Journal of Machine Learning Research 22 (2021) 1-56 Submitted 11/20; Revised 8/21; Published 11/21

# Reproducing kernel Hilbert $C^*$-module and kernel mean embeddings

**Yuka Hashimoto** YUKA.HASHIMOTO.RW@HCO.NTT.CO.JP
*NTT Network Service Systems Laboratories, NTT Corporation*
*3-9-11, Midori-cho, Musashinoshi, Tokyo, 180-8585, Japan /*
*Graduate School of Science and Technology, Keio University*
*3-14-1, Hiyoshi, Kohoku, Yokohama, Kanagawa, 223-8522, Japan*

**Isao Ishikawa** ISHIKAWA.ISAO.ZX@EHIME-U.AC.JP
*Center for Data Science, Ehime University*
*2-5, Bunkyo-cho, Matsuyama, Ehime, 790-8577, Japan /*
*Center for Advanced Intelligence Project, RIKEN*
*1-4-1, Nihonbashi, Chuo-ku, Tokyo 103-0027, Japan*

**Masahiro Ikeda** MASAHIRO.IKEDA@RIKEN.JP
*Center for Advanced Intelligence Project, RIKEN*
*1-4-1, Nihonbashi, Chuo-ku, Tokyo 103-0027, Japan /*
*Faculty of Science and Technology, Keio University*
*3-14-1, Hiyoshi, Kohoku, Yokohama, Kanagawa, 223-8522, Japan*

**Fuyuta Komura** FUYUTA.K@KEIO.JP
*Faculty of Science and Technology, Keio University*
*3-14-1, Hiyoshi, Kohoku, Yokohama, Kanagawa, 223-8522, Japan /*
*Center for Advanced Intelligence Project, RIKEN*
*1-4-1, Nihonbashi, Chuo-ku, Tokyo 103-0027, Japan*

**Takeshi Katsura** KATSURA@MATH.KEIO.AC.JP
*Faculty of Science and Technology, Keio University*
*3-14-1, Hiyoshi, Kohoku, Yokohama, Kanagawa, 223-8522, Japan /*
*Center for Advanced Intelligence Project, RIKEN*
*1-4-1, Nihonbashi, Chuo-ku, Tokyo 103-0027, Japan*

**Yoshinobu Kawahara** KAWAHARA@IMI.KYUSHU-U.AC.JP
*Institute of Mathematics for Industry, Kyushu University*
*744, Motooka, Nishi-ku, Fukuoka, 819-0395, Japan /*
*Center for Advanced Intelligence Project, RIKEN*
*1-4-1, Nihonbashi, Chuo-ku, Tokyo 103-0027, Japan*

**Editor:** Corinna Cortes

## Abstract

Kernel methods have been among the most popular techniques in machine learning, where learning tasks are solved using the property of reproducing kernel Hilbert space (RKHS). In this paper, we propose a novel data analysis framework with reproducing kernel Hilbert $C^*$-module (RKHM) and kernel mean embedding (KME) in RKHM. Since RKHM contains richer information than RKHS or vector-valued RKHS (vvRKHS), analysis with RKHM enables us to capture and extract structural properties in such as functional data. We show a branch of theories for RKHM to apply to data analysis, including the representer theorem,

and the injectivity and universality of the proposed KME. We also show RKHM generalizes RKHS and vvRKHS. Then, we provide concrete procedures for employing RKHM and the proposed KME to data analysis.

**Keywords:** reproducing kernel Hilbert $C^*$-module, kernel mean embedding, structured data, kernel PCA, interaction effects

## 1. Introduction

Kernel methods have been among the most popular techniques in machine learning (Schölkopf and Smola, 2001), where learning tasks are solved using the property of reproducing kernel Hilbert space (RKHS). RKHS is the space of complex-valued functions equipped with an inner product determined by a positive-definite kernel. One of the important tools with RKHS is kernel mean embedding (KME). In KME, a probability distribution (or measure) is embedded as a function in an RKHS (Smola et al., 2007; Muandet et al., 2017; Sriperumbudur et al., 2011), which enables us to analyze distributions in RKHSs.

Whereas much of the classical literature on RKHS approaches has focused on complex-valued functions, RKHSs of vector-valued functions, i.e., vector-valued RKHSs (vvRKHSs), have also been proposed (Micchelli and Pontil, 2005; Álvarez et al., 2012; Lim et al., 2015; Minh et al., 2016; Kadri et al., 2016). This allows us to learn vector-valued functions rather than complex-valued functions.

In this paper, we develop a branch of theories on reproducing kernel Hilbert $C^*$-module (RKHM) and propose a generic framework for data analysis with RKHM. RKHM is a generalization of RKHS and vvRKHS in terms of $C^*$-algebra, and we show that RKHM is a powerful tool to analyze structural properties in such as functional data. An RKHM is constructed by a $C^*$-algebra-valued positive definite kernel and characterized by a $C^*$-algebra-valued inner product (see Definition 2.21). The theory of $C^*$-algebra has been discussed in mathematics, especially in operator algebra theory. An important example of $C^*$-algebra is $L^\infty(\Omega)$, where $\Omega$ is a compact measure space. Another important example is $\mathcal{B}(\mathcal{W})$, which denotes the space of bounded linear operators on a Hilbert space $\mathcal{W}$. Note that $\mathcal{B}(\mathcal{W})$ coincides with the space of matrices $\mathbb{C}^{m\times m}$ if the Hilbert space $\mathcal{W}$ is finite dimensional.

Although there are several advantages for studying RKHM compared with RKHS and vvRKHS, those can be summarized into two points as follows: First, an RKHM is a "Hilbert $C^*$-module", which is mathematically more general than a "Hilbert space". The inner product in an RKHM is $C^*$-algebra-valued, which captures more information than the complex-valued one in an RKHS or vvRKHS and enables us to extract richer information. For example, if we set $L^\infty(\Omega)$ as a $C^*$-algebra, we can control and extract features of functional data such as derivatives, total variation, and frequency components. Also, if we set $\mathcal{B}(\mathcal{W})$ as a $C^*$-algebra and the inner product is described by integral operators, we can control and extract features of continuous relationships between pairs of functional data. This cannot be achieved, in principle, by RKHSs and vv-RKHSs. This is because their inner products are complex-valued, where such information degenerates into one complex value or is lost by discretizations of function into complex values. Therefore, we cannot reconstruct the information from a vector in an RKHS or vvRKHS. Second, RKHM generalizes RKHS and vvRKHS, that is, it can be shown that we can reconstruct RKHSs and vvRKHSs from

RKHMs. This implies that existing algorithms with RKHSs and vvRKHSs are reconstructed by using the framework of RKHM.

The theory of RKHM has been studied in mathematical physics and pure mathematics (Itoh, 1990; Heo, 2008; Szafraniec, 2010). On the other hand, to the best of our knowledge, as for the application of RKHM to data analysis, we can find the only literature by Ye (2017), where only the case of setting the space of matrices as a $C^*$-algebra is discussed. In this paper, we develop a branch of theories on RKHM and propose a generic framework for data analysis with RKHM. We show a theoretical property on minimization with respect to orthogonal projections and give a representer theorem in RKHMs. These properties are fundamental for data analysis that have been investigated and applied in the cases of RKHS and vvRKHS, which has made RKHS and vvRKHS widely-accepted tools for data analysis (Schölkopf et al., 2001). Moreover, we define a KME in an RKHM, and provide theoretical results about the injectivity of the proposed KME and the connection with universality of RKHM. Note that, as is well known for RKHSs, these two properties have been actively studied to theoretically guarantee the validity of kernel-based algorithms (Steinwart, 2001; Gretton et al., 2006; Fukumizu et al., 2007; Sriperumbudur et al., 2011). Then, we apply the developed theories to generalize kernel PCA (Schölkopf and Smola, 2001), analyze time-series data with the theory of dynamical system, and analyze interaction effects for infinite dimensional data.

The remainder of this paper is organized as follows. First, in Section 2, we briefly review RKHS, vvRKHS, and the definition of RKHM. In Section 3, we provide an overview of the motivation of studying RKHM for data analysis. In Section 4, we show general properties of RKHM for data analysis and the connection of RKHMs with RKHSs and vvRKHSs. In Sections 5, we propose a KME in RKHMs, and show the connection between the injectivity of the KME and the universality of RKHM. Then, in Section 6, we discuss applications of the developed results to kernel PCA, time-series data analysis, and the analysis of interaction effects in finite or infinite dimensional data. Finally, in Section 7, we discuss the connection of RKHMs and the proposed KME with the existing notions, and conclude the paper in Section 8.

**Notations** Lowercase letters denote $\mathcal{A}$-valued coefficients (often by $a, b, c, d$), vectors in a Hilbert $C^*$-module $\mathcal{M}$ (often by $p, q, u, v$), or vectors in a Hilbert space $\mathcal{W}$ (often by $w, h$). Lowercase Greek letters denote measures (often by $\mu, \nu, \lambda$) or complex-valued coefficients (often by $\alpha, \beta$). Calligraphic capital letters denote sets. And, bold lowercase letters denote vectors in $\mathcal{A}^n$ for $n \in \mathbb{N}$ (a finite dimensional Hilbert $C^*$-module). Also, we use $\sim$ for objects related to RKHSs. Moreover, an inner product, an absolute value, and a norm in a space or a module $\mathcal{S}$ (see Definitions 2.12 and 2.13) are denoted as $\langle \cdot, \cdot \rangle_{\mathcal{S}}$, $|\cdot|_{\mathcal{S}}$, and $\|\cdot\|_{\mathcal{S}}$, respectively.

The typical notations in this paper are listed in Table 1.

## 2. Background

We briefly review RKHS and vvRKHS in Subsections 2.1 and 2.2, respectively. Then, we review $C^*$-algebra and $C^*$-module in Subsection 2.3, Hilbert $C^*$-module in Subsection 2.4, and RKHM in Subsection 2.5.

Table 1: Notation table

| | |
|---|---|
| $\mathcal{A}$ | A $C^*$-algebra |
| $1_{\mathcal{A}}$ | The multiplicative identity in $\mathcal{A}$ |
| $\mathcal{A}_+$ | The subset of $\mathcal{A}$ composed of all positive elements in $\mathcal{A}$ |
| $\leq_{\mathcal{A}}$ | For $c, d \in \mathcal{A}$, $c \leq_{\mathcal{A}} d$ means $d - c$ is positive. |
| $<_{\mathcal{A}}$ | For $c, d \in \mathcal{A}$, $c < d$ means $d - c$ is strictly positive, i.e., $d - c$ is positive and invertible. |
| $L^\infty(\Omega)$ | The space of complex-valued $L^\infty$ functions on a measure space $\Omega$ |
| $\mathcal{B}(\mathcal{W})$ | The space of bounded linear operators on a Hilbert space $\mathcal{W}$ |
| $\mathbb{C}^{m\times m}$ | A set of all complex-valued $m \times m$ matrix |
| $\mathcal{M}$ | A Hilbert $\mathcal{A}$-module |
| $\mathcal{X}$ | A nonempty set for data |
| $C(\mathcal{X}, \mathcal{Y})$ | The space of $\mathcal{Y}$-valued continuous functions on $\mathcal{X}$ for topological spaces $\mathcal{X}$ and $\mathcal{Y}$ |
| $n$ | A natural number that represents the number of samples |
| $k$ | An $\mathcal{A}$-valued positive definite kernel |
| $\phi$ | The feature map endowed with $k$ |
| $\mathcal{M}_k$ | The RKHM associated with $k$ |
| $\mathcal{S}^{\mathcal{X}}$ | The set of all functions from a set $\mathcal{X}$ to a space $\mathcal{S}$ |
| $\tilde{k}$ | A complex-valued positive definite kernel |
| $\tilde{\phi}$ | The feature map endowed with $\tilde{k}$ |
| $\mathcal{H}_{\tilde{k}}$ | The RKHS associated with $\tilde{k}$ |
| $\mathcal{H}_k^{\mathrm{v}}$ | The vvRKHS associated with $k$ |
| $\mathcal{D}(\mathcal{X}, \mathcal{A})$ | The set of all $\mathcal{A}$-valued finite regular Borel measures |
| $\Phi$ | The proposed KME in an RKHM |
| $\delta_x$ | The $\mathcal{A}$-valued Dirac measure defined as $\delta_x(E) = 1_{\mathcal{A}}$ for $x \in E$ and $\delta_x(E) = 0$ for $x \notin E$ |
| $\tilde{\delta}_x$ | The complex-valued Dirac measure defined as $\tilde{\delta}_x(E) = 1$ for $x \in E$ and $\tilde{\delta}_x(E) = 0$ for $x \notin E$ |
| $\chi_E$ | The indicator function of a Borel set $E$ on $\mathcal{X}$ |
| $C_0(\mathcal{X}, \mathcal{A})$ | The space of all continuous $\mathcal{A}$-valued functions on $\mathcal{X}$ vanishing at infinity |
| $\mathbf{G}$ | The $\mathcal{A}$-valued Gram matrix defined as $\mathbf{G}_{i,j} = k(x_i, x_j)$ for given samples $x_1, \ldots, x_n \in \mathcal{X}$ |
| $p_j$ | The $j$-th principal axis generated by kernel PCA with an RKHM |
| $r$ | A natural number that represents the number of principal axes |
| $Df_{\mathbf{c}}$ | The Gâteaux derivative of a function $f : \mathcal{M} \to \mathcal{A}$ at $\mathbf{c} \in \mathcal{M}$ |
| $\nabla f_{\mathbf{c}}$ | The gradient of a function $f : \mathcal{M} \to \mathcal{A}$ at $\mathbf{c} \in \mathcal{M}$ |

### 2.1 Reproducing kernel Hilbert space (RKHS)

We review the theory of RKHS. An RKHS is a Hilbert space and useful for extracting non-linearity or higher-order moments of data (Schölkopf and Smola, 2001; Saitoh and Sawano, 2016).

We begin by introducing positive definite kernels. Let $\mathcal{X}$ be a non-empty set for data, and $\tilde{k}$ be a positive definite kernel, which is defined as follows:

**Definition 2.1 (Positive definite kernel)** *A map $\tilde{k} : \mathcal{X} \times \mathcal{X} \to \mathbb{C}$ is called a* positive definite kernel *if it satisfies the following conditions:*

1. $\tilde{k}(x, y) = \overline{\tilde{k}(y, x)}$ *for* $x, y \in \mathcal{X}$,
2. $\sum_{i,j=1}^{n} \overline{\alpha}_i \alpha_j \tilde{k}(x_i, x_j) \geq 0$ *for* $n \in \mathbb{N}$, $\alpha_i \in \mathbb{C}$, $x_i \in \mathcal{X}$.

Let $\tilde{\phi} : \mathcal{X} \to \mathbb{C}^{\mathcal{X}}$ be a map defined as $\tilde{\phi}(x) = \tilde{k}(\cdot, x)$. With $\tilde{\phi}$, the following space as a subset of $\mathbb{C}^{\mathcal{X}}$ is constructed:

$$\mathcal{H}_{\tilde{k},0} := \left\{ \sum_{i=1}^{n} \alpha_i \tilde{\phi}(x_i) \middle| \ n \in \mathbb{N},\ \alpha_i \in \mathbb{C},\ x_i \in \mathcal{X} \right\}.$$

Then, a map $\langle \cdot, \cdot \rangle_{\mathcal{H}_{\tilde{k}}} : \mathcal{H}_{\tilde{k},0} \times \mathcal{H}_{\tilde{k},0} \to \mathbb{C}$ is defined as follows:

$$\left\langle \sum_{i=1}^{n} \alpha_i \tilde{\phi}(x_i), \sum_{j=1}^{l} \beta_j \tilde{\phi}(y_j) \right\rangle_{\mathcal{H}_{\tilde{k}}} := \sum_{i=1}^{n} \sum_{j=1}^{l} \overline{\alpha_i} \beta_j \tilde{k}(x_i, y_j).$$

By the properties in Definition 2.1 of $\tilde{k}$, $\langle \cdot, \cdot \rangle_{\mathcal{H}_{\tilde{k}}}$ is well-defined, satisfies the axiom of inner products, and has the reproducing property, that is,

$$\langle \tilde{\phi}(x), v \rangle_{\mathcal{H}_{\tilde{k}}} = v(x)$$

for $v \in \mathcal{H}_{\tilde{k},0}$ and $x \in \mathcal{X}$.

The completion of $\mathcal{H}_{\tilde{k},0}$ is called the *RKHS* associated with $\tilde{k}$ and denoted as $\mathcal{H}_{\tilde{k}}$. It can be shown that $\langle \cdot, \cdot \rangle_{\mathcal{H}_{\tilde{k}}}$ is extended continuously to $\mathcal{H}_{\tilde{k}}$ and the map $\mathcal{H}_{\tilde{k}} \ni v \mapsto (x \mapsto \langle \tilde{\phi}(x), v \rangle_{\mathcal{H}_{\tilde{k}}}) \in \mathbb{C}^{\mathcal{X}}$ is injective. Thus, $\mathcal{H}_{\tilde{k}}$ is regarded to be a subset of $\mathbb{C}^{\mathcal{X}}$ and has the reproducing property. Also, $\mathcal{H}_{\tilde{k}}$ is determined uniquely.

The map $\tilde{\phi}$ maps data into $\mathcal{H}_{\tilde{k}}$ and is called the *feature map*. Since the dimension of $\mathcal{H}_{\tilde{k}}$ is higher (often infinite dimensional) than that of $\mathcal{X}$, complicated behaviors of data in $\mathcal{X}$ are expected to be transformed into simple ones in $\mathcal{H}_{\tilde{k}}$ (Schölkopf and Smola, 2001).

### 2.2 Vector-valued RKHS (vvRKHS)

We review the theory of vvRKHS. Complex-valued functions in RKHSs are generalized to vector-valued functions in vvRKHSs. Similar to the case of RKHS, we begin by introducing positive definite kernels. Let $\mathcal{X}$ be a non-empty set for data and $\mathcal{W}$ be a Hilbert space. In addition, let $k$ be an operator-valued positive definite kernel, which is defined as follows:

**Definition 2.2 (Operator-valued positive definite kernel)** *A map $k : \mathcal{X} \times \mathcal{X} \to \mathcal{B}(\mathcal{W})$ is called an* operator-valued positive definite kernel *if it satisfies the following conditions:*

1. $k(x, y) = k(y, x)^*$ *for* $x, y \in \mathcal{X}$,
2. $\sum_{i,j=1}^{n} \langle w_i, k(x_i, x_j) w_j \rangle_{\mathcal{W}} \geq 0$ *for* $n \in \mathbb{N}$, $w_i \in \mathcal{W}$, $x_i \in \mathcal{X}$.

*Here, $^*$ represents the adjoint.*

Let $\phi : \mathcal{X} \to \mathcal{B}(\mathcal{W})^{\mathcal{X}}$ be a map defined as $\phi(x) = k(\cdot, x)$. With $\phi$, the following space as a subset of $\mathcal{W}^{\mathcal{X}}$ is constructed:

$$\mathcal{H}_{k,0}^{\mathrm{v}} := \left\{ \sum_{i=1}^{n} \phi(x_i) w_i \,\middle|\, n \in \mathbb{N},\ w_i \in \mathcal{W},\ x_i \in \mathcal{X} \right\}.$$

Then, a map $\langle \cdot, \cdot \rangle_{\mathcal{H}_k^{\mathrm{v}}} : \mathcal{H}_{k,0}^{\mathrm{v}} \times \mathcal{H}_{k,0}^{\mathrm{v}} \to \mathbb{C}$ is defined as follows:

$$\left\langle \sum_{i=1}^{n} \phi(x_i) w_i, \sum_{j=1}^{l} \phi(y_j) h_j \right\rangle_{\mathcal{H}_k^{\mathrm{v}}} := \sum_{i=1}^{n} \sum_{j=1}^{l} \langle w_i, k(x_i, y_j) h_j \rangle_{\mathcal{W}}.$$

By the properties in Definition 2.2 of $k$, $\langle \cdot, \cdot \rangle_{\mathcal{H}_k^{\mathrm{v}}}$ is well-defined, satisfies the axiom of inner products, and has the reproducing property, that is,

$$\langle \phi(x) w, u \rangle_{\mathcal{H}_k^{\mathrm{v}}} = \langle w, u(x) \rangle_{\mathcal{W}} \tag{1}$$

for $u \in \mathcal{H}_{k,0}^{\mathrm{v}}$, $x \in \mathcal{X}$, and $w \in \mathcal{W}$.

The completion of $\mathcal{H}_{k,0}^{\mathrm{v}}$ is called the *vvRKHS* associated with $k$ and denoted as $\mathcal{H}_k^{\mathrm{v}}$. Note that since an inner product in $\mathcal{H}_k^{\mathrm{v}}$ is defined with the complex-valued inner product in $\mathcal{W}$, it is complex-valued.

### 2.3 $C^*$-algebra and Hilbert $C^*$-module

A $C^*$-algebra and a $C^*$-module are generalizations of the space of complex numbers $\mathbb{C}$ and a vector space, respectively. In this paper, we denote a $C^*$-algebra by $\mathcal{A}$ and a $C^*$-module by $\mathcal{M}$, respectively. As we see below, many complex-valued notions can be generalized to $\mathcal{A}$-valued.

A $C^*$-algebra is defined as a Banach space equipped with a product structure and an involution $(\cdot)^* : \mathcal{A} \to \mathcal{A}$. We denote the norm of $\mathcal{A}$ by $\| \cdot \|_{\mathcal{A}}$.

**Definition 2.3 (Algebra)** *A set $\mathcal{A}$ is called an* algebra *on a filed $\mathbb{F}$ if it is a vector space equipped with an operation $\cdot : \mathcal{A} \times \mathcal{A} \to \mathcal{A}$ which satisfies the following conditions for $b, c, d \in \mathcal{A}$ and $\alpha \in \mathbb{F}$:*

- $(b + c) \cdot d = c \cdot d + c \cdot d$,
- $b \cdot (c + d) = b \cdot c + b \cdot d$,
- $(\alpha c) \cdot d = \alpha (c \cdot d) = c \cdot (\alpha d)$.

*The symbol $\cdot$ is omitted when it does not cause confusion.*

**Definition 2.4 ($C^*$-algebra)** *A set $\mathcal{A}$ is called a* $C^*$-algebra *if it satisfies the following conditions:*

1. *$\mathcal{A}$ is an algebra over $\mathbb{C}$, and there exists a bijection $(\cdot)^* : \mathcal{A} \to \mathcal{A}$ that satisfies the following conditions for $\alpha, \beta \in \mathbb{C}$ and $c, d \in \mathcal{A}$:*
   - $(\alpha c + \beta d)^* = \overline{\alpha} c^* + \overline{\beta} d^*$, $\quad$ - $(cd)^* = d^* c^*$, $\quad$ - $(c^*)^* = c$.
2. *$\mathcal{A}$ is a normed space with $\| \cdot \|_{\mathcal{A}}$, and for $c, d \in \mathcal{A}$, $\|cd\|_{\mathcal{A}} \leq \|c\|_{\mathcal{A}} \|d\|_{\mathcal{A}}$ holds. In addition, $\mathcal{A}$ is complete with respect to $\| \cdot \|_{\mathcal{A}}$.*
3. *For $c \in \mathcal{A}$, $\|c^* c\|_{\mathcal{A}} = \|c\|_{\mathcal{A}}^2$ holds.*

**Definition 2.5 (Multiplicative identity and unital $C^*$-algebra)** *The* multiplicative identity *of $\mathcal{A}$ is the element $a \in \mathcal{A}$ which satisfies $ac = ca = c$ for any $c \in \mathcal{A}$. We denote by $1_{\mathcal{A}}$ the multiplicative identity of $\mathcal{A}$. If a $C^*$-algebra $\mathcal{A}$ has the multiplicative identity, then it is called a* unital $C^*$-algebra.

**Example 2.6** *Important examples of (unital) $C^*$-algebras are $L^\infty(\Omega)$ and $\mathcal{B}(\mathcal{W})$, i.e., the space of complex-valued $L^\infty$ functions on a compact measure space $\Omega$ and the space of bounded linear operators on a Hilbert space $\mathcal{W}$, respectively.*

1. *For $\mathcal{A} = L^\infty(\Omega)$, the product of two functions $c, d \in \mathcal{A}$ is defined as $(cd)(t) = c(t)d(t)$ for any $t \in \Omega$, the involution is defined as $c(t) = \overline{c(t)}$, the norm is the $L^\infty$-norm, and the multiplicative identity is the constant function whose value is $1$ at almost everywhere $t \in \Omega$.*
2. *For $\mathcal{A} = \mathcal{B}(\mathcal{W})$, the product structure is the product (the composition) of operators, the involution is the adjoint, the norm $\| \cdot \|_{\mathcal{A}}$ is the operator norm, and the multiplicative identity is the identity map.*

In fact, by the Gelfand–Naimark theorem (see, for example, Murphy (1990)), any $C^*$-algebra can be regarded as a subalgebra of $\mathcal{B}(\mathcal{W})$ for some Hilbert space $\mathcal{W}$. Therefore, considering the case of $\mathcal{A} = \mathcal{B}(\mathcal{W})$ is sufficient for applications.

*The positiveness* is also important in $C^*$-algebras.

**Definition 2.7 (Positive)** *An element $c$ of $\mathcal{A}$ is called* positive *if there exists $d \in \mathcal{A}$ such that $c = d^* d$ holds. For a unital $C^*$-algebra $\mathcal{A}$, if a positive element $c \in \mathcal{A}$ is invertible, i.e., there exists $d \in \mathcal{A}$ such that $cd = dc = 1_{\mathcal{A}}$, then $c$ is called* strictly positive. *For $c, d \in \mathcal{A}$, we denote $c \leq_{\mathcal{A}} d$ if $d - c$ is positive and $c <_{\mathcal{A}} d$ if $d - c$ is strictly positive. We denote by $\mathcal{A}_+$ the subset of $\mathcal{A}$ composed of all positive elements in $\mathcal{A}$.*

**Example 2.8**
1. *For $\mathcal{A} = L^\infty(\Omega)$, a function $c \in \mathcal{A}$ is positive if and only if $c(t) \geq 0$ for almost everywhere $t \in \Omega$, and strictly positive if and only if $c(t) > 0$ for almost everywhere $t \in \Omega$.*
2. *For $\mathcal{A} = \mathcal{B}(\mathcal{W})$, the positiveness is equivalent to the positive semi-definiteness of operators and the strictly positiveness is equivalent to the positive definiteness of operators.*

The positiveness provides us the (pre) order in $\mathcal{A}$ and, thus, enables us to consider optimization problems in $\mathcal{A}$.

**Definition 2.9 (Supremum and infimum)** *1. For a subset $\mathcal{S}$ of $\mathcal{A}$, $a \in \mathcal{A}$ is said to be an* upper bound *with respect to the order $\leq_{\mathcal{A}}$, if $d \leq_{\mathcal{A}} a$ for any $d \in \mathcal{S}$. Then, $c \in \mathcal{A}$ is said to be a* supremum *of $\mathcal{S}$, if $c \leq_{\mathcal{A}} a$ for any upper bound $a$ of $\mathcal{S}$.*

*2. For a subset $\mathcal{S}$ of $\mathcal{A}$, $a \in \mathcal{A}$ is said to be a* lower bound *with respect to the order $\leq_{\mathcal{A}}$, if $a \leq_{\mathcal{A}} d$ for any $d \in \mathcal{S}$. Then, $c \in \mathcal{A}$ is said to be a* infimum *of $\mathcal{S}$, if $a \leq_{\mathcal{A}} c$ for any lower bound $a$ of $\mathcal{S}$.*

We now introduce a $C^*$-module over $\mathcal{A}$, which is a generalization of the vector space.

**Definition 2.10 (Right multiplication)** *Let $\mathcal{M}$ be an abelian group with operation $+$. For $c, d \in \mathcal{A}$ and $u, v \in \mathcal{M}$, if an operation $\cdot : \mathcal{M} \times \mathcal{A} \to \mathcal{M}$ satisfies*

1. $(u + v) \cdot c = u \cdot c + v \cdot c$,
2. $u \cdot (c + d) = u \cdot c + u \cdot d$,
3. $u \cdot (cd) = (u \cdot d) \cdot c$,
4. $u \cdot 1_{\mathcal{A}} = u$ *if $\mathcal{A}$ is unital,*

*then, $\cdot$ is called a (right)* $\mathcal{A}$-multiplication. *The multiplication $u \cdot c$ is usually denoted as $uc$.*

**Definition 2.11 ($C^*$-module)** *Let $\mathcal{M}$ be an abelian group with operation $+$. If $\mathcal{M}$ has the structure of a (right) $\mathcal{A}$-multiplication, $\mathcal{M}$ is called a (right) $C^*$-*module *over $\mathcal{A}$.*

In this paper, we consider column vectors rather than row vectors for representing $\mathcal{A}$-valued coefficients, and column vectors act on the right. Therefore, we consider right multiplications. However, considering row vectors and left multiplications instead of column vectors and right multiplications is also possible.

### 2.4 Hilbert $C^*$-module

A Hilbert $C^*$-module is a generalization of a Hilbert space. We first consider an $\mathcal{A}$-valued inner product, which is a generalization of a complex-valued inner product, and then, introduce the definition of a Hilbert $C^*$-module.

**Definition 2.12 ($\mathcal{A}$-valued inner product)** *A map $\langle \cdot, \cdot \rangle_{\mathcal{M}} : \mathcal{M} \times \mathcal{M} \to \mathcal{A}$ is called an $\mathcal{A}$-valued* inner product *if it satisfies the following properties for $u, v, p \in \mathcal{M}$ and $c, d \in \mathcal{A}$:*

1. $\langle u, vc + pd \rangle_{\mathcal{M}} = \langle u, v \rangle_{\mathcal{M}} c + \langle u, p \rangle_{\mathcal{M}} d$,
2. $\langle v, u \rangle_{\mathcal{M}} = \langle u, v \rangle_{\mathcal{M}}^*$,
3. $\langle u, u \rangle_{\mathcal{M}} \geq_{\mathcal{A}} 0$,
4. *If* $\langle u, u \rangle_{\mathcal{M}} = 0$ *then* $u = 0$.

**Definition 2.13 ($\mathcal{A}$-valued absolute value and norm)** *For $u \in \mathcal{M}$, the* $\mathcal{A}$-valued absolute value *$|u|_{\mathcal{M}}$ on $\mathcal{M}$ is defined by the positive element $|u|_{\mathcal{M}}$ of $\mathcal{A}$ such that $|u|_{\mathcal{M}}^2 = \langle u, u \rangle_{\mathcal{M}}$. The (real-valued) norm $\| \cdot \|_{\mathcal{M}}$ on $\mathcal{M}$ is defined by $\|u\|_{\mathcal{M}} = \big\| |u|_{\mathcal{M}} \big\|_{\mathcal{A}}$.*

Since the absolute value $|\cdot|_{\mathcal{M}}$ takes values in $\mathcal{A}$, it behaves more complicatedly. For example, the triangle inequality does not hold for the absolute value. However, it provides us with more information than the norm $\|\cdot\|_{\mathcal{M}}$ (which is real-valued). For example, let $\mathcal{M} = \mathcal{A} = \mathbb{C}^{m\times m}$, $c = \mathrm{diag}\{\alpha, 0, \ldots, 0\}$, and $d = \mathrm{diag}\{\alpha, \ldots, \alpha\}$, where $\alpha \in \mathbb{C}$. Then, $\|c\|_{\mathcal{M}} = \|d\|_{\mathcal{M}}$, but $|c|_{\mathcal{M}} \neq |d|_{\mathcal{M}}$. For a self-adjoint matrix, the absolute value describes the whole spectrum of it, but the norm only describes the largest eigenvalue.

**Definition 2.14 (Hilbert $C^*$-module)** *Let $\mathcal{M}$ be a (right) $C^*$-module over $\mathcal{A}$ equipped with an $\mathcal{A}$-valued inner product defined in Definition 2.12. If $\mathcal{M}$ is complete with respect to the norm $\|\cdot\|_{\mathcal{M}}$, it is called a* Hilbert $C^*$-module *over $\mathcal{A}$ or* Hilbert $\mathcal{A}$-module.

**Example 2.15** *A simple example of Hilbert $C^*$ modules over $\mathcal{A}$ is $\mathcal{A}^n$ for a natural number $n$. The $\mathcal{A}$-valued inner product between $\mathbf{c} = [c_1, \ldots, c_n]^T$ and $\mathbf{d} = [d_1, \ldots, d_n]^T$ is defined as $\langle \mathbf{c}, \mathbf{d}\rangle_{\mathcal{A}^n} = \sum_{i=1}^n c_i^* d_i$. The absolute value and norm in $\mathcal{A}^n$ are given as $|\mathbf{c}|^2_{\mathcal{A}^n} = (\sum_{i=1}^n c_i^* c_i)$ and $\|\mathbf{c}\|_{\mathcal{A}^n} = \|\sum_{i=1}^n c_i^* c_i\|_{\mathcal{A}}^{1/2}$, respectively.*

Similar to the case of Hilbert spaces, the following Cauchy–Schwarz inequality for $\mathcal{A}$-valued inner products is available (Lance, 1995, Proposition 1.1).

**Lemma 2.16 (Cauchy–Schwarz inequality)** *For $u, v \in \mathcal{M}$, the following inequality holds:*

$$|\langle u, v\rangle_{\mathcal{M}}|^2_{\mathcal{A}} \leq_{\mathcal{A}} \|u\|^2_{\mathcal{M}} \langle v, v\rangle_{\mathcal{M}}.$$

An important property associated with an inner product is the orthonormality. The orthonormality plays an important role in data analysis. For example, an orthonormal basis constructs orthogonal projections and an orthogonally projected vector minimizes the deviation from its original vector in the projected space. Therefore, we also introduce the orthonormality in Hilbert $C^*$-module. See, for example, Definition 1.2 in (Bakić and Guljaš, 2001) for more details.

**Definition 2.17 (Normalized)** *A vector $q \in \mathcal{M}$ is* normalized *if $0 \neq \langle q, q\rangle_{\mathcal{M}} = \langle q, q\rangle^2_{\mathcal{M}}$.*

Note that in the case of a general $C^*$-valued inner product, for a normalized vector $q$, $\langle q, q\rangle_{\mathcal{M}}$ is not always equal to the identity of $\mathcal{A}$ in contrast to the case of a complex-valued inner product.

**Definition 2.18 (Orthonormal system and basis)** *Let $\mathcal{I}$ be an index set. A set $\mathcal{S} = \{q_i\}_{i\in\mathcal{I}} \subseteq \mathcal{M}$ is called an* orthonormal system (ONS) *of $\mathcal{M}$ if $q_i$ is normalized for any $i \in \mathcal{I}$ and $\langle q_i, q_j\rangle_{\mathcal{M}} = 0$ for $i \neq j$. We call $\mathcal{S}$ an* orthonormal basis (ONB) *if $\mathcal{S}$ is an ONS and dense in $\mathcal{M}$.*

In Hilbert $C^*$-modules, $\mathcal{A}$-linear is often used instead of $\mathbb{C}$-linear.

**Definition 2.19 ($\mathcal{A}$-linear operator)** *Let $\mathcal{M}_1, \mathcal{M}_2$ be Hilbert $\mathcal{A}$-modules. A linear map $L : \mathcal{M}_1 \to \mathcal{M}_2$ is referred to as* $\mathcal{A}$-linear *if it satisfies $L(uc) = (Lu)c$ for any $u \in \mathcal{M}$ and $c \in \mathcal{A}$.*

**Definition 2.20 ($\mathcal{A}$-linearly independent)** *The set $\mathcal{S}$ of $\mathcal{M}$ is said to be* $\mathcal{A}$-linearly independent *if it satisfies the following condition: For any finite subset $\{v_1, \ldots, v_n\}$ of $\mathcal{S}$, if $\sum_{i=1}^n v_i c_i = 0$ for $c_i \in \mathcal{A}$, then $c_i = 0$ for $i = 1, \ldots, n$.*

For further details about $C^*$-algebra, $C^*$-module, and Hilbert $C^*$-module, refer to Murphy (1990); Lance (1995).

### 2.5 Reproducing kernel Hilbert $C^*$-module (RKHM)

We summarize the theory of RKHM, which is discussed, for example, in Heo (2008).

Similar to the case of RKHS, we begin by introducing an $\mathcal{A}$-valued generalization of a positive definite kernel on a non-empty set $\mathcal{X}$ for data.

**Definition 2.21 ($\mathcal{A}$-valued positive definite kernel)** *An $\mathcal{A}$-valued map $k: \mathcal{X} \times \mathcal{X} \to \mathcal{A}$ is called a* positive definite kernel *if it satisfies the following conditions:*

1. $k(x, y) = k(y, x)^*$ *for* $x, y \in \mathcal{X}$,
2. $\sum_{i,j=1}^n c_i^* k(x_i, x_j) c_j \geq_{\mathcal{A}} 0$ *for* $n \in \mathbb{N}$, $c_i \in \mathcal{A}$, $x_i \in \mathcal{X}$.

**Example 2.22** *1. Let $\mathcal{X} = C([0,1]^m)$. Let $\mathcal{A} = L^\infty([0,1])$ and let $k : \mathcal{X} \times \mathcal{X} \to \mathcal{A}$ be defined as $k(x,y)(t) = \int_{[0,1]^m} (t - x(s))(t - y(s))ds$ for $t \in [0,1]$. Then, for $x_1, \ldots, x_n \in \mathcal{X}$, $c_1, \ldots, c_n \in \mathcal{A}$ and $t \in [0,1]$, we have*

$$\begin{aligned}\sum_{i,j=1}^n c_i^*(t) k(x_i, x_j)(t) c_j(t) &= \int_{[0,1]^m} \sum_{i,j=1}^n \overline{c_i(t)}(t - x_i(s))(t - x_j(s)) c_j(t) ds \\ &= \int_{[0,1]^m} \sum_{i=1}^n \overline{c_i(t)}(t - x_i(s)) \sum_{j=1}^n (t - x_j(s)) c_j(t) ds \geq 0\end{aligned}$$

*for $t \in [0,1]$. Thus, $k$ is an $\mathcal{A}$-valued positive definite kernel.*

2. *Let $\mathcal{A} = L^\infty([0,1])$ and $k : \mathcal{X} \times \mathcal{X} \to \mathcal{A}$ be defined such that $k(x,y)(t)$ is a complex-valued positive definite kernel for any $t \in [0,1]$. Then, $k$ is an $\mathcal{A}$-valued positive definite kernel.*

3. *Let $\mathcal{W}$ be a separable Hilbert space and let $\{e_i\}_{i=1}^\infty$ be an orthonormal basis of $\mathcal{W}$. Let $\mathcal{A} = \mathcal{B}(\mathcal{W})$ and let $k : \mathcal{X} \times \mathcal{X} \to \mathcal{A}$ be defined as $k(x,y)e_i = k_i(x,y)e_i$, where $k_i : \mathcal{X} \times \mathcal{X} \to \mathbb{C}$ is a complex-valued positive definite kernel for any $i = 1, 2, \ldots$. Then, for $x_1, \ldots, x_n \in \mathcal{X}$, $c_1, \ldots, c_n \in \mathcal{A}$ and $w \in \mathcal{W}$, we have*

$$\begin{aligned}\left\langle w, \left( \sum_{i,j=1}^n c_i^* k(x_i, x_j) c_j \right) w \right\rangle_{\mathcal{W}} &= \sum_{i,j=1}^n \sum_{l=1}^\infty \langle \alpha_{i,l} e_l, k(x_i, x_j) \alpha_{j,l} e_l \rangle_{\mathcal{W}} \\ &= \sum_{l=1}^\infty \sum_{i,j=1}^n \overline{\alpha_{i,l}} \alpha_{j,l} \tilde{k}_l(x_i, x_j) \geq 0,\end{aligned}$$

*where $c_i w = \sum_{l=1}^\infty \alpha_{i,l} e_l$ is the expansion with respect to $\{e_i\}_{i=1}^\infty$. Thus, $k$ is an $\mathcal{A}$-valued positive definite kernel.*

4. *Let $\mathcal{X} = C(\Omega, \mathcal{Y})$ and $\mathcal{W} = L^2(\Omega)$ for a compact measure space $\Omega$ and a topological space $\mathcal{Y}$. Let $\mathcal{A} = \mathcal{B}(\mathcal{W})$, and $\tilde{k} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{C}$ be a complex-valued continuous positive definite kernel. Moreover, let $k : \mathcal{X} \times \mathcal{X} \to \mathcal{A}$ be defined as $(k(x,y)w)(s) =$*

$\int_{t\in\Omega} \tilde{k}(x(s), y(t))w(t)dt$. *Then, for* $x_1, \ldots, x_n \in \mathcal{X}$, $c_1, \ldots, c_n \in \mathcal{A}$ *and* $w \in \mathcal{W}$, *we have*

$$\left\langle w, \left( \sum_{i,j=1}^{n} c_i^* k(x_i, x_j) c_j \right) w \right\rangle_{\mathcal{W}} = \int_{t\in\Omega} \int_{s\in\Omega} \sum_{i,j=1}^{n} \overline{d_i(s)} \tilde{k}(x_i(s), x_j(t)) d_j(t) ds dt \geq 0,$$

*where* $d_i = c_i w$. *Thus,* $k$ *is an* $\mathcal{A}$*-valued positive definite kernel.*

Let $\phi : \mathcal{X} \to \mathcal{A}^{\mathcal{X}}$ be the *feature map* associated with $k$, which is defined as $\phi(x) = k(\cdot, x)$ for $x \in \mathcal{X}$. Similar to the case of RKHS, we construct the following $C^*$-module composed of $\mathcal{A}$-valued functions by means of $\phi$:

$$\mathcal{M}_{k,0} := \left\{ \sum_{i=1}^{n} \phi(x_i) c_i \,\middle|\, n \in \mathbb{N},\ c_i \in \mathcal{A},\ x_i \in \mathcal{X} \right\}.$$

An $\mathcal{A}$-valued map $\langle \cdot, \cdot \rangle_{\mathcal{M}_k} : \mathcal{M}_{k,0} \times \mathcal{M}_{k,0} \to \mathcal{A}$ is defined as follows:

$$\left\langle \sum_{i=1}^{n} \phi(x_i) c_i, \sum_{j=1}^{l} \phi(y_j) d_j \right\rangle_{\mathcal{M}_k} := \sum_{i=1}^{n} \sum_{j=1}^{l} c_i^* k(x_i, y_j) d_j.$$

By the properties in Definition 2.21 of $k$, $\langle \cdot, \cdot \rangle_{\mathcal{M}_k}$ is well-defined and has the reproducing property

$$\langle \phi(x), v \rangle_{\mathcal{M}_k} = v(x)$$

for $v \in \mathcal{M}_{k,0}$ and $x \in \mathcal{X}$. Also, it satisfies the properties in Definition 2.12. As a result, $\langle \cdot, \cdot \rangle_{\mathcal{M}_k}$ is shown to be an $\mathcal{A}$-valued inner product.

The *reproducing kernel Hilbert* $\mathcal{A}$*-module (RKHM)* associated with $k$ is defined as the completion of $\mathcal{M}_{k,0}$. We denote by $\mathcal{M}_k$ the RKHM associated with $k$.

Heo (2008) focused on the case where a group acts on $\mathcal{X}$ and investigated corresponding actions on RKHMs. Moreover, he considered the space of operators on Hilbert $\mathcal{A}$-module and proved that for each operator-valued positive definite kernel associated with a group and cocycle, there is a corresponding representation on the Hilbert $C^*$-module associated with the positive definite kernel.

## 3. Application of RKHM to functional data

In this section, we provide an overview of the motivation for studying RKHM for data analysis. We especially focus on the application of RKHM to functional data.

Analyzing functional data has been researched to take advantage of the additional information implied by the smoothness of functions underlying data (Ramsay and Silverman, 2005; Levitin et al., 2007; Wang et al., 2016). By describing data as functions, we obtain information as functions such as derivatives. Applying kernel methods to functional data is also proposed (Kadri et al., 2016). In these frameworks, the functions are assumed to be vectors in a Hilbert space such as $L^2(\Omega)$ for a measure space $\Omega$, or they are embedded in an RKHS or vvRKHS. Then, analyses are addressed in these Hilbert spaces.

However, since functional data itself is infinite-dimensional data, Hilbert spaces are not always sufficient for extracting its continuous behavior. This is because the inner products

in Hilbert spaces are complex-valued, degenerating or failing to capture the continuous behavior of the functional data. We compare algorithms in Hilbert spaces and those in Hilbert $C^*$-modules and show advantages of algorithms in Hilbert $C^*$-modules over those in Hilbert spaces, which are summarized in Figure 1. We first consider algorithms in Hilbert spaces for analyzing functional data $x_1, x_2, \ldots \in C(\Omega, \mathcal{X})$, where $\Omega$ is a compact measure space and $\mathcal{X}$ is a Hilbert space. There are two possible typical patterns of algorithms in Hilbert spaces. The first pattern (Pattern 1 in Fig. 1) is regarding each function $x_i$ as a vector in a Hilbert space $\mathcal{H}$ containing $C(\Omega, \mathcal{X})$. In this case, the inner product $\langle x_i, x_j \rangle_{\mathcal{H}}$ between two functions $x_i$ and $x_j$ is single complex-valued although $x_i$ and $x_j$ are functions. Therefore, information of the value of functions at each point degenerates into a complex value. The second pattern (Pattern 2 in Fig. 1) is discretizing each function $x_i$ as $x_i(t_0), x_i(t_1), \ldots$ for $t_0, t_1, \ldots \in \Omega$ and regarding each discretized value $x_i(t_l)$ as a vector in the Hilbert space $\mathcal{X}$. In this case, we obtain the complex-valued inner product $\langle x_i(t_l), x_j(t_l) \rangle_{\mathcal{X}}$ at each point $t_l \in \Omega$. However, because of the discretization, continuous behaviors, for example, derivatives, total variation, and frequency components, of the function $x_i$ are lost. Algorithms of both patterns in the Hilbert spaces proceed by using the computed complex-valued inner products. As a result, capturing features of functions with the algorithms in the Hilbert spaces is difficult. On the other hand, if we regard each function $x_i$ as a vector in a Hilbert $C^*$-module $\mathcal{M}$ (the rightmost picture in Fig. 1), then the inner product $\langle x_i, x_j \rangle_{\mathcal{M}}$ between two functions $x_i$ and $x_j$ in the Hilbert $C^*$-module is $C^*$-algebra-valued. Thus, if we set the $C^*$-algebra as a function space such as $L^\infty(\Omega)$, the inner product $\langle x_i, x_j \rangle_{\mathcal{M}}$ is function-valued. Therefore, algorithms in Hilbert $C^*$-modules enable us to capture and extract continuous behaviors of functions. Moreover, in the case of the outputs are functions, we can control the outputs according to the features of the functions.

Since RKHM is a generalization of RKHS and vvRKHS (see Subsection 4.2 for further details), the framework of RKHMs (Hilbert $C^*$-modules) allows us to generalize kernel methods in RKHSs and vvRKHSs (Hilbert spaces) to those in Hilbert $C^*$-modules. Therefore, by using RKHM, we can capture and extract features of functions in kernel methods. The remainder of this paper is devoted to developing the theory of applying RKHMs to data analysis and showing examples of practical applications of data analysis in RKHMs (PCA, time-series data analysis, and analysis of interaction effects).

## 4. RKHM for data analysis

As we mentioned in Section 1, RKHM has been studied in mathematical physics and pure mathematics. In existing studies, mathematical properties of RKHM such as the relationship between group actions and RKHMs (see the last paragraph of Subsection 2.5) have been discussed. However, these studies have not been focused on data and algorithms for analyzing it. Therefore, we fill the gaps between the existing theory of RKHM and its application to data analysis in this section. We develop theories for the validity to applying it to data analysis in Subsection 4.1. Also, we investigate the connection of RKHM with RKHS and vvRKHS in Subsection 4.2.

Generalizations of theories of Hilbert space and RKHS are quite nonobvious for general $C^*$-algebras since fundamental properties in Hilbert spaces such as the Riesz representation theorem and orthogonal complementedness are not always obtained in Hilbert $C^*$-modules.

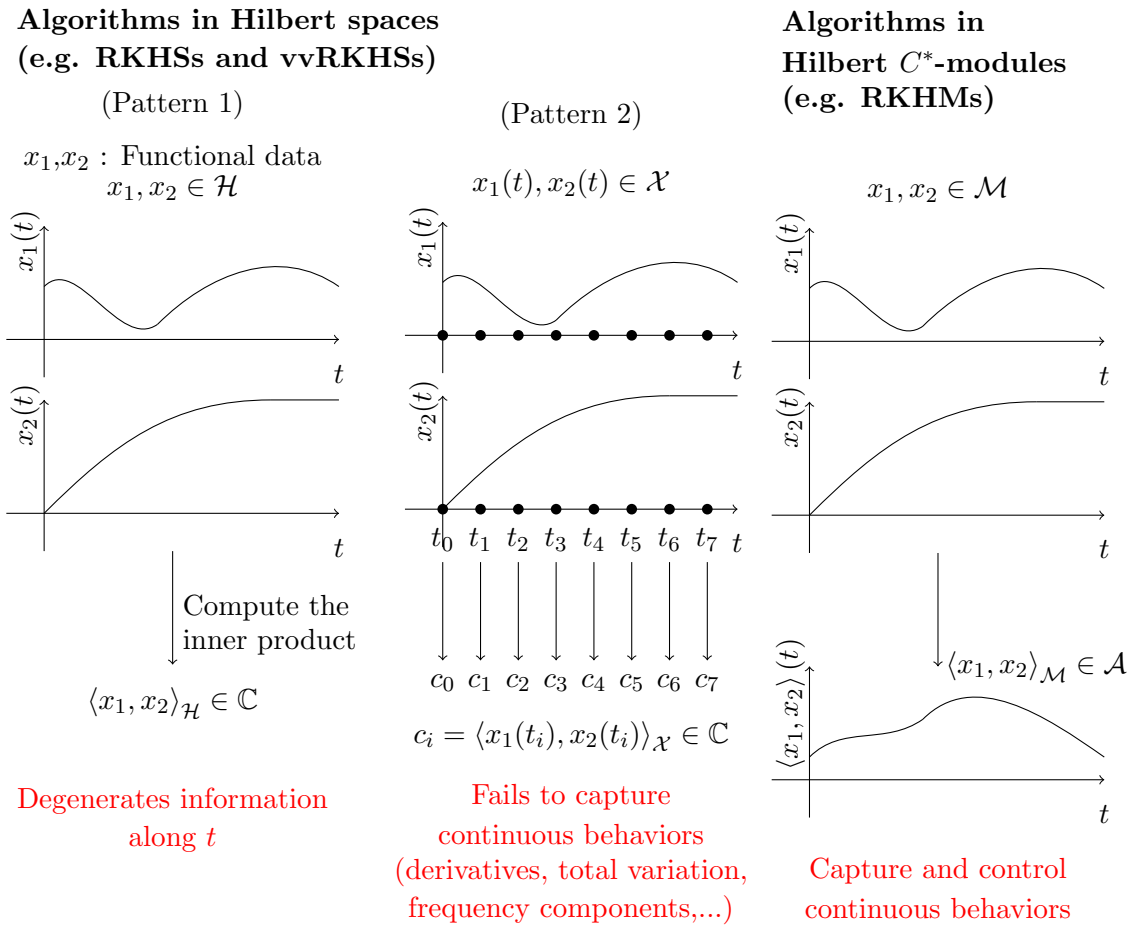



Figure 1: Advantages of algorithms in Hilbert $C^*$-modules over those in Hilbert spaces

Therefore, we consider limiting $C^*$-algebras to an appropriate class of $C^*$-algebras. In fact, von Neumann-algebras satisfy desired properties.

**Definition 4.1 (von Neumann-algebra)** *A $C^*$-algebra $\mathcal{A}$ is called a von Neumann-algebra if $\mathcal{A}$ is isomorphic to the dual Banach space of some Banach space.*

The following propositions are fundamental for deriving useful properties for data analysis in Hilbert $C^*$-modules and RKHMs (Skeide, 2000, Theorem 4.16), (Manuilov and Troitsky, 2000, Proposition 2.5.4).

**Proposition 4.2 (The Riesz representation theorem for Hilbert $\mathcal{A}$-modules)** *Let $\mathcal{A}$ be a von Neumann algebra. Let $\mathcal{H} = \mathcal{M} \otimes_{\mathcal{B}(\mathcal{W})} \mathcal{W}$ (see Definition 4.12 for the definition of the product $\mathcal{M} \otimes_{\mathcal{B}(\mathcal{W})} \mathcal{W}$). Then, every $v \in \mathcal{M}$ can be regarded as an operator in $\mathcal{B}(\mathcal{W}, \mathcal{H})$, the set of bounded linear operators from $\mathcal{W}$ to $\mathcal{H}$. If $\mathcal{M} \subseteq \mathcal{B}(\mathcal{W}, \mathcal{H})$ is strongly closed (in this case, we say that $\mathcal{M}$ is a von Neumann $\mathcal{A}$-module), then for a bounded $\mathcal{A}$-linear map $L : \mathcal{M} \to \mathcal{A}$ (see Definition 2.19), there exists a unique $u \in \mathcal{M}$ such that $Lv = \langle u, v \rangle_{\mathcal{M}}$ for all $v \in \mathcal{M}$.*

Let $\mathcal{A}$ be a von Neumann-algebra. We remark that the Hilbert $\mathcal{A}$-module $\mathcal{A}^n$ for some $n \in \mathbb{N}$ is a von Neumann $\mathcal{A}$-module. Moreover, for an $\mathcal{A}$-valued positive definite kernel defined as $\tilde{k}1_{\mathcal{A}}$, where $\tilde{k}$ is a (standard) positive definite kernel, the RKHM $\mathcal{M}_k$ is a von Neumann $\mathcal{A}$-module. (Generally, the Hilbert $\mathcal{A}$-module represented as $\mathcal{H} \otimes \mathcal{A}$ for a Hilbert space $\mathcal{H}$ is a von Neumann $\mathcal{A}$-module. Here, $\otimes$ represents the tensor product of a Hilbert space and $C^*$-module. See Lance (1995, p.6) for further details about the tensor product.)

**Proposition 4.3 (Orthogonal complementedness in Hilbert $\mathcal{A}$-modules)** *Let $\mathcal{A}$ be a von Neumann algebra and let $\mathcal{M}$ be a Hilbert $\mathcal{A}$-module. Let $\mathcal{V}$ be a closed submodule of $\mathcal{M}$. Then, any $u \in \mathcal{M}$ is decomposed into $u = u_1 + u_2$ where $u_1 \in \mathcal{V}$ and $u_2 \in \mathcal{V}^\perp$. Here, $\mathcal{V}^\perp$ is the orthogonal complement of $\mathcal{V}$ defined as $\{u \in \mathcal{M} \mid \langle u, v \rangle_{\mathcal{M}} = 0\}$.*

Therefore, we set $\mathcal{A}$ as a von Neumann-algebra to derive useful properties of RKHM for data analysis. Note that every von Neumann-algebra is unital (see Definitions 2.5).

**Assumption 4.4** *We assume $\mathcal{A}$ is a von Neumann-algebra throughout this paper.*

$C^*$-algebras in Example 2.6 are also von Neumann algebras. As we noted after Example 2.6, any $C^*$-algebra can be regarded as a subalgebra of $\mathcal{B}(\mathcal{W})$. Thus, this fact implies setting the range of the positive definite kernel as $\mathcal{B}(\mathcal{W})$ rather than general $C^*$-algebras is effective for data analysis.

## 4.1 General properties of RKHM for data analysis

### 4.1.1 Fundamental properties of RKHM

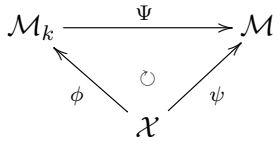
Similar to the cases of RKHSs, we show RKHMs constructed by $\mathcal{A}$-valued positive definite kernels have the reproducing property. Also, we show that the RKHM associated with an $\mathcal{A}$-valued positive definite kernel $k$ is uniquely determined.

**Proposition 4.5** *The map $\langle \cdot, \cdot \rangle_{\mathcal{M}_k}$ defined on $\mathcal{M}_{k,0}$ is extended continuously to $\mathcal{M}_k$ and the map $\mathcal{M}_k \ni v \mapsto (x \mapsto \langle \phi(x), v \rangle_{\mathcal{M}_k}) \in \mathcal{A}^{\mathcal{X}}$ is injective. Thus, $\mathcal{M}_k$ is regarded to be the subset of $\mathcal{A}^{\mathcal{X}}$ and has the reproducing property.*

**Proposition 4.6** *Assume a Hilbert $C^*$-module $\mathcal{M}$ over $\mathcal{A}$ and a map $\psi : \mathcal{X} \to \mathcal{M}$ satisfy the following conditions:*

1. $^{\forall} x, y \in \mathcal{X}$, $\langle \psi(x), \psi(y) \rangle_{\mathcal{M}} = k(x, y)$
2. $\overline{\{\sum_{i=1}^{n} \psi(x_i) c_i \mid \ x_i \in \mathcal{X},\ c_i \in \mathcal{A}\}} = \mathcal{M}$

*Then, there exists a unique $\mathcal{A}$-linear bijection map $\Psi : \mathcal{M}_k \to \mathcal{M}$ that preserves the inner product and satisfies the following commutative diagram:*




We give the proofs for the above propositions in Appendix A.

### 4.1.2 Minimization property and representer theorem in RKHMs

We now develop some theories for the validity to apply RKHM to data analysis. First, we show a minimization property of orthogonal projection operators, which is a fundamental property in Hilbert spaces, is also available in Hilbert $C^*$-modules.

**Theorem 4.7 (Minimization property of orthogonal projection operators)** *Let $\mathcal{I}$ be an index set. Let $\{q_i\}_{i \in \mathcal{I}}$ be an ONS of $\mathcal{M}$ and $\mathcal{V}$ be the completion of the space spanned by $\{q_i\}_{i \in \mathcal{I}}$. For $u \in \mathcal{M}_k$, let $P : \mathcal{M} \to \mathcal{V}$ be the projection operator defined as $Pu := \sum_{i \in \mathcal{I}} q_i \langle q_i, u \rangle_{\mathcal{M}}$. Then $Pu$ is the unique solution of the following minimization problem, where the minimum is taken with respect to a (pre) order in $\mathcal{A}$ (see Definition 2.9):*

$$\min_{v \in \mathcal{V}} |u - v|^2_{\mathcal{M}}. \tag{2}$$

**Proof** By Proposition 4.3, $u \in \mathcal{M}$ is decomposed into $u = u_1 + u_2$, where $u_1 = Pu \in \mathcal{V}$ and $u_2 = u - u_1 \in \mathcal{V}^{\perp}$. Let $v \in \mathcal{V}$. Since $u_1 - v \in \mathcal{V}$, the identity $\langle u_2, u_1 - v \rangle_{\mathcal{M}} = 0$ holds. Therefore, we have

$$|u - v|^2_{\mathcal{M}} = |u_2 + (u_1 - v)|^2_{\mathcal{M}} = |u_2|^2_{\mathcal{M}} + |u_1 - v|^2_{\mathcal{M}}, \tag{3}$$

which implies $|u - v|^2_{\mathcal{M}} - |u - u_1|^2_{\mathcal{M}} \geq_{\mathcal{A}} 0$. Since $v \in \mathcal{V}$ is arbitrary, $u_1$ is a solution of $\min_{v \in \mathcal{V}} |u - v|_{\mathcal{M}}$.

Moreover, if there exists $u' \in \mathcal{V}$ such that $|u - u_1|^2_{\mathcal{M}} = |u - u'|^2_{\mathcal{M}}$, then letting $v = u'$ in Eq. (3) derives $|u - u'|^2_{\mathcal{M}} = |u_2|^2_{\mathcal{M}} + |u_1 - u'|^2_{\mathcal{M}}$, which implies $|u_1 - u'|^2_{\mathcal{M}} = 0$. As a result, $u_1 = u'$ holds and the uniqueness of $u_1$ has been proved. ■

Proposition 4.7 shows the orthogonally projected vector uniquely minimizes the deviation from an original vector in $\mathcal{V}$. Thus, we can generalize methods related to orthogonal projections in Hilbert spaces to Hilbert $C^*$-modules.

Next, we show the representer theorem in RKHMs.

**Theorem 4.8 (Representer theorem)** *Let $x_1, \ldots, x_n \in \mathcal{X}$ and $a_1, \ldots, a_n \in \mathcal{A}$. Let $h : \mathcal{X} \times \mathcal{A}^2 \to \mathcal{A}_+$ be an error function and let $g : \mathcal{A}_+ \to \mathcal{A}_+$ satisfy $g(c) \leq_{\mathcal{A}} g(d)$ for $c \leq_{\mathcal{A}} d$. Then, any $u \in \mathcal{M}_k$ minimizing $\sum_{i=1}^n h(x_i, a_i, u(x_i)) + g(|u|_{\mathcal{M}_k})$ admits a representation of the form $\sum_{i=1}^n \phi(x_i)c_i$ for some $c_1, \ldots, c_n \in \mathcal{A}$.*

**Proof** Let $\mathcal{V}$ be the space spanned by $\{\phi(x_i)\}_{i=1}^n$. By Proposition 4.3, $u \in \mathcal{M}_k$ is decomposed into $u = u_1 + u_2$, where $u_1 \in \mathcal{V}$, $u_2 \in \mathcal{V}^\perp$. By the reproducing property of $\mathcal{M}_k$, the following equalities are derived for $i = 1, \ldots, n$:

$$u(x_i) = \langle \phi(x_i), u \rangle_{\mathcal{M}_k} = \langle \phi(x_i), u_1 + u_2 \rangle_{\mathcal{M}_k} = \langle \phi(x_i), u_1 \rangle_{\mathcal{M}_k}.$$

Thus, $\sum_{i=1}^n h(x_i, a_i, u(x_i))$ is independent of $u_2$. As for the term $g(|u|_{\mathcal{M}_k})$, since $g$ satisfies $g(c) \leq_{\mathcal{A}} g(d)$ for $c \leq_{\mathcal{A}} d$, we have

$$g(|u|_{\mathcal{M}_k}) = g(|u_1 + u_2|_{\mathcal{M}_k}) = g\Big( \big( |u_1|^2_{\mathcal{M}_k} + |u_2|^2_{\mathcal{M}_k} \big)^{1/2} \Big) \geq_{\mathcal{A}} g(|u_1|_{\mathcal{M}_k}).$$

Therefore, setting $u_2 = 0$ does not affect the term $\sum_{i=1}^n h(x_i, a_i, u(x_i))$, while strictly reducing the term $g(|u|_{\mathcal{M}_k})$, which implies any minimizer must have $u_2 = 0$. As a result, any minimizer takes the form $\sum_{i=1}^n \phi(x_i)c_i$. ■

### 4.2 Connection with RKHSs and vvRKHSs

We show that the framework of RKHM is more general than those of RKHS and vvRKHS. Let $\tilde{k}$ be a complex-valued positive definite kernel and let $\mathcal{H}_{\tilde{k}}$ be the RKHS associated with $\tilde{k}$. In addition, let $k$ be an $\mathcal{A}$-valued positive definite kernel and $\mathcal{M}_k$ be the RKHM associated with $k$. The following proposition is derived by the definitions of RKHSs and RKHMs.

**Proposition 4.9 (Connection between RKHMs with RKHSs)** *If $\mathcal{A} = \mathbb{C}$ and $k = \tilde{k}$, then $\mathcal{H}_{\tilde{k}} = \mathcal{M}_k$.*

As for the connection between vvRKHSs and RKHMs, we first remark that in the case of $\mathcal{A} = \mathcal{B}(\mathcal{W})$, Definition 2.21 is equivalent to the operator valued positive definite kernel (Definition 2.2) for the theory of vv-RKHSs.

**Lemma 4.10 (Connection between Definition 2.21 and Definition 2.2)** *If $\mathcal{A} = \mathcal{B}(\mathcal{W})$, then, the $\mathcal{A}$-valued positive definite kernel defined in Definition 2.21 is equivalent to the operator valued positive definite kernel defined in Definition 2.2.*

The proof for Lemma 4.10 is given in Appendix A.

Let $\mathcal{A} = \mathcal{B}(\mathcal{W})$ and let $\mathcal{H}_k^{\mathrm{v}}$ be the vvRKHS associated with $k$. To investigate further connections between vvRKHSs and RKHMs, we introduce the notion of interior tensor (Lance, 1995, Chapter 4).

**Proposition 4.11** *Let $\mathcal{M}$ be a Hilbert $\mathcal{B}(\mathcal{W})$-module and let $\mathcal{M} \otimes \mathcal{W}$ be the tensor product of $\mathcal{M}$ and $\mathcal{W}$ as vector spaces. The map $\langle \cdot, \cdot \rangle_{\mathcal{M}\otimes\mathcal{W}} : \mathcal{M} \otimes \mathcal{W} \times \mathcal{M} \otimes \mathcal{W} \to \mathbb{C}$ defined as*

$$\langle v \otimes w, u \otimes h \rangle_{\mathcal{M}\otimes\mathcal{W}} = \langle w, \langle v, u \rangle_{\mathcal{M}} h \rangle_{\mathcal{W}}$$

*is a complex-valued pre inner product on $\mathcal{M} \otimes \mathcal{W}$.*

**Definition 4.12 (Interior tensor)** *The completion of $\mathcal{M} \otimes \mathcal{W}$ with respect to the pre inner product $\langle \cdot, \cdot \rangle_{\mathcal{M} \otimes \mathcal{W}}$ is referred to as the* interior tensor *between $\mathcal{M}$ and $\mathcal{W}$, and denoted as $\mathcal{M} \otimes_{\mathcal{B}(\mathcal{W})} \mathcal{W}$.*

Note that $\mathcal{M} \otimes_{\mathcal{B}(\mathcal{W})} \mathcal{W}$ is a Hilbert space. We now show vvRKHSs are reconstructed by the interior tensor between RKHMs and $\mathcal{W}$.

**Theorem 4.13 (Connection between RKHMs and vvRKHSs)** *If $\mathcal{A} = \mathcal{B}(\mathcal{W})$, then two Hilbert spaces $\mathcal{H}_k^{\mathrm{v}}$ and $\mathcal{M} \otimes_{\mathcal{B}(\mathcal{W})} \mathcal{W}$ are isomorphic.*

Theorem 4.13 is derived by the following lemma.

**Lemma 4.14** *There exists a unique unitary map $U \colon \mathcal{M}_k \otimes_{\mathcal{B}(\mathcal{W})} \mathcal{W} \to \mathcal{H}_k^{\mathrm{v}}$ such that $U(\phi(x)c \otimes w) = \phi(x)(cw)$ holds for all $x \in \mathcal{X}$, $c \in \mathcal{B}(\mathcal{W})$ and $w \in \mathcal{W}$.*

**Proof** First, we show that

$$\left\langle \sum_{i=1}^{n} \phi(x_i)c_i \otimes w_i, \sum_{j=1}^{l} \phi(y_j)d_j \otimes h_j \right\rangle_{\mathcal{M}_k \otimes \mathcal{W}} = \left\langle \sum_{i=1}^{n} \phi(x_i)(c_i w_i), \sum_{j=1}^{l} \phi(y_j)(d_j h_j) \right\rangle_{\mathcal{H}_k^{\mathrm{v}}}$$

holds for all $\sum_{i=1}^{n} \phi(x_i)c_i \otimes w_i, \sum_{j=1}^{l} \phi(y_j)d_j \otimes h_j \in \mathcal{M}_k \otimes_{\mathcal{B}(\mathcal{W})} \mathcal{W}$. This follows from the straightforward calculation. Indeed, we have

$$\begin{aligned}
\left\langle \sum_{i=1}^{n} \phi(x_i)c_i \otimes w_i, \sum_{j=1}^{l} \phi(y_j)d_j \otimes h_j \right\rangle_{\mathcal{M}_k \otimes \mathcal{W}} &= \sum_{i=1}^{n} \sum_{j=1}^{l} \left\langle w_i, \langle \phi(x_i)c_i, \phi(y_j)d_j \rangle_k h_j \right\rangle_{\mathcal{W}} \\
&= \sum_{i=1}^{n} \sum_{j=1}^{l} \langle w_i, c_i^* k(x_i, y_j) d_j h_j \rangle_{\mathcal{W}} = \sum_{i=1}^{n} \sum_{j=1}^{l} \langle c_i w_i, k(x_i, y_j) d_j h_j \rangle_{\mathcal{W}} \\
&= \left\langle \sum_{i=1}^{n} \phi(x_i)(c_i w_i), \sum_{j=1}^{l} \phi(y_j)(d_j h_j) \right\rangle_{\mathcal{H}_k^{\mathrm{v}}}.
\end{aligned}$$

Therefore, by the standard functional analysis argument, it turns out that there exists an isometry $U \colon \mathcal{M}_k \otimes_{\mathcal{B}(\mathcal{W})} \mathcal{W} \to \mathcal{H}_k^{\mathrm{v}}$ such that $U(\phi(x)c \otimes w) = \phi(x)(cw)$ holds for all $x \in \mathcal{X}$, $c \in \mathcal{B}(\mathcal{W})$ and $w \in \mathcal{W}$. Since the image of $U$ is closed and dense in $\mathcal{H}_k^{\mathrm{v}}$, $U$ is surjective. Thus $U$ is a unitary map. ■

## 5. Kernel mean embedding in RKHM

We generalize KME in RKHSs, which is widely used in analyzing distributions, to RKHMs. By using the framework of RKHM, we can embed $\mathcal{A}$-valued measures instead of probability measures (more generally, complex-valued measures). We provide a brief review of $\mathcal{A}$-valued measures and the integral with respect to $\mathcal{A}$-valued measures in Appendix B. We define a KME in RKHMs in Subsection 5.1 and show its theoretical properties in Subsection 5.2.

To define a KME by using $\mathcal{A}$-valued measures and integrals, we first define $c_0$-kernels.

**Definition 5.1 (Function space $C_0(\mathcal{X}, \mathcal{A})$)** *For a locally compact Hausdorff space $\mathcal{X}$, the set of all $\mathcal{A}$-valued continuous functions on $\mathcal{X}$ vanishing at infinity is denoted as $C_0(\mathcal{X}, \mathcal{A})$. Here, an $\mathcal{A}$-valued continuous function $u$ is said to vanish at infinity if the set $\{x \in \mathcal{X} \mid \|u(x)\|_{\mathcal{A}} \geq \epsilon\}$ is compact for any $\epsilon > 0$. The space $C_0(\mathcal{X}, \mathcal{A})$ is a Banach $\mathcal{A}$-module with respect to the sup norm.*

Note that if $\mathcal{X}$ is compact, any continuous function is contained in $C_0(\mathcal{X}, \mathcal{A})$.

**Definition 5.2 ($c_0$-kernel)** *Let $\mathcal{X}$ be a locally compact Hausdorff space. An $\mathcal{A}$-valued positive definite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathcal{A}$ is referred to as a* $c_0$-kernel *if $k$ is bounded and $\phi(x) = k(\cdot, x) \in C_0(\mathcal{X}, \mathcal{A})$ for any $x \in \mathcal{X}$.*

In this section, we impose the following assumption.

**Assumption 5.3** *We assume $\mathcal{X}$ is a locally compact Hausdorff space and $k$ is an $\mathcal{A}$-valued $c_0$-positive definite kernel. In addition, we assume $\mathcal{M}_k$ is a von Neumann $\mathcal{A}$-module (see Proposition 4.2).*

For example, we often consider $\mathcal{X} = \mathbb{R}^d$ in practical situations. Also, we provide examples of $c_0$-kernels as follows.

**Example 5.4** *1. Let $\mathcal{A} = L^\infty([0,1])$ and $k$ is an $\mathcal{A}$-valued positive definite kernel defined such that $k(x,y)(t)$ is a complex-valued $c_0$-positive definite kernel for $t \in [0,1]$ (see Example 2.22.2). If $\|k(x,y)\|_{\mathcal{A}}$ is continuous with respect to $y$ for any $x \in \mathcal{X}$, then the inclusion*

$$\{y \in \mathcal{X} \mid \ \|k(x,y)\|_{\mathcal{A}} \geq \epsilon\} \subseteq \{y \in \mathcal{X} \mid \ k(x,y)(t_0) \geq \epsilon\}$$

*holds for some $t_0 \in [0,1]$ and any $x \in \mathcal{X}$ and $\epsilon > 0$. Since $k(\cdot,\cdot)(t_0)$ is a $c_0$-kernel, the set $\{y \in \mathcal{X} \mid \ k(x,y)(t_0) \geq \epsilon\}$ is compact (see Definition 5.1). Thus, $\{y \in \mathcal{X} \mid \ \|k(x,y)\|_{\mathcal{A}} \geq \epsilon\}$ is also compact and $k$ is an $\mathcal{A}$-valued $c_0$-positive definite kernel. Examples of complex-valued $c_0$-positive definite kernels are Gaussian, Laplacian and $B_{2n+1}$-spline kernels.*

*2. Let $\mathcal{W}$ be a separable Hilbert space and let $\{e_i\}_{i=1}^\infty$ be an orthonormal basis of $\mathcal{W}$. Let $\mathcal{A} = \mathcal{B}(\mathcal{W})$ and let $k : \mathcal{X} \times \mathcal{X} \to \mathcal{A}$ be defined as $k(x,y)e_i = k_i(x,y)e_i$, where $k_i : \mathcal{X} \times \mathcal{X} \to \mathbb{C}$ is a complex-valued positive definite kernel for any $i = 1, 2, \ldots$ (see Example 2.22.3). If $\|k(x,y)\|_{\mathcal{A}}$ is continuous with respect to $y$ for any $x \in \mathcal{X}$, then $k$ is shown to be an $\mathcal{A}$-valued $c_0$-positive definite kernel in the same manner as the above example.*

We introduce $\mathcal{A}$-valued measure and integral in preparation for defining a KME in RKHMs. They are special cases of vector measure and integral (Dinculeanu, 1967, 2000), respectively. We review vector measure and integral as $\mathcal{A}$-valued ones in Appendix B. The notions of measure and the Lebesgue integral are generalized to $\mathcal{A}$-valued.

### 5.1 Kernel mean embedding of $C^*$-algebra-valued measures

We now define a KME in RKHMs.

**Definition 5.5 (KME in RKHMs)** *Let $\mathcal{D}(\mathcal{X}, \mathcal{A})$ be the set of all $\mathcal{A}$-valued finite regular Borel measures. A* kernel mean embedding *in an RKHM $\mathcal{M}_k$ is a map $\Phi : \mathcal{D}(\mathcal{X}, \mathcal{A}) \to \mathcal{M}_k$ defined by*

$$\Phi(\mu) := \int_{x \in \mathcal{X}} \phi(x) d\mu(x). \tag{4}$$

We emphasize that the well-definedness of $\Phi$ is not trivial, and von Neumann-$\mathcal{A}$-module is adequate to show it. More precisely, the following theorem derives the well-definedness.

**Theorem 5.6 (Well-definedness for the KME in RKHMs)** *Let $\mu \in \mathcal{D}(\mathcal{X}, \mathcal{A})$. Then, $\Phi(\mu) \in \mathcal{M}_k$. In addition, the following equality holds for any $v \in \mathcal{M}_k$:*

$$\langle \Phi(\mu), v \rangle_{\mathcal{M}_k} = \int_{x \in \mathcal{X}} d\mu^*(x) v(x). \tag{5}$$

To show Theorem 5.6, we use the Riesz representation theorem for Hilbert $\mathcal{A}$-modules (Proposition 4.2).

**Proof** Let $L_\mu : \mathcal{M}_k \to \mathcal{A}$ be an $\mathcal{A}$-linear map defined as $L_\mu v := \int_{x \in \mathcal{X}} d\mu^*(x) v(x)$. The following inequalities are derived by the reproducing property and the Cauchy–Schwarz inequality (Lemma 2.16):

$$\begin{aligned} \|L_\mu v\|_{\mathcal{A}} &\leq \int_{x \in \mathcal{X}} \|v(x)\|_{\mathcal{A}} d|\mu|(x) = \int_{x \in \mathcal{X}} \| \langle \phi(x), v \rangle_{\mathcal{M}_k} \|_{\mathcal{A}} d|\mu|(x) \\ &\leq \|v\|_{\mathcal{M}_k} \int_{x \in \mathcal{X}} \|\phi(x)\|_{\mathcal{M}_k} d|\mu|(x) \leq |\mu|(\mathcal{X}) \|v\|_{\mathcal{M}_k} \sup_{x \in \mathcal{X}} \|\phi(x)\|_{\mathcal{M}_k}, \end{aligned} \tag{6}$$

where the first inequality is easily checked for a step function $s(x) := \sum_{i=1}^n c_i \chi_{E_i}(x)$ as follows:

$$\begin{aligned} \left\| \int_{x \in \mathcal{X}} d\mu^*(x) s(x) \right\|_{\mathcal{A}} &= \left\| \sum_{i=1}^n \mu(E_i)^* c_i \right\|_{\mathcal{A}} \leq \sum_{i=1}^n \|\mu(E_i)\|_{\mathcal{A}} \|c_i\|_{\mathcal{A}} \\ &\leq \sum_{i=1}^n |\mu|(E_i) \|c_i\|_{\mathcal{A}} = \int_{x \in \mathcal{X}} \|s(x)\|_{\mathcal{A}} d|\mu|(x). \end{aligned}$$

Thus, it holds for any totally measurable functions. Since both $|\mu|(\mathcal{X})$ and $\sup_{x \in \mathcal{X}} \|\phi(x)\|_{\mathcal{M}_k}$ are finite, inequality (6) means $L_\mu$ is bounded. Thus, by the Riesz representation theorem for Hilbert $\mathcal{A}$-modules (Proposition 4.2), there exists $u_\mu \in \mathcal{M}_k$ such that $L_\mu v = \langle u_\mu, v \rangle_{\mathcal{M}_k}$. By setting $v = \phi(y)$, for $y \in \mathcal{X}$, we have $u_\mu(y) = L_\mu \phi(y)^* = \int_{x \in \mathcal{X}} k(y, x) d\mu(x)$. Therefore, $\Phi(\mu) = u_\mu \in \mathcal{M}_k$ and $\langle \Phi(\mu), v \rangle_{\mathcal{M}_k} = \int_{x \in \mathcal{X}} d\mu^*(x) v(x)$. ■

**Corollary 5.7** *For $\mu, \nu \in \mathcal{D}(\mathcal{X}, \mathcal{A})$, the inner product between $\Phi(\mu)$ and $\Phi(\nu)$ is given as follows:*

$$\langle \Phi(\mu), \Phi(\nu) \rangle_{\mathcal{M}_k} = \int_{x \in \mathcal{X}} \int_{y \in \mathcal{X}} d\mu^*(x) k(x, y) d\nu(y).$$

Moreover, many basic properties for the existing KME in RKHS are generalized to the proposed KME as follows.

**Proposition 5.8 (Basic properties of the KME $\Phi$)** *For $\mu, \nu \in \mathcal{D}(\mathcal{X}, \mathcal{A})$ and $c \in \mathcal{A}$, $\Phi(\mu + \nu) = \Phi(\mu) + \Phi(\nu)$ and $\Phi(\mu c) = \Phi(\mu)c$ (i.e., $\Phi$ is $\mathcal{A}$-linear, see Definition 2.19) hold. In addition, for $x \in \mathcal{X}$, $\Phi(\delta_x) = \phi(x)$ (see Definition B.2 for the definition of the $\mathcal{A}$-valued Dirac measure $\delta_x$).*

This is derived from Eqs. (4) and (5). Note that if $\mathcal{A} = \mathbb{C}$, then the proposed KME (4) is equivalent to the existing KME in RKHS considered in Sriperumbudur et al. (2011).

### 5.2 Injectivity and universality

Here, we show the connection between the injectivity of the KME and the universality of RKHM. The proofs of the propositions in this subsection are given in Appendix C.

#### 5.2.1 injectivity

In practice, the injectivity of $\Phi$ is important to transform problems in $\mathcal{D}(\mathcal{X}, \mathcal{A})$ into those in $\mathcal{M}_k$. This is because if a KME $\Phi$ in an RKHM is injective, then $\mathcal{A}$-valued measures are embedded into $\mathcal{M}_k$ through $\Phi$ without loss of information. Note that, for probability measures, the injectivity of the existing KME is also referred to as the "characteristic" property. The injectivity of the existing KME in RKHS has been discussed in, for example, Fukumizu et al. (2007); Sriperumbudur et al. (2010, 2011). These studies give criteria for the injectivity of the KMEs associated with important complex-valued kernels such as transition invariant kernels and radial kernels. Typical examples of these kernels are Gaussian, Laplacian, and inverse multiquadratic kernels. Here, we define the transition invariant kernels and radial kernels for $\mathcal{A}$-valued measures, and generalize their criteria to RKHMs associated with $\mathcal{A}$-valued kernels.

To characterize transition invariant kernels, we first define a Fourier transform and support of an $\mathcal{A}$-valued measure.

**Definition 5.9 (Fourier transform and support of an $\mathcal{A}$-valued measure)** *For an $\mathcal{A}$-valued measure $\lambda$ on $\mathbb{R}^d$, the* Fourier transform *of $\lambda$, denoted as $\hat{\lambda}$, is defined as*

$$\hat{\lambda}(x) = \int_{\omega \in \mathbb{R}^d} e^{-\sqrt{-1}x^T\omega} d\lambda(\omega).$$

*In addition, the* support *of $\lambda$ is defined as*

$$\mathrm{supp}(\lambda) = \{x \in \mathbb{R}^d \mid \ \lambda(\mathcal{U}) >_{\mathcal{A}} 0 \text{ for any open set } \mathcal{U} \text{ such that } x \in \mathcal{U}\}.$$

**Definition 5.10 (Transition invariant kernel and radial kernel)** *1. An $\mathcal{A}$-valued positive definite kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathcal{A}$ is called a* transition invariant kernel *if it is represented as $k(x, y) = \hat{\lambda}(y - x)$ for a positive $\mathcal{A}$-valued measure $\lambda$.*

*2. An $\mathcal{A}$-valued positive definite kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathcal{A}$ is called a* radial kernel *if it is represented as $k(x, y) = \int_{[0,\infty)} e^{-t\|x-y\|^2} d\eta(t)$ for a positive $\mathcal{A}$-valued measure $\eta$.*

*Here, an $\mathcal{A}$-valued measure $\mu$ is said to be positive if $\mu(E) \geq_{\mathcal{A}} 0$ for any Borel set $E$.*

We show transition invariant kernels and radial kernels induce injective KMEs.

**Proposition 5.11 (The injectivity for transition invariant kernels)** *Let $\mathcal{A} = \mathbb{C}^{m\times m}$ and $\mathcal{X} = \mathbb{R}^d$. Assume $k : \mathcal{X} \times \mathcal{X} \to \mathcal{A}$ is a transition invariant kernel with a positive $\mathcal{A}$-valued measure $\lambda$ that satisfies* $\mathrm{supp}(\lambda) = \mathcal{X}$. *Then, KME $\Phi : \mathcal{D}(\mathcal{X}, \mathcal{A}) \to \mathcal{M}_k$ defined as Eq.* (4) *is injective.*

**Proposition 5.12 (The injectivity for radial kernels)** *Let $\mathcal{A} = \mathbb{C}^{m\times m}$ and $\mathcal{X} = \mathbb{R}^d$. Assume $k : \mathcal{X} \times \mathcal{X} \to \mathcal{A}$ is a radial kernel with a positive definite $\mathcal{A}$-valued measure $\eta$ that satisfies* $\mathrm{supp}(\eta) \neq \{0\}$. *Then, KME $\Phi : \mathcal{D}(\mathcal{X}, \mathcal{A}) \to \mathcal{M}_k$ defined as Eq.* (4) *is injective.*

**Example 5.13**

1. *If $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{C}^{m\times m}$ is a matrix-valued kernel whose diagonal elements are Gaussian, Laplacian, or $B_{2n+1}$-spline and nondiagonal elements are* 0, *then $k$ is a $c_0$-kernel (See Example 2.22.1). There exists a matrix-valued measure $\lambda$ that satisfies $k(x, y) = \hat{\lambda}(y - x)$ and whose diagonal elements are nonnegative and supported by $\mathbb{R}^d$ (c.f. Table 2 in Sriperumbudur et al. (2010)) and nondiagonal elements are* 0. *Thus, by Proposition 5.11, $\Phi$ is injective.*

2. *If $k$ is a matrix-valued kernel whose diagonal elements are inverse multiquadratic and nondiagonal elements are* 0, *then $k$ is a $c_0$-kernel. There exists a matrix-valued measure $\eta$ that satisfies $k(x, y) = \int_{[0,\infty)} e^{-t\|x-y\|^2} d\eta(t)$, and whose diagonal elements are nonnegative and* $\mathrm{supp}(\eta) \neq \{0\}$ *and nondiagonal elements are* 0 *(c.f. Theorem 7.15 in Wendland (2004)). Thus, by Proposition 5.12, $\Phi$ is injective.*

### 5.2.2 Connection with universality

Another important property for kernel methods is universality, which ensures that kernel-based algorithms approximate each continuous target function arbitrarily well. For RKHS, Sriperumbudur et al. (2011) showed the equivalence of the injectivity of the existing KME in RKHSs and universality of RKHSs. We define a universality of RKHMs as follows.

**Definition 5.14 (Universality)** *An RKHM is said to be* universal *if it is dense in $C_0(\mathcal{X}, \mathcal{A})$.*

We show the above equivalence holds also for RKHM in the case of $\mathcal{A} = \mathbb{C}^{m\times m}$.

**Proposition 5.15 (Equivalence of the injectivity and universality for $\mathcal{A} = \mathbb{C}^{m\times m}$)** *Let $\mathcal{A} = \mathbb{C}^{m\times m}$. Then, $\Phi : \mathcal{D}(\mathcal{X}, \mathcal{A}) \to \mathcal{M}_k$ is injective if and only if $\mathcal{M}_k$ is dense in $C_0(\mathcal{X}, \mathcal{A})$.*

By Proposition 5.15, if $k$ satisfies the condition in Proposition 5.11 or 5.12, then $\mathcal{M}_k$ is universal.

For the case where $\mathcal{A}$ is infinite dimensional, the universality of $\mathcal{M}_k$ in $C_0(\mathcal{X}, \mathcal{A})$ is a sufficient condition for the injectivity of the proposed KME.

**Theorem 5.16 (Connection between the injectivity and universality for general $\mathcal{A}$)** *If $\mathcal{M}_k$ is dense in $C_0(\mathcal{X}, \mathcal{A})$, then $\Phi : \mathcal{D}(\mathcal{X}, \mathcal{A}) \to \mathcal{M}_k$ is injective.*

However, the equivalence of the injectivity and universality, and the injectivity for transition invariant kernels and radial kernels are open problems. This is because their proofs strongly depend on the Hahn–Banach theorem and Riesz–Markov representation theorem, and generalizations of these theorems to $\mathcal{A}$-valued functions and measures are challenging problems due to the situation peculiar to the infinite dimensional spaces. Further details of the proofs of propositions in this section are given in Appendix C.

## 6. Applications

We apply the framework of RKHM described in Sections 4 and 5 to problems in data analysis. We propose kernel PCA in RKHMs in Subsection 6.1, time-series data analysis in RKHMs in Subsection 6.2, and analysis of interaction effects in finite or infinite dimensional data with the proposed KME in RKHMs in Subsection 6.3. Then, we discuss further applications in Subsection 6.4.

### 6.1 PCA in RKHMs

Principal component analysis (PCA) is a fundamental tool for describing data in a low dimensional space. Its implementation in RKHSs has also been proposed (c.f. Schölkopf and Smola (2001)). It enables us to deal with the nonlinearlity of data by virtue of the high expressive power of RKHSs. Here, we generalize the PCA in RKHSs to capture more information in data, such as multivariate data and functional data, by using the framework of RKHM.

**Applying RKHM to PCA** In the existing framework of PCA in Hilbert spaces, the following reconstruction error is minimized with respect to vectors $p_1, \ldots, p_r$:

$$\sum_{i=1}^{n} \left\| x_i - \sum_{j=1}^{r} p_j \langle p_j, x_i \rangle \right\|^2, \tag{7}$$

where $x_1, \ldots, x_n$ are given samples in a Hilbert space and $p_1, \ldots, p_r$ are called principal axes. Here, the complex-valued inner product $\langle p_j, x_i \rangle$ is the weight with respect to the principal axis $p_j$ for representing the sample $x_i$. PCA for functional data (functional PCA) has also investigated (Ramsay and Silverman, 2005). For example, in standard functional PCA settings, we set the Hilbert space as $L^2(\Omega)$ for a compact measure space $\Omega$. However, if samples $x_1, \ldots, x_n$ are finite dimensional vectors or functions, Eq. (7) fails to describe their element wise or continuous dependencies on the principal axes. For $d$-dimensional (finite dimensional) vectors, we can just split $x_i = [x_{i,1}, \ldots, x_{i,d}]$ into $d$ vectors $[x_{i,1}, 0, \ldots, 0], \ldots, [0, \ldots, 0, x_{i,d}]$. Then, we can understand which element is dominant for representing $x_i$ by using the principal axis $p_j$. On the other hand, for functional data, the situation is completely different. For example, assume samples are in $L^2(\Omega)$. Since delta functions are not contained in $L^2(\Omega)$, we cannot split a sample $x_i = x_i(t)$ into discrete functions. In this case, how can we understand the continuous dependencies on the principal axes with respect to the variable $t \in \Omega$? One possible way to answer this question is to employ Hilbert $C^*$-modules instead of Hilbert spaces. We consider the same type of reconstruction error as Eq. (7) in Hilbert $C^*$-modules. In this case, the inner product $\langle p_j, x_i \rangle_{\mathcal{W}}$ is $C^*$-algebra-valued, which allows us to provide more information than the complex-valued

one. If we set the $C^*$-algebra as the function space on $\Omega$ such as $L^\infty(\Omega)$ and define a $C^*$-algebra-valued inner product which depends on $t \in \Omega$, then, the weight $\langle p_j, x_i \rangle_{\mathcal{W}}$ depends on $t$. As a result, we can extract continuous dependencies of samples on the principal axes. More generally, PCA is often considered in an RKHS $\mathcal{H}_{\tilde{k}}$. In this case, $x_i$ in Eq. (7) is replaced with $\tilde{\phi}(x_i)$, where $\tilde{\phi}$ is the feature map, and the inner product and norm are replaced with those in the RKHS. We can extract continuous dependencies of samples on the principal axes by generalizing RKHS to RKHM.

### 6.1.1 Generalization of the PCA in RKHSs to RKHMs

Let $x_1, \ldots, x_n \in \mathcal{X}$ be given samples. Let $k : \mathcal{X} \times \mathcal{X} \to \mathcal{A}$ be an $\mathcal{A}$-valued positive definite kernel on $\mathcal{X}$ and let $\mathcal{M}_k$ be the RKHM associated with $k$. We explore a useful set of axes $p_1, \ldots, p_r$ in $\mathcal{M}_k$, which are referred to as principal axes, to describe the feature of given samples $x_1, \ldots, x_n$. The corresponding components $p_j \langle p_j, \phi(x_i) \rangle_{\mathcal{M}_k}$ are referred to as principal components. We emphasize our proposed PCA in RKHM provides weights of principal components contained in $\mathcal{A}$, not in complex numbers. This is a remarkable difference between our method and existing PCAs. When samples have some structures such as among variables or in functional data, $\mathcal{A}$-valued weights provide us richer information than complex-valued ones. For example, if $\mathcal{X}$ is the space of functions of multi-variables and if we set $\mathcal{A}$ as $L^\infty([0,1])$, then we can reduce multi-variable functional data to functions in $L^\infty([0,1])$, functions of single variable (as illustrated in Section 6.1.4).

To obtain $\mathcal{A}$-valued weights of principal components, we consider the following minimization problem regarding the following reconstruction error (see Definition 2.18 for the definition of ONS):

$$\inf_{\{p_j\}_{j=1}^r \subseteq \mathcal{M}_k:\ \mathrm{ONS}} \sum_{i=1}^n \left| \phi(x_i) - \sum_{j=1}^r p_j \langle p_j, \phi(x_i) \rangle_{\mathcal{M}_k} \right|_{\mathcal{M}_k}^2, \tag{8}$$

where the infimum is taken with respect to a (pre) order in $\mathcal{A}$ (see Definition 2.9). Since the identity $|\phi(x_i) - \sum_{j=1}^r p_j \langle p_j, \phi(x_i) \rangle_{\mathcal{M}_k}|_{\mathcal{M}_k}^2 = k(x_i, x_i) - \sum_{j=1}^r \langle \phi(x_i), p_j \rangle_{\mathcal{M}_k} \langle p_j, \phi(x_i) \rangle_{\mathcal{M}_k}$ holds and $\langle \phi(x_i), p_j \rangle_{\mathcal{M}_k}$ is represented as $p_j(x_i)$ by the reproducing property, the problem (8) can be reduced to the minimization problem

$$\inf_{\{p_j\}_{j=1}^r \subseteq \mathcal{M}_k:\ \mathrm{ONS}} \sum_{i=1}^n \sum_{j=1}^r -p_j(x_i) p_j(x_i)^*. \tag{9}$$

In the case of RKHS, i.e., $\mathcal{A} = \mathbb{C}$, the solution of the problem (9) is obtained by computing eigenvalues and eigenvectors of Gram matrices (see, for example, Schölkopf and Smola (2001)). Unfortunately, we cannot extend their procedure to RKHM straightforwardly. Therefore, we develop two methods to obtain approximate solutions of the problem (9): by gradient descents on Hilbert $C^*$-modules, and by the minimization of the trace of the $\mathcal{A}$-valued objective function.

### 6.1.2 Gradient descent on Hilbert $C^*$-modules

We propose a gradient descent method on Hilbert $\mathcal{A}$-module for the case where $\mathcal{A}$ is commutative. An important example of commutative von Neumann-algebra is $L^\infty([0,1])$. The

gradient descent for a real-valued function on a Hilbert space has been proposed (Smyrlis and Zisis, 2004). However, in our situation, the objective function of the problem (9) is an $\mathcal{A}$-valued function in a Hilbert $C^*$-module $\mathcal{A}^n$. Thus, the existing gradient descent is not applicable to our situation. Therefore, we generalize the existing gradient descent algorithm to $\mathcal{A}$-valued functions on Hilbert $C^*$-modules.

Let $\mathcal{A}$ be a commutative von Neumann-algebra. Assume the positive definite kernel $k$ takes its values in $\mathcal{A}_r := \{c-d \in \mathcal{A} \mid \ c, d \in \mathcal{A}_+\}$. For example, for $\mathcal{A} = L^\infty([0,1])$, $\mathcal{A}_r$ is the space of real-valued $L^\infty$ functions on $[0,1]$. By the representer theorem (Theorem 4.8), if there is a solution of the problem (9), it is represented as $p_j = \sum_{i=1}^n \phi(x_i)c_{j,i}$ for some $c_{j,i} \in \mathcal{A}$. Moreover, since $\mathcal{A}$ is commutative, $p_j(x_i)p_j(x_i)^*$ is equal to $p_j(x_i)^*p_j(x_i)$. Therefore, the problem (8) on $\mathcal{M}_k$ is equivalent to the following problem on the Hilbert $\mathcal{A}$-module $\mathcal{A}^n$ (see Example 2.15 about $\mathcal{A}^n$):

$$\inf_{\mathbf{c}_j \in \mathcal{A}^n,\ \{\sqrt{\mathbf{G}}\mathbf{c}_j\}_{j=1}^r:\ \mathrm{ONS}} -\sum_{j=1}^r \mathbf{c}_j^* \mathbf{G}^2 \mathbf{c}_j, \tag{10}$$

where $\mathbf{G}$ is the $\mathcal{A}$-valued Gram matrix defined as $\mathbf{G}_{i,j} = k(x_i, x_j)$. For simplicity, we assume $r = 1$, i.e., the number of principal axes is 1. We rearrange the problem (10) to the following problem by adding a penalty term:

$$\inf_{\mathbf{c} \in \mathcal{A}^n} (-\mathbf{c}^*\mathbf{G}^2\mathbf{c} + \lambda |\mathbf{c}^*\mathbf{G}\mathbf{c} - 1_{\mathcal{A}}|_{\mathcal{A}}^2), \tag{11}$$

where $\lambda$ is a real positive weight for the penalty term. For $r > 1$, let $\mathbf{c}_1$ be a solution of the problem (10). Then, we solve the same problem in the orthogonal complement of the module spanned by $\{\mathbf{c}_1\}$ and set the solution of this problem as $\mathbf{c}_2$. Then, we solve the same problem in the orthogonal complement of the module spanned by $\{\mathbf{c}_1, \mathbf{c}_2\}$ and repeat this procedure to obtain solutions $\mathbf{c}_1, \dots \mathbf{c}_r$. The problem (11) is the minimization problem of an $\mathcal{A}$-valued function defined on the Hilbert $\mathcal{A}$-module $\mathcal{A}^n$. We search a solution of the problem (11) along the steepest descent directions. To calculate the steepest descent directions, we introduce a derivative $Df_{\mathbf{c}}$ of an $\mathcal{A}$-valued function $f$ on a Hilbert $C^*$-module at $\mathbf{c} \in \mathcal{M}$. It is defined as the derivative on Banach spaces (c.f. Blanchard and Brüning (2015)). The definition of the derivative is included in Appendix D. The following gives the derivative of the objective function in problem (11).

**Proposition 6.1 (Derivative of the objective function)** *Let $f : \mathcal{A}^n \to \mathcal{A}$ be defined as*

$$f(\mathbf{c}) = -\mathbf{c}^*\mathbf{G}^2\mathbf{c} + \lambda |\mathbf{c}^*\mathbf{G}\mathbf{c} - 1_{\mathcal{A}}|_{\mathcal{A}}^2. \tag{12}$$

*Then, $f$ is infinitely differentiable and the first derivative of $f$ is calculated as*

$$Df_{\mathbf{c}}(u) = -2\mathbf{c}^*\mathbf{G}^2 u - 4\lambda \mathbf{c}^*\mathbf{G}u + 4\lambda \mathbf{c}^*\mathbf{G}\mathbf{c}\mathbf{c}^*\mathbf{G}u.$$

*Moreover, for each $\mathbf{c} \in \mathcal{A}^n$, there exists a unique $\mathbf{d} \in \mathcal{A}^n$ such that $\langle \mathbf{d}, u \rangle_{\mathcal{A}^n} = Df_{\mathbf{c}}(u)$ for any $u \in \mathcal{A}^n$. The vector $\mathbf{d}$ is calculated as*

$$\mathbf{d} = -2\mathbf{G}^2\mathbf{c} - 4\lambda \mathbf{G}\mathbf{c} + 4\lambda \mathbf{G}\mathbf{c}\mathbf{c}^*\mathbf{G}\mathbf{c}. \tag{13}$$

**Proof** The derivative of $f$ is calculated by the definition and the assumption that $\mathcal{A}$ is commutative. Since $Df_{\mathbf{c}}$ is a bounded $\mathcal{A}$-linear operator, by the Riesz representation theorem (Proposition 4.2), there exists a unique $\mathbf{d} \in \mathcal{A}^n$ such that $\langle \mathbf{d}, u \rangle_{\mathcal{A}^n} = Df_{\mathbf{c}}(u)$. ■

**Definition 6.2 (Gradient of $\mathcal{A}$-valued functions on Hilbert $C^*$-modules)** *Let $f : \mathcal{M} \to \mathcal{A}$ be a differentiable function. Assume for each $\mathbf{c} \in \mathcal{M}$, there exists a unique $\mathbf{d} \in \mathcal{M}$ such that $\langle \mathbf{d}, u \rangle_{\mathcal{A}^n} = Df_{\mathbf{c}}(u)$ for any $u \in \mathcal{M}$. In this case, we denote $\mathbf{d}$ by $\nabla f_{\mathbf{c}}$ and call it the* gradient *of $f$ at $\mathbf{c}$.*

We now develop an $\mathcal{A}$-valued gradient descent scheme.

**Theorem 6.3** *Assume $f : \mathcal{M} \to \mathcal{A}$ is differentiable. Moreover, assume there exists $\nabla f_{\mathbf{c}}$ for any $\mathbf{c} \in \mathcal{M}$. Let $\eta_t > 0$. Let $\mathbf{c}_0 \in \mathcal{M}$ and*

$$\mathbf{c}_{t+1} = \mathbf{c}_t - \eta_t \nabla f_{\mathbf{c}_t} \tag{14}$$

*for $t = 0, 1, \ldots$. Then, we have*

$$f(\mathbf{c}_{t+1}) = f(\mathbf{c}_t) - \eta_t |\nabla f_{\mathbf{c}_t}|^2_{\mathcal{M}} + S(\mathbf{c}_t, \eta_t), \tag{15}$$

*where $S(\mathbf{c}, \eta)$ satisfies $\lim_{\eta \to 0} \|S(\mathbf{c}, \eta)\|_{\mathcal{A}}/\eta = 0$.*

The statement is derived by the definition of the derivative (Definition D.1). The following examples show the scheme (14) is valid to solve the problem (11).

**Example 6.4** *Let $\mathcal{A} = L^\infty([0,1])$, let $a_t = |\nabla f_{\mathbf{c}_t}|^2_{\mathcal{A}^n} \in \mathcal{A}$ and let $b_{t,\eta} = S(\mathbf{c}_t, \eta) \in \mathcal{A}$. If $a_t \geq_{\mathcal{A}} \delta 1_{\mathcal{A}}$ for some positive real value $\delta$, then the function $a_t$ on $[0,1]$ satisfies $a_t(s) > 0$ for almost everywhere $s \in [0,1]$. On the other hand, since $b_{t,\eta}$ satisfies $\lim_{\eta \to 0} \|b_{t,\eta}\|_{\mathcal{A}}/\eta^2 = 0$, there exists sufficiently small positive real value $\eta_{t,0}$ such that for almost everywhere $s \in [0,1]$, $b_{t,\eta_{t,0}}(s) \leq \|b_{t,\eta_{t,0}}\|_{\mathcal{A}} \leq \eta^2_{t,0}\delta \leq \eta_{t,0}(1-\xi_1)\delta$ hold for some positive real value $\xi_1$. As a result, $-\eta_{t,0}|\nabla f_{\mathbf{c}_t}|^2_{\mathcal{A}^n} + S(\mathbf{c}_t, \eta_{t,0}) \leq_{\mathcal{A}} -\eta_{t,0}\xi_1 |\nabla f_{\mathbf{c}_t}|^2_{\mathcal{A}^n}$ holds and by the Eq. (15), we have*

$$f(\mathbf{c}_{t+1}) <_{\mathcal{A}} f(\mathbf{c}_t) \tag{16}$$

*for $t = 0, 1, \ldots$. As we mentioned in Example 2.8, the inequality (16) means the function $f(\mathbf{c}_{t+1}) \in L^\infty([0,1])$ is smaller than the function $f(\mathbf{c}_t) \in L^\infty([0,1])$ at almost every points on $[0,1]$, i.e.,*

$$f(\mathbf{c}_{t+1})(s) < f(\mathbf{c}_t)(s)$$

*for almost every $s \in [0,1]$.*

**Example 6.5** *Assume $\mathcal{A}$ is a finite dimensional space. If $|\nabla f_{\mathbf{c}_t}|^2_{\mathcal{A}} \geq_{\mathcal{A}} \delta 1_{\mathcal{A}}$ for some positive real value $\delta$, the inequality $f(\mathbf{c}_{t+1}) \leq_{\mathcal{A}} f(\mathbf{c}_t) - \eta_t \xi_1 |\nabla f_{\mathbf{c}_t}|^2_{\mathcal{A}^n}$ holds for $t = 0, 1, \ldots$ and some $\eta_t$ and $\xi_1$ in the same manner as Example 6.4. Moreover, the function $f$ defined as Eq. (12) is bounded below and $\nabla f_{\mathbf{c}_t}$ is Lipschitz continuous on the set $\{\mathbf{c} \in \mathcal{A}^n \mid f(\mathbf{c}) \leq_{\mathcal{A}} f(\mathbf{c}_0)\}$. In this case, if there exists a positive real value $\xi_2$ such that $\|\nabla f_{\mathbf{c}_{t+1}} - \nabla f_{\mathbf{c}_t}\|_{\mathcal{A}^n} \geq \xi_2 \|\nabla f_{\mathbf{c}_t}\|_{\mathcal{A}^n}$, then we have*

$$\xi_2 \|\nabla f_{\mathbf{c}_t}\|_{\mathcal{A}^n} \leq L \|\mathbf{c}_{t+1} - \mathbf{c}_t\|_{\mathcal{A}^n} \leq L\eta_t \|\nabla f_{\mathbf{c}_t}\|_{\mathcal{A}^n},$$

*where $L$ is a Liptschitz constant of $\nabla f_{\mathbf{c}_t}$. As a result, we have*

$$f(\mathbf{c}_{t+1}) \leq_{\mathcal{A}} f(\mathbf{c}_t) - \eta_t \xi_1 |\nabla f_{\mathbf{c}_t}|^2_{\mathcal{A}^n} \leq_{\mathcal{A}} f(\mathbf{c}_t) - \frac{\xi_1 \xi_2}{L} |\nabla f_{\mathbf{c}_t}|^2_{\mathcal{A}^n},$$

*which implies $\sum_{t=1}^{T} |\nabla f_{\mathbf{c}_t}|^2_{\mathcal{A}^n} \leq_{\mathcal{A}} L/(\xi_1\xi_2)(f(\mathbf{c}_1) - f(\mathbf{c}_{T+1}))$. Since $f$ is bounded below, the sum $\sum_{t=1}^{\infty} |\nabla f_{\mathbf{c}_t}|^2_{\mathcal{A}^n}$ converges. Therefore, $|\nabla f_{\mathbf{c}_t}|^2_{\mathcal{A}^n} \to 0$ as $t \to \infty$, i.e., the gradient $\nabla f_{\mathbf{c}_t}$ in Eq. (14) converges to 0.*

**Remark 6.6** *It is possible to generalize the above method to the case where the objective function $f$ has the form $f(\mathbf{c}) = \mathbf{c}^*\mathbf{G}\mathbf{c}$ for $\mathbf{G} \in \mathcal{A}^{n\times n}$ and $\mathcal{A}$ is noncommutative. In this case, the derivative $Df_{\mathbf{c}}$ is calculated as*

$$Df_{\mathbf{c}}(u) = u^*\mathbf{G}\mathbf{c} + \mathbf{c}^*\mathbf{G}u.$$

*Therefore, defining the gradient $\nabla f_{\mathbf{c}}$ as $\nabla f_{\mathbf{c}} = \mathbf{G}\mathbf{c}$ results in $Df_{\mathbf{c}}(-\eta\nabla f_{\mathbf{c}}) = -2\eta\mathbf{c}^*\mathbf{G}^2\mathbf{c} \leq_{\mathcal{A}} 0$ for a real positive value $\eta$, which allows us to derive the same result as Theorem 6.3.*

**Remark 6.7** *The computational complexity of the PCA in RKHMs is higher than the standard PCA in RKHSs. Indeed, in the case of RKHSs, the minimization problem is reduced to an eigenvalue problem of the Gram matrix with respect to given samples. On the other hand, we solve the minimization problem (8) by the gradient descent, and in each iteration step, we compute the gradient $\mathbf{d}$ in Eq. (13). Since the elements of $\mathbf{G}$ and $\mathbf{c}$ are in $\mathcal{A}$, the computation of $\mathbf{d}$ involves the multiplication in $\mathcal{A}$ such as multiplication of functions. Even though we compute the multiplication in $\mathcal{A}$ approximately in practice (see Subsection 6.1.4), its computational cost is much higher than the multiplication in $\mathbb{C}$.*

### 6.1.3 Minimization of the trace

In the case of $\mathcal{A} = \mathcal{B}(\mathcal{W})$, $p_j(x_i)$ and $p_j(x_i)^*$ in the problem (9) do not always commute. Therefore, we restrict the solution to the form $p_j(x_i) = \sum_{i=1}^{n} \phi(x_i)c_i$ where each $c_i$ is a Hilbert–Schmidt operator and minimize the trace of the objective function of the problem (9) as follows:

$$\inf_{\mathbf{c}_j \in F,\ \{\sqrt{\mathbf{G}}\mathbf{c}_j\}_{j=1}^{r}:\ \mathrm{ONS}} -\mathrm{tr}\left(\sum_{j=1}^{r} \mathbf{c}_j^*\mathbf{G}^2\mathbf{c}_j\right), \tag{17}$$

where $F = \{\mathbf{c} = [c_1, \ldots, c_n] \in \mathcal{A}^n \mid \ c_i \text{ is a Hilbert–Schmidt operator for } i = 1, \ldots, n\}$. If $\mathcal{A} = \mathbb{C}^{m\times m}$, i.e., $\mathcal{W}$ is a finite dimensional space, then we solve the problem (17) by regarding $\mathbf{G}$ as an $mn \times mn$ matrix and computing the eigenvalues and eigenvectors of $\mathbf{G}$.

**Proposition 6.8** *Let $\mathcal{A} = \mathbb{C}^{m\times m}$. Let $\lambda_1, \ldots, \lambda_r \in \mathbb{C}$ and $\mathbf{v}_1, \ldots, \mathbf{v}_r \in \mathbb{C}^{mn}$ be the largest $r$ eigenvalues and the corresponding orthonormal eigenvectors of $\mathbf{G} \in \mathbb{C}^{mn\times mn}$. Then, $\mathbf{c}_j = [\mathbf{v}_j, 0, \ldots, 0]\lambda_j^{-1/2}$ is a solution of the problem (17).*

**Proof** Since the identity $\sum_{j=1}^{r} \mathbf{c}_j^*\mathbf{G}^2\mathbf{c}_j = \sum_{j=1}^{r} (\sqrt{\mathbf{G}}\mathbf{c}_j)^*\mathbf{G}(\sqrt{\mathbf{G}}\mathbf{c}_j)$ holds, any solution $\mathbf{c}_j$ of the problem (17) satisfies $\sqrt{\mathbf{G}}\mathbf{c}_j = \mathbf{v}_j u^*$ for a normalized vector $u \in \mathbb{C}^m$. Thus,

$p_j = \sum_{i=1}^n \phi(x_i)c_{i,j}$, where $c_{i,j}$ is the $i$-th element of $\lambda_j^{-1/2}[\mathbf{v}_j, 0, \ldots, 0]$, is a solution of the problem. ■

If $\mathcal{W}$ is an infinite dimensional space, we rewrite the problem (17) with the Hilbert–Schmidt norm as follows:

$$\inf_{\mathbf{c_j}\in F,\ \{\sqrt{\mathbf{G}}\mathbf{c}_j\}_{j=1}^r:\ \mathrm{ONS}} -\sum_{j=1}^r \|\mathbf{G}\mathbf{c}_j\|_F^2, \tag{18}$$

where $\|\mathbf{c}\|_F^2 = \sum_{i=1}^n \|c_i\|_{\mathrm{HS}}^2$ and $\|\cdot\|_{\mathrm{HS}}$ is the Hilbert–Schmidt norm for Hilbert–Schmidt operators. Similar to Eq. (11), we rearrange the problem (18) to the following problem by adding a penalty term:

$$\inf_{\mathbf{c}\in F} -\|\mathbf{G}\mathbf{c}\|_F^2 + \lambda\left|\left\|\sqrt{\mathbf{G}\mathbf{c}}\right\|_F^2 - 1\right|, \tag{19}$$

where $\lambda$ is a real positive weight for the penalty term. Then, we can apply the standard gradient descent method in Hilbert spaces to the problem in $F$ (Smyrlis and Zisis, 2004) since $F$ is the Hilbert space equipped with the Hilbert–Schmidt inner product. Similar to the case of Eq. (11), for $r > 1$, let $\mathbf{c}_1$ be a solution of the problem (19). Then, we solve the same problem in the orthogonal complement of the space spanned by $\{\mathbf{c}_1\}$ and set the solution of this problem as $\mathbf{c}_2$. Then, we solve the same problem in the orthogonal complement of the space spanned by $\{\mathbf{c}_1, \mathbf{c}_2\}$ and repeat this procedure to obtain solutions $\mathbf{c}_1, \ldots \mathbf{c}_r$.

### 6.1.4 Numerical examples

**Experiments with synthetic data** We applied the above PCA with $\mathcal{A} = L^\infty([0,1])$ to functional data. We randomly generated three kinds of sample-sets from the following functions of two variables on $[0,1]\times[0,1]$:

$$y_1(s,t) = e^{10(s-t)}, \quad y_2(s,t) = 10st, \quad y_3(s,t) = \cos(10(s-t)).$$

Each sample-set $i$ is composed of 20 samples with random noise. We denote these samples by $x_1, \ldots, x_{60}$. The noise was randomly drawn from the Gaussian distribution with mean 0 and standard deviation 0.3. Since $L^\infty([0,1])$ is commutative, we applied the gradient descent proposed in Subsection 6.1.2 to solve the problem (8). The parameters were set as $\lambda = 0.1$ and $\eta_t = 0.01$. We set the $L^\infty([0,1])$-valued positive definite kernel $k$ as $(k(x_i,x_j))(t) = \int_0^1\int_0^1 (t - x_i(s_1,s_2))(t - x_j(s_1,s_2))ds_1ds_2$ (see Example 2.22.1). Since $(k(x_i,x_j))(t)$ is a polynomial of $t$, all the computations on $\mathcal{A}$ result in polynomials. Thus, the results are obtained by keeping coefficients of the polynomials. Moreover, we set $\mathbf{c}_0$ as the constant function $[1,\ldots,1]^T \in \mathcal{A}^n$ and computed $\mathbf{c}_1, \mathbf{c}_2, \ldots$ according to Eq. (14). For comparison, we also vectorized the samples by discretizing $y_i$ at $121 = 11\times 11$ points composed of 11 equally spaced points in $[0,1]$ $(0, 0.1, \ldots, 1)$ and applied the standard kernel PCA in the RKHS associated with the Laplacian kernel on $\mathbb{R}^{121}$. The results are illustrated in Figure 2. Since the samples are contaminated by the noise, the PCA in the RKHS cannot separate three sample-sets. On the other hand, the $L^\infty([0,1])$-valued weights of principal components obtained by the proposed PCA in the RKHM reduce the information of the samples as functions. As a result, it clearly separates three sample-sets.

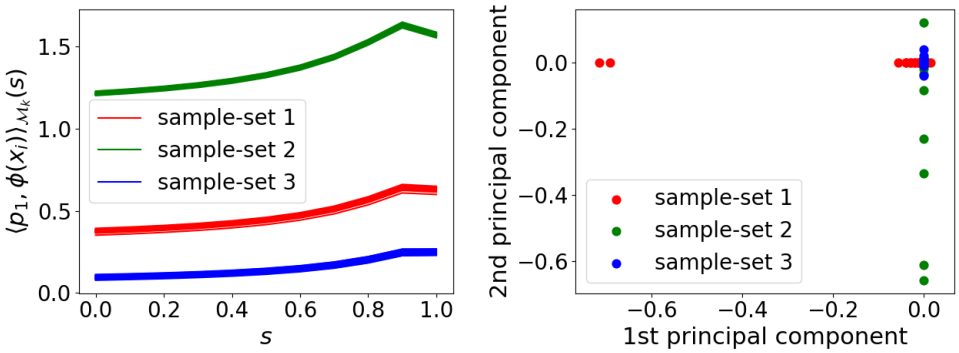



Figure 2: The $L^\infty([0,1])$-valued first principal components obtained by the proposed PCA in an RKHM (left) and the real-valued first and second principal components obtained by the standard PCA in an RKHS (right)

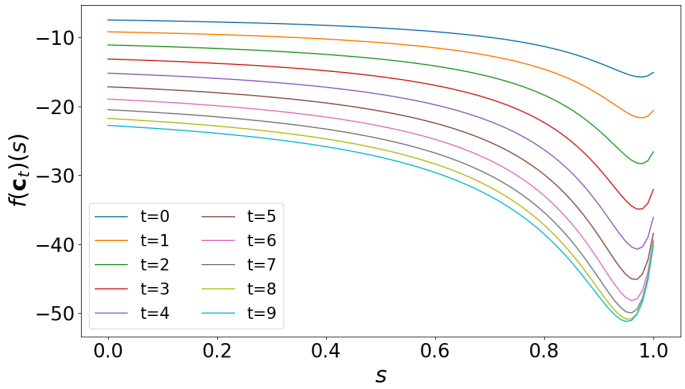



Figure 3: The convergence of the function $f(\mathbf{c}_t)$ along $t$.

Figure 3 shows the convergence of the proposed gradient descent. In this example, we only compute the first principal components, hence $r$ is set as 1. For the objective function $f$ defined as $f(\mathbf{c}) = -\mathbf{c}^*\mathbf{G}^2\mathbf{c} + \lambda\mathbf{c}\mathbf{G}\mathbf{c}\mathbf{c}^*\mathbf{G}\mathbf{c} + \lambda\mathbf{c}^*\mathbf{G}\mathbf{c}$, functions $f(\mathbf{c}_t) \in L^\infty([0,1])$ for $t = 0, \ldots, 9$ are illustrated. We can see $f(\mathbf{c}_{t+1}) < f(\mathbf{c}_t)$ and $f(\mathbf{c}_t)$ gradually approaches a certain function as $t$ grows.

**Experiments with real-world data** To show the proposed PCA with RKHMs extracts the continuous dependencies of samples on the principal axes as we insisted in Section 3, we conducted experiments with climate data in Japan[1]. The data is composed of the maximum and minimum daily temperatures at 47 prefectures in Japan in 2020. The original data is illustrated in Figure 4. The red line represents the temperature at Hokkaido, the northernmost prefecture in Japan and the blue line represents that at Okinawa, the southernmost prefecture in Japan. We respectively fit the maximum and minimum temperatures at each location to the Fourier series $a_0 + \sum_{i=1}^{10}(a_i \cos(it) + b_i \sin(it))$. The fitted functions

1. available at `https://www.data.jma.go.jp/gmd/risk/obsdl/`

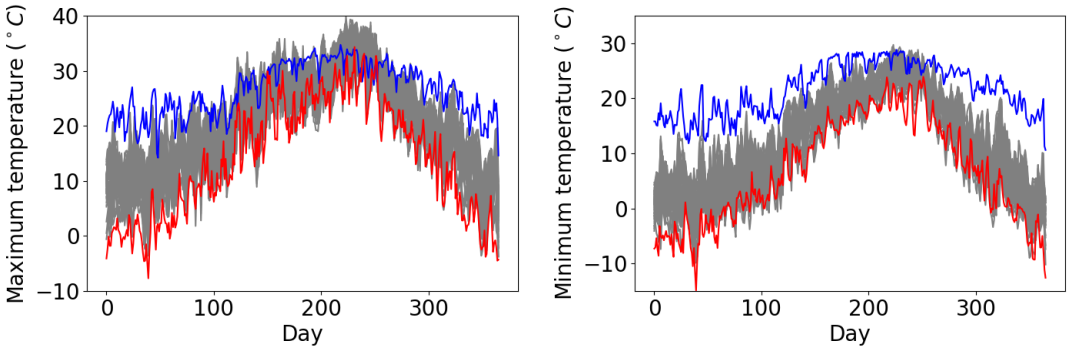



Figure 4: Original climate data at 47 locations

$x_1, \ldots, x_{47} \in C([0, 366], \mathbb{R}^2)$ are illustrated in Figure 5. Then, we applied the PCA with the RKHM associated with the $L^\infty([0, 366])$-valued positive definite kernel $(k(x, y))(t) = e^{-\|x(t)-y(t)\|_2^2}$. Let $\mathcal{F} = \{a_0 + \sum_{i=1}^{10}(a_i \cos(it) + b_i \sin(it)) \mid \ a_i, b_i \in \mathbb{R}\} \subseteq L^2([0, 366])$. We project $k(x, y)$ onto $\mathcal{F}$. Then, for $c, d \in \mathcal{F}$, $c + d \in \mathcal{F}$ is satisfied, but $cd \in \mathcal{F}$ is not always satisfied. Thus, we approximate $cd$ with $a_0 + \sum_{i=1}^{N}(a_i \cos(it) + b_i \sin(it))$ for $N \leq 10$ to restrict all the computations in $\mathcal{F}$ in practice. Here, to remove high frequency components corresponding to noise and extract essential information, we set $N = 3$. Figure 6(a) shows the computed $L^\infty([0, 366])$-valued weights of the first principal axis in the RKHM, which continuously depends on time. The red and blue lines correspond to Hokkaido and Okinawa, respectively. We see these lines are well-separated from other lines corresponding to other prefectures. For comparison, we also applied the PCA in RKHSs to discrete time data. First, we respectively applied the standard kernel PCA with RKHSs to the original temperature each day and obtained real-valued weights of the first principal components. Here, we used the complex-valued Gaussian kernel $\tilde{k}(x, y) = e^{-\|x-y\|_2^2}$. Then, we connected the results and obtained Figure 6(b). Since the original data is not smooth, the PCA amplifies the non-smoothness, which provides meaningless results. Next, we respectively applied the standard kernel PCA with the RKHS to the value of the fitted Fourier series each day and obtained real-valued weights of the first principal components. Then, similar to the case of Figure 6(b), we connected the results and obtained Figure 6(c). In this case, the extracted features somewhat capture the continuous behaviors of the temperatures. However, the PCA in the RKHS amplifies high frequency components, which correspond to noise. Therefore, the result fails to separate the temperatures of Hokkaido and Okinawa, whose behaviors are significantly different as illustrated in Figure 4. On the other hand, the PCA in the RKHM captures the feature of each sample as a function and removes nonessential high frequency components, which results in separating functional data properly.

### 6.2 Time-series data analysis

The problem of analyzing dynamical systems from data by using Perron–Frobenius operators and their adjoints (called Koopman operators), which are linear operators expressing the time evolution of dynamical systems, has recently attracted attention in various fields (Budišić et al., 2012; Črnjarić-Žic et al., 2020; Takeishi et al., 2017a,b; Lusch et al.,

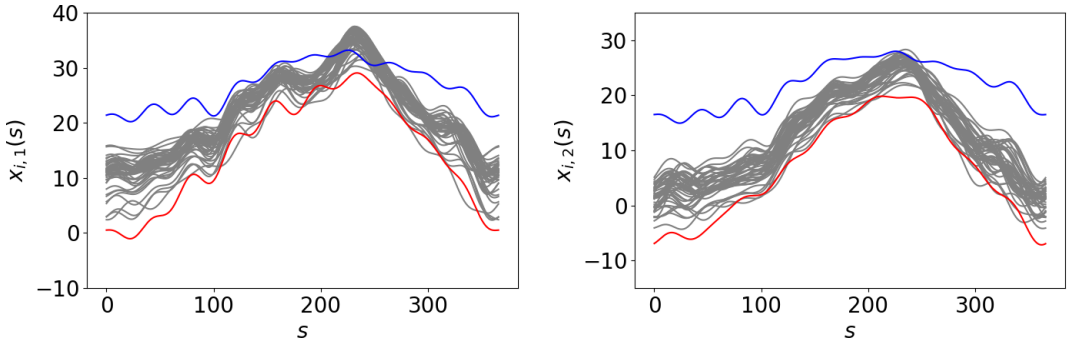
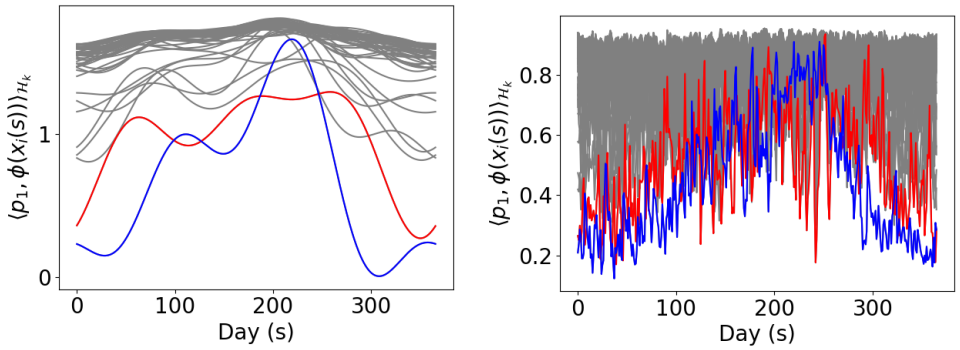
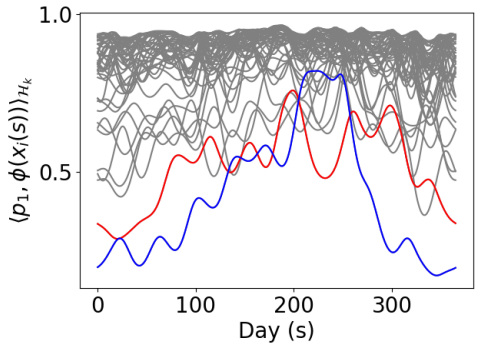



Figure 5: Fitted Fourier series




(a) PCA with RKHMs for the fitted Fourier series

(b) PCA with RKHSs for the original data




(c) PCA with RKHSs for the fitted Fourier series

Figure 6: Principal components of PCA for climate data

2018). And, several methods for this problem using RKHSs have also been proposed (Kawahara, 2016; Klus et al., 2020; Ishikawa et al., 2018; Hashimoto et al., 2020; Fujii & Kawahara, 2019). In these methods, sequential data is supposed to be generated from dynamical systems and is analyzed through Perron–Frobenius operators in RKHSs. To analyze the time evolution of functional data, we generalize Perron–Frobenius operators defined in RKHSs to

those in RKHMs by using an operator-valued positive definite kernel describing similarities between pairs of functions.

**Defining Perron–Frobenius operators in RKHMs** We consider the RKHM and vvRKHS associated with an operator-valued positive definite kernel. VvRKHSs are associated with operator-valued kernels, and as we stated in Lemma 4.10, those operator-valued kernels are special cases of $C^*$-algebra-valued positive definite kernels. Here, we discuss the advantage of RKHMs over vvRKHSs. Comparing with vvRKHSs, RKHMs have enough representation power for preserving continuous behaviors of infinite dimensional operator-valued kernels, while vvRKHSs are not sufficient for preserving such behaviors. Let $\mathcal{W}$ be a Hilbert space, let $k : \mathcal{X} \times \mathcal{X} \to \mathcal{B}(\mathcal{W})$ be an operator-valued positive definite kernel on a data space $\mathcal{X}$, and let $\mathcal{H}_k^{\mathrm{v}}$ be the vvRKHS associated with $k$. Since the inner products in vvRKHSs have the form $\langle w, k(x,y)h\rangle$ for $w, h \in \mathcal{W}$ and $x, y \in \mathcal{X}$, if $\mathcal{W}$ is a $d$-dimensional space, putting $w$ as $d$ linearly independent vectors in $\mathcal{W}$ reconstructs $k(x,y)$. However, if $\mathcal{W}$ is an infinite dimensional space, we need infinitely many $w$ to reconstruct $k(x,y)$, and we cannot recover the continuous behavior of the operator $k(x,y)$ with finitely many $w$. For example, let $\mathcal{X} = C(\Omega, \mathcal{Y})$ and $\mathcal{W} = L^2(\Omega)$ for a compact measure space $\Omega$ and a topological space $\mathcal{Y}$. Let $(k(x,y)w)(s) = \int_{t\in\Omega} \tilde{k}(x(s), y(t))w(t)dt$, where $\tilde{k}$ is a complex-valued positive definite kernel on $\mathcal{Y}$ (see Example 2.22.4). The operator $k(x,y)$ for functional data $x$ and $y$ describes the continuous changes of similarities between function $x$ and $y$. However, the estimation or prediction of the operator $k(x,y)$ in vvRKHSs fails to extract the continuous behavior of the function $\tilde{k}(x(s), y(t))$ in the operator $k(x,y)$ since vectors in vvRKHSs have the form $k(\cdot, y)w$ and we cannot completely recover $k(x,y)$ with finitely many vectors in the vvRKHS. On the other hand, RKHMs have enough information to recover $k(x,y)$ since it is just the inner product between two vectors $\phi(x)$ and $\phi(y)$.

### 6.2.1 Perron–Frobenius operator in RKHSs

We briefly review the definition of the Perron-Frobenius operator on RKHS and existing methods for analysis of time-series data through Perron–Frobenius operators and construction of their estimations (Kawahara, 2016; Hashimoto et al., 2020) . First, we define Perron–Frobenius operators in RKHSs. Let $\{x_0, x_1, \ldots\} \subseteq \mathcal{X}$ be time-series data. We assume it is generated from the following deterministic dynamical system:

$$x_{i+1} = f(x_i), \tag{20}$$

where $f : \mathcal{X} \to \mathcal{X}$ is a map. By embedding $x_i$ and $f(x_i)$ in an RKHS $\mathcal{H}_{\tilde{k}}$ associated with a positive definite kernel $\tilde{k}$ and the feature map $\tilde{\phi}$, dynamical system (20) in $\mathcal{X}$ is transformed into that in the RKHS as

$$\tilde{\phi}(x_{i+1}) = \tilde{\phi}(f(x_i)).$$

The Perron–Frobenius operator $\tilde{K}$ in the RKHS is defined as a linear operator on $\mathcal{H}_{\tilde{k}}$ satisfying

$$\tilde{K}\tilde{\phi}(x) := \tilde{\phi}(f(x))$$

for $x \in \mathcal{X}$. If $\{\tilde{\phi}(x) \mid \ x \in \mathcal{X}\}$ is linearly independent, $\tilde{K}$ is well-defined as a linear map in the RKHS. For example, if $\tilde{k}$ is a universal kernel (Sriperumbudur et al., 2011) such as the Gaussian or Laplacian kernel on $\mathcal{X} = \mathbb{R}^d$, $\{\tilde{\phi}(x) \mid \ x \in \mathcal{X}\}$ is linearly independent.

By considering eigenvalues and the corresponding eigenvectors of $\tilde{K}$, we can understand the long-time behavior of the dynamical system. For example, let $v_1, \ldots, v_m$ be the eigenvectors with respect to eigenvalue 1 of $\tilde{K}$. We project the vector $\tilde{\phi}(x_0)$ onto the subspace spanned by $v_1, \ldots, v_m$. We denote the projected vector by $v$. Then, for $\alpha = 1, 2, \ldots$, we have

$$\tilde{\phi}(x_\alpha) = \tilde{K}^\alpha (v + v^\perp) = v + \tilde{K}^\alpha v^\perp,$$

where $v^\perp = \tilde{\phi}(x_0) - v$. Therefore, by calculating a pre-image of $v$, we can extract the time-invariant component of the dynamical system with the initial value $x_0$.

For practical uses of the above discussion, we construct an estimation of $\tilde{K}$ only with observed data $\{x_0, x_1, \ldots\} \subseteq \mathcal{X}$ as follows: We project $\tilde{K}$ onto the finite dimensional subspace spanned by $\{\tilde{\phi}(x_0), \ldots, \tilde{\phi}(x_{T-1})\}$. Let $\tilde{W}_T := [\tilde{\phi}(x_0), \ldots, \tilde{\phi}(x_{T-1})]$ and $\tilde{W}_T = \tilde{Q}_T \tilde{\mathbf{R}}_T$ be the QR decomposition of $\tilde{W}_T$ in the RKHS. Then, the Perron–Frobenius operator $\tilde{K}$ is estimated by projecting $\tilde{K}$ onto the space spanned by $\{\tilde{\phi}(x_0), \ldots, \tilde{\phi}(x_{T-1})\}$. Since $\tilde{K}\tilde{\phi}(x_i) := \tilde{\phi}(f(x_i)) = \tilde{\phi}(x_{i+1})$ holds, we construct an estimation $\tilde{\mathbf{K}}_T$ of $\tilde{K}$ as follows:

$$\tilde{\mathbf{K}}_T := \tilde{Q}_T^* \tilde{K} \tilde{Q}_T = \tilde{Q}_T^* \tilde{K} \tilde{W}_T \tilde{\mathbf{R}}_T^{-1} = \tilde{Q}_T^* [\tilde{\phi}(x_1), \ldots, \tilde{\phi}(x_T)] \tilde{\mathbf{R}}_T^{-1},$$

which can be computed only with observed data.

### 6.2.2 Perron–Frobenius operator in RKHMs

Existing analyses (Kawahara, 2016; Hashimoto et al., 2020) of time-series data with Perron–Frobenius operators are addressed only in RKHSs. In the remaining parts of this section, we generalize the existing analyses to RKHM to extract continuous behaviors of functional data. We consider the case where time-series is functional data. Let $\Omega$ be a compact measure space, $\mathcal{Y}$ be a topological space, $\mathcal{X} = C(\Omega, \mathcal{Y})$, $\mathcal{A} = \mathcal{B}(L^2(\Omega))$, and $\{x_0, x_1, \ldots\} \subseteq \mathcal{X}$ be functional time-series data. Let $k : \mathcal{X} \times \mathcal{X} \to \mathcal{A}$ be defined as $(k(x, y)w)(s) = \int_{t \in \Omega} \tilde{k}(x(s), y(t))w(t)dt$, where $\tilde{k} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{C}$ is a complex-valued positive definite kernel (see Example 2.22.4 and the last paragraph of Section 3). The operator $k(x, y)$ is the integral operator whose integral kernel is $\tilde{k}(x(s), y(t))$. We define a Perron–Frobenius operator in the RKHM $\mathcal{M}_k$ associated with the above kernel $k$ as an $\mathcal{A}$-linear operator satisfying

$$K\phi(x) = \phi(f(x))$$

for $x \in \mathcal{X}$. We assume $K$ is well-defined on a dense subset of $\mathcal{M}_k$. Then, for $\alpha, \beta = 1, 2, \ldots$, we have

$$k(x_\alpha, x_\beta) = \langle \phi(x_\alpha), \phi(x_\beta) \rangle_{\mathcal{M}_k} = \left\langle K^\alpha \phi(x_0), K^\beta \phi(x_0) \right\rangle_{\mathcal{M}_k}.$$

Therefore, by estimating $K$ in the RKHM $\mathcal{M}_k$, we can extract the similarity between arbitrary points of functions $x_\alpha$ and $x_\beta$. Moreover, the eigenvalues and eigenvectors of $K$ provide us a decomposition of the similarity $k(x_\alpha, x_\beta)$ into a time-invariant term and time-dependent term. Since $K$ is a linear operator on a Banach space $\mathcal{M}_k$, eigenvalues and eigenvectors of $K$ are available. Let $v_1, \ldots, v_m \in \mathcal{M}_k$ be the eigenvectors with respect to eigenvalue 1 of $K$. We project the vector $\phi(x_0)$ onto the submodule spanned by $v_1, \ldots, v_m$, which is denoted by $\mathcal{V}$. Let $\{q_1, \ldots, q_m\} \subseteq \mathcal{M}_k$ be an orthonormal basis of $\mathcal{V}$ and let $v = \sum_{i=1}^m q_i \langle q_i, \phi(x_0) \rangle_{\mathcal{M}_k}$. Then, we have

$$k(x_\alpha, x_\beta) = \left\langle K^\alpha (v + v^\perp), K^\beta (v + v^\perp) \right\rangle_{\mathcal{M}_k} = \langle v, v \rangle_{\mathcal{M}_k} + r(\alpha, \beta), \tag{21}$$

where $v^\perp = \phi(x_0) - v$ and $r(\alpha,\beta) = \left\langle K^\alpha v, K^\beta v^\perp \right\rangle_{\mathcal{M}_k} + \left\langle K^\alpha v^\perp, K^\beta v \right\rangle_{\mathcal{M}_k} + \left\langle K^\alpha v^\perp, K^\beta v^\perp \right\rangle_{\mathcal{M}_k}$. Therefore, the term $\langle v, v \rangle_{\mathcal{M}_k}$ provides us with the information about time-invariant similarities.

**Remark 6.9** *We can also consider the vvRKHS $\mathcal{H}_k^{\mathrm{v}}$ with respect to the operator-valued kernel $k$. Here, we discuss the difference between the case of vvRKHS and RKHM. The Perron–Frobenius operator $K^{\mathrm{v}}$ in a vvRKHS $\mathcal{H}_k^{\mathrm{v}}$ (Fujii & Kawahara, 2019) is defined as a linear operator satisfying*

$$K^{\mathrm{v}} \phi(x) w = \phi(f(x)) w$$

*for $x \in \mathcal{X}$ and $w \in \mathcal{W}$. However, with finitely many vectors in $\mathcal{H}_k^{\mathrm{v}}$, we can only recover an projected operator $UU^* k(x_\alpha, x_\beta) UU^*$, where $N \in \mathbb{N}$, $U = [u_1, \ldots, u_N]$, and $\{u_1, \ldots, u_N\}$ is an orthonormal system on $\mathcal{W}$ as follows:*

$$U^* k(x_\alpha, x_\beta) U = \left[ \langle \phi(x_s) u_i, \phi(x_t) u_j \rangle_{\mathcal{H}_k^{\mathrm{v}}} \right]_{i,j} = \left[ \left\langle (K^{\mathrm{v}})^\alpha \phi(x_0) u_i, (K^{\mathrm{v}})^\beta \phi(x_0) u_j \right\rangle_{\mathcal{H}_k^{\mathrm{v}}} \right]_{i,j}. \quad (22)$$

*Furthermore, let $v_1, \ldots, v_m \in \mathcal{M}_k$ be the eigenvectors with respect to eigenvalue $1$ of $K^{\mathrm{v}}$. Let $\{q_1, \ldots, q_m\} \subseteq \mathcal{H}_k^{\mathrm{v}}$ be an orthonormal basis of the subspace spanned by $v_1, \ldots, v_m$ and let $\tilde{v}_j = \sum_{i=1}^m q_i \langle q_i, \phi(x_0) u_j \rangle_{\mathcal{H}_k^{\mathrm{v}}}$. Then, we have*

$$U^* k(x_\alpha, x_\beta) U = \left[ \langle (K^{\mathrm{v}})^\alpha (\tilde{v}_i + \tilde{v}_i^\perp), (K^{\mathrm{v}})^\beta (\tilde{v}_j + \tilde{v}_j^\perp) \rangle_{\mathcal{H}_k^{\mathrm{v}}} \right]_{i,j} = [\langle \tilde{v}_i, \tilde{v}_j \rangle_{\mathcal{H}_k^{\mathrm{v}}}]_{i,j} + \tilde{r}(\alpha,\beta), \quad (23)$$

*where $\tilde{v}_i^\perp = \phi(x_0) u_i - \tilde{v}_i$ and $\tilde{r}(\alpha,\beta) = [\langle (K^{\mathrm{v}})^\alpha \tilde{v}_i, (K^{\mathrm{v}})^\beta \tilde{v}_j^\perp \rangle_{\mathcal{H}_k^{\mathrm{v}}} + \langle (K^{\mathrm{v}})^\alpha \tilde{v}_i^\perp, (K^{\mathrm{v}})^\beta \tilde{v}_j \rangle_{\mathcal{H}_k^{\mathrm{v}}} + \langle (K^{\mathrm{v}})^\alpha \tilde{v}_i^\perp, (K^{\mathrm{v}})^\beta \tilde{v}_j^\perp \rangle_{\mathcal{H}_k^{\mathrm{v}}}]_{i,j}$. Therefore, with vvRKHSs, we cannot recover the continuous behavior of the operator $k(x, y)$ which encodes similarities between functions $x$ and $y$.*

### 6.2.3 Estimation of Perron–Frobenius operators in RKHMs

In practice, we only have time-series data but do not know the underlying dynamical system and its Perron–Frobenius operator in an RKHM. Therefore, we consider estimating the Perron–Frobenius operator only with the data. To do so, we generalize the Gram–Schmidt orthonormalization algorithm to Hilbert $C^*$-modules to apply the QR decomposition and project Perron–Frobenius operators onto the submodule spanned by $\{\phi(x_0), \ldots, \phi(x_{T-1})\}$. The Gram–Schmidt orthonormalization in Hilbert modules is theoretically investigated by Cnops (1992). Here, we develop a practical method for our settings. Then, we can apply the decomposition (21), proposed in Subsection 6.2.2, of the estimated operator regarding eigenvectors. Since we are considering the RKHM associated with the integral operator-valued positive definite kernel defined in the first part of Subsection 6.2.2, we assume $\mathcal{A} = \mathcal{B}(\mathcal{W})$ and we denote by $\mathcal{M}$ a Hilbert $C^*$-module over $\mathcal{A}$ throughout this subsection. Note that integral operators are compact.

We first develop a normalization method for Hilbert $C^*$-modules. In $C^*$-algebras, nonzero elements are not always invertible, which is the main difficulty of the normalization in Hilbert $C^*$-modules. However, by carefully applying the definition of normalized (see Definition 2.17), we can construct a normalization method.

**Proposition 6.10 (Normalization)** *Let $\epsilon \geq 0$ and let $\hat{q} \in \mathcal{M}$ satisfy $\|\hat{q}\|_{\mathcal{M}} > \epsilon$. Assume $\langle \hat{q}, \hat{q} \rangle_{\mathcal{M}}$ is compact. Then, there exists $\hat{b} \in \mathcal{A}$ such that $\|\hat{b}\|_{\mathcal{A}} < 1/\epsilon$ and $q := \hat{q}\hat{b}$ is normalized. In addition, there exists $b \in \mathcal{A}$ such that $\|\hat{q} - qb\|_{\mathcal{M}} \leq \epsilon$.*

**Proof** Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$ be the eigenvelues of the compact operator $\langle \hat{q}, \hat{q} \rangle_{\mathcal{M}}$, and $m' := \max\{j \mid \lambda_j > \epsilon^2\}$. Since $\langle \hat{q}, \hat{q} \rangle_{\mathcal{M}}$ is positive and compact, it admits the spectral decomposition $\langle \hat{q}, \hat{q} \rangle_{\mathcal{M}} = \sum_{i=1}^{\infty} \lambda_i v_i v_i^*$, where $v_i$ is the orthonormal eigenvector with respect to $\lambda_i$. Also, since $\lambda_1 = \|\hat{q}\|_{\mathcal{M}}^2 > \epsilon^2$, we have $m' \geq 1$. Let $\hat{b} = \sum_{i=1}^{m'} 1/\sqrt{\lambda_i} v_i v_i^*$. By the definition of $\hat{b}$, $\|\hat{b}\|_{\mathcal{A}} = 1/\sqrt{\lambda_{m'}} < 1/\epsilon$ holds. Also, we have

$$\langle \hat{q}\hat{b}, \hat{q}\hat{b} \rangle_{\mathcal{M}} = \hat{b}^* \langle \hat{q}, \hat{q} \rangle_{\mathcal{M}} \hat{b} = \sum_{i=1}^{m'} \frac{1}{\sqrt{\lambda_i}} v_i v_i^* \sum_{i=1}^{\infty} \lambda_i v_i v_i^* \sum_{i=1}^{m'} \frac{1}{\sqrt{\lambda_i}} v_i v_i^* = \sum_{i=1}^{m'} v_i v_i^*.$$

Thus, $\langle \hat{q}\hat{b}, \hat{q}\hat{b} \rangle_{\mathcal{M}}$ is a nonzero orthogonal projection.

In addition, let $b = \sum_{i=1}^{m'} \sqrt{\lambda_i} v_i v_i^*$. Since $\hat{b}b = \sum_{i=1}^{m'} v_i v_i^*$, the identity $\langle \hat{q}, \hat{q}\hat{b}b \rangle = \langle \hat{q}\hat{b}b, \hat{q}\hat{b}b \rangle$ holds, and we obtain

$$\begin{aligned} \langle \hat{q} - qb, \hat{q} - qb \rangle_{\mathcal{M}} &= \langle \hat{q} - \hat{q}\hat{b}b, \hat{q} - \hat{q}\hat{b}b \rangle_{\mathcal{M}} = \langle \hat{q}, \hat{q} \rangle - \langle \hat{q}\hat{b}b, \hat{q}\hat{b}b \rangle_{\mathcal{M}} \\ &= \sum_{i=1}^{\infty} \lambda_i v_i v_i^* - \sum_{i=1}^{m'} \lambda_i v_i v_i^* = \sum_{i=m'+1}^{\infty} \lambda_i v_i v_i^*. \end{aligned}$$

Thus, $\|\hat{q} - q\hat{b}\|_{\mathcal{M}} = \sqrt{\lambda_{m'+1}} \leq \epsilon$ holds, which completes the proof of the proposition. ■

Proposition 6.10 and its proof provide a concrete procedure to obtain normalized vectors in $\mathcal{M}$. This enables us to compute an orthonormal basis practically by applying Gram-Schmidt orthonormalization with respect to $\mathcal{A}$-valued inner product.

**Proposition 6.11 (Gram-Schmidt orthonormalization)** *Let $\{w_i\}_{i=1}^{\infty}$ be a sequence in $\mathcal{M}$. Assume $\langle w_i, w_j \rangle_{\mathcal{M}}$ is compact for any $i, j = 1, 2, \ldots$. Consider the following scheme for $i = 1, 2, \ldots$ and $\epsilon \geq 0$:*

$$\begin{aligned} &\hat{q}_j = w_j - \sum_{i=1}^{j-1} q_i \langle q_i, w_j \rangle_{\mathcal{M}}, \quad q_j = \hat{q}_j \hat{b}_j \quad \text{if } \|\hat{q}_j\|_{\mathcal{M}} > \epsilon, \\ &q_j = 0 \quad o.w., \end{aligned} \tag{24}$$

*where $\hat{b}_j$ is defined as $\hat{b}$ in Proposition 6.10 by setting $\hat{q} = \hat{q}_j$. Then, $\{q_j\}_{j=1}^{\infty}$ is an orthonormal basis in $\mathcal{M}$ such that any $w_j$ is contained in the $\epsilon$-neighborhood of the space spanned by $\{q_j\}_{j=1}^{\infty}$.*

**Remark 6.12** *We give some remarks about the role of $\epsilon$ in Propositions 6.10. The vector $\hat{q}_i$ can always be reconstructed by $w_i$ only when $\epsilon = 0$. This is because the information of the spectrum of $\langle \hat{q}_i, \hat{q}_i \rangle_{\mathcal{M}}$ may be lost if $\epsilon > 0$. However, if $\epsilon$ is sufficiently small, we can reconstruct $\hat{q}_i$ with a small error. On the other hand, the norm of $\hat{b}_i$ can be large if $\epsilon$ is small, and the computation of $\{q_i\}_{i=1}^{\infty}$ can become numerically unstable. This corresponds to the trade-off between the theoretical accuracy and numerical stability.*

To prove Proposition 6.11, we first prove the following lemmas.

**Lemma 6.13** *For $c \in \mathcal{A}$ and $v \in \mathcal{M}$, if $\langle v, v \rangle_{\mathcal{M}} c = \langle v, v \rangle_{\mathcal{M}}$, then $vc = v$ holds.*

**Proof** If $\langle v, v\rangle_{\mathcal{M}} c = \langle v, v\rangle_{\mathcal{M}}$, then $c^* \langle v, v\rangle_{\mathcal{M}} = \langle v, v\rangle_{\mathcal{M}}$ and we have

$$\langle vc - v, vc - v\rangle_{\mathcal{M}} = c^* \langle v, v\rangle_{\mathcal{M}} c - c^* \langle v, v\rangle_{\mathcal{M}} - \langle v, v\rangle_{\mathcal{M}} c + \langle v, v\rangle_{\mathcal{M}} = 0,$$

which implies $vc = v$. ■

**Lemma 6.14** *If $q \in \mathcal{M}$ is normalized, then $q \langle q, q\rangle_{\mathcal{M}} = q$ holds.*

**Proof** Since $\langle q, q\rangle_{\mathcal{M}}$ is a projection, $\langle q, q\rangle_{\mathcal{M}} \langle q, q\rangle_{\mathcal{M}} = \langle q, q\rangle_{\mathcal{M}}$ holds. Therefore, letting $c = \langle q, q\rangle_{\mathcal{M}}$ and $v = q$ in Lemma 6.13 completes the proof of the lemma. ■

**Proof of Proposition 6.11** By Proposition 6.10, $q_j$ is normalized, and for $\epsilon \geq 0$, there exists $b_j \in \mathcal{A}$ such that $\|\hat{q}_j - q_j b_j\|_{\mathcal{M}} \leq \epsilon$. Therefore, by the definition of $\hat{q}_j$, $\|w_j - v_j\|_{\mathcal{M}} \leq \epsilon$ holds, where $v_j$ is a vector in the space spanned by $\{q_j\}_{j=0}^{\infty}$ which is defined as $v_j = \sum_{i=1}^{j-1} q_i \langle q_i, w_j\rangle_{\mathcal{M}} - q_j b_j$. This means that the $\epsilon$-neighborhood of the space spanned by $\{q_j\}_{j=1}^{\infty}$ contains $\{w_j\}_{j=1}^{\infty}$. Next, we show the orthogonality of $\{q_j\}_{j=1}^{\infty}$. Assume $q_1, \ldots, q_{j-1}$ are orthogonal to each other. For $i < j$, the following identities are deduced by Lemma 6.14:

$$\begin{aligned}\langle q_j, q_i\rangle_{\mathcal{M}} &= \hat{b}_t^* \langle \hat{q}_j, q_i\rangle_{\mathcal{M}} = \hat{b}_j^* \left\langle w_j - \sum_{l=1}^{j-1} q_l \langle q_l, w_j\rangle, q_i \right\rangle_{\mathcal{M}} \\ &= \hat{b}_j^* \left(\langle w_j, q_i\rangle_{\mathcal{M}} - \langle q_i \langle q_i, w_j\rangle_{\mathcal{M}}, q_i\rangle\right) = \hat{b}_j^* \left(\langle w_j, q_i\rangle_{\mathcal{M}} - \langle w_j, q_i\rangle_{\mathcal{M}}\right) = 0.\end{aligned}$$

Therefore, $q_1, \ldots, q_j$ are also orthogonal to each other, which completes the proof of the proposition. ■

In practical computations, the scheme (24) should be represented with matrices. For this purpose, we derive the following QR decomposition from Proposition 6.11. This is a generalization of the QR decomposition in Hilbert spaces.

**Corollary 6.15 (QR decomposition)** *For $n \in \mathbb{N}$, let $W := [w_1, \ldots, w_n]$ and $Q := [q_1, \ldots, q_n]$. Let $\epsilon \geq 0$. Then, there exist $\mathbf{R}, \mathbf{R}_{\mathrm{inv}} \in \mathcal{A}^{n\times n}$ that satisfy*

$$Q = W\mathbf{R}_{\mathrm{inv}}, \quad \|W - Q\mathbf{R}\| \leq \epsilon. \tag{25}$$

*Here, $\|W\|$ for a $\mathcal{A}$-linear map $W : \mathcal{A}^n \to \mathcal{M}$ is defined as $\|W\| := \sup_{\|v\|_{\mathcal{A}^n}=1} \|Wv\|_{\mathcal{M}}$.*

**Proof** Let $\mathbf{R} = [r_{i,j}]_{i,j}$ be an $n \times n$ $\mathcal{A}$-valued matrix. Here, $r_{i,j}$ is defined by $r_{i,j} = \langle q_i, w_j\rangle_{\mathcal{M}} \in \mathcal{A}$ for $i < j$, $r_{i,j} = 0$ for $i > j$, and $r_{j,j} = b_j$, where $b_j$ is defined as $b$ in Proposition 6.10 by setting $\hat{q} = \hat{q}_j$. In addition, let $\hat{\mathbf{B}} = \mathrm{diag}\{\hat{b}_1, \ldots, \hat{b}_n\}$, $\mathbf{B} = \mathrm{diag}\{b_1, \ldots, b_n\}$, and $\mathbf{R}_{\mathrm{inv}} = \hat{\mathbf{B}}(I + (\mathbf{R} - \mathbf{B})\hat{\mathbf{B}})^{-1}$ be $n \times n$ $\mathcal{A}$-valued matrices. The equality $Q = W\mathbf{R}_{\mathrm{inv}}$ is derived directly from scheme (24). In addition, by the scheme (24), for $t = 1, \ldots, n$, we have

$$w_j = \sum_{i=1}^{j-1} q_i \langle q_i, w_j\rangle_{\mathcal{M}} + \hat{q}_j = \sum_{i=1}^{j-1} q_i \langle q_i, w_j\rangle_{\mathcal{M}} + q_j b_j + \hat{q}_j - q_j b_j = Q\mathbf{r}_j + \hat{q}_j - q_j b_j,$$

where $\mathbf{r}_j \in \mathcal{A}^n$ is the $i$-th column of $\mathbf{R}$. Therefore, by Proposition 6.10, $\|w_j - Q\mathbf{r}_j\|_{\mathcal{M}} = \|\hat{q}_j - q_j b_j\|_{\mathcal{M}} \leq \epsilon$ holds for $j = 1, \ldots, n$, which implies $\|W - Q\mathbf{R}\| \leq \epsilon$. ■

We call the decomposition (25) as the QR decomposition in Hilbert $C^*$-modules. Although we are handling vectors in $\mathcal{M}$, by applying the QR decomposition, we only have to compute $\mathbf{R}_{\text{inv}}$ and $\mathbf{R}$.

We now consider estimating the Perron–Frobenius operator $K$ with observed time-series data $\{x_0, x_1, \ldots\}$. Let $W_T = [\phi(x_0), \ldots, \phi(x_{T-1})]$. We are considering an integral operator-valued positive definite kernel (see the first part of Subsection 6.2.2 and the last paragraph in Section 3). Since integral operators are compact, $W_T$ satisfies the assumption in Corollary 6.11. Thus, let $W_T\mathbf{R}_{\text{inv},T} = Q_T$ be the QR decomposition (25) of $W_T$ in the RKHM $\mathcal{M}_k$. The Perron–Frobenius operator $K$ is estimated by projecting $K$ onto the space spanned by $\{\phi(x_0), \ldots, \phi(x_{T-1})\}$. We define $\mathbf{K}_T$ as the estimation of $K$. Since $K\phi(x_i) = \phi(f(x_i)) = \phi(x_{i+1})$ hold, $\mathbf{K}_T$ can be computed only with observed data as follows:

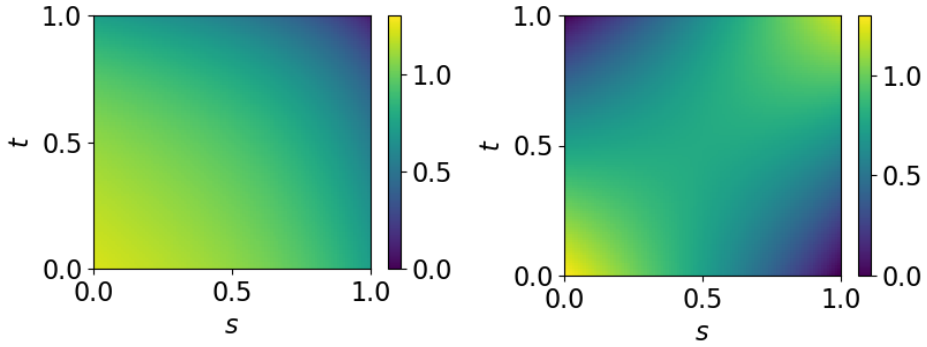
$$\mathbf{K}_T = Q_T^* K Q_T = Q_T^* K W_T \mathbf{R}_{\text{inv},T} = Q_T^* [\phi(x_1), \ldots, \phi(x_T)] \mathbf{R}_{\text{inv},T}.$$

**Remark 6.16** *In practical computations, we only need to keep the integral kernels to implement the Gram–Schmidt orthonormalization algorithm and estimate Perron–Frobenius operators in the RKHM associated with the integral operator-valued kernel $k$. Therefore, we can directly access integral kernel functions of operators, which is not achieved by vvRKHS as we stated in Remark 4.13. Indeed, the operations required for estimating Perron–Frobenius operators are explicitly computed as follows: Let $c, d \in \mathcal{B}(L^2(\Omega))$ be integral operators whose integral kernels are $f(s,t)$ and $g(s,t)$. Then, the integral kernels of the operator $c+d$ and $cd$ are $f(s,t)+g(s,t)$ and $\int_{r\in\Omega} f(s,r)g(r,t)dr$, respectively. And that of $c^*$ is $f(t,s)$. Moreover, if $c$ is positive, let $c_\epsilon^+$ be $\sum_{\lambda_i>\epsilon} 1/\sqrt{\lambda_i} v_i {v_i}^*$, where $\lambda_i$ are eigenvalues of the compact positive operator $c$ and $v_i$ are corresponding orthonormal eigenvectors. Then, the integral kernel of the operator $c_\epsilon^+$ is $\sum_{\lambda_i>\epsilon} 1/\sqrt{\lambda_i} v_i(s)\overline{v_i(t)}$.*

### 6.2.4 Numerical examples

To show the proposed analysis with RKHMs captures continuous changes of values of kernels along functional data as we insisted in Section 3, we conducted experiments with river flow data of the Thames River in London[2]. The data is composed of daily flow at 10 stations. We used the data for 51 days beginning from January first, 2018. We regard every daily flow as a function of the ratio of the distance from the most downstream station and fit it to a polynomial of degree 5 to obtain time series $x_0, \ldots, x_{50} \in C([0,1], \mathbb{R})$. Then, we estimated the Perron–Frobenius operator which describes the time evolution of the series $x_0, \ldots, x_{50}$ in the RKHM associated with the $\mathcal{B}(L^2([0,1]))$-valued positive definite kernel $k(x,y)$ defined as the integral operator whose integral kernel is $\tilde{k}(s,t) = e^{-|x(s)-y(t)|^2}$ for $x, y \in C([0,1], \mathbb{R})$. In this case, $T = 50$. As we noted in Remark 6.16, all the computations in $\mathcal{A} = \mathcal{B}(L^2([0,1]))$ are implemented by keeping integral kernels of operators. Let $\mathcal{F}$ be the set of polynomials of the form $x_i(s,t) = \sum_{j,l=0}^{5} \eta_{j,l} s^j t^l$, where $\eta_{j,l} \in \mathbb{R}$. We project $\tilde{k}$

2. available at `https://nrfa.ceh.ac.uk/data/search`




(a) RKHM (b) vvRKHS

Figure 7: Heat maps representing time-invariant similarities

onto $\mathcal{F}$. Then, for $c, d \in \mathcal{F}$, $c + d \in \mathcal{F}$ is satisfied, but $cd \in \mathcal{F}$ is not always satisfied. Thus, we project $cd$ onto $\mathcal{F}$ to restrict all the computations in $\mathcal{F}$ in practice. We computed the time-invariant term $\langle v, v \rangle_{\mathcal{M}_k}$ in Eq. (22). Regarding the computation of eigenvectors with respect to the eigenvalue 1, we consider the following minimization problem for the estimated Perron–Frobenius operator $\mathbf{K}_T$:

$$\inf_{\mathbf{v} \in \mathcal{A}^T} |\mathbf{K}_T \mathbf{v} - \mathbf{v}|^2_{\mathcal{A}^T} - \lambda |\mathbf{v}|^2_{\mathcal{A}^T}. \tag{26}$$

Here, $-\lambda |\mathbf{v}|^2_{\mathcal{A}^T}$ is a penalty term to keep $\mathbf{v}$ not going to 0. Since the objective function of the problem (26) is represented as $\mathbf{v}^*(\mathbf{K}_T^* \mathbf{K}_T - \mathbf{K}_T^* - \mathbf{K}_T + (1-\lambda)\mathbf{I})\mathbf{v}$, where $\mathbf{I}$ is the identity operator on $\mathcal{A}^T$, we apply the gradient descent on $\mathcal{A}^T$ (see Remark 6.6). Figure 7(a) shows the heat map representing the integral kernel of $\langle v, v \rangle_{\mathcal{M}_k}$.

For comparison, we also applied the similar analysis in a vvRKHS. We computed the time-invariant term $[\langle \tilde{v}_i, \tilde{v}_j \rangle_{\mathcal{H}_k^{\mathrm{v}}}]_{i,j}$ in Eq. (23) by setting $u_i$ as orthonormal polynomials of the form $u_i(s) = \sum_{j=1}^5 \eta_j s^j$, where $\eta_j \in \mathbb{R}$. Let $c_{\mathrm{inv}} = [\langle \tilde{v}_i, \tilde{v}_j \rangle_{\mathcal{H}_k^{\mathrm{v}}}]_{i,j}$. In this case, we cannot obtain the integral kernel of the time-invariant term of the operator $k(x_\alpha, x_\beta)$, which is denoted by $\tilde{k}_{\mathrm{inv}}$ here. Instead, by approximating $k(x_\alpha, x_\beta)$ by $UU^* k(x_\alpha, x_\beta) UU^*$ and computing $U c_{\mathrm{inv}} U^* \chi_{[0,t]}$, we obtain an approximation of $\int_0^t \tilde{k}_{\mathrm{inv}}(s, r) dr$ for $s \in [0, 1]$. Here, $\chi_E : [0, 1] \to \{0, 1\}$ is the indicator function for a Borel set $E$ on $[0, 1]$. Therefore, by numerically differentiating $U c_{\mathrm{inv}} U^* \chi_{[0,t]}$ by $t$, we obtain an approximation of $\tilde{k}_{\mathrm{inv}}$. Figure 7(b) shows the heat map representing the approximation of $\tilde{k}_{\mathrm{inv}}$.

Around the upstream stations, there are many branches and the flow is affected by them. Thus, the similarity between flows at two points would change along time. While, around the downstream stations, the flow is supposed not to be affected by other rivers. Thus, the similarity between flows at two points would be invariant along time. The values around the diagonal part of Figure 7(a) (RKHM) become small as $s$ and $t$ become large (as going up the river). On the other hand, those of Figure 7(b) (vvRKHS) are also large for large $s$ and $t$. Therefore, RKHM captures the aforementioned fact more properly.

### 6.3 Analysis of interaction effects

Polynomial regression is a classical problem in statistics (Hastie et al., 2009) and analyzing interacting effects by the polynomial regression has been investigated (for its recent improvements, see, for example, Suzumura et al. (2017)). Most of the existing methods focus on the case of finite dimensional (discrete) data. However, in practice, we often encounter situations where we cannot fix the dimension of data. For example, observations are obtained at multiple locations and the locations are not fixed. It may be changed depending on time. Therefore, analysing interaction effects of infinite dimensional (continuous) data is essential. We show the KMEs of $\mathcal{A}$-valued measures in RKHMs provide us with a method for the analysis of infinite dimensional data by setting $\mathcal{A}$ as an infinite dimensional space such as $\mathcal{B}(\mathcal{W})$. Moreover, the proposed method does not need the assumption that interaction effects are described by a polynomial. We first develop the analysis in RKHMs for the case of finite dimensional data in Subsection 6.3.1. Then, we show the analysis is naturally generalized to the infinite dimensional data in Subsection 6.3.2.

**Applying $\mathcal{A}$-valued measures and KME in RKHMs** Using $\mathcal{A}$-valued measures, we can describe the measure corresponding to each point of functional data as functions or operators. For example, let $\mathcal{X}$ be a locally compact Hausdorff space and let $x_1, x_2, \ldots \in C([0,1], \mathcal{X})$ be samples. Let $\mathcal{A} = L^\infty([0,1])$ and let $\mu$ be the $\mathcal{A}$-valued measure defined as $\mu(t) = \tilde{\mu}_t$, where $\tilde{\mu}_t$ is the distribution which samples $x_1(t), x_2(t), \ldots$ follow. Then, $\mu$ describes continuous behaviors of the distribution of samples $x_1(t), x_2(t), \ldots$ with respect to $t$. Moreover, let $\mathcal{A} = \mathcal{B}(L^2([0,1]))$ and let $\mu$ be the $\mathcal{A}$-valued measure defined as $(\mu(E)v)(s) = \int_{t\in[0,1]} \tilde{\mu}(E)_{s,t} v(t) dt$ for a Borel set $E$, where $\tilde{\mu}_{s,t}$ is the joint distribution of the distributions which samples $x_1(s), x_2(s), \ldots$ and samples $x_1(t), x_2(t), \ldots$ follow. Then, $\mu$ describes continuous dependencies of samples $x_1(s), x_2(s), \ldots$ and samples $x_1(t), x_2(t), \ldots$ with respect to $s$ and $t$. Using the KME in RKHMs, we can embed $\mathcal{A}$-valued measures into RKHMs, which enables us to compute inner products between $\mathcal{A}$-valued measures. Then, we can generalize algorithms in Hilbert spaces to $\mathcal{A}$-valued measures.

#### 6.3.1 The case of finite dimensional data

In this subsection, we assume $\mathcal{A} = \mathbb{C}^{m\times m}$. Let $\mathcal{X}$ be a locally compact Hausdorff space and let $x_1, \ldots, x_n \in \mathcal{X}^{m\times m}$ and $y_1, \ldots, y_n \in \mathcal{A}$ be given samples. We assume there exist functions $f_{j,l} : \mathcal{X} \to \mathcal{A}$ such that

$$y_i = \sum_{j,l=1}^{m} f_{j,l}((x_i)_{j,l})$$

for $i = 1, \ldots, n$. For example, the $(j,l)$-element of each $x_i$ describes an effect of the $l$-th element on the $j$-th element of $x_i$ and $f_{j,l}$ is a nonlinear function describing an impact of the effect to the value $y_i$. If the given samples $y_i$ are real or complex-valued, we can regard them as $y_i 1_{\mathcal{A}}$ to meet the above setting. Let $\mu_x \in \mathcal{D}(\mathcal{X}, \mathbb{C}^{m\times m})$ be a $\mathbb{C}^{m\times m}$-valued measure defined as $(\mu_x)_{j,l} = \tilde{\delta}_{x_{j,l}}$, where $\tilde{\delta}_x$ for $x \in \mathcal{X}$ is the standard (complex-valued) Dirac measure centered at $x$. Note that the $(j,l)$-element of $\mu_x$ describes a measure regarding the element $x_{j,l}$. Let $k$ be an $\mathcal{A}$-valued $c_0$-kernel (see Definition 5.2), let $\mathcal{M}_k$ be the RKHM associated with $k$, and let $\Phi$ be the KME defined in Section 5.1. In addition, let $\mathcal{V}$ be the submodule

of $\mathcal{M}_k$ spanned by $\{\Phi(\mu_{x_1}), \ldots, \Phi(\mu_{x_n})\}$, and let $P_f : \mathcal{V} \to \mathbb{C}^{m\times m}$ be a $\mathbb{C}^{m\times m}$-linear map (see Definition 2.19) which satisfies

$$P_f \Phi(\mu_{x_i}) = \sum_{j,l=1}^{m} f_{j,l}((x_i)_{j,l})$$

for $i = 1, \ldots, n$. Here, we assume the vectors $\Phi(\mu_{x_1}), \ldots, \Phi(\mu_{x_n})$ are $\mathbb{C}^{m\times m}$-linearly independent (see Definition 2.20).

### 6.3.2 Generalization to the continuous case

We generalize the setting mentioned in Subsection 6.3.1 to the case of functional data. We assume Assumption 5.3 in this subsection. We set $\mathcal{A}$ as $\mathcal{B}(L^2[0,1])$ instead of $\mathbb{C}^{m\times m}$ in this subsection. Let $x_1, \ldots, x_n \in C([0,1]\times[0,1], \mathcal{X})$ and $y_1, \ldots, y_n \in \mathcal{A}$ be given samples. We assume there exists an integrable function $f : [0,1]\times[0,1]\times\mathcal{X} \to \mathcal{A}$ such that

$$y_i = \int_0^1 \int_0^1 f(s,t,x_i(s,t))dsdt$$

for $i = 1, \ldots, n$. We consider an $\mathcal{A}$-valued positive definite kernel $k$ on $\mathcal{X}$, the RKHM $\mathcal{M}_k$ associated with $k$, and the KME $\Phi$ in $\mathcal{M}_k$. Let $\mu_x \in \mathcal{D}(\mathcal{X}, \mathcal{B}(L^2([0,1])))$ be a $\mathcal{B}(L^2([0,1]))$-valued measure defined as $\mu_x(E)v = \langle \chi_E(x(s,\cdot)), v\rangle_{L^2([0,1])}$ for a Borel set $E$ on $\mathcal{X}$. Here, $\chi_E : \mathcal{X} \to \{0,1\}$ is the indicator function for $E$. Note that $\mu_x(E)$ is an integral operator whose integral kernel is $\chi_E(x(s,t))$, which corresponds to the Dirac measure $\tilde{\delta}_{x(s,t)}(E)$. Let $\mathcal{V}$ be the submodule of $\mathcal{M}_k$ spanned by $\{\Phi(\mu_{x_1}), \ldots, \Phi(\mu_{x_n})\}$, and let $P_f : \mathcal{V} \to \mathcal{B}(L^2([0,1]))$ be a $\mathcal{B}(L^2([0,1]))$-linear map (see Definition 2.19) which satisfies

$$P_f \Phi(\mu_{x_i}) = \int_0^1 \int_0^1 f(s,t,x_i(s,t))dsdt$$

for $i = 1, \ldots, n$. Here, we assume the vectors $\Phi(\mu_{x_1}), \ldots, \Phi(\mu_{x_n})$ are $\mathcal{B}(L^2([0,1]))$-linearly independent (see Definition 2.20).

We estimate $P_f$ by restricting it to a submodule of $\mathcal{V}$. For this purpose, we apply the PCA in RKHMs proposed in Section 6.1 and obtain principal axes $p_1, \ldots, p_r$ to construct the submodule. We replace $\phi(x_i)$ in the problem (8) with $\Phi(\mu_{x_i})$ and consider the problem

$$\inf_{\{p_j\}_{j=1}^r \subseteq \mathcal{M}_k:\ \mathrm{ONS}} \sum_{i=1}^{n} \left\| \Phi(\mu_{x_i}) - \sum_{j=1}^{r} p_j \langle p_j, \Phi(\mu_{x_i})\rangle_{\mathcal{M}_k} \right\|_{\mathcal{M}_k}^2. \tag{27}$$

The projection operator onto the submodule spanned by $p_1, \ldots, p_r$ is represented as $QQ^*$, where $Q = [p_1, \ldots, p_r]$. Therefore, we estimate $P_f$ by $P_f QQ^*$. We can compute $P_f QQ^*$ as follows.

**Proposition 6.17** *The solution of the problem* (27) *is represented as* $p_j = \sum_{i=1}^n \Phi(\mu_{x_i}) c_{i,j}$ *for some* $c_{i,j} \in \mathcal{A}$. *Let* $C = [c_{i,j}]_{i,j}$. *Then, the estimation* $P_f QQ^*$ *is computed as*

$$P_f QQ^* = [y_1, \ldots, y_n] CQ^*.$$

The following proposition shows we can obtain a vector which attains the largest transformation by $P_f$.

**Proposition 6.18** *Let $u \in \mathcal{M}_k$ be a unique vector satisfying for any $v \in \mathcal{M}_k$, $\langle u, v\rangle_{\mathcal{M}_k} = P_f QQ^* v$. For $\epsilon > 0$, let $b_\epsilon = (|u|_{\mathcal{M}_k} + \epsilon 1_{\mathcal{A}})^{-1}$ and let $v_\epsilon = ub_\epsilon$. Then, $P_f QQ^* v_\epsilon$ converges to*

$$\sup_{v \in \mathcal{M}_k,\ \|v\|_{\mathcal{M}_k} \leq 1} P_f QQ^* v \tag{28}$$

*as $\epsilon \to 0$, where the supremum is taken with respect to a (pre) order in $\mathcal{A}$ (see Definition 2.9). If $\mathcal{A} = \mathbb{C}^{m\times m}$, then the supremum is replaced with the maximum. In this case, let $|u|^2_{\mathcal{M}_k} = a^* d a$ be the eigenvalue decomposition of the positive semi-definite matrix $|u|^2_{\mathcal{M}_k}$ and let $b = a^* d^+ a$, where the $i$-th diagonal element of $d^+$ is $d_{i,i}^{-1/2}$ if $d_{i,i} \neq 0$ and 0 if $d_{i,i} = 0$. Then, $ub$ is the solution of the maximization problem.*

**Proof** By the Riesz representation theorem (Proposition 4.2), there exists a unique $u \in \mathcal{M}_k$ satisfying for any $v \in \mathcal{M}_k$, $\langle u, v\rangle_{\mathcal{M}_k} = P_f QQ^* v$. Then, for $v \in \mathcal{M}_k$ which satisfies $\|v\|_{\mathcal{M}_k} = 1$, by the Cauchy–Schwarz inequality (Lemma 2.16), we have

$$P_f QQ^* v = \langle u, v\rangle_{\mathcal{M}_k} \leq_{\mathcal{A}} |u|_{\mathcal{M}_k} \|v\|_{\mathcal{M}_k} \leq_{\mathcal{A}} |u|_{\mathcal{M}_k}. \tag{29}$$

The vector $v_\epsilon$ satisfies $\|v_\epsilon\|_{\mathcal{M}_k} \leq 1$. In addition, we have

$$|u|^2_{\mathcal{M}_k} - (|u|^2_{\mathcal{M}_k} - \epsilon^2 1_{\mathcal{A}}) \geq_{\mathcal{A}} 0.$$

By multiplying $(|u|_{\mathcal{M}_k} + \epsilon 1_{\mathcal{A}})^{-1}$ on the both sides, we have $\langle u, v_\epsilon\rangle_{\mathcal{M}_k} + \epsilon 1_{\mathcal{A}} - |u|_{\mathcal{M}_k} \geq_{\mathcal{A}} 0$, which implies $\| |u|_{\mathcal{M}_k} - \langle u, v_\epsilon\rangle_{\mathcal{M}_k} \|_{\mathcal{A}} \leq \epsilon$, and $\lim_{\epsilon\to 0} P_f QQ^* v_\epsilon = \lim_{\epsilon\to 0} \langle u, v_\epsilon\rangle_{\mathcal{M}_k} = |u|_{\mathcal{M}_k}$. Since $\langle u, v_\epsilon\rangle_{\mathcal{M}_k} \leq_{\mathcal{A}} d$ for any upper bound $d$ of $\{\langle u, v\rangle_{\mathcal{M}_k} \mid \|v\|_{\mathcal{M}_k} \leq 1\}$, $|u|_{\mathcal{M}_k} \leq_{\mathcal{A}} d$ holds. As a result, $|u|_{\mathcal{M}_k}$ is the supremum of $P_f QQ^* v$. In the case of $\mathcal{A} = \mathbb{C}^{m\times m}$, the inequality (29) is replaced with the equality by setting $v = ub$. ■

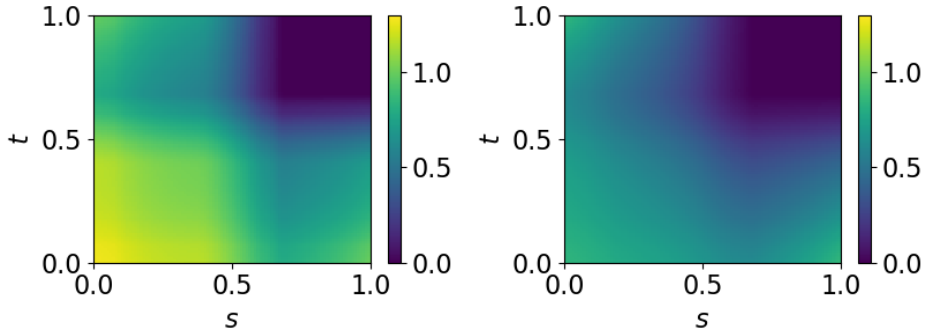
The vector $ub_\epsilon$ is represented as $ub_\epsilon = QC^*[y_1, \ldots, y_n]^T b_\epsilon = \sum_{i=1}^n \Phi(\mu_{x_i}) d_i$, where $d_i \in \mathcal{A}$ is the $i$-th element of $CC^*[y_1, \ldots, y_n]^T b_\epsilon \in \mathcal{A}^n$, and $\Phi$ is $\mathcal{A}$-linear (see Proposition 5.8). Therefore, the vector $ub_\epsilon$ corresponds to the $\mathcal{A}$-valued measure $\sum_{i=1}^n \mu_{x_i} d_i$, and if $\Phi$ is injective (see Example 5.13), the corresponding measure is unique. This means that if we transform the samples $x_i$ according to the measure $\sum_{i=1}^n \mu_{x_i} d_i$, then the transformation makes a large impact to $y_i$.

### 6.3.3 Numerical examples

We applied our method to functional data $x_1, \ldots, x_n \in C([0,1]\times[0,1], [0,1])$, where $n = 30$, $x_i$ are polynomials of the form $x_i(s,t) = \sum_{j,l=0}^5 \eta_{j,l} s^j t^l$. The coefficients $\eta_{j,l}$ of $x_i$ are randomly and independently drawn from the uniform distribution on $[0, 0.1]$. Then, we set $y_i \in \mathbb{R}$ as

$$y_i = \int_0^1 \int_0^1 x_i(s,t)^{-\alpha + \alpha|s+t|} ds dt$$

for $\alpha = 3, 0.5$. We set $\mathcal{A} = \mathcal{B}(L^2([0,1]))$ and $k(x_1, x_2) = \tilde{k}(x_1, x_2) 1_{\mathcal{A}}$, where $\tilde{k}$ is a complex-valued positive definite kernel on $[0,1]$ defined as $\tilde{k}(x_1, x_2) = e^{-\|x_1 - x_2\|_2^2}$. We applied the




(a) $\alpha = 3$ (b) $\alpha = 0.5$

Figure 8: Heat map representing the value the integral kernel of $\nu([0, 1])$

PCA proposed in Subsection 6.1.3 with $r = 3$, and then computed $\lim_{\epsilon\to 0} ub_\epsilon \in \mathcal{M}_k$ in Proposition 6.17, which can be represented as $\Phi(\sum_{i=1}^n \mu_{x_i} d_i)$ for some $d_i \in \mathcal{A}$. The parameter $\lambda$ in the objective function of the PCA was set as 0.5. Figure 8 shows the heat map representing the value related to the integral kernel of the $\mathcal{A}$-valued measure $\sum_{i=1}^n \mu_{x_i}(E)d_i$ for $E = [0, 0.1]$. We denote $\sum_{i=1}^n \mu_{x_i}(E)d_i$ by $\nu(E)$ and the integral kernel of the integral operator $\nu(E)$ by $\tilde{k}_{\nu(E)}$. As we stated in Section 6.3.2, if we transform the samples $x_i$ according to the measure $\nu$, then the transformation makes a large impact to $y_i$. Moreover, the value of $\tilde{k}_{\nu(E)}$ at $(s, t)$ corresponds to the measure at $(s, t)$. Therefore, the value of $\tilde{k}_{\nu(E)}$ at $(s, t)$ describes the impact of the effect of $t$ on $s$ to $y_i$. To additionally take the effect of $s$ on $t$ into consideration, we show the value of $\tilde{k}_{\nu(E)}(s, t) + \tilde{k}_{\nu(E)}(t, s)$ in Figure 8. The values for $\alpha = 3$ are larger than those for $\alpha = 0.5$, which implies the overall impacts to $y_i$ for $\alpha = 3$ are larger than that for $\alpha = 0.5$. Moreover, the value is large if $s + t$ is small. This is because for $x_i(s, t) \in [0, 0.1]$, $x_i(s, t)^{-\alpha+\alpha|s+t|}$ is large if $s+t$ is small. Furthermore, the values around $(s, t) = (1, 0)$ and $(0, 1)$ are also large since $x_i$ has the form $x_i(s, t) = \sum_{j,l=0}^5 \eta_{j,l} s^j t^l$ for $\eta_{j,l} \in [0, 0.1]$ and $x_i(s, t)$ itself is large around $(s, t) = (1, 0)$ and $(0, 1)$, which results in $x_i(s, t)^{-\alpha+\alpha|s+t|} \approx x_i(s, t)$ being large.

## 6.4 Other applications

### 6.4.1 Maximum mean discrepancy with kernel mean embedding

Maximum mean discrepancy (MMD) is a metric of measures according to the largest difference in means over a certain subset of a function space. It is also known as integral probability metric (IPM). For a set $\mathcal{U}$ of real-valued bounded measurable functions on $\mathcal{X}$ and two real-valued probability measures $\mu$ and $\nu$, MMD $\gamma(\mu, \nu, \mathcal{U})$ is defined as follows (Müller, 1997; Gretton et al., 2012):

$$\sup_{u\in\mathcal{U}} \left| \int_{x\in\mathcal{X}} u(x)d\mu(x) - \int_{x\in\mathcal{X}} u(x)d\nu(x) \right|.$$

For example, if $\mathcal{U}$ is the unit ball of an RKHS, denoted as $\mathcal{U}_{\mathrm{RKHS}}$, the MMD can be represented using the KME $\tilde{\Phi}$ in the RKHS as $\gamma(\mu, \nu, \mathcal{U}_{\mathrm{RKHS}}) = \|\tilde{\Phi}(\mu) - \tilde{\Phi}(\nu)\|_{\mathcal{H}_{\tilde{k}}}$. In addition, let $\mathcal{U}_{\mathrm{K}} = \{u \mid \|u\|_L \leq 1\}$ and let $\mathcal{U}_{\mathrm{D}} = \{u \mid \|u\|_\infty + \|u\|_L \leq 1\}$, where,

$\|u\|_L := \sup_{x \neq y} |u(x) - u(y)|/|x - y|$, and $\|u\|_\infty$ is the sup norm of $u$. The MMDs with $\mathcal{U}_{\mathrm{K}}$ and $\mathcal{U}_{\mathrm{D}}$ are also discussed in Rachev (1985); Dudley (2002); Sriperumbudur et al. (2012).

Let $\mathcal{X}$ be a locally compact Hausdorff space, let $\mathcal{U}_{\mathcal{A}}$ be a set of $\mathcal{A}$-valued bounded and measurable functions, and let $\mu, \nu \in \mathcal{D}(\mathcal{X}, \mathcal{A})$. We generalize the MMD to that for $\mathcal{A}$-valued measures as follows:

$$\gamma_{\mathcal{A}}(\mu, \nu, \mathcal{U}_{\mathcal{A}}) := \sup_{u \in \mathcal{U}_{\mathcal{A}}} \left| \int_{x \in \mathcal{X}} u(x) d\mu(x) - \int_{x \in \mathcal{X}} u(x) d\nu(x) \right|_{\mathcal{A}},$$

where the supremum is taken with respect to a (pre) order in $\mathcal{A}$ (see Definition 2.9). Let $k$ be an $\mathcal{A}$-valued positive definite kernel and let $\mathcal{M}_k$ be the RKHM associated with $k$. We assume Assumption 5.3. Let $\Phi$ be the KME defined in Section 5.1. The following theorem shows that similar to the case of RKHS, if $\mathcal{U}_{\mathcal{A}}$ is the unit ball of an RKHM, the generalized MMD $\gamma_{\mathcal{A}}(\mu, \nu, \mathcal{U}_{\mathcal{A}})$ can also be represented using the proposed KME in the RKHM.

**Proposition 6.19** *Let* $\mathcal{U}_{\mathrm{RKHM}} := \{u \in \mathcal{M}_k \mid \|u\|_{\mathcal{M}_k} \leq 1\}$. *Then, for* $\mu, \nu \in \mathcal{D}(\mathcal{X}, \mathcal{A})$, *we have*

$$\gamma_{\mathcal{A}}(\mu, \nu, \mathcal{U}_{\mathrm{RKHM}}) = |\Phi(\mu) - \Phi(\nu)|_{\mathcal{M}_k}.$$

**Proof** By the Cauchy–Schwarz inequality (Lemma 2.16), we have

$$\begin{aligned} \left| \int_{x \in \mathcal{X}} d\mu^* u(x) - \int_{x \in \mathcal{X}} d\nu^* u(x) \right|_{\mathcal{A}} &= |\langle \Phi(\mu - \nu), u \rangle_{\mathcal{M}_k}|_{\mathcal{A}} \\ \leq_{\mathcal{A}} \|u\|_{\mathcal{M}_k} |\Phi(\mu - \nu)|_{\mathcal{M}_k} &\leq_{\mathcal{A}} |\Phi(\mu - \nu)|_{\mathcal{M}_k} \end{aligned}$$

for any $u \in \mathcal{M}_k$ such that $\|u\|_{\mathcal{M}_k} \leq 1$. Let $\epsilon > 0$. We put $v = \Phi(\mu - \nu)$ and $u_\epsilon = v(|v|_{\mathcal{M}_k} + \epsilon 1_{\mathcal{A}})^{-1}$. In the same manner as Proposition 6.18, $|\Phi(\mu - \nu)|_{\mathcal{M}_k}$ is shown to be the supremum of $|\int_{x \in \mathcal{X}} d\mu^* u(x) - \int_{x \in \mathcal{X}} d\nu^* u(x)|_{\mathcal{A}}$. ■

Various methods with the existing MMD of real-valued probability measures are generalized to $\mathcal{A}$-valued measures by applying our MMD. Using our MMD of $\mathcal{A}$-valued measures instead of the existing MMD allows us to evaluate discrepancies between measures regarding each point of structured data such as multivariate data and functional data. For example, the following existing methods can be generalized:

**Two-sample test:** In two-sample test, samples from two distributions (measures) are compared by computing the MMD of these measures (Gretton et al., 2012).

**Kernel mean matching for generative models:** In generative models, MMD is used in finding points whose distribution is as close as that of input points (Jitkrittum et al., 2019).

**Domain adaptation:** In domain adaptation, MMD is used in describing the difference between the distribution of target domain data and that of source domain data (Li et al., 2019).

### 6.4.2 Time-series data analysis with random noise

Recently, random dynamical systems, which are (nonlinear) dynamical systems with random effects, have been extensively researched. Analyses of them by generalizing the discussion

mentioned in Subsection 6.2.1 using the existing KME in RKHSs have been proposed (Klus et al., 2020; Hashimoto et al., 2020). We can apply our KME of $\mathcal{A}$-valued measures to generalize the analysis proposed in Subsection 6.2.2 to random dynamical systems. Then, we can extract continuous behaviors of the time evolution of functions with consideration of random noise.

## 7. Connection with existing methods

In this section, we discuss connections between the proposed methods and existing methods. We show the connection with the PCA in vvRKHSs in Subsection 7.1 and an existing notion in quantum mechanics.

### 7.1 Connection with PCA in vvRKHSs

We show that PCA in vvRKHSs is a special case of the proposed PCA in RKHMs. Let $\mathcal{W}$ be a Hilbert space and we set $\mathcal{A} = \mathcal{B}(\mathcal{W})$. Let $k : \mathcal{X} \times \mathcal{X} \to \mathcal{B}(\mathcal{W})$ be a $\mathcal{B}(\mathcal{W})$-valued positive definite kernel. In addition, let $x_1, \ldots, x_n \in \mathcal{X}$ be given data and $w_{1,1}, \ldots, w_{1,N}, \ldots, w_{n,1}, \ldots, w_{n,N} \in \mathcal{W}$ be fixed vectors in $\mathcal{W}$. The following proposition shows that we can reconstruct principal components of PCA in vvRKHSs by using the proposed PCA in RKHMs.

**Proposition 7.1** *Let $W_j : \mathcal{X} \to \mathcal{W}$ be a map satisfying $W_j(x_i) = w_{i,j}$ for $j = 1, \ldots, N$, let $W = [W_1, \ldots, W_N]$, and let $\hat{k} : \mathcal{X} \times \mathcal{X} \to \mathbb{C}^{N \times N}$ be defined as $\hat{k}(x, y) = W(x)^* k(x, y) W(y)$. Let $\{q_1, \ldots, q_r\} \subseteq \mathcal{F}_{\hat{k}}$ is a solution of the minimization problem*

$$\min_{\{q_j\}_{j=1}^r \subseteq \mathcal{F}_{\hat{k}} :\ ONS} \sum_{i=1}^{n} \mathrm{tr}\left( \left| \phi(x_i) - \sum_{j=1}^{r} q_j \left\langle q_j, \phi(x_i) \right\rangle_{\mathcal{M}_{\hat{k}}} \right|^2_{\mathcal{M}_{\hat{k}}} \right), \tag{30}$$

*where $\mathcal{F}_k = \{v \in \mathcal{M}_k \mid \ v(x)$ is a rank 1 operator for any $x \in \mathcal{X}\}$. In addition, let $p_1, \ldots, p_r \in \mathcal{H}_k^{\mathrm{v}}$ be the solution of the minimization problem*

$$\min_{\{p_j\}_{j=1}^r \subseteq \mathcal{H}_k^{\mathrm{v}} :\ ONS} \sum_{i=1}^{n} \sum_{l=1}^{N} \left\| \phi(x_i) w_{i,l} - \sum_{j=1}^{r} p_j \left\langle p_j, \phi(x_i) w_{i,l} \right\rangle_{\mathcal{H}_k^{\mathrm{v}}} \right\|^2_{\mathcal{H}_k^{\mathrm{v}}}. \tag{31}$$

*Then, $\|(\langle q_j, \hat{\phi}(x_i) \rangle_{\mathcal{M}_{\hat{k}}})_l\|_{\mathbb{C}^N} = \langle p_j, \phi(x_i) w_{i,l} \rangle_{\mathcal{H}_k^{\mathrm{v}}}$ for $i = 1, \ldots, n$, $j = 1, \ldots, r$, and $l = 1, \ldots, N$. Here, $(\langle q_j, \hat{\phi}(x_i) \rangle_{\mathcal{M}_{\hat{k}}})_l$ is the $l$-th column of the matrix $\langle q_j, \hat{\phi}(x_i) \rangle_{\mathcal{M}_{\hat{k}}} \in \mathbb{C}^{N \times N}$.*

**Proof** Let $\mathbf{G} \in (\mathbb{C}^{N \times N})^{n \times n}$ be defined as $\mathbf{G}_{i,j} = \hat{k}(x_i, x_j)$. By Proposition 6.8, any solution of the problem (30) is represented as $q_j = \sum_{i=1}^{n} \hat{\phi}(x_i) c_{i,j}$, where $j = 1, \ldots, r$ and $[c_{1,j}, \ldots, c_{n,j}]^T = \lambda_j^{-1/2} \mathbf{v}_j u^*$ for any normalized vector $u \in \mathbb{C}^N$. Here, $\lambda_j$ are the largest $r$ eigenvalues and $\mathbf{v}_j$ are the corresponding orthonormal eigenvectors of the matrix $\mathbf{G}$. Therefore, by the definition of $\hat{k}$, the principal components are calculated as

$$\langle q_j, \hat{\phi}(x_i) \rangle^*_{\mathcal{M}_{\hat{k}}} = \lambda_j^{-1/2} W(x_i)^* [k(x_i, x_1) W(x_1), \ldots, k(x_i, x_n) W(x_n)] \mathbf{v}_j u^*.$$

On the other hand, in the same manner as Proposition 6.8, the solution of the problem (31) is shown to be represented as $p_j = \sum_{i=1}^{n} \sum_{l=1}^{N} \phi(x_i) w_{i,l} \alpha_{(i-1)N+l,j}$, where $j = 1, \ldots, r$ and $[\alpha_{1,j}, \ldots, \alpha_{Nn,j}]^T = \lambda_j^{-1/2} \mathbf{v}_j$. Therefore, the principal components are calculated as

$$\overline{\langle p_j, \phi(x_i) w_{i,l} \rangle_{\mathcal{H}_k^{\mathrm{v}}}} = \lambda_j^{-1/2} W_l(x_i)^* [k(x_i, x_1) W(x_1), \ldots, k(x_i, x_n) W(x_n)] \mathbf{v}_j,$$

which completes the proof of the proposition. ■

### 7.2 Connection with quantum mechanics

Positive operator-valued measures play an important role in quantum mechanics. A positive operator-valued measure is defined as an $\mathcal{A}$-valued measure $\mu$ such that $\mu(\mathcal{X}) = I$ and $\mu(E)$ is positive for any Borel set $E$. It enables us to extract information of the probabilities of outcomes from a state (Peres and Terno, 2004; Holevo, 2011). We show that the existing inner product considered for quantum states (Balkir, 2014; Deb, 2016) is generalized with our KME of positive operator-valued measures.

Let $\mathcal{X} = \mathbb{C}^m$ and $\mathcal{A} = \mathbb{C}^{m \times m}$. Let $\rho \in \mathcal{A}$ be a positive semi-definite matrix with unit trace, called a density matrix. A density matrix describes the states of a quantum system, and information about outcomes is described as measure $\mu\rho \in \mathcal{D}(\mathcal{X}, \mathcal{A})$. We have the following proposition. Here, we use the bra-ket notation, i.e., $|\alpha\rangle \in \mathcal{X}$ represents a (column) vector in $\mathcal{X}$, and $\langle \alpha|$ is defined as $\langle \alpha| = |\alpha\rangle^*$:

**Proposition 7.2** *Assume $\mathcal{X} = \mathbb{C}^m$, $\mathcal{A} = \mathbb{C}^{m \times m}$, and $k : \mathcal{X} \times \mathcal{X} \to \mathcal{A}$ is a positive definite kernel defined as $k(|\alpha\rangle, |\beta\rangle) = |\alpha\rangle\langle\alpha|\beta\rangle\langle\beta|$. If $\mu$ is represented as $\mu = \sum_{i=1}^{m} \delta_{|\psi_i\rangle} |\psi_i\rangle\langle\psi_i|$ for an orthonormal basis $\{|\psi_1\rangle, \ldots, |\psi_m\rangle\}$ of $\mathcal{X}$, then for any $\rho_1, \rho_2 \in \mathcal{A}$, $\mathrm{tr}(\langle \Phi(\mu\rho_1), \Phi(\mu\rho_2) \rangle_{\mathcal{M}_k}) = \langle \rho_1, \rho_2 \rangle_{\mathrm{HS}}$ holds. Here, $\langle \cdot, \cdot \rangle_{\mathrm{HS}}$ is the Hilbert–Schmidt inner product.*

**Proof** Let $M_i = |\psi_i\rangle\langle\psi_i|$ for $i = 1, \ldots, m$. The inner product between $\Phi(\mu\rho_1)$ and $\Phi(\mu\rho_2)$ is calculated as follows:

$$\langle \Phi(\mu\rho_1), \Phi(\mu\rho_2) \rangle_{\mathcal{M}_k} = \int_{x \in \mathcal{X}} \int_{y \in \mathcal{X}} \rho_1^* \mu^*(x) k(x, y) \mu\rho_2(y) = \sum_{i,j=1}^{m} \rho_1^* M_i k(|\psi_i\rangle, |\psi_j\rangle) M_j \rho_2.$$

Since the identity $k(|\psi_i\rangle, |\psi_j\rangle) = M_i M_j$ holds and $\{|\psi_1\rangle, \ldots, |\psi_m\rangle\}$ is orthonormal, we have $\langle \Phi(\mu\rho_1), \Phi(\mu\rho_2) \rangle_{\mathcal{M}_k} = \sum_{i=1}^{m} \rho_1^* M_i \rho_2$. By using the identity $\sum_{i=1}^{m} M_i = I$, we have

$$\mathrm{tr}\left( \sum_{i=1}^{m} \rho_1^* M_i \rho_2 \right) = \mathrm{tr}\left( \sum_{i=1}^{m} M_i \rho_2 \rho_1^* \right) = \mathrm{tr}(\rho_2 \rho_1^*),$$

which completes the proof of the proposition. ■

In previous studies (Balkir, 2014; Deb, 2016), the Hilbert–Schmidt inner product between density matrices was considered to represent similarities between two quantum states. Liu and Rebentrost (2018) considered the Hilbert–Schmidt inner product between square roots of density matrices. Theorem 7.2 shows that these inner products are represented via our KME in RKHMs.

## 8. Conclusions and future works

In this paper, we proposed a new data analysis framework with RKHM and developed a KME in RKHMs for analyzing distributions. We showed the theoretical validity for applying those to data analysis. Then, we applied it to kernel PCA, time-series data analysis, and analysis of interaction effects in finite or infinite dimensional data. RKHM is a generalization of RKHS in terms of $C^*$-algebra, and we can extract rich information about structures in data such as functional data by using $C^*$-algebras. For example, we can reduce multivariable functional data to functions of single variable by considering the space of functions of single variables as a $C^*$-algebra and then by applying the proposed PCA in RKHMs. Moreover, we can extract information of interaction effects in continuously distributed spatio data by considering the space of bounded linear operators on a function space as a $C^*$-algebra.

As future works, we will address $C^*$-algebra-valued supervised problems on the basis of the representer theorem (Theorem 4.8) and apply the proposed KME in RKHMs to quantum mechanics.

## Acknowledgments

We would like to thank the anonymous referees and action editor Corinna Cortes, whose comments improve the manuscript significantly. This work was partially supported by JST CREST Grant Number JPMJCR1913.

## Appendix A. Proofs of the lemmas and propositions in Section 2.5

### Proof of Proposition 4.5

(Existence) For $u, v \in \mathcal{M}_k$, there exist $u_i, v_i \in \mathcal{M}_{k,0}$ $(i = 1, 2, \ldots)$ such that $v = \lim_{i\to\infty} v_i$ and $w = \lim_{i\to\infty} w_i$. By the Cauchy-Schwarz inequality (Lemma 2.16), the following inequalities hold:

$$\begin{aligned}\| \langle u_i, v_i\rangle_{\mathcal{M}_k} - \langle u_j, v_j\rangle_{\mathcal{M}_k} \|_{\mathcal{A}} &\leq \| \langle u_i, v_i - v_j\rangle_{\mathcal{M}_k} \|_{\mathcal{A}} + \| \langle u_i - u_j, u_j\rangle_{\mathcal{M}_k} \|_{\mathcal{A}} \\ &\leq \|u_i\|_{\mathcal{M}_k} \|v_i - v_j\|_{\mathcal{M}_k} + \|u_i - u_j\|_{\mathcal{M}_k} \|v_j\|_{\mathcal{M}_k} \\ &\to 0 \ (i, j \to \infty),\end{aligned}$$

which implies $\{\langle u_i, v_i\rangle_{\mathcal{M}_k}\}_{i=1}^{\infty}$ is a Cauchy sequence in $\mathcal{A}$. By the completeness of $\mathcal{A}$, there exists a limit $\lim_{i\to\infty} \langle u_i, v_i\rangle_{\mathcal{M}_k}$.

(Well-definedness) Assume there exist $u_i', v_i' \in \mathcal{M}_{k,0}$ $(i = 1, 2, \ldots)$ such that $u = \lim_{i\to\infty} u_i = \lim_{i\to\infty} u_i'$ and $v = \lim_{i\to\infty} v_i = \lim_{i\to\infty} v_i'$. By the Cauchy-Schwarz inequality (Lemma 2.16), we have

$$\| \langle u_i, v_i\rangle_{\mathcal{M}_k} - \langle u_i', v_i'\rangle_{\mathcal{M}_k} \|_{\mathcal{A}} \leq \|u_i\|_{\mathcal{M}_k}\|v_i - v_i'\|_{\mathcal{M}_k} + \|u_i - u_i'\|_{\mathcal{M}_k}\|v_i'\|_{\mathcal{M}_k} \to 0 \ (i \to \infty),$$

which implies $\lim_{i\to\infty} \langle u_i, v_i\rangle_{\mathcal{M}_k} = \lim_{i\to\infty} \langle u_i', v_i'\rangle_{\mathcal{M}_k}$.

(Injectivity) For $u, v \in \mathcal{M}_k$, we assume $\langle \phi(x), u\rangle_{\mathcal{M}_k} = \langle \phi(x), v\rangle_{\mathcal{M}_k}$ for $x \in \mathcal{X}$. By the linearity of $\langle \cdot, \cdot\rangle_{\mathcal{M}_k}$, $\langle p, u\rangle_{\mathcal{M}_k} = \langle p, v\rangle_{\mathcal{M}_k}$ holds for $p \in \mathcal{M}_{k,0}$. For $p \in \mathcal{M}_k$, there

exist $p_i \in \mathcal{M}_{k,0}$ $(i = 1, 2, \ldots)$ such that $p = \lim_{i\to\infty} p_i$. Therefore, $\langle p, u - v\rangle_{\mathcal{M}_k} = \lim_{i\to\infty} \langle p_i, u - v\rangle_{\mathcal{M}_k} = 0$. As a result, $\langle u - v, u - v\rangle_{\mathcal{M}_k} = 0$ holds by setting $p = u - v$, which implies $u = v$.

**Proof of Proposition 4.6**

We define $\Psi : \mathcal{M}_{k,0} \to \mathcal{M}$ as an $\mathcal{A}$-linear map that satisfies $\Psi(\phi(x)) = \psi(x)$. We show $\Psi$ can be extended to a unique $\mathcal{A}$-linear bijection map on $\mathcal{M}_k$ , which preserves the inner product.
(Uniqueness) The uniqueness follows by the definition of $\Psi$.

(Inner product preservation) For $x, y \in \mathcal{X}$, we have

$$\langle \Psi(\phi(x)), \Psi(\phi(y))\rangle_{\mathcal{M}_k} = \langle \psi(x), \psi(y)\rangle_{\mathcal{M}} = k(x, y) = \langle \phi(x), \phi(y)\rangle_{\mathcal{M}_k}.$$

Since $\Psi$ is $\mathcal{A}$-linear, $\Psi$ preserves the inner products between arbitrary $u, v \in \mathcal{M}_{k,0}$.

(Well-definedness) Since $\Phi$ preserves the inner product, if $\{v_i\}_{i=1}^{\infty} \subseteq \mathcal{M}_k$ is a Cauchy sequence, $\{\Psi(v_i)\}_{i=1}^{\infty} \subseteq \mathcal{M}$ is also a Cauchy sequence. Therefore, by the completeness of $\mathcal{M}$, $\Psi$ also preserves the inner product in $\mathcal{M}_k$, and for $v \in \mathcal{M}_k$, $\|\Psi(v)\|_{\mathcal{M}} = \|v\|_{\mathcal{M}_k}$ holds. As a result, for $v \in \mathcal{M}_k$, if $v = 0$, $\|\Psi(v)\|_{\mathcal{M}} = \|v\|_{\mathcal{M}_k} = 0$ holds. This implies $\Psi(v) = 0$.

(Injectivity) For $u, v \in \mathcal{M}_k$, if $\Psi(u) = \Psi(v)$, then $0 = \|\Psi(u) - \Psi(v)\|_{\mathcal{M}} = \|u - v\|_{\mathcal{M}_k}$ holds since $\Psi$ preserves the inner product, which implies $u = v$.

(Surjectivity) It follows directly by the condition $\overline{\{\sum_{i=0}^{n} \psi(x_i)c_i \mid \ x_i \in \mathcal{X},\ c_i \in \mathcal{A}\}} = \mathcal{M}$.

**Proof of Lemma 4.10**

Let $k$ be an $\mathcal{A}$-valued positive definite kernel defined in Definition 2.21. Let $w \in \mathcal{W}$. For $n \in \mathbb{N}$, $w_1, \ldots, w_n \in \mathcal{W}$, let $c_i \in \mathcal{B}(\mathcal{W})$ be defined as $c_i h := \langle w, h\rangle_{\mathcal{W}} / \langle w, w\rangle_{\mathcal{W}} w_i$ for $h \in \mathcal{W}$. Since $w_i = c_i w$ holds, the following equalities are derived for $x_1, \ldots, x_n \in \mathcal{X}$:

$$\sum_{i,j=1}^{n} \langle w_i, k(x_i, x_j)w_j\rangle_{\mathcal{W}} = \sum_{i,j=1}^{n} \langle c_i w, k(x_i, x_j)c_j w\rangle_{\mathcal{W}} = \left\langle w, \sum_{i,j=1}^{n} c_i^* k(x_i, x_j)c_i w \right\rangle_{\mathcal{W}}.$$

By the positivity of $\sum_{i,j=1}^{n} c_i^* k(x_i, x_j)c_j$, $\langle w, \sum_{i,j=1}^{n} c_i^* k(x_i, x_j)c_j w\rangle_{\mathcal{W}} \geq 0$ holds, which implies $k$ is an operator valued positive definite kernel defined in Definition 2.2.

On the other hand, let $k$ be an operator valued positive definite kernel defined in Definition 2.2. Let $v \in \mathcal{W}$. For $n \in \mathbb{N}$, $c_1, \ldots, c_n \in \mathcal{A}$ and $x_1, \ldots, x_n \in \mathcal{X}$, the following equality is derived:

$$\left\langle w, \sum_{i,j=1}^{n} c_i^* k(x_i, x_j)c_j w \right\rangle_{\mathcal{W}} = \sum_{i,j=1}^{n} \langle c_i w, k(x_i, x_j)c_j w\rangle_{\mathcal{W}}.$$

By Definition 2.2, $\sum_{i,j=1}^{n} \langle c_i w, k(x_i, x_j)c_j w\rangle_{\mathcal{W}} \geq 0$ holds, which implies $k$ is an $\mathcal{A}$-valued positive definite kernel defined in Definition 2.21.

## Appendix B. $\mathcal{A}$-valued measure and integral

We introduce $\mathcal{A}$-valued measure and integral in preparation for defining a KME in RKHMs. $\mathcal{A}$-valued measure and integral are special cases of vector measure and integral (Dinculeanu, 1967, 2000), respectively. Here, we review these notions especially for the case of $\mathcal{A}$-valued ones. The notions of measures and the Lebesgue integrals are generalized to $\mathcal{A}$-valued. The *left and right integral of an $\mathcal{A}$-valued function $u$ with respect to an $\mathcal{A}$-valued measure $\mu$* is defined through $\mathcal{A}$-valued step functions.

**Definition B.1 ($\mathcal{A}$-valued measure)** *Let $\Sigma$ be a $\sigma$-algebra on $\mathcal{X}$.*

1. *An $\mathcal{A}$-valued map $\mu : \Sigma \to \mathcal{A}$ is called a* (countably additive) $\mathcal{A}$-vaued measure *if $\mu(\bigcup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} \mu(E_i)$ for all countable collections $\{E_i\}_{i=1}^{\infty}$ of pairwise disjoint sets in $\Sigma$.*

2. *An $\mathcal{A}$-valued measure $\mu$ is said to be finite if $|\mu|(E) := \sup\{\sum_{i=1}^{n} \|\mu(E_i)\|_{\mathcal{A}} \mid \ n \in \mathbb{N},\ \{E_i\}_{i=1}^{n}$ is a finite partition of $E \in \Sigma\} < \infty$. We call $|\mu|$ the total variation of $\mu$.*

3. *An $\mathcal{A}$-valued measure $\mu$ is said to be regular if for all $E \in \Sigma$ and $\epsilon > 0$, there exist a compact set $K \subseteq E$ and an open set $G \supseteq E$ such that $\|\mu(F)\|_{\mathcal{A}} \leq \epsilon$ for any $F \subseteq G \backslash K$. The regularity corresponds to the continuity of $\mathcal{A}$-valued measures.*

4. *An $\mathcal{A}$-valued measure $\mu$ is called a* Borel measure *if $\Sigma = \mathcal{B}$, where $\mathcal{B}$ is the Borel $\sigma$-algebra on $\mathcal{X}$ ($\sigma$-algebra generated by all compact subsets of $\mathcal{X}$).*

*The set of all $\mathcal{A}$-valued finite regular Borel measures is denoted as $\mathcal{D}(\mathcal{X}, \mathcal{A})$.*

**Definition B.2 ($\mathcal{A}$-valued Dirac measure)** *For $x \in \mathcal{X}$, we define $\delta_x \in \mathcal{D}(\mathcal{X}, \mathcal{A})$ as $\delta_x(E) = 1_{\mathcal{A}}$ for $x \in E$ and $\delta_x(E) = 0$ for $x \notin E$. The measure $\delta_x$ is referred to as the* $\mathcal{A}$-valued Dirac measure *at $x$.*

Similar to the Lebesgue integrals, an integral of an $\mathcal{A}$-valued function with respect to an $\mathcal{A}$-valued measure is defined through $\mathcal{A}$-valued step functions.

**Definition B.3 (Step function)** *An $\mathcal{A}$-valued map $s : \mathcal{X} \to \mathcal{A}$ is called a* step function *if $s(x) = \sum_{i=1}^{n} c_i \chi_{E_i}(x)$ for some $n \in \mathbb{N}$, $c_i \in \mathcal{A}$ and finite partition $\{E_i\}_{i=1}^{n}$ of $\mathcal{X}$, where $\chi_E : \mathcal{X} \to \{0, 1\}$ is the indicator function for $E \in \mathcal{B}$. The set of all $\mathcal{A}$-valued step functions on $\mathcal{X}$ is denoted as $\mathcal{S}(\mathcal{X}, \mathcal{A})$.*

**Definition B.4 (Integrals of functions in $\mathcal{S}(\mathcal{X}, \mathcal{A})$)** *For $s \in \mathcal{S}(\mathcal{X}, \mathcal{A})$ and $\mu \in \mathcal{D}(\mathcal{X}, \mathcal{A})$, the* left and right integrals of $s$ with respect to $\mu$ *are respectively defined as*

$$\int_{x \in \mathcal{X}} s(x) d\mu(x) := \sum_{i=1}^{n} c_i \mu(E_i), \quad \int_{x \in \mathcal{X}} d\mu(x) s(x) := \sum_{i=1}^{n} \mu(E_i) c_i.$$

As we explain below, the integrals of step functions are extended to those of "integrable functions". For a real positive finite measure $\nu$, let $L^1_\nu(\mathcal{X}, \mathcal{A})$ be the set of all $\mathcal{A}$-valued $\nu$-Bochner integrable functions on $\mathcal{X}$, i.e., if $u \in L^1_\nu(\mathcal{X}, \mathcal{A})$, there exists a sequence $\{s_i\}_{i=1}^{\infty} \subseteq \mathcal{S}(\mathcal{X}, \mathcal{A})$ of step functions such that $\lim_{i \to \infty} \int_{x \in \mathcal{X}} \|u(x) - s_i(x)\|_{\mathcal{A}} d\nu(x) = 0$ (Diestel, 1984,

Chapter IV). Note that $u \in L^1_\nu(\mathcal{X}, \mathcal{A})$ if and only if $\int_{x\in\mathcal{X}} \|u(x)\|_{\mathcal{A}} d\nu(x) < \infty$, and $L^1_\nu(\mathcal{X}, \mathcal{A})$ is a Banach $\mathcal{A}$-module (i.e., a Banach space equipped with an $\mathcal{A}$-module structure) with respect to the norm defined as $\|u\|_{L^1_\nu(\mathcal{X},\mathcal{A})} = \int_{x\in\mathcal{X}} \|u(x)\|_{\mathcal{A}} d\nu(x)$.

**Definition B.5 (Integrals of functions in $L^1_{|\mu|}(\mathcal{X}, \mathcal{A})$)** *For $u \in L^1_{|\mu|}(\mathcal{X}, \mathcal{A})$, the* left and right integrals of $u$ with respect to $\mu$ *is respectively defined as*

$$\lim_{i\to\infty} \int_{x\in\mathcal{X}} d\mu(x) s_i(x), \quad \lim_{i\to\infty} \int_{x\in\mathcal{X}} s_i(x) d\mu(x),$$

*where $\{s_i\}_{i=1}^\infty \subseteq \mathcal{S}(\mathcal{X}, \mathcal{A})$ is a sequence of step functions whose $L^1_\nu(\mathcal{X}, \mathcal{A})$-limit is $u$.*

Note that since $\mathcal{A}$ is not commutative in general, the left and right integrals do not always coincide.

There is also a stronger notion for integrability. An $\mathcal{A}$-valued function $u$ on $\mathcal{X}$ is said to be totally measurable if it is a uniform limit of a step function, i.e., there exists a sequence $\{s_i\}_{i=1}^\infty \subseteq \mathcal{S}(\mathcal{X}, \mathcal{A})$ of step functions such that $\lim_{i\to\infty} \sup_{x\in\mathcal{X}} \|u(x) - s_i(x)\|_{\mathcal{A}} = 0$. We denote by $\mathcal{T}(\mathcal{X}, \mathcal{A})$ the set of all $\mathcal{A}$-valued totally measurable functions on $\mathcal{X}$. Note that if $u \in \mathcal{T}(\mathcal{X}, \mathcal{A})$, then $u \in L^1_{|\mu|}(\mathcal{X}, \mathcal{A})$ for any $\mu \in \mathcal{D}(\mathcal{X}, \mathcal{A})$. In fact, the continuous functions in $C_0(\mathcal{X}, \mathcal{A})$ is totally measurable (see Definition 5.1 for the definition of $C_0(\mathcal{X}, \mathcal{A})$).

**Proposition B.6** *The space $C_0(\mathcal{X}, \mathcal{A})$ is contained in $\mathcal{T}(\mathcal{X}, \mathcal{A})$. Moreover, for any real positive finite regular measure $\nu$, it is dense in $L^1_\nu(\mathcal{X}, \mathcal{A})$ with respect to $\|\cdot\|_{L^1_\nu(\mathcal{X},\mathcal{A})}$.*

For further details, refer to Dinculeanu (1967, 2000).

## Appendix C. Proofs of the propositions and theorem in Section 5.2

Before proving the propositions and theorem, we introduce some definitions and show fundamental properties which are related to the propositions and theorem.

**Definition C.1 ($\mathcal{A}$-dual)** *For a Banach $\mathcal{A}$-module $\mathcal{M}$, the* $\mathcal{A}$-dual of $\mathcal{M}$ *is defined as* $\mathcal{M}' := \{f : \mathcal{M} \to \mathcal{A} \mid \ f \text{ is bounded and } \mathcal{A}\text{-linear}\}$.

Note that for a right Banach $\mathcal{A}$-module $\mathcal{M}$, $\mathcal{M}'$ is a left Banach $\mathcal{A}$-module.

**Definition C.2 (Orthogonal complement)** *For an $\mathcal{A}$-submodule $\mathcal{M}_0$ of a Banach $\mathcal{A}$-module $\mathcal{M}$, the* orthogonal complement of $\mathcal{M}_0$ *is defined as a closed submodule* $\mathcal{M}_0^\perp := \bigcap_{u\in\mathcal{M}_0}\{f \in \mathcal{M}' \mid \ f(u) = 0\}$ *of $\mathcal{M}'$. In addition, for an $\mathcal{A}$-submodule $\mathcal{N}_0$ of $\mathcal{M}'$, the* orthogonal complement of $\mathcal{N}_0$ *is defined as a closed submodule* $\mathcal{N}_0^\perp := \bigcap_{f\in\mathcal{N}_0}\{u \in \mathcal{M} \mid f(u) = 0\}$ *of $\mathcal{M}$.*

Note that for a von Neumann $\mathcal{A}$-module $\mathcal{M}$, by Proposition 4.2, $\mathcal{M}'$ and $\mathcal{M}$ are isomorphic. The following lemma shows a connection between an orthogonal complement and the density property.

**Lemma C.3** *For a Banach $\mathcal{A}$-module $\mathcal{M}$ and its submodule $\mathcal{M}_0$, $\mathcal{M}_0^\perp = \{0\}$ if $\mathcal{M}_0$ is dense in $\mathcal{M}$.*

**Proof** We first show $\overline{\mathcal{M}_0} \subseteq (\mathcal{M}_0^\perp)^\perp$. Let $u \in \mathcal{M}_0$. By the definition of orthogonal complements, $u \in (\mathcal{M}_0^\perp)^\perp$. Since $(\mathcal{M}_0^\perp)^\perp$ is closed, $\overline{\mathcal{M}_0} \subseteq (\mathcal{M}_0^\perp)^\perp$. If $\mathcal{M}_0$ is dense in $\mathcal{M}$, $\mathcal{M} \subseteq (\mathcal{M}_0^\perp)^\perp$ holds, which means $\mathcal{M}_0^\perp = \{0\}$. ■

Moreover, in the case of $\mathcal{A} = \mathbb{C}^{m\times m}$, a generalization of the Riesz–Markov representation theorem for $\mathcal{D}(\mathcal{X}, \mathcal{A})$ holds.

**Proposition C.4 (Riesz–Markov representation theorem for $\mathbb{C}^{m\times m}$-valued measures)** *Let $\mathcal{A} = \mathbb{C}^{m\times m}$. There exists an isomorphism between $\mathcal{D}(\mathcal{X}, \mathcal{A})$ and $C_0(\mathcal{X}, \mathcal{A})'$.*

**Proof** For $f \in C_0(\mathcal{X}, \mathcal{A})'$, let $f_{i,j} \in C_0(\mathcal{X}, \mathbb{C})'$ be defined as $f_{i,j}(u) = (f(u1_{\mathcal{A}}))_{i,j}$ for $u \in C_0(\mathcal{X}, \mathbb{C})$. Then, by the Riesz–Markov representation theorem for complex-valued measure, there exists a unique finite complex-valued regular measure $\mu_{i,j}$ such that $f_{i,j}(u) = \int_{x\in\mathcal{X}} u(x)d\mu_{i,j}(x)$. Let $\mu(E) := [\mu_{i,j}(E)]_{i,j}$ for $E \in \mathcal{B}$. Then, $\mu \in \mathcal{D}(\mathcal{X}, \mathcal{A})$, and we have

$$\begin{aligned} f(u) &= f\left(\sum_{l,l'=1}^m u_{l,l'}e_{l,l'}\right) = \sum_{l,l'=1}^m [f_{i,j}(u_{l,l'})]_{i,j}e_{l,l'} \\ &= \sum_{l,l'=1}^m \left[\int_{x\in\mathcal{X}} u_{l,l'}(x)d\mu_{i,j}(x)\right]_{i,j} e_{l,l'} = \int_{x\in\mathcal{X}} d\mu(x)u(x), \end{aligned}$$

where $e_{i,j}$ is an $m \times m$ matrix whose $(i, j)$-element is 1 and all the other elements are 0. Therefore, if we define $h' : C_0(\mathcal{X}, \mathcal{A})' \to \mathcal{D}(\mathcal{X}, \mathcal{A})$ as $f \mapsto \mu$, $h'$ is the inverse of $h$, which completes the proof of the proposition. ■

### C.1 Proofs of Propositions 5.11 and 5.12

To show Propositions 5.11 and 5.12, the following lemma is used.

**Lemma C.5** *$\Phi : \mathcal{D}(\mathcal{X}, \mathcal{A}) \to \mathcal{M}_k$ is injective if and only if $\langle\Phi(\mu), \Phi(\mu)\rangle_{\mathcal{M}_k} \neq 0$ for any nonzero $\mu \in \mathcal{D}(\mathcal{X}, \mathcal{A})$.*

**Proof** ($\Rightarrow$) Suppose there exists a nonzero $\mu \in \mathcal{D}(\mathcal{X}, \mathcal{A})$ such that $\langle\Phi(\mu), \Phi(\mu)\rangle_{\mathcal{M}_k} = 0$. Then, $\Phi(\mu) = \Phi(0) = 0$ holds, and thus, $\Phi$ is not injective.

($\Leftarrow$) Suppose $\Phi$ is not injective. Then, there exist $\mu, \nu \in \mathcal{D}(\mathcal{X}, \mathcal{A})$ such that $\Phi(\mu) = \Phi(\nu)$ and $\mu \neq \nu$, which implies $\Phi(\mu - \nu) = 0$ and $\mu - \nu \neq 0$. ■

We now show Propositions 5.11 and 5.12.

**Proof of Theorem 5.11** Let $\mu \in \mathcal{D}(\mathcal{X}, \mathcal{A})$, $\mu \neq 0$. We have

$$\begin{aligned} \langle\Phi(\mu), \Phi(\mu)\rangle &= \int_{x\in\mathbb{R}^d}\int_{y\in\mathbb{R}^d} d\mu^*(x)k(x, y)d\mu(y) \\ &= \int_{x\in\mathbb{R}^d}\int_{y\in\mathbb{R}^d} d\mu^*(x)\int_{\omega\in\mathbb{R}^d} e^{-\sqrt{-1}(y-x)^T\omega}d\lambda(\omega)d\mu(y) \\ &= \int_{\omega\in\mathbb{R}^d}\int_{x\in\mathbb{R}^d} e^{\sqrt{-1}x^T\omega}d\mu^*(x)d\lambda(\omega)\int_{y\in\mathbb{R}^d} e^{-\sqrt{-1}y^T\omega}d\mu(y) \\ &= \int_{\omega\in\mathbb{R}^d} \hat{\mu}(\omega)^* d\lambda(\omega)\hat{\mu}(\omega). \end{aligned}$$

Assume $\hat{\mu} = 0$. Then, $\int_{x\in\mathcal{X}} u(x)d\mu(x) = 0$ for any $u \in C_0(\mathcal{X}, \mathcal{A})$ holds, which implies $\mu \in C_0(\mathcal{X}, \mathcal{A})^\perp = \{0\}$ by Proposition C.4 and Lemma C.3. Thus, $\mu = 0$. In addition, by the assumption, $\mathrm{supp}(\lambda) = \mathbb{R}^d$ holds. As a result, $\int_{\omega\in\mathbb{R}^d} \hat{\mu}(\omega)^* d\lambda(\omega)\hat{\mu}(\omega) \neq 0$ holds. By Lemma C.5, $\Phi$ is injective. ■

**Proof of Theorem 5.12** Let $\mu \in \mathcal{D}(\mathcal{X}, \mathcal{A})$, $\mu \neq 0$. We have

$$\begin{aligned}
\langle \Phi(\mu), \Phi(\mu)\rangle &= \int_{x\in\mathbb{R}^d}\int_{y\in\mathbb{R}^d} d\mu^*(x)k(x,y)d\mu(y) \\
&= \int_{x\in\mathbb{R}^d}\int_{y\in\mathbb{R}^d} d\mu^*(x)\int_{t\in[0,\infty)} e^{-t\|x-y\|^2} d\eta(t)d\mu(y) \\
&= \int_{x\in\mathbb{R}^d}\int_{y\in\mathbb{R}^d} d\mu^*(x)\int_{t\in[0,\infty)} \frac{1}{(2t)^{d/2}}\int_{\omega\in\mathbb{R}^d} e^{-\sqrt{-1}(y-x)^T\omega - \frac{\|\omega\|^2}{4t}} d\omega d\eta(t)d\mu(y) \\
&= \int_{\omega\in\mathbb{R}^d} \hat{\mu}(\omega)^* \int_{t\in[0,\infty)} \frac{1}{(2t)^{d/2}} e^{\frac{-\|\omega\|^2}{4t}} d\eta(t)\hat{\mu}(\omega)d\omega, \qquad (32)
\end{aligned}$$

where we applied a formula $e^{-t\|x\|^2} = (2t)^{-d/2}\int_{\omega\in\mathbb{R}^d} e^{-\sqrt{-1}x^T\omega - \|\omega\|^2/(4t)} d\omega$ in the third equality. In the same manner as the proof of Theorem 5.11, $\hat{\mu} \neq 0$ holds. In addition, since $\mathrm{supp}(\eta) \neq \{0\}$ holds, $\int_{t\in[0,\infty)} (2t)^{-d/2} e^{-\|\omega\|^2/(4t)} d\eta(t)$ is positive definite. As a result, the last formula in Eq. (32) is nonzero. By Lemma C.5, $\Phi$ is injective. ■

### C.2 Proofs of Proposition 5.15 and Theorem 5.16

Let $\mathcal{R}_+(\mathcal{X})$ be the set of all real positive-valued regular measures, and $\mathcal{D}_\nu(\mathcal{X}, \mathcal{A})$ the set of all finite regular Borel $\mathcal{A}$-valued measures $\mu$ whose total variations are dominated by $\nu \in \mathcal{R}_+(\mathcal{X})$ (i.e., $|\mu| \leq \nu$). We apply the following representation theorem to derive Theorem 5.16.

**Proposition C.6** *For $\nu \in \mathcal{R}_+(\mathcal{X})$, there exists an isomorphism between $\mathcal{D}_\nu(\mathcal{X}, \mathcal{A})$ and $L^1_\nu(\mathcal{X}, \mathcal{A})'$.*

**Proof** For $\mu \in \mathcal{D}_\nu(\mathcal{X}, \mathcal{A})$ and $u \in L^1_\nu(\mathcal{X}, \mathcal{A})$, we have

$$\left\| \int_{x\in\mathcal{X}} d\mu(x)u(x) \right\|_{\mathcal{A}} \leq \int_{x\in\mathcal{X}} \|u(x)\|_{\mathcal{A}} d|\mu|(x) \leq \int_{x\in\mathcal{X}} \|u(x)\|_{\mathcal{A}} d\nu(x).$$

Thus, we define $h : \mathcal{D}_\nu(\mathcal{X}, \mathcal{A}) \to L^1_\nu(\mathcal{X}, \mathcal{A})'$ as $\mu \mapsto (u \mapsto \int_{x\in\mathcal{X}} d\mu(x)u(x))$.

Meanwhile, for $f \in L^1_\nu(\mathcal{X}, \mathcal{A})'$ and $E \in \mathcal{B}$, we have

$$\|f(\chi_E 1_{\mathcal{A}})\|_{\mathcal{A}} \leq C\int_{x\in\mathcal{X}} \|\chi_E 1_{\mathcal{A}}\|_{\mathcal{A}} d\nu(x) = C\nu(E)$$

for some $C > 0$ since $f$ is bounded. Here, $\chi_E$ is an indicator function for a Borel set $E$. Thus, we define $h' : L^1_\nu(\mathcal{X}, \mathcal{A})' \to \mathcal{D}_\nu(\mathcal{X}, \mathcal{A})$ as $f \mapsto (E \mapsto f(\chi_E 1_{\mathcal{A}}))$.

By the definitions of $h$ and $h'$, $h(h'(f))(s) = f(s)$ holds for $s \in \mathcal{S}(\mathcal{X}, \mathcal{A})$. Since $\mathcal{S}(\mathcal{X}, \mathcal{A})$ is dense in $L^1_\nu(\mathcal{X}, \mathcal{A})$, $h(h'(f))(u) = f(u)$ holds for $u \in L^1_\nu(\mathcal{X}, \mathcal{A})$. Moreover,

$h'(h(\mu))(E) = \mu(E)$ holds for $E \in \mathcal{B}$. Therefore, $\mathcal{D}_\nu(\mathcal{X}, \mathcal{A})$ and $L^1_\nu(\mathcal{X}, \mathcal{A})'$ are isomorphic. ■

**Proof of Theorem 5.16** Assume $\mathcal{M}_k$ is dense in $C_0(\mathcal{X}, \mathcal{A})$. Since $C_0(\mathcal{X}, \mathcal{A})$ is dense in $L^1_\nu(\mathcal{X}, \mathcal{A})$ for any $\nu \in \mathcal{R}_+(\mathcal{X})$, $\mathcal{M}_k$ is dense in $L^1_\nu(\mathcal{X}, \mathcal{A})$ for any $\nu \in \mathcal{R}_+(\mathcal{X})$. By Proposition C.3, $\mathcal{M}_k^\perp = \{0\}$ holds. Let $\mu \in \mathcal{D}(\mathcal{X}, \mathcal{A})$. There exists $\nu \in \mathcal{R}_+(\mathcal{X})$ such that $\mu \in \mathcal{D}_\nu(\mathcal{X}, \mathcal{A})$. By Proposition C.6, if $\int_{x\in\mathcal{X}} d\mu(x)u(x) = 0$ for any $u \in \mathcal{M}_k$, $\mu = 0$. Since $\int_{x\in\mathcal{X}} d\mu(x)u(x) = \langle u, \Phi(\mu)\rangle_{\mathcal{M}_k}$, $\int_{x\in\mathcal{X}} d\mu(x)u(x) = 0$ means $\Phi(\mu) = 0$. Therefore, by Lemma C.5, $\Phi$ is injective. ■

For the case of $\mathcal{A} = \mathbb{C}^{m\times m}$, we apply the following extension theorem to derive the converse of Theorem 5.16.

**Proposition C.7 (c.f. Theorem in Helemskii (1994))** *Let $\mathcal{A} = \mathbb{C}^{m\times m}$. Let $\mathcal{M}$ be a Banach $\mathcal{A}$-module, $\mathcal{M}_0$ be a closed submodule of $\mathcal{M}$, and $f_0 : \mathcal{M}_0 \to \mathcal{A}$ be a bounded $\mathcal{A}$-linear map. Then, there exists a bounded $\mathcal{A}$-linear map $f : \mathcal{M} \to \mathcal{A}$ that extends $f_0$ (i.e., $f(u) = f_0(u)$ for $u \in \mathcal{M}_0$).*

**Proof** Von Neumann-algebra $\mathcal{A}$ itself is regarded as an $\mathcal{A}$-module and is normal. Also, $\mathbb{C}^{m\times m}$ is Connes injective. By Theorem in Helemskii (1994), $\mathcal{A}$ is an injective object in the category of Banach $\mathcal{A}$-module. The statement is derived by the definition of injective objects in category theory. ■

We derive the following lemma and proposition by Proposition C.7.

**Lemma C.8** *Let $\mathcal{A} = \mathbb{C}^{m\times m}$. Let $\mathcal{M}$ be a Banach $\mathcal{A}$-module and $\mathcal{M}_0$ be a closed submodule of $\mathcal{M}$. For $u_1 \in \mathcal{M} \setminus \mathcal{M}_0$, there exists a bounded $\mathcal{A}$-linear map $f : \mathcal{M} \to \mathcal{A}$ such that $f(u_0) = 0$ for $u_0 \in \mathcal{M}_0$ and $f(u_1) \neq 0$.*

**Proof** Let $q : \mathcal{M} \to \mathcal{M}/\mathcal{M}_0$ be the quotient map to $\mathcal{M}/\mathcal{M}_0$, and $\mathcal{U}_1 := \{q(u_1)c \mid c \in \mathcal{A}\}$. Note that $\mathcal{M}/\mathcal{M}_0$ is a Banach $\mathcal{A}$-module and $\mathcal{U}_1$ is its closed submodule. Let $\mathcal{V} := \{c \in \mathcal{A} \mid \ q(u_1)c = 0\}$, which is a closed subspace of $\mathcal{A}$. Since $\mathcal{V}$ is orthogonally complemented (Manuilov and Troitsky, 2000, Proposition 2.5.4), $\mathcal{A}$ is decomposed into $\mathcal{A} = \mathcal{V} + \mathcal{V}^\perp$. Let $p : \mathcal{A} \to \mathcal{V}^\perp$ be the projection onto $\mathcal{V}^\perp$ and $f_0 : \mathcal{U}_1 \to \mathcal{A}$ defined as $q(u_1)c \mapsto p(c)$. Since $p$ is $\mathcal{A}$-linear, $f_0$ is also $\mathcal{A}$-linear. Also, for $c \in \mathcal{A}$, we have

$$\begin{aligned}\|q(u_1)c\|_{\mathcal{M}/\mathcal{M}_0} &= \|q(u_1)(c_1 + c_2)\|_{\mathcal{M}/\mathcal{M}_0} = \|q(u_1)c_1\|_{\mathcal{M}/\mathcal{M}_0} \\ &\geq \inf_{d\in\mathcal{V}^\perp, \|d\|_{\mathcal{A}}=1} \|q(u_1)d\|_{\mathcal{M}/\mathcal{M}_0} \, \|c_1\|_{\mathcal{A}} = \inf_{d\in\mathcal{V}^\perp, \|d\|_{\mathcal{A}}=1} \|q(u_1)d\|_{\mathcal{M}/\mathcal{M}_0} \, \|p(c)\|_{\mathcal{A}},\end{aligned}$$

where $c_1 = p(c)$ and $c_2 = c_1 - p(c)$. Since $\inf_{d\in\mathcal{V}^\perp, \|d\|_{\mathcal{A}}=1} \|q(u_1)d\|_{\mathcal{M}/\mathcal{M}_0} \, \|p(c)\|_{\mathcal{A}} > 0$, $f_0$ is bounded. By Proposition C.7, $f_0$ is extended to a bounded $\mathcal{A}$-linear map $f_1 : \mathcal{M}/\mathcal{M}_0 \to \mathcal{A}$. Setting $f := f_1 \circ q$ completes the proof of the lemma. ■

Then we prove the converse of Lemma C.3.

**Proposition C.9** *Let $\mathcal{A} = \mathbb{C}^{m\times m}$. For a Banach $\mathcal{A}$-module $\mathcal{M}$ and its submodule $\mathcal{M}_0$, $\mathcal{M}_0$ is dense in $\mathcal{M}$ if $\mathcal{M}_0^\perp = \{0\}$.*

**Proof** Assume $u \notin \overline{\mathcal{M}_0}$. We show $\overline{\mathcal{M}_0} \supseteq (\mathcal{M}_0^\perp)^\perp$. By Lemma C.8, there exists $f \in \mathcal{M}'$ such that $f(u) \neq 0$ and $f(u_0) = 0$ for any $u_0 \in \overline{\mathcal{M}_0}$. Thus, $u \notin (\mathcal{M}_0^\perp)^\perp$. As a result, $\overline{\mathcal{M}_0} \supseteq (\mathcal{M}_0^\perp)^\perp$. Therefore, if $\mathcal{M}_0^\perp = \{0\}$, then $\overline{\mathcal{M}_0} \supseteq \mathcal{M}$, which implies $\mathcal{M}_0$ is dense in $\mathcal{M}$. ■

As a result, we derive Proposition 5.15 as follows.

**Proof of Proposition 5.15** Let $\mu \in \mathcal{D}(\mathcal{X}, \mathcal{A})$. Then, "$\Phi(\mu) = 0$" is equivalent to "$\int_{x\in\mathcal{X}} d\mu^*(x)u(x) = \langle \Phi(\mu), u\rangle_{\mathcal{M}_k} = 0$ for any $u \in \mathcal{M}_k$". Thus, by Proposition C.4, "$\Phi(\mu) = 0 \Rightarrow \mu = 0$" is equivalent to "$f \in C_0(\mathcal{X}, \mathcal{A})'$, $f(u) = 0$ for any $u \in \mathcal{M}_k \Rightarrow f = 0$". By the definition of $\mathcal{M}_k^\perp$ and Proposition C.9, $\mathcal{M}_k$ is dense in $C_0(\mathcal{X}, \mathcal{A})$. ■

## Appendix D. Derivative on Banach spaces

**Definition D.1 (Fréchet derivative)** *Let $\mathcal{M}$ be a Banach space. Let $f : \mathcal{M} \to \mathcal{A}$ be an $\mathcal{A}$-valued function defined on $\mathcal{M}$. The function $f$ is referred to as* (Fréchet) differentiable *at a point $\mathbf{c} \in \mathcal{M}$ if there exists a continuous $\mathbb{R}$-linear operator $l$ such that*

$$\lim_{u\to 0,\ u\in\mathcal{M}\setminus\{0\}} \frac{\|f(\mathbf{c}+u) - f(\mathbf{c}) - l(u)\|_{\mathcal{A}}}{\|u\|_{\mathcal{M}}} = 0$$

*for any $u \in \mathcal{M}$. In this case, we denote $l$ as $Df_{\mathbf{c}}$.*

## References


M. Álvarez, L. Rosasco, and N. Lawrence. Kernels for vector-valued functions: A review. *Foundations and Trends in Machine Learning*, 4, 2012.

D. Bakić and B. Guljaš. Operators on Hilbert $H^*$-modules. *Journal of Operator Theory*, 46:123–137, 2001.

E. Balkir. *Using Density Matrices in a Compositional Distributional Model of Meaning*. Master's thesis, University of Oxford, 2014.

P. Blanchard and E. Brüning. *Mathematical Methods in Physics*. Birkhäuser, 2nd edition, 2015.

M. Budišić, R. Mohr, and I. Mezić. Applied Koopmanism. *Chaos*, 22:047510, 2012.

J. Cnops. A Gram–Schmidt method in Hilbert modules. *Clifford Algebras and their Applications in Mathematical Physics*, 47:193–203, 1992.

N. Črnjarić-Žic, S. Maćešić, and I. Mezić. Koopman operator spectrum for random dynamical systems. *Journal of Nonlinear Science*, 30:2007–2056, 2020.

P. Deb. Geometry of quantum state space and quantum correlations. *Quantum Information Processing*, 15:1629–1638, 2016.

J. Diestel. *Sequences and Series in Banach spaces*. Graduate texts in mathematics ; Volume 92. Springer-Verlag, 1984.

N. Dinculeanu. *Vector Measures*. International Series of Monographs on Pure and Applied Mathematics ; Volume 95. Pergamon Press, 1967.

N. Dinculeanu. *Vector Integration and Stochastic Integration in Banach Spaces*. John Wiley & Sons, 2000.

R. M. Dudley. *Real Analysis and Probability*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2nd edition, 2002.

Fujii, K. and Kawahara, Y. Dynamic mode decomposition in vector-valued reproducing kernel Hilbert spaces for extracting dynamical structure among observables. *Neural Networks*, 117:94–103, 2019.

K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf. Kernel measures of conditional dependence. *Advances in Neural Information Processing Systems 20*, 489–496, 2007.

A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. J. Smola. A kernel method for the two-sample-problem. *Advances in Neural Information Processing Systems 19*, 513–520, 2006.

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. J. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012.

Y. Hashimoto, I. Ishikawa, M. Ikeda, Y. Matsuo, and Y. Kawahara. Krylov subspace method for nonlinear dynamical systems with random noise. *Journal of Machine Learning Research*, 21(172):1–29, 2020.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2nd edition, 2009.

A. Helemskii. The spatial flatness and injectiveness of connes operator algebras. *Extracta mathematicae*, 9:75–81, 1994.

J. Heo. Reproducing kernel Hilbert $C^*$-modules and kernels associated with cocycles. *Journal of Mathematical Physics*, 49:103507, 2008.

A. S. Holevo. *Probabilistic and Statistical Aspects of Quantum Theory*. Scuola Normale Superiore, 2011.

I. Ishikawa, K. Fujii, M. Ikeda, Y. Hashimoto, and Y. Kawahara. Metric on nonlinear dynamical systems with Perron–Frobenius operators. In *Advances in Neural Information Processing Systems 31*, 2856–2866, 2018.

S. Itoh. Reproducing kernels in modules over $C^*$-algebras and their applications. *Journal of Mathematics in Nature Science*, 37:1–20, 1990.

W. Jitkrittum, P. Sangkloy, M. W.Gondal, A. Raj, J. Hays, and B. Schölkopf. Kernel mean matching for content addressability of GANs. In *Proceedings of the 36th International Conference on Machine Learning*, 3140–3151, 2019.

H. Kadri, E. Duflos, P. Preux, S. Canu, A. Rakotomamonjy, and J. Audiffren. Operator-valued kernels for learning from functional response data. *Journal of Machine Learning Research*, 17(20):1–54, 2016.

Y. Kawahara. Dynamic mode decomposition with reproducing kernels for Koopman spectral analysis. In *Advances in Neural Information Processing Systems 29*, 911–919, 2016.

S. Klus, I. Schuster, and K. Muandet. Eigendecompositions of transfer operators in reproducing kernel Hilbert spaces. *Journal of Nonlinear Science*, 30:283–315, 2020.

E. C. Lance. *Hilbert $C^*$-modules – a toolkit for operator algebraists*. London Mathematical Society Lecture Note Series; Volume 210. Cambridge University Press, 1995.

D. J. Levitin, R. L. Nuzzo, B. W. Vines, and J. O. Ramsay. Introduction to functional data analysis. *Canadian Psychology*, 48:135–155, 2007.

H. Li, S. J. Pan, S. Wang, and A. C. Kot. Heterogeneous domain adaptation via nonlinear matrix factorization. *IEEE Transactions on Neural Networks and Learning Systems*, 31: 984–996, 2019.

N. Lim, F. Buc, C. Auliac, and G. Michailidis. Operator-valued kernel-based vector autoregressive models for network inference. *Machine Learning*, 99(3):489–513, 2015.

N. Liu and P. Rebentrost. Quantum machine learning for quantum anomaly detection. *Physical Review A*, 97:042315, 2018.

B. Lusch, J. N. Kutz, and S. L. Brunton. Deep learning for universal linear embeddings of nonlinear dynamics. *Nature Communications*, 9:4950, 2018.

V. M. Manuilov and E. V. Troitsky. Hilbert $C^*$- and $W^*$-modules and their morphisms. *Journal of Mathematical Sciences*, 98:137–201, 2000.

C. A. Micchelli and M. Pontil. On learning vector-valued functions. *Neural Computation*, 17:177–204, 2005.

H. Q. Minh, L. Bazzani, and V. Murino. A unifying framework in vector-valued reproducing kernel Hilbert spaces for manifold regularization and co-regularized multi-view learning. *Journal of Machine Learning Research*, 17(25):1–72, 2016.

K. Muandet, K. Fukumizu, B. K. Sriperumbudur, and B. Schölkopf. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends in Machine Learning*, 10(1–2), 2017.

A. Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29:429–443, 1997.

G. J. Murphy. *C*-Algebras and Hilbert Space Operators*. Academic Press, 1990.

A. Peres and D. R. Terno. Quantum information and relativity theory. *Reviews of Modern Physics*, 76:93–123, 2004.

S. T. Rachev. On a class of minimal functionals on a space of probability measures. *Theory of Probability & Its Applications*, 29:41–49, 1985.

J. O. Ramsay and B. W. Silverman. *Functional Data Analysis*. Springer, 2nd edition, 2005.

S. Saitoh and Y. Sawano. *Theory of Reproducing Kernels and Applications*. Springer, 2016.

B. Schölkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2001.

M. Skeide. Generalised matrix $C^*$-algebras and representations of Hilbert modules. *Mathematical Proceedings of the Royal Irish Academy*, 100A(1):11–38, 2000.

A. J. Smola, A. Gretton, L. Song, and B. Schölkopf. A Hilbert space embedding for distributions. In *Proceedings of the 18th International Conference on Algorithmic Learning Theory*, 13–31, 2007.

B. Schölkopf, R. Herbrich, and A. J. Smola. A generalized representer theorem. In *Proceedings of the 14th Annual Conference on Computational Learning Theory*, 416–426, 2001.

G. Smyrlis and V. Zisis. Local convergence of the steepest descent method in Hilbert spaces. *Journal of Mathematical Analysis and Applications*, 300:436–453, 2004.

B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G. R. G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11:1517–1561, 2010.

B. K. Sriperumbudur, K. Fukumizu, and G. R. G. Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, 12 (70):2389–2410, 2011.

B. K. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf, and G. R. G. Lanckriet. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6:1550–1599, 2012.

I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, 2:67–93, 2001.

S. Suzumura, K. Nakagawa, Y. Umezu, K. Tsuda, and I. Takeuchi. Selective inference for sparse high-order interaction models. In *Proceedings of the 34th International Conference on Machine Learning*, 3338–3347, 2017.

F. H. Szafraniec. Murphy's positive definite kernels and Hilbert $C^*$-modules reorganized. *Noncommutative Harmonic Analysis with applications to probability II*, 89:275–295, 2010.

N. Takeishi, Y. Kawahara, and T. Yairi. Subspace dynamic mode decomposition for stochastic Koopman analysis. *Physical Review E*, 96:033310, 2017a.

N. Takeishi, Y. Kawahara, and T. Yairi. Learning Koopman invariant subspaces for dynamic mode decomposition. In *Advances in Neural Information Processing Systems 30*, 1130–1140, 2017b.

J. L. Wang, J. M. Chiou, and H. G. Müller. Functional data analysis. *Annual Review of Statistics and Its Application*, 3:257–295, 2016.

H. Wendland. *Scattered Data Approximation*. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, 2004.

Y. Ye. The matrix Hilbert space and its application to matrix learning. *arXiv:1706.08110v2*, 2017.