^{(n)}\left(\widehat{C}_{XX}^{(n)}+\varepsilon_n I\right)^{-1}\widehat{m}_\Pi^{(n)} - C_{YX}\left(C_{XX}+\varepsilon_n I\right)^{-1}m_\Pi\right\|_{\mathcal{H}_{\mathcal{Y}}} = O_p\left(n^{-\alpha}\varepsilon_n^{-1/2}\right), \tag{24}$$

as $n\to\infty$. The left hand side of Equation (24) is upper bounded by

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

In a similar manner to derivation of the bound of Equation (22), we obtain Equation (24).

Next, we show the rate for the approximation error

$$\left\|C_{YX}\left(C_{XX}+\varepsilon_n I\right)^{-1}m_\Pi - m_{Q_{\mathcal{Y}}}\right\|_{\mathcal{H}_{\mathcal{Y}}} = O(\varepsilon_n^{\min\{(1+2\beta)/2,1\}}) \qquad (n\to\infty). \tag{25}$$

Let $C_{YX} = C_{YY}^{1/2}W_{YX}C_{XX}^{1/2}$ be the decomposition with $\|W_{YX}\|\leq 1$. It follows from Equation (23) and the relation

$$\begin{aligned}
m_{Q_{\mathcal{Y}}} = \int\int k_{\mathcal{Y}}(\cdot,y)dP_{\mathcal{Y}|x}(y)d\Pi(x) &= \int\int k(\cdot,y)\left(\frac{d\Pi}{dP_{\mathcal{X}}}\right)(x)dP_{\mathcal{Y}|x}(y)dP_{\mathcal{X}}(x)\\
&= \int\int k(\cdot,y)\left(\frac{d\Pi}{dP_{\mathcal{X}}}\right)(x)dP(x,y) = C_{YX}C_{XX}^\beta\eta
\end{aligned}$$

that the left hand side of Equation (25) is upper bounded by

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

For $\beta \geq 1/2$, it follows from

$$\varepsilon_n\|(C_{XX}+\varepsilon_n I)^{-1}C_{XX}^{\beta+1/2}\| \leq \varepsilon_n\|(C_{XX}+\varepsilon_n I)^{-1}C_{XX}\|\,\|C_{XX}^{\beta-1/2}\|$$

that the left hand side of Equation (25) converges to zero in $O(\varepsilon_n)$. If $0 \leq \beta < 1/2$, we have

$$\varepsilon_n\|(C_{XX}+\varepsilon_n I)^{-1}C_{XX}^{\beta+1/2}\| \\ = \varepsilon_n^{\beta+1/2}\|\varepsilon_n^{1/2-\beta}(C_{XX}+\varepsilon_n I)^{-(1/2-\beta)}\|\,\|(C_{XX}+\varepsilon_n I)^{-\beta-1/2}C_{XX}^{\beta+1/2}\| \leq \varepsilon_n^{\beta+1/2},$$

which proves Equation (25).

With the order of ε_n to balance Equations (24) and (25), the asserted rate of consistency is obtained. ■

The following theorem shows the convergence rate of the estimator used in the second step of KBR.

Theorem 12 *Let f be a function in $\mathcal{H}_X$, 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 that $\widehat{C}_{WZ}^{(n)} : \mathcal{H}_X \to \mathcal{H}_Y$ and $\widehat{C}_{WW}^{(n)} : \mathcal{H}_Y \to \mathcal{H}_Y$ are bounded operators 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 $n\to\infty$*

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

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

$$\big\|\widehat{C}_{WW}^{(n)}\big((\widehat{C}_{WW}^{(n)})^2+\delta_n I\big)^{-1}\widehat{C}_{WZ}^{(n)}f - C_{WW}(C_{WW}^2+\delta_n I)^{-1}C_{WZ}f\big\|_{\mathcal{H}_Y} = O_p(n^{-\gamma}\delta_n^{-5/4}). \qquad (26)$$

The left hand side of Equation (26) is upper bounded by

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

For $A = \widehat{C}_{WW}^{(n)}$, we have

$$\|A(A^2+\delta_n I)^{-1}\| = \|\{A^2(A^2+\delta_n I)^{-1}\}^{1/2}(A^2+\delta_n I)^{-1/2}\| \leq \delta_n^{-1/2},$$

and thus the first term of the above bound is of $O_p(n^{-\gamma}\delta_n^{-1/2})$. A similar argument to 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 Equation (26).

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

$$\left\|C_{WW}(C_{WW}^2 + \delta_n I)^{-1} C_{WZ} f - E[f(Z)|W=\cdot]\right\|_{\mathcal{H}_\mathcal{Y}} = O(\delta_n^{\min\{1,\frac{\nu}{2}\}}). \tag{27}$$

can be proved in the same way as Equation (25).

Finally, combining Equations (26) and (27) proves the assertion. ■

6.3 Consistency Results in L^2

Recall that $\tilde{C}_{WW}$ is the integral operator on $L^2(Q_\mathcal{Y})$ defined by Equation (15). The following theorem shows the convergence rate on average. Here $\mathcal{R}(\tilde{C}_{WW}^0)$ means $L^2(Q_\mathcal{Y})$. In the following the canonical mapping from $\mathcal{H}_\mathcal{Y}$ to $L^2(Q_\mathcal{Y})$ is denoted by $J_\mathcal{Y}$. The mapping $J_\mathcal{X} : \mathcal{H}_\mathcal{X} \to L^2(Q_\mathcal{X})$ is defined similarly.

Theorem 13 *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 $E[f(Z)|W=\cdot] \in \mathcal{H}_\mathcal{Y}$ and $J_\mathcal{Y}E[f(Z)|W=\cdot] \in \mathcal{R}(\tilde{C}_{WW}^{\nu})$ for some $\nu > 0$, and that $\widehat{C}_{WZ}^{(n)} : \mathcal{H}_\mathcal{X} \to \mathcal{H}_\mathcal{Y}$ and $\widehat{C}_{WW}^{(n)} : \mathcal{H}_\mathcal{Y} \to \mathcal{H}_\mathcal{Y}$ are bounded operators 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 $n \to \infty$*

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

where $Q_\mathcal{Y}$ is the marginal distribution of W.

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

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

First, by a similar argument to the proof of Equation (26), it is easy to show that the rate of the estimation error is given by

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

It suffices then to prove

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

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

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

A similar argument to the proof of Equation (27) shows the assertion. ■

The convergence rate of KBR follows by combining the above theorems.

Theorem 14 *Let f be a function in $\mathcal{H}_{\mathcal{X}}$, (Z,W) be a random variable that has the distribution Q defined by Equation (5), 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 the Radon Nikodym derivative $d\Pi/dP_{\mathcal{X}}$ is 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 Equation (13).

Theorem 15 *Let f be a function in $\mathcal{H}_{\mathcal{X}}$, (Z,W) be a random variable that has the distribution Q defined by Equation (5), 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 the Radon Nikodym derivative $d\Pi/dP_{\mathcal{X}}$ is in $\mathcal{R}(C_{XX}^{\beta})$ with $\beta \geq 0$, $E[f(Z)|W = \cdot] \in \mathcal{H}_{\mathcal{Y}}$, and $J_{\mathcal{Y}} E[f(Z)|W = \cdot] \in \mathcal{R}(\tilde{C}_{WW}^{\nu})$ for some $\nu > 0$. With 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*

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

as n goes to infinity.

Acknowledgments

We would like to express our gratitude to Action Editor and anonymous referees for their helpful feedback and suggestions. We also thank Arnaud Doucet, Lorenzo Rosasco, Yee Whye Teh and Shuhei Mano for their valuable comments. KF has been supported in part by JSPS KAKENHI (B) 22300098 and MEXT KAKENHI on Innovative Areas 25120012. LS is supported in part by NSF IIS1218749 and NIH BIGDATA 1R01GM108341-01. AG has been supported in part by the MPI for Intelligent Systems.

References

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

Francis R. Bach and Michael I. Jordan. Kernel independent component analysis. *Journal of Machine Learning Research*, 3:1–48, 2002.

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

Alain Berlinet and Christine Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic Publisher, 2004.

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

Byron Boots, Arthur Gretton, and Geoffrey J. Gordon. Hilbert space embeddings of predictive state representations. In *Proceedings of Conference on Uncertainty in Artificial Intelligence (UAI2013)*, pages 92–101, 2013.

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

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

Arnaud Doucet, Nando De Freitas, and Neil J. Gordon. *Sequential Monte Carlo Methods in Practice*. Springer, 2001.

Mona Eberts and Ingo Steinwart. Optimal learning rates for least squares svms using gaussian kernels. In J. Shawe-Taylor, R.S. Zemel, P. Bartlett, F.C.N. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 1539–1547. Curran Associates, Inc., 2011.

Heinz W. Engl, Martin Hanke, and Andreas Neubauer. *Regularization of Inverse Problems*. Kluwer Academic Publishers, 2000.

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

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

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.

Kenji Fukumizu, Francis R. Bach, and Michael I. Jordan. Kernel dimension reduction in regression. *Annals of Statistics*, 37(4):1871–1905, 2009a.

Kenji Fukumizu, Bhrath Sriperumbudur, Arthur Gretton, and Bernhard Schölkopf. Characteristic kernels on groups and semigroups. In *Advances in Neural Information Processing Systems 21*, pages 473–480, Red Hook, NY, 2009b. Curran Associates Inc.

Arthur Gretton, Karsten M. Borgwardt, Malte Rasch, Bernhard 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, Cambridge, MA, 2007. MIT Press.

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

Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

Steffan Grünewälder, Guy Lever, Luca Baldassarre, Sam Patterson, Arthur Gretton, and Massimiliano Pontil. Conditional mean embeddings as regressors. In *Proceedings of the 29th International Conference on Machine Learning (ICML2012)*, pages 1823–1830, 2012.

Steffen Grünewälder, Arthur Gretton, and John Shawe-Taylor. Smooth operators. In *Proceedings of the 30th International Conference on Machine Learning (ICML2013)*, pages 1184–1192, 2013.

Zaid Harchaoui, Francis Bach, and Eric Moulines. Testing for homogeneity with kernel Fisher discriminant analysis. In *Advances in Neural Information Processing Systems 21*, pages 609–616, Cambridge, MA, 2008. MIT Press.

Simon J. Julier and Jeffrey 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.

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

Steven 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.

Florence Merlevéde, Magda Peligrad, and Sergey Utev. Sharp conditions for the clt of linear processes in a hilbert space. *Journal of Theoretical Probability*, 10:681–693, 1997.

Charles A. Micchelli and Massimiliano Pontil. On learning vector-valued functions. *Neural Computation*, 17(1):177–204, 2005.

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.

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

Peter Müller and Fernando A. Quintana. Nonparametric Bayesian data analysis. *Statistical Science*, 19(1):95–110, 2004.

Shigeki Nakagome, Kenji Fukumizu, and Shuhei Mano. Kernel approximate Bayesian computation for population genetic inferences. *Statistical Applications in Genetics and Molecular Biology*, 2013. Accepted.

Yu Nishiyama, Abdeslam Boularias, Arthur Gretton, and Kenji Fukumizu. Hilbert space embeddings of POMDPs. In *Proceedigns of Conference on Uncertainty in Artificial Intelligence (UAI2012)*, pages 644–653, 2012.

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

Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels*. MIT Press, 2002.

Albert N. Shiryaev. *Probability*. Springer, 2nd edition, 1995.

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

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

Le Song, Jonathan Huang, Alexander Smola, and Kenji 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.

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

Le Song, Sajid M. Siddiqi, Geoffrey Gordon, and Alexander Smola. Hilbert space embeddings of hidden Markov models. In *Proceedings of the 27th International Conference on Machine Learning (ICML2010)*, pages 991–998, 2010b.

Le Song, Arthur Gretton, Danny Bickson, Yucheng Low, and Carlos 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.

John Stachurski. A Hilbert space central limit theorem for geometrically ergodic Markov chains. Working paper, Australian National University, 2012.

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

Charles J. Stone. Optimal global rates of convergence for nonparametric regression. *Annals of Statistics*, 10(4):1040–1053, 1982.

Simon Tavaré, David J. Balding, Robert C. Griffithis, and Peter Donnelly. Inferring coalescence times from dna sequece data. *Genetics*, 145:505–518, 1997.

Sebastian Thrun, John Langford, and Dieter Fox. Monte Carlo hidden Markov models: Learning non-parametric models of partially observable stochastic processes. In *Proceedings of International Conference on Machine Learning (ICML 1999)*, pages 415–424, 1999.

Larry Wasserman. *All of Nonparametric Statistics*. Springer, 2006.

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. Wiley, 1994.

Harold Widom. Asymptotic behavior of the eigenvalues of certain integral equations. *Transactions of the American Mathematical Society*, 109:278–295, 1963.

Harold Widom. Asymptotic behavior of the eigenvalues of certain integral equations II. *Archive for Rational Mechanics and Analysis*, 17:215–229, 1964.

Kun Zhang, Jan Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *27th Conference on Uncertainty in Artificial Intelligence*, pages 804–813, 2011.