Remarks on kernel Bayes' rule

Hisashi Johno[1], Kazunori Nakamoto[2], and Tatsuhiko Saigo[3]

[1]Department of Anatomy and Cell Biology, Faculty of Medicine, University of Yamanashi
Shimokato 1110, Chuo, Yamanashi 409-3898, Japan.   `hisashi.johno@gmail.com`

[2]Center for Medical Education and Sciences, Faculty of Medicine, University of Yamanashi
Shimokato 1110, Chuo, Yamanashi 409-3898, Japan.   `nakamoto@yamanashi.ac.jp`

[3]Center for Medical Education and Sciences, Faculty of Medicine, University of Yamanashi
Shimokato 1110, Chuo, Yamanashi 409-3898, Japan.   `tsaigoh@yamanashi.ac.jp`

Abstract

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 prediction result by
kernel Bayes' rule is in some cases unnatural. We consider that this phenomenon
is in part due to the fact that the assumptions in kernel Bayes' rule do not hold in
general.

Keywords: Kernel method, Bayes' rule, reproducing kernel Hilbert space

§1   Introduction

Kernel Bayes' rule has recently emerged as a novel framework for Bayesian infer-
ence [1–3]. It is generally agreed that, in this framework, we can estimate the ker-
nel mean of the posterior distribution, given kernel mean expressions of the prior and
likelihood distributions. Since the distributions are mapped and nonparametrically ma-
nipulated in infinite-dimensional feature spaces called reproducing kernel hilbert spaces
(RKHS), it is believed that kernel Bayes' rule can accurately evaluate the statistical fea-
tures of high-dimensional data and enable Bayesian inference even if there were no
appropriate parametric models. To date, several applications of kernel Bayes' rule have
been reported [2, 4]. However, the basic theory and the algorithm of kernel Bayes' rule
might need to be modified because of the following reasons:

1. The posterior in kernel Bayes' rule is in some cases completely unaffected by the
   prior.

2. The posterior in 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 kernel Bayes' rule.

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

§2 Kernel Bayes' rule

In this section, we briefly review kernel Bayes' rule following [2]. Let $\mathcal{X}$ and $\mathcal{Y}$ be measurable spaces, (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 Π 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 Π and the family $\{P_{\mathcal{Y}|x} \mid x \in \mathcal{X}\}$, where $P_{\mathcal{Y}|x}$ denotes the conditional distribution of Y given $X = x$. For each $y \in \mathcal{Y}$, let $Q_{\mathcal{X}|y}$ represent the posterior distribution of Z given $W = y$. The aim of kernel Bayes' rule is to derive the kernel mean of $Q_{\mathcal{X}|y}$.

Definition 2.1. Let $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$ be measurable positive definite kernels on $\mathcal{X}$ and $\mathcal{Y}$ such that $E[k_{\mathcal{X}}(X, X)] < \infty$ and $E[k_{\mathcal{Y}}(Y, Y)] < \infty$, respectively, where $E[\cdot]$ denotes the expectation operator. Let $\mathcal{H}_{\mathcal{X}}$ and $\mathcal{H}_{\mathcal{Y}}$ be the RKHS defined by $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$, respectively. We consider two bounded linear operators $C_{YX} : \mathcal{H}_{\mathcal{X}} \to \mathcal{H}_{\mathcal{Y}}$ and $C_{XX} : \mathcal{H}_{\mathcal{X}} \to \mathcal{H}_{\mathcal{X}}$ such that

$$\langle g, C_{YX}f \rangle_{\mathcal{H}_{\mathcal{Y}}} = E\left[f(X)g(Y)\right] \quad \text{and} \quad \langle f_1, C_{XX}f_2 \rangle_{\mathcal{H}_{\mathcal{X}}} = E\left[f_1(X)f_2(X)\right] \quad (1)$$

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

$$(C_{YX}f)(\cdot) = \int_{\mathcal{X} \times \mathcal{Y}} k_{\mathcal{Y}}(\cdot, y)f(x)\, dP(x, y) \quad \text{and} \quad (C_{XX}f)(\cdot) = \int_{\mathcal{X}} k_{\mathcal{X}}(\cdot, x)f(x)\, dP_X(x),$$

where P_X denotes the marginal distribution of X. Let C_{XY} be the bounded linear operator defined by

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

for any $f \in \mathcal{H}_{\mathcal{X}}$ and $g \in \mathcal{H}_{\mathcal{Y}}$. Then C_{XY} is the adjoint of C_{YX}.

Theorem 2.2. *([2], Theorem 1) If $E[g(Y)|X = \cdot] \in \mathcal{H}_{\mathcal{X}}$ for $g \in \mathcal{H}_{\mathcal{Y}}$, then $C_{XX}E[g(Y)|X = \cdot] = C_{XY}g$.*

2

Definition 2.3. Let $Q_{\mathcal{Y}}$ denote the marginal distribution of W. Assuming that $E[k_\mathcal{X}(U, U)] < \infty$ and $E[k_\mathcal{Y}(W, W)] < \infty$, we can define the kernel means of Π and $Q_{\mathcal{Y}}$ by

$$m_\Pi = E[k_\mathcal{X}(\cdot, U)] \quad \text{and} \quad m_{Q_{\mathcal{Y}}} = E[k_\mathcal{Y}(\cdot, W)],$$

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

Theorem 2.4. *([2], Theorem 2) If C_{XX} is injective, $m_\Pi \in \mathrm{Ran}(C_{XX})$, and $E[g(Y) \,|\, X = \cdot] \in \mathcal{H}_\mathcal{X}$ for any $g \in \mathcal{H}_\mathcal{Y}$, then*

$$m_{Q_{\mathcal{Y}}} = C_{YX} C_{XX}^{-1} m_\Pi, \tag{2}$$

where $\mathrm{Ran}(C_{XX})$ denotes the range of C_{XX}.

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

$$E\left[k_\mathcal{Y}(\cdot, Y) \,|\, X = x\right] = C_{YX} C_{XX}^{-1} k_\mathcal{X}(\cdot, x) \tag{3}$$

by replacing m_Π in Equation (2) for $k_\mathcal{X}(\cdot, x)$. It is noted in [2] that the assumption $m_\Pi \in \mathrm{Ran}(C_{XX})$ 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 ϵ is a regularization constant and I is the identity operator. Thus, the approximations of Equations (2) and (3) are respectively given by

$$m_{Q_{\mathcal{Y}}}^{reg} = C_{YX} (C_{XX} + \epsilon I)^{-1} m_\Pi \quad \text{and} \quad E^{reg}\left[k_\mathcal{Y}(\cdot, Y) \,|\, X = x\right] = C_{YX} (C_{XX} + \epsilon I)^{-1} k_\mathcal{X}(\cdot, x).$$

Similarly, for any $y \in \mathcal{Y}$, the approximation of $m_{Q_{\mathcal{X}|y}}$ is provided by

$$m_{Q_{\mathcal{X}|y}}^{reg} = E^{reg}\left[k_\mathcal{X}(\cdot, Z) \,|\, W = y\right] = C_{ZW} (C_{WW} + \delta I)^{-1} k_\mathcal{Y}(\cdot, y), \tag{4}$$

where δ is a regularization constant and the linear operators C_{ZW} and C_{WW} will be defined below.

Definition 2.5. We consider the kernel means $m_Q = m_{(ZW)}$ and $m_{(WW)}$ such that

$$\langle m_{(ZW)}, g \otimes f \rangle_{\mathcal{H}_\mathcal{Y} \otimes \mathcal{H}_\mathcal{X}} = E\left[f(Z) g(W)\right] \quad \text{and} \quad \langle m_{(WW)}, g_1 \otimes g_2 \rangle_{\mathcal{H}_\mathcal{Y} \otimes \mathcal{H}_\mathcal{Y}} = E\left[g_1(W) g_2(W)\right]$$

for any $f \in \mathcal{H}_\mathcal{X}$ and $g, g_1, g_2 \in \mathcal{H}_\mathcal{Y}$, where $\otimes$ denotes the tensor product. Let $C_{(YX)X} : \mathcal{H}_\mathcal{X} \to \mathcal{H}_\mathcal{Y} \otimes \mathcal{H}_\mathcal{X}$ and $C_{(YY)X} : \mathcal{H}_\mathcal{X} \to \mathcal{H}_\mathcal{Y} \otimes \mathcal{H}_\mathcal{Y}$ be bounded linear operators which respectively satisfy

$$
\begin{aligned}
\langle g \otimes f, C_{(YX)X} h \rangle_{\mathcal{H}_\mathcal{Y} \otimes \mathcal{H}_\mathcal{X}} &= E\left[g(Y) f(X) h(X)\right], \\
\langle g_1 \otimes g_2, C_{(YY)X} f \rangle_{\mathcal{H}_\mathcal{Y} \otimes \mathcal{H}_\mathcal{Y}} &= E\left[g_1(Y) g_2(Y) f(X)\right]
\end{aligned}
\tag{5}
$$

for any $f, h \in \mathcal{H}_\mathcal{X}$ and $g, g_1, g_2 \in \mathcal{H}_\mathcal{Y}$.

From Theorem 2.4, Fukumizu et al. [2] proposed that $m_{(ZW)}$ and $m_{(WW)}$ can be given by

$$m_{(ZW)} = C_{(YX)X}C_{XX}^{-1}m_\Pi \in \mathcal{H}_\mathcal{Y} \otimes \mathcal{H}_\mathcal{X} \quad \text{and} \quad m_{(WW)} = C_{(YY)X}C_{XX}^{-1}m_\Pi \in \mathcal{H}_\mathcal{Y} \otimes \mathcal{H}_\mathcal{Y}.$$

In case m_Π is not included in $\mathrm{Ran}(C_{XX})$, they suggested that $m_{(ZW)}$ and $m_{(WW)}$ could be approximated by

$$m_{(ZW)}^{reg} = C_{(YX)X}\,(C_{XX} + \epsilon I)^{-1}\,m_\Pi \quad \text{and} \quad m_{(WW)}^{reg} = C_{(YY)X}\,(C_{XX} + \epsilon I)^{-1}\,m_\Pi.$$

Remark 2.6. ([2], page 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 $m_{Q_{X|y}}$ following [2].

Definition 2.7. 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 $\{U_j\}_{j=1}^{l}$ from the prior distribution Π on $\mathcal{X}$. The prior kernel mean m_Π is estimated by

$$\widehat{m}_\Pi = \sum_{j=1}^{l} \gamma_j k_\mathcal{X}(\cdot, U_j), \tag{6}$$

where $\gamma_1, \ldots, \gamma_l$ are weights. Let us put $\widehat{\boldsymbol{m}}_\Pi = (\widehat{m}_\Pi(X_1), \ldots, \widehat{m}_\Pi(X_n))^\mathrm{T}$, $G_X = (k_\mathcal{X}(X_i, X_j))_{1 \le i, j \le n}$, and $G_Y = (k_\mathcal{Y}(Y_i, Y_j))_{1 \le i, j \le n}$.

Proposition 2.8. ([2], Proposition 3, revised) Let I_n denote the identity matrix of size n. The estimates of C_{ZW} and C_{WW} are given by

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

respectively, where $\widehat{\boldsymbol{\mu}} = (\widehat{\mu}_1, \ldots, \widehat{\mu}_n)^\mathrm{T} = (G_X + n\epsilon I_n)^{-1}\widehat{\boldsymbol{m}}_\Pi$.

The proof of this revised proposition is given in §6.1. It is suggested in [2] that Equation (4) can be empirically estimated by

$$\widehat{m}_{Q_{X|y}} = \widehat{C}_{ZW}\left(\widehat{C}_{WW}^2 + \delta I_n\right)^{-1}\widehat{C}_{WW}k_\mathcal{Y}(\cdot, y).$$

Theorem 2.9. ([2], Proposition 4) Given an observation $y \in \mathcal{Y}$, $\widehat{m}_{Q_{X|y}}$ can be calculated by

$$\widehat{m}_{Q_{X|y}} = \boldsymbol{k}_X^\mathrm{T} R_{X|Y}\boldsymbol{k}_Y(y), \quad R_{X|Y} = \Lambda G_Y\left((\Lambda G_Y)^2 + \delta I\right)^{-1}\Lambda,$$

where $\Lambda = \mathrm{diag}(\widehat{\boldsymbol{\mu}})$ is the diagonal matrix with the elements of $\widehat{\boldsymbol{\mu}}$, $\boldsymbol{k}_X = (k_\mathcal{X}(\cdot, X_1), \ldots, k_\mathcal{X}(\cdot, X_n))^\mathrm{T}$, and $\boldsymbol{k}_Y = (k_\mathcal{Y}(\cdot, Y_1), \ldots, k_\mathcal{Y}(\cdot, Y_n))^\mathrm{T}$.

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

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

where $\boldsymbol{f}_X = (f(X_1), \ldots, f(X_n))^\mathrm{T}$.

4

§3 Theoretical arguments

In this section, we theoretically support the three arguments raised in §1. First, we show in §3.1 that the posterior kernel mean $\widehat{m}_{Q_{X|y}}$ is completely unaffected by the prior distribution Π under the condition that Λ and G_Y are non-singular. This implies that, at least in some cases, Π does not properly affect $\widehat{m}_{Q_{X|y}}$. Second, we mention in §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 ϵ and δ. Third, we demonstrate in §3.3 that conditional expectation functions are not generally contained in the RKHS, which means that Theorems 1, 2, 5, 6, 7, and 8 in [2] do not work in some situations.

§3.1 Relations between the posterior $\widehat{m}_{Q_{X|y}}$ and the prior Π

Let us review Theorem 2.9. Assume that G_Y and Λ are non-singular matrices. (This assumption is not so strange, as shown in §6.2.) The matrix $R_{X|Y} = \Lambda G_Y((\Lambda G_Y)^2 + \delta I)^{-1}\Lambda$ tends to G_Y^{-1} as δ tends to 0. Furthermore, if we set $\delta = 0$ from the beginning, we obtain $R_{X|Y} = G_Y^{-1}$. This implies that the posterior kernel mean $\widehat{m}_{Q_{X|y}} = k_X^{\mathrm{T}} R_{X|Y} k_Y(y)$ never depends on the prior distribution Π on $\mathcal{X}$, which seems to be a contradiction to the nature of Bayes' rule. This result is numerically confirmed in §4.2.

§3.2 The inverse of the operators C_{XX} and C_{WW}

As noted by Fukumizu et al. [2], the linear operators C_{XX} and C_{WW} are not surjective in some usual cases, the proof of which is given in §6.3. Therefore, they proposed an alternative way of obtaining a solution $f \in \mathcal{H}_X$ of the equation $C_{XX}f = m_{\Pi}$, that is, a regularized inversion $f = (C_{XX} + \epsilon I)^{-1}m_{\Pi}$ as an analog of ridge regression, where ϵ is a regularization parameter and I is an identity operator. One of the disadvantages of this method is that the solution $f = (C_{XX} + \epsilon I)^{-1}m_{\Pi}$ depends upon the choice of ϵ. In §4.2, we numerically show that the prediction using kernel Bayes' rule considerably depends on the regularization parameters ϵ and δ. Theorems 5, 6, 7, and 8 in [2] 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 several cases as proved in §3.3. Furthermore, since we need to assume sufficiently slow decay of the regularization constants ϵ and δ in these theorems, it is practically difficult to set appropriate values for ϵ and δ. 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 ΛG_Y, given a sample $\{(X_i, Y_i)\}_{i=1}^{n}$. Below, we briefly introduce a generalization of a matrix inverse. For more details, see [6].

Definition 3.1. Let A be a matrix of size $m \times n$ over the complex number space $\mathbb{C}$. We say that a matrix $A^{\times}$ of size $n \times m$ is a generalized inverse matrix of A if $AA^{\times}A = A$. We also say that a matrix $A^{\dagger}$ of size $n \times m$ is the Moore-Penrose generalized inverse matrix of A if $AA^{\dagger}$ and $A^{\dagger}A$ are Hermitian, $AA^{\dagger}A = A$, and $A^{\dagger}AA^{\dagger} = A^{\dagger}$.

Remark 3.2. In fact, any matrix A has the Moore-Penrose generalized inverse matrix $A^\dagger$. Note that $A^\dagger$ is uniquely determined by A. If A is square and non-singular, then $A^\times = A^\dagger = A^{-1}$. For a generalized inverse matrix $A^\times$ of size $n \times m$, $AA^\times v = v$ for any vector $v \in \mathbb{C}^m$ if v is contained in the image of A. In particular, $A^\times v$ is a vector contained in the preimage of v under A.

In the calculation of $\widehat{m}_{Q_{X|y}} = k_X^{\mathrm{T}} R_{X|Y} k_Y(y)$, we numerically compare the case $R_{X|Y} = (\Lambda' G_Y)^\dagger \Lambda'$ with the original case $R_{X|Y} = \Lambda G_Y((\Lambda G_Y)^2 + \delta I)^{-1}\Lambda$ in §4.2, where $\Lambda' = \mathrm{diag}(G_X^\dagger \widehat{m}_\Pi)$.

§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 3.3. 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

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

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

$$L^p(\mathbb{R}^2, \mathbb{R}) \quad := \quad \left\{ f : \mathbb{R}^2 \to \mathbb{R} \,\bigg|\, \int_{\mathbb{R}^2} |f(x_1, x_2)|^p \, dx_1 dx_2 < \infty \right\}.$$

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

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

We can uniquely extend the Fourier transform to an isometry $\hat{} : L^2(\mathbb{R}, \mathbb{C}) \to L^2(\mathbb{R}, \mathbb{C})$. We also define the inverse Fourier transform $\check{} : L^2(\mathbb{R}, \mathbb{C}) \to L^2(\mathbb{R}, \mathbb{C})$ as an isometry uniquely determined by

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

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

Definition 3.5. Let us define a Gaussian kernel k_G on $\mathbb{R}$ by

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

As described in [5], the RKHS of real-valued functions and complex-valued functions corresponding to the positive definite kernel k_G are given by

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

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

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

$$\langle f, g \rangle = \int_{-\infty}^{\infty} \hat{f}(t)\overline{\hat{g}(t)} \exp\left(\frac{\sigma^2}{2}t^2\right) dt, \tag{7}$$

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

Fukumizu et al. [2] mentioned that the conditional expectation function $E[g(Y)\,|\,X = \cdot]$ is not always included in $\mathcal{H}_X$. Indeed, if the variables X and Y are independent, then $E[g(Y)\,|\,X = \cdot]$ becomes a constant function on X, the value of which might be non-zero. In the case that $X = \mathbb{R}$ and $k_X = k_G$, the constant function with non-zero value is not contained in $\mathcal{H}_X = \mathcal{H}_G$.

Additionally, in order to prove Theorems 5 and 8 in [2], they made the assumption that $E[k_\mathcal{Y}(Y, \tilde{Y})\,|\,X = x, \tilde{X} = \tilde{x}] \in \mathcal{H}_X \otimes \mathcal{H}_X$ and $E[k_X(Z, \tilde{Z})\,|\,W = y, \tilde{W} = \tilde{y}] \in \mathcal{H}_\mathcal{Y} \otimes \mathcal{H}_\mathcal{Y}$, where $(\tilde{X}, \tilde{Y})$ and $(\tilde{Z}, \tilde{W})$ are independent copies of the random variables (X, Y) and (Z, W) on $X \times \mathcal{Y}$, respectively. We also see that this assumption does not hold in general. Suppose that X and Y are independent and that so are $\tilde{X}$ and $\tilde{Y}$. Then $E[k_\mathcal{Y}(Y, \tilde{Y})\,|\,X = x, \tilde{X} = \tilde{x}]$ is a constant function of $(x, \tilde{x})$, the value of which might be non-zero. In the case that $X = \mathbb{R}$ and $k_X = k_G$, the constant function having non-zero value is not contained in $\mathcal{H}_X \otimes \mathcal{H}_X = \mathcal{H}_G \otimes \mathcal{H}_G$. Note that $\mathcal{H}_G \otimes \mathcal{H}_G$ is isomorphic to the RKHS corresponding to the kernel $k((x_1, x_2), (\tilde{x}_1, \tilde{x}_2)) = k_G(x_1, \tilde{x}_1)k_G(x_2, \tilde{x}_2)$ on $\mathbb{R}^2$, that is,

$$\mathcal{H}_G \otimes \mathcal{H}_G = \left\{ f \in L^2(\mathbb{R}^2, \mathbb{R}) \;\middle|\; \int_{\mathbb{R}^2} |\hat{f}(t_1, t_2)|^2 \exp\left(\frac{\sigma^2}{2}(t_1^2 + t_2^2)\right) dt_1 dt_2 < \infty \right\},$$

where the Fourier transform of $f : \mathbb{R}^2 \to \mathbb{R}$ is defined by

$$\hat{f}(t_1, t_2) := \mathop{\mathrm{l.i.m.}}_{n\to\infty} \left(\frac{1}{\sqrt{2\pi}}\right)^2 \int_{x_1^2 + x_2^2 < n} f(x_1, x_2) \exp\left(-\sqrt{-1}(t_1 x_1 + t_2 x_2)\right) dx_1 dx_2.$$

Thus, the assumption that conditional expectation functions are included in the RKHS does not hold in general. Since most of the theorems in [2] require this assumption, 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 §3.1 and §3.2. We first introduce probabilistic classifiers in §4.1 based on conventional Bayes' rule assuming Gaussian distributions (BR), original kernel Bayes' rule (KBR1), and kernel Bayes' rule using Moore-Penrose generalized inverse matrices (KBR2). In §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 Π 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, $Q_{\mathcal{X}|y}(C_1), \ldots, Q_{\mathcal{X}|y}(C_g)$.

§4.1.1 The algorithm of BR

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

$$\widehat{Q}_{\mathcal{X}|y}(C_j) = \frac{\widehat{P}_{\mathcal{Y}|C_j}(y)\, \Pi(C_j)}{\sum_{k=1}^g \widehat{P}_{\mathcal{Y}|C_k}(y)\, \Pi(C_k)},$$

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

$$\widehat{P}_{\mathcal{Y}|C_j}(\cdot) = \frac{1}{\sqrt{(2\pi)^d \left| \widehat{S}_j \right|}} \exp\left(-\frac{1}{2}(\cdot - \widehat{M}_j)^{\mathrm{T}} \widehat{S}_j^{-1}(\cdot - \widehat{M}_j) \right).$$

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

§4.1.2 The algorithm of KBR1

Let us define positive definite kernels $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$ as

$$k_{\mathcal{X}}(X, X') = \begin{cases} 1 & (X = X') \\ 0 & (X \neq X') \end{cases} \quad \text{and} \quad k_{\mathcal{Y}}(Y, Y') = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\|Y - Y'\|^2}{2\sigma^2} \right)$$

for $X, X' \in \mathcal{X}$ and $Y, Y' \in \mathcal{Y}$, and the corresponding RKHS as $\mathcal{H}_{\mathcal{X}}$ and $\mathcal{H}_{\mathcal{Y}}$, respectively. Here we set $\|Y\| = \sqrt{\sum_{i=1}^d y_i^2}$ for $Y = (y_1, y_2, \ldots, y_d)^{\mathrm{T}} \in \mathcal{Y} = \mathbb{R}^d$. Then, the prior kernel

mean is given by

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

where $\sum_{j=1}^{g} \Pi(C_j) = 1$. Let us put $G_X = (k_\mathcal{X}(X_i, X_j))_{1 \le i,j \le n}$, $G_Y = (k_\mathcal{Y}(Y_i, Y_j))_{1 \le i,j \le n}$, $D = (\mathbb{1}_{\{C_i\}}(X_j))_{1 \le i \le g, 1 \le j \le n} \in \{0,1\}^{g \times n}$, $\widehat{\boldsymbol{m}}_\Pi = (\widehat{m}_\Pi(X_1), \dots, \widehat{m}_\Pi(X_n))^\mathrm{T}$, $\widehat{\boldsymbol{\mu}} = (\widehat{\mu}_1, \dots, \widehat{\mu}_n)^\mathrm{T} = (G_X + n\epsilon I_n)^{-1} \widehat{\boldsymbol{m}}_\Pi$, $\Lambda = \mathrm{diag}(\widehat{\boldsymbol{\mu}})$, $\boldsymbol{k}_X(\cdot) = (k_\mathcal{X}(\cdot, X_1), \dots, k_\mathcal{X}(\cdot, X_n))^\mathrm{T}$, $\boldsymbol{k}_Y(\cdot) = (k_\mathcal{Y}(\cdot, Y_1), \dots, k_\mathcal{Y}(\cdot, Y_n))^\mathrm{T}$, and $R_{X|Y} = \Lambda G_Y((\Lambda G_Y)^2 + \delta I_n)^{-1} \Lambda$, where I_n is the identity matrix of size n and $\epsilon, \delta \in \mathbb{R}$ are heuristically set regularization parameters. Note that $\mathbb{1}_A$ stands for the indicator function of a set A described as

$$\mathbb{1}_A(t) := \begin{cases} 1 & (t \in A) \\ 0 & (t \notin A) \end{cases}.$$

Following Theorem 2.9, the posterior kernel mean given a test value $y \in \mathcal{Y}$ is estimated by

$$\widehat{m}_{Q_{X|y}} = \boldsymbol{k}_X^\mathrm{T} R_{X|Y} \boldsymbol{k}_Y(y).$$

Here, we estimate the posterior probabilities for classes given a test value $y \in \mathcal{Y}$ by

$$\begin{pmatrix} \widehat{Q}_{X|y}(C_1) \\ \vdots \\ \widehat{Q}_{X|y}(C_g) \end{pmatrix} = \begin{pmatrix} \left\langle \mathbb{1}_{\{C_1\}}, \widehat{m}_{Q_{X|y}} \right\rangle_{\mathcal{H}_\mathcal{X}} \\ \vdots \\ \left\langle \mathbb{1}_{\{C_g\}}, \widehat{m}_{Q_{X|y}} \right\rangle_{\mathcal{H}_\mathcal{X}} \end{pmatrix} = DR_{X|Y} \boldsymbol{k}_Y(y).$$

§4.1.3   The algorithm of KBR2

Let $G_X^\dagger$ denote the Moore-Penrose generalized inverse matrix of G_X. Let us put $\widehat{\boldsymbol{\mu}}' = (\widehat{\mu}_1', \dots, \widehat{\mu}_n')^\mathrm{T} = G_X^\dagger \widehat{\boldsymbol{m}}_\Pi$, $\Lambda' = \mathrm{diag}(\widehat{\boldsymbol{\mu}}')$, and $R_{X|Y}' = (\Lambda' G_Y)^\dagger \Lambda'$. Replacing $R_{X|Y}$ in §4.1.2 for $R_{X|Y}'$, the posterior probabilities for classes given a test value $y \in \mathcal{Y}$ is estimated by

$$\left(\widehat{Q}_{X|y}(C_1), \dots, \widehat{Q}_{X|y}(C_g) \right)^\mathrm{T} = DR_{X|Y}' \boldsymbol{k}_Y(y).$$

§4.2   Probabilistic predictions by the three classifiers

In this subsection, we apply the three classifiers defined in §4.1 to a binary classification problem using computer-simulated data sets, where $\mathcal{X} = \{C_1, C_2\}$ and $\mathcal{Y} = \mathbb{R}^2$. 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 $1 \le i \le 50$, $X_i = C_2$ and $Y_i \sim \mathcal{N}(M_2, S_2)$ if $51 \le i \le 100$, $M_1 = (1,0)^\mathrm{T}$, $M_2 = (0,1)^\mathrm{T}$, and $S_1 = S_2 = \mathrm{diag}(0.1, 0.1)$. 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.
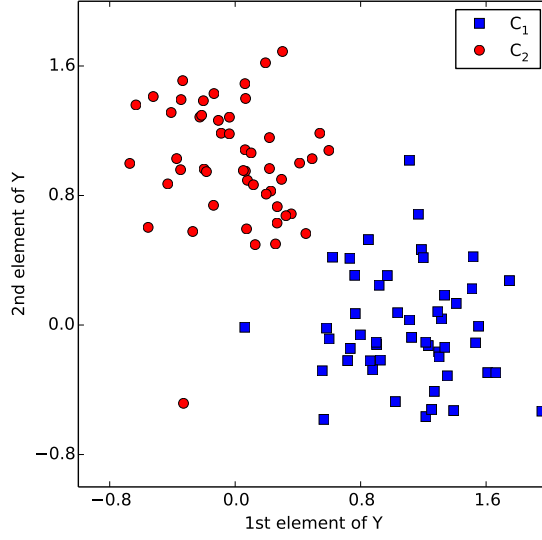
9

Figure 1: Individual Y-values of a training sample

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 §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, $Q_{\mathcal{X}|y}(C_1)$. Figures 2-5 show the mean (plus or minus standard error of the mean, SEM) of the 100 values of $\widehat{Q}_{\mathcal{X}|y}(C_1)$ calculated by each of the classifiers, BR, KBR1, and KBR2. Here we show the case where σ 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), $\epsilon = \delta = 10^{-5}$ (Figure 3), $\epsilon = \delta = 10^{-3}$ (Figure 4), and $\epsilon = \delta = 10^{-1}$ (Figure 5). In Figures 2-5, BR_th illustrates the theoretical result of BR, where $\widehat{M}_1$, $\widehat{M}_2$, $\widehat{S}_1$, and $\widehat{S}_2$ in BR are set to be M_1, M_2, S_1, and S_2, respectively.

Consistent to §3.1, $\widehat{Q}_{\mathcal{X}|y}(C_1)$ calculated by KBR1 is poorly influenced by $\Pi(C_1)$ compared with that by BR when ϵ and δ are set to be small (see Figures 2 and 3). In addition, $\widehat{Q}_{\mathcal{X}|y}(C_1)$ calculated by KBR2 also seems to be uninfluenced by $\Pi(C_1)$. When ϵ and δ are set to be larger, the effect of $\Pi(C_1)$ on $\widehat{Q}_{\mathcal{X}|y}(C_1)$ becomes apparent in KBR1, however, the value of $\widehat{Q}_{\mathcal{X}|y}(C_1)$ becomes too small (see Figures 4 and 5). These results suggest that in kernel Bayes' rule, the posterior does not depend on the prior if ϵ and δ are negligible, which might be a contradiction to the nature of Bayes' theorem. Moreover, even though the prior affects the posterior when ϵ and δ become larger, the posterior seems too much dependent on ϵ and δ, 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: $\epsilon \in \{10^{-1}, 10^{-3}, 10^{-5}, 10^{-7}, 10^{-9}, 10^{-11}, 10^{-13}, 10^{-15}\}$, $\delta \in \{10^{-1}, 10^{-3}, 10^{-5}, 10^{-7}, 10^{-9}, 10^{-11}, 10^{-13}, 10^{-15}\}$, and $\sigma \in \{0.01, 0.1, 1, 10, 100\}$.
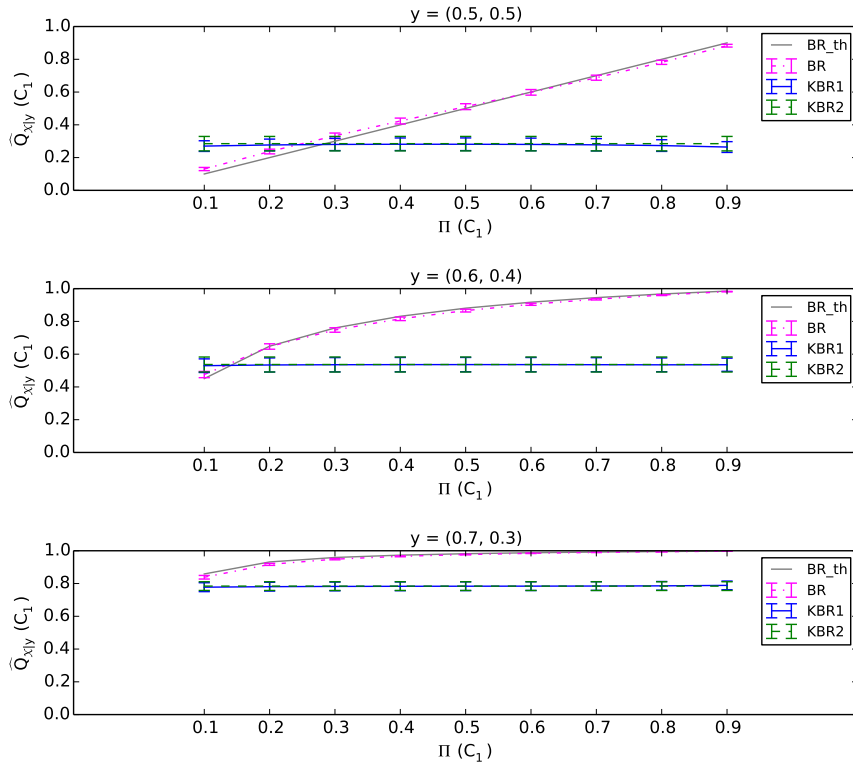
Figure 2: The case $\epsilon = \delta = 10^{-7}$

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.

§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. 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 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 kernel Bayes' rule, at least in the context of Bayesian inference.

First, as shown in §3.1, when Λ and G_Y are non-singular matrices and so we set $\delta = 0$, the posterior kernel mean $\widehat{m}_{Q_{X|y}}$ is entirely unaffected by the prior distribution Π on $\mathcal{X}$. This means that, in Bayesian inference with kernel Bayes' rule, prior beliefs
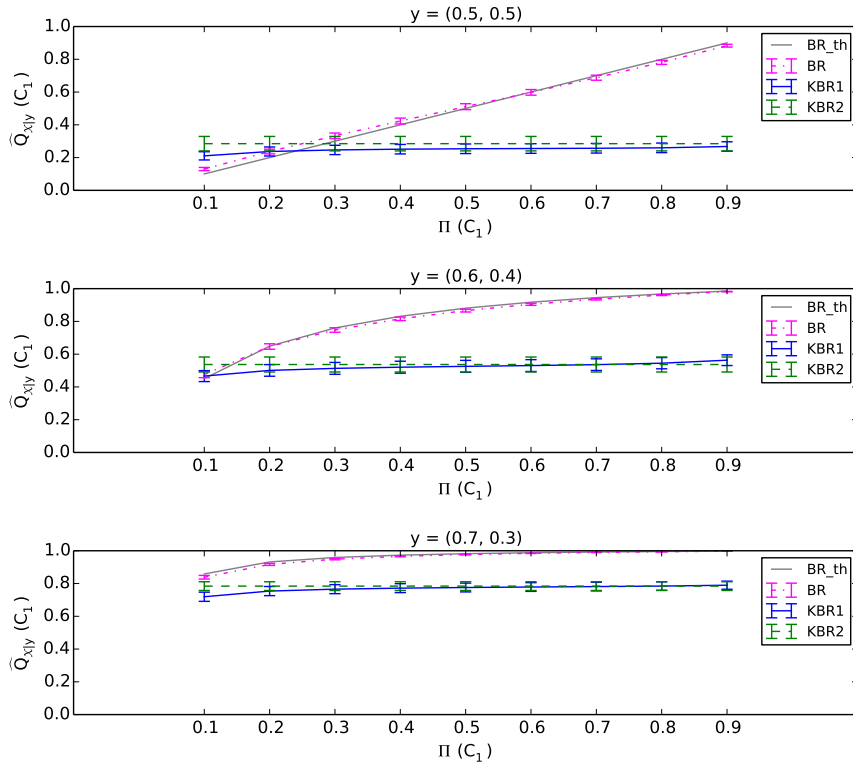
Figure 3: The case $\epsilon = \delta = 10^{-5}$

are in some cases completely neglected in calculating the kernel mean of the posterior distribution. Numerical evidence is also presented in §4.2. When the regularization parameters ϵ and δ are set to be small, the posterior probability calculated by 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.

Second, as discussed in §3.2 and §4.2, the posterior estimated by kernel Bayes' rule considerably depends upon the regularization parameters ϵ and δ, which are originally introduced just for the regularization of matrices. A cross-validation approach is proposed in [2] 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
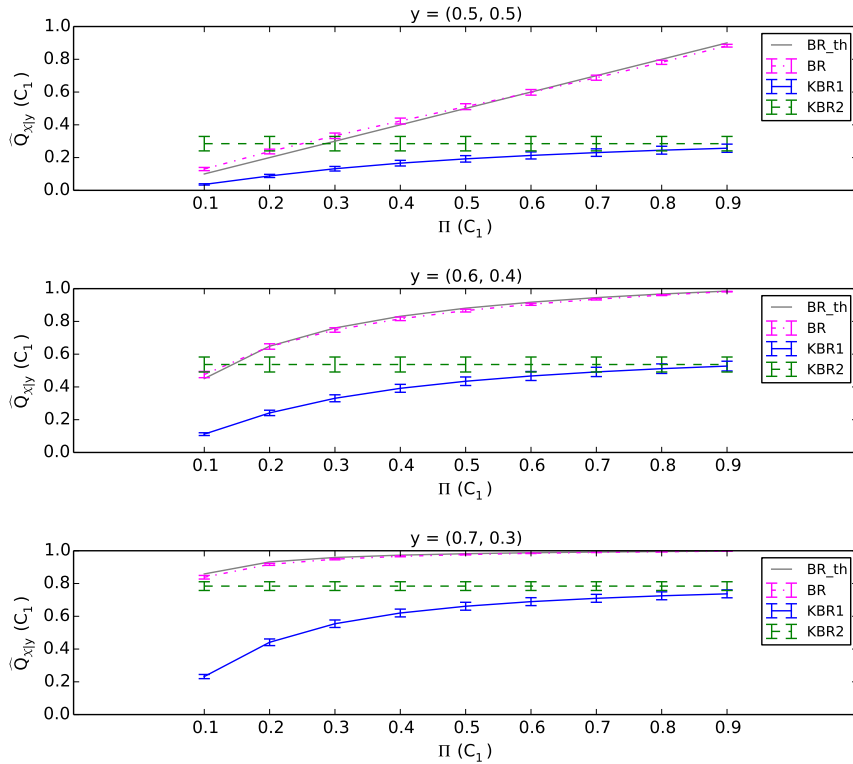
Figure 4: The case $\epsilon = \delta = 10^{-3}$

problem that C_{XX} and C_{WW} are not surjective by just adding regularization parameters.

Third, as shown in §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 [2], we believe that the assumption itself may need to be reconsidered.

In summary, even though current research efforts are focused on the application of kernel Bayes' rule [2, 4], it might be necessary to reexamine its basic framework of combining new evidence with prior beliefs.

§6 Appendix

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

§6.1 Estimation of C_{ZW} and C_{WW}

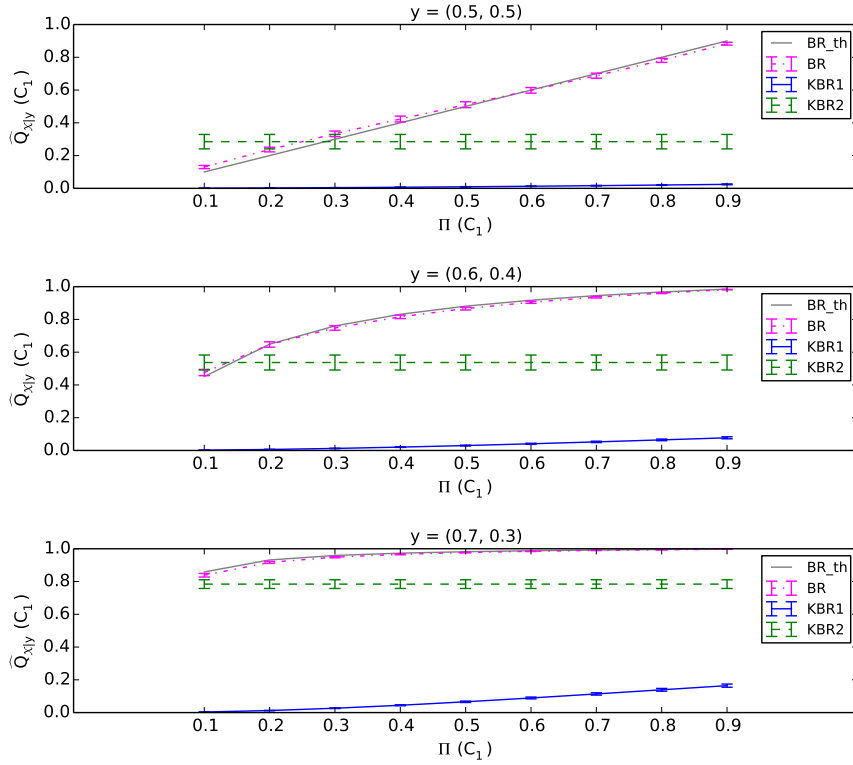Here we give the proof of Proposition 2.8.

13

Figure 5: The case $\epsilon = \delta = 10^{-1}$

Proof. Let $\widehat{C}_{XX}$, $\widehat{C}_{(YX)X}$, and $\widehat{C}_{(YY)X}$ 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

$$\widehat{m}_{(ZW)} = \widehat{C}_{(YX)X}\widehat{C}_{XX}^{-1}\widehat{m}_{\Pi} \quad \text{and} \quad \widehat{m}_{(WW)} = \widehat{C}_{(YY)X}\widehat{C}_{XX}^{-1}\widehat{m}_{\Pi},$$

and put $h = \widehat{C}_{XX}^{-1}\widehat{m}_{\Pi} \in \mathcal{H}_X$. According to Equations (5), for any $f \in \mathcal{H}_X$ and $g \in \mathcal{H}_Y$,

$$\langle \widehat{m}_{(ZW)}, g \otimes f \rangle_{\mathcal{H}_Y \otimes \mathcal{H}_X} = \left\langle \widehat{C}_{(YX)X} h, g \otimes f \right\rangle_{\mathcal{H}_Y \otimes \mathcal{H}_X} = \widehat{E}[f(X)g(Y)h(X)]$$

$$= \frac{1}{n}\sum_{i=1}^{n} f(X_i)g(Y_i)h(X_i) = \left\langle \frac{1}{n}\sum_{i=1}^{n} h(X_i)k_X(\cdot, X_i) \otimes k_Y(\cdot, Y_i), f \otimes g \right\rangle_{\mathcal{H}_X \otimes \mathcal{H}_Y},$$

where $\widehat{E}[\cdot]$ represents the empirical expectation operator. Thus, from Remark 2.6,

$$\widehat{C}_{ZW} = \widehat{m}_{(ZW)} = \frac{1}{n}\sum_{i=1}^{n} h(X_i)k_X(\cdot, X_i) \otimes k_Y(\cdot, Y_i). \tag{8}$$

14

Similarly, for any $g_1, g_2 \in \mathcal{H}_y$,

$$\langle \widehat{m}_{(WW)}, g_1 \otimes g_2 \rangle_{\mathcal{H}_y \otimes \mathcal{H}_y} = \langle \widehat{C}_{(YY)X} h, g_1 \otimes g_2 \rangle_{\mathcal{H}_y \otimes \mathcal{H}_y} = \widehat{E}[g_1(Y)g_2(Y)h(X)]$$

$$= \frac{1}{n} \sum_{i=1}^{n} g_1(Y_i)g_2(Y_i)h(X_i) = \left\langle \frac{1}{n} \sum_{i=1}^{n} h(X_i)k_y(\cdot, Y_i) \otimes k_y(\cdot, Y_i), g_1 \otimes g_2 \right\rangle_{\mathcal{H}_y \otimes \mathcal{H}_y}.$$

Thus, from Remark 2.6,

$$\widehat{C}_{WW} = \widehat{m}_{(WW)} = \frac{1}{n} \sum_{i=1}^{n} h(X_i)k_y(\cdot, Y_i) \otimes k_y(\cdot, Y_i). \tag{9}$$

Next, we will derive $h(X_1), \ldots, h(X_n)$. Since C_{XX} is a self-adjoint operator,

$$\langle h, \widehat{C}_{XX}f \rangle_{\mathcal{H}_X} = \langle \widehat{C}_{XX}h, f \rangle_{\mathcal{H}_X} = \langle \widehat{m}_{\Pi}, f \rangle_{\mathcal{H}_X} = \sum_{j=1}^{l} \gamma_j f(U_j)$$

for any $f \in \mathcal{H}_X$. On the other hand, from Equations (1),

$$\langle h, \widehat{C}_{XX}f \rangle_{\mathcal{H}_X} = \widehat{E}[f(X)h(X)] = \frac{1}{n} \sum_{i=1}^{n} f(X_i)h(X_i)$$

for any $f \in \mathcal{H}_X$. Hence, we have

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

for any $f \in \mathcal{H}_X$. Replacing f in Equation (10) for $k_X(X_1, \cdot), \ldots, k_X(X_n, \cdot) \in \mathcal{H}_X$, we have

$$\begin{pmatrix} k_X(X_1, U_1) & \cdots & k_X(X_1, U_l) \\ \vdots & \ddots & \vdots \\ k_X(X_n, U_1) & \cdots & k_X(X_n, U_l) \end{pmatrix} \begin{pmatrix} \gamma_1 \\ \vdots \\ \gamma_l \end{pmatrix} = \frac{1}{n} G_X \begin{pmatrix} h(X_1) \\ \vdots \\ h(X_n) \end{pmatrix}. \tag{11}$$

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

$$\begin{pmatrix} \langle \sum_{j=1}^{l} \gamma_j k_X(\cdot, U_j), k_X(\cdot, X_1) \rangle_{\mathcal{H}_X} \\ \vdots \\ \langle \sum_{j=1}^{l} \gamma_j k_X(\cdot, U_j), k_X(\cdot, X_n) \rangle_{\mathcal{H}_X} \end{pmatrix} = \begin{pmatrix} \sum_{j=1}^{l} \gamma_j k_X(X_1, U_j) \\ \vdots \\ \sum_{j=1}^{l} \gamma_j k_X(X_n, U_j) \end{pmatrix} = \begin{pmatrix} \widehat{m}_{\Pi}(X_1) \\ \vdots \\ \widehat{m}_{\Pi}(X_n) \end{pmatrix}.$$

Therefore, we have

$$\frac{1}{n} \begin{pmatrix} h(X_1) \\ \vdots \\ h(X_n) \end{pmatrix} = G_X^{-1} \begin{pmatrix} \widehat{m}_{\Pi}(X_1) \\ \vdots \\ \widehat{m}_{\Pi}(X_n) \end{pmatrix} \approx (G_X + n\epsilon I)^{-1} \widehat{m}_{\Pi} = \widehat{\mu}.$$

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

$$\widehat{C}_{ZW} = \sum_{i=1}^{n} \widehat{\mu}_i k_X(\cdot, X_i) \otimes k_y(\cdot, Y_i) \quad \text{and} \quad \widehat{C}_{WW} = \sum_{i=1}^{n} \widehat{\mu}_i k_y(\cdot, Y_i) \otimes k_y(\cdot, Y_i),$$

respectively. $\qquad\qquad\square$

§6.2 Non-singularity of G_Y and Λ

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

Definition 6.1. Let f be a real-valued function defined on a non-empty open domain $\mathrm{Dom}(f) \subseteq \mathbb{R}^d$. We say that f is analytic if f can be described by a Taylor expansion on a neighborhood of each point of $\mathrm{Dom}(f)$.

Proposition 6.2. *Let k be a positive definite kernel on $\mathbb{R}^d$. Let ν 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 ν, 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

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

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

From Proposition 6.2, we easily obtain the following corollary.

Corollary 6.3. *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 6.4. *Let k be a positive definite kernel on $\mathcal{X} = \mathbb{R}^d$, ν 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 $\mathcal{H}$ corresponding to k is infinite dimensional. Then for any $(\epsilon, \gamma_1, \gamma_2, \ldots, \gamma_l, U_1, U_2, \ldots, U_l) \in \mathbb{R}_+ \times \mathbb{R}^l \times (\mathbb{R}^d)^l$ except Lebesgue measure zero, and for any i.i.d. random variables $X_1, X_2, \ldots, X_n$ with the same distribution ν, each μ_i for $i = 1, 2, \ldots, n$ is non-zero almost surely, where $(\mu_1, \mu_2, \ldots, \mu_n)^T = (G_X + n\epsilon I_n)^{-1} \widehat{\boldsymbol{m}}_\Pi$, $\widehat{\boldsymbol{m}}_\Pi = (\widehat{m}_\Pi(X_1), \widehat{m}_\Pi(X_2), \ldots, \widehat{m}_\Pi(X_n))^T$, and $\widehat{m}_\Pi(\cdot) = \sum_{j=1}^l \gamma_j k(\cdot, U_j)$. Here $\mathbb{R}_+$ denotes the set of positive real numbers.*

Proof. Let us put $\mathcal{S} := \mathbb{R}_+ \times \mathbb{R}^l \times (\mathbb{R}^d)^l$, $\mathcal{T} := \mathcal{X}^n \times \mathcal{S}$, and

$$f_i(x_1, x_2, \ldots, x_n, \epsilon, \gamma_1, \gamma_2, \ldots, \gamma_l, U_1, U_2, \ldots, U_l) := \mu_i \quad (i = 1, 2, \ldots, n)$$

for $(x_1, x_2, \ldots, x_n) \in \mathcal{X}^n$ and $(\epsilon, \gamma_1, \gamma_2, \ldots, \gamma_l, U_1, U_2, \ldots, U_l) \in \mathcal{S}$. We can verify that $G_X + n\epsilon I_n = (k(x_i, x_j))_{1 \leq i, j \leq n} + n\epsilon I_n$ is non-singular almost everywhere on $\mathcal{T}$ in the same way as in the proof of Proposition 6.2. Let us define a closed measure-zero set

$$\mathcal{V} := \{(x_1, x_2, \ldots, x_n, \epsilon, \gamma_1, \gamma_2, \ldots, \gamma_l, U_1, U_2, \ldots, U_l) \in \mathcal{T} \mid \det(G_X + n\epsilon I_n) = 0\} \subset \mathcal{T}.$$

Then f_i is defined on $\mathcal{T} \setminus \mathcal{V}$ for each $i \in \{1, 2, \ldots, n\}$. Using Cramer's rule,

$$\mu_i = \frac{\det\left(\eta_1, \eta_2, \ldots, \eta_{i-1}, \widehat{\boldsymbol{m}}_\Pi, \eta_{i+1}, \ldots, \eta_n\right)}{\det\left(G_X + n\epsilon I_n\right)},$$

where η_m stands for the m-th column vector of $G_X + n\epsilon I_n$. Here we denote by g_i the numerator of μ_i, that is, $g_i = \mu_i \det(G_X + n\epsilon I_n)$. Let us choose $\xi_1, \xi_2, \ldots, \xi_n \in \mathcal{X}$ such that $\{k(\cdot, \xi_i)\}_{1 \le i \le n}$ are linearly independent in $\mathcal{H}$. It is easy to see that $g_i(\xi_1, \xi_2, \ldots, \xi_n, *)$ is a non-zero analytic function of $*$ on $\mathcal{S}$. Indeed, if $\epsilon \to +0$, $U_1 = \xi_i$, $\gamma_1 = 1$, and $\gamma_2 = \gamma_3 = \cdots = \gamma_l = 0$, then $g_i \to \det(\langle k(\cdot, \xi_i), k(\cdot, \xi_j) \rangle_\mathcal{H})_{1 \le i, j \le n} \ne 0$. Hence $\mathcal{Z}_i := \{* \in \mathcal{S} \mid g_i(\xi_1, \xi_2, \ldots, \xi_n, *) = 0\}$ is a closed subset of $\mathcal{S}$ with Lebesgue measure zero for each $i \in \{1, 2, \ldots, n\}$. Thus, since $g_i(*, \epsilon, \gamma_1, \gamma_2, \ldots, \gamma_l, U_1, U_2, \ldots, U_l)$ is a non-zero analytic function of $*$ on $\mathcal{X}^n$ for any $(\epsilon, \gamma_1, \gamma_2, \ldots, \gamma_l, U_1, U_2, \ldots, U_l) \in \mathcal{S} \setminus (\cup_{i=1}^n \mathcal{Z}_i)$,

$$\mathcal{F}_i := \left\{ * \in \mathcal{X}^n \mid g_i(*, \epsilon, \gamma_1, \gamma_2, \ldots, \gamma_l, U_1, U_2, \ldots, U_l) = 0 \right\}$$

is a closed subset of $\mathcal{X}^n$ with Lebesgue measure zero for each $i \in \{1, 2, \ldots, n\}$. Therefore $\mu_i = f_i(*, \epsilon, \gamma_1, \gamma_2, \ldots, \gamma_l, U_1, U_2, \ldots, U_l)$ is non-zero almost surely on $\mathcal{X}^n$ for any $(\epsilon, \gamma_1, \gamma_2, \ldots, \gamma_l, U_1, U_2, \ldots, U_l) \in \mathcal{S} \setminus (\cup_{i=1}^n \mathcal{Z}_i)$ because the subset $\{* \in \mathcal{X}^n \mid f_i(*, \epsilon, \gamma_1, \gamma_2, \ldots, \gamma_l, U_1, U_2, \ldots, U_l) = 0\}$ is contained in $\mathcal{F}_i$ for each $i \in \{1, 2, \ldots, n\}$. This completes the proof. $\qquad\square$

The following corollary directly follows from Proposition 6.4.

Corollary 6.5. *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 6.4. Then $\Lambda := \mathrm{diag}(\mu_1, \mu_2, \ldots, \mu_n)$ is non-singular almost surely for any $(\epsilon, \gamma_1, \gamma_2, \ldots, \gamma_l, U_1, U_2, \ldots, U_l) \in \mathbb{R}_+ \times \mathbb{R}^l \times (\mathbb{R}^d)^l$ 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 not 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 $\mathcal{X} = \mathbb{R}$. Let X be a random variable on $\mathbb{R}$ with a normal distribution $\mathcal{N}(\mu, \sigma_0^2)$. 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 §3.3 and several proven results on function spaces and Fourier transforms (see [7], for example). Note that the following three propositions are introduced without proofs.

Proposition 6.6. *Let us put $f(x) = \exp(-(ax^2 + bx + c))$ for $a, b, c \in \mathbb{R}$, where $a > 0$. Then*

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

Proposition 6.7. *For $f \in L^2(\mathbb{R}, \mathbb{C})$, $\overline{\hat{f}(t)} = \overline{\hat{f}(-t)}$ almost everywhere. In particular, if $f \in L^2(\mathbb{R})$, then $\overline{\hat{f}(t)} = \hat{f}(-t)$ almost everywhere.*

Proposition 6.8. *For $f \in L^2(\mathbb{R}, \mathbb{C})$, put $f_a(x) := f(x-a)$. Then $\hat{f}_a(t) = \exp\left(-\sqrt{-1}at\right)\hat{f}(t)$.*

Definition 6.9. Let $p(\cdot)$ denote the density function of the normal distribution $\mathcal{N}(\mu, \sigma_0^2)$ on $\mathbb{R}$, that is,

$$p(\cdot) = \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left(-\frac{(\cdot - \mu)^2}{2\sigma_0^2}\right).$$

Let X be a random variable on $\mathbb{R}$ with $\mathcal{N}(\mu, \sigma_0^2)$. The linear operator $C_{XX} : \mathcal{H}_G \to \mathcal{H}_G$ is defined by $\langle C_{XX}f, g\rangle_{\mathcal{H}_G} = E[f(X)g(X)]$ for any $f, g \in \mathcal{H}_G$, which is also described as

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

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

Proposition 6.10. *If $f, g \in \mathcal{H}_G$, then $\langle f, g\rangle_{\mathcal{H}_G} \in \mathbb{R}$*

Proof. From Proposition 6.7, $\overline{\hat{f}(t)} = \hat{f}(-t)$ and $\overline{\hat{g}(t)} = \hat{g}(-t)$ for any $f, g \in \mathcal{H}_G$. Then, using Equation (7), we have

$$\overline{\langle f, g\rangle_{\mathcal{H}_G}} = \overline{\int_{-\infty}^{\infty} \hat{f}(t)\overline{\hat{g}(t)}\exp\left(\frac{\sigma^2}{2}t^2\right)dt} = \int_{-\infty}^{\infty} \overline{\hat{f}(t)}\hat{g}(t)\exp\left(\frac{\sigma^2}{2}t^2\right)dt$$

$$= \int_{-\infty}^{\infty} \hat{f}(-t)\hat{g}(t)\exp\left(\frac{\sigma^2}{2}t^2\right)dt = \int_{-\infty}^{\infty} \hat{f}(t)\hat{g}(-t)\exp\left(\frac{\sigma^2}{2}t^2\right)dt$$

$$= \int_{-\infty}^{\infty} \hat{f}(t)\overline{\hat{g}(t)}\exp\left(\frac{\sigma^2}{2}t^2\right)dt = \langle f, g\rangle_{\mathcal{H}_G}.$$

Therefore, $\langle f, g\rangle_{\mathcal{H}_G} \in \mathbb{R}$. $\qquad\square$

Proposition 6.11. *If $f \in \mathcal{H}_G(\mathbb{R}, \mathbb{C})$, then $\overline{f} \in \mathcal{H}_G(\mathbb{R}, \mathbb{C})$.*

Proof. From Proposition 6.7, $\overline{\hat{f}(t)} = \overline{\hat{f}(-t)}$ for $f \in L^2(\mathbb{R}, \mathbb{C})$. Then, using Equation (7), we have

$$\left\|\overline{f}\right\|_{\mathcal{H}_G(\mathbb{R}, \mathbb{C})}^2 = \left\langle \overline{f}, \overline{f}\right\rangle_{\mathcal{H}_G(\mathbb{R}, \mathbb{C})}$$

$$= \int_{-\infty}^{\infty} \left|\overline{\hat{f}(t)}\right|^2 \exp\left(\frac{\sigma^2}{2}t^2\right)dt = \int_{-\infty}^{\infty} \left|\overline{\hat{f}(-t)}\right|^2 \exp\left(\frac{\sigma^2}{2}t^2\right)dt$$

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

Therefore, $\overline{f} \in \mathcal{H}_G(\mathbb{R}, \mathbb{C})$. $\qquad\square$

Here, we denote by Re, Im, and Cl the real part of a complex number, the imaginary part of a complex number, and the closure operator, respectively.

Corollary 6.12. *If $f \in \mathcal{H}_G(\mathbb{R}, \mathbb{C})$, then $\mathrm{Re}(f), \mathrm{Im}(f) \in \mathcal{H}_G$.*

Proof. If $f \in \mathcal{H}_G(\mathbb{R}, \mathbb{C})$, then $\overline{f} \in \mathcal{H}_G(\mathbb{R}, \mathbb{C})$ by Proposition 6.11. Hence we see that

$$\mathrm{Re}(f) = \frac{f + \overline{f}}{2} \in \mathcal{H}_G, \quad \mathrm{Im}(f) = \frac{f - \overline{f}}{2\sqrt{-1}} \in \mathcal{H}_G.$$

This completes the proof. □

Remark 6.13. If $f \in \mathcal{H}_G(\mathbb{R}, \mathbb{C})$, then there uniquely exist $f_1, f_2 \in \mathcal{H}_G$ such that $f = f_1 + \sqrt{-1} f_2$ by Corollary 6.12. This means that $\mathcal{H}_G(\mathbb{R}, \mathbb{C}) = \mathcal{H}_G \oplus \sqrt{-1}\mathcal{H}_G$, where $\oplus$ denotes the direct sum.

Proposition 6.14. *For any $f \in L^2(\mathbb{R}, \mathbb{C})$ and for any $\epsilon > 0$, there exists $g \in \mathcal{H}_G(\mathbb{R}, \mathbb{C})$ such that $\|f - g\|_2 < \epsilon$. In other words, $\mathcal{H}_G(\mathbb{R}, \mathbb{C})$ is dense in $L^2(\mathbb{R}, \mathbb{C})$.*

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

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

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

The following corollary has also been shown in Theorem 4.63 in [8].

Corollary 6.15. $\mathrm{Cl}(\mathcal{H}_G) = L^2(\mathbb{R})$.

Proof. From Proposition 6.14, for any $f \in L^2(\mathbb{R}) \subset L^2(\mathbb{R}, \mathbb{C})$ and for any $\epsilon > 0$, there exists $g \in \mathcal{H}_G(\mathbb{R}, \mathbb{C})$ such that $\|f - g\|_2 < \epsilon$. By Remark 6.13, there exist $g_1, g_2 \in \mathcal{H}_G$ such that $g = g_1 + \sqrt{-1} g_2$. Thus,

$$\epsilon^2 > \|f - g\|_2^2 = \int_{-\infty}^{\infty} |f - g|^2 \, dx = \int_{-\infty}^{\infty} \left|(f - g_1) - \sqrt{-1} g_2\right|^2 \, dx$$

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

Therefore $\|f - g_1\|_2 < \epsilon$. This completes the proof. □

Definition 6.16. Let us define $r, r_n \in L^2(\mathbb{R})$ as

$$r(t) := \frac{1}{|t|}\mathbb{1}_{(1,\infty)}(|t|), \quad r_n(t) := \frac{1}{|t|}\mathbb{1}_{(1,n)}(|t|),$$

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

$$\lim_{n\to\infty} \|r_n - r\|_2^2 = 2 \lim_{n\to\infty} \int_n^\infty \frac{1}{x^2}\, dx = 0.$$

Proposition 6.17. $h_n, h \in L^2(\mathbb{R})$.

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

$$\overline{h_n(x)} = \overline{\check{r}_n(x)} = \overline{\frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty r_n(t) \exp\left(\sqrt{-1}tx\right) dt}$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty r_n(t) \exp\left(-\sqrt{-1}tx\right) dt = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty r_n(-t') \exp\left(\sqrt{-1}t'x\right) dt'$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty r_n(t') \exp\left(\sqrt{-1}t'x\right) dt' = \check{r}_n(x) = h_n(x),$$

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

$$
\begin{aligned}
h(x) &= \operatorname*{l.i.m.}_{n\to\infty} \frac{1}{\sqrt{2\pi}} \int_{-n}^n r(t) \exp\left(\sqrt{-1}tx\right) dt \\
&= \operatorname*{l.i.m.}_{n\to\infty} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty r_n(t) \exp\left(\sqrt{-1}tx\right) dt \\
&= \operatorname*{l.i.m.}_{n\to\infty} h_n(x).
\end{aligned}
$$

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

Let us define $k_a(\cdot) := \sqrt{2\pi}\sigma k_G(\cdot, a) = \exp\left(-\frac{(\cdot - a)^2}{2\sigma^2}\right) \in \mathcal{H}_G$ for $a \in \mathbb{R}$. Now, we prove that $k_a \notin \operatorname{Ran}(C_{XX})$ for any $a \in \mathbb{R}$. This implies that C_{XX} is not surjective.

Proposition 6.18. *For any* $a \in \mathbb{R}$, $k_a \in \mathcal{H}_G \setminus \operatorname{Ran}(C_{XX})$.

Proof. Suppose that there exists $g \in \mathcal{H}_G$ such that $C_{XX}g = k_a$. Then, for any $f \in \mathcal{H}_G$,

$$\langle k_a, f \rangle_{\mathcal{H}_G} = \langle C_{XX}g, f \rangle_{\mathcal{H}_G}. \tag{12}$$

Let us put $k(\cdot) = \sqrt{2\pi}\sigma k_G(\cdot, 0) = \exp\left(-\frac{(\cdot - 0)^2}{2\sigma^2}\right)$. From Proposition 6.6, $\hat{k}(t) = \sigma \exp\left(-\frac{\sigma^2}{2}t^2\right)$. Then, using Equation (7) and Proposition 6.8, the left hand side of Equation (12) equals

$$\int_{-\infty}^\infty \hat{k}_a(t)\overline{\hat{f}(t)} \exp\left(\frac{\sigma^2}{2}t^2\right) dt = \int_{-\infty}^\infty \exp\left(-\sqrt{-1}at\right)\hat{k}(t)\overline{\hat{f}(t)} \exp\left(\frac{\sigma^2}{2}t^2\right) dt$$

$$= \sigma \int_{-\infty}^\infty \exp\left(-\sqrt{-1}at\right)\overline{\hat{f}(t)}\, dt.$$

20

The right hand side of Equation (12) is equal to

$$E\left[g(X)f(X)\right] = \int_{-\infty}^{\infty} g(x)f(x)p(x)\,dx = \langle gp, f\rangle_{L^2(\mathbb{R})}.$$

Thus, Equation (12) is equivalent to the following equation:

$$\langle gp, f\rangle_{L^2(\mathbb{R})} = \sigma \int_{-\infty}^{\infty} \exp\left(-\sqrt{-1}at\right)\overline{\hat{f}(t)}\,dt. \tag{13}$$

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 $\|h_{n,a} - h_a\|_2 = \|h_n - h\|_2 = \|r_n - r\|_2 \to 0$ as $n \to \infty$. Hence $\lim_{n\to\infty} h_{n,a} = h_a$ in $L^2(\mathbb{R})$. Since $\hat{h}_{n,a}(t) = \exp(-\sqrt{-1}at)\hat{h}_n(t)$ by Proposition 6.8, we have

$$\int_{-\infty}^{\infty} \left|\hat{h}_{n,a}(t)\right|^2 \exp\left(\frac{\sigma^2}{2}t^2\right)\,dt = \int_{-\infty}^{\infty} \left|\hat{h}_n(t)\right|^2 \exp\left(\frac{\sigma^2}{2}t^2\right)\,dt = \int_{-\infty}^{\infty} |r_n(t)|^2 \exp\left(\frac{\sigma^2}{2}t^2\right)\,dt$$

$$= 2\int_1^n \frac{1}{t^2} \exp\left(\frac{\sigma^2}{2}t^2\right)\,dt < \infty,$$

which indicates that $h_{n,a} \in \mathcal{H}_G$. Substituting $h_{n,a}$ for f, Equation (13) becomes

$$\langle gp, h_{n,a}\rangle_{L^2(\mathbb{R})} = \sigma \int_{-\infty}^{\infty} \exp\left(-\sqrt{-1}at\right)\overline{\hat{h}_{n,a}(t)}\,dt. \tag{14}$$

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

$$\sigma \int_{-\infty}^{\infty} \exp\left(-\sqrt{-1}at\right)\overline{\exp(-\sqrt{-1}at)\hat{h}_n(t)}\,dt = \sigma \int_{-\infty}^{\infty} \overline{\hat{h}_n(t)}\,dt$$

$$= \sigma \int_{-\infty}^{\infty} \overline{r_n(t)}\,dt$$

$$= 2\sigma \int_1^n \frac{1}{t}\,dt \to \infty \quad (n \to \infty).$$

This is a contradiction. Therefore, there exists no $g \in \mathcal{H}_G$ such that $C_{XX}g = k_a$. This completes the proof. $\qquad\square$

§7 Disclosures

References

[1] Le Song, Jonathan Huang, Alex Smola, Kenji Fukumizu. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 961-968, 2009.

[2] Kenji Fukumizu, Le Song, Arthur Gretton. Kernel Bayes' rule: Bayesian inference with positive definite kernels. *Journal of Machine Learning Research*, 14:3753-3783, 2013.

[3] Le Song, Kenji Fukumizu, Arthur Gretton. Kernel embeddings of conditional distributions. *IEEE Signal Processing Magazine*, 30:98-111, 2014.

[4] Motonobu Kanagawa, Yu Nishiyama, Arthur Gretton, Kenji Fukumizu. Monte Carlo filtering using kernel embedding of distributions. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, pages 1897-1903, 2014.

[5] Kenji Fukumizu. Introduction to kernel methods (in Japanese). *Asakura Shoten, Tokyo*, 2010.

[6] Roger A. Horn, Charles R Johnson. Matrix analysis, second edition. *Cambridge University Press, Cambridge*, 2013.

[7] Walter Rudin. Real and complex analysis, third edition. *McGraw-Hill Book Co., New York*, 1987.

[8] Ingo Steinwart, Andreas Christmann. Support vector machines. Information Science and Statistics. *Springer, New York*, 2008.