Notes on RKHS

Rui Wang

Friday 5th October, 2018

1 Introduction

This is a (hopefully) self-contained note on RKHS and its application. Good references include Wahba (1990), Gu (2013), Hofmann et al. (2008), Andreas Christmann (2008), Andreas Christmann (2008), Muller et al. (2001), van der Vaart and van Zanten (2008), Amini (2013) and Hsing and Eubank (2015). Compared with the materials in these documents, there's nothing new in this note except for some possible errors.

2 Positive semi-definite kernel

This section is adapted from Hofmann et al. (2008).

We consider a general index set $\mathcal{T}$. A *kernel* is a function $R(\cdot,\cdot)$ from $\mathcal{T}\times\mathcal{T}$ to $\mathbb{R}$ which is symmetric: $R(s,t)=R(t,s)$. A kernel $R(\cdot,\cdot)$ is said to be *positive semi-definite* (PSD) if, for any real $a_1,\ldots,a_n$ and $t_1,\ldots,t_n\in\mathcal{T}$,

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

In other words, for any $t_1,\ldots,t_n\in\mathcal{T}$ the matrix

$$\begin{pmatrix} R(t_1,t_1) & \ldots & R(t_1,t_n) \\ \ldots & \ldots & \ldots \\ R(t_n,t_1) & \ldots & R(t_n,t_n) \end{pmatrix}$$

is PSD.

Let $(\mathbb{H},(\cdot,\cdot))$ be an inner product space. It is clear that for any function $\phi:\mathcal{T}\to\mathbb{H}$, $R(s,t):=(\phi(s),\phi(t))$ is a PSD kernel. If $R(\cdot,\cdot)$ is so defined, $\phi(\cdot)$ is called its *feature map* and $\mathbb{H}$ is called the *feature space*. We note that the feature map of a kernel is not unique. We shall return to this point later.

The following proposition gives other ways to construct PSD kernels.

Proposition 1. *Suppose $R_i(\cdot,\cdot)$, $i=1,2,\ldots$, are PSD kernels on $\mathcal{T}\times\mathcal{T}$. Then*

(i) *If $\alpha_1, \alpha_2 \geq 0$, then $\alpha_1 R_1(\cdot,\cdot) + \alpha_2 R_2(\cdot,\cdot)$ is a PSD kernel.*

(ii) *If $R(s,t) := \lim_{n\to\infty} R_n(s,t)$ exists for all $s,t \in \mathcal{T}$, then $R(\cdot,\cdot)$ is a PSD kernel.*

(iii) *The pointwise product $R_1 \circ R_2(s,t) := R_1(s,t)R_2(s,t)$ is a PSD kernel.*

(iv) *Assume that for $i = 1, 2$, $R_i(\cdot,\cdot)$ is a PSD kernel on $\mathcal{T}_1 \times \mathcal{T}_2$. Then the tensor product, which is a function defined on $(\mathcal{T}_1 \times \mathcal{T}_2) \times (\mathcal{T}_1 \times \mathcal{T}_2)$ as*

$$R_1 \otimes R_2\left((s_1,s_2),(t_1,t_2)\right) := R_1(s_1,t_1)R_2(s_2,t_2)$$

is a PSD kernel. The direct sum

$$R_1 \oplus R_2\left((s_1,s_2),(t_1,t_2)\right) := R_1(s_1,t_1) + R_2(s_2,t_2)$$

is a PSD kernel.

Proof. These results follows directly from the analogous results in matrix theory. For example, (iii) follows from that the Hadamard product of two PSD matrix is PSD. As for (iv), for any real $a_1, \ldots, a_n$ and $(t_{11}, t_{12}), \ldots, (t_{n1}, t_{n2}) \in \mathcal{T}_1 \times \mathcal{T}_2$,

$$\begin{aligned}\sum_{i,j=1}^n a_i a_j R_1 \oplus R_2((t_{i1},t_{i2}),(t_{j1},t_{j2})) &= \sum_{i,j=1}^n a_i a_j \left(R_1(t_{i1},t_{j1}) + R_2(t_{i2},t_{j2})\right) \\ &= \sum_{i,j=1}^n a_i a_j R_1(t_{i1},t_{j1}) + \sum_{i,j=1}^n a_i a_j R_2(t_{i2},t_{j2}) \geq 0.\end{aligned}$$

On the other hand

$$\sum_{i,j=1}^n a_i a_j R_1 \otimes R_2((t_{i1},t_{i2}),(t_{j1},t_{j2})) = \sum_{i,j=1}^n a_i a_j R_1(t_{i1},t_{j1})R_2(t_{i2},t_{j2}) \geq 0$$

since it corresponds to the Hadamard product of

$$\begin{pmatrix} R_1(t_{11},t_{11}) & \ldots & R_1(t_{11},t_{n1}) \\ \ldots & \ldots & \ldots \\ R_1(t_{n1},t_{11}) & \ldots & R_1(t_{n1},t_{n1}) \end{pmatrix}$$

and

$$\begin{pmatrix} R_2(t_{12},t_{12}) & \ldots & R_2(t_{12},t_{n2}) \\ \ldots & \ldots & \ldots \\ R_2(t_{n2},t_{12}) & \ldots & R_1(s_{n2},t_{n2}) \end{pmatrix}.$$

□

Example 1 (Gaussian kernel)**.** *Let $\mathcal{T} = \mathbb{R}^p$. Then $R(s,t) = s^\top t$ is a PSD kernel. By (i) and (iii) of Proposition 1,*

$$R_n(s,t) := \sum_{i=0}^n \frac{R(s,t)^n}{n!}$$

is a PSD kernel. Then by (ii) of Proposition 1, $\exp(R(s,t)) := \exp(s^\top t)$ *is a PSD kernel. By (v) of Proposition 1,* $\exp(-\|s\|^2 - \|t\|^2)$ *is a PSD kernel. Thus, again by (iii) of Proposition 1,* $\exp(-\|t-s\|^2)$ *is a PSD kernel.*

A PSD kernel $R(\cdot,\cdot)$ *is call* radial *if* $R(x,y) = g(\|x-y\|)$ *for some function* $g : [0,+\infty) \to \mathbb{R}$.

Example 2 (Polynomial kernels)**.** *Let* $\mathcal{T} = \mathbb{R}^p$*. From (iii) of Proposition 1 it is clear that homogeneous polynomial kernels* $R(s,t) = (s^\top t)^n$ *are PSD for* $n \in \mathbb{N}$*. We can also explicitly give the corresponding feature map:*

$$R(s,t) = (s^\top t)^n = (\sum_{i=1}^p s_i t_i)^n = \sum_{i_1=1}^p \sum_{i_2=1}^p \cdots \sum_{i_n=1}^p (s_{i_1} \cdots s_{i_n}) \cdot (t_{i_1} \cdots t_{i_n}) = (C_{n,p}(s), C_{n,p}(t)),$$

where $C_{n,p}(\cdot)$ *maps* $t \in \mathbb{R}^p$ *to a* p^n *dimensional vector whose entries are all possible nth degree ordered products of the entries of* t*. Other useful kernels include the inhomogeneous polynomial* $R(s,t) = (s^\top t + c)^n$ *where* $n \in \mathbb{N}$ *and* $c \geq 0$*.*

3 Reproducing kernel Hilbert space

This section is adapted from Wahba (1990), Andreas Christmann (2008), Hsing and Eubank (2015). A *Hilbert space* $(\mathbb{H}, (\cdot,\cdot))$ is a complete vector space with an inner product. An important example of Hilbert space is the class of all square integrable measurable functions $L^2(\mathbb{X}, \mathcal{B}, \mu)$ on a measurable space $(\mathbb{X}, \mathcal{B}, \mu)$. A continuous linear functional (or bounded linear functional) L is a linear map from $\mathbb{H}$ into $\mathbb{R}$ such that

$$|Lf| \leq M\|f\| \text{ for all } f \in \mathcal{H}.$$

For each $y \in \mathbb{H}$, the map $x \mapsto (x,y)$ is a continuous linear functional, denoted as $(\cdot, y)$. A fundamental result in real analysis says all continuous linear functional can be represented by $(\cdot, y)$ for some $y \in \mathbb{H}$.

Theorem 1 (Riesz-Fréchet)**.** *A map* L *from a Hilbert space* $\mathbb{H}$ *into* $\mathbb{R}$ *is a continuous linear functional if and only if for some* $y \in \mathbb{H}$*,* $Lx = (x,y)$ *for all* $x \in \mathbb{H}$*. If so, then* y *is unique.*

Let $\mathbb{R}^\mathcal{T}$ denote the space of all real functions from $\mathcal{T}$ to $\mathbb{R}$. Suppose $\mathbb{H}$ is a subset of $\mathbb{R}^\mathcal{T}$ and $(\mathbb{H}, (\cdot,\cdot)_\mathbb{H})$ is a Hilbert space. Then it can be seen that the for each $t \in \mathcal{T}$, the coordinate projection L_t, defined as $L_t f(\cdot) := f(t)$, is a linear functional. Note that L_t is not necessarily continuous. In fact, if $\mathbb{H} = L^2(\mathbb{R}, \mathcal{B}, \mu)$, then L_t is not continuous since $f(t)$ can be arbitrarily defined.

Definition 1. Suppose $\mathbb{H}$ is a subset of $\mathbb{R}^\mathcal{T}$ and $(\mathbb{H}, (\cdot,\cdot)_\mathbb{H})$ is a Hilbert space. Then $\mathbb{H}$ is called a reproducing kernel Hilber space (RKHS) if for each $t \in \mathcal{T}$, the coordinate projection L_t is a continuous linear functional.

Every RKHS has a unique reproducing kernel, as stated by the following theorem.

Theorem 2. *Let $\mathbb{H} \subset \mathbb{R}^{\mathcal{T}}$ be an RKHS. Then there is a unique PSD kernel $R(\cdot,\cdot) : \mathcal{T}\times\mathcal{T} \to \mathbb{R}$, such that*

1. *for every $t \in \mathcal{T}$, $R(\cdot, t) \in \mathbb{H}$ and*
2. *R satisfies the reproducing property that for every $f(\cdot) \in \mathbb{H}$ and $t \in \mathcal{T}$*

$$f(t) = (f(\cdot), R(\cdot, t))_{\mathbb{H}}.$$

Remark 1. A PSD kernel satisfying $R(\cdot,\cdot)$ satisfying 1 and 2 is called the *reproducing kernel* of $\mathbb{H}$. Hence the last theorem says that every RKHS has a unique reproducing kernel.

Proof. For each $t \in \mathcal{T}$ there exists, by Theorem 1, a unique element $R_t(\cdot) \in \mathbb{H}$ with the property

$$L_t f(\cdot) = (f(\cdot), R_t(\cdot))_{\mathbb{H}} = f(t), \quad \forall f \in \mathbb{H}.$$

Let $f(\cdot) = R_s(\cdot)$ in the above equality, we have

$$(R_s(\cdot), R_t(\cdot))_{\mathbb{H}} = R_s(t), \quad \forall s, t \in \mathcal{T}.$$

Since the inner product is symmetric, $(R_s(\cdot), R_t(\cdot))_{\mathbb{H}} = (R_t(\cdot), R_s(\cdot))_{\mathbb{H}}$. Hence

$$(R_s(\cdot), R_t(\cdot))_{\mathbb{H}} = R_t(s), \quad \forall s, t \in \mathcal{T}.$$

Thus, $R_s(t) = R_t(s)$. For any real $a_1, \ldots, a_n$ and $t_1, \ldots, t_n \in \mathcal{T}$,

$$\sum_{i,j=1}^{n} a_i a_j R_{t_i}(t_j) = \sum_{i,j=1}^{n} a_i a_j \left(R_{t_i}(\cdot), R_{t_j}(\cdot)\right)_{\mathbb{H}} = \left(\sum_{i}^{n} a_i R_{t_i}(\cdot), \sum_{i}^{n} a_j R_{t_j}(\cdot)\right)_{\mathbb{H}} \geq 0.$$

It follows that $R_{\cdot}(\cdot)$ is a positive semi-definite kernel. Since $R_{\cdot}(\cdot)$ is symmetric, it can be written as $R(\cdot,\cdot)$. □

Conversely, given a positive-definite kernel $R(\cdot,\cdot)$, there is a unique RKHS with R as its reproducing kernel.

Theorem 3 (Moore-Aronszajn)**.** *Suppose that $R(\cdot,\cdot) : \mathcal{T} \times \mathcal{T} \to \mathbb{R}$ is a PSD kernel. Then there is a unique RKHS $\mathbb{H} \subset \mathbb{R}^{\mathcal{T}}$ with $R(\cdot,\cdot)$ as its reproducing kernel.*

Proof. Set

$$\mathbb{H}_0 := \text{span}\{R(\cdot, t) : t \in \mathcal{T}\} = \left\{ \sum_{i=1}^{n} a_i R(\cdot, t_i) \Big| n = 1, 2, \ldots, a_i \in \mathbb{R}, t_i \in \mathcal{T} \right\}.$$

Clearly $\mathbb{H}_0$ is a linear space. Define inner product $(\cdot,\cdot)_{\mathbb{H}_0}$ on $\mathbb{H}_0$ as

$$\left(\sum_{i=1}^{n_1} a_i R(\cdot, s_i), \sum_{j=1}^{n_2} b_j R(\cdot, t_j)\right)_{\mathbb{H}_0} = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} a_i b_j R(s_i, t_j). \tag{1}$$

This definition is indeed feasible since if $\sum_{i=1}^{n_1} a_i R(\cdot, s_i) = \sum_{i=1}^{n_1'} a_i' R(\cdot, s_i')$ and $\sum_{j=1}^{n_2} b_j R(\cdot, t_j) = \sum_{j=1}^{n_2'} b_j' R(\cdot, t_j')$, then

$$\begin{aligned}\left(\sum_{i=1}^{n_1'} a_i R(\cdot, s_i'), \sum_{j=1}^{n_2'} b_j' R(\cdot, t_j')\right)_{\mathbb{H}_0} &= \sum_{i=1}^{n_1'}\sum_{j=1}^{n_2'} a_i' b_j' R(s_i', t_j') = \sum_{j=1}^{n_2'} b_j' \left(\sum_{i=1}^{n_1'} a_i' R(s_i', t_j')\right) \\ &= \sum_{j=1}^{n_2'} b_j' \left(\sum_{i=1}^{n_1} a_i R(s_i, t_j')\right) = \sum_{i=1}^{n_1} a_i \left(\sum_{j=1}^{n_2'} b_j' R(s_i, t_j')\right) \\ &= \sum_{i=1}^{n_1} a_i \left(\sum_{j=1}^{n_2} b_j R(s_i, t_j)\right) = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} a_i b_j R(s_i, t_j).\end{aligned}$$

Now we check that $(\cdot,\cdot)_{\mathbb{H}_0}$ so defined is indeed an inner product. Clearly, $(\cdot,\cdot)_{\mathbb{H}_0}$ is bilinear and symmetric. The assumption that $R(\cdot,\cdot)$ is PSD ensures that $(f,f)_{\mathbb{H}_0} \geq 0$ for $f \in \mathbb{H}_0$. Hence $(\cdot,\cdot)_{\mathbb{H}_0}$ is a semi inner product. So it suffices to verify that $(f,f)_{\mathbb{H}_0} = 0$ implies $f = 0$. Note that the definition (1) implies that $(f, R(\cdot,t))_{\mathbb{H}_0} = f(t)$ for every $t \in \mathcal{T}$. Then for every $t \in \mathcal{T}$,

$$|f(t)| = |(f, R(\cdot,t))_{\mathbb{H}_0}| \leq \sqrt{(f,f)_{\mathbb{H}_0}(R(\cdot,t), R(\cdot,t))_{\mathbb{H}_0}} = 0.$$

Thus, $(\mathbb{H}_0, (\cdot,\cdot)_{\mathbb{H}_0})$ is an inner product space with reproducing kernel $R(\cdot,\cdot)$. However, $\mathbb{H}_0$ itself may not be complete and hence may not be a Hilbert space.

Now we proceed to complete $\mathbb{H}_0$. Suppose $\{f_n(\cdot)\}_{n=1}^{\infty}$ is a Cauchy sequence in $\mathbb{H}_0$. Since

$$|f_n(t) - f_m(t)| = |(f_n - f_m, R(\cdot,t))_{\mathbb{H}_0}| \leq \|f_n - f_m\|_{\mathbb{H}_0} \|R(\cdot,t)\|_{\mathbb{H}_0},$$

$\{f_n(t)\}$ is a Cauchy sequence in $\mathbb{R}$. Therefore, $\{f_n\}_{n=1}^{\infty}$ has a pointwise limit. Define

$$\mathbb{H} = \{f(\cdot) |\ f \text{ is the pointwise limit of some Cauchy sequence in } \mathbb{H}_0\}.$$

Clearly $\mathbb{H}$ is a linear space. Let $f(\cdot), g(\cdot) \in \mathbb{H}$, then there exist Cauchy sequences $\{f_n(\cdot)\}$ and $\{g_n(\cdot)\}$ such that $f(t) = \lim f_n(t)$, $g(t) = \lim g_n(t)$, for all $t \in \mathcal{T}$. Define $(f(\cdot), g(\cdot))_{\mathbb{H}} = \lim_{n\to\infty} (f_n(\cdot), g_n(\cdot))_{\mathbb{H}_0}$. This limit exists since $(f_n(\cdot), g_n(\cdot))_{\mathbb{H}_0}$ is a Cauchy sequence in $\mathbb{R}$:

$$\begin{aligned}&\left|(f_n(\cdot), g_n(\cdot))_{\mathbb{H}_0} - (f_m(\cdot), g_m(\cdot))_{\mathbb{H}_0}\right| \\ =&\left|(f_n(\cdot), g_n(\cdot))_{\mathbb{H}_0} - (f_n(\cdot), g_m(\cdot))_{\mathbb{H}_0}\right| + \left|(f_n(\cdot), g_m(\cdot))_{\mathbb{H}_0} - (f_m(\cdot), g_m(\cdot))_{\mathbb{H}_0}\right| \\ \leq&\|f_n(\cdot)\|_{\mathbb{H}_0}\|g_n - g_m\|_{\mathbb{H}_0} + \|g_m(\cdot)\|_{\mathbb{H}_0}\|f_n - f_m\|_{\mathbb{H}_0}.\end{aligned}$$

Also, this limit only depends on the limits $f(\cdot)$ and $g(\cdot)$. To see this, suppose there exist other Cauchy sequences $\{f_n'(\cdot)\}$ and $\{g_n'(\cdot)\}$ such that $f(t) = \lim f_n'(t)$, $g(t) = \lim g_n'(t)$, for all $t \in \mathcal{T}$. Let $f_n^{(d)}(t) = f_n(t) - f_n'(t)$. Then $\{f_n^{(d)}(\cdot)\}_{n=1}^{\infty}$ is also a Cauchy sequence in $\mathbb{H}_0$ and $f_n^{(d)}(t) \to 0$ for all $t \in \mathcal{T}$. We would like to show that $\|f_n^{(d)}(\cdot)\|_{\mathbb{H}_0} \to 0$. In fact,

$$\limsup_{n\to\infty} \limsup_{m\to\infty} \|f_n^{(d)}(\cdot) - f_m^{(d)}(\cdot)\|_{\mathbb{H}_0}^2 \to 0.$$

But

$$\begin{aligned}\limsup_{m\to\infty}\|f_n^{(d)}(\cdot)-f_m^{(d)}(\cdot)\|_{\mathbb{H}_0}^2 &= \limsup_{m\to\infty}\left(\|f_n^{(d)}(\cdot)\|_{\mathbb{H}_0}^2+\|f_m^{(d)}(\cdot)\|_{\mathbb{H}_0}^2-2(f_n^{(d)}(\cdot),f_m^{(d)}(\cdot))_{\mathbb{H}_0}\right)\\ &= \|f_n^{(d)}(\cdot)\|_{\mathbb{H}_0}^2+\limsup_{m\to\infty}\|f_m^{(d)}(\cdot)\|_{\mathbb{H}_0}^2 \qquad (2)\\ &\geq \|f_n^{(d)}(\cdot)\|_{\mathbb{H}_0}^2,\end{aligned}$$

where the second last equality holds since for fixed $f_n^{(d)}(\cdot):=\sum_{i=1}^k a_i R(\cdot,t_i)\in\mathbb{H}_0$,

$$\limsup_{m\to\infty}(f_n^{(d)}(\cdot),f_m^{(d)}(\cdot))_{\mathbb{H}_0}=\limsup_{m\to\infty}\sum_{i=1}^k a_i(R(\cdot,t_i),f_m^{(d)}(\cdot))_{\mathbb{H}_0}=\limsup_{m\to\infty}\sum_{i=1}^k a_i f_m^{(d)}(t_i)=0.$$

Let n tends to infinity in (2), we have $\|f_n^{(d)}(\cdot)\|_{\mathbb{H}_0}=\|f_n-f_n'\|_{\mathbb{H}_0}\to 0$. Similarly, we have $\|g_n-g_n'\|_{\mathbb{H}_0}\to 0$. Thus,

$$|(f_n(\cdot),g_n(\cdot))_{\mathbb{H}_0}-(f_n'(\cdot),g_n'(\cdot))_{\mathbb{H}_0}|\leq\|f_n(\cdot)\|_{\mathbb{H}_0}\|g_n-g_n'\|_{\mathbb{H}_0}+\|g_n'(\cdot)\|_{\mathbb{H}_0}\|f_n-f_n'\|_{\mathbb{H}_0}\to 0.$$

Hence

$$\lim_{n\to\infty}(f_n(\cdot),g_n(\cdot))_{\mathbb{H}_0}=\lim_{n\to\infty}(f_n'(\cdot),g_n'(\cdot))_{\mathbb{H}_0}$$

and $(f(\cdot),g(\cdot))_{\mathbb{H}}$ is well defined.

Its not hard to verify that $(\cdot,\cdot)_{\mathbb{H}}$ is an inner product in $\mathbb{H}$. By definition, for any $f(\cdot)\in\mathbb{H}$, there exsits a Cauchy sequence $\{f_n(\cdot)\}_{i=1}^\infty$ in $\mathbb{H}_0$ such that $f_n(t)\to f(t)$ for all $t\in\mathcal{T}$. We would like to show that $\|f_n-f\|_{\mathbb{H}}\to 0$. In fact

$$\begin{aligned}\|f_n-f\|_{\mathbb{H}}^2 =&\|f_n\|_{\mathbb{H}_0}^2+\|f\|_{\mathbb{H}}^2-2(f_n(\cdot),f(\cdot))_{\mathbb{H}}\\ =&\|f_n\|_{\mathbb{H}_0}^2+\lim_{m\to\infty}\|f_m\|_{\mathbb{H}_0}^2-2\lim_{m\to\infty}(f_n(\cdot),f_m(\cdot))_{\mathbb{H}_0}\\ =&\lim_{m\to\infty}\|f_n-f_m\|_{\mathbb{H}_0}^2.\end{aligned}$$

Let n tends to infty, we have $\lim_{n\to\infty}\|f_n-f\|_{\mathbb{H}}^2=0$.

For each Cauchy sequence $\{f_n(\cdot)\}_{i=1}^\infty$ in $\mathbb{H}$, there exists, by the above argument, $\{f_n'(\cdot)\}_{i=1}^\infty$ in $\mathbb{H}_0$ such that $\|f_n'-f_n\|\leq n^{-1}$. Hence $\{f_n'(\cdot)\}_{i=1}^\infty$ is also Cauchy, and thus converges to some $f(\cdot)$ in $\|\cdot\|_{\mathbb{H}}$. And thus $\{f_n(\cdot)\}_{i=1}^\infty$ also converges to $f(\cdot)$ in $\|\cdot\|_{\mathbb{H}}$.

We have proved that $(\mathbb{H},(\cdot,\cdot)_{\mathbb{H}})$ is complete. Its easy to verify that for every $f(\cdot)\in\mathbb{H}$, $(f(\cdot),R(\cdot,t))=f(t)$ for $t\in\mathcal{T}$. Thus $R(\cdot,\cdot)$ is the reproducing kernel of $\mathbb{H}$.

We turn to the uniqueness of $\mathbb{H}$. Any RKHS $\tilde{\mathbb{H}}$ must contain $R(\cdot,t)$, hence contain $\mathbb{H}_0$. $\mathbb{H}$ is the closure of $\mathbb{H}_0$ in $\tilde{\mathbb{H}}$. Hence also $\mathbb{H}\subset\tilde{\mathbb{H}}$. We have $\tilde{\mathbb{H}}=\mathbb{H}\oplus\mathbb{H}^\perp$. If $\mathbb{H}^\perp\neq\emptyset$, let $f\in\mathbb{H}^\perp$. Then $f(t)=(f,R(\cdot,t))_{\mathbb{H}}=0$ for every $t\in\mathbb{T}$, a contradiction. This completes the proof. □

We have shown that there is a one-to-one correspondence between kernels and RKHSs. For any kernel $R(\cdot,\cdot):\mathcal{T}\times\mathcal{T}\to\mathbb{R}$, let $\mathbb{H}$ be its RKHS. Define $\phi(\cdot):\mathcal{T}\to\mathbb{R}$ as $\phi(t)=R(\cdot,t)$. Then by reproducing property, $(\phi(s),\phi(t))_{\mathbb{H}}=R(s,t)$. That is, $\phi(\cdot)$ is a feature map of kernel $R(\cdot,\cdot)$. We call $\phi(\cdot)$ the *cononical feature map* Andreas Christmann (2008).

4 Kernel methods in machine learning

The earliest application of kernels and RKHS in machine learning is *kernel support vector machine*. A rich and comprehensive reference book of this topic is Andreas Christmann (2008). Since the introduction of kernel support vector machine, many kernels for specific learning tasks have been developed. See Hofmann et al. (2008) for a review.

4.1 Kernel principal component analysis

Principal component analysis (PCA) is a powerful technique for extracting structure from possibly high-dimensional data sets. Let $X_1, \dots, X_n \in \mathbb{R}^p$ be a sample drawn from distribution P with zero mean and covariance matrix Σ. Then $\mathbf{S} = n^{-1} \sum_{i=1}^n X_i X_i^\top$ is an estimator of Σ. Denote by λ_i $(i = 1, \dots, p)$ the ith largest eigenvalue of $\mathbf{S}$, and by v_i the corresponding eigenvector. We have

$$\mathbf{S} v_i = \lambda_i v_i \quad i = 1, \dots, p.$$

Then the first few eigenvectors $v_1, \dots, v_k$ $(k < p)$ capture the most variation. For a new data X_{new}, its principal component is $(v_1^\top X_{\text{new}}, \dots, v_k^\top X_{\text{new}})^\top \in \mathbb{R}^k$.

The kernel PCA algorithm is proposed by Schölkopf et al. (1998). It compute principal components in feature spaces. Consider the (measurable) feature map $\phi : \mathbb{R}^p \to \mathbb{H}$ where $\mathbb{H}$ is a (possibly infinite-dimensional) Hilbert space. Then $R(s,t) := (\phi(s), \phi(t))_{\mathbb{H}}$ is a PSD kernel. To simplify the problem, we assume that $\mathrm{E}\, \phi(X) = 0$. The sample covariance function of $\phi(X_1), \dots, \phi(X_n)$ is defined as $\mathbf{S} = n^{-1} \sum_{i=1}^n \phi(X_i) \otimes \phi(X_i)$. By definition, the tensor product $\phi(X_i) \otimes \phi(X_i)$ is an operator from $\mathbb{H}$ to $\mathbb{H}$ and $\phi(X_i) \otimes \phi(X_i)(x) = (\phi(X_i), x)_{\mathbb{H}} \phi(X_i)$ for every $x \in \mathbb{H}$. The image of $\mathbf{S}$ is finite dimensional and hence $\mathbf{S}$ is a self-adjoint compact operator Hsing and Eubank (2015). It follows that $\mathbf{S}$ has an eigen decomposition

$$\mathbf{S} = \sum_{i=1}^p \lambda_i v_i \otimes v_i,$$

where $\lambda_1 \geq \cdots \geq \lambda_p$ and v_i's are orthonormal: $(v_i, v_j)_{\mathbb{H}} = 0$ if $i \neq j$ and 1 if $i = j$. For a new data X_{new}, its principal component is $(v_1^\top \phi(X_{\text{new}}), \dots, v_k^\top \phi(X_{\text{new}}))^\top \in \mathbb{R}^k$.

Since $\mathbf{S}$ is infinite dimensional, it is hard to directly compute its eigen decomposition. Note that for $i = 1, \dots, p$,

$$(\mathbf{S}, v_i)_{\mathbb{H}} = \lambda_i v_i. \tag{3}$$

The left hand of (3) is

$$n^{-1} \sum_{k=1}^n (\phi(X_k) \otimes \phi(X_k), v_i)_{\mathbb{H}} = n^{-1} \sum_{k=1}^n (\phi(X_k), v_i)_{\mathbb{H}} \phi(X_k)$$

which is a linear combinition of $\phi(X_k)$. Hence we can write $v_i = \sum_{j=1}^n a_{ij} \phi(X_j)$. Then (3) becomes

$$n^{-1} \sum_{j=1}^n a_{ij} \sum_{k=1}^n R(X_k, X_j) \phi(X_k) = \lambda_i \sum_{j=1}^n a_{ij} \phi(X_j).$$

Taking the inner product with $\phi(X_l)$ $(l = 1, \ldots, n)$ on both sides yields

$$n^{-1} \sum_{j=1}^{n} a_{ij} \sum_{k=1}^{n} R(X_k, X_j) R(X_k, X_l) = \lambda_i \sum_{j=1}^{n} a_{ij} R(X_j, X_l).$$

Define the $n \times n$ *Gram matrix* G as $G_{ij} = R(X_i, X_j)$. Let $\mathbf{a}_i = (a_{i1}, \ldots, a_{in})^\top$. Then the last equality can be written as

$$n^{-1} G^2 \mathbf{a}_i = \lambda_i G \mathbf{a}_i.$$

So we only need to calculate the first k eigenvalues and eigenvectors of G. Then v_i is $\sum_{j=1}^{n} a_{ij} \phi(X_j)$ divided by its norm.

For a new data X_{new}, its ith $(i \leq k)$ principal component is

$$v_i^\top \phi(X_{\text{new}}) = \sum_{j=1}^{n} a_{ij} R(X_j, X_{\text{new}}).$$

Note that the algorithm of kernel PCA only depends on the kernel $R(\cdot, \cdot)$ and the feature map $\phi(\cdot)$ is not directly need. This makes the algorithm feasible and is called *kernel trick*.

4.2 Kernel Fisher discriminent analysis

Suppose $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ are a sample of training data, where $X_i \in \mathbb{R}^p$ is predictor and $Y_i \in \{1, 2\}$ is the corresponding label. Fisher's linear discriminant aims at finding a linear projections such that two classes are well separated. This is achieved by maximizing the Rayleigh quotient

$$J(w) = \frac{w^\top S_B w}{w^\top S_W w}$$

of between and within class variance with respect to $w \neq 0$ where

$$S_B = (\bar{X}_2 - \bar{X}_1)(\bar{X}_2 - \bar{X}_1)^\top, \quad S_W = \sum_{k=1}^{2} \sum_{\{i: Y_i = k\}} (X_i - \bar{X}_k)(X_i - \bar{X}_k)^\top.$$

Let w^* be the solution of this optimization problem (which in fact can be solved analytically). Suppose a new data X_{new} is given. If $|w^{*\top}(X_{\text{new}} - \bar{X}_1)| > |w^{*\top}(X_{\text{new}} - \bar{X}_2)|$, then we predict that it is from class 1. Otherwise, we predict that it is from class 2.

The idea Kernel Fisher discriminent analysis is to solve the problem of Fisher's linear discriminant in a feature space Muller et al. (2001). Suppose $\phi(\cdot)$ is a feature map from $\mathbb{R}^p$ to some inner product space $\mathbb{H}$ and $R(s, t) = (\phi(s), \phi(t))$. Suppose the linear span of $\{X_i\}$ is exactly $\mathbb{H}$. So for any $w \in \mathbb{H}$, there exists a_i, $i = 1, \ldots, n$, such that $w = \sum_{i=1}^{n} a_i \phi(X_i)$. We would like to maximize the Rayleigh quetiont

$$J(\omega) = \frac{\left(\omega, n_2^{-1} \sum_{\{i: Y_i = 2\}} \phi(X_i) - n_1^{-1} \sum_{\{i: Y_i = 1\}} \phi(X_i) \right)^2}{\sum_{k=1}^{2} \sum_{\{i: Y_i = k\}} \left(\omega, \phi(X_i) - n_k^{-1} \sum_{\sum_{\{l: Y_l = k\}}} \phi(X_{kl}) \right)^2}.$$

Let $\mathbf{1}_1$ be an n dimensional vector whose ith element equals 1 if $Y_i = 1$ and 0 otherwise. Let $\mathbf{1}_2$ be an n dimensional vector whose ith element equals 1 if $Y_i = 2$ and 0 otherwise. Let G be the Gram matrix with elements $G_{ij} = R(X_i, X_j)$. Then

$$J(\omega) = \frac{\left(\mathbf{a}^\top G \left(n_2^{-1}\mathbf{1}_2 - n_1^{-1}\mathbf{1}_1\right)\right)^2}{\mathbf{a}^\top \left(G^2 - \sum_{k=1}^{2} n_k^{-1} G \mathbf{1}_k \mathbf{1}_k^\top G\right) \mathbf{a}}. \tag{4}$$

Note that the matrix in the denominator is not full-rank. The reson is because that in the representation $w = \sum_{i=1}^n a_i \phi(X_i)$, there exists nonzero $\mathbf{a}$ corresponds to zero w. Hence (4) should be maximized for $\mathbf{a}$ such that $w \neq 0$. Some researchers proposed to regularize (4) by, e.g., adding a positive matrix in the denominator. See Muller et al. (2001) and the references therein.

5 Further theoretical results

5.1 Mercer's theorem

This subsection is mainly adapted from Hsing and Eubank (2015), Section 4.6.

Let $(\mathcal{T}, \mathcal{B}, \mu)$ be a measure space where $\mathcal{T}$ is a compact metric space, $\mathcal{B}$ is the Borel σ-field and μ is a finite measure. Suppose that $K(\cdot, \cdot)$ is a continuous function on $\mathcal{T} \times \mathcal{T}$ such that $\iint_{\mathcal{T}\times\mathcal{T}} K^2(s,t) d\mu(s) d\mu(t)$ is finite and define the integral operator $\mathcal{K}$ by

$$(\mathcal{K}f)(\cdot) := \int_{\mathcal{T}} K(s, \cdot) f(s) d\mu(s)$$

for $f \in \mathcal{L}^2(\mathcal{T}, \mathcal{B}, \mu)$. The function K si referred to as the *kernel* of $\mathcal{K}$. By Fubini's theorem, for $f \in \mathcal{L}^2(\mathcal{T}, \mathcal{B}, \mu)$, $(\mathcal{K}f)(\cdot)$ is measurable and satisfies

$$\int_{\mathcal{T}} (\mathcal{K}f)^2(t) d\mu(t) \leq \iint_{\mathcal{T}\times\mathcal{T}} K^2(s,t) d\mu(s) d\mu(t) \int_{\mathcal{T}} f^2(s) d\mu(s).$$

That is,

$$\|\mathcal{K}f\| \leq \left(\iint_{\mathcal{T}\times\mathcal{T}} K^2(s,t) d\mu(s) d\mu(t)\right)^{1/2} \|f\|.$$

Hence $\mathcal{K}$ is a continuous operator on $\mathcal{L}^2(\mathcal{T}, \mathcal{B}, \mu)$.

Lemma 1. *For each $f \in \mathcal{L}^2(\mathcal{T}, \mathcal{B}, \mu)$, $(\mathcal{K}f)(\cdot)$ is uniformly continuous.*

Proof. As K is uniformly continuous, for any given $\epsilon > 0$, there exists $\delta > 0$ such that $|K(s, t_2) - K(s, t_1)| < \epsilon$ for all $s, t_2, t_1 \in \mathcal{T}$ with $d(s_2, s_1) < \delta$. Thus,

$$\left| \int_{\mathcal{T}} K(s, t_2) f(s) d\mu(s) - \int_{\mathcal{T}} K(s, t_1) f(s) d\mu(s) \right| \leq \epsilon \sqrt{\mu(\mathcal{T})} \|f\|$$

□

Theorem 4. *$\mathcal{K}$ is a compact operator.*

Assume that K is symmetric. Then it can be seen that $\mathcal{K}$ is self-adjoint. A compact self-adjoint operator admits an eigenvalue-eigenvector decomposition $\mathcal{K} = \sum_{j=1}^{\infty} \lambda_j e_j \otimes e_j$. Since $e_j(t) = \lambda_j^{-1} \int_{\mathcal{T}} K(s,t) e_j(s) d\mu(s)$ for almost all t, Lemma 1 implies that $e_j(t)$ can be chosen to be continuous in t.

Theorem 5. *An integral operator is nonnegative definite if and only if its kernel is nonnegative definite.*

The following result is the celebrated Mercer's theorem.

Theorem 6. *Let $(\mathcal{T}, \mathcal{B}, \mu)$ be a measure space where $\mathcal{T}$ is a compact metric space, $\mathcal{B}$ is the Borel σ-field and μ is a finite measure. Suppose that the support of μ is the entire space $\mathcal{T}$. Let the continuous kernel K be symmetric and nonnegative definite and $\mathcal{K}$ the corresponding integral operator. If (λ_j, e_j) are the eigenvalue and eigenfunction pairs of $\mathcal{K}$, then K has the representation*

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

for all s, t, with the sum converging absolutely and uniformly.

Mercer's theorem gives conditions under which a PSD kernel has a series representation. This representation is very useful for many purposes. In particular, if a PSD kernel has such representation, its RKHS has a corresponding representation, as we shall see.

5.2 A representation of RKHS

The following result is adapted from van der Vaart and van Zanten (2008), Theorem 4.1.

Suppose $R(\cdot, \cdot) : \mathcal{T} \times \mathcal{T} \to \mathbb{R}$ is a PSD kernel and can be written in the form

$$R(s,t) = \sum_{j=1}^{\infty} \lambda_j f_j(s) f_j(t) \tag{5}$$

for positive numbers $\lambda_1, \lambda_2, \ldots$ and **arbitrary** functions $f_j : \mathcal{T} \to \mathbb{R}$, where the series is assumed to converge pointwise on $\mathcal{T} \times \mathcal{T}$. The convergence on the diagonal implies that $\sum_j \lambda_j f_j^2(t) < \infty$ for all $t \in \mathcal{T}$. Then by the Cauchy-Schwarz inequality the series $\sum_{j=1}^{\infty} w_j f_j(t)$ converges absolutely for every sequence $\{w_j\}$ of numbers with $\sum_j w_j^2/\lambda_j < \infty$, for every t, and hence defines a function from $\mathcal{T}$ to $\mathbb{R}$. We assume that the functions $\{f_j\}$ are linearly independent in the sense that $\sum_j w_j f_j(t) = 0$ for every $t \in \mathcal{T}$ for some sequence $\{w_j\}$ with $\sum_j w_j^2/\lambda_j < \infty$ implying that $w_j = 0$ for every $j \in \mathbb{N}$.

Theorem 7. *If the PSD kernel $R(\cdot,\cdot)$ can be represented as in* (5) *for linearly independent* $\{f_j(\cdot)\}$. *Then the RKHS of* $R(\cdot,\cdot)$ *is*

$$\mathbb{H} = \left\{ \sum_{j=1}^{\infty} w_j f_j(\cdot) \,\Big|\, \sum_{j=1}^{\infty} w_j^2/\lambda_j < \infty \right\},$$

and the inner product is given by

$$\left(\sum_{i=1}^{\infty} v_i f_i, \sum_{j=1}^{\infty} w_j f_j \right)_{\mathbb{H}} = \sum_{j=1}^{\infty} \frac{v_j w_j}{\lambda_j}. \tag{6}$$

Proof. Note that the functions in $\mathbb{H}$ are all well defined since $\sum_{j=1}^{\infty} w_j f_j(t)$ converges for every $t \in \mathcal{T}$. Also, by the assumed linear independence of the functions $\{f_i(\cdot)\}$, the coefficients $\{w_j\}$ are identifiable from the corresponding functions $\sum_{j=1}^{\infty} w_j f_j(\cdot) \in \mathbb{H}$. Therefore we can define a bijection $\imath : \mathbb{H} \to \ell_2$ by $\imath : \sum_{j=1}^{\infty} w_j f_j(\cdot) \mapsto \{w_j/\sqrt{\lambda_j}\}_{j=1}^{\infty}$. The set $\mathbb{H}$ becomes a Hilbert space under the inner product induced from ℓ_2, which is given on the right side of (6), and which we denote by $(\cdot,\cdot)_{\mathbb{H}}$.

To prove that $\mathbb{H}$ is the RKHS of $R(\cdot,\cdot)$, we need to show that (i) $R(\cdot,t) \in \mathbb{H}$ for every $t \in \mathcal{T}$ and (ii) for every $\sum_{j=1}^{\infty} w_j f_j(\cdot) \in \mathbb{H}$,

$$\left(\sum_{j=1}^{\infty} w_j f_j(\cdot), R(\cdot,t) \right)_{\mathbb{H}} = \sum_{j=1}^{\infty} w_j f_j(t).$$

For (i), note that the function $R(\cdot,t)$ has a representation $\sum_{j=1}^{\infty} w_j f_j(\cdot)$ for $w_j = \lambda_j f_j(t)$, and hence is contained in $\mathbb{H}$. For (ii), we have

$$\left(\sum_{j=1}^{\infty} w_j f_j(\cdot), R(\cdot,t) \right)_{\mathbb{H}} = \sum_{j=1}^{\infty} \frac{w_j \lambda_j f_j(t)}{\lambda_j} = \sum_{j=1}^{\infty} w_j f_j(t).$$

This completes the proof. □

5.3 Regularization and RKHS

This subsection is adapted from T. Hastie (2003), Section 5.8.1. See Wahba (1990) for more thorough treatment.

Consider a general (nonparametric) supervised learning problem with data $\{(X_i, Y_i)\}_{i=1}^n$, where $X_i \in \mathcal{T}$. We would like to learn some function of X. For example, if $Y_i = f(X_i) + \epsilon_i$ with $f(\cdot)$ unknown, we would like to estimate $f(\cdot)$.

A general learning strategy is to minimize the objective function

$$\sum_{i=1}^{n} L(Y_i, f(X_i)) + \lambda J(f),$$

where $L(\cdot,\cdot)$ is some Loss function, e.g., negative log likelihood, $J(\cdot)$ is a penalty function to regularize the solution and λ ia a tuning parameter.

Now we consider the following specific strategy. Let $R(\cdot,\cdot)$ be a PSD kernel which can be represented as in (5) for linearly independent $\{f_j(\cdot)\}$. Then its RKHS $\mathbb{H}$ is given by Theorem 7. Define the penalty function $J(\cdot)$ as $J(f) = \|f\|_{\mathbb{H}}^2$. Then we would like to solve the optimization problem

$$\min_{f\in\mathbb{H}} \left[\sum_{i=1}^{n} L(Y_i, f(X_i)) + \lambda\|f\|_{\mathbb{H}}^2\right]. \tag{7}$$

By theorem 7, the last display can be written as

$$\min_{\{w_j\}_{j=1}^{\infty}} \left[\sum_{i=1}^{n} L(Y_i, \sum_{j=1}^{\infty} w_j f_j(X_i)) + \lambda\sum_{j=1}^{\infty} w_j^2/\lambda_j\right] \quad \text{such that} \quad \sum_{j=1}^{\infty} w_j^2/\lambda_j < \infty.$$

This is an optimization over an infinite dimensional space. It turns out the solution will always be finite dimensional.

Theorem 8. *The solution of* (7) *takes the form*

$$f(x) = \sum_{i=1}^{n} \alpha_i R(x, x_i).$$

Proof. Let $\mathbb{H}_0 = \text{span}\{R(\cdot, x_i)\}_{i=1}^n$. Then $\mathbb{H}_0$ is a finite dimensional subspace of $\mathbb{H}$, and hence is also a Hilbert space. We need to show that the solution of (7) falls in $\mathbb{H}_0$. Let $f(\cdot)$ be any element of $\mathbb{H}$. By orthogonal decomposition theorem, we can write $f(\cdot) = g(\cdot) + h(\cdot)$ where $g(\cdot) \in \mathbb{H}_0$ and $h(\cdot) \in \mathbb{H}_0^{\perp}$. Hence we have $(h(\cdot), K(\cdot, x_i))_{\mathbb{H}} = 0$, $i = 1, \ldots, n$. Then reproducing property implies that $h(x_i) = 0$, $i = 1, \ldots, n$. It follows that $f(x_i) = g(x_i)$, $i = 1, \ldots, n$. On the other hand, $\|f\|_{\mathbb{H}}^2 = \|g\|_{\mathbb{H}}^2 + \|h\|_{\mathbb{H}}^2 \geq \|g\|_{\mathbb{H}}^2$. Thus,

$$\sum_{i=1}^{n} L(Y_i, f(X_i)) + \lambda\|f\|_{\mathbb{H}}^2 \geq \sum_{i=1}^{n} L(Y_i, g(X_i)) + \lambda\|g\|_{\mathbb{H}}^2.$$

This completes the proof.

□

6 Gaussian processes

A zero-mean Gaussian stochastic process $W = \{W_t : t \in \mathcal{T}\}$ is a set of random variables W_t indexed by an arbitrary set $\mathbb{T}$ and defined on a common probability space $(\Omega, \mathcal{U}, P)$ such that each finite subset possess a zero-mean multivariate normal distribution. The finite-dimensional distributions of such a process are determined by the covariance function $\Sigma(\cdot,\cdot) : \mathcal{T} \times \mathcal{T} \to \mathbb{R}$, defined by

$$\Sigma(s,t) = \mathrm{E}\, W_s W_t.$$

Clearly, for any Gaussian process, the associated covariance function is a PSD kernel. It follows from Kolmogorov's extension theorem that every PSD kernel is the covariance function of some Gaussian process. By Moore-Aronszajn theorem, there is a unique RKHS $(\mathbb{H}, (\cdot,\cdot)_{\mathbb{H}})$ with $\Sigma(\cdot,\cdot)$ being its reproducing kernel. We also call $\mathbb{H}$ the RKHS of Gaussian process $\{W_t\}$. Define a map $U : \mathbb{H} \to \mathcal{L}^2(\Omega,\mathcal{U}, P)$ by

$$U\Sigma(\cdot, t) = W_t,$$

and extending linearly and continuously. This map is an Hilbert space isometry of $\mathbb{H}$ and a closed subspace of $\mathcal{L}^2(\Omega,\mathcal{U}, P)$ since

$$\mathrm{E}\, U\Sigma(\cdot, s)U\Sigma(\cdot, t) = \mathrm{E}\, W_s W_t = \Sigma(s, t) = (\Sigma(\cdot, s), \Sigma(\cdot, t))_{\mathbb{H}}.$$

If $\mathcal{T}$ is finite dimensional, then $\{W_t\}$ is the familiar normal distribution. Suppose $\mathcal{T} = \{1, \ldots, p\}$. Then $W \sim N(0, \Sigma)$ for some $n \times n$ PSD matrix Σ. Then $\mathbb{H}$ is generated by the columns of Σ. That is $\mathbb{H} = \{\Sigma a : a \in \mathbb{R}^p\}$. The RKHS norm $(\cdot,\cdot)_{\mathbb{H}}$ is

$$(\Sigma a, \Sigma b)_{\mathbb{H}} = a^\top \Sigma b.$$

The map U is $U(\Sigma a) = a^\top W$.

If Σ is positive definite, then $\mathbb{H} = \mathbb{R}^p$, $(a, b)_{\mathbb{H}} = a^\top \Sigma^{-1} b$ and $Ua = a^\top \Sigma^{-1} W$. Let $f = \{f_i\}_{i=1}^p$ be a nonrandom p dimensional vector. Then $W + f \sim N(f, \Sigma)$. The likelihood ratio between $W + f$ and W is

$$\frac{dP^{W+f}}{dP^W}(W) = \exp\left(f^\top \Sigma^{-1} W - \frac{1}{2} f^\top \Sigma^{-1} f\right) = \exp\left(Uf - \frac{1}{2}\|f\|_{\mathbb{H}}^2\right).$$

This expression is also true not only for finite $\mathbb{T}$, but also for arbitrary $\mathbb{T}$. See Section 3 of van der Vaart and van Zanten (2008).

Such phenomenon may give a (loose) Bayes explanation of (7).

7 Other topics

- Karhunen-Loève theorem.
- Infinite dimensional exponential model Sriperumbudur et al. (2017):

  $$p_f(x) = \exp[f(x) - A(f)]q_0(x),$$

  where $f(\cdot)$ is the unknown parameter in a RKHS.
- Functional data analysis: Functional regression, Functional PCA.
- Bayesian nonparametrics.
- Consistency of Bayes factor: H_0 is a parametric model, H_1 is a semi-parametric model.

7.1 Results from Wahba's book

Sobolev space.

Product of RKHS.

One of the useful properties of RKHS is that from them one can obtain the representer of any bounded linear functional. Let η_i be the representer for bounded linear functional L, that is,

$$(\eta, f) = Lf, \quad \text{for all } f \in \mathbb{H}_R.$$

Then

$$\eta(s) = (\eta, R(s, \cdot)) = LR(s, \cdot).$$

References

Amini, A. A. (2013). High-dimensional principal component analysis. *Dissertations & Theses - Gradworks*, 61(3):464–473.

Andreas Christmann, I. S. a. (2008). *Support Vector Machines*. Information science and statistics. Springer-Verlag New York, 1 edition.

Gu, C. (2013). *Smoothing Spline ANOVA Models*. Springer Series in Statistics 297. Springer-Verlag New York, 2 edition.

Hofmann, T., Schölkopf, B., and Smola, A. J. (2008). Kernel methods in machine learning. *Ann. Statist.*, 36(3):1171–1220.

Hsing, T. and Eubank, R. (2015). *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. Wiley Series in Probability and Statistics. Wiley, 1 edition.

Muller, K. ., Mika, S., Ratsch, G., Tsuda, K., and Schölkopf, B. (2001). An introduction to kernel-based learning algorithms. *IEEE Transactions on Neural Networks*, 12(2):181–201.

Schölkopf, B., Smola, A., and Mller, K. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319.

Sriperumbudur, B., Fukumizu, K., Gretton, A., Hyvärinen, A., and Kumar, R. (2017). Density estimation in infinite dimensional exponential families. *Journal of Machine Learning Research*, 18(57):1–59.

T. Hastie, R. Tibshirani, J. H. F. (2003). *The Elements of Statistical Learning*. Springer, corrected edition.

van der Vaart, A. W. and van Zanten, J. H. (2008). *Reproducing kernel Hilbert spaces of Gaussian priors*, volume Volume 3 of *Collections*, pages 200–222. Institute of Mathematical Statistics, Beachwood, Ohio, USA.

Wahba, G. (1990). *Spline Models for Observational Data (CBMS-NSF Regional Conference Series in Applied Mathematics)*. SIAM: Society for Industrial and Applied Mathematics.