Remarks on kernel Bayes’ rule

Hisashi Johno¹, Kazunori Nakamoto²* and Tatsuhiko Saigo²

Abstract: The kernel Bayes’ rule has been proposed as a nonparametric kernel-based method to realize Bayesian inference in reproducing kernel Hilbert spaces. However, we demonstrate both theoretically and experimentally that the way of incorporating the prior in the kernel Bayes’ rule is unnatural. In particular, we show that under some reasonable conditions, the posterior in the kernel Bayes’ rule is completely unaffected by the prior, which seems to be irrelevant in the context of Bayesian inference. We consider that this phenomenon is in part due to the fact that the assumptions in the kernel Bayes’ rule do not hold in general.

Subjects: Mathematics & Statistics

Keywords: Kernel method; Bayes’ rule; reproducing kernel Hilbert space

1. Introduction

The kernel Bayes’ rule has recently emerged as a novel framework for Bayesian inference (Fukumizu, Song, & Gretton, 2013; Song, Fukumizu, & Gretton, 2014; Song, Huang, Smola, & Fukumizu, 2009). It is generally agreed that, in this framework, we can estimate the kernel mean of the posterior distribution, given kernel mean expressions of the prior and likelihood distributions. Since the distributions are mapped and nonparametrically manipulated in infinite-dimensional feature spaces called reproducing kernel Hilbert spaces (RKHS), it is believed that the kernel Bayes’ rule can accurately evaluate the statistical features of high-dimensional data and enable Bayesian inference even if there were no appropriate parametric models. To date, several applications of the kernel Bayes’ rule have been reported (Fukumizu et al., 2013; Kanagawa et al., 2014). However, the basic theory and the algorithm of the kernel Bayes’ rule might need to be modified because of the following reasons:

ABOUT THE AUTHORS

Hisashi Johno is a Ph.D student at the Department of Mathematical Sciences, Faculty of Medicine, University of Yamanashi, Japan. His current research interests include probability theory and interpretable machine learning.

Kazunori Nakamoto is a professor of mathematics at Center for Medical Education and Sciences, Faculty of Medicine, University of Yamanashi, Japan. His main research interests include algebraic geometry, invariant theory, and the moduli of representations.

Tatsuhiko Saigo is an associate professor of probability and statistics at Center for Medical Education and Sciences, Faculty of Medicine, University of Yamanashi, Japan. His research fields include probability theory, statistics, and applied mathematics.

PUBLIC INTEREST STATEMENT

This paper examines the validity of the kernel Bayes’ rule, a recently proposed nonparametric framework for Bayesian inference. The researchers on the kernel Bayes’ rule are aiming to apply this method to a wide range of Bayesian inference problems. However, as we demonstrate in this paper, the way of incorporating the prior in the kernel Bayes’ rule seems wrong in the context of Bayesian inference. Several theorems of the kernel Bayes’ rule rely on a strong assumption which does not hold in general.

The problems of the kernel Bayes’ rule seem to be nontrivial and difficult, and we have currently no idea to solve them. We hope that this study would trigger reexamination and correction of the basic framework of the kernel Bayes’ rule.
(1) The posterior in the kernel Bayes’ rule is in some cases completely unaffected by the prior.

(2) The posterior in the kernel Bayes’ rule considerably depends upon the choice of the parameters to regularize covariance operators.

(3) It does not hold in general that conditional expectation functions are included in the RKHS, which is an essential assumption of the kernel Bayes’ rule.

This paper is organized as follows. We begin in Section 2 with a brief review of the kernel Bayes’ rule. In Section 3, we theoretically address the three arguments described above. Numerical experiments are performed in Section 4 to confirm the theoretical results in Section 3. In Section 5, we summarize the theoretical and experimental results and present our conclusions. Some of the proofs for Sections 2 and 3 are given in Section 6.

2. Kernel Bayes’ rule

In this section, we briefly review the kernel Bayes’ rule following (Fukumizu et al., 2013). Let $\mathcal{X}$ and $\mathcal{Y}$ be measurable spaces, $(\mathcal{X}, Y)$ be a random variable with an observed distribution $P$ on $\mathcal{X} \times \mathcal{Y}$, $U$ be a random variable with the prior distribution $\Pi$ on $\mathcal{X}$, and $(Z, W)$ be a random variable with the joint distribution $Q$ on $\mathcal{X} \times \mathcal{Y}$. Note that $Q$ is defined by the prior $\Pi$ and the family $\{P_{Y|x} | x \in \mathcal{X}\}$, where $P_{Y|x}$ denotes the conditional distribution of $Y$ given $X = x$. For each $y \in \mathcal{Y}$, let $Q_{x|y}$ represent the posterior distribution of $Z$ given $W = y$. The aim of the kernel Bayes’ rule is to derive the kernel mean of $Q_{x|y}$.

**Definition 1** Let $k_x$ and $k_y$ be measurable positive definite kernels on $\mathcal{X}$ and $\mathcal{Y}$ such that $E[k_x(X, X)] < \infty$ and $E[k_y(Y, Y)] < \infty$, respectively, where $E[\cdot]$ denotes the expectation operator. Let $\mathcal{H}_x$ and $\mathcal{H}_y$ be the RKHS defined by $k_x$ and $k_y$, respectively. We consider two bounded linear operators $C_{\mathcal{X}X}: \mathcal{H}_x \to \mathcal{H}_y$ and $C_{\mathcal{X}X}: \mathcal{H}_y \to \mathcal{H}_x$ such that

$$\langle g, C_{\mathcal{X}X} f \rangle_{\mathcal{H}_y} = E[f(X)g(Y)] \quad \text{and} \quad \langle f_1, C_{\mathcal{X}X} f_2 \rangle_{\mathcal{H}_x} = E[f_1(X)f_2(X)]$$

for any $f, f_1, f_2 \in \mathcal{H}_x$ and $g \in \mathcal{H}_y$, where $\langle \cdot, \cdot \rangle_{\mathcal{H}_x}$ and $\langle \cdot, \cdot \rangle_{\mathcal{H}_y}$ denote inner products on $\mathcal{H}_x$ and $\mathcal{H}_y$, respectively. The integral expressions for $C_{\mathcal{X}X}$ and $C_{\mathcal{X}X}$ are given by

$$C_{\mathcal{X}X} f(x) = \int_{\mathcal{X} \times \mathcal{Y}} k_x(x, y)f(y)dP(x, y) \quad \text{and} \quad C_{\mathcal{X}X} g(y) = \int_{\mathcal{X}} k_x(x, y)g(y)dP_x(x),$$

where $P_x$ denotes the marginal distribution of $X$. Let $C_{\mathcal{X}X}$ be the bounded linear operator defined by

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

for any $f \in \mathcal{H}_x$ and $g \in \mathcal{H}_y$. Then $C_{\mathcal{X}X}$ is the adjoint of $C_{\mathcal{X}X}$.

**Theorem 1** (Fukumizu et al., 2013, Theorem 1) If $E[g(Y) | X = \cdot] \in \mathcal{H}_x$ for $g \in \mathcal{H}_y$, then $C_{\mathcal{X}X} E[g(Y) | X = \cdot] = C_{\mathcal{X}X} g$.

**Definition 2** Let $Q_x$ denote the marginal distribution of $W$. Assuming that $E[k_x(U, U)] < \infty$ and $E[k_y(W, W)] < \infty$, we can define the kernel means of $\Pi$ and $Q_y$ by

$$m_{\Pi} = E[k_x(\cdot, U)] \quad \text{and} \quad m_{Q_y} = E[k_y(\cdot, W)],$$

respectively. Due to the reproducing properties of $\mathcal{H}_x$ and $\mathcal{H}_y$, the kernel means satisfy $\langle f, m_{\Pi} \rangle_{\mathcal{H}_y} = E[f(U)]$ and $\langle g, m_{Q_y} \rangle_{\mathcal{H}_x} = E[g(W)]$ for any $f \in \mathcal{H}_x$ and $g \in \mathcal{H}_y$.

**Theorem 2** (Fukumizu et al., 2013, Theorem 2) If $C_{\mathcal{X}X}$ is injective, $m_{\Pi} \in \text{Ran}(C_{\mathcal{X}X})$, and $E[g(Y) | X = \cdot] \in \mathcal{H}_x$ for any $g \in \mathcal{H}_y$.
\[ m_{0,j} = c_{ix}c_{xx}^{-1}m_{i}, \]  \tag{2}  

where Ran(C_{xx}) denotes the range of C_{xx}.  

Here we have, for any \( x \in \mathcal{X} \),

\[ E[k_{j}(:, Y)|X = x] = c_{xx}c_{xx}^{-1}k_{x}(:, x) \]  \tag{3}  

by replacing \( m_{ij} \) in Equation (2) for \( k_{x}(:, x) \). It is noted in Fukumizu et al. (2013) that the assumption \( m_{ij} \in \text{Ran}(C_{xx}) \) does not hold in general. In order to remove this assumption, \((C_{xx} + \epsilon I)^{-1}\) has been suggested to be used instead of \( C_{xx}^{-1} \), where \( \epsilon \) is a regularization constant and \( I \) is the identity operator. Thus, the approximations of Equations (2) and (3) are respectively given by

\[ m_{ij}^{\text{reg}} = c_{xx}(C_{xx} + \epsilon I)^{-1}m_{i} \quad \text{and} \quad E^{\text{reg}}[k_{j}(:, Y)|X = x] = c_{xx}(C_{xx} + \epsilon I)^{-1}k_{x}(:, x). \]

Similarly, for any \( y \in \mathcal{Y} \), the approximation of \( m_{iy} \) is provided by

\[ m_{iy}^{\text{reg}} = E^{\text{reg}}[k_{x}(:, Z)|W = y] = c_{zw}(C_{ww} + \delta I)^{-1}k_{y}(:, y), \]  \tag{4}  

where \( \delta \) is a regularization constant and the linear operators \( C_{zw} \) and \( C_{ww} \) will be defined below.

**Definition 3** We consider the kernel means \( m_{a} = m_{zw} \) and \( m_{ww} \) such that

\[ \langle m_{zw}, g \otimes f \rangle_{H_{y} \otimes H_{x}} = E[f(Z)g(W)] \quad \text{and} \quad \langle m_{ww}, g_{1} \otimes g_{2} \rangle_{H_{y} \otimes H_{y}} = E[g_{1}(W)g_{2}(W)] \]

for any \( f, g, g_{1}, g_{2} \in H_{y} \), and \( g, g_{1}, g_{2} \in H_{y} \), where \( \otimes \) denotes the tensor product. Let \( C_{yy} : H_{y} \rightarrow H_{y} \otimes H_{y} \) and \( C_{y} : H_{x} \rightarrow H_{y} \otimes H_{y} \) be bounded linear operators which respectively satisfy

\[ \langle g \otimes f, c_{yx}(\cdot, h) \rangle_{H_{y} \otimes H_{x}} = E[g(Y)f(X)h(X)], \]

\[ \langle g, c_{yy}(\cdot, h) \rangle_{H_{y} \otimes H_{y}} = E[g(Y)h(Y)f(X)] \]  \tag{5}  

for any \( f, h \in H_{x} \) and \( g, g_{1}, g_{2} \in H_{y} \).

From Theorem 2, Fukumizu et al. (2013) proposed that \( m_{zw} \) and \( m_{ww} \) can be given by

\[ m_{zw} = c_{yx}c_{xx}^{-1}m_{i} \in H_{y} \otimes H_{x} \quad \text{and} \quad m_{ww} = c_{yy}c_{xx}^{-1}m_{i} \in H_{y} \otimes H_{y}. \]

In case \( m_{i} \) is not included in \( \text{Ran}(C_{xx}) \), they suggested that \( m_{zw} \) and \( m_{ww} \) could be approximated by

\[ m_{zw}^{\text{reg}} = c_{yx}(C_{xx} + \epsilon I)^{-1}m_{i} \quad \text{and} \quad m_{ww}^{\text{reg}} = c_{yy}(C_{xx} + \epsilon I)^{-1}m_{i}. \]

**Remark 1** (Fukumizu et al., 2013, p. 3760) \( m_{zw} \) and \( m_{ww} \) can respectively be identified with \( C_{zw} \) and \( C_{ww} \).

Here, we introduce the empirical method for estimating the posterior kernel mean \( m_{a,\text{est}} \) following (Fukumizu et al., 2013).

**Definition 4** Suppose we have an independent and identically distributed (i.i.d.) sample \( \{(X_{i}, Y_{i})\}_{i=1}^{n} \) from the observed distribution \( P \) on \( \mathcal{X} \times \mathcal{Y} \) and a sample \( \{(U_{j})\}_{j=1}^{m} \) from the prior distribution \( \Pi \) on \( \mathcal{X} \). The prior kernel mean \( m_{i} \) is estimated by

\[ \hat{m}_{i} = \sum_{j=1}^{m} j_{i}k_{x}(:, U_{j}), \]  \tag{6}  

\[ m_{a,\text{est}} = \sum_{i=1}^{n} \sum_{j=1}^{m} (i, j)k_{x}(:, U_{j}), \]

\[ m_{a,\text{est}} = \sum_{j=1}^{m} j_{i}k_{x}(:, U_{j}), \]  \tag{6}  

\[ m_{a,\text{est}} = \sum_{j=1}^{m} (i, j)k_{x}(:, U_{j}), \]  \tag{6}  

\[ m_{a,\text{est}} = \sum_{j=1}^{m} (i, j)k_{x}(:, U_{j}), \]  \tag{6}
where \( y_1, \ldots, y_\ell \) are weights. Let us put \( \hat{m}_n = (\hat{m}_n(X_1), \ldots, \hat{m}_n(X_\ell))^T \), \( G_x = (k_x(X_i, X_j))_{1:\ell,1:\ell} \) and 
\( G_y = (k_y(Y_i, Y_j))_{1:\ell,1:\ell} \).

**Proposition 1** (Fukumizu et al., 2013, Proposition 3, revised) Let \( I_n \) denote the identity matrix of size \( n \). The estimates of \( C_{zw} \) and \( C_{ww} \) are given by

\[
\hat{C}_{zw} = \sum_{i=1}^{\ell} \hat{\mu}_i k_x(\cdot, X_i) \otimes k_y(\cdot, Y_i) \quad \text{and} \quad \hat{C}_{ww} = \sum_{i=1}^{\ell} \hat{\mu}_i k_y(\cdot, Y_i) \otimes k_y(\cdot, Y_i),
\]

respectively, where \( \hat{\mu} = (\hat{\mu}_1, \ldots, \hat{\mu}_\ell)^T = (G_x + nI_n)^{-1}\hat{m}_n \).

The proof of this revised proposition is given in Section 6.1. It is suggested in Fukumizu et al. (2013) that Equation (4) can be empirically estimated by

\[
\hat{m}_{Q_{xy}} = \hat{C}_{zw} \left( \hat{C}_{ww}^{-1} + \delta I_n \right)^{-1} \hat{C}_{ww} k_y(\cdot, Y).
\]

**THEOREM 3** (Fukumizu et al., 2013, Proposition 4) Given an observation \( y \in \mathcal{Y} \), \( \hat{m}_{Q_{xy}} \) can be calculated by

\[
\hat{m}_{Q_{xy}} = k_x \Lambda R_{xy} k_y(y), \quad R_{xy} = \Lambda (\Lambda G_y + \delta I)^{-1} \Lambda,
\]

where \( \Lambda = \text{diag}(\hat{\mu}) \) is the diagonal matrix with the elements of \( \hat{\mu} \), \( k_x = (k_x(\cdot, X_1), \ldots, k_x(\cdot, X_\ell))^T \), \( k_y = (k_y(\cdot, Y_1), \ldots, k_y(\cdot, Y_\ell))^T \).

If we want to know the posterior expectation of a function \( f \in \mathcal{H}_x \) given an observation \( y \in \mathcal{Y} \), it is estimated by

\[
\langle f, \hat{m}_{Q_{xy}} \rangle_{\mathcal{H}_x} = f_X^T R_{xy} k_y(y),
\]

where \( f_X = (f(X_1), \ldots, f(X_\ell))^T \).

### 3. Theoretical arguments

In this section, we theoretically support the three arguments raised in Section 1. First, we show in Section 3.1 that the posterior kernel mean \( \hat{m}_{Q_{xy}} \) is completely unaffected by the prior distribution \( \Pi \) under the condition that \( \Lambda \) and \( G_y \) are non-singular. This implies that, at least in some cases, \( \Pi \) does not properly affect \( \hat{m}_{Q_{xy}} \). Second, we mention in Section 3.2 that the linear operators \( C_{xx} \) and \( C_{ww} \) are not always surjective, and address the problems associated with the setting of the regularization parameters \( \epsilon \) and \( \delta \). Third, we demonstrate in Section 3.3 that conditional expectation functions are not generally contained in the RKHS, which means that Theorems 1, 2, and 5–8 in Fukumizu et al. (2013) do not work in some situations.

#### 3.1. Relations between the posterior \( \hat{m}_{Q_{xy}} \) and the prior \( \Pi \)

Let us review Theorem 3. Assume that \( G_x \) and \( \Lambda \) are non-singular matrices. (This assumption is not so strange, as shown in Section 6.2.) The matrix \( R_{xy} = \Lambda G_y \Lambda^{-1} \Lambda G_y \Lambda^{-1} \) tends to \( G_y \) as \( \delta \) tends to 0. Furthermore, if we set \( \delta = 0 \) from the beginning, we obtain \( R_{xy} = G_y^{-1} \). This implies that the posterior kernel mean \( \hat{m}_{Q_{xy}} = k_x^T R_{xy} k_y(y) \) never depends on the prior distribution \( \Pi \) on \( \mathcal{X} \), which seems to be a contradiction to the nature of Bayes’ rule.

Some readers may argue that, even in this case, we should not set \( \delta = 0 \). Then, however, there is ambiguity about why and how the regularization parameters are introduced in the kernel Bayes’ rule, since Fukumizu et al. originally used the regularization parameters just to solve inverse problems as an analog of ridge regression (Fukumizu et al., 2013, p. 3758). They seem to support the validity of the regularization parameters by Theorems 5, 6, 7, and 8 in Fukumizu et al. (2013),
however, these theorems do not work without the strong assumption that conditional expectation functions are included in the RKHS, as will be discussed in Section 3.2. In addition, since the theorems work only when $\delta_\epsilon$ etc. decay to zero sufficiently slowly, it seems that we have no principled way to choose values for the regularization parameters, except for cross-validation or similar techniques. It is worth mentioning that, in our simple experiments in Section 4.2, we could not obtain a reasonable result with the kernel Bayes’ rule using any combination of values for the regularization parameters.

### 3.2. The inverse of the operators $C_{XX}$ and $C_{WW}$

As noted by Fukumizu et al. (2013), the linear operators $C_{XX}$ and $C_{WW}$ are not surjective in some usual cases, the proof of which is given in Section 6.3. Therefore, they proposed an alternative way of obtaining a solution $f \in H_X$ of the equation $C_{XX}f = m_\|$, that is, a regularized inversion $f = (C_{XX} + \epsilon I)^{-1}m_\|$ as an analog of ridge regression, where $\epsilon$ is a regularization parameter and $I$ is an identity operator. One of the disadvantages of this method is that the solution $f = (C_{XX} + \epsilon I)^{-1}m_\|$ depends upon the choice of $\epsilon$. In Section 4.2, we numerically show that the prediction using the kernel Bayes’ rule considerably depends on the regularization parameters $\epsilon$ and $\delta$. Theorems 5–8 in Fukumizu et al. (2013) seem to support the appropriateness of the regularized inversion. However, these theorems work under the condition that conditional expectation functions are contained in the RKHS, which does not hold in some cases as proved in Section 3.3. Furthermore, since we need to assume sufficiently slow decay of the regularization constants $\epsilon$ and $\delta$ in these theorems, it is practically difficult to set appropriate values for $\epsilon$ and $\delta$. A cross-validation procedure seems to be useful for tuning the parameters and we may obtain good experimental results, however, it seems to lack theoretical background.

Instead of the regularized inversion method, we can compute generalized inverse matrices of $G_X$ and $\Delta G_\nu$, given a sample $(\{X_i, Y_i\})_{i=1}^n$. Below, we briefly introduce a generalization of a matrix inversion. For more details, see Horn and Johnson (2013).

**Definition 5** Let $A$ be a matrix of size $m \times n$ over the complex number space $\mathbb{C}$. We say that a matrix $A^*$ of size $n \times m$ is a generalized inverse matrix of $A$ if $AA^*A = A$. We also say that a matrix $A^*$ of size $n \times m$ is the Moore-Penrose generalized inverse matrix of $A$ if $AA^*A$ and $A^*AA^*$ are Hermitian, $AA^*A = A$, and $A^*AA^* = A^*$.

**Remark 2** In fact, any matrix $A$ has the Moore-Penrose generalized inverse matrix $A^*$. Note that $A^*$ is uniquely determined by $A$. If $A$ is square and non-singular, then $A^* = A^T = A^{-1}$. For a generalized inverse matrix $A^*$ of size $n \times m$, $AA^*v = v$ for any vector $v \in \mathbb{C}^m$ if $v$ is contained in the image of $A$. In particular, $A^*v$ is a vector contained in the preimage of $v$ under $A$.

In the calculation of $\hat{m}_{G_X} = k^\dagger_{X,Y} \hat{k}_Y(y)$, we numerically compare the case $\hat{R}_{XY} = (\Lambda'G_Y)^{-1}\Lambda'$ with the original case $\hat{R}_{XY} = \Delta G_Y((\Delta G_Y)^2 + \delta I)^{-1}\Delta$ in Section 4.2, where $\Lambda' = \text{diag}(G_\nu^\dagger \hat{m}_\|)$.

### 3.3. Conditional expectation functions and RKHS

In this subsection, we show that conditional expectation functions are in some cases not contained in the RKHS.

**Definition 6** For $p \in [1, \infty)$, we define the spaces $L^p(\mathbb{R})$, $L^p(\mathbb{R}, \mathbb{C})$, and $L^p(\mathbb{R}^2, \mathbb{R})$ as

$$L^p(\mathbb{R}): = \left\{ f: \mathbb{R} \to \mathbb{R} \mid \int_{-\infty}^{\infty} |f(x)|^p \, dx < \infty \right\},$$

$$L^p(\mathbb{R}, \mathbb{C}): = \left\{ f: \mathbb{R} \to \mathbb{C} \mid \int_{-\infty}^{\infty} |f(x)|^p \, dx < \infty \right\},$$

$$L^p(\mathbb{R}^2, \mathbb{R}): = \left\{ f: \mathbb{R}^2 \to \mathbb{R} \mid \int_{\mathbb{R}^2} |f(x_1, x_2)|^p \, dx_1 dx_2 < \infty \right\}.$$
We also define the $L^p$ norm for $f \in L^p(\mathbb{R})$ or $f \in L^p(\mathbb{R}, \mathbb{C})$ as

$$
\|f\|_p := \left( \int_{-\infty}^{\infty} |f(x)|^p \, dx \right)^{\frac{1}{p}},
$$

and the $L^p$ norm for $f \in L^p(\mathbb{R}^2, \mathbb{R})$ as

$$
\|f\|_p := \left( \int_{\mathbb{R}^2} |f(x_1, x_2)|^p \, dx_1 \, dx_2 \right)^{\frac{1}{p}}.
$$

**Definition 7** For a function $f \in L^1(\mathbb{R}, \mathbb{C}) \cap L^2(\mathbb{R}, \mathbb{C})$, we define its Fourier transform as

$$
\mathcal{F}(f)(t) := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(x) \exp \left(-\sqrt{-1}tx \right) \, dx.
$$

We can uniquely extend the Fourier transform to an isometry $\mathcal{F}: L^2(\mathbb{R}, \mathbb{C}) \to L(\mathbb{R}, \mathbb{C})$. We also define the inverse Fourier transform $\mathcal{F}^{-1}$ on the RKHS as an isometry uniquely determined by

$$
\tilde{f}(t) := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(x) \exp \left(\sqrt{-1}tx \right) \, dx
$$

for $f \in L^1(\mathbb{R}, \mathbb{C}) \cap L^2(\mathbb{R}, \mathbb{C})$.

**Definition 8** Let us define a Gaussian kernel $k_\sigma$ on $\mathbb{R}$ by

$$
k_\sigma(x, y) := \frac{1}{\sqrt{2\pi}\sigma} \exp \left(-\frac{(x-y)^2}{2\sigma^2} \right).
$$

As described in Fukumizu (2014), the RKHS of real-valued functions and complex-valued functions corresponding to the positive definite kernel $k_\sigma$ are given by

$$
\mathcal{H}_G = \left\{ f \in L^2(\mathbb{R}) \left| \int_{-\infty}^{\infty} |\mathcal{F}(f)(t)|^2 \exp \left(\frac{\sigma^2 t^2}{2} \right) \, dt < \infty \right. \right\},
$$

$$
\mathcal{H}_{\mathbb{C}}(\mathbb{R}, \mathbb{C}) = \left\{ f \in L^2(\mathbb{R}, \mathbb{C}) \left| \int_{-\infty}^{\infty} |\mathcal{F}(f)(t)|^2 \exp \left(\frac{\sigma^2 t^2}{2} \right) \, dt < \infty \right. \right\},
$$

respectively, and the inner product of $f$, $g \in \mathcal{H}_G$ or $f$, $g \in \mathcal{H}_{\mathbb{C}}(\mathbb{R}, \mathbb{C})$ on the RKHS is calculated by

$$
\langle f, g \rangle = \int_{-\infty}^{\infty} \bar{f}(t)\tilde{g}(t) \exp \left(\frac{\sigma^2 t^2}{2} \right) \, dt,
$$

where the overline denotes the complex conjugate. Remark that $\mathcal{H}_G$ is a real Hilbert subspace contained in the complex Hilbert space $\mathcal{H}_{\mathbb{C}}(\mathbb{R}, \mathbb{C})$.

Fukumizu et al. (2013) mentioned that the conditional expectation function $\mathbb{E}[g(Y) \mid X = \cdot]$ is not always included in $\mathcal{H}_{y^*}$. Indeed, if the variables $X$ and $Y$ are independent, then $\mathbb{E}[g(Y) \mid X = \cdot]$ becomes a constant function on $\mathcal{X}$, the value of which might be non-zero. In the case that $\mathcal{X} = \mathbb{R}$ and $k_{X^*} = k_G$, the constant function with non-zero value is not contained in $\mathcal{H}_{\mathcal{X}} = \mathcal{H}_G$.

Additionally, in order to prove Theorems 5 and 8 in Fukumizu et al. (2013), they made the assumption that $\mathbb{E}[k_{x^*}(Y, Y) \mid X = x, \bar{X} = \bar{x}] \in \mathcal{H}_{X^*} \otimes \mathcal{H}_{\mathcal{X}}$ and $\mathbb{E}[k_{x^*}(Z, Z) \mid W = y, \bar{W} = \bar{y}] \in \mathcal{H}_{Z} \otimes \mathcal{H}_Z$, where $(X, Y)$ and $(\bar{Z}, \bar{W})$ are independent copies of the random variables $(X, Y)$ and $(Z, W)$ on $\mathcal{X} \times \mathcal{Y}$, respectively. We also see that this assumption does not hold in general. Suppose that $X$ and $Y$ are independent and that so are $\bar{X}$ and $\bar{Y}$. Then $\mathbb{E}[k_{x^*}(Y, Y) \mid X = x, \bar{X} = \bar{x}]$ is a constant function of $(x, \bar{x})$, the value of which might be non-zero. In the case that $\mathcal{X} = \mathbb{R}$ and $k_{x^*} = k_G$, the constant function having non-zero value is not contained in $\mathcal{H}_{\mathcal{X}} = \mathcal{H}_G \otimes \mathcal{H}_G$. Note that $\mathcal{H}_G \otimes \mathcal{H}_G$ is isomorphic to the RKHS corresponding to the kernel $k((x_1, x_2), (\bar{x}_1, \bar{x}_2)) = k_G(x_1, \bar{x}_1)k_G(x_2, \bar{x}_2)$ on $\mathbb{R}^2$, that is,
\[ \mathcal{H}_G \otimes \mathcal{H}_G = \left\{ f \in L^2(\mathbb{R}^2, \mathbb{R}) \mid \int_{\mathbb{R}^2} |\hat{f}(t_1, t_2)|^2 \exp \left( \frac{1}{2} (t_1^2 + t_2^2) \right) dt_1 dt_2 < \infty \right\}, \]

where the Fourier transform of \( f: \mathbb{R}^2 \rightarrow \mathbb{R} \) is defined by

\[ \hat{f}(t_1, t_2) = \lim_{n \to \infty} \left( \frac{1}{\sqrt{2\pi}} \right)^2 \int_{x_1, x_2 \in \mathbb{R}^2} f(x_1, x_2) \exp \left( -\sqrt{-1} (t_1 x_1 + t_2 x_2) \right) dx_1 dx_2. \]

Thus, the assumption that conditional expectation functions are included in the RKHS does not hold in general. Since most of the theorems in Fukumizu et al. (2013) require this assumption, the kernel Bayes’ rule may not work in several cases.

### 4. Numerical experiments

In this section, we perform numerical experiments to illustrate the theoretical results in Sections 3.1 and 3.2. We first introduce probabilistic classifiers in Section 4.1 based on conventional Bayes’ rule assuming Gaussian distributions (BR), the original kernel Bayes’ rule (KBR1), and the kernel Bayes’ rule using Moore-Penrose generalized inverse matrices (KBR2). In Section 4.2, we apply the three classifiers to a binary classification problem with computer-simulated data sets. Numerical experiments are implemented in version 2.7.6 of the Python software (Python Software Foundation, Wolfeboro Falls, NH, USA).

#### 4.1. Algorithms of the three classifiers, BR, KBR1, and KBR2

Let \((X, Y)\) be a random variable with a distribution \(P\) on \(\mathcal{X} \times \mathcal{Y}\), where \(\mathcal{X} = \{C_1, \ldots, C_g\}\) is a family of classes and \(\mathcal{Y} = \mathbb{R}^d\). Let \(\Pi\) and \(Q\) be the prior and the joint distributions on \(\mathcal{X}\) and \(\mathcal{X} \times \mathcal{Y}\), respectively. Suppose we have an i.i.d training sample \([(X_i, Y_i)]_{i=1}^n\) from the distribution \(P\). The aim of this subsection is to derive algorithms of the three classifiers, BR, KBR1, and KBR2, which respectively calculate the posterior probability for each class given an observation \(y \in \mathcal{Y}\), that is, 

\[ Q_{XY}(C_j), \ldots, Q_{XY}(C_g). \]

##### 4.1.1. The algorithm of BR

In BR, we estimate the posterior probability of \(\ell\)-th class \((\ell = 1, \ldots, g)\) given a test value \(y \in \mathcal{Y}\) by

\[ \hat{Q}_{XY}(C_j) = \frac{\hat{P}_{Y|C_j}(y) \Pi(C_j)}{\sum_{k=1}^g \hat{P}_{Y|C_k}(y) \Pi(C_k)}, \]

where \(\hat{P}_{Y|C_j}(\cdot)\) is the density function of the \(d\)-dimensional normal distribution \(\mathcal{N}(\hat{M}_j, \hat{S}_j)\) defined by

\[ \hat{P}_{Y|C_j}(\cdot) = \frac{1}{\sqrt{2\pi}^d |\hat{S}_j|} \exp \left( -\frac{1}{2} (\cdot - \hat{M}_j) \hat{S}_j^{-1} (\cdot - \hat{M}_j) \right). \]

The mean vector \(\hat{M}_j \in \mathbb{R}^d\) and the covariance matrix \(\hat{S}_j \in \mathbb{R}^{d \times d}\) are calculated from the training data of the class \(C_j\).

##### 4.1.2. The algorithm of KBR1

Let us define positive definite kernels \(k_x\) and \(k_y\) as

\[ k_x(x, x') = \begin{cases} 1 & (x = x') \\ 0 & (x \neq x') \end{cases} \quad \text{and} \quad k_y(y, y') = \frac{1}{\sqrt{2\pi} \sigma} \exp \left( -\frac{\|Y - Y'\|^2}{2\sigma^2} \right) \]

for \(x, x' \in \mathcal{X}\) and \(y, y' \in \mathcal{Y}\), and the corresponding RKHS as \(\mathcal{H}_x\) and \(\mathcal{H}_y\), respectively. Here we set \(\|Y\| = \sqrt{\sum_{i=1}^d y_i^2}\) for \(Y = (y_1, y_2, \ldots, y_d)^T \in \mathcal{Y} = \mathbb{R}^d\). Then, the prior kernel mean is given by

\[ \hat{M}_\Pi(\cdot) = \sum_{j=1}^g \Pi(C_j) k_x(\cdot, C_j). \]
where \( \sum_{j=1}^{d} \mathbb{I}(C_j) = 1 \). Let us put \( G_X = (k_X(X_i, X_j))_{i,j \in \mathcal{S}} \) and \( G_Y = (k_Y(Y_i, Y_j))_{i,j \in \mathcal{S}} \).

\[
D = (1_{(t)}(C_i(t)|X_i))_{i,j \in \mathcal{S}} \in \{0, 1\}^{\mathcal{S} \times \mathcal{S}}, \quad \hat{m}_{w} = (\hat{m}_{w}(X_i), \ldots, \hat{m}_{w}(X_n))^\top, \quad \hat{\mu} = (\hat{\mu}_1, \ldots, \hat{\mu}_n)^\top = (G_X + \epsilon I_\mathcal{S})^{-1} \hat{m}_w, \\
\Lambda = \text{diag}(\hat{\mu}), \quad k_X(\cdot) = (k_X(\cdot, X_1), \ldots, k_X(\cdot, X_n))^\top, \quad k_Y(\cdot) = (k_Y(\cdot, Y_1), \ldots, k_Y(\cdot, Y_n))^\top, \quad \text{and} \\
R_{XY} = \Lambda G_X ((G_X)^2 + \epsilon I_\mathcal{S})^{-1} \Lambda, \text{ where } I_\mathcal{S} \text{ is the identity matrix of size } n. \]

Here, we apply the three classifiers defined in Section 4.1 to a binary classification problem using computer-simulated data-sets, where \( \mathcal{X} \) and \( \mathcal{Y} \) are \( \mathbb{R}^2 \). In the first step, we independently generate 100 sets of training samples with each training sample being \((X_i, Y_i) \in \mathcal{X} \times \mathcal{Y}\) \(i = 1^{100}\), where \( X_i = C_1 \) and \( Y_i \sim \mathcal{N}(M_1, S_1) \) if \( 1 \leq i \leq 50 \), \( X_i = C_2 \) and \( Y_i \sim \mathcal{N}(M_2, S_2) \) if \( 51 \leq i \leq 100 \), \( M_1 = (1, 0)^\top \), \( M_2 = (0, 1)^\top \), and \( S_1 = S_2 = \text{diag}(0.1, 0.1) \). Here, \( \{Y_i\}_{i=1}^{50} \) and \( \{Y_i\}_{i=51}^{100} \) are sampled i.i.d. from \( \mathcal{N}(M_1, S_1) \) and \( \mathcal{N}(M_2, S_2) \), respectively. Individual \( Y \)-values of one of the training samples are plotted in Figure 1.

With each of the 100 training samples and a simulated prior probability of \( C_1 \) or \( \Pi(C_1) \in \{0.1, 0.2, \ldots, 0.9\} \), the classifiers defined in Section 4.1 estimate the posterior probability of \( C_1 \) given a test value \( y \in ((0.5, 0.5), (0.6, 0.4), (0.7, 0.3)) \), that is, \( \hat{Q}_{XY}(C_1) \). Figures 2–5 show the mean (plus or minus standard error of the mean, SEM) of the 100 values of \( \hat{Q}_{XY}(C_1) \) calculated by each of the classifiers, BR, KBR1, and KBR2. Here, we show the case where \( \epsilon \) in KBR1 and KBR2 is fixed to 0.1, and the regularization parameters of KBR1 are set to be \( \epsilon = \delta = 10^{-7} \) (Figure 2), \( \epsilon = \delta = 10^{-3} \) (Figure 3), \( \epsilon = \delta = 10^{-1} \) (Figure 4), and \( \epsilon = \delta = 10^{-1} \) (Figure 5). In Figures 2–5, BR_th represents the theoretical value of BR, which coincides with BR if the parameters \( \hat{M}_v, \hat{M}_2, \hat{S}_v, \) and \( \hat{S}_2 \) are set to be \( M_v, M_2, S_v, \) and \( S_2 \), respectively.

Consistent to Section 3.1, \( \hat{Q}_{XY}(C_1) \) calculated by KBR1 is poorly influenced by \( \Pi(C_1) \) compared with that by BR when \( \epsilon \) and \( \delta \) are set to be small (see Figures 2 and 3). In addition, \( \hat{Q}_{XY}(C_1) \) calculated by KBR2 also seems to be uninfluenced by \( \Pi(C_1) \). When \( \epsilon \) and \( \delta \) are set to be larger, the effect of \( \Pi(C_1) \) on \( \hat{Q}_{XY}(C_1) \) becomes apparent in KBR1, however, the value of \( \hat{Q}_{XY}(C_1) \) becomes too small.
These results suggest that in the kernel Bayes’ rule, the posterior does not depend on the prior if \( \epsilon \) and \( \delta \) are negligible, which might be a contradiction to the nature of Bayes’ theorem. Moreover, even though the prior affects the posterior when \( \epsilon \) and \( \delta \) become larger, the posterior seems too much dependent on \( \epsilon \) and \( \delta \), which are initially defined just for the regularization of matrices.

We have also tested all possible combinations of the following values for the parameters in KBR1 and/or KBR2: \( \epsilon \in \{10^{-1}, 10^{-3}, 10^{-5}, 10^{-7}, 10^{-11}, 10^{-13}, 10^{-15}\} \), \( \delta \in \{10^{-1}, 10^{-3}, 10^{-5}, 10^{-7}, 10^{-9}, 10^{-11}, 10^{-13}, 10^{-15}\} \), and \( \sigma \in \{0.01, 0.1, 1, 10, 100\} \). All the experimental results have been evaluated in a similar manner as above, and none of the results are found to be reasonable in the context of Bayesian inference (see Supplementary material).

5. Conclusions

One of the important features of Bayesian inference is that it provides a reasonable way of updating the probability for a hypothesis as additional evidence is acquired. The kernel Bayes’ rule has been expected to enable Bayesian inference in RKHS. In other words, the posterior kernel mean has been considered to be reasonably estimated by the kernel Bayes’ rule, given kernel mean expressions of the prior and likelihood. What is “reasonable” depends on circumstances, however, some of the results in this paper seem to show obviously unreasonable aspects of the kernel Bayes’ rule, at least in the context of Bayesian inference.

First, as shown in Section 3.1, when \( \Lambda \) and \( G_p \) are non-singular matrices and so we set \( \delta = 0 \), the posterior kernel mean \( \hat{m}_Q \) is entirely unaffected by the prior distribution \( \Pi \) on \( \mathcal{X} \). This means that, in Bayesian inference with the kernel Bayes’ rule, prior beliefs are in some cases completely neglected in calculating the kernel mean of the posterior distribution. Numerical evidence is also presented in Section 4.2. When the regularization parameters \( \epsilon \) and \( \delta \) are set to be small, the posterior probability calculated by the kernel Bayes’ rule (KBR1) is almost unaffected by the prior probability in comparison with that by conventional Bayes’ rule (BR). Consistently, when the regularized inverse matrices in KBR1 are replaced for the Moore-Penrose generalized inverse matrices (KBR2), the posterior probability is also uninfluenced by the prior probability, which seems to be unsuitable in the context of Bayesian updating of a probability distribution.
Figure 2. The case $\epsilon = \delta = 10^{-7}$.

Figure 3. The case $\epsilon = \delta = 10^{-5}$.
Figure 4. The case $\epsilon = \delta = 10^{-3}$.

Figure 5. The case $\epsilon = \delta = 10^{-1}$. 
Second, as discussed in Sections 3.2 and 4.2, the posterior estimated by the kernel Bayes' rule considerably depends upon the regularization parameters $\epsilon$ and $\delta$, which are originally introduced just for the regularization of matrices. A cross-validation approach is proposed in Fukumizu et al. (2013) to search for the optimal values of the parameters. However, theoretical foundations seem to be insufficient for the correct tuning of the parameters. Furthermore, in our experimental settings, we are not able to obtain a reasonable result using any combination of the parameter values, suggesting the possibility that there are no appropriate values for the parameters in general. Thus, we consider it difficult to solve the problem that $C_{xx}$ and $C_{ww}$ are not surjective by just adding regularization parameters.

Third, as shown in Section 3.3, the assumption that conditional expectation functions are included in the RKHS does not hold in general. Since this assumption is necessary for most of the theorems in Fukumizu et al. (2013), we believe that the assumption itself may need to be reconsidered.

In summary, even though current research efforts are focused on the application of the kernel Bayes' rule (Fukumizu et al., 2013; Kanagawa et al., 2014), it might be necessary to reexamine its basic framework of combining new evidence with prior beliefs.

6. Proofs

In this section, we provide some proofs for Sections 2 and 3.

6.1. Estimation of $C_{zw}$ and $C_{ww}$

Here we give the proof of Proposition 1.

**Proof** Let $\hat{C}_{xx}$, $\hat{C}_{yx,x}$, and $\hat{C}_{yy,x}$ denote the estimates of $C_{xx}$, $C_{yx,x}$, and $C_{yy,x}$, respectively. We define the estimates of $m_{zw}$ and $m_{ww}$ as

$$\hat{m}_{zw} = \hat{C}_{yx,x} \hat{C}_{xx}^{-1} \hat{m}_x,$$

and $\hat{m}_{ww} = \hat{C}_{yy,x} \hat{C}_{xx}^{-1} \hat{m}_x$, and put $h = \hat{C}_{xx}^{-1} \hat{m}_x \in H_x$. According to Equation (5), for any $f \in H_x$ and $g \in H_y$,

$$\langle \hat{m}_{zw}, g \otimes f \rangle_{H_x \otimes H_y} = \langle \hat{C}_{yx,x} h, g \otimes f \rangle_{H_x \otimes H_y} = \tilde{E} [f(X)g(Y)h(X)],$$

where $\tilde{E}[-]$ represents the empirical expectation operator. Thus, from Remark 1,

$$\hat{C}_{zw} = \hat{m}_{zw} = \frac{1}{n} \sum_{i=1}^{n} h(X_i) k_x(., X_i) \otimes k_y(., Y_i).$$

Similarly, for any $g_1, g_2 \in H_y$,

$$\langle \hat{m}_{ww}, g_1 \otimes g_2 \rangle_{H_x \otimes H_y} = \langle \hat{C}_{yy,x} h, g_1 \otimes g_2 \rangle_{H_x \otimes H_y} = \tilde{E} [g_1(Y)g_2(Y)h(X)],$$

where $\tilde{E}[-]$ represents the empirical expectation operator. Thus, from Remark 1,

$$\hat{C}_{ww} = \hat{m}_{ww} = \frac{1}{n} \sum_{i=1}^{n} h(X_i) k_y(., Y_i) \otimes k_y(., Y_i).$$

Next, we will derive $h(X_1), \ldots, h(X_n)$. Since $C_{xx}$ is a self-adjoint operator,
\[
\langle h, \hat{c}_{xx}f \rangle_{\mathcal{H}_x} = \langle \hat{c}_{xx}h, f \rangle_{\mathcal{H}_x} = \langle \hat{m}_n, f \rangle_{\mathcal{H}_x} = \sum_{j=1}^{l} \gamma_j f(U_j)
\]
for any \( f \in \mathcal{H}_x\). On the other hand, from Equation (1),

\[
\langle h, \hat{c}_{xx}f \rangle_{\mathcal{H}_x} = \hat{E}[f(X)h(X)] = \frac{1}{n} \sum_{i=1}^{n} f(X_i)h(X_i)
\]
for any \( f \in \mathcal{H}_x\). Hence, we have

\[
\sum_{j=1}^{l} \gamma_j f(U_j) = \frac{1}{n} \sum_{i=1}^{n} f(X_i)h(X_i)
\]
(10)

for any \( f \in \mathcal{H}_x\). Replacing \( f \) in Equation (10) for \( k_\chi(X_1, \cdot), \ldots, k_\chi(X_n, \cdot) \in \mathcal{H}_x \) we have

\[
\begin{pmatrix}
  k_\chi(X_1, U_1) & \cdots & k_\chi(X_1, U_l) \\
  \vdots & \ddots & \vdots \\
  k_\chi(X_n, U_1) & \cdots & k_\chi(X_n, U_l)
\end{pmatrix}
\begin{pmatrix}
  \gamma_1 \\
  \vdots \\
  \gamma_l
\end{pmatrix}
= \frac{1}{n} G_n
\begin{pmatrix}
  h(X_1) \\
  \vdots \\
  h(X_n)
\end{pmatrix}
\]
(11)

Using Equation (6), the left hand side of Equation (11) is given by

\[
\begin{pmatrix}
  \langle \sum_{j=1}^{l} \gamma_j k_\chi(\cdot, U_j), k_\chi(\cdot, X_1) \rangle_{\mathcal{H}_x} \\
  \vdots \\
  \langle \sum_{j=1}^{l} \gamma_j k_\chi(\cdot, U_j), k_\chi(\cdot, X_n) \rangle_{\mathcal{H}_x}
\end{pmatrix}
= \begin{pmatrix}
  \langle \hat{m}_n(X_1) \rangle \\
  \vdots \\
  \langle \hat{m}_n(X_n) \rangle
\end{pmatrix}
\]

Therefore, we have

\[
\frac{1}{n} \begin{pmatrix}
  h(X_1) \\
  \vdots \\
  h(X_n)
\end{pmatrix}
= G_n^{-1} \begin{pmatrix}
  \hat{m}_n(X_1) \\
  \vdots \\
  \hat{m}_n(X_n)
\end{pmatrix}
\approx (G_n + \rho I)^{-1} \hat{m}_n = \hat{\mu}.
\]

Replacing \( \frac{1}{n} (h(X_1), \ldots, h(X_n))^T \) for \( \hat{\mu} = (\hat{\mu}_1, \ldots, \hat{\mu}_n)^T \), Equations (8) and (9) become

\[
\hat{C}_{xx} = \sum_{i=1}^{n} \hat{\mu}_i k_\chi(\cdot, X_i) \otimes k_\chi(\cdot, Y_i) \text{ and } \hat{C}_{ww} = \sum_{i=1}^{n} \hat{\mu}_i k_\chi(\cdot, Y_i) \otimes k_\chi(\cdot, X_i)
\]
respectively. \( \square \)

### 6.2. Non-singularity of \( G_x \) and \( \Lambda \)

Here we show that the assumption in Section 3.1 holds under reasonable conditions.

**Definition 9** Let \( f \) be a real-valued function defined on a non-empty open domain \( \text{Dom}(f) \subseteq \mathbb{R}^d \). We say that \( f \) is analytic if \( f \) can be described by a Taylor expansion on a neighborhood of each point of \( \text{Dom}(f) \).

**Proposition 2** Let \( k \) be a positive definite kernel on \( \mathbb{R}^d \). Let \( \nu \) be a probability measure on \( \mathbb{R}^d \) which is absolutely continuous with respect to Lebesgue measure. Assume that \( k \) is an analytic function on \( \mathbb{R}^d \times \mathbb{R}^d \) and that the RKHS corresponding to \( k \) is infinite dimensional. Then for any i.i.d. random variables \( X_1, X_2, \ldots, X_n \) with the same distribution \( \nu \), the Gram matrix \( G_x = (k(X_i, X_j))_{1 \leq i, j \leq n} \) is non-singular almost surely with respect to \( \nu^n = \nu \times \nu \times \cdots \times \nu \) (n times).
Proof Let us put \( f(x_1, x_2, \ldots, x_n) := \det(k(x_i, x_j))_{1 \leq i, j \leq n} \). Since the RKHS corresponding to \( k \) is infinite dimensional, there are \( \xi_1, \xi_2, \ldots, \xi_n \in \mathbb{R}^d \) such that \((k(\cdot, \xi_i))_{1 \leq i \leq n}\) are linearly independent. Then \( f(\xi_1, \xi_2, \ldots, \xi_n) \neq 0 \) and hence \( f \) is a non-zero analytic function. Note that any non-trivial subvarieties of the euclidean spaces defined by analytic functions have Lebesgue measure zero. By this fact, the subvariety

\( \mathcal{V}(f) := \{ (x_1, x_2, \ldots, x_n) \in (\mathbb{R}^d)^n \mid f(x_1, x_2, \ldots, x_n) = 0 \} \subset (\mathbb{R}^d)^n \)

has Lebesgue measure zero. Since \( \nu \) is absolutely continuous, \( \nu(\mathcal{V}(f)) = 0 \). This completes the proof. \( \square \)

From Proposition 2, we easily obtain the following corollary.

Corollary 1 Let \( k \) be a Gaussian kernel on \( \mathbb{R}^d \) and let \( X_1, X_2, \ldots, X_n \) be i.i.d. random variables with the same normal distribution on \( \mathbb{R}^d \). Then the Gram matrix \( G_X = (k(X_i, X_j))_{1 \leq i, j \leq n} \) is non-singular almost surely.

Proposition 3 Let \( k \) be a positive definite kernel on \( \mathcal{X} = \mathbb{R}^d \), \( \nu \) a probability measure on \( \mathcal{X} \) which is absolutely continuous with respect to Lebesgue measure. Assume that \( k \) is an analytic function on \( \mathcal{X} \times \mathcal{X} \) and that the RKHS \( H \) corresponding to \( k \) is infinite dimensional. Then for any \((\xi, \gamma_1, \gamma_2, \ldots, \gamma_n, \nu) \in \mathbb{R}^d \times \mathcal{X}^n \times (\mathbb{R}^d)^n \) except Lebesgue measure zero, and for any i.i.d. random variables \( X_1, X_2, \ldots, X_n \) with the same distribution \( \nu \), each \( \mu_i \) for \( i = 1, 2, \ldots, n \) is (defined almost surely and) non-zero almost surely, where \((\mu_1, \mu_2, \ldots, \mu_n) = (G_X + n I_n)^{-1} \hat{\mathbf{m}}_n\).

For \((\nu_1, \nu_2, \ldots, \nu_n, \nu) \in S\), let us define a closed measure-zero set

\( \mathcal{V}_\nu := \{ (x_1, x_2, \ldots, x_n) \in \mathcal{X}^n \mid \det(G_X + n I_n) = 0 \} \). Then \( f_\nu(\nu_1, \nu_2, \ldots, \nu_n, \nu) \) is defined on \( \mathcal{X}^n \setminus \mathcal{V}_\nu \) for each \( i = 1, 2, \ldots, n \). Using Cramer’s rule,

\[
\mu_i = \frac{\det(n_{11}, n_{21}, \ldots, n_{ni}, \hat{\mathbf{m}}_{i1}, \hat{\mathbf{m}}_{i2}, \ldots, \hat{\mathbf{m}}_{in})}{\det(G_X + n I_n)},
\]

where \( n_{ij} \) stands for the \( m \)-th column vector of \( G_X + n I_n \). Here we denote by \( g \) the numerator of \( \mu_i \) that is, \( g_i = \mu_i \det(G_X + n I_n) \). It is easy to see that \( g_i(\xi_1, \xi_2, \ldots, \xi_n, \nu) \) is a non-zero analytic function of \( \nu \) on \( \mathcal{S} \). Indeed, if \( \nu \rightarrow \nu_i \), \( \nu_i = \nu \), and \( \nu_2 = \cdots = \nu_n = 0 \), then \( g_i \rightarrow \det((k(\cdot, \xi_i), k(\cdot, \xi_j))_{1 \leq i, j \leq n}) \). Hence \( \mathcal{S} = \{ \nu \mid g_i(\xi_1, \xi_2, \ldots, \xi_n, \nu) = 0 \} \) is a closed subset of \( \mathcal{S} \) with Lebesgue measure zero for each \( i = 1, 2, \ldots, n \). For any \((\nu_1, \nu_1, \ldots, \nu_n, \nu) \in \mathcal{S} \setminus (\cup_{i=1}^n \mathcal{S}_i)\),

\[
T_i(\nu_1, \nu_2, \ldots, \nu_n, \nu) := \{ \nu \in \mathcal{X}^n \mid g_i(\nu, \nu_1, \nu_2, \ldots, \nu_n) = 0 \}.
\]
is a closed subset of $X^n$ with Lebesgue measure zero for each $i \in \{1, 2, \ldots, n\}$, since $g_i(\star, e_1, \ldots, e_i, U_1, \ldots, U_i)$ is a non-zero analytic function of $\star$ on $X^n$. Therefore, $\mu_i = f_i(\star, e_1, \ldots, e_i, U_1, \ldots, U_i)$ is defined and non-zero for $i = 1, 2, \ldots, n$ and for $\star \in X^n \setminus \{P(e_1, e_1, \ldots, e_i, U_1, \ldots, U_i) \cup \{ (C_{\text{diagonal}}, e_1, \ldots, e_i, U_1, \ldots, U_i) \} \}$ if $(e_1, e_1, \ldots, e_i, U_1, \ldots, U_i) \in S \setminus \{ 0 \}$. This completes the proof. 

The following corollary directly follows from Proposition 3.

**Corollary 2.** Let $k$ be a Gaussian kernel on $\mathbb{R}^d$ and let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables with the same normal distribution on $\mathbb{R}^d$. All other notations are as in Proposition 3. Then $\Lambda = \text{diag}(\mu_1, \mu_2, \ldots, \mu_n)$ is non-singular almost surely for any $i, j, k, k', U_1, U_2, \ldots, U_i \in \mathbb{R}_+ \times \mathbb{R}^1 \times (\mathbb{R}^d)^i$ except for those in a set of Lebesgue measure zero.

### 6.3. Non-surjectivity of $C_{XX}$ and $C_{WW}$

The covariance operators $C_{XX}$ and $C_{WW}$ are non-surjective in general. This can be verified by the fact that they are compact operators. (If the operators are surjective on the corresponding RKHS which is infinite-dimensional, then they cannot be compact because of the open mapping theorem.) Here we present some easy examples where $C_{XX}$ and $C_{WW}$ are not surjective. Let us consider for simplicity the case $X = \mathbb{R}$. Let $X$ be a random variable on $\mathbb{R}$ with a normal distribution $\mathcal{N}(\mu, \sigma^2)$. We prove that $C_{XX}$ is not surjective under the usual assumption that the positive definite kernel on $\mathbb{R}$ is Gaussian. In order to demonstrate this, we use the symbols defined in Section 3.3 and several proven results on function spaces and Fourier transforms (see Rudin, 1987, for example). Note that the following three propositions are introduced without proofs.

**Proposition 4.** Let us put $f(x) = \exp(-ax^2 + bx + c)$ for $a, b, c \in \mathbb{R}$, where $a > 0$. Then

$$f(t) = \frac{1}{\sqrt{2\pi}} \exp \left( \frac{-t^2 - 2\sqrt{-1}bt - b^2 + 4ac}{4a} \right).$$

**Proposition 5.** For $f \in L^2(\mathbb{R}, C)$, $\hat{f}(t) = \tilde{f}(-t)$ almost everywhere. In particular, if $f \in L^2(\mathbb{R})$, then $\hat{f}(t) = \tilde{f}(-t)$ almost everywhere.

**Proposition 6.** For $f \in L^2(\mathbb{R}, C)$, put $f_0(x) = f(x - a)$. Then $\hat{f}_0(t) = \exp(-\sqrt{-1}at)\hat{f}(t)$.

**Definition 10.** Let $p(\cdot)$ denote the density function of the normal distribution $\mathcal{N}(\mu, \sigma^2)$ on $\mathbb{R}$, that is,

$$p(\cdot) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left( \frac{- (\cdot - \mu)^2}{2\sigma^2} \right).$$

Let $X$ be a random variable on $\mathbb{R}$ with $\mathcal{N}(\mu, \sigma^2)$. The linear operator $C_{XX} : \mathcal{H}_0 \to \mathcal{H}_0$ is defined by $\langle C_{XX} f, g \rangle_{\mathcal{H}_0} = E[ f(X) g(X) ]$ for any $f, g \in \mathcal{H}_0$, which is also described as

$$\langle C_{XX} f, g \rangle(\cdot) = \int_{-\infty}^{\infty} f(x) k(\cdot, x) p(x) \, dx$$

for any $f \in \mathcal{H}_0$. 

**Proposition 7.** If $f, g \in \mathcal{H}_0$, then $\langle f, g \rangle_{\mathcal{H}_0} \in \mathbb{R}$

**Proof.** From Proposition 5, $\hat{f}(t) = \hat{f}(-t)$ and $\hat{g}(t) = \hat{g}(-t)$ for any $f, g \in \mathcal{H}_0$. Then, using Equation (7), we have
\[
\langle f, g \rangle_{\mathcal{H}} = \int_{-\infty}^{\infty} f(t) \bar{g}(t) \exp \left(\frac{\sigma^2}{2} \right) dt = \int_{-\infty}^{\infty} f(-t) \bar{g}(t) \exp \left(\frac{\sigma^2}{2} \right) dt
\]

This completes the proof. □

**Proposition 8** If \( f \in \mathcal{H}_f(\mathbb{R}, C) \), then \( \bar{f} \in \mathcal{H}_f(\mathbb{R}, C) \).

**Proof** From Proposition 5, \( \bar{f}(t) = \bar{f}(-t) \) for \( f \in L^2(\mathbb{R}, C) \). Then, using Equation (7), we have

\[
\| \bar{f} \|_{\mathcal{H}_f(\mathbb{R}, C)}^2 = \langle \bar{f}, \bar{f} \rangle_{\mathcal{H}_f(\mathbb{R}, C)} = \int_{-\infty}^{\infty} |\bar{f}(t)|^2 \exp \left(\frac{\sigma^2}{2} \right) dt = \int_{-\infty}^{\infty} |\bar{f}(-t)|^2 \exp \left(\frac{\sigma^2}{2} \right) dt
\]

Therefore, \( \bar{f} \in \mathcal{H}_f(\mathbb{R}, C) \). □

Here, we denote by \( \text{Re} \) and \( \text{Im} \) the real part and the imaginary part of a complex number, respectively. We also denote by \( \text{Cl}(A) \) the closure of a subset \( A \) in a topological space.

**Corollary 3** If \( f \in \mathcal{H}_f(\mathbb{R}, C) \) then \( \text{Re}(f), \text{Im}(f) \in \mathcal{H}_f \).

**Proof** If \( f \in \mathcal{H}_f(\mathbb{R}, C) \), then \( \bar{f} \in \mathcal{H}_f(\mathbb{R}, C) \) by Proposition 8. Hence we see that

\[
\text{Re}(f) = \frac{f + \bar{f}}{2} \in \mathcal{H}_f, \quad \text{Im}(f) = \frac{f - \bar{f}}{2\sqrt{-1}} \in \mathcal{H}_f.
\]

This completes the proof. □

**Remark 3** If \( f \in \mathcal{H}_f(\mathbb{R}, C) \), then there uniquely exist \( f_1, f_2 \in \mathcal{H}_f \) such that \( f = f_1 + \sqrt{-1}f_2 \) by Corollary 3. This means that \( \mathcal{H}_f(\mathbb{R}, C) = \mathcal{H}_f \oplus \sqrt{-1}\mathcal{H}_f \) where \( \oplus \) denotes the direct sum.

**Proposition 9** For any \( f \in L^2(\mathbb{R}, C) \) and for any \( \varepsilon > 0 \), there exists \( g \in \mathcal{H}_f(\mathbb{R}, C) \) such that \( \| f - g \|_2 < \varepsilon \).

In other words, \( \mathcal{H}_f(\mathbb{R}, C) \) is dense in \( L^2(\mathbb{R}, C) \).

**Proof** Let \( C_0(\mathbb{R}, C) \) denote the space of continuous complex-valued functions with compact support on \( \mathbb{R} \). Let us define \( \tilde{\mathcal{H}}_f(\mathbb{R}, C) \) by

\[
\tilde{\mathcal{H}}_f(\mathbb{R}, C) = \left\{ h \in L^2(\mathbb{R}, C) \mid \int_{-\infty}^{\infty} |h(t)|^2 \exp \left(\frac{\sigma^2}{2} \right) dt < \infty \right\}.
\]

Note that \( \tilde{\mathcal{H}}_f(\mathbb{R}, C) \) coincides with the image of \( \mathcal{H}_f(\mathbb{R}, C) \) by the Fourier transform. Then, \( C_0(\mathbb{R}, C) \subset \tilde{\mathcal{H}}_f(\mathbb{R}, C) \subset L^2(\mathbb{R}, C) \) and \( \text{Cl}(C_0(\mathbb{R}, C)) = L^2(\mathbb{R}, C) \). Hence \( \text{Cl}(\tilde{\mathcal{H}}_f(\mathbb{R}, C)) = L^2(\mathbb{R}, C) \). In other words, for any \( f \in L^2(\mathbb{R}, C) \) and for any \( \varepsilon > 0 \), there exists \( \tilde{g} \in \tilde{\mathcal{H}}_f(\mathbb{R}, C) \) such that \( \| f - \tilde{g} \|_2 < \varepsilon \) because \( \tilde{f} \in L^2(\mathbb{R}, C) \), which implies that there exists \( g \in \mathcal{H}_f(\mathbb{R}, C) \) such that \( \| f - g \|_2 < \varepsilon \). This completes the proof. □

The following corollary has also been shown in Theorem 4.63 in Steinwart and Christmann (2008).
Corollary 4  \( \text{Cl}(H_0) = L^2(\mathbb{R}). \)

Proof From Proposition 9, for any \( f \in L^1(\mathbb{R}) \subset L^1(\mathbb{R}, \mathbb{C}) \) and for any \( \varepsilon > 0 \), there exists \( g \in H_0(\mathbb{R}, \mathbb{C}) \) such that \( \| f - g \|_2 < \varepsilon \). By Remark 3, there exist \( g_1, g_2 \in H_0 \) such that \( g = g_1 + \sqrt{-1} g_2 \). Thus,

\[
\varepsilon^2 > \| f - g \|_2^2 = \int_{-\infty}^{\infty} | f - g |^2 \, dx = \int_{-\infty}^{\infty} | f - g_1 - \sqrt{-1} g_2 |^2 \, dx
\]

\[
\geq \int_{-\infty}^{\infty} | f - g_1 |^2 \, dx = \| f - g_1 \|_2^2.
\]

Therefore, \( \| f - g_1 \|_2 < \varepsilon \). This completes the proof.  \( \square \)

Definition 11  Let us define \( r, r_n \in L^2(\mathbb{R}) \) as

\[
r(t) = \frac{1}{|t|} 1_{(1, \infty)}(|t|), \quad r_n(t) = \frac{1}{|t|} 1_{(1, n)}(|t|),
\]

where \( 1_{(1, \infty)} \) and \( 1_{(1, n)} \) denote the indicator functions of the intervals \((1, \infty)\) and \((1, n)\), respectively. We also put \( h_n = r_n \) and \( h = r \). Note that \( \lim_{n \to \infty} r_n = r \in L^2(\mathbb{R}) \), because

\[
\lim_{n \to \infty} \| r_n - r \|_2^2 = 2 \lim_{n \to \infty} \int_{|t| > 1} \frac{1}{k^2} \, dx = 0.
\]

Proposition 10  \( h_n, h \in L^2(\mathbb{R}). \)

Proof It is obvious that \( h_n, h \in L^2(\mathbb{R}, \mathbb{C}). \) Since \( r_n \in L^1(\mathbb{R}) \cap L^2(\mathbb{R}) \), we see that

\[
\overline{h_n}(x) = \overline{r_n}(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} r_n(t) \exp \left( -\frac{1}{2} t x \right) \, dt
\]

\[
= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} r_n(t) \exp \left( -\frac{1}{2} t (x - n) \right) \, dt = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} r_n(t') \exp \left( -\frac{1}{2} t' x \right) \, dt'
\]

\[
= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} r_n(t') \exp \left( -\frac{1}{2} t' x \right) \, dt' = \overline{r_n}(x) = h_n(x),
\]

where \( t' = -t \). Hence, \( h_n \in L^2(\mathbb{R}). \) On the other hand,

\[
h(x) = \lim_{n \to \infty} \frac{1}{\sqrt{2\pi}} \int_{-n}^{n} r(t) \exp \left( -\frac{1}{2} t x \right) \, dt
\]

\[
= \lim_{n \to \infty} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} r_n(t) \exp \left( -\frac{1}{2} t x \right) \, dt = \lim_{n \to \infty} h_n(x).
\]

Therefore \( h \in L^2(\mathbb{R}). \)  \( \square \)

Let us define \( k_\sigma(\cdot) = \sqrt{2\pi} \sigma k_\sigma(\cdot), \) \( a = \exp \left( -\frac{\cdot - \sigma^2}{2\sigma^2} \right) \) \( \in H_0 \) for \( a \in \mathbb{R}. \) Now, we prove that \( k_\sigma \notin \text{Ran}(C_{xa}) \) for any \( a \in \mathbb{R}. \) This implies that \( C_{xa} \) is not surjective.

Proposition 11  For any \( a \in \mathbb{R}, \) \( k_\sigma \in H_0 \setminus \text{Ran}(C_{xa}). \)

Proof Suppose that there exists \( g \in H_0 \) such that \( C_{xa} g = k_\sigma \). Then, for any \( f \in H_0, \)

\[
\langle k_\sigma, f \rangle_{H_0} = \langle C_{xa} g, f \rangle_{H_0}.
\]

Let us put \( k(\cdot) = \sqrt{2\pi} \sigma k_\sigma(\cdot, 0) = \exp \left( -\frac{\cdot - \sigma^2 t^2}{2} \right) \). From Proposition 4, \( \tilde{k}(t) = \sigma \exp \left( -\frac{\sigma^2 t^2}{2} \right) \). Then, using Equation (7) and Proposition 6, the left hand side of Equation (12) equals
\[ \int_{-\infty}^{\infty} k_\sigma(t) \bar{f}(t) \exp \left( \frac{\sigma^2}{2} t^2 \right) dt = \int_{-\infty}^{\infty} \exp \left( -\sqrt{-1}at \right) k_\sigma(t) \bar{f}(t) \exp \left( \frac{\sigma^2}{2} t^2 \right) dt \\
= \sigma \int_{-\infty}^{\infty} \exp \left( -\sqrt{-1}at \right) \bar{f}(t) dt. \]

The right hand side of Equation (12) is equal to
\[
E[gX \bar{f}(X)] = \int_{-\infty}^{\infty} g(x)f(x)p(x) dx = \langle gp, f \rangle_{L^2(\mathbb{R})}. 
\]

Thus, Equation (12) is equivalent to the following equation:
\[ \langle gp, f \rangle_{L^2(\mathbb{R})} = \sigma \int_{-\infty}^{\infty} \exp \left( -\sqrt{-1}at \right) \bar{f}(t) dt. \] (13)

Let us define \( h_{n,a}(x) := h_n(x-a) \) and \( h_a(x) = h(x-a) \). Then \( h_{n,a}, h_a \in L^2(\mathbb{R}) \). It is easy to see that \( \|h_{n,a} - h_a\|_2 = \|h_n - h\|_2 = \|n - r\|_2 \to 0 \) as \( n \to \infty \). Hence \( \lim_{n \to \infty} h_{n,a} = h_a \) in \( L^2(\mathbb{R}) \). Since \( \hat{h}_{n,a}(t) = \exp(-\sqrt{-1}at)h_n(t) \) by Proposition 6, we have
\[
\int_{-\infty}^{\infty} |\hat{h}_{n,a}(t)|^2 \exp \left( \frac{\sigma^2}{2} t^2 \right) dt = \int_{-\infty}^{\infty} |\hat{h}_a(t)|^2 \exp \left( \frac{\sigma^2}{2} t^2 \right) dt = \int_{-\infty}^{\infty} |r_n(t)|^2 \exp \left( \frac{\sigma^2}{2} t^2 \right) dt \\
= 2 \int_{1}^{\infty} \frac{1}{t} \exp \left( \frac{\sigma^2}{2} t^2 \right) dt < \infty, 
\]

which indicates that \( h_{n,a} \in \mathcal{H}_0 \). Substituting \( h_{n,a} \) for \( f \), Equation (13) becomes
\[ \langle gp, h_{n,a} \rangle_{L^2(\mathbb{R})} = \sigma \int_{-\infty}^{\infty} \exp \left( -\sqrt{-1}at \right) \hat{h}_{n,a}(t) dt. \] (14)

If \( n \) goes to infinity, the left hand side of Equation (14) becomes \( \langle gp, h_a \rangle_{L^2(\mathbb{R})} \in \mathbb{R} \). On the other hand, the right hand side of Equation (14) becomes
\[
\sigma \int_{-\infty}^{\infty} \exp \left( -\sqrt{-1}at \right) \exp(-\sqrt{-1}at)\hat{h}_a(t) dt = \sigma \int_{-\infty}^{\infty} \hat{h}_a(t) dt \\
= \sigma \int_{-\infty}^{\infty} r_n(t) dt \\
= 2\sigma \int_{1}^{\infty} \frac{1}{t} dt \to \infty \quad (n \to \infty). 
\]

This is a contradiction. Therefore, there exists no \( g \in \mathcal{H}_d \) such that \( C_{ax}g = k_a \). This completes the proof.

Supplementary material
Supplementary material for this article can be accessed https://doi.org/10.1080/23311835.2018.1447220.

Funding
Kazunori Nakamoto was partially supported by JSPS KAKENHI [grant numbers JP23540044, JP15K046814].

Author details
Hisashi Johno1
E-mail: hisashi.johno@gmail.com
Kazunori Nakamoto1
E-mail: nakamoto@yamanashi.ac.jp
Tatsuhiko Saigo2
E-mail: tsaiqgh@yamanashi.ac.jp

1 Faculty of Medicine, Department of Mathematical Sciences, University of Yamanashi, Shimokata 1110, Chuo, Yamanashi 409-3898, Japan.
2 Center for Medical Education and Sciences, Faculty of Medicine, University of Yamanashi, Shimokata 1110, Chuo, Yamanashi, 409-3898, Japan.

Citation information
Cite this article as: Remarks on kernel Bayes’ rule, Hisashi Johno, Kazunori Nakamoto & Tatsuhiko Saigo, Cogent Mathematics & Statistics (2018), 5: 1447220.

References
Fukumizu, K. (2014). Introduction to kernel methods (in Japanese). Tokyo: Asakura Shoten.
Fukumizu, K., Song, L., & Gretton, A. (2013). Kernel bayes’ rule: Bayesian inference with positive definite kernels. Journal of Machine Learning Research, 14, 3753–3783.
Horn, R. A., & Johnson, C. R. (2013). Matrix analysis (2nd ed.). Cambridge: Cambridge University Press.
Kanagawa, M., Nishiyama, Y., Gretton, A., & Fukumizu, K. (2014). Monte carlo filtering using kernel embedding of
distributions. "Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence" (p. 1897-1903).
Rudin, W. (1987). "Real and complex analysis" (3rd ed.). New York, NY: McGraw-Hill Book Co.
Song, L., Fukumizu, K., & Gretton, A. (2014). Kernel embeddings of conditional distributions. "IEEE Signal Processing Magazine", 30, 98–111.

Song, L., Huang, J., Smola, A., & Fukumizu, K. (2009). Hilbert space embeddings of conditional distributions with applications to dynamical systems. In "Proceedings of the 26th Annual International Conference on Machine Learning" (p. 961-968).
Steinwart, I., & Christmann, A. (2008). "Support vector machines". New York, NY: Springer.

© 2018 The Author(s). This open access article is distributed under a Creative Commons Attribution (CC-BY) 4.0 license.
You are free to:
Share — copy and redistribute the material in any medium or format
Adapt — remix, transform, and build upon the material for any purpose, even commercially.
The licensor cannot revoke these freedoms as long as you follow the license terms.
Under the following terms:
Attribution — You must give appropriate credit, provide a link to the license, and indicate if changes were made.
You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.
No additional restrictions
You may not apply legal terms or technological measures that legally restrict others from doing anything the license permits.

Cogent Mathematics & Statistics (ISSN: 2574-2558) is published by Cogent OA, part of Taylor & Francis Group.
Publishing with Cogent OA ensures:
• Immediate, universal access to your article on publication
• High visibility and discoverability via the Cogent OA website as well as Taylor & Francis Online
• Download and citation statistics for your article
• Rapid online publication
• Input from, and dialog with, expert editors and editorial boards
• Retention of full copyright of your article
• Guaranteed legacy preservation of your article
• Discounts and waivers for authors in developing regions
Submit your manuscript to a Cogent OA journal at www.CogentOA.com