Chapter 1
Nonparametric Bayesian Inference with Kernel Mean Embedding

Kenji Fukumizu

Abstract Kernel methods have been successfully used in many machine learning problems with favorable performance in extracting nonlinear structure of high-dimensional data. Recently, nonparametric inference methods with positive definite kernels have been developed, employing the kernel mean expression of distributions. In this approach, the distribution of a variable is represented by the kernel mean, which is the mean element of the random feature vector defined by the kernel function, and relation among variables is expressed by covariance operators. This article gives an introduction to this new approach called *kernel Bayesian inference*, in which the Bayes' rule is realized with the computation of kernel means and covariance expressions to estimate the kernel mean of posterior [11]. This approach provides a novel nonparametric way of Bayesian inference, expressing a distribution with weighted sample, and computing posterior with simple matrix calculation. As an example of problems for which this kernel Bayesian inference is applied effectively, nonparametric state-space model is discussed, in which it is assumed that the state transition and observation model are neither known nor estimable with a simple parametric model. This article gives detailed explanations on intuitions, derivations, and implementation issues of kernel Bayesian inference.

1.1 Introduction

Recent data analysis often involves voluminous high-dimensional data, which may include continuous and complex-structured variables. Classical toolboxes of statistical data analysis may not be sufficient to derive useful information or make reliable predictions in such problems, since the methods often assume low-dimensional simple structure for data such as Gaussian distributions in Euclidean space. It is highly desirable to develop more flexible approaches to tackle those modern data analysis.

K. Fukumizu (✉)
The Institute of Statistical Mathematics, Tokyo, Japan
e-mail: fukumizu@ism.ac.jp

G.W. Peters and T. Matsui (eds.), *Modern Methodology and Applications in Spatial-Temporal Modeling*, JSS Research Series in Statistics,
DOI 10.1007/978-4-431-55339-7_1

Kernel methods have been developed as useful tools for generalizing linear statistical approaches to nonlinear settings. The main idea of kernel methods is to embed original data to a high-dimensional feature space, called a reproducing kernel Hilbert space (RKHS), and apply some linear methods of data analysis for the embedded feature vectors. With this approach, nonlinear features of data can be efficiently handled by virtue of the special way of computing the inner product, which is often called kernel trick. Since the proposal of support vector machines, a number of methods, such as kernel principle component analysis and kernel ridge regression, have been proposed along this discipline and successfully applied in many fields.

The aim of this article is to review recent development of kernel methods for nonparametric statistical inference. In the methods, the mean of the feature vector in the RKHS is considered as a summary for the distribution of feature vectors. We call it *kernel mean*. Although it might be thought that taking the mean loses information of the underlying distribution of data, if a kernel is chosen appropriately, the kernel mean maintains all the information of the distribution. This is possible by the fact that the kernel mean is a function with infinite degree of freedom in an infinite-dimensional RKHS. With this kernel mean approach, probability distributions are expressed by the corresponding kernel means, and linear operations with Gram matrices yield various algorithms for statistical inference, which includes homogeneity test [13–15, 26], independence test [16, 17], conditional independence test [9], and Bayes' theorem [11]. See [29] for a gentle introduction to these researches.

This article focuses on nonparametric kernel methods for Bayesian inference. In Bayesian inference, the sum rule, product rule, and Bayes' rule are important building blocks of inference procedures. The general kernel implementation of these three rules is first presented to realize a nonparametric method for Bayesian inference. As a basis, the conditional kernel mean is introduced and a new theoretical result on the convergence rate of its estimator is shown. A particularly important building block is the kernel implementation of Bayes' rule, called *Kernel Bayes' Rule* [11]. The KBR has special properties in comparison with other methods for Bayesian computation: (a) unlike other popular methods of computing posterior distributions such as Markov Chain Monte Carlo and sequential Monte Carlo, the KBR computes the kernel mean of posterior simply with linear operations of Gram matrices with no need of numerical integration or advanced approximate inference, (b) the ingredients for the Bayesian inference, prior and conditional probability (likelihood), are provided in the form of samples nonparametrically. Thus, this KBR approach is a purely nonparametric Bayesian inference.

A particularly useful application of the kernel Bayes' rule is nonparametric state-space model, for which sequential application of Bayes' rule realizes filtering, prediction, and smoothing. This paper particularly focuses on filtering with nonparametric state-space models, in which it is assumed that the state transition $p(x_{t+1}|x_t)$ and the observation model $p(y_t|x_t)$ are unknown but paired data for the state and observation variables are available for training. The detailed derivation of the kernel filtering algorithm based on the kernel Bayes' rule is presented.

The purpose of this article is to explain the kernel Bayesian inference with details together with some new results. In particular, as building blocks, kernel sum rule, kernel product rule, and kernel Bayes' rule are explained in detail including intuitions and derivations. A new theoretical result on the convergence rate of the conditional kernel mean estimator is presented using the decay rate of eigenvalues of the covariance operator. Additionally, as a typical application, details on the KBR filter are discussed including efficient low-rank approximation.

1.2 Representing Distributions with Kernel Mean Embedding

1.2.1 Preliminary: General Kernel Methods

We first give a brief review of positive definite kernels and kernel methods. A standard reference for readers unfamiliar with kernel methods is [28].

Given a set Ω, a ($\mathbb{R}$-valued) *positive definite kernel* k on Ω is a symmetric kernel $k : \Omega \times \Omega \to \mathbb{R}$ that satisfies positive semidefiniteness, i.e., $\sum_{i,j=1}^{n} c_i c_j k(x_i, x_j) \geq 0$ 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 [1] that a positive definite kernel on Ω uniquely defines a Hilbert space $\mathscr{H}$ consisting of functions on Ω such that the following three conditions hold: (i) $k(\cdot, x) \in \mathscr{H}$ for any $x \in \Omega$, (ii) $\mathrm{Span}\{k(\cdot, x) \mid x \in \Omega\}$ is dense in $\mathscr{H}$, and (iii) $\langle f, k(\cdot, x) \rangle = f(x)$ for any $x \in \Omega$ and $f \in \mathscr{H}$ (the reproducing property), where $\langle \cdot, \cdot \rangle$ is the inner product of $\mathscr{H}$. The Hilbert space $\mathscr{H}$ is called the *reproducing kernel Hilbert space* (RKHS) associated with k.

In kernel methods, Ω is a space where data exist, and a positive definite kernel k is prepared for Ω. The corresponding RKHS $\mathscr{H}$ is used as a feature space, and a nonlinear mapping (feature mapping) from data space Ω to the feature space $\mathscr{H}$ is defined by

$$\Phi : \Omega \to \mathscr{H}, \quad x \to k(\cdot, x),$$

where $k(\cdot, x) \in \mathscr{H}$ should be interpreted as a function of the first argument with x fixed. A data is thus mapped to a function, and this functional representation of data extracts various nonlinear features of data. From computational side, the reproducing property provides an efficient way of extracting nonlinear features in data analysis, without expanding the original variables with basis functions, which causes an intractably large number of components for high-dimensional original variables.

The traditional way of kernel methods considers the mapping of data $X_1, \ldots, X_n$ in the original space Ω to feature vectors $\Phi(X_1), \ldots, \Phi(X_n)$ in the RKHS, and apply some linear method of data analysis, such as principal component analysis, to those

feature vectors. By the reproducing property, the inner product of two feature vectors is reduced to evaluation of the kernel, that is

$$\langle \Phi(x), \Phi(y) \rangle = k(x, y).$$

This fact is sometimes referred to as *kernel trick*, providing one of the essential elements in kernel methods. More generally, given two linear combinations of the feature vectors, say $f = \sum_{i=1}^{n} \alpha_i \Phi(X_i)$ and $g = \sum_{j=1}^{n} \beta_j \Phi(X_j)$, then the inner product between f and g is given by

$$\langle f, g \rangle = \alpha^T G_X \beta,$$

where $G_{X,ij} = k(X_i, X_j)$ is the Gram matrix. Given that computation of an analysis method for Euclidean data relies on the inner product among data points, the method can be extended to a nonlinear version with the above inner products among feature vectors. The computational cost thus does not depend on the dimensionality of data, once the Gram matrices are computed. This is computational advantage of kernel methods for handling high-dimensional data.

Computation with Gram matrices is obviously expensive if the sample size is large. It is known, however, that low-rank approximation of a Gram matrix reduces the size of the involved matrix drastically, while maintaining the approximation accuracy reasonably. As typical methods for low-rank approximation, the incomplete Cholesky decomposition [6] and Nyström approximation [38] approximate a Gram matrix G of size n to the form $G \approx RR^T$ with $n \times r$ matrix R in computational time proportional to n. Once the low-rank approximation is done, inversion $(G + \lambda I_n)^{-1}$ can be approximated by $I_n - R(R^T R + \lambda I_r)^{-1} R^T$ (Woodbury's formula), in which the inverse is taken for a matrix of size r. Here I_m denotes the $m \times m$ identity matrix. The merit of this approximation will be discussed in Sect. 1.4.2.

1.2.2 Kernel Mean Representation of Probability Distributions

In the recent development of kernel methods for nonparametric inference, the mean of the random feature vector $\Phi(X) = k(\cdot, X)$ is considered to represent a probability distribution on the random variable X.

More formally, let $(\mathscr{X}, \mathscr{B}_{\mathscr{X}})$ be a measurable space, X be a random variable taking values in $\mathscr{X}$ with probability distribution P on $\mathscr{X}$, and k be a measurable positive definite kernel on $\mathscr{X}$ such that $E[\sqrt{k(X, X)}] < \infty$. The associated RKHS is denoted by $\mathscr{H}$. The *kernel mean* m_X (also written by m_P) of X in $\mathscr{H}$ is defined by the mean $E[k(\cdot, X)]$ of the $\mathscr{H}$-valued random variable $\Phi(X)$.[1] Here, the mean

[1] As the kernel mean depends on k, it should be written by m_X^k rigorously. We will, however, generally write m_X for simplicity, where there is no ambiguity.

is interpreted as Bochner integral, which exists by the assumption $E[\|k(\cdot, X)\|] = E[\sqrt{k(X, X)}] < \infty$.

By the reproducing property, the kernel mean satisfies the relation

$$\langle f, m_X \rangle = E[f(X)] \tag{1.1}$$

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

$$m_X(u) = E[k(u, X)] = \int k(u, \tilde{x}) dP(\tilde{x}), \tag{1.2}$$

which is an explicit integral form of the kernel mean.

To represent probabilities, an important notion is the characteristic property. A positive definite kernel k is called *bounded* if $\sup_{x \in \mathscr{X}} k(x, x) < \infty$. A bounded measurable positive definite kernel k on a measurable space $(\Omega, \mathscr{B})$ is called *characteristic* if the mapping from a probability Q on $(\Omega, \mathscr{B})$ to the kernel mean $m_Q \in \mathscr{H}$ is injective [7, 8, 32]. This is equivalent to assuming that $E_{X \sim P}[k(\cdot, X)] = E_{X' \sim Q}[k(\cdot, X')]$ implies $P = Q$ by definition: probabilities are uniquely determined by their kernel means on the associated RKHS. 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 characteristic kernel provides a RKHS that contains a rich class of functions so that the moments $E[f(X)]$ for all $f \in \mathscr{H}$ can identify the underlying distribution. Various conditions for a kernel to be characteristic can be found in [12, 31, 32].

By the unique representation property of characteristic kernels, statistical inference problems on probability distribution can be converted to the inference problems on the kernel means, which are easier to handle by the special properties of RKHS. This is the principle of the nonparametric inference with kernel means. Various inference methods have been proposed under this discipline. If we consider a two-sample problem, which aims at determining whether or not given two samples come from the same distribution, it can be cast as the problem of comparing the corresponding two kernel means in a RKHS [13]. The problem of independence test can be solved by comparing the kernel means of joint distributions and the product of the marginals [15].

When the relation of two random variables is discussed, covariance is useful in addition to means. In the kernel mean framework, covariance of the two feature vectors on the RKHS's is considered, and it is called covariance operator. More precisely, let $(\mathscr{X}, \mathscr{B}_\mathscr{X})$ and $(\mathscr{Y}, \mathscr{B}_\mathscr{Y})$ be measurable spaces, (X, Y) be a random variable on $\mathscr{X} \times \mathscr{Y}$ with distribution P, and $k_\mathscr{X}$ and $k_\mathscr{Y}$ be measurable positive definite kernels with respective RKHS $\mathscr{H}_\mathscr{X}$ and $\mathscr{H}_\mathscr{Y}$ such that $E[k_\mathscr{X}(X, X)] < \infty$ and $E[k_\mathscr{Y}(Y, Y)] < \infty$.[2] The (uncentered) *covariance operator* $C_{YX} : \mathscr{H}_\mathscr{X} \to \mathscr{H}_\mathscr{Y}$

[2]These conditions guarantee existence of the covariance operator. Note also $E[k(X, X)] < \infty$ is stronger than the condition for kernel mean, $E[\sqrt{k(X, X)}] < \infty$; this is obvious from Cauchy–Schwarz inequality.

is defined by

$$C_{YX} = E[k_{\mathscr{Y}}(\cdot, Y)\langle k_{\mathscr{X}}(X, \cdot), *\rangle],$$

or equivalently, for $f \in \mathscr{H}_{\mathscr{X}}$,

$$(C_{YX} f)(y) = E[k_{\mathscr{Y}}(y, Y) f(X)] = \int k_{\mathscr{Y}}(y, \tilde{y}) f(\tilde{x}) dP(\tilde{x}, \tilde{y}). \tag{1.3}$$

From the reproducing property, the covariance operator is a linear operator that satisfies

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

for all $f \in \mathscr{H}_{\mathscr{X}}$, $g \in \mathscr{H}_{\mathscr{Y}}$. We also define C_{XX} by the operator on $\mathscr{H}_{\mathscr{X}}$ that satisfies $\langle f_2, C_{XX} f_1\rangle = E[f_2(X) f_1(X)]$ for any $f_1, f_2 \in \mathscr{H}_{\mathscr{X}}$.

The covariance operator is a natural extension of an ordinary covariance matrix: given two random vectors Z and W on Euclidean spaces, the covariance matrix can be regarded as a linear mapping $a \mapsto E[WZ^T]a$. Replacing Z and W with $k_{\mathscr{X}}(\cdot, X)$ and $k_{\mathscr{Y}}(\cdot, Y)$, respectively, yields the covariance operator $E[k_{\mathscr{Y}}(\cdot, Y)\langle k_{\mathscr{X}}(\cdot, X), *\rangle]$. Readers who are unfamiliar with the notion of operators can simply think of linear mappings on infinite-dimensional vector spaces to grasp the general ideas in this article.

Note also that by identifying the dual element $\langle k_{\mathscr{X}}(\cdot, X), *\rangle$ with $k_{\mathscr{X}}(\cdot, X)$, the covariance operator C_{YX} can be identified with the kernel mean $m_{YX} = E[k_{\mathscr{Y}}(\cdot, Y)$ $k_{\mathscr{X}}(\cdot, X)]$ in the direct product $\mathscr{H}_{\mathscr{Y}} \otimes \mathscr{H}_{\mathscr{X}}$, which is given by the product kernel $k_{\mathscr{Y}} k_{\mathscr{X}}$ on $\mathscr{Y} \times \mathscr{X}$ [1]. This fact will be used in deriving kernel Bayes' rule.

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 the empirical mean and covariance as

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

where $\widehat{C}_{YX}^{(n)}$ is written in the tensor form. These estimators are known to be $\sqrt{n}$-consistent in appropriate norms, and $\sqrt{n}(\widehat{m}_X - m_X)$ converges to a Gaussian process on $\mathscr{H}_{\mathscr{X}}$ [3].

1.3 Bayesian Inference with Kernel Means

There are three basic operations used in general Bayesian inference: sum rule, product rule, and Bayes' rule, which are summarized in Table 1.1. Correspondingly, in the framework of Bayesian inference with kernel means, these operations are realized

Table 1.1 Operations for Bayesian inference.

Density form		Kernel version $\widehat{m}_\Pi = \sum_j \gamma_j k_{\mathscr{X}}(\cdot, U_j)$, $(X_i, Y_i) \sim P$
Sum rule	$q_{\mathscr{Y}}(y) = \int p(y\|x)\pi(x)dx$	$\widehat{m}_{Q_{\mathscr{Y}}} = \sum_i w_i k_{\mathscr{Y}}(\cdot, Y_i)$, $w = (G_X + n\varepsilon_n I_n)^{-1} G_{XU}\gamma$
Product rule	$q(x, y) = p(y\|x)\pi(x)$	$\widehat{m}_Q = \sum_i w_i k_{\mathscr{X}}(\cdot, X_i) \otimes k_{\mathscr{Y}}(\cdot, Y_i)$, $w = (G_X + n\varepsilon_n I_n)^{-1} G_{XU}\gamma$
Bayes' rule	$q(x\|y_{\text{obs}}) = \frac{p(y_{\text{obs}}\|x)\pi(x)}{\int p(y_{\text{obs}}\|x)\pi(x)dx}$	$\widehat{m}_{Q_{x\|y_{\text{obs}}}} = \sum_i w_i k_{\mathscr{X}}(\cdot, X_i)$, $\Lambda = \text{Diag}\{(G_X + n\varepsilon_n I_n)^{-1} G_{XU}\gamma\}$, $w = \Lambda G_Y((\Lambda G_Y)^2 + \delta_n I_n)^{-1}\Lambda \mathbf{k}_Y(y_{\text{obs}})$

In the kernel version, $G_X = (k(X_i, X_j))$, $G_Y = (k(Y_i, Y_j))$, and $G_{XU} = (k(X_i, U_j))$

in terms of kernel means. This section first provides an intuitive explanation for the population version of the kernel realization, which may not be rigorous in handling operator inversion, and then shows rigorous empirical expressions, which can be proved to be consistent.

In the framework, each distribution is represented by the corresponding kernel mean or its empirical estimate. An empirical estimator of the kernel mean of a probability P is, in general, given by a weighted sum of feature vectors

$$\widehat{m}_P = \sum_{i=1}^{n} w_i k(\cdot, X_i),$$

where $(X_i)_{i=1}^n$ is some sample, which may not be generated by P.

1.3.1 Conditional Kernel Mean

For Bayesian inference with kernels, a basis is how to express or estimate the conditional kernel mean. It is not straightforward, however, to have an empirical expression of kernel mean of the conditional probability of Y given X. If we had many samples of Y for each value of x, we could just use the samples or their feature vectors to represent the kernel mean of Y given x. It is unlikely, however, that we have such *conditional samples*, if the variable X is continuous and random. We then need an alternative way of expressing the kernel mean of a conditional probability. We assume that there is a probability P with density $p(x, y)$ that gives a conditional density $p(y|x)$, and we have data (X_i, Y_i) generated by P.

The theoretical basis of the conditional kernel mean is the following theorem.

Theorem 1.3.1 *([7]) If, for $g \in \mathscr{H}_{\mathscr{Y}}$, $E[g(Y)|X = x]$ is included in $\mathscr{H}_{\mathscr{X}}$ as a function of x, then*

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

The proof is easy from the fact $\langle C_{XX} E[g(Y)|X = \cdot], f\rangle = E[g(Y) f(X)] = \langle C_{XY} g, f\rangle$ for any $f \in \mathscr{H}_{\mathscr{X}}$. From this theorem, if C_{XX} is invertible, we have

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

Taking the inner product with $k_{\mathscr{X}}(\cdot, x)$ derives

$$\langle E[g(Y)|X = \cdot], k_{\mathscr{X}}(\cdot, x)\rangle_{\mathscr{H}_{\mathscr{X}}} = \langle C_{XX}^{-1} C_{XY} g, k_{\mathscr{X}}(\cdot, x)\rangle_{\mathscr{H}_{\mathscr{X}}},$$

which implies

$$\langle g, E[k_{\mathscr{Y}}(\cdot, Y)|X = x]\rangle_{\mathscr{H}_{\mathscr{Y}}} = \langle g, C_{YX} C_{XX}^{-1} k_{\mathscr{X}}(\cdot, x)\rangle_{\mathscr{H}_{\mathscr{Y}}}.$$

If $E[g(Y)|X = \cdot] \in \mathscr{H}_{\mathscr{X}}$ holds for any $g \in \mathscr{H}_{\mathscr{Y}}$, it follows that

$$E[k_{\mathscr{Y}}(\cdot, Y)|X = x] = C_{YX} C_{XX}^{-1} k_{\mathscr{X}}(\cdot, x). \tag{1.4}$$

Since the left-hand side of Eq. (1.4) is exactly the kernel mean of conditional probability of Y given $X = x$, this equation provides an expression of its kernel mean in terms of the covariance operator of the joint distribution (X, Y). Note, however, that the above reasoning involves a strong assumption: C_{XX} is invertible. In fact, this does not hold if the dimensionality of $\mathscr{H}_{\mathscr{X}}$ is infinite and C_{XX} has arbitrarily small or zero eigenvalues. This occurs in typical cases with a bounded kernel of infinite-dimensional RKHS, since the trace of the infinite-dimensional linear map C_{XX} is finite [10].

Nonetheless, from the expression Eq. (1.4), we can introduce an empirical estimator of the kernel mean of $p(y|x)$, namely, given i.i.d. sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ following the joint distribution P, an estimator is defined by

$$\widehat{m}_{Y|X=x} := \widehat{C}_{YX}^{(n)} \big(\widehat{C}_{XX}^{(n)} + \varepsilon_n I\big)^{-1} k_{\mathscr{X}}(\cdot, x), \tag{1.5}$$

where I is the identity operator and ε_n is a regularization constant so that the operator can be inverted. This estimator is rigorously defined and proved to be consistent to $E[k_{\mathscr{Y}}(\cdot, Y)|X = x]$ under the sufficient condition in the following Theorem 1.3.2.

To describe the following convergence result, decay rate of eigenvalues is introduced. The eigenvalues of a positive compact operator C are said to *decay at rate b* if there is a constant $\beta > 0$ such that $\lambda_\ell \le \beta \ell^{-b}$ for all ℓ, where (λ_ℓ) is the positive eigenvalues of C in descending order. (See [4]). The following theorem shows the convergence rate of the conditional kernel mean estimator.

Theorem 1.3.2 *Assume that* $E[k(X, \tilde{X})|Y = \cdot, \tilde{Y} = *] \in \mathscr{R}(C_{YY} \otimes C_{YY})$, *where* $(\tilde{X}, \tilde{Y})$ *is an independent copy of* (X, Y), *and that the eigenvalues of* C_{YY} *decay at rate* b $(1 < b < +\infty)$. *Then, with* $\varepsilon_n = n^{-b/(4b+1)}$,

$$\left\| \widehat{C}_{XY}^{(n)} (\widehat{C}_{YY}^{(n)} + \varepsilon_n I)^{-1} k_{\mathscr{Y}}(\cdot, y_0) - E[k_{\mathscr{X}}(\cdot, X)|Y = y_0] \right\|_{\mathscr{H}_{\mathscr{X}}} = O_p(n^{-b/(4b+1)})$$

as $n \to \infty$.

See the appendix for the proof. The decay rates of eigenvalues of a covariance operator are known in some typical cases; see [36, 37]. Note that the assumption of the decay rate of the covariance operator is related to the entropy number, and standard in discussing the behavior of kernel regression [33]. The assumption $E[k(X, \tilde{X})|Y = \cdot, \tilde{Y} = *] \in \mathscr{R}(C_{YY} \otimes C_{YY})$ requires the smoothness of the conditional expectation when the kernel is smooth such as Gaussian kernel; the range space consists of smoother functions by the smoothing effect of the integral in Eq. (1.3). To the best of our knowledge, the convergence rate of the conditional kernel mean in the above form has not been presented in existing literatures.[3]

1.3.2 Kernel Sum Rule and Kernel Product Rule

For the sum and product rules, this subsection gives intuitive explanation rather than rigorous convergence results. See [11] for the results.

For the kernel mean implementation of the sum rule, let Π be a probability on $\mathscr{X}$ with density $\pi(x)$. As in the previous subsection, we assume that there is a joint distribution P on $\mathscr{X} \times \mathscr{Y}$ with density $p(x, y)$ of which the conditional p.d.f. is equal to the given $p(y|x)$. Suppose that the sum rule gives $Q_{\mathscr{Y}}$ with density $q_{\mathscr{Y}}(y)$, i.e.,

$$q_{\mathscr{Y}}(y) = \int p(y|x)\pi(x)dx.$$

The kernel mean of $Q_{\mathscr{Y}}$ is then given by

$$m_{Q_y} = \int \int k_{\mathscr{Y}}(\cdot, y) p(y|x)\pi(x)dxdy.$$

[3] Some previous literatures derived a convergence rate at unrealistic assumptions. For example, Theorem 6 in [30] assumes $k(\cdot, y_0) \in \mathscr{R}(C_{YY})$ to achieve the rate $n^{-1/4}$, but in typical cases there is no function $f \in \mathscr{H}_{\mathscr{Y}}$ that satisfies $\int k(y, z) f(z) dP_Y(z) = k(y, y_0)$. Theorem 1.3.2 shows that if the eigenvalues decay sufficiently fast the rate approaches $n^{-1/4}$. As a relevant result, Theorem 11 in [11] shows a convergence rate of the kernel sum rule. While the conditional kernel mean is a special case of kernel sum rule with prior given by Dirac's delta function at x, the faster rate ($n^{-1/3}$ at best) is not achievable by Theorem 1.3.2, since the former assumes that π / p_X is a function in the RKHS and smooth enough.

From Eq. (1.4), we already know the (non-rigorous) expression

$$\int k_{\mathscr{Y}}(\cdot, y)p(y|x)dy = C_{YX}C_{XX}^{-1}k_{\mathscr{X}}(\cdot, x).$$

Plugging this into the previous equation, we have (the population version of) *kernel sum rule*:

$$m_{Q_{\mathscr{Y}}} = \int C_{YX}C_{XX}^{-1}k_{\mathscr{X}}(\cdot, x)\pi(x)dx = C_{YX}C_{XX}^{-1}m_{\Pi}. \tag{1.6}$$

There is another way to derive Eq. (1.6) in terms of density functions. Suppose that the density ratio π/p_X is included in $\mathscr{H}_{\mathscr{X}}$. From Eqs. (1.2) and (1.3), we see

$$m_{\Pi} = \int k_{\mathscr{X}}(\cdot, x)\pi(x)dx = \int k_{\mathscr{X}}(\cdot, x)\frac{\pi(x)}{p_X(x)}dP_X(x) = C_{XX}\left(\frac{\pi}{p_X}\right),$$

from which we obtain

$$C_{XX}^{-1}m_{\Pi} = \frac{\pi}{p_X}.$$

It follows from Eq. (1.3) that

$$\begin{aligned} C_{YX}C_{XX}^{-1}m_{\Pi} = C_{YX}\left(\frac{\pi}{p_X}\right) &= \int k_{\mathscr{Y}}(\cdot, y)\frac{\pi(x)}{p_X(x)}dP(x, y) \\ &= \int\int k_{\mathscr{Y}}(\cdot, y)p(y|x)\pi(x)dxdy = m_{Q_{\mathscr{Y}}}, \end{aligned}$$

which agrees with Eq. (1.6).

Given a consistent estimator $\widehat{m}_{\Pi}$ of m_{Π} and i.i.d. sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ from P, the empirical version of the kernel sum rule is defined based on Eq. (1.6);

$$\widehat{m}_{Q_{\mathscr{Y}}} = \widehat{C}_{YX}^{(n)}\left(\widehat{C}_{XX}^{(n)} + \varepsilon_n I\right)^{-1}\widehat{m}_{\Pi}. \tag{1.7}$$

In a Gram matrix expression, given

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

we have

$$\widehat{m}_{Q_{\mathscr{Y}}} = \sum_{i=1}^{n} w_i k_{\mathscr{Y}}(\cdot, Y_i), \quad w = (G_X + n\varepsilon_n I_n)^{-1}G_{XU}\gamma,$$

where $G_X = (k_{\mathscr{X}}(X_i, X_j))_{ij}$ and $G_{XU} = (k_{\mathscr{X}}(X_i, U_j))_{ij}$. The convergence of this estimator to the true $m_{Q_{\mathscr{Y}}}$ and its convergence rate are shown in Theorems 8 and

11 of [11]. For the convergence, it is assumed that the sample size ℓ for the prior increases as $n \to \infty$.

The kernel version of product rule can be derived as a special case of the kernel sum rule. Consider the conditional density $\tilde{p}(y, \tilde{x}|x) = p(y|x)\delta_x(\tilde{x})$ on the product space $\mathscr{Y} \times \mathscr{X}$, where δ_x is Dirac's delta function with mass concentrated at x. Let Q be a probability distribution on $\mathscr{Y} \times \mathscr{X}$ with density $p(y|x)\pi(x)$, i.e., the density given by the product rule. The population version of kernel sum rule applied to $\tilde{p}(y, \tilde{x}|x)$ and $\pi(x)$ with the product kernel then yields

$$m_Q = \int\int\int k_{\mathscr{Y}}(\cdot, y) \otimes k_{\mathscr{X}}(\cdot, \tilde{x})\tilde{p}(y, \tilde{x}|x)\pi(x)d\tilde{x}dydx = C_{(YX)X}C_{XX}^{-1}m_{\Pi},$$

where $C_{(YX)X} : \mathscr{H}_{\mathscr{X}} \to \mathscr{H}_{\mathscr{Y}} \otimes \mathscr{H}_{\mathscr{X}}$ is the covariance operator for the random variable $(X, (X, Y))$. Based on the (non-rigorous) population expression, we define the empirical kernel product rule by

$$\widehat{m}_Q := \widehat{C}_{(YX)X}^{(n)}\big(\widehat{C}_{XX}^{(n)} + \varepsilon_n I\big)^{-1}\widehat{m}_{\Pi}, \tag{1.8}$$

or in Gram matrix expression

$$\widehat{m}_Q = \sum_{i=1}^{n} w_i k_{\mathscr{Y}}(\cdot, Y_i) \otimes k_{\mathscr{X}}(\cdot, X_i), \quad w = (G_X + n\varepsilon_n I_n)^{-1}G_{XU}\gamma, \tag{1.9}$$

Note that the weight vectors of Eqs. (1.7) and (1.9) are exactly the same, while the feature vectors or the spaces of interest are different.

1.3.3 Kernel Bayes' Rule

As demonstrated in this subsection, by combining the kernel product rule and conditional kernel mean, we can easily derive the kernel Bayes' rule. As in the previous subsection, let Π be the prior and P be a probability on $\mathscr{X} \times \mathscr{Y}$ with conditional density $p(y|x)$. The distribution of the variable (X, Y) is P. The posterior distribution given y_{obs} is denoted by $Q_{x|y_{\text{obs}}}$.

From the expression of Bayes' rule

$$q(x|y_{\text{obs}}) = \frac{p(y|x)\pi(x)}{\int p(y|x)\pi(x)dx},$$

we see that the posterior is simply the conditional distribution of x given y_{obs} with the joint distribution Q given by the product rule. Once we have covariance operators for Q, Theorem 1.3.1 tells how to derive the conditional kernel mean, that is the kernel mean of posterior. The remaining task is thus to construct the covariance operators for Q.

Let (Z, W) denote a random variable taking values on $\mathscr{X} \times \mathscr{Y}$ with distribution Q. Then, from Eq. (1.8),

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

where the second relation can be obtained in a similar way to the first one. Recall that the covariance operators C_{WZ} and C_{WW} are identified with the kernel means m_{WZ} and m_{WW}, respectively, on the product spaces, as discussed in Sect. 1.2.2. We can therefore obtain the estimator of $\widehat{C}^{(n)}_{WZ}$ and $\widehat{C}^{(n)}_{WW}$ from the above empirical version of kernel product rule. Namely, when the kernel product rule provides the empirical expressions

$$\widehat{m}_{(WZ)} = \sum_{i=1}^{n} \widehat{\mu}_i k(\cdot, Y_i) \otimes k(\cdot, X_i) \quad \text{and} \quad \widehat{m}_{(WW)} = \sum_{i=1}^{n} \widehat{\mu}_i k(\cdot, X_i) \otimes k(\cdot, X_i)$$

with

$$\widehat{\mu} = (G_X + n\varepsilon_n I_n)^{-1} G_{XU}\gamma, \tag{1.10}$$

the empirical estimators of covariance operators for Q are given by

$$\widehat{C}^{(n)}_{WZ} = \sum_{i=1}^{n} \widehat{\mu}_i k_{\mathscr{Y}}(\cdot, Y_i)\langle k_{\mathscr{X}}(\cdot, X_i), * \rangle, \quad \widehat{C}^{(n)}_{WW} = \sum_{i=1}^{n} \widehat{\mu}_i k_{\mathscr{Y}}(\cdot, Y_i)\langle k_{\mathscr{X}}(\cdot, Y_i), * \rangle.$$

Note that the coefficients to the feature vectors are the same for $\widehat{C}^{(n)}_{WZ}$ and $\widehat{C}^{(n)}_{WW}$.

In applying Eq. (1.5), there is another technical point. The estimated covariance operator $\widehat{C}^{(n)}_{WW}$ may not be positive definite, since the coefficients $\widehat{\mu}_i$ are not necessarily positive as the solution of the matrix operation Eq. (1.10). We use a more involved regularization to make the operator inversion possible, and introduce

$$\widehat{m}_{Q_x|y_{\mathrm{obs}}} := \widehat{C}_{ZW}\big(\widehat{C}^2_{WW} + \delta_n I\big)^{-1}\widehat{C}_{WW} k_{\mathscr{Y}}(\cdot, y_{\mathrm{obs}}). \tag{1.11}$$

This gives an estimator of the posterior kernel mean, and is called *Kernel Bayes' Rule* (KBR).

Theorem 1.3.3 (Kernel Bayes' Rule [11]) *For any $y_{\mathrm{obs}} \in \mathscr{Y}$, the estimator $\widehat{m}_{Q_x|y_{\mathrm{obs}}}$ of the posterior kernel mean is given by*

$$\widehat{m}_{Q_x|y_{\mathrm{obs}}} = \sum_{i=1}^{n} w_i k(\cdot, X_i), \qquad w = \Lambda G_Y((\Lambda G_Y)^2 + \delta_n I_n)^{-1}\Lambda \mathbf{k}_Y(y_{\mathrm{obs}}), \tag{1.12}$$

where $\Lambda = \mathrm{diag}(\widehat{\mu})$ is a diagonal matrix with elements $\widehat{\mu}_i$ in Eq. (1.10), and $\mathbf{k}_Y(y_{\mathrm{obs}}) = (k_{\mathscr{Y}}(y_{\mathrm{obs}}, Y_1), \ldots, k_{\mathscr{Y}}(y_{\mathrm{obs}}, Y_n))^T \in \mathbb{R}^n$.

It is known that under some conditions the estimator $\widehat{m}_{Q_x|y_{\mathrm{obs}}}$ converges to the true kernel mean of the posterior in probability, and an upper bound of its convergence rate is also known (Theorem 4, [11]).

The expression Eq. (1.12) takes the form of a weighted sum of feature vectors $k(\cdot, X_i)$, and is regarded as the kernel mean of the signed measure $\sum_{i=1}^{n} w_i \delta_{X_i}$. The KBR thus provides a weighted sample expression $(w_i, X_i)_{i=1}^{n}$ of the posterior. Note again that the weights may include negative values, which is different from ordinary weighted sample expression used popularly in importance sampling and particle filters. Figure 1.1 illustrates the procedure of KBR.

The above estimator provides the kernel mean of the posterior, and not the posterior itself. We need to develop methods for decoding necessary information of posterior from the kernel mean expression. Two methods are discussed below: estimation of expectation with respect to posterior and point estimation with the posterior.

If our aim is to estimate the expectation of a function $f \in \mathscr{H}_{\mathscr{X}}$ with respect to the posterior, the reproducing property of Eq. (1.1) gives an estimator

$$\langle f, \widehat{m}_{Q_x|y_{\mathrm{obs}}} \rangle = \sum_{i=1}^{n} w_i f(X_i). \tag{1.13}$$

In fact, it is known that, under some conditions, the estimator Eq. (1.13) for any $f \in \mathscr{H}_{\mathscr{X}}$ converges to the expectation of f w.r.t. the true posterior, and its convergence rate is also known (Theorems 6 and 7, [11]). A recent work has shown that the consistency of $\sum_{i=1}^{n} w_i f(X_i)$ to $\int f(x) q_{x|y_{\mathrm{obs}}}(x) dx$ is true for a wider class of functions than $\mathscr{H}_{\mathscr{Y}}$ [19]. This fact confirms similarity of (w_i, X_i) in KBR to the standard weighted sample expression.

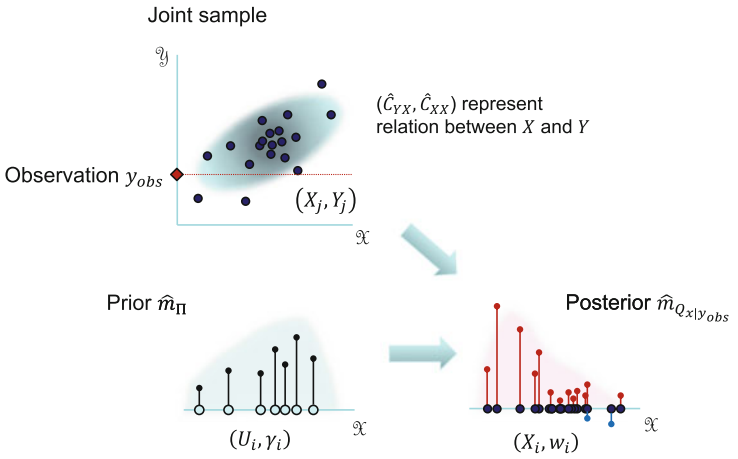

Fig. 1.1 Kernel Bayes' rule

If our aim is to obtain a point estimate based on the posterior, such as MAP, we can use a point $x \in \mathcal{X}$ such that the feature vector is the closest to the kernel mean of posterior [11, 30], i.e.,

$$\widehat{x} = \arg\min_{x \in \mathcal{X}} \left\| k_{\mathcal{X}}(\cdot, x) - \widehat{m}_{Q_x|y_{\mathrm{obs}}} \right\|^2.$$

In the case of Gaussian kernel $k(x, x') = \exp(-\frac{1}{2\sigma^2}\|x - x'\|^2)$, from $\|k(\cdot, x)\| = 1$, the above minimization is equivalent to

$$\widehat{x} = \arg\max_{x \in \mathcal{X}} \sum_{i=1}^{n} w_i \exp\Big(-\frac{1}{2\sigma^2}\|x - X_i\|^2\Big),$$

which is similar to the MAP estimation, though $\sum_i w_i k(x, X_i)$ may not be a density function.

The above optimization problem can be solved in the same manner as the pre-image problem [24]. Taking the derivative of the squared norm in the right-hand side, we obtain the consistence equation

$$\widehat{x} = \frac{\sum_{i=1}^{n} w_i \exp(-\frac{1}{2\sigma^2}\|\widehat{x} - X_i\|^2)}{\sum_{i=1}^{n} \exp(-\frac{1}{2\sigma^2}\|\widehat{x} - X_i\|^2)},$$

which yields an iterative method for solving the point estimate:

$$\widehat{x}^{(t+1)} = \frac{\sum_{i=1}^{n} w_i \exp(-\frac{1}{2\sigma^2}\|\widehat{x}^{(t)} - X_i\|^2)}{\sum_{i=1}^{n} \exp(-\frac{1}{2\sigma^2}\|\widehat{x}^{(t)} - X_i\|^2)}.$$

Note that the objective function of pre-image problem is not necessarily convex and there may be local optima. The initial point of the above iteration must be chosen carefully. One possible method for initialization is to use the posterior mean. In the filtering problem discussed in Sect. 1.4, the estimate in the previous time step can serve as an initial point. Other pre-image methods [21] can be also applied to the above point estimation problem.

1.4 Kernel Bayesian Inference for State-Space Models

We discuss applications of KBR to the sequential Bayesian inference with state-space models. A time-invariant state-space model is defined by

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

where Y_t is an observation and X_t is a hidden state variable. The index t indicates time. The conditional probability $q(x_{t+1}|x_t)$ and $p(y_t|x_t)$ are called the state transition and observation model, respectively. With this model of time series, given $Y_1, \ldots, Y_t$, we wish to estimate the posteriors $p(X_s|Y_1, \ldots, Y_t)$. Filtering, prediction, and smoothing refer to as the case $s = t$, $s > t$, and $s < t$, respectively. This article discusses only the filtering problem for simplicity, while the other cases can be solved similarly.

1.4.1 KBR Filter

It is well known that, under the assumption of state-space models, application of Bayes' rule derives a sequential algorithm of filtering, which consists of two steps: prediction and correction steps.

Prediction step: Given an estimate of $p(x_t|y_1, \ldots, y_t)$, the conditional probability $p(x_{t+1}|y_1, \ldots, y_t)$ is estimated. This is done by the sum rule,

$$p(x_{t+1}|y_1, \ldots, y_t) = \int q(x_{t+1}|x_t)p(x_t|y_1, \ldots, y_t)dx_t. \tag{1.14}$$

Correction step: Given a new observation y_{t+1}, Bayes' rule derives the estimate of $p(x_{t+1}|y_1, \ldots, y_{t+1})$ with the prior $p(x_{t+1}|y_1, \ldots, y_t)$ and likelihood $p(y_t|x_t)$,

$$p(x_{t+1}|y_1, \ldots, y_{t+1}) = \frac{p(y_{t+1}|x_{t+1})p(x_{t+1}|y_1, \ldots, y_t)}{\int p(y_{t+1}|x_{t+1})p(x_{t+1}|y_1, \ldots, y_t)dx_{t+1}}. \tag{1.15}$$

If the state transition and observation model are given by linear mapping plus Gaussian noise, Kalman filter is the well-known filtering procedure. If they are written by known nonlinear dynamics, nonlinear extensions of Kalman filter, such as the extended Kalman filter (EKF) and unscented Kalman filter (UKF, [35]), are popular choices. In more general setting, given the state transition and observation model are known upto constant, the particle filter or sequential Monte Carlo [5] gives a weighted sample expression of the sequential update. These methods, however, require the precise knowledge on the functional form of the state transition and observation model, and not applicable unless they are known.

The KBR can be effectively applied to inference with the nonparametric setting of state-space models. In the nonparametric state-space models, it is not assumed that the conditional probabilities $p(Y_t|X_t)$ and $q(X_{t+1}|X_t)$ are known explicitly, nor estimated them with simple parametric models. Rather, it is assumed that training data $(X_1, Y_1), \ldots, (X_{T+1}, Y_{T+1})$ are given for both the observable and state variables in the *training phase*. In the *testing phase*, the state x_t is inferred based on a different sequence of observations $\tilde{y}_1, \ldots, \tilde{y}_t$ without knowing the corresponding state variables.

In the training phase, given the training sample, the observation model $p(y_t|x_t)$ and the state transition $q(x_{t+1}|x_t)$ are represented using the empirical covariances operators[4]: $\widehat{C}_{YX} = \frac{1}{T}\sum_{i=1}^{T} k_{\mathscr{Y}}(\cdot, Y_i) \otimes k_{\mathscr{X}}(\cdot, X_i)$, $\widehat{C}_{YY} = \frac{1}{T}\sum_{i=1}^{T} k_{\mathscr{Y}}(\cdot, Y_i) \otimes k_{\mathscr{Y}}(\cdot, Y_i)$, $\widehat{C}_{XX} = \frac{1}{T}\sum_{i=1}^{T} k_{\mathscr{X}}(\cdot, X_i) \otimes k_{\mathscr{X}}(\cdot, X_i)$, and $\widehat{C}_{X_{+1}X} = \frac{1}{T}\sum_{i=1}^{T} k_{\mathscr{X}}(\cdot, X_{i+1}) \otimes k_{\mathscr{X}}(\cdot, X_i)$. In practice, we compute

$$G_X = (k_X(X_i, X_j))_{i,j=1}^{T}, \quad G_Y = (k_Y(Y_i, Y_j))_{i,j=1}^{T}, \quad \text{and} \quad G_{XX_{+1}} = (k_X(X_i, X_{j+1}))_{i,j=1}^{T},$$

where $G_{XX_{+1}}$ is the "transfer" matrix.

In the testing phase, given new observations $\tilde{y}_1, \ldots, \tilde{y}_t$, the prediction and correction steps are kernelized. Suppose we already have an estimate of the kernel mean of $p(x_t|\tilde{y}_1, \ldots, \tilde{y}_t)$ in the form

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

where $\alpha_i^{(t)} = \alpha_i^{(t)}(\tilde{y}_1, \ldots, \tilde{y}_t)$ are the coefficients at time t. For the prediction step (1.14), we can simply apply the kernel sum rule (1.7) to estimate the kernel mean of $p(x_{t+1}|\tilde{y}_1, \ldots, \tilde{y}_t)$:

$$\widehat{m}_{x_{t+1}|\tilde{y}_1,\ldots,\tilde{y}_t} = \widehat{C}^{(n)}_{X_{+1}X}\big(\widehat{C}^{(n)}_{XX} + \varepsilon_T I\big)^{-1}\widehat{m}_{x_t|\tilde{y}_1,\ldots,\tilde{y}_t} =: \sum_{j=1}^{T} \beta_j^{(t+1)} k(\cdot, X_{j+1}),$$
$$\text{where} \quad \beta^{(t+1)} = (G_X + T\varepsilon_T I_T)^{-1} G_X \alpha^{(t)}, \tag{1.16}$$

In the correction step (1.15), the kernel Bayes' rule first computes $\widehat{m}_{(y_{t+1}x_{t+1})|\tilde{y}_1,\ldots,\tilde{y}_t} = \widehat{C}^{(n)}_{(YX)X}(\widehat{C}^{(n)}_{XX} + \varepsilon_T I)^{-1}\widehat{m}_{x_{t+1}|\tilde{y}_1,\ldots,\tilde{y}_t}$ and $\widehat{m}_{(y_{t+1}y_{t+1})|\tilde{y}_1,\ldots,\tilde{y}_t} = \widehat{C}^{(n)}_{(YY)X}(\widehat{C}^{(n)}_{XX} + \varepsilon_T I)^{-1}$ $\widehat{m}_{x_{t+1}|\tilde{y}_1,\ldots,\tilde{y}_t}$, of which the coefficients are given by

$$\mu^{(t+1)} = \big(G_X + T\varepsilon_T I_T\big)^{-1} G_{XX_{+1}} \beta^{(t+1)}, \tag{1.17}$$

and next takes the conditioning, which yields

$$\alpha^{(t+1)} = \Lambda^{(t+1)} G_Y \big((\Lambda^{(t+1)} G_Y)^2 + \delta_T I_T\big)^{-1} \Lambda^{(t+1)} \mathbf{k}_Y(\tilde{y}_{t+1}), \tag{1.18}$$

where $\Lambda^{(t+1)} = \mathrm{diag}(\mu_1^{(t+1)}, \ldots, \mu_T^{(t+1)})$. Equations (1.16)–(1.18) describe the sequential update rule of the KBR filtering. The initial estimate $\widehat{m}^{(1)}_{x_1|\tilde{y}_1} = \sum_{i=1}^{T} \alpha_i^{(1)} k(\cdot, X_i)$ can be computed by applying the KBR. We can also use the estimate of the

[4] Although the samples are not i.i.d., we assume an appropriate mixing condition and thus the empirical covariances converge to the covariances with respect to the stationary distribution as $T \to \infty$.

Table 1.2 Algorithm of the KBR filter

Input: Training data $(X_1, Y_1), \ldots, (X_T, Y_T)$, regularization constants ε_T, δ_T, kernels k_X, k_Y.

Training phase:

- Compute $G_X = (k_X(X_i, X_j))_{i,j=1}^T$, $G_Y = (k_Y(Y_i, Y_j))_{i,j=1}^T$, $G_{XX_{+1}} = (k_X(X_i, X_{j+1}))_{i,j=1}^T$.

Testing phase:

- Compute the initial estimate $\alpha^{(1)}$ given $\tilde{y}_1$.
- For $t = 1, 2, \ldots$, given $\tilde{y}_{t+1}$, do the following
  1. $\beta^{(t+1)} = (G_X + T\varepsilon_T I_T)^{-1} G_X \alpha^{(t)}$.
  2. $\mu^{(t+1)} = \left(G_X + T\varepsilon_T I_T\right)^{-1} G_{XX_{+1}} \beta^{(t+1)}$, $\Lambda^{(t+1)} = \mathrm{Diag}(\mu^{(t+1)})$.
  3. $\alpha^{(t+1)} = \Lambda^{(t+1)} G_Y \left((\Lambda^{(t+1)} G_Y)^2 + \delta_T I_T\right)^{-1} \Lambda^{(t+1)} \mathbf{k}_Y(\tilde{y}_{t+1})$

conditional kernel mean $E[k(\cdot, X)|Y = \tilde{y}_1]$ if the prior $\pi(X_1)$ is not available. The computation for the sequential filtering is summarized in Table 1.2.

Applications of the KBR filter to artificial data and camera-angle estimation problems are shown in [11], which demonstrates favorable performance of the KBR filter in comparison with other methods.

1.4.2 Discussions

The matrix inversion $(G_X + T\varepsilon_T I_T)^{-1}$ can be computed only once before the testing phase, while $((\Lambda^{(t+1)} G_Y)^2 + \delta_T I_T)^{-1}$ must be computed every time step in the testing phase, since it depends on $\widehat{\mu}^{(t+1)}$. Direct matrix inversion would cost $O(T^3)$, which is not feasible for large T. Substantial reduction in computational cost can be achieved by low-rank matrix approximations such as incomplete Cholesky factorization. Given an approximation of rank r for the Gram matrices and transfer matrix, the Woodbury identity yields the computation costs just $O(Tr^2)$ for each time step. In fact, let $G_X \approx R_X R_X^T$, $G_Y \approx R_Y R_Y^T$, and $G_{XX_+} \approx A_X B_{X_+}^T$ be the low-rank approximations, where the rank of R_X, R_Y, A_X and B_{X_+} is r at most. It is easy to see from the Woodbury identity that

$$\beta^{(t+1)} \approx \frac{1}{T\varepsilon_T}\left\{R_X R_X^T \alpha^{(t)} - R_X (R_X^T R_X + T\varepsilon_T I_r)^{-1} (R_X^T R_X) R_X^T \alpha^{(t)}\right\},$$

$$\mu^{(t+1)} \approx \frac{1}{T\varepsilon_T}\left\{A_X B_{X_+}^T \beta^{(t+1)} - R_X (R_X^T R_X + T\varepsilon_T I_r)^{-1} (R_X^T A_X) B_{X_+}^T \beta^{(t+1)}\right\},$$

and

$$\alpha^{(t+1)} \approx \frac{1}{\delta_T}\Big\{\Lambda^{(t+1)} R_Y R_Y^T \Lambda^{(t+1)} \mathbf{k}_Y(\tilde{y}_{t+1}) - \Lambda^{(t+1)} R_Y H_Y \big(H_Y^2 + \delta_T I_r\big)^{-1} H_Y R_Y^T \Lambda^{(t+1)} \mathbf{k}_Y(\tilde{y}_{t+1})\Big\},$$

where $H_Y = R_Y^T \Lambda^{(t+1)} R_Y$ $(r \times r)$. Since $\Lambda^{(t+1)}$ is a diagonal matrix, all of the above computation can be done at the cost $O(Tr^2)$.

For the nonparametric state-space models, where training data are given as in the KBR filter, an alternative method is the conditional density estimation, including kernel density estimation or partitioning of the space [25, 34]. It is known, however, that these estimators have low estimation accuracy if the dimensionality is more than several. Empirical studies have shown that the KBR approach gives better estimation accuracy than the density estimation approach for large-dimensional cases; see [11].

Another possible Bayesian method for the nonparametric setting is Gaussian processes. An advantage of Gaussian processes is that one can use standard techniques of Bayesian inference such as hyperparameter selection with the marginal likelihood. Also, direct computation of the posterior is possible. On the other hand, the obtained posterior is unimodal by the nature of Gaussian distribution so that it may not be suitable for problems where multimodal posteriors are essential [22, 23]. In addition, since Gaussian processes are basically a model with univariate response, it is difficult to handle the correlation among a large number of response variables.

A possible limitation of the KBR filter is the assumption that training data exist including the state variable. While one might think it unrealistic, there are indeed some problems where one can obtain training data. One of such cases is expensive measurement: although one can observe the state variable, the measurement is very expensive, and one wishes to use a limited number of training data for inference. For instance, in sensor-based localization problems, pairs of sensor and location data can be once measured with some expensive devises and used for location estimation based solely on new sensor information [18, 27]. Another situation is that states are observed with considerable time delay. In this case, we can use the observed state variables for training, but the current state variable is not known and to be estimated.

It is true that performance of any kernel methods depends on the choice of a kernel. Additionally, in the KBR there are two regularization parameters to be chosen as hyperparameters. In the KBR filter, since we have training data for state variables, we can evaluate the prediction accuracy and thus use the validation approach by dividing the training data into the data for training and evaluation. This method for hyperparameter choice has been successfully used in the filtering applications of KBR in [11, 20].

This article discusses only the fully nonparametric setting of state-space models; both of the state transition and observation model are unknown and estimated nonparametrically. There are, however, semiparametric situations, where one of them is known. Consider vision-based robot localization problems, where the state x_t is the location and orientation of a robot, while the observation y_t is a movie image taken

by video camera mounted on the robot. In this case, it is easy to provide a reasonable parametric model for the dynamics of robot move. On the other hand, the observation model from the location/orientation to the image is too complex and environment-dependent. It is thus preferable to apply a nonparametric method based on data for this observation model. Since the kernel method is purely a nonparametric method expressing the information with Gram matrices, it is not straightforward to combine the kernel Bayesian approach with parametric models. Reference [20] has proposed the kernel Monte Carlo filter, which is a combination of sampling and KBR method for the semiparametric situation, and demonstrated the preferable performance of the proposed method for the vision-based robot localization problem.

1.5 Conclusions

This article has provided detailed explanations of recently proposed kernel mean approach to Bayesian inference. The basic ideas, intuitions, and implementation issues have been discussed in details. A new result on the convergence rate of the estimator of conditional kernel mean has been also presented. As an application of the KBR approach, nonparametric state-space models are discussed focusing the algorithm and efficient computation.

Acknowledgments The author has been supported in part by MEXT Grant-in-Aid for Scientific Research on Innovative Areas 25120012.

Appendix: Proof of Theorem 1.3.2

First, we show a lemma to derive a convergence rate of conditional kernel mean.

Lemma 1.5.1 *Assume that the kernels are measurable and bounded. Let* $N(\varepsilon) := \mathrm{Tr}[C_{YY}(C_{YY}+\varepsilon I)^{-1}]$ *and* ε_n *be a constant such that* $\varepsilon_n \to 0$ *as* $n \to \infty$. *Then,*

$$\left\| (\widehat{C}_{YY}^{(n)} - C_{YY})(C_{YY}+\varepsilon_n I)^{-1} \right\| = O_p\left(\frac{1}{\varepsilon_n n} + \sqrt{\frac{N(\varepsilon_n)}{\varepsilon_n n}} \right)$$

and

$$\left\| (\widehat{C}_{XY}^{(n)} - C_{XY})(C_{YY}+\varepsilon_n I)^{-1} \right\| = O_p\left(\frac{1}{\varepsilon_n n} + \sqrt{\frac{N(\varepsilon_n)}{\varepsilon_n n}} \right)$$

as $n \to \infty$.

Proof The first result is shown in [4] (page 349). While the proof of the second one is similar, it is shown below for completeness.

Let ξ_{yx} be an element in $\mathscr{H}_{\mathscr{Y}} \otimes \mathscr{H}_{\mathscr{X}}$ defined by

$$\xi_{yx} := \left\{(C_{YY} + \varepsilon_n I)^{-1} k(\cdot, y)\right\} \otimes k(\cdot, x).$$

With identification between $Hy \otimes \mathscr{H}_{\mathscr{X}}$ and the Hilbert–Schmidt operators from $\mathscr{H}_{\mathscr{X}}$ to $\mathscr{H}_{\mathscr{Y}}$,

$$E[\xi_{YX}] = (C_{YY} + \varepsilon_n I)^{-1} C_{YX}.$$

Take $a > 0$ such that $k(x, x) \leq a^2$ and $k(y, y) \leq a^2$. It follows from $\|f \otimes g\| = \|f\| \, \|g\|$ and $\|(C_{YY} + \varepsilon_n I)^{-1}\| \leq 1/\varepsilon_n$ that

$$\|\xi_{yx}\| = \left\|(C_{YY} + \varepsilon_n I)^{-1} k(\cdot, y)\right\| \left\|k(\cdot, x)\right\| \leq \frac{1}{\varepsilon_n} \|k(\cdot, y)\| \, \|k(\cdot, x)\| \leq \frac{a^2}{\varepsilon_n},$$

and

$$\begin{aligned}
E\|\xi_{YX}\|^2 &= E\left\|\{(C_{YY} + \varepsilon_n I)^{-1} k(\cdot, Y)\} \otimes k(\cdot, X)\right\|^2 \\
&= E\|k(\cdot, X)\|^2 \left\|(C_{YY} + \varepsilon_n I)^{-1} k(\cdot, Y)\right\|^2 \\
&\leq a^2 E\left\|(C_{YY} + \varepsilon_n I)^{-1} k(\cdot, Y)\right\|^2 \\
&= a^2 E\left\langle (C_{YY} + \varepsilon_n I)^{-2} k(\cdot, Y), k(\cdot, Y)\right\rangle \\
&= a^2 E \mathrm{Tr}\left[(C_{YY} + \varepsilon_n I)^{-2} (k(\cdot, Y) \otimes k(\cdot, Y)^*)\right] \\
&= a^2 \mathrm{Tr}\left[(C_{YY} + \varepsilon_n I)^{-2} C_{YY}\right] \\
&\leq \frac{a^2}{\varepsilon_n} \mathrm{Tr}\left[(C_{YY} + \varepsilon_n I)^{-1} C_{YY}\right] = \frac{a^2}{\varepsilon_n} N(\varepsilon_n).
\end{aligned}$$

Here $k(\cdot, Y)^*$ is the dual element of $k(\cdot, Y)$ and $k(\cdot, Y) \otimes k(\cdot, Y)^*$ is regarded as an operator on $\mathscr{H}_{\mathscr{Y}}$. In the last inequality, $(C_{YY} + \varepsilon_n I)^{-1}$ in the trace is replaced by its upper bound $\varepsilon_n^{-1} I$. Since $\frac{1}{n} \sum_{i=1}^n (C_{YY} + \varepsilon_n I)^{-1} \xi_{Y_i X_i} = (C_{YY} + \varepsilon_n I)^{-1} \widehat{C}_{YX}^{(n)}$, it follows from Proposition 2 in [4] that for all $n \in \mathbb{N}$ and $0 < \eta < 1$

$$\begin{aligned}
\Pr\Bigg(&\left\|(C_{YY} + \varepsilon_n I)^{-1} \widehat{C}_{YX}^{(n)} - (C_{YY} + \varepsilon_n I)^{-1} C_{YX}\right\| \\
&\geq 2\left(\frac{2a^2}{n\varepsilon_n} + \sqrt{\frac{a^2 N(\varepsilon_n)}{\varepsilon_n n}}\right) \log \frac{2}{\eta} \Bigg) \leq \eta,
\end{aligned}$$

which proves the assertion. □

Proof of Theorem 1.3.2 First, we have

$$\begin{aligned}
&\left\|\widehat{C}_{XY}^{(n)}(\widehat{C}_{YY}^{(n)}+\varepsilon_n I)^{-1}k_{\mathscr{Y}}(\cdot,y_0)-E[k_{\mathscr{X}}(\cdot,X)|Y=y_0]\right\|_{\mathscr{H}_{\mathscr{X}}}\\
&\le\left\|\widehat{C}_{XY}^{(n)}(\widehat{C}_{YY}^{(n)}+\varepsilon_n I)^{-1}k_{\mathscr{Y}}(\cdot,y_0)-C_{XY}(C_{YY}+\varepsilon_n I)^{-1}k_{\mathscr{Y}}(\cdot,y_0)\|_{\mathscr{H}_{\mathscr{X}}}\right. \qquad (1.19)\\
&\quad+\left\|C_{XY}(C_{YY}+\varepsilon_n I)^{-1}k_{\mathscr{Y}}(\cdot,y_0)-E[k_{\mathscr{X}}(\cdot,X)|Y=y_0]\right\|_{\mathscr{H}_{\mathscr{X}}}. \qquad (1.20)
\end{aligned}$$

Using the general formula $A^{-1}-B^{-1}=A^{-1}(B-A)B^{-1}$ for any invertible operators A, B, the first term in the right-hand side of the above inequality is upper bounded by

$$\begin{aligned}
&\left\|(\widehat{C}_{XY}^{(n)}-C_{XY})(\widehat{C}_{YY}^{(n)}+\varepsilon_n I)^{-1}k_{\mathscr{Y}}(\cdot,y_0)\right\|_{\mathscr{H}_{\mathscr{X}}}\\
&\quad+\left\|C_{XY}(C_{YY}+\varepsilon_n I)^{-1}(C_{YY}-\widehat{C}_{YY}^{(n)})(\widehat{C}_{YY}^{(n)}+\varepsilon_n I)^{-1}k_{\mathscr{Y}}(\cdot,y_0)\right\|_{\mathscr{H}_{\mathscr{X}}}\\
&\le\left\|(\widehat{C}_{XY}^{(n)}-C_{XY})(\widehat{C}_{YY}^{(n)}+\varepsilon_n I)^{-1}\right\|\left\|k_{\mathscr{Y}}(\cdot,y_0)\right\|_{\mathscr{H}_{\mathscr{Y}}}\\
&\quad+\frac{1}{\sqrt{\varepsilon_n}}\|C_{XX}\|^{1/2}\left\|(\widehat{C}_{YY}^{(n)}-C_{YY})(\widehat{C}_{YY}^{(n)}+\varepsilon_n I)^{-1}\right\|\left\|k_{\mathscr{Y}}(\cdot,y_0)\right\|_{\mathscr{H}_{\mathscr{Y}}},
\end{aligned}$$

where in the second inequality the decomposition $C_{XY}=C_{XX}^{1/2}W_{XY}C_{YY}^{1/2}$ with some $W_{XY}:\mathscr{H}_{\mathscr{Y}}\to\mathscr{H}_{\mathscr{X}}$ ($\|W_{XY}\|\le 1$) [2] is used. It follows from Lemma 1.5.1 that

$$\begin{aligned}
&\left\|\widehat{C}_{XY}^{(n)}(\widehat{C}_{YY}^{(n)}+\varepsilon_n I)^{-1}k_{\mathscr{Y}}(\cdot,y_0)-C_{XY}(C_{YY}+\varepsilon_n I)^{-1}k_{\mathscr{Y}}(\cdot,y_0)\|_{\mathscr{H}_{\mathscr{X}}}\right.\\
&\qquad=O_p\left(\varepsilon_n^{-1/2}\left\{\frac{1}{\varepsilon_n n}+\sqrt{\frac{N(\varepsilon_n)}{\varepsilon_n n}}\right\}\right),
\end{aligned}$$

as $n\to\infty$. It is known (Proposition 3, [4]) that, under the assumption on the decay rate of the eigenvalues, $N(\varepsilon)\le\frac{b\beta}{b-1}\varepsilon^{-1/b}$ holds with some $\beta\ge 0$. Since $\varepsilon_n^{-3/2}n^{-1}\ll\varepsilon_n^{-1-\frac{1}{2b}}n^{-1/2}$ for $b>1$ and $n\varepsilon_n\to\infty$, we have

$$\begin{aligned}
&\left\|\widehat{C}_{XY}^{(n)}(\widehat{C}_{YY}^{(n)}+\varepsilon_n I)^{-1}k_{\mathscr{Y}}(\cdot,y_0)-C_{XY}(C_{YY}+\varepsilon_n I)^{-1}k_{\mathscr{Y}}(\cdot,y_0)\|_{\mathscr{H}_{\mathscr{X}}}\right.\\
&\qquad=O_p\left(\varepsilon_n^{-1-\frac{1}{2b}}n^{-1/2}\right), \qquad (1.21)
\end{aligned}$$

as $n\to\infty$.

For the second term of Eq. (1.19), let $\Theta:=E[k(X,\tilde{X})|Y=\cdot,\tilde{Y}=*]\in\mathscr{R}(C_{YY}\otimes C_{YY})$. Note that for any $\varphi\in\mathscr{H}_{\mathscr{Y}}$ we have

$$\begin{aligned}
\langle C_{XY}\varphi,C_{XY}\varphi\rangle&=E[k(X,\tilde{X})\varphi(Y)\varphi(\tilde{Y})]\\
&=E\big[E[k(X,\tilde{X})|Y,\tilde{Y}]\varphi(Y)\varphi(\tilde{Y})\big]=\langle(C_{YY}\otimes C_{YY})\Theta,\varphi\otimes\varphi\rangle_{\mathscr{H}_{\mathscr{Y}}\otimes\mathscr{H}_{\mathscr{Y}}}.
\end{aligned}$$

Similarly,

$$\begin{aligned}\langle C_{XY}\varphi, E[k(\cdot,X)|Y=y_0]\rangle_{\mathscr{H}_\mathscr{X}} &= \langle E[k(X,\tilde{X})|Y=y_0,\tilde{Y}=*], C_{YY}\varphi\rangle_{\mathscr{H}_\mathscr{Y}} \\ &= \langle (I\otimes C_{YY})\Theta, k(\cdot,y_0)\otimes\varphi\rangle_{\mathscr{H}_\mathscr{Y}\otimes\mathscr{H}_\mathscr{Y}}.\end{aligned}$$

It follows form these equalities with $\varphi = (C_{YY}+\varepsilon_n I)^{-1}k_\mathscr{Y}(\cdot,y_0)$ that

$$\begin{aligned}&\left\|C_{XY}(C_{YY}+\varepsilon_n I)^{-1}k_\mathscr{Y}(\cdot,y_0) - E[k_\mathscr{X}(\cdot,X)|Y=y_0]\right\|^2_{\mathscr{H}_\mathscr{X}} \\ &= \Big\langle \Big\{(C_{YY}+\varepsilon_n I)^{-1}C_{YY}\otimes(C_{YY}+\varepsilon_n I)^{-1}C_{YY} - I\otimes(C_{YY}+\varepsilon_n I)^{-1}C_{YY} \\ &\quad - (C_{YY}+\varepsilon_n I)^{-1}C_{YY}\otimes I + I\otimes I\Big\}\Theta, k_\mathscr{Y}(\cdot,y_0)\otimes k_\mathscr{Y}(*,y_0)\Big\rangle_{\mathscr{H}_\mathscr{Y}\otimes\mathscr{H}_\mathscr{Y}}.\end{aligned}$$

From the assumption $\Theta\in\mathscr{R}(\mathbb{C}_{YY}\otimes C_{YY})$, there is $\Psi\in\mathscr{H}_\mathscr{Y}\otimes\mathscr{H}_\mathscr{Y}$ such that $\Theta = (C_{YY}\otimes C_{YY})\Psi$. Let $\{\phi_i\}$ be the eigenvectors of C_{YY} with eigenvalues $\lambda_1\geq\lambda_2\geq\cdots 0$. Since the eigenvectors and eigenvalues of $C_{YY}\otimes C_{YY}$ are given by $\{\phi_i\otimes\phi_j\}_{ij}$ and $\lambda_i\lambda_j$, respectively, with the fact $(C_{YY}+\varepsilon_n I)^{-1}C_{YY}^2\phi_i = (\lambda_i^2/(1+\lambda_i))\phi_i$ and Parseval's theorem we have

$$\begin{aligned}&\Big\|\Big\{(C_{YY}+\varepsilon_n I)^{-1}C_{YY}\otimes(C_{YY}+\varepsilon_n I)^{-1}C_{YY} - I\otimes(C_{YY}+\varepsilon_n I)^{-1}C_{YY} \\ &\quad - (C_{YY}+\varepsilon_n I)^{-1}C_{YY}\otimes I + I\otimes I\Big\}\Theta\Big\|^2_{\mathscr{H}_\mathscr{Y}\otimes\mathscr{H}_\mathscr{Y}} \\ &= \sum_{i,j}\left\{\frac{\lambda_i^2}{\lambda_i+\varepsilon_n}\frac{\lambda_j^2}{\lambda_j+\varepsilon_n} - \frac{\lambda_i^2\lambda_j}{\lambda_i+\varepsilon_n} - \frac{\lambda_i\lambda_j^2}{\lambda_j+\varepsilon_n} + \lambda_i\lambda_j\right\}^2\langle\phi_i\otimes\phi_j,\Psi\rangle^2_{\mathscr{H}_\mathscr{X}\otimes\mathscr{H}_\mathscr{X}} \\ &= \varepsilon_n^4\sum_{i,j}\left\{\frac{\lambda_i\lambda_j}{(\lambda_i+\varepsilon_n)(\lambda_j+\varepsilon_n)}\right\}^2\langle\phi_i\otimes\phi_j,\Psi\rangle^2_{\mathscr{H}_\mathscr{X}\otimes\mathscr{H}_\mathscr{X}} \leq \varepsilon_n^4\|\Psi\|^2_{\mathscr{H}_\mathscr{X}\otimes\mathscr{H}_\mathscr{X}},\end{aligned}$$

which shows

$$\left\|C_{XY}(C_{YY}+\varepsilon_n I)^{-1}k_\mathscr{Y}(\cdot,y_0) - E[k_\mathscr{X}(\cdot,X)|Y=y_0]\right\|_{\mathscr{H}_\mathscr{X}} = O(\varepsilon_n). \tag{1.22}$$

By balancing Eqs. (1.21) and (1.22), the assertion is obtained with $\varepsilon_n = n^{-b/(4b+1)}$. □

References

1. Aronszajn, N.: Theory of reproducing kernels. Trans. Am. Math. Soc. **68**(3), 337–404 (1950)
2. Baker, C.: Joint measures and cross-covariance operators. Trans. Am. Math. Soc. **186**, 273–289 (1973)
3. Berlinet, A., Thomas-Agnan, C.: Reproducing kernel Hilbert Spaces in Probability and Statistics. Kluwer Academic Publisher (2004)
4. Caponnetto, A., De Vito, E.: Optimal rates for regularized least-squares algorithm. Found. Comput. Math. **7**(3), 331–368 (2007)

5. Doucet, A., Freitas, N.D., Gordon, N.: Sequential Monte Carlo Methods in Practice. Springer (2001)
6. Fine, S., Scheinberg, K.: Efficient SVM training using low-rank kernel representations. J. Mach. Learn. Res. **2**, 243–264 (2001)
7. Fukumizu, K., Bach, F., Jordan, M.: Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. J. Mach. Learn. Res. **5**, 73–99 (2004)
8. Fukumizu, K., Bach, F., Jordan, M.: Kernel dimension reduction in regression. Ann. Stat. **37**(4), 1871–1905 (2009)
9. Fukumizu, K., Gretton, A., Sun, X., Schölkopf, B.: Kernel measures of conditional dependence. In: Advances in Neural Information Processing Systems 20, pp. 489–496. MIT Press (2008)
10. Fukumizu, K., R.Bach, F., Jordan, M.I.: Kernel dimension reduction in regression. Technical Report 715, Department of Statistics, University of California, Berkeley (2006)
11. Fukumizu, K., Song, L., Gretton, A.: Kernel Bayes' rule: Bayesian inference with positive definite kernels. J. Mach. Learn. Res. **14**, 3753–3783 (2013)
12. Fukumizu, K., Sriperumbudur, B.K., Gretton, A., Schölkopf, B.: Characteristic kernels on groups and semigroups. Adv. Neural Inf. Proc. Syst. **20**, 473–480 (2008)
13. Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., Smola, A.: A kernel method for the two-sample-problem. In: Advances in Neural Information Processing Systems 19, pp. 513–520. MIT Press (2007)
14. Gretton, A., Borgwardt, K.M., Rasch, M.J., Schölkopf, B., Smola, A.: A kernel two-sample test. J. Mach. Learn. Res. **13**, 723–773 (2012)
15. Gretton, A., Fukumizu, K., Harchaoui, Z., Sriperumbudur, B.: A fast, consistent kernel two-sample test. Adv. Neural Inf. Process. Syst. **22**, 673–681 (2009)
16. Gretton, A., Fukumizu, K., Sriperumbudur, B.: Discussion of: brownian distance covariance. Ann. Appl. Stat. **3**(4), 1285–1294 (2009)
17. Gretton, A., Fukumizu, K., Teo, C.H., Song, L., Schölkopf, B., Smola, A.: A kernel statistical test of independence. In: Advances in Neural Information Processing Systems 20, pp. 585–592. MIT Press (2008)
18. Haeberlen, A., Flannery, E., Ladd, A.M., Rudys, A., Wallach, D.S., Kavraki, L.E.: Practical robust localization over large-scale 802.11 wireless networks. In: Proceedings of 10th International Conference on Mobile computing and networking (MobiCom '04), pp. 70–84 (2004)
19. Kanagawa, M., Fukumizu, K.: Recovering distributions from gaussian rkhs embeddings. J. Mach. Learn. Res. W&CP **3**, 457–465 (2014)
20. Kanagawa, M., Nishiyama, Y., Gretton, A., Fukumizu, K.: Monte carlo filtering using kernel embedding of distributions. In: Proceedings of 28th AAAI Conference on Artificial Intelligence (AAAI-14), pp. 1987–1903 (2014)
21. Kwok, J.Y., Tsang, I.: The pre-image problem in kernel methods. IEEE Trans. Neural Networks **15**(6), 1517–1525 (2004)
22. McCalman, L.: Function embeddings for multi-modal bayesian inference. Ph.D. thesis. School of Information Technology. The University of Sydney (2013)
23. McCalman, L., O'Callaghan, S., Ramos, F.: Multi-modal estimation with kernel embeddings for learning motion models. In: IEEE International Conference on Robotics and Automation (ICRA), pp. 2845–2852 (2013)
24. Mika, S., Schölkopf, B., Smola, A., Müller, K.R., Scholz, M., Rätsch, G.: Kernel PCA and de-noising in feature spaces. In: Advances in Neural Information Pecessing Systems 11, pp. 536–542. MIT Press (1999)
25. Monbet, V., Ailliot, P., Marteau, P.: l^1-convergence of smoothing densities in non-parametric state space models. Stat. Infer. Stoch. Process. **11**, 311–325 (2008)
26. Moulines, E., Bach, F.R., Harchaoui, Z.: Testing for homogeneity with kernel Fisher discriminant analysis. In: Advances in Neural Information Processing Systems 20, pp. 609–616. Curran Associates, Inc. (2008)
27. Quigley, M., Stavens, D., Coates, A., Thrun, S.: Sub-meter indoor localization in unmodified environments with inexpensive sensors. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2010), pp. 2039 – 2046 (2010)

28. Schölkopf, B., Smola, A.: Learning with Kernels. MIT Press (2002)
29. Song, L., Fukumizu, K., Gretton, A.: Kernel embeddings of conditional distributions: a unified kernel framework for nonparametric inference in graphical models. IEEE Sig. Process. Mag. **30**(4), 98–111 (2013)
30. Song, L., Huang, J., Smola, A., Fukumizu, K.: Hilbert space embeddings of conditional distributions with applications to dynamical systems. In: Proceedings of the 26th International Conference on Machine Learning (ICML2009), pp. 961–968 (2009)
31. Sriperumbudur, B.K., Fukumizu, K., Lanckriet, G.: Characteristic kernels and rkhs embedding of measures. J. Mach. Learn. Res. Universality **12**, 2389–2410 (2011)
32. Sriperumbudur, B.K., Gretton, A., Fukumizu, K., Schölkopf, B., Lanckriet, G.: Hilbert space embeddings and metrics on probability measures. J. Mach. Learn. Res. **11**, 1517–1561 (2010)
33. Steinwart, I., Hush, D., Scovel, C.: Optimal rates for regularized least squares regression. Proc. COLT **2009**, 79–93 (2009)
34. Thrun, S., Langford, J., Fox, D.: Monte carlo hidden markov models: Learning non-parametric models of partially observable stochastic processes. In: Proceedings of International Conference on Machine Learning (ICML 1999), pp. 415–424 (1999)
35. Wan, E., and van der Merwe, R.: The unscented Kalman filter for nonlinear estimation. In: Adaptive Systems for Signal Processing, Communications, and Control Symposium (AS-SPCC 2000), pp. 153–158. IEEE (2000)
36. Widom, H.: Asymptotic behavior of the eigenvalues of certain integral equations. Trans. Am. Math. Soc. **109**, 278–295 (1963)
37. Widom, H.: Asymptotic behavior of the eigenvalues of certain integral equations II. Arch. Ration. Mech. Anal. **17**, 215–229 (1964)
38. Williams, C.K.I., Seeger, M.: Using the Nyström method to speed up kernel machines. In: Advances in Neural Information Processing Systems, vol. 13, pp. 682–688. MIT Press (2001)