Kernels in Functional Data Analysis

Ruoxu Tan

January 22, 2025

This note collects a few basic concepts and results on different but related topics in functional data analysis where "kernels" play a fundamental role. This note can be used as a first glance before learning the related method and theory in detail. There are many different kernels in machine learning. For example, in the kernel regression, a kernel is typically a symmetric density. Besides, in the kernel support vector machines, a kernel is a bivariate nonlinear mapping from the feature space to a higher dimensional space. The kernel discussed in this note is more closely related to the latter, but we focus on the kernel-deduced functional space, which is different from the kernel support vector machines.

1 Reproducing Kernel Hilbert Space

We start from introducing the Reproducing Kernel Hilbert Space (RKHS) following Berlinet and Thomas-Agnan (2011). Let E be an abstract non-empty set. Consider a Hilbert space $\mathcal{H}$ of real-valued functions defined on E endowed with the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$. In the kernel support vector machines, E is the d-dim Euclidean space $\mathbb{R}^d$ where covariates lie in. However, in this note, we mainly concerns E as a compact interval, say $[0, 1]$.

Definition 1. *A bivariate function $K : E \times E \to \mathbb{R}$ is a reproducing kernel if (1) $\forall t \in E$, $K(\cdot, t) \in \mathcal{H}$; and (2) $\forall t \in E$ and $\forall h \in \mathcal{H}$, $\langle h(\cdot), K(\cdot, t) \rangle_{\mathcal{H}} = h(t)$.*

The (2) property above is the so-called "reproducing property": the value of h at t is reproduced by the inner product of h and the kernel induced function $K(\cdot, t)$. Since $K(\cdot, s)$ itself is a function in $\mathcal{H}$ by (1), we deduce from (2) that $\langle K(\cdot, s), K(\cdot, t) \rangle_{\mathcal{H}} = K(t, s) = K(s, t)$. A Hilbert space of real-valued functions which possesses a reproducing kernel is called a RKHS.

Example 1. The Hilbert space $L^2([a,b])$ is *not* a RKHS, because there is no reproducing kernel satisfying

$$\int_{[a,b]} K(s,t)h(s)\,ds = h(t),\ \forall h \in L^2([a,b]);$$

see Example 3 in Chapter 1 of Berlinet and Thomas-Agnan (2011) for details. The example shows that for any kernel on $[a,b] \times [a,b]$, the spanned RKHS is a strict subspace of $L^2([a,b])$.

To give a first characterization of a RKHS, we need the concept of continuous or bounded functional. We say a functional $f : \mathcal{H} \to \mathbb{R}$ is continuous if $\forall \epsilon > 0$, $\exists \delta > 0$ such that if $\|h_1 - h_2\|_{\mathcal{H}} < \delta$ then $|f(h_1) - f(h_2)| < \epsilon$. We say a functional $f : \mathcal{H} \to \mathbb{R}$ is bounded if $\exists M > 0$ such that $|f(h)| \leq M\|h\|_{\mathcal{H}}$, $\forall h \in \mathcal{H}$. A well known property of a *linear* functional is that continuity is equivalent to boundedness. For any $t \in E$, let $e_t : \mathcal{H} \to \mathbb{R}$ denote the evaluation functional at t: $e_t(h) = h(t)$, for any $h \in \mathcal{H}$. It is easy to see that the evaluation functional is linear.

The Riesz representation theorem tells us that for any continuous linear functional f, there exist a function $g_f \in \mathcal{H}$ such that $f(h) = \langle h, g_f\rangle_{\mathcal{H}}$. An application of the Riesz representation theorem leads to the following result.

Theorem 1. *A Hilbert space of real-valued functions on E has a reproducing kernel if and only if all the evaluation functionals $e_t, t \in E$, are continuous.*

Proof outline. ($\Rightarrow$) The Cauchy-Schwarz inequality applied on $e_t(h) = \langle h, K(\cdot,t)\rangle_{\mathcal{H}}$.
($\Leftarrow$) The Riesz representation theorem applied on e_t. □

Next, we discuss a basic characterization of a reproducing kernel.

Definition 2 (Positive type function)**.** *A bivariate function $K : E \times E \to \mathbb{R}$ is a positive type function if*

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j K(t_i, t_j) \geq 0\,, \forall n \geq 1, a_i \in \mathbb{R}, t_i \in E.$$

For any reproducing kernel K, we have

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j K(t_i, t_j) = \|\sum_{i=1}^{n} a_i K(t_i, \cdot)\|_{\mathcal{H}}^2 \geq 0\,,$$

implying that K is positive type. The converse is the famous Moore-Aronszajn theorem.

Theorem 2 (Moore-Aronszajn)**.** *Let K be a positive type function on $E \times E$. There exists a unique RKHS $\mathcal{H}$ of functions on E with K as the reproducing kernel. Specifically, $\mathcal{H}$ is spanned by the functions $\{K(\cdot, t)\}_{t \in E}$ with the inner product*

$$\langle f, g \rangle_{\mathcal{H}} = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \alpha_i \beta_j K(y_j, x_i)\,,$$

where $f = \sum_{i=1}^{\infty} \alpha_i K(\cdot, x_i)$ and $g = \sum_{j=1}^{\infty} \beta_i K(\cdot, y_j)$ are the Cauchy sequences.

An important representation theorem of a continuous symmetric positive type function is the famous Mercer theorem.

Theorem 3 (Mercer)**.** *Let K be a continuous symmetric positive type function on $[a, b] \times [a, b]$. There exists an orthonormal basis $\{\phi_i\}_{i=1}^{\infty}$ of $L^2([a, b])$ such that*

$$K(s, t) = \sum_{j=1}^{\infty} \lambda_j \phi_j(s) \phi_j(t),$$

where $\lambda_1 \geq \lambda_2 \geq \ldots \geq 0$ are called the eigenvalues of K, and ϕ_j are called the eigenfunctions of K.

A continuous symmetric positive type function is referred as a Mercer kernel. For any Mercer kernel $K : [a, b] \times [a, b] \to \mathbb{R}$, we can associate it with a linear operator $K : L^2([a, b]) \to L^2([a, b])$ defined by $K(\phi_j)(t) = \int_a^b K(s, t) \phi_j(s)\, ds$. The first application of the Mercer theorem is the following example.

Example 2 (RKHS generated by a Mercer kernel)**.** Let K be a Mercer kernel on $[a, b] \times [a, b]$, then there exists a unique RKHS $\mathcal{H}$ of functions on $[a, b]$ with K as the reproducing kernel by the Moore-Aronszajn theorem. According to the Mercer theorem, we can express K by

$$K(s, t) = \sum_{j=1}^{\infty} \lambda_j \phi_j(s) \phi_j(t), \ \forall s, t \in [a, b],$$

where $\{\phi_i\}_{i=1}^{\infty}$ is an orthonormal basis of $L^2([a, b])$. Next, we try to find an expression of the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ of $\mathcal{H}$ in terms of $\langle \cdot, \cdot \rangle_{L^2}$, the inner product of $L^2([a, b])$.

For any $f \in \mathcal{H}$, we can write $f(t) = \sum_{j=1}^{\infty} \langle f, \phi_j \rangle_{L^2} \phi_j(t)$. Using the reproducing property and the expression of K above, we have $f(t) = \langle f(\cdot), K(\cdot, t) \rangle_{\mathcal{H}} =$

$\sum_{j=1}^{\infty} \lambda_j \langle f, \phi_j \rangle_{\mathcal{H}} \phi_j(t)$. It follows that for all j, $\langle f, \phi_j \rangle_{L^2} = \lambda_j \langle f, \phi_j \rangle_{\mathcal{H}}$. Now, for any $g \in \mathcal{H}$, we write $g(t) = \sum_{j=1}^{\infty} \langle g, \phi_j \rangle_{L^2} \phi_j(t)$, then we conclude that

$$\begin{aligned}\langle f, g \rangle_{\mathcal{H}} &= \sum_{j=1}^{\infty} \langle g, \phi_j \rangle_{L^2} \langle f, \phi_j \rangle_{\mathcal{H}} \\ &= \sum_{j=1}^{\infty} \frac{\langle f, \phi_j \rangle_{L^2} \langle g, \phi_j \rangle_{L^2}}{\lambda_j}.\end{aligned}$$

Therefore, the RKHS generated by K is $H = \{f \in L^2[a,b]; \sum_{j=1}^{\infty} \frac{\langle f, \phi_j \rangle_{L^2}^2}{\lambda_j} < \infty\}$, which is a strict subspace of $L^2[a,b]$, confirming the argument in Example 1.

Now we are able to introduce the random element in functional data analysis. Given a sample space Ω, let X be a measurable mapping from Ω to $\mathcal{H}$, where $\mathcal{H}$ is a RKHS of functions on a compact interval E. The random element X is the random variable of interest in functional data analysis, also known as the random function/process on E, the functional datum, the functional variable, etc. In fact, the set E is allowed to not be a compact interval, which leads to a general functional variable. When the set E is restrictive to a compact interval, the corresponding X is sometimes referred to as a one-dimensional functional variable. As we only discuss one-dimensional functional data in this note, we omit the term “one-dimensional”.

To study the random function X, just similar to what we learned in elementary statistics, we focus on the first two moments of X. Here, the moments also become functions introducing obscurity indeed. The first moment function of X, the mean function $m = E(X)$, is relatively easy to understand, while the centered second moment function of X, the covariance function $C_X(s,t) = E\{X(s)X(t)\} - E\{X(s)\}E\{X(t)\}$ is more difficult and more important to be investigated. We assume that X has finite second moment, i.e., $\sup_{t \in E} C_X(t,t) < \infty$, under which X is referred to as a second order random process.

Remark 1. *In some literature, it is occasional to define the covariance function as $E\{X(s)X(t)\}$, i.e., the second moment. The covariance operator defined below needs to be modified accordingly. Both definitions are useful under certain applications.*

We first introduce a crucial concept related to the covariance function.

Definition 3 (Covariance operator)**.** *The covariance operator $C_X : \mathcal{H} \to \mathcal{H}$ of X is defined as $C_X(h) = E\{\langle X - m, h \rangle_{\mathcal{H}} (X - m)\}$.*

Definition 4 (Kernel of an operator)**.** *Let $\mathcal{H}$ be a Hilbert space of functions defined on E, and let u be an operator in $\mathcal{H}$. A function $U : E \times E \to \mathbb{R}$ is a kernel of u if (1) $\forall t \in E$, $U(\cdot, t) \in \mathcal{H}$; and (2) $\forall t \in E, \forall h \in \mathcal{H}$, $u(h)(t) = \langle U(\cdot, t), h(\cdot)\rangle_{\mathcal{H}}$.*

Remark 2. *The kernel of an operator here is a slight generalization to the Mecer kernel and its associated operator defined below Theorem 3, in the sense that the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ here can be different from $\langle \cdot, \cdot \rangle_{\mathcal{L}^2}$.*

The covariance operator C_X is self-adjoint, positive, continuous and compact. The covariance function $C_X : E \times E \to \mathbb{R}$ and the covariance operator $C_X : \mathcal{H} \to \mathcal{H}$ use the same notation is a common slight abuse of notation. Because $\forall t \in E$,

$$\begin{aligned} C_X(h)(t) &= E[\langle X - m, h\rangle_{\mathcal{H}}\{X(t) - m(t)\}] \\ &= \langle E[\{X(\cdot) - m(\cdot)\}\{X(t) - m(t)\}], h(\cdot)\rangle_{\mathcal{H}} = \langle C_X(\cdot, t), h(\cdot)\rangle_{\mathcal{H}}\,, \end{aligned}$$

we see that the covariance function is the kernel of the covariance operator. Further, by taking $h(\cdot) = K(\cdot, s)$, $\forall s \in E$, the reproducing kernel, we conclude

$$C_X(s, t) = [C_X\{K(\cdot, s)\}](t)\,.$$

The second application of the Mercer theorem is applying it on the covariance function C_X, which yields the Karhunen–Loève expansion of X, the foundation of the functional principal component analysis.

Theorem 4 (Karhunen–Loève)**.** *Let X be a second order random process on $[a, b]$, then the following representation of X holds,*

$$X(t) = m(t) + \sum_{j=1}^{\infty} \xi_j \phi_j(t),$$

where ϕ_j are the eigenfunctions of C_X (also orthonormal bases of $L^2([a, b])$), ξ_j are mean zero and variance λ_j uncorrelated random variables. The values $\lambda_1 \geq \lambda_2 \geq \ldots \geq 0$ are the eigenvalues of C_X.

The descending order of λ_j is crucial to the functional principal component analysis, because it implies that the truncation to the first d principal components $X_d(t) = m(t) + \sum_{j=1}^{d} \xi_j \phi_j(t)$ capture the most variability of X, i.e., minimal information is lost.

2 Functional Regression

2.1 Function-to-scalar Linear Regression Using RKHS

We follow Yuan and Cai (2010) to introduce the RKHS technique used in function-to-scalar linear regression. Let $\{(X_i, Y_i)\}_{i=1}^n$ be an i.i.d. copy of (X, Y), where $X = X(\cdot)$ is a second order random process on a compact interval $\mathcal{T}$ and Y is a scalar variable satisfying

$$Y = \alpha_0 + \int_{\mathcal{T}} X(t)\beta_0(t)\, dt + \epsilon.$$

Here α_0 is the intercept, β_0 is the slope function and ϵ is the noise variable satisfying $E(\epsilon) = 0$ and $E(\epsilon^2) < \infty$. The key assumption is that the slope function β_0 lies in a RKHS $\mathcal{H}$.

The method of regularization to estimate α_0 and β_0 is given by

$$(\widehat{\alpha}, \widehat{\beta}) = \operatorname*{argmin}_{\alpha\in\mathbb{R},\beta\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n}\left\{Y_i - \alpha - \int_{\mathcal{T}} X_i(t)\beta(t)\, dt\right\}^2 + \lambda J(\beta)\,, \tag{1}$$

where $J(\beta) = \int_{\mathcal{T}}\{\beta^{(m)}(t)\}^2\, dt$ with $\beta^{(m)}$ denoting the m-th order derivative of β. With this choice of the penalty $J(\beta)$, the RKHS $\mathcal{H}$ is in fact the m-order Sobolev space $W_2^m(\mathcal{T})$ defined as

$$W_2^m(\mathcal{T}) = \{\beta : \mathcal{T} \to \mathbb{R},\, \beta, \ldots, \beta^{(m-1)} \text{ are absolutely continuous and } \beta^{(m)} \in L^2(\mathcal{T})\}\,,$$

where the norm is given by

$$\|\beta\|_{W_2^m}^2 = \sum_{q=0}^{m-1}\left(\int \beta^{(q)}\right)^2 + \int (\beta^{(m)})^2\,.$$

The most appealing property of the RKHS estimator given in (1) is that the minimization problem in (1) has a closed-form solution. It is easy to see that

$$\widehat{\alpha} = \bar{Y} - \int_{\mathcal{T}} \bar{X}(t)\widehat{\beta}(t)\, dt\,, \tag{2}$$

where $\bar{Y} = \sum_{i=1}^n Y_i/n$ and $\bar{X} = \sum_{i=1}^n X_i/n$. Next, we derive $\widehat{\beta}$.

Consider the null space $\mathcal{H}_0$ of the penalty functional J,

$$\mathcal{H}_0 = \{\beta \in \mathcal{H} : J(\beta) = 0\}\,,$$

which is a finite-dimensional linear subspace of $\mathcal{H}$ with the orthonormal basis $\{\xi_1, \ldots, \xi_N\}$. Let $\mathcal{H}_1$ be its orthogonal complement such that $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$.

For any $f \in \mathcal{H}$, there exists a unique decomposition $f = f_0 + f_1$, where $f_0 \in \mathcal{H}_0$ and $f_1 \in \mathcal{H}_1$. Note that $\mathcal{H}_1$ is also a RKHS with the inner product of $\mathcal{H}$ restricted to $\mathcal{H}_1$. Letting K be the reproducing kernel of $\mathcal{H}_1$, we have, for any $f_1 \in \mathcal{H}_1$, $J(f_1) = \|f_1\|_K^2 = \|f_1\|_{\mathcal{H}}^2$. We assume that K is continuous and positive-type, i.e., a Mercer kernel. As noted below Theorem 3, K is also a linear operator given by

$$(Kf)(t) = \int_{\mathcal{T}} K(t,s)f(s)\,ds,\ \forall t \in \mathcal{T}\,.$$

It is known that $Kf \in \mathcal{H}_1$, for any $f \in L^2$. Also, note that, for any $f \in \mathcal{H}$

$$\langle Kf, \beta\rangle_{\mathcal{H}} = \int_{\mathcal{T}} \langle K(\cdot, s), \beta\rangle_{\mathcal{H}} f(s)\,ds = \int_{\mathcal{T}} \beta(s)f(s)\,ds.$$

The observations above lead to the following important representer theorem, which is a generalization of the representer lemma for smoothing splines.

Theorem 5. *The estimator $\widehat{\beta}$ has the following finite-dimensional representation,*

$$\widehat{\beta} = \sum_{k=1}^{N} d_k \xi_k(t) + \sum_{i=1}^{n} c_i (KX_i)(t)\,,$$

where $c_1, \ldots, c_n, d_1, \ldots, d_N \in \mathbb{R}$.

With $\widehat{\alpha}$ given at (2), $\widehat{\beta}$ in (1) can be written by

$$\widehat{\beta} = \operatorname{argmin}_{\beta \in \mathcal{H}} \left[\frac{1}{n}\sum_{i=1}^{n} \left\{Y_i - \bar{Y} - \int_{\mathcal{T}} \{X_i(t) - \bar{X}(t)\}\beta(t)\,dt\right\}^2 + \lambda J(\beta)\right]. \quad (3)$$

Consider the case $\mathcal{H} = W_2^2$ and thus $J(\beta) = \int_{\mathcal{T}} \{\beta^{(2)}(t)\}^2\,dt$. It follows that $\mathcal{H}_0$ is spanned by $\{\xi_1(t) = 1, \xi_2(t) = t\}$. A popular reproducing kernel K of $\mathcal{H}_1$ is given by

$$K(s,t) = \frac{1}{4}B_2(s)B_2(t) - \frac{1}{4!}B_4(|s-t|)\,.$$

Using Theorem 5, we have

$$\widehat{\beta} = d_1 + d_2 t + \sum_{i=1}^{n} c_i \int_{\mathcal{T}} \{X_i(s) - \bar{X}(s)\}K(t,s)\,ds.$$

Letting $\mathbf{c} = (c_1, \ldots, c_n)^\top, \mathbf{d} = (d_1, d_2)^\top$, the minimization problem in (3) is equivalent to

$$(\widehat{\mathbf{c}}, \widehat{\mathbf{d}}) = \operatorname{argmin}_{\mathbf{c}\in\mathbb{R}^n, \mathbf{d}\in\mathbb{R}^2} \frac{1}{n}\|\mathbf{Y} - (T\mathbf{d} + \Sigma\mathbf{c})\|_{\ell^2}^2 + \lambda \mathbf{c}^\top \Sigma \mathbf{c}\,,$$

where $\mathbf{Y} = (Y_1 - \bar{Y}, \ldots, Y_n - \bar{Y})^\top$, T is an $n \times 2$ matrix with the (i, j)-th entry given by

$$T_{ij} = \int_{\mathcal{T}} \{X_i(t) - \bar{X}(t)\} t^{j-1}\, dt\,,$$

and Σ is an $n \times n$ matrix with the (i, j)-th entry given by

$$\Sigma_{ij} = \int_{\mathcal{T}} \int_{\mathcal{T}} \{X_i(s) - \bar{X}(s)\} K(s, t) \{X_i(t) - \bar{X}(t)\}\, ds\, dt\,.$$

The weighted least squared problem for $(\widehat{\mathbf{c}}, \widehat{\mathbf{d}})$ above has the following explicit solution,

$$\begin{aligned} \widehat{\mathbf{d}} &= (T^\top W^{-1} T)^{-1} T^\top W^{-1} \mathbf{Y}\,, \\ \widehat{\mathbf{c}} &= W^{-1} \{I_n - T(T^\top W^{-1} T)^{-1} T^\top W^{-1}\} \mathbf{Y}\,, \end{aligned}$$

where $W = \Sigma + n\lambda I_n$ with I_n the n-order identity matrix.

2.2 Function-to-function Nonlinear Regression Using RKHS

In this subsection, we follow Kadri et al. (2016) to introduce the function-to-function nonlinear regression using RKHS. Here the function-to-function operator is assumed to lie in a RKHS of operators, i.e., function-valued functions. Such generalizations require further technical development as well as computational algorithms.

2.2.1 RKHS of Operators

Let $\mathcal{X} = \{x(\cdot) : \mathcal{T}_X \to \mathbb{R}\}$ and $\mathcal{Y} = \{y(\cdot) : \mathcal{T}_Y \to \mathbb{R}\}$ be the Hilbert spaces of real-valued functions where the random processes X and Y are valued in. Let $\mathcal{L}(\mathcal{Y})$ denote the set of bounded linear operators from $\mathcal{Y} \to \mathcal{Y}$. To investigate the RKHS of operators, we first define operator-related concepts.

Definition 5 (adjoint, self-adjoint, and positive operators)**.** *Let $A \in \mathcal{L}(\mathcal{Y})$, then*
(1) A^*, the adjoint operator of A, is the unique operator in $\mathcal{L}(\mathcal{Y})$ that satisfies

$$\langle Ay, z\rangle_{\mathcal{Y}} = \langle y, A^* z\rangle_{\mathcal{Y}}, \forall y, z \in \mathcal{Y};$$

(2) A is self-adjoint if $A = A^*$;
(3) A is positive if it is self-adjoint and $\forall y \in \mathcal{Y}$, $\langle Ay, y\rangle_{\mathcal{Y}} \geq 0$;
(4) A is larger than or equal to $B \in \mathcal{L}(\mathcal{Y})$, if $A - B$ is positive, i.e., $\forall y \in \mathcal{Y}$, $\langle Ay, y\rangle_{\mathcal{Y}} \geq \langle By, y\rangle_{\mathcal{Y}}$.

Definition 6 (Operator-valued kernel)**.** *An $\mathcal{L}(\mathcal{Y})$-valued kernel K on $\mathcal{X} \times \mathcal{X}$ is*

(1) Hermitian if $K(w, z) = K(z, w)^*$, where $K(\cdot, \cdot)^*$ denotes the adjoint operator;

(2) nonnegative on $\mathcal{X}$ if it is Hermitian and $\forall n \geq 1, w_i \in \mathcal{X}, u_i \in \mathcal{Y}$, the $n \times n$ matrix $\langle K(w_i, w_j)u_i, u_j\rangle_{\mathcal{Y}}$ is positive-definite.

Now we are able to define a function-valued RKHS.

Definition 7 (Function-valued RKHS)**.** *A Hilbert space $\mathcal{F}$ of functions from $\mathcal{X}$ to $\mathcal{Y}$ is called a RKHS if there is a nonnegative $\mathcal{L}(\mathcal{Y})$-valued kernel K on $\mathcal{X} \times \mathcal{X}$ such that:*
(1) the function $z \mapsto K(w, z)g$ belongs to $\mathcal{F}$, $\forall z, w \in \mathcal{X}$ and $g \in \mathcal{Y}$;
(2) $\forall F \in \mathcal{F}$, $w \in \mathcal{X}$ and $g \in \mathcal{Y}$, $\langle F, K(w, \cdot)g\rangle_{\mathcal{F}} = \langle F(w), g\rangle_{\mathcal{Y}}$.

If the reproducing kernel K is locally bounded and separately continuous, we call it a Mercer kernel. The following theorem is an extension of the Moore-Aronszajn theorem to the case of function-valued RHKS.

Theorem 6. *A $\mathcal{L}(\mathcal{Y})$-valued Mercer kernel K on $\mathcal{X}^2$ is the reproducing kernel of some Hilbert space F if and only if it is nonnegative.*

Although we have defined the operator-valued kernel, it remains to explicitly design a few operator-valued kernels. To this end, we first present a result that produces new kernels by combing existing ones.

Theorem 7. *Let H and G be two nonnegative operator-valued kernels from $\mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{Y})$, then*
(1) $K = H + G$ is a nonnegative kernel;
(2) if $H(w, z)G(w, z) = G(w, z)H(w, z)$, $\forall w, z \in \mathcal{X}$, then $K = HG$ is a nonnegative kernel;
(3) $K = THT^*$ is a nonnegative kernel for any $T : \mathcal{X} \to \mathcal{L}(\mathcal{Y})$.

As for application to functional data, consider $\mathcal{Y}$ as the Hilbert space $L^2(\mathcal{T})$ of square integrable functions on a compact interval $\mathcal{T}$ endowed with the usual inner product $\langle \phi, \psi\rangle_{\mathcal{Y}} = \int_{\mathcal{T}} \phi(t)\psi(t)\, dt$. Next, we present three examples of operator-valued kernels.

1. (Multiplication operator) For any $h \in \mathcal{Y}$, a multipilication operator T^h on $\mathcal{Y}$ is defined as $T^h : \mathcal{Y} \to \mathcal{Y}$, $T^h(y)(t) = h(t)y(t)$, $\forall t \in \mathcal{T}$. The associated operator-valued kernel K is defined as $K : \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{Y})$,

$K(x_1, x_2)(y)(\cdot) = k_x(x_1, x_2)T^{k_y}(y)(\cdot)$, where $k_x : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a positive definite scalar-valued kernel and k_y is a positive real function. It is easy to see that K is Hermitian and positive.

2. (Hilbert-Schmidt integral operator) A Hilbert-Schmidt integral operator T^h associated with a kernel $h : \mathcal{T} \times \mathcal{T} \to \mathbb{R}$ is defined as: $T^h : \mathcal{Y} \to \mathcal{Y}$, $T^h(y)(t) = \int h(s,t)y(s)\,ds$, $\forall t \in \mathcal{T}$. An operator-valued kernel K associated with positive kernels $k_x : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $k_y : \mathcal{T} \times \mathcal{T} \to \mathbb{R}$ is defined as $K : \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{Y})$, $K(x_1, x_2)(y)(\cdot) = k_x(x_1, x_2) \int k_y(s, \cdot)f(s)\,ds$. If k_y is Hermitian, then K is also Hermitian.

3. (Composition operator) Let $\phi : \mathcal{T} \to \mathcal{T}$ be an analytic map, the associated composition operator $C_\phi : \mathcal{Y} \to \mathcal{Y}$ is defined as $C_\phi(y) = y \circ \phi$. In the case that $\mathcal{Y}$ is a RKHS with the kernel k, we have

$$\begin{aligned} \langle f, C_\phi^* k(t,\cdot)\rangle_{\mathcal{Y}} &= \langle C_\phi(f), k(t,\cdot)\rangle_{\mathcal{Y}} = \langle f \circ \phi, k(t,\cdot)\rangle_{\mathcal{Y}} = f(\phi)(t) \\ &= \langle f, k(\phi(t),\cdot)\rangle_{\mathcal{Y}}\,, \end{aligned}$$

implying that $C_\phi^* k(t,\cdot) = k(\phi(t),\cdot)$. Similarly, we have $C_\phi^*(f)(t) = \langle f, k(t, \phi(\cdot))\rangle_{\mathcal{Y}}$. Once a composition operator and its adjoint operator are well expressed in a RKHS, we can define a operator-valued kernel $K : \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{Y})$ as $K(x_1, x_2) = C_{\phi(x_1)} C_{\phi(x_2)}^*$, where $\phi(x_1)$ and $\phi(x_2)$ are analytic maps from $\mathcal{T}$ to $\mathcal{T}$. Using Theorem 7 (3), we see that K is nonnegative.

2.2.2 Function-valued Functional Learning

Consider the function-to-function regression of estimating $F(x) = E(Y|X = x)$ from observed data $(x_i, y_i)_{i=1}^n$, where $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y} = L^2(\Omega_x) \times L^2(\Omega_y)$, the Hilbert spaces of square integrable functions on Ω_x and Ω_y.

The estimator of F via the method of regularization is defined as

$$\widetilde{F}_\lambda = \operatorname{argmin}_{F \in \mathcal{F}} \|y_i - F(x_i)\|_{\mathcal{Y}}^2 + \lambda \|F\|_{\mathcal{F}}^2\,, \tag{4}$$

where $\lambda > 0$ is a regularization parameter.

In the case that $\mathcal{F}$ is a real-valued RKHS, the solution of the minimization problem above has the following form,

$$\widetilde{F}(x) = \sum_{i=1}^n \alpha_i K(x_i, x)\,,$$

where $\alpha_i \in \mathbb{R}$ and K is the reproducing kernel. An extension to the case of function-valued RKHS leads to the following form,

$$\widetilde{F}(\cdot) = \sum_{i=1}^{n} K(x_i, \cdot) u_i \,,$$

where $u_i(\cdot) \in \mathcal{Y}$ are functions and K is the nonnegative operator-valued reproducing kernel. Plugging the expression above into (4), we obtain the following minimization problem over n scalar-valued functions $u_i \in \mathcal{Y}$,

$$\begin{aligned}\widetilde{\mathbf{u}}_\lambda &= (\widetilde{u}_{\lambda 1}, \ldots, \widetilde{u}_{\lambda n})^\top \\ &= \operatorname*{argmin}_{\mathbf{u} \in \mathcal{Y}^n} \sum_{i=1}^{n} \|y_i - \sum_{j=1}^{n} K(x_i, x_j) u_j\|_{\mathcal{Y}}^2 + \lambda \sum_{i=1}^{n} \sum_{j=1}^{n} \langle K(x_i, x_j) u_i, u_j \rangle_{\mathcal{Y}}.\end{aligned}$$

Setting the directional derivative of $\mathbf{u}$ above to zero yields that the vector of functions $\mathbf{u} \in \mathcal{Y}^n$ satisfies the following system of linear operator equations,

$$(\mathbf{K} + \lambda I)\mathbf{u} = \mathbf{y} \,,$$

where $\mathbf{K} = [K(x_i, x_j)]_{i,j=1,\ldots,n}$ is a $n \times n$ block operator kernel matrix and $\mathbf{y} = (y_1, \ldots, y_n)^\top$.

To overcome the problem of finding the inverse of the block operator kernel matrix $\mathbf{K}$, we assume that the operator-valued kernel K has the following form,

$$K(x_i, x_j) = g(x_i, x_j) T \,,$$

where $g : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a scalar-valued kernel and T is a linear operator in $\mathcal{L}(\mathcal{Y})$. For example, if T is a integral operator with the kernel $e^{-|t-s|}$, then

$$K(x_i, x_j)(y)(t) = g(x_i, x_j) \int_{\Omega_y} e^{-|t-s|} y(s)\, ds \,,$$

for $y \in \mathcal{Y}$. It follows that the block operator kernel matrix $\mathbf{K}$ can be expressed as

$$\mathbf{K} = \begin{pmatrix} g(x_1, x_1)T & \cdots & g(x_1, x_n)T \\ & \cdots & \\ g(x_n, x_1)T & \cdots & g(x_n, x_n)T \end{pmatrix} = G \otimes T \,,$$

where $G = [g(x_i, x_j)]_{i,j=1,\ldots,n}$. Using the basic property of the Kronecker product, we obtain that $\mathbf{K}^{-1} = G^{-1} \otimes T^{-1}$, and thus we only need to find G^{-1} and T^{-1}.

The inversion of $n \times n$ matrix G can be expressed by $G^{-1} = V \Gamma V^\top$, where $V = (v_1, \ldots, v_n)$ consists of eigenvectors v_i of G, and $\Gamma = \operatorname{diag}(\alpha_1^{-1}, \ldots, \alpha_n^{-1})$

with α_i the eigenvalue of G. In terms of inversion of a linear operator T, we utilize the result that if T is compact and normal, then

$$T(y) = \sum_{i=1}^{\infty} \delta_i \langle y, \omega_i \rangle_{\mathcal{Y}} \omega_i \,,$$

where δ_i and ω_i are referred to as the eigenvalues and the eigenfunctions of T. It follows that $T^{-1}(y) = \sum_{i=1}^{\infty} \delta_i^{-1} \langle y, \omega_i \rangle_{\mathcal{Y}} \omega_i$. For fixed G, $\mathbf{K} + \lambda I$ can also be seen as an operator in $\mathcal{L}(\mathcal{Y})$. Therefore, to compute $\mathbf{u} = (\mathbf{K} + \lambda I)^{-1}\mathbf{y}$, we fix a truncation parameter κ, and obtain

$$\widehat{\mathbf{u}} = \sum_{i=1}^{n\kappa} (\theta_i + \lambda)^{-1} \langle \mathbf{z}_i, \mathbf{y} \rangle_{\mathcal{Y}^n} \mathbf{z}_i \,,$$

where $\theta = (\theta_1, \ldots, \theta_{n\kappa})^\top = (\alpha_1, \ldots, \alpha_n)^\top \otimes (\delta_1, \ldots, \delta_\kappa)^\top$, $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_{n\kappa})^\top = (v_1, \ldots, v_n)^\top \otimes (\omega_1, \ldots, \omega_\kappa)^\top$ and $\langle \mathbf{a}, \mathbf{b} \rangle_{\mathcal{Y}^n} = \sum_{i=1}^{n} \langle a_i, b_i \rangle_{\mathcal{Y}}$.

3 Gaussian Measure in Hilbert Space

An obstacle to study random variables in an infinite dimensional space is that the Lebesgue measure generally does not exist. Yet, a Gaussian measure can be well defined in a separable Banach space, which can serve as an alternative to the Lebesgue measure. Therefore, a Gaussian measure, induced by a Gaussian process, plays a fundamental role in studying infinite dimensional random elements. In this section, we follow Kuo (1975) and Williams and Rasmussen (2006) to introduce the concepts of Gaussian processes particularly in a Hilbert space.

Let $\mathcal{H}$ be a separable Hilbert space with norm $|\cdot| = \sqrt{\langle \cdot, \cdot \rangle_{\mathcal{H}}}$. Let A be a linear operator in $\mathcal{H}$.

Definition 8 (Hilbert-Schmidt operator)**.** *A linear operator A in $\mathcal{H}$ is a Hilbert-Schmidt operator if for some orthonormal basis $\{e_n\}_{n=1}^{\infty}$ of $\mathcal{H}$,*

$$\sum_{n=1}^{\infty} |A(e_n)|^2 < \infty.$$

The Hilbert-Schmidt norm of A is defined as $\|A\|_2 = \left(\sum_{n=1}^{\infty} |A(e_n)|^2 \right)^{1/2}$.

Note that the Hilbert-Schmidt norm does not depend on the choice of $\{e_n\}_{n=1}^{\infty}$. An example of Hilbert-Schmidt operator is the integral operator defined in Section 2.2.1.

A operator in $\mathcal{H}$ is compact if it maps any bounded subset of $\mathcal{H}$ into a set whose closure is compact. The following expression for a self-adjoint compact operator has actually been used in Section 2.2.2.

Theorem 8. *If A is a self-adjoint compact operator in $\mathcal{H}$, then there exists an orthonormal basis $\{e_n\}_{n=1}^{\infty}$ of $\mathcal{H}$ such that*

$$A(x) = \sum_{n=1}^{\infty} \lambda_n \langle x, e_n \rangle_{\mathcal{H}} e_n \,,$$

where the λ_n's and e_n's are called the eigenvalues and eigenfunctions of A. If A is positive definite, then $\lambda_n \geq 0$.

Definition 9 (Trace class operator)**.** *A compact operator A in $\mathcal{H}$ is called a trace class operator if $\sum_{n=1}^{\infty} \lambda_n < \infty$, where the λ_n's are the eigenvalues of $(A^*A)^{1/2}$.*

Proposition 1. *A Hilbert-Schmidt operator is compact. An operator A is Hilbert-Schmidt if and only if $\sum_{n=1}^{\infty} \lambda_n^2 < \infty$, where the λ_n's are the eigenvalues of $(A^*A)^{1/2}$. In this case, $\|A\|_2 = (\sum_{n=1}^{\infty} \lambda_n^2)^{1/2}$.*

If A is a trace class operator, the trace of A is defined as $\sum_{n=1}^{\infty} \langle A(e_n), e_n \rangle_{\mathcal{H}}$, where $\{e_n\}_{n=1}^{\infty}$ is any orthonormal basis of $\mathcal{H}$. Any trace class operator can be written as a product of two Hilbert-Schmidt operators.

For a random function X valued in $\mathcal{H}$, recall from Definition 3 that the covariance operator $C_X : \mathcal{H} \to \mathcal{H}$ of X is defined as $C_X(h) = E\{\langle X - m, h \rangle_{\mathcal{H}} (X - m)\}$. Now we define a covariance operator of a measure in $\mathcal{H}$. All measures considered in this section are Borel measures, i.e., measures defined on the σ-filed generated by the open subsets of $\mathcal{H}$.

Definition 10 (Covariance operator)**.** *For a measure μ in $\mathcal{H}$, the covariance operator S_μ of μ is defined such that*

$$\langle S_\mu(x), y \rangle_{\mathcal{H}} = \int_{\mathcal{H}} \langle x, z \rangle_{\mathcal{H}} \langle y, z \rangle_{\mathcal{H}} \, \mu(dz) \,.$$

Remark 3. *Compared to the covariance operator of a random variable Z in Definition 3, $\langle C_Z(x), y \rangle_{\mathcal{H}} = \int_{\mathcal{H}} \langle z - m, x \rangle_{\mathcal{H}} \langle z - m, y \rangle_{\mathcal{H}} \, \mu_Z(dz)$, we see that Definition 10 is actually the* uncentered *second moment. Such difference does not affect the main idea.*

A covariance operator is necessarily positive definite and self-adjoint. If a covariance operator has finite trace, we call it a S-operator. In fact, we have trace $S_\mu = \int_{\mathcal{H}} |x|^2 \, \mu(dx)$.

Definition 11 (Mean)**.** *For a measure μ in $\mathcal{H}$, the mean of μ is an element m_μ in $\mathcal{H}$ such that*

$$\langle m_\mu, x \rangle_{\mathcal{H}} = \int_{\mathcal{H}} \langle z, x \rangle_{\mathcal{H}} \, \mu(dz) \,.$$

Definition 12 (Characteristic functional)**.** *The characteristic functional ϕ of a measure μ in $\mathcal{H}$ is defined as*

$$\phi(x) = \int_{\mathcal{H}} \exp\{i\langle x, y\rangle_{\mathcal{H}}\}\mu(dy)\,,$$

for $x \in \mathcal{H}$.

Finally, we are able to define Gaussian measure in $\mathcal{H}$.

Definition 13 (Gaussian measure)**.** *A measure μ in $\mathcal{H}$ is Gaussian if for each $x \in \mathcal{H}$, the measurable functional $\langle x, \cdot\rangle_{\mathcal{H}}$ is normally distributed, i.e., there exists real numbers m_x and σ_x^2 such that for all $a \in \mathbb{R}$,*

$$P(y \in \mathcal{H}; \langle x, y\rangle_{\mathcal{H}} \le a) = \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi\sigma_x^2}} \exp\Big\{-\frac{(t - m_x)^2}{2\sigma_x^2}\Big\}\,dt\,.$$

Recall that the characteristic function of a $\mathbb{R}$-valued normal variable $X \sim N(\mu, \sigma^2)$ is $\psi(t) = E\{\exp(itX)\} = \exp(i\mu t - \sigma^2 t^2/2)$. Using a change of variable, we have, for a Gaussian measure μ in $\mathcal{H}$, its characteristic functional ϕ is given by

$$\phi(x) = \int_{\mathcal{H}} \exp\{i\langle x, y\rangle_{\mathcal{H}}\}\mu(dy) = \int_{\mathcal{H}} \exp(is)\mu_x(ds) = \psi_x(1)\,,$$

where $\psi_x(t) = \exp(i\mu_x t - \sigma_x^2 t^2/2)$ is the characteristic function of the normal variable $X = \langle x, \cdot\rangle_{\mathcal{H}} \sim N(\mu_x, \sigma_x^2)$. In particular, we can show by change of variables that

$$\mu_x = \langle m_\mu, x\rangle_{\mathcal{H}}\,, \quad \sigma_x^2 = \langle S_\mu(x), x\rangle_{\mathcal{H}}\,,$$

and thus the characteristic functional ϕ of the Gaussian measure μ is given by

$$\phi(x) = \exp\Big\{i\langle m_\mu, x\rangle_{\mathcal{H}} - \frac{\langle S_\mu(x), x\rangle_{\mathcal{H}}}{2}\Big\}\,.$$

We also conclude from the expression above that a Gaussian measure is uniquely determined by its mean and covariance operator.

To simplify exposition, we assume that $\mathcal{H}$ is a Hilbert space of functions on E and $m \equiv 0$ from now on. A random element X valued in $\mathcal{H}$ is called a Gaussian process if the push-forward measure $\mu\big(X(\cdot)\big)$ is a Gaussian measure in $\mathcal{H}$. Recall from Definition 4 below that the covariance function $S_X : E \times E \to \mathbb{R}$ of X, $S_X(s,t) = E\{X(s)X(t)\}$ is associated with the covariance operator through

$$S_X(h)(t) = \langle S_X(\cdot, t), h\rangle_{\mathcal{H}}\,,$$

for $t \in E$. The bivariate function S_X is also referred as the kernel, or covariance kernel of X.

Next, we present several examples of the covariance function. A stationary covariance function is only a function of $s-t$, and an isotropic covariance function is only a function of $r = |s-t|$. The squared exponential covariance function has the following form

$$S_{\mathrm{SE}}(r) = \exp\left(\frac{-r^2}{2\ell^2}\right),$$

where ℓ is called the characteristic length-scale. Since the squared exponential covariance function is infinite differentiable, the induced Gaussian process is very smooth, which may be unrealistic.

A less smooth class of covariance function is the Matérn class,

$$S_{\mathrm{Matern}}(r) = \frac{2^{1-\nu}}{\Gamma(v)}\Big(\frac{\sqrt{2\nu}r}{\ell}\Big)^2 K_\nu\Big(\frac{\sqrt{2\nu}r}{\ell}\Big),$$

where $\nu > 0, \ell > 0$ are parameters and K_ν is a modified Bessel function. The induced Gaussian process with the Matérn class covariance function is $\lfloor \nu \rfloor$-times mean squared differentiable. When ν is half-integer, the Matérn class covariance function has a more explicit form.

The rational quadratic covariance function has the following form,

$$S_{\mathrm{RQ}}(r) = \left(1 + \frac{r^2}{2\alpha\ell^2}\right)^{-\alpha},$$

where α and $\ell > 0$ are parameters. It can be seen as a scale mixture (an infinite sum) of squared exponential covariance functions with different characteristic length-scales. The covariance functions mentioned above are all stationary and non-degenerate, while other non-stationary covariance functions include polynomial and neural network classes; see Williams and Rasmussen (2006, Ch. 4.2) for details.

References

Berlinet, A. and Thomas-Agnan, C. (2011). *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media.

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

Kuo, H.-H. (1975). *Gaussian measures in Banach spaces*. Springer.

Williams, C. K. and Rasmussen, C. E. (2006). *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA.

Yuan, M. and Cai, T. T. (2010). A reproducing kernel hilbert space approach to functional linear regression. *The Annals of Statistics*, 38(1):3412–3444.