Kernel Bayes' Rule

Kenji Fukumizu
The Institute of Statistical Mathematics, Tokyo
fukumizu@ism.ac.jp

Le Song
College of Computing
Georgia Institute of Technology
lsong@cc.gatech.edu

Arthur Gretton
Gatsby Unit, UCL
MPI for Intelligent Systems
arthur.gretton@gmail.com

Abstract

A nonparametric kernel-based method for realizing Bayes' rule is proposed, based on kernel representations of probabilities in reproducing kernel Hilbert spaces. The prior and conditional probabilities are expressed as empirical kernel mean and covariance operators, respectively, and the kernel mean of the posterior distribution is computed in the form of a weighted sample. The kernel Bayes' rule can be applied to a wide variety of Bayesian inference problems: we demonstrate Bayesian computation without likelihood, and filtering with a nonparametric state-space model. A consistency rate for the posterior estimate is established.

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) [16]. The inner product between feature mappings need never be computed explicitly, but is given by a positive definite kernel function, which permits efficient computation without the need to deal explicitly with the feature representation. More recently, the *mean* of the RKHS feature map has been used to represent probability distributions, rather than mapping single points: we will refer to these representations of probability distributions as *kernel means*. With an appropriate choice of kernel, the feature mapping becomes rich enough that its expectation uniquely identifies the distribution: the associated RKHSs are termed *characteristic* [6, 7, 22]. Kernel means in characteristic RKHSs have been applied successfully in a number of statistical tasks, including the two sample problem [9], independence tests [10], and conditional independence tests [8]. An advantage of the kernel approach is that these tests apply immediately to any domain on which kernels may be defined.

We propose a general nonparametric framework for Bayesian inference, expressed entirely in terms of kernel means. The goal of Bayesian inference is to find the posterior of x given observation y;

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

where $\pi(x)$ and $p(y|x)$ are respectively the density function of the prior, and the conditional density or likelihood of y given x. In our framework, the posterior, prior, and likelihood are all expressed as kernel means: the update from prior to posterior is called the Kernel Bayes' Rule (KBR). To implement KBR, the kernel means are learned nonparametrically from training data: the prior and likelihood means are expressed in terms of samples from the prior and joint probabilities, and the posterior as a kernel mean of a weighted sample. The resulting updates are straightforward matrix operations. This leads to the main advantage of the KBR approach: in the absence of a specific parametric model or an analytic form for the prior and likelihood densities, we can still perform Bayesian inference by making sufficient observations on the system. Alternatively, we may have a parametric model, but it might be complex and require time-consuming sampling techniques for inference. By contrast, KBR is simple to implement, and is amenable to well-established approximation techniques which yield an overall computational cost linear in the training sample size [5]. We further establish the rate of consistency of the estimated posterior kernel mean to the true posterior, as a function of training sample size.

The proposed kernel realization of Bayes' rule is an extension of the approach used in [20] for state-space models. This earlier work applies a heuristic, however, in which the kernel mean of the previous hidden state and the observation are assumed to combine additively to update the hidden state estimate. More recently, a method for belief propagation using kernel means was proposed [18, 19]: unlike the present work, this directly estimates conditional densities assuming the prior to be uniform. An alternative to kernel means would be to use nonparametric density estimates. Classical approaches include finite distribution estimates on a partitioned domain or kernel density estimation, which perform poorly on high dimensional data. Alternatively, direct estimates of the density ratio may be used in estimating the conditional p.d.f. [24]. By contrast with density estimation approaches, KBR makes it easy to compute posterior expectations (as an RKHS inner product) and to perform conditioning and marginalization, without requiring numerical integration.

2 Kernel expression of Bayes' rule

2.1 Positive definite kernel and probabilities

We begin with a review of some basic concepts and tools concerning statistics on RKHS [1, 3, 6, 7]. 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 $\sum_{i,j=1}^n c_ic_jk(x_i,x_j)\geq 0$ for arbitrary points $x_1,\ldots,x_n$ in Ω and real numbers $c_1,\ldots,c_n$. It is known [1] that a positive definite kernel on Ω uniquely defines a Hilbert space $\mathcal{H}$ (RKHS) consisting of functions on Ω, where $\langle f,k(\cdot,x)\rangle=f(x)$ for any $x\in\Omega$ and $f\in\mathcal{H}$ (reproducing property).

Let $(\mathcal{X},\mathcal{B}_\mathcal{X},\mu_\mathcal{X})$ and $(\mathcal{Y},\mathcal{B}_\mathcal{Y},\mu_\mathcal{Y})$ be measure spaces, and (X,Y) be a random variable on $\mathcal{X}\times\mathcal{Y}$ with probability P. Throughout this paper, it is assumed that positive definite kernels on the measurable spaces are measurable and bounded, where boundedness is defined as $\sup_{x\in\Omega}k(x,x)<\infty$. Let $k_\mathcal{X}$ be a positive definite kernel on a measurable space $(\mathcal{X},\mathcal{B}_\mathcal{X})$, with RKHS $\mathcal{H}_\mathcal{X}$. The *kernel mean* m_X of X on $\mathcal{H}_\mathcal{X}$ is defined by the mean of the $\mathcal{H}_\mathcal{X}$-valued random variable $k_\mathcal{X}(\cdot,X)$, namely

$$m_X=\int k_\mathcal{X}(\cdot,x)dP_X(x). \tag{2}$$

For notational simplicity, the dependence on $k_\mathcal{X}$ in m_X is not shown. Since the kernel mean depends only on the distribution of X (and the kernel), it may also be written m_{P_X}; we will use whichever of these equivalent notations is clearest in each context. From the reproducing property, we have

$$\langle f,m_X\rangle=E[f(X)]\qquad(\forall f\in\mathcal{H}_\mathcal{X}). \tag{3}$$

Let $k_\mathcal{X}$ and $k_\mathcal{Y}$ be positive definite kernels on $\mathcal{X}$ and $\mathcal{Y}$ with respective RKHS $\mathcal{H}_\mathcal{X}$ and $\mathcal{H}_\mathcal{Y}$. The (uncentered) *covariance operator* $C_{YX}:\mathcal{H}_\mathcal{X}\to\mathcal{H}_\mathcal{Y}$ is defined by the relation

$$\langle g,C_{YX}f\rangle_{\mathcal{H}_\mathcal{Y}}=E[f(X)g(Y)]\quad(=\langle g\otimes f,m_{(YX)}\rangle_{\mathcal{H}_\mathcal{Y}\otimes\mathcal{H}_\mathcal{X}})\qquad(\forall f\in\mathcal{H}_\mathcal{X},g\in\mathcal{H}_\mathcal{Y}).$$

It should be noted that C_{YX} is identified with the mean $m_{(YX)}$ in the tensor product space $\mathcal{H}_\mathcal{Y}\otimes\mathcal{H}_\mathcal{X}$, which is given by the product kernel $k_\mathcal{Y}k_\mathcal{X}$ [1]. The identification is standard: the tensor product is isomorphic to the space of linear maps by the correspondence $\psi\otimes\phi\leftrightarrow[h\mapsto\psi\langle\phi,h\rangle]$. We also define $C_{XX}:\mathcal{H}_\mathcal{X}\to\mathcal{H}_\mathcal{X}$ by $\langle f_2,C_{XX}f_1\rangle=E[f_2(X)f_1(X)]$ for any $f_1,f_2\in\mathcal{H}_\mathcal{X}$.

We next introduce the notion of a characteristic RKHS, which is essential when using kernels to manipulate probability measures. A bounded measurable positive definite kernel k is called *characteristic* if $E_{X\sim P}[k(\cdot,X)]=E_{X'\sim Q}[k(\cdot,X')]$ implies $P=Q$: probabilities are uniquely determined by their kernel means [7, 22]. With this property, problems of statistical inference can be cast in terms of inference on the kernel means. A widely used characteristic kernel on $\mathbb{R}^m$ is the Gaussian kernel, $\exp(-\|x-y\|^2/(2\sigma^2))$.

Empirical estimates of the kernel mean and covariance operator are straightforward to obtain. Given an i.i.d. sample $(X_1,Y_1),\ldots,(X_n,Y_n)$ with law P, the empirical kernel mean and covariance operator are respectively

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

where $\widehat{C}_{YX}^{(n)}$ is written in the tensor product form. These are known to be $\sqrt{n}$-consistent in norm.

2.2 Kernel Bayes' rule

We now derive the kernel mean implementation of Bayes' rule. Let Π be a *prior* distribution on $\mathcal{X}$ with p.d.f. $\pi(x)$. In the following, Q and $Q_{\mathcal{Y}}$ denote the probabilities with p.d.f. $q(x,y) = p(y|x)\pi(x)$ and $q_{\mathcal{Y}}(y)$ in Eq. (1), respectively. Our goal is to obtain an estimator of the kernel mean of posterior $m_{Q_{\mathcal{X}}|y} = \int k_{\mathcal{X}}(\cdot,x)q(x|y)d\mu_{\mathcal{X}}(x)$. The following theorem is fundamental in manipulating conditional probabilities with positive definite kernels.

Theorem 1 ([6])**.** *If $E[g(Y)|X=\cdot] \in \mathcal{H}_{\mathcal{X}}$ holds for $g \in \mathcal{H}_{\mathcal{Y}}$, then*

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

If C_{XX} is injective, the above relation can be expressed as

$$E[g(Y)|X=\cdot] = C_{XX}{}^{-1}C_{XY}g. \tag{4}$$

Using Eq. (4), we can obtain an expression for the kernel mean of $Q_{\mathcal{Y}}$.

Theorem 2 ([20])**.** *Assume C_{XX} is injective, and 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 $m_\Pi \in \mathcal{R}(C_{XX})$ and $E[g(Y)|X=\cdot] \in \mathcal{H}_{\mathcal{X}}$ for any $g \in \mathcal{H}_{\mathcal{Y}}$, then*

$$m_{Q_{\mathcal{Y}}} = C_{YX}C_{XX}{}^{-1}m_\Pi. \tag{5}$$

As discussed in [20], the operator $C_{YX}C_{XX}^{-1}$ implements forward filtering of the prior π with the conditional density $p(y|x)$, as in Eq. (1). Note, however, that the assumptions $E[g(Y)|X=\cdot] \in \mathcal{H}_{\mathcal{X}}$ and injectivity of C_{XX} may not hold in general; we can easily provide counterexamples. In the following, we nonetheless derive a population expression of Bayes' rule under these strong assumptions, use it as a prototype for an empirical estimator expressed in terms of Gram matrices, and prove its consistency subject to appropriate smoothness conditions on the distributions.

In deriving kernel realization of Bayes' rule, we will also use Theorem 2 to obtain a kernel mean representation of the *joint* probability Q:

$$m_Q = C_{(YX)X}C_{XX}^{-1}m_\Pi \quad \in \mathcal{H}_{\mathcal{Y}} \otimes \mathcal{H}_{\mathcal{X}}. \tag{6}$$

In the above equation, $C_{(YX)X}$ is the covariance operator from $\mathcal{H}_{\mathcal{X}}$ to $\mathcal{H}_{\mathcal{Y}} \otimes \mathcal{H}_{\mathcal{X}}$ with p.d.f. $\tilde{p}((y,x),x') = p(x,y)\delta_x(x')$, where $\delta_x(x')$ is the point measure at x.

In many applications of Bayesian inference, the probability conditioned on a particular value should be computed. By plugging the point measure at x into Π in Eq. (5), we have a population expression

$$E[k_{\mathcal{Y}}(\cdot,Y)|X=x] = C_{YX}C_{XX}{}^{-1}k_{\mathcal{X}}(\cdot,x), \tag{7}$$

which was used by [20, 18, 19] as the kernel mean of the conditional probability $p(y|x)$. Let (Z,W) be a random variable on $\mathcal{X} \times \mathcal{Y}$ with law Q. Replacing P by Q and x by y in Eq. (7), we obtain

$$E[k_{\mathcal{X}}(\cdot,Z)|W=y] = C_{ZW}C_{WW}^{-1}k_{\mathcal{Y}}(\cdot,y). \tag{8}$$

This is exactly the kernel mean of the posterior which we want to obtain. The next step is to derive the covariance operators in Eq. (8). Recalling that the mean $m_Q = m_{(ZW)} \in \mathcal{H}_{\mathcal{X}} \otimes \mathcal{H}_{\mathcal{Y}}$ can be identified with the covariance operator $C_{ZW} : \mathcal{H}_{\mathcal{Y}} \to \mathcal{H}_{\mathcal{X}}$, and $m_{(WW)} \in \mathcal{H}_{\mathcal{Y}} \otimes \mathcal{H}_{\mathcal{Y}}$ with C_{WW}, we use Eq. (6) to obtain the operators in Eq. (8), and thus the kernel mean expression of Bayes' rule.

The above argument can be rigorously implemented for empirical estimates of the kernel means and covariances. Let $(X_1,Y_1),\ldots, (X_n,Y_n)$ be an i.i.d. sample with law P, and assume a consistent estimator for m_Π given by

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

where $U_1,\ldots,U_\ell$ is the sample that defines the estimator (which need not be generated by Π), and γ_j are the weights. Negative values are allowed for γ_j. The empirical estimators for C_{ZW} and C_{WW} are identified with $\widehat{m}_{(ZW)}$ and $\widehat{m}_{(WW)}$, respectively. From Eq. (6), they are given by

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

where I is the identity and ε_n is the coefficient of Tikhonov regularization for operator inversion.

The next two propositions express these estimators using Gram matrices. The proofs are simple matrix manipulation and shown in Supplementary material. 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.

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 $\widehat{\mathbf{m}}_\Pi = (\sum_{j=1}^\ell \gamma_j k_\mathcal{X}(X_i, U_j))_{i=1}^n \in \mathbb{R}^n$.
2. Compute $\widehat{\mu} = n(G_X + n\varepsilon_n I_n)^{-1}\widehat{\mathbf{m}}_\Pi$.
3. Compute $R_{X|Y} = \Lambda G_Y((\Lambda G_Y)^2 + \delta_n I_n)^{-1}\Lambda$, where $\Lambda = \mathrm{Diag}(\widehat{\mu})$.

Output: $n \times n$ matrix $R_{X|Y}$.

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

Figure 1: Kernel Bayes' Rule Algorithm

Proposition 3. *The Gram matrix expressions of $\widehat{C}_{ZW}$ and $\widehat{C}_{WW}$ are given by*

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

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

$$\widehat{\mu} = n(G_X + n\varepsilon_n I_n)^{-1}\widehat{\mathbf{m}}_\Pi, \qquad \widehat{\mathbf{m}}_{\Pi,i} = \widehat{m}_\Pi(X_i) = \textstyle\sum_{j=1}^\ell \gamma_j k_\mathcal{X}(X_i, U_j). \tag{9}$$

Prop. 3 implies 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 weights $\widehat{\mu}_i$ may be negative, we use another type of Tikhonov regularization in computing Eq. (8),

$$\widehat{m}_{Q_\mathcal{X}|y} := \widehat{C}_{ZW}\big(\widehat{C}_{WW}^2 + \delta_n I\big)^{-1}\widehat{C}_{WW} k_\mathcal{Y}(\cdot, y). \tag{10}$$

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

$$\widehat{m}_{Q_\mathcal{X}|y} = \mathbf{k}_X^T R_{X|Y}\mathbf{k}_Y(y), \qquad R_{X|Y} := \Lambda G_Y((\Lambda G_Y)^2 + \delta_n I_n)^{-1}\Lambda, \tag{11}$$

where $\Lambda = \mathrm{Diag}(\widehat{\mu})$ is a diagonal matrix with elements $\widehat{\mu}_i$ given by Eq. (9), $\mathbf{k}_X = (k_\mathcal{X}(\cdot, X_1), \ldots, k_\mathcal{X}(\cdot, X_n))^T \in {\mathcal{H}_\mathcal{X}}^n$, and $\mathbf{k}_Y = (k_\mathcal{Y}(\cdot, Y_1), \ldots, k_\mathcal{Y}(\cdot, Y_n))^T \in {\mathcal{H}_\mathcal{Y}}^n$.

We call Eqs.(10) or (11) the *kernel Bayes' rule* (KBR): i.e., the expression of Bayes' rule entirely in terms of kernel means. The algorithm to implement KBR is summarized in Fig. 1. If our aim is to estimate $E[f(Z)|W = y]$, that is, the expectation of a function $f \in \mathcal{H}_\mathcal{X}$ with respect to the posterior, then based on Eq. (3) an estimator is given by

$$\langle f, \widehat{m}_{Q_\mathcal{X}|y}\rangle_{\mathcal{H}_\mathcal{X}} = \mathbf{f}_X^T R_{X|Y}\mathbf{k}_\mathcal{Y}(y), \tag{12}$$

where $\mathbf{f}_X = (f(X_1), \ldots, f(X_n))^T \in \mathbb{R}^n$. In using a weighted sample to represent the posterior, KBR has some similarity to Monte Carlo methods such as importance sampling and sequential Monte Carlo ([4]). The KBR method, however, does not generate samples from the posterior, but updates the weights of a sample via matrix operations. We will provide experimental comparisons between KBR and sampling methods in Sec. 4.1.

2.3 Consistency of KBR estimator

We now demonstrate the consistency of the KBR estimator in Eq. (12). We show only the best rate that can be derived under the assumptions, and leave more detailed discussions and proofs to the Supplementary material. We assume 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 the theoretical results, we assume all Hilbert spaces are separable. In the following, $\mathcal{R}(A)$ denotes the range of A.

Theorem 5. *Let $f \in \mathcal{H}_\mathcal{X}$, (Z, W) be a random vector on $\mathcal{X} \times \mathcal{Y}$ such that its law is Q with p.d.f. $p(y|x)\pi(x)$, and $\widehat{m}_\Pi^{(\ell_n)}$ be an estimator of m_Π such that $\|\widehat{m}_\Pi^{(\ell_n)} - m_\Pi\|_{\mathcal{H}_\mathcal{X}} = O_p(n^{-\alpha})$ as $n \to \infty$ for some $0 < \alpha \leq 1/2$. Assume that $\pi/p_X \in \mathcal{R}(C_{XX}^{1/2})$, where p_X is the p.d.f. of P_X, and $E[f(Z)|W = \cdot] \in \mathcal{R}(C_{WW}^2)$. For $\varepsilon_n = n^{-\frac{2}{3}\alpha}$ 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 the estimator of $E[f(Z)|W = y]$ given by Eq. (12).

The condition $\pi/p_X \in \mathcal{R}(C_{XX}^{1/2})$ requires the prior to be smooth. If $\ell_n = n$, and if $\widehat{m}_\Pi^{(n)}$ is a direct empirical kernel mean with an i.i.d. sample of size n from Π, typically $\alpha = 1/2$ and the theorem implies $n^{4/27}$-consistency. While this might seem to be a slow rate, in practice the convergence may be much faster than the above theoretical guarantee.

3 Bayesian inference with Kernel Bayes' Rule

In Bayesian inference, tasks of interest include finding properties of the posterior (MAP value, moments), and computing the expectation of a function under the posterior. We now demonstrate the use of the kernel mean obtained via KBR in solving these problems.

First, we have already seen from Theorem 5 that we may obtain a consistent estimator under the posterior for the expectation of some $f \in \mathcal{H}_\mathcal{X}$. This covers a wide class of functions when characteristic kernels are used (see also experiments in Sec. 4.1).

Next, regarding a point estimate of x, [20] proposes to use the preimage $\widehat{x} = \arg\min_x \|k_\mathcal{X}(\cdot, x) - \mathbf{k}_X^T R_{X|Y} \mathbf{k}_Y(y)\|^2_{\mathcal{H}_\mathcal{X}}$, which represents the posterior mean most effectively by one point. We use this approach in the present paper where point estimates are considered. In the case of the Gaussian kernel, a fixed point method can be used to sequentially optimize x [13].

In KBR the prior and likelihood are expressed in terms of samples. Thus unlike many methods for Bayesian inference, exact knowledge on their densities is not needed, once samples are obtained. The following are typical situations where the KBR approach is advantageous:

- The relation among variables is difficult to realize with a simple parametric model, however we can obtain samples of the variables (e.g. nonparametric state-space model in Sec. 3).
- The p.d.f of the prior and/or likelihood is hard to obtain explicitly, but sampling is possible: (a) In population genetics, branching processes are used for the likelihood to model the split of species, for which the explicit density is hard to obtain. Approximate Bayesian Computation (ABC) is a popular sampling method in these situations [25, 12, 17]. (b) In nonparametric Bayesian inference (e.g. [14]), the prior is typically given in the form of a process without a density.

  The KBR approach can give alternative ways of Bayesian computation for these problems. We will show some experimental comparisons between KBR approach and ABC in Sec. 4.2.
- If a standard sampling method such as MCMC or sequential MC is applicable, the computation given y may be time consuming, and real-time applications may not be feasible. Using KBR, the expectation of the posterior given y is obtained simply by the inner product as in Eq. (12), once $\mathbf{f}_X^T R_{X|Y}$ has been computed.

The KBR approach nonetheless has a weakness common to other nonparametric methods: if a new data point appears far from the training sample, the reliability of the output will be low. Thus, we need sufficient diversity in training sample to reliably estimate the posterior.

In KBR computation, Gram matrix inversion is necessary, which would cost $O(n^3)$ for sample size n if attempted directly. Substantial cost reductions can be achieved by low rank matrix approximations such as the incomplete Cholesky decomposition [5], which approximates a Gram matrix in the form of $\Gamma\Gamma^T$ with $n \times r$ matrix Γ. Computing Γ costs $O(nr^2)$, and with the Woodbury identity, the KBR can be approximately computed with cost $O(nr^2)$.

Kernel choice or model selection is key to the effectiveness of KBR, as in other kernel methods. KBR involves three model parameters: the kernel (or its parameters), 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 applied in a supervised setting, we can use standard cross-validation (CV). A more general approach requires constructing a related supervised problem. Suppose the prior is given by the marginal P_X of P. The posterior density $q(x|y)$ averaged with P_Y is then equal to the marginal density p_X. We are then able to compare the discrepancy of the kernel mean of P_X and the average of the estimators $\widehat{Q}_{\mathcal{X}|y=Y_i}$ over Y_i. This leads to application of K-fold CV approach. Namely, for a partition of $\{1, \ldots, n\}$ into K disjoint subsets $\{T_a\}_{a=1}^K$, let $\widehat{m}_{Q_{\mathcal{X}|y}}^{[-a]}$ be the kernel mean of posterior estimated with data $\{(X_i, Y_i)\}_{i \notin T_a}$, and the prior mean $\widehat{m}_X^{[-a]}$ with data $\{X_i\}_{i \notin T_a}$. We use $\sum_{a=1}^K \big\| \frac{1}{|T_a|} \sum_{j \in T_a} \widehat{m}_{Q_{\mathcal{X}|y=Y_j}}^{[-a]} - \widehat{m}_X^{[a]} \big\|^2_{\mathcal{H}_\mathcal{X}}$ for CV, where $\widehat{m}_X^{[a]} = \frac{1}{|T_a|} \sum_{j \in T_a} k_\mathcal{X}(\cdot, X_j)$.

Application to nonparametric state-space model. Consider the state-space model,

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

where Y_t is observable and X_t is a hidden state. 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. This problem has already been considered in [20], but we give a more principled approach based on KBR. The conditional probability for the transition $q(x_{t+1}|x_t)$ and observation process $p(y|x)$ are represented by the covariance operators as computed with the training sample; $\widehat{C}_{X,X_{+1}}=\frac{1}{T}\sum_{i=1}^{T}k_{\mathcal{X}}(\cdot,X_i)\otimes k_{\mathcal{X}}(\cdot,X_{i+1})$, $\widehat{C}_{XY}=\frac{1}{T}\sum_{i=1}^{T}k_{\mathcal{X}}(\cdot,X_i)\otimes k_{\mathcal{Y}}(\cdot,Y_i)$, and $\widehat{C}_{YY}$ and $\widehat{C}_{XX}$ are defined similarly. Note that though the data are not i.i.d., consistency is achieved by the mixing property of the Markov model.

For simplicity, we focus on the filtering problem, but smoothing and prediction can be done similarly. In filtering, we wish to estimate the current hidden state x_t, given observations $\tilde{y}_1,\ldots,\tilde{y}_t$. The sequential estimate of $p(x_t|\tilde{y}_1,\ldots,\tilde{y}_t)$ can be derived using KBR (we give only a sketch below; see Supplementary material for the detailed derivation). Suppose we already have an estimator of the kernel mean of $p(x_t|\tilde{y}_1,\ldots,\tilde{y}_t)$ in the form

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

where $\alpha_i^{(t)}=\alpha_i^{(t)}(\tilde{y}_1,\ldots,\tilde{y}_t)$ are the coefficients at time t. By applying Theorem 2 twice, the kernel mean of $p(y_{t+1}|\tilde{y}_1,\ldots,\tilde{y}_t)$ is estimated by $\widehat{m}_{y_{t+1}|\tilde{y}_1,\ldots,\tilde{y}_t}=\sum_{i=1}^{T}\widehat{\mu}_i^{(t+1)}k_{\mathcal{Y}}(\cdot,Y_i)$, where

$$\widehat{\mu}^{(t+1)}=(G_X+T\varepsilon_T I_T)^{-1}G_{X,X_{+1}}(G_X+T\varepsilon_T I_T)^{-1}G_X\alpha^{(t)}. \qquad (13)$$

Here $G_{X_{+1}X}$ is the "transfer" matrix defined by $\left(G_{X_{+1}X}\right)_{ij}=k_{\mathcal{X}}(X_{i+1},X_j)$. With the notation $\Lambda^{(t+1)}=\mathrm{Diag}(\widehat{\mu}_1^{(t+1)},\ldots,\widehat{\mu}_T^{(t+1)})$, kernel Bayes' rule yields

$$\alpha^{(t+1)}=\Lambda^{(t+1)}G_Y\left((\Lambda^{(t+1)}G_Y)^2+\delta_T I_T\right)^{-1}\Lambda^{(t+1)}\mathbf{k}_Y(\tilde{y}_{t+1}). \qquad (14)$$

Eqs. (13) and (14) describe the update rule of $\alpha^{(t)}(\tilde{y}_1,\ldots,\tilde{y}_t)$. By contrast with [20], where the estimates of the previous hidden state and observation are assumed to combine additively, the above derivation is based only on applying KBR. In sequential filtering, a substantial reduction of computational cost can be achieved by low rank approximations for the matrices of a training phase: given rank r, the computation costs only $O(Tr^2)$ for each step in filtering.

Bayesian computation without likelihood. When the likelihood and/or prior is not obtained in an analytic form but sampling is possible, the ABC approach [25, 12, 17] is popular for Bayesian computation. The ABC *rejection method* generates a sample from $q(X|Y=y)$ as follows: (1) generate X_t from the prior Π, (2) generate Y_t from $p(y|X_t)$, (3) if $D(y,Y_t)<\rho$, accept X_t; otherwise reject, (4) go to (1). In Step (3), D is a distance on $\mathcal{X}$, and ρ is the tolerance to acceptance.

In the exactly the same situation as the above, the KBR approach gives the following method: (i) generate $X_1,\ldots,X_n$ from the prior Π, (ii) generate a sample Y_t from $p(y|X_t)$ $(t=1,\ldots,n)$, (iii) compute Gram matrices G_X and G_Y with $(X_1,Y_1),\ldots,(X_n,Y_n)$, and $R_{X|Y}\mathbf{k}_Y(y)$.

The distribution of a sample given by ABC approaches the true posterior if $\rho\to 0$, while the empirical posterior estimate of KBR converges to the true one as $n\to\infty$. The computational efficiency of ABC, however, can be arbitrarily low for a small ρ, since X_t is then rarely accepted in Step (3). Finally, ABC generates a sample, which allows any statistic of the posterior to be approximated. In the case of KBR, certain statistics of the posterior (such as confidence intervals) can be harder to obtain, since consistency is guaranteed only for expectations of RKHS functions. In Sec. 4.2, we provide experimental comparisons addressing the trade-off between computational time and accuracy for ABC and KBR.

4 Experiments

4.1 Nonparametric inference of posterior

First we compare KBR and the standard kernel density estimation (KDE). Let $\{(X_i,Y_i)\}_{i=1}^{n}$ be an i.i.d. sample from P on $\mathbb{R}^d\times\mathbb{R}^r$. With p.d.f. $K(x)$ on $\mathbb{R}^d$ and $H(y)$ on $\mathbb{R}^r$, the conditional p.d.f. $p(y|x)$ is estimated by $\widehat{p}(y|x) = \sum_{j=1}^{n} K_{h_X}(x - X_j)H_{h_Y}(y - Y_j)/\sum_{j=1}^{n} K_{h_X}(x - X_j)$, where $K_{h_X}(x) = h_X^{-d}K(x/h_X)$ and $H_{h_Y}(x) = h_Y^{-r}H(y/h_Y)$. Given an i.i.d. sample $\{U_j\}_{j=1}^{\ell}$ from the prior Π, the posterior $q(x|y)$ is represented by the weighted sample (U_i, w_i) with $w_i = \widehat{p}(y|U_i)/\sum_{j=1}^{\ell} \widehat{p}(y|U_j)$ as importance weight (IW).

We compare the estimates of $\int xq(x|y)dx$ obtained by KBR and KDE + IW, using Gaussian kernels for both the methods. Note that with Gaussian kernel, the function $f(x) = x$ does not belong to $\mathcal{H}_{\mathcal{X}}$, and the consistency of the KBR method is not rigorously guaranteed (*c.f.* Theorem 5). Gaussian kernels, however, are known to be able to approximate any continuous function on a compact subset with arbitrary accuracy [23]. We can thus expect that the posterior mean can be estimated effectively.

In the experiments, the dimensionality was given by $r = d$ ranging form 2 to 64. The distribution P of (X, Y) was $N((0, 1_d)^T, V)$ with V randomly generated for each run. The prior Π was $P_X = N(0, V_{XX}/2)$, where V_{XX} is the X-component of V. The sample sizes were $n = \ell = 200$. The bandwidth parameter h_X, h_Y in KDE were set $h_X = h_Y$ and chosen by two ways, the least square cross-validation [15] and the best mean performance, over the set $\{2 * i \mid i = 1, \ldots, 10\}$. For the KBR, we used use two methods to choose the deviation parameter in Gaussian kernel: the median over the pairwise distances in the data [10] and the 10-fold CV described in Sec. 3. Fig. 2 shows the MSE of the estimates over 1000 random points $y \sim N(0, V_{YY})$. While the accuracy of the both methods decrease for larger dimensionality, the KBR significantly outperforms the KDE+IW.

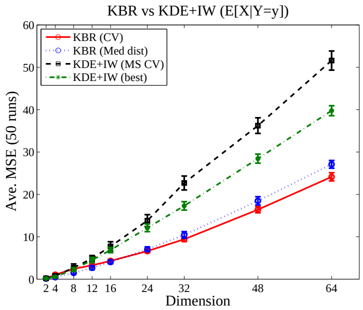


Figure 2: KBR v.s. KDE+IW.

4.2 Bayesian computation without likelihood

We compare KBR and ABC in terms of the estimation accuracy and computational time. To compute the estimation accuracy rigorously, Gaussian distributions are used for the true prior and likelihood. The samples are taken from the same model as in Sec. 4.1, and $\int xq(x|y)dx$ is evaluated at 10 different points of y. We performed 10 runs with different covariance.

For ABC, we used only the rejection method; while there are more advanced sampling schemes [12, 17], implementation is not straightforward. Various parameters for the acceptance are used, and the accuracy and computational time are shown in Fig.3 together with total sizes of generated samples. For the KBR method, the sample sizes n of the likelihood and prior are varied. The regularization parameters are given by $\varepsilon_n = 0.01/n$ and $\delta_n = 2\varepsilon_n$. In KBR, Gaussian kernels are used and the incomplete Cholesky decomposition is employed. The results indicate that KBR achieves more accurate results than ABC in the same computational time.

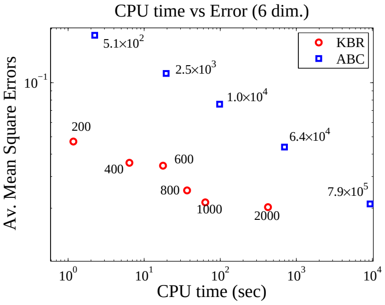
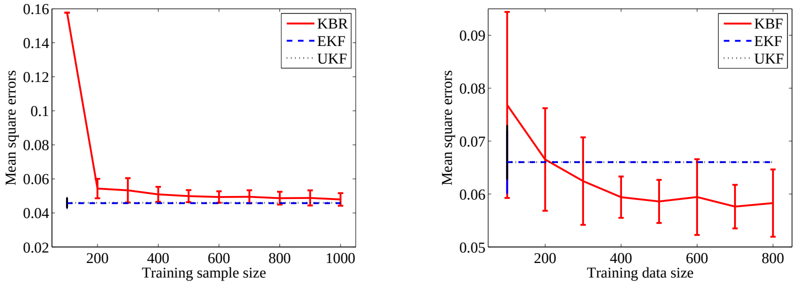


Figure 3: Estimation accuracy and computational time with KBR and ABC.

4.3 Filtering problems

The KBR filter proposed in Sec. 3 is applied. Alternative strategies for state-space models with complex dynamics involve the extended Kalman filter (EKF) and unscented Kalman filter (UKF, [11]). There are some works on nonparametric state-space model or HMM which use nonparametric estimation of conditional p.d.f. such as KDE or partitions [27, 26] and, more recently, kernel method [20, 21]. In the following, the KBR method is compared with linear and nonlinear Kalman filters.

KBR has the regularization parameters ε_T, δ_T, and kernel parameters for $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$ (*e.g.*, the deviation parameter for Gaussian kernel). The validation approach is applied for selecting them by dividing the training sample into two. To reduce the search space, we set $\delta_T = 2\varepsilon_T$ and use the Gaussian kernel deviation $\beta\sigma_{\mathcal{X}}$ and $\beta\sigma_{\mathcal{Y}}$, where $\sigma_{\mathcal{X}}$ and $\sigma_{\mathcal{Y}}$ are the median of pairwise distances among the training samples ([10]), leaving only two parameters β and ε_T to be tuned.



Data (a) Data (b)

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

	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 MSEs and SEs of camera angle estimates (10 runs).

We first use two synthetic data sets with KBR, EKF, and UKF, assuming that EKF and UKF *know* the exact dynamics. The dynamics has a hidden state $X_t = (u_t, v_t)^T \in \mathbb{R}^2$, and is given by

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

where $Z_t \sim N(0, \sigma_h^2 I_2)$ is independent noise. Note that the dynamics of (u_t, v_t) is nonlinear even for $b = 0$. The observation Y_t follows $Y_t = X_t + W_t$, where $W_t \sim N(0, \sigma_o^2 I)$. The two dynamics are defined as follows: (a) (noisy rotation) $\eta = 0.3$, $b = 0$, $\sigma_h = \sigma_o = 0.2$, (b) (noisy oscillatory rotation) $\eta = 0.4$, $b = 0.4$, $M = 8$, $\sigma_h = \sigma_o = 0.2$. The results are shown in Fig. 4. In all the cases, EKF and UKF show unrecognizably small difference. The dynamics in (a) has weak nonlinearity, and KBR shows slightly worse MSE than EKF and UKF. For dataset (b) of strong nonlinearity, KBR outperforms for $T \geq 200$ the nonlinear Kalman filters, which know the true dynamics.

Next, we applied the KBR filter to the camera rotation problem used in [20][1], where the angle of a camera is the hidden variable and the movie frames of a room taken by the camera are observed. We are given 3600 frames of 20×20 RGB pixels ($Y_t \in [0, 1]^{1200}$), where the first 1800 frames are used for training, and the second half are used for test. For the details on the data, see [20]. We make the data noisy by adding Gaussian noise $N(0, \sigma^2)$ to Y_t. Our experiments cover two settings. In the first, we assume we do not know the hidden state X_t is included in $SO(3)$, but is a general 3×3 matrix. In this case, we use the Kalman filter by estimating the relations under a linear assumption, and the KBR filter with Gaussian kernels for X_t and Y_t. In the second setting, we exploit the fact $X_t \in SO(3)$: for the Kalman filter, X_t is represented by a quanternion, and for the KBR filter the kernel $k(A, B) = \mathrm{Tr}[AB^T]$ is used for X_t. Table 1 shows the Frobenius norms between the estimated matrix and the true one. The KBR filter significantly outperforms the Kalman filter, since KBR has the advantage in extracting the complex nonlinear dependence of the observation on the hidden state.

5 Conclusion

We have proposed a general, novel framework for implementing Bayesian inference, where the prior, likelihood, and posterior are expressed as kernel means in reproducing kernel Hilbert spaces. The model is expressed in terms of a set of training samples, and inference consists of a small number of straightforward matrix operations. Our approach is well suited to cases where simple parametric models or an analytic forms of density are not available, but samples are easily obtained. We have addressed two applications: Bayesian inference without likelihood, and sequential filtering with nonparametric state-space model. Future studies could include more comparisons with sampling approaches like advanced Monte Carlo, and applications to various inference problems such as nonparametric Bayesian models and Bayesian reinforcement learning.

Acknowledgements. KF was supported in part by JSPS KAKENHI (B) 22300098.

[1] Due to some difference in noise model, the results here are not directly comparable with those of [20].

References

[1] N. Aronszajn. Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, 68(3):337–404, 1950.

[2] C.R. Baker. Joint measures and cross-covariance operators. *Trans. Amer. Math. Soc.*, 186:273–289, 1973.

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

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

[5] S. Fine and K. Scheinberg. Efficient SVM training using low-rank kernel representations. *JMLR*, 2:243–264, 2001.

[6] K. Fukumizu, F.R. Bach, and M.I. Jordan. Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *JMLR*, 5:73–99, 2004.

[7] K. Fukumizu, F.R. Bach, and M.I. Jordan. Kernel dimension reduction in regression. *Anna. Stat.*, 37(4):1871–1905, 2009.

[8] K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf. Kernel measures of conditional dependence. In *Advances in NIPS 20*, pages 489–496. MIT Press, 2008.

[9] A. Gretton, K.M. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola. A kernel method for the two-sample-problem. In *Advances in NIPS 19*, pages 513–520. MIT Press, 2007.

[10] A. Gretton, K. Fukumizu, C. H. Teo, L. Song, B. Schölkopf, and A. Smola. A kernel statistical test of independence. In *Advances in NIPS 20*, pages 585–592. MIT Press, 2008.

[11] S.J. Julier and J.K. Uhlmann. A new extension of the Kalman filter to nonlinear systems. In *Proc. AeroSense: The 11th Intern. Symp. Aerospace/Defence Sensing, Simulation and Controls*, 1997.

[12] P. Marjoram, Jo. Molitor, V. Plagnol, and S. Tavare. Markov chain monte carlo without likelihoods. *PNAS*, 100(26):15324–15328, 2003.

[13] S. Mika, B. Schölkopf, A. Smola, K.-R. Müller, M. Scholz, and G. Rätsch. Kernel pca and de-noising in feature spaces. In *Advances in NIPS 11*, pages 536–542. MIT Press, 1999.

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

[15] M. Rudemo. Empirical choice of histograms and kernel density estimators. *Scandinavian J. Statistics*, 9(2):pp. 65–78, 1982.

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

[17] S. A. Sisson, Y. Fan, and M. M. Tanaka. Sequential monte carlo without likelihoods. *PNAS*, 104(6):1760–1765, 2007.

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

[19] L. Song, A. Gretton, D. Bickson, Y. Low, and C. Guestrin. Kernel belief propagation. In *AISTATS 2011*.

[20] L. Song, J. Huang, A. Smola, and K. Fukumizu. Hilbert space embeddings of conditional distributions with applications to dynamical systems. *Proc ICML2009*, pages 961–968. 2009.

[21] L. Song and S. M. Siddiqi and G. Gordon and A. Smola. Hilbert Space Embeddings of Hidden Markov Models. *Proc. ICML2010*, 991–998. 2010.

[22] B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G. R.G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *JMLR*, 11:1517–1561, 2010.

[23] I. Steinwart. On the Influence of the Kernel on the Consistency of Support Vector Machines. *JMLR*, 2:67–93, 2001.

[24] M. Sugiyama, I. Takeuchi, T. Suzuki, T. Kanamori, H. Hachiya, and D. Okanohara. Conditional density estimation via least-squares density ratio estimation. In *AISTATS 2010*, pages 781–788, 2010.

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

[26] S. Thrun, J. Langford, and D. Fox. Monte carlo hidden markov models: Learning non-parametric models of partially observable stochastic processes. In *ICML 1999*, pages 415–424, 1999.

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