To-do list

Contents

2 Vector space of functions **2**
2.1 Some functional analysis . . . 3
2.2 Reproducing kernel Hilbert space theory . . . 12
2.3 Reproducing kernel Kreĭn space theory . . . 17
2.4 RKHS building blocks . . . 20
2.4.1 The RKHS of constant functions . . . 20
2.4.2 The canonical (linear) RKHS . . . 21
2.4.3 The fractional Brownian motion RKHS . . . 23
2.4.4 The squared exponential RKHS . . . 26
2.4.5 The Pearson RKHS . . . 28
2.5 Constructing RKKSs from existing RKHSs . . . 29
2.5.1 Sums, products and scaling of RKHS . . . 30
2.5.2 The polynomial RKKS . . . 32
2.5.3 The ANOVA RKKS . . . 33
2.6 Summary . . . 39

Bibliography **42**

Figures **43**

Tables **44**

Theorems **45**

Definitions **47**

Nomenclature **51**

Abbreviations **52**

Haziq Jamil
Department of Statistics
London School of Economics and Political Science
PhD thesis: 'Regression modelling using Fisher information covariance kernels (I-priors)'

Chapter 2

Vector space of functions

chapter2

For regression modelling with I-priors, it is assumed that the regression functions lie in some vector space of functions. The purpose of this chapter is to provide a concise review of functional analysis leading up to the theory of reproducing kernel Hilbert and Kreĭn spaces (RKHS/RKKS). The interest with these RKHSs and RKKSs is that these spaces have well established mathematical structure and offer desirable topologies. In particular, it allows the possibility of deriving the Fisher information for regression functions—this will be covered in Chapter 3. As we shall see, RKHSs are also extremely convenient in that they may be specified completely via their reproducing kernels. Several of these function spaces are of interest to us, for example, spaces of linear functions, smoothing functions, and functions whose inputs are nominal values and even functions themselves. RKHSs are widely studied in the applied statistical and machine learning literature, but perhaps RKKSs are less so. To provide an early insight, RKKSs are simply a generalisation of RKHSs, and are defined as the difference between two RKHSs. The flexibility provided by RKKSs will prove both useful and necessary, especially when considering sums and products of scaled function spaces, as is done in I-prior modelling.

It is emphasised that a deep knowledge of functional analysis, including RKHS and RKKS theory, is not at all necessary for I-prior modelling, so perhaps the advanced reader may wish to skip Sections 2.1 to 2.3. Section 2.4 describes the fundamental RKHS of interest for I-prior regression, which we refer to as the "building block" RKHSs. The reason for this is that it is possible to construct new function spaces from existing ones, and this is described in Section 2.5.

Two remarks before starting. Firstly, on notation: sets and vector spaces are denoted by calligraphic letters, and as much as possible, we shall stick to the convention that $\mathcal{F}$ denotes a function space, and $\mathcal{X}$ denotes the set of covariates or function inputs. Occasionally, we will describe a generic Hilbert space denoted by $\mathcal{H}$. Elements of the vector space of real functions over a set $\mathcal{X}$ are denoted $f(\cdot)$, but more commonly and

simply f. This distinguishes them from the actual evaluation of the function at an input point $x \in \mathcal{X}$, denoted $f(x) \in \mathbb{R}$. For a much cleaner read, we dispense with boldface notation for vectors and matrices when talking about them, without ambiguity, in the abstract sense. Secondly, on bibliography: references will be minimised throughout the presentation of this chapter, but a complete annotated bibliography at the end in Section 2.6.

2.1 Some functional analysis

sec:funcanalysis

The core study of functional analysis revolves around the treatment of functions as objects in vector spaces over a field[1]. Vector spaces, or linear spaces as they are sometimes known, may be endowed with some kind of structure so as to allow ideas such as closeness and limits to be conceived. Of particular interest to us is the structure brought about by *inner products*, which allow the rigorous mathematical study of various geometrical concepts such as lengths, directions, and orthogonality, among other things. We begin with the definition of an inner product.

def:innerprod

Definition 2.1 (Inner products)**.** Let $\mathcal{F}$ be a vector space over $\mathbb{R}$. A function $\langle \cdot, \cdot \rangle_{\mathcal{F}} : \mathcal{F} \times \mathcal{F} \to \mathbb{R}$ is said to be an inner product on $\mathcal{F}$ if all of the following are satisfied:

- **Symmetry**. $\langle f, g \rangle_{\mathcal{F}} = \langle g, f \rangle_{\mathcal{F}}$, $\forall f, g \in \mathcal{F}$.
- **Linearity**. $\langle \lambda_1 f_1 + \lambda_2 f_2, g \rangle_{\mathcal{F}} = \lambda_1 \langle f_1, g \rangle_{\mathcal{F}} + \lambda_2 \langle f_2, g \rangle_{\mathcal{F}}$, $\forall f_1, f_2, g \in \mathcal{F}$, $\forall \lambda_1, \lambda_2 \in \mathbb{R}$.
- **Non-degeneracy**. $\langle f, f \rangle_{\mathcal{F}} = 0 \Leftrightarrow f = 0$.

Additionally, an inner product is said to be *positive definite* if $\langle f, f \rangle_{\mathcal{F}} \geq 0$, $\forall f \in \mathcal{F}$. Inner products need not necessarily be positive definite, and we shall revisit this fact later when we cover Kreĭn spaces. However, for the purposes of the forthcoming discussion, the inner products that are referenced are the positive definite kind, unless otherwise stated.

We can always define a *norm* on $\mathcal{F}$ using the inner product as

$$\|f\|_{\mathcal{F}} = \sqrt{\langle f, f \rangle_{\mathcal{F}}}. \tag{2.1}$$

{eq:normip}

Norms are another form of structure that specifically captures the notion of length. This is defined below.

Definition 2.2 (Norms)**.** Let $\mathcal{F}$ be a vector space over $\mathbb{R}$. A non-negative function $||\cdot||_{\mathcal{F}} : \mathcal{F} \times \mathcal{F} \to [0, \infty)$ is said to be a norm on $\mathcal{F}$ if all of the following are satisfied:

- **Absolute homogeneity**. $||\lambda f||_{\mathcal{F}} = |\lambda| \, ||f||_{\mathcal{F}}$, $\forall \lambda \in \mathbb{R}$, $\forall f \in \mathcal{F}$

[1] In this thesis, this will be $\mathbb{R}$ exclusively.

- **Subadditivity**. $||f+g||_{\mathcal{F}} \leq ||f||_{\mathcal{F}} + ||g||_{\mathcal{F}}$, $\forall f, g \in \mathcal{F}$
- **Point separating**. $||f||_{\mathcal{F}} = 0 \Leftrightarrow f = 0$

The subadditivity property is also known as the *triangle inequality*. Also note that since $\|-f\|_{\mathcal{F}} = \|f\|_{\mathcal{F}}$, and by the triangle inequality and point separating property, we have that $\|f\|_{\mathcal{F}} = \frac{1}{2}\|f\|_{\mathcal{F}} + \frac{1}{2}\|-f\|_{\mathcal{F}} \geq \frac{1}{2}\|f-f\|_{\mathcal{F}} = 0$, thus implying non-negativity of norms. Several important relationships between norms and inner products hold in linear spaces, namely, the *Cauchy-Schwarz inequality*

$$|\langle f, g\rangle_{\mathcal{F}}| \leq \|f\|_{\mathcal{F}}\, \|g\|_{\mathcal{F}};$$

the *parallelogram law*

$$\|f+g\|_{\mathcal{F}}^2 + \|f-g\|_{\mathcal{F}}^2 = 2\|f\|_{\mathcal{F}}^2 + 2\|g\|_{\mathcal{F}}^2;$$

and the *polarisation identity* (in various forms)

$$\begin{aligned} \|f+g\|_{\mathcal{F}}^2 - \|f-g\|_{\mathcal{F}}^2 &= 4\langle f, g\rangle_{\mathcal{F}}, \\ \|f+g\|_{\mathcal{F}}^2 - \|f\|_{\mathcal{F}}^2 - \|g\|_{\mathcal{F}}^2 &= 2\langle f, g\rangle_{\mathcal{F}}, \\ -\|f-g\|_{\mathcal{F}}^2 + \|f\|_{\mathcal{F}}^2 + \|g\|_{\mathcal{F}}^2 &= 2\langle f, g\rangle_{\mathcal{F}}, \end{aligned}$$

for any $f, g \in \mathcal{F}$.

A vector space endowed with an inner product (c.f. norm) is called an inner product space (c.f. normed vector space). As a remark, inner product spaces can always be equipped with a norm using (2.1), but not always the other way around. A norm needs to satisfy the parallelogram law for an inner product to be properly defined.

The norm $||\cdot||_{\mathcal{F}}$, in turn, induces a metric (a notion of distance) on $\mathcal{F}$, i.e. $D(f, g) = ||f-g||_{\mathcal{F}}$, for $f, g \in \mathcal{F}$. With these notions of distances, one may talk about sequences of functions in $\mathcal{F}$ which are *convergent*, and sequences whose elements become arbitrarily close to one another as the sequence progresses (*Cauchy*).

Definition 2.3 (Convergent sequence)**.** A sequence $\{f_n\}_{n=1}^{\infty}$ of elements of a normed vector space $(\mathcal{F}, ||\cdot||_{\mathcal{F}})$ is said to *converge* to some $f \in \mathcal{F}$, if for every $\epsilon > 0$, $\exists N = N(\epsilon) \in \mathbb{N}$, such that $\forall n > N$, $||f_n - f||_{\mathcal{F}} < \epsilon$.

Definition 2.4 (Cauchy sequence)**.** A sequence $\{f_n\}_{n=1}^{\infty}$ of elements of a normed vector space $(\mathcal{F}, ||\cdot||_{\mathcal{F}})$ is said to be a Cauchy sequence if for every $\epsilon > 0$, $\exists N = N(\epsilon) \in \mathbb{N}$, such that $\forall n, m > N$, $||f_n - f_m||_{\mathcal{F}} < \epsilon$.

Every convergent sequence is Cauchy (from the triangle inequality), but the converse is not true. If the limit of the Cauchy sequence exists within the vector space, then the

sequence converges to it. A vector space is said to be *complete* if it contains the limits of all Cauchy sequences, or in other words, if every Cauchy sequence converges. There are special names given to complete vector spaces. A complete inner product space is known as a *Hilbert space*, while a complete normed space is called a *Banach space*. Out of interest, an inner product space that is not complete is sometimes known as a *pre-Hilbert space*, since its completion with respect to the norm induced by the inner product is a Hilbert space.

A subset $\mathcal{G} \subseteq \mathcal{F}$ is a *closed subspace* of $\mathcal{F}$ if it is closed under addition and multiplication by a scalar. That is, for any $g, g' \in \mathcal{G}$, $\lambda_1 g + \lambda_2 g'$ is also in $\mathcal{G}$, for $\lambda_1, \lambda_2 \in \mathbb{R}$. For Hilbert spaces, each closed subspace is also complete, and thus a Hilbert space in its own right. Although, as a remark, not every Hilbert subspace need be closed, and therefore complete.

Being vectors in a vector space, we can discuss mapping of vectors onto a another space, or in essence, having a function acted upon them. To establish terminology, we define linear and bilinear maps (operators).

Definition 2.5 (Linear map/operator)**.** Let $\mathcal{F}$ and $\mathcal{G}$ be two Hilbert spaces over $\mathbb{R}$. An operator A is a map from $\mathcal{F}$ to $\mathcal{G}$, and we denote its action on a function $f \in \mathcal{F}$ as $A(f) \in \mathcal{G}$, or simply $Af \in \mathcal{G}$. A *linear operator* satisfies $A(f + f') = A(f) + A(f')$ and $A(\lambda f) = \lambda A(f)$, for all $f, f' \in \mathcal{F}$ and $\lambda \in \mathbb{R}$. If $\mathcal{G}$ is the base field ($\mathbb{R}$ in our case), then the linear operator A is called a *linear functional*.

Definition 2.6 (Bilinear map/operator)**.** Let $\mathcal{F}$, $\mathcal{G}$ and $\mathcal{H}$ be Hilbert spaces over $\mathbb{R}$. A *bilinear operator* $B : \mathcal{F} \times \mathcal{G} \to \mathcal{H}$ is linear in each argument separately, i.e.

- $B(\lambda_1 f + \lambda_2 f', h) = \lambda_1 B(f, h) + \lambda_2 B(f', h)$; and
- $B(f, \lambda_1 g + \lambda_2 g') = \lambda_1 B(f, g) + \lambda_2 B(f, g')$,

for all $f, f' \in \mathcal{F}$, $g, g' \in \mathcal{G}$ and $\lambda_1, \lambda_2 \in \mathbb{R}$. In other words, the mappings $B_g : f \mapsto B(f, g)$ for any $g \in \mathcal{G}$, and $B_f : g \mapsto B(f, g)$ for any $f \in \mathcal{F}$, are both linear maps. If $\mathcal{F} \equiv \mathcal{G}$, then the bilinear map is *symmetric*. If $\mathcal{H}$ is the base field ($\mathbb{R}$ in our case), then B is called a *bilinear form*.

An interesting property of these operators to look at, besides linearity, is whether or not they are *continuous*.

def:continuity

Definition 2.7 (Continuity)**.** Let $\mathcal{F}$ and $\mathcal{G}$ be two Hilbert spaces. A function $A : \mathcal{F} \to \mathcal{G}$ is said to be *continuous at* $g \in \mathcal{F}$, if for every $\epsilon > 0$, $\exists \delta = \delta(\epsilon, g) > 0$ such that

$$\|f - g\|_{\mathcal{F}} < \delta \quad \Rightarrow \quad \|Af - Ag\|_{\mathcal{G}} < \epsilon.$$

A is *continuous* on $\mathcal{F}$, if it is continuous at every point $g \in \mathcal{F}$. If, in addition, δ depends on ϵ only, A is said to be *uniformly continuous*.

Continuity in the sense of linear operators here means that a convergent sequence in $\mathcal{F}$ can be mapped to a convergent sequence in $\mathcal{G}$. For a particular linear operator, the evaluation functional, this means that closeness in norm implies pointwise closeness—this relates to RKHSs, which is discussed in Section 2.2. There is an even stronger notion of continuity called *Lipschitz continuity*.

Definition 2.8 (Lipschitz continuity)**.** Let $\mathcal{F}$ and $\mathcal{G}$ be two Hilbert spaces. A function $A : \mathcal{F} \to \mathcal{G}$ is *Lipschitz continuous* if $\exists M > 0$ such that $\forall f, f' \in \mathcal{F}$,

$$\|Af - Af'\|_{\mathcal{G}} \leq M\|f - f'\|_{\mathcal{F}}.$$

Clearly, Lipschitz continuity implies uniform continuity: choose $\delta = \delta(\epsilon) := \epsilon/M$ and replace this in Definition 2.7. A continuous, linear operator is also one that is bounded.

def:boundedop

Definition 2.9 (Bounded operator)**.** The linear operator $A : \mathcal{F} \to \mathcal{G}$ between two Hilbert spaces $\mathcal{F}$ and $\mathcal{G}$ is said to be *bounded* if there exists some $M > 0$ such that

$$\|Af\|_{\mathcal{G}} \leq M\|f\|_{\mathcal{F}}.$$

The smallest such M is defined to be the *operator norm*, denoted $\|A\| := \sup_{f\in\mathcal{F}} \frac{\|Af\|_{\mathcal{G}}}{\|f\|_{\mathcal{F}}}$.

thm:boundcont

Lemma 2.1 (Equivalence of boundedness and continuity)**.** *Let $\mathcal{F}$ and $\mathcal{G}$ be two Hilbert spaces, and $A : \mathcal{F} \to \mathcal{G}$ a linear operator. A is bounded if and only if it is continuous.*

Proof. Suppose that A is bounded. Then, $\forall f, f' \in \mathcal{F}$, $\exists M > 0$ such that $\|A(f - f')\|_{\mathcal{G}} \leq M\|f - f'\|_{\mathcal{G}}$, so A is Lipschitz continuous. Conversely, let A be a continuous linear operator, especially at the zero vector. In other words, $\exists\delta > 0$ such that $\|A(f)\|_{\mathcal{G}} = \|A(f+0-0)\|_{\mathcal{G}} = \|A(f)-A(0)\| \leq 1$, $\forall f \in \mathcal{F}$ whenever $\|f\|_{\mathcal{F}} \leq \delta$. Thus, for all non-zero $f \in \mathcal{F}$,

$$\begin{aligned}
\|A(f)\|_{\mathcal{G}} &= \left\| \frac{\|f\|_{\mathcal{F}}}{\delta} A\left(\frac{\delta}{\|f\|_{\mathcal{F}}} f\right) \right\|_{\mathcal{G}} \\
&= \left| \frac{\|f\|_{\mathcal{F}}}{\delta} \right| \left\| A\left(\frac{\delta}{\|f\|_{\mathcal{F}}} f\right) \right\|_{\mathcal{G}} \\
&\leq \frac{\|f\|_{\mathcal{F}}}{\delta} \cdot 1,
\end{aligned}$$

and therefore A is bounded. ■

So important is the concept of linearity and continuity, that there are specially named spaces which contain linear and continuous functionals.

Definition 2.10 (Dual spaces)**.** Let $\mathcal{F}$ be a Hilbert space. The space $\mathcal{F}^\vee$ of *linear functionals* is called the *algebraic dual space* of $\mathcal{F}$. The space $\mathcal{F}^*$ of *continuous linear functionals* is called the *continuous dual space* or alternatively, the *topological dual space*, of $\mathcal{F}$.

As it turns out, the algebraic dual space and continuous dual space coincide in finite-dimensional Hilbert spaces: take any $A \in \mathcal{F}^\vee$; since A is finite-dimensional, it is bounded, and therefore continuous (see Lemma 2.1), so $A \in \mathcal{F}^*$ and $\mathcal{F}^\vee \subseteq \mathcal{F}^*$; but $\mathcal{F}^* \subseteq \mathcal{F}^\vee$ trivially, so $\mathcal{F}^\vee \equiv \mathcal{F}^*$. For infinite-dimensional Hilbert spaces, this is not so, but in any case, we will only be considering the continuous dual space in this thesis. The following result is an important one, which states that continuous linear functionals of an inner product space are nothing more than inner products.

Theorem 2.2 (Riesz-Fréchet)**.** *Let $\mathcal{F}$ be a Hilbert space. Every element A of the continuous dual space $\mathcal{F}^*$, i.e. all continuous linear functionals $A : \mathcal{F} \to \mathbb{R}$, can be uniquely written in the form $\langle \cdot, g \rangle_{\mathcal{F}} =: A_g \in \mathcal{F}^*$, for some $g \in \mathcal{F}$. Moreover, $\|g\|_{\mathcal{F}} = \|A_g\|_{\mathcal{F}^*}$.*

Proof. Omitted—see Yamamoto (2012, Theorem 4.2.1) for a proof. ■

Remark 2.1. The Riesz-Fréchet theorem is also commonly referred to as the *Riesz representation theorem* for Hilbert spaces.

The notion of isometry (transformation that preserves distance) is usually associated with metric spaces; two metric spaces being isometric means that they identical as far as their metric properties are concerned. For Hilbert spaces (and more generally, for normed spaces), there is an analogous concept as well in *isometric isomorphism* (a bijective isometry), such that two Hilbert spaces being isometrically isomorphic imply that they have exactly the same geometric structure, but may very well contain fundamentally different objects.

Definition 2.11 (Isometric isomorphism)**.** Two Hilbert spaces $\mathcal{F}$ and $\mathcal{G}$ are said to be *isometrically isomorphic*, symbolised $\mathcal{F} \cong \mathcal{G}$, if there is a linear bijective map $U : \mathcal{F} \to \mathcal{G}$ which preserves the inner product, i.e. for any $f, f' \in \mathcal{F}$,

$$\langle f, f' \rangle_{\mathcal{F}} = \langle Uf, Uf' \rangle_{\mathcal{G}}.$$

A consequence of the Riesz-Fréchet theorem is that it gives us a canonical isometric isomorphism $U : g \mapsto \langle \cdot, g \rangle_{\mathcal{F}} =: A_g$ between $\mathcal{F}$ and its continuous dual $\mathcal{F}^*$: A_g is obviously linear (bilinearity of inner products), and using the polarisation identity,

$$\begin{aligned} 2\langle Ug, Ug' \rangle_{\mathcal{F}^*} &= \|U(g)\|^2_{\mathcal{F}^*} + \|U(g')\|^2_{\mathcal{F}^*} - \|U(g - g')\|^2_{\mathcal{F}^*} \\ &= \|g\|^2_{\mathcal{F}} + \|g'\|^2_{\mathcal{F}} - \|g - g'\|^2_{\mathcal{F}} \\ &= 2\langle g, g' \rangle_{\mathcal{F}}. \end{aligned}$$

Implicitly, this means that $\mathcal{F}^*$ is a Hilbert space as well.

Another important type of mapping is the mapping P of an element in $\mathcal{F}$ onto a closed subspace $\mathcal{G} \subset \mathcal{F}$, such that $Pf \in \mathcal{G}$ is closest to f. This mapping is called the *orthogonal projection*, due to the fact that such projections yield perpendicularity in the sense that $\langle f - Pf, g\rangle_{\mathcal{F}} = 0$ for any $g \in \mathcal{G}$. Consequently, we see that $\|f\|_{\mathcal{F}}^2 = \|Pf\|_{\mathcal{F}}^2 + \|f - Pf\|_{\mathcal{F}}^2$ from the polarisation identity. The remainder $f - Pf$ belongs to the *orthogonal complement* of $\mathcal{G}$.

Definition 2.12 (Orthogonal complement)**.** Let $\mathcal{F}$ be a Hilbert space and $\mathcal{G} \subset \mathcal{F}$ be a closed subspace. The linear subspace $\mathcal{G}^\perp = \{f \mid \langle f, g\rangle_{\mathcal{F}} = 0, \forall g \in \mathcal{G}\}$ is called the orthogonal complement of $\mathcal{G}$ in $\mathcal{F}$.

thm:orthdecomp

Theorem 2.3 (Orthogonal decomposition)**.** *Let $\mathcal{F}$ be a Hilbert space and $\mathcal{G} \subset \mathcal{F}$ be a closed subspace. For every $f \in \mathcal{F}$, we can write $f = g + g^c$, where $g \in \mathcal{G}$ and $g^c \in \mathcal{G}^\perp$, and this decomposition is unique.*

Proof. Omitted—see Rudin (1987, Theorem 4.11) for a proof. ■

We can write $\mathcal{F} = \mathcal{G} \oplus \mathcal{G}^\perp$, where the $\oplus$ symbol denotes the *direct sum*, and such a decomposition is called a *tensor sum decomposition*. In infinite-dimensional Hilbert spaces, some subspaces are not closed, but all orthogonal complements are closed. In such spaces, the orthogonal complement of the orthogonal complement of $\mathcal{G}$ is the closure of $\mathcal{G}$, i.e. $(\mathcal{G}^\perp)^\perp =: \overline{\mathcal{G}}$, and we say that $\mathcal{G}$ is *dense* in $\overline{\mathcal{G}}$. Another interesting fact regarding the orthogonal complement is that $\mathcal{G} \cap \mathcal{G}^\perp = \{0\}$, since any $g \in \mathcal{G} \cap \mathcal{G}^\perp$ must be orthogonal to itself, i.e. $\langle g, g\rangle_{\mathcal{G}} = 0$ implying that $g = 0$.

The following theorem states that orthogonal decompositions are unique.

thm:orthdecomp2

Corollary 2.3.1. *Let $\mathcal{G}$ be a subspace of a Hilbert space $\mathcal{F}$. Then, $\mathcal{G}^\perp = \{0\}$ if and only if $\mathcal{G}$ is dense in $\mathcal{F}$.*

Proof. If $\mathcal{G}^\perp = \{0\}$ then $(\mathcal{G}^\perp)^\perp = \overline{\mathcal{G}} = \mathcal{F}$. Conversely, since $\mathcal{G}$ is dense in $\mathcal{F}$, we have $\mathcal{G}^\perp = \overline{\mathcal{G}}^\perp = \mathcal{F}^\perp = \{0\}$. ■

Besides tensor sums, of importance is the concept of *tensor products*, which can be thought of as a generalisation of the outer product in Euclidean space.

def:tensorprod

Definition 2.13 (Tensor products)**.** Let $x_1 \in \mathcal{H}_1$ and $x_2 \in \mathcal{H}_2$ be two elements of two real Hilbert spaces. Then, the tensor product $x_1 \otimes x_2 : \mathcal{H}_1 \times \mathcal{H}_2 \to \mathbb{R}$, is a bilinear form defined as

$$(x_1 \otimes x_2)(y_1, y_2) = \langle x_1, y_1\rangle_{\mathcal{H}_1} \langle x_2, y_2\rangle_{\mathcal{H}_2}$$

for any $(y_1, y_2) \in \mathcal{H}_1 \times \mathcal{H}_2$.

Correspondingly, we may also define the *tensor product space*.

def:tensprod space

Definition 2.14 (Tensor product space)**.** The tensor product space $\mathcal{H}_1 \otimes \mathcal{H}_2$ is the completion of the space

$$\mathcal{A} = \left\{ \sum_{j=1}^{J} x_{1j} \otimes x_{2j} \,\middle|\, x_{1j} \in \mathcal{H}_1, x_{2j} \in \mathcal{H}_2, J \in \mathbb{N} \right\}.$$

with respect to the norm induced by the inner product

$$\left\langle \sum_{j=1}^{J} x_{1j} \otimes x_{2j}, \sum_{k=1}^{K} y_{1k} \otimes y_{2k} \right\rangle_{\mathcal{A}} = \sum_{j=1}^{J} \sum_{k=1}^{K} \langle x_{1j}, y_{1k} \rangle_{\mathcal{H}_1} \langle x_{2j}, y_{2k} \rangle_{\mathcal{H}_2}.$$

Interestingly, the tensor product can be viewed as an operator between two Hilbert spaces. That is, for each pair of elements $(x_1, x_2) \in \mathcal{H}_1 \times \mathcal{H}_2$, we define the operator $A_{x_1,x_2} : \mathcal{H}_1 \to \mathcal{H}_2$ in the following way:

$$\begin{aligned} A_{x_1,x_2} : \mathcal{H}_1 &\to \mathcal{H}_2 \\ y_1 &\mapsto \langle x_1, y_1 \rangle_{\mathcal{H}_1} x_2. \end{aligned}$$

Incidentally, an operator defined in such a way is called a *rank one* operator. Indeed, for some $y_1 \in \mathcal{H}_1$ and $y_2 \in \mathcal{H}_2$, we have that

$$\begin{aligned} \langle A_{x_1,x_2}(y_1), y_2 \rangle_{\mathcal{H}_2} &= \big\langle \langle x_1, y_1 \rangle_{\mathcal{H}_1} x_2, y_2 \big\rangle_{\mathcal{H}_2} \\ &= \langle x_1, y_1 \rangle_{\mathcal{H}_1} \langle x_2, y_2 \rangle_{\mathcal{H}_2} \\ &= (x_1 \otimes x_2)(y_1, y_2). \end{aligned}$$

We now have three distinct interpretations of the tensor product. For $x_1, y_1 \in \mathcal{H}_1$ and $x_2, y_2 \in \mathcal{H}_2$, these are:

- **General form** (as an element in the tensor product space).

$$x_1 \otimes x_2 \in \mathcal{H}_1 \otimes \mathcal{H}_2.$$

- **Operator**.

$$\begin{aligned} x_1 \otimes x_2 : \mathcal{H}_1 &\to \mathcal{H}_2 \\ y_1 &\mapsto \langle x_1, y_1 \rangle_{\mathcal{H}_1} x_2 \end{aligned}$$

- **Bilinear form** (as per Definition 2.13).

$$\begin{aligned} x_1 \otimes x_2 : \mathcal{H}_1 \times \mathcal{H}_2 &\to \mathbb{R} \\ (y_1, y_2) &\mapsto \langle x_1, y_1 \rangle_{\mathcal{H}_1} \langle x_2, y_2 \rangle_{\mathcal{H}_2} \end{aligned}$$

Remark 2.2. As explained by Kokoszka and Reimherr (2017, §10.5, p. 227), tensors are often thought of as generalisations of matrices. For example, in Euclidean space, a matrix $\mathbf{A} \in \mathbb{R}^{n\times m}$, formed by two vectors $x_1 \in \mathbb{R}^n$ and $x_2 \in \mathbb{R}^m$ via $\mathbf{A} = x_1 x_2^\top =: x_1 \otimes x_2$, can be viewed in at least three ways: 1) as a traditional matrix in the space $\mathbb{R}^n \otimes \mathbb{R}^m = \mathbb{R}^{n\times m}$; 2) as a linear transformation in Euclidean space $\mathbf{A} : \mathbb{R}^n \to \mathbb{R}^m$ (or the reverse) by multiplying $\mathbf{A}$ from the left or right by a vector; or 3) as a bilinear mapping $\mathbf{A} : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ in the form of $\mathbf{A}(y_1, y_2) = y_1^\top \mathbf{A} y_2 = y_1^\top x_1 x_2^\top y_2 = (y_1^\top x_1)(y_2^\top x_2)$, for some $y_1 \in \mathbb{R}^n$ and $y_2 \in \mathbb{R}^m$, arising often in the study of quadratic forms.

For the last part of this introductory section on functional analysis, we discuss measures on Hilbert spaces, and in particular, a probability measure. Let $\mathcal{H}$ be a real Hilbert space. As discussed earlier, we can define a metric on $\mathcal{H}$ using $D(x, x') = \|x - x'\|_{\mathcal{H}}$, where the norm on $\mathcal{H}$ is the norm induced by the inner product. A collection Σ of subsets of $\mathcal{H}$ is called a *σ-algebra* if $\emptyset \in \Sigma$, $S \in \Sigma$ implies its complement $S^c \in \Sigma$, and $S_j \in \Sigma$, $j \geq 1$ implies $\bigcup_{j=1}^\infty S_j \in \Sigma$. The smallest σ-algebra containing all open subsets of $\mathcal{H}$ is called the *Borel σ-algebra*, and its members the Borel sets. Denote by $\mathcal{B}(\mathcal{H})$ the Borel σ-algebra of $\mathcal{H}$.

Recall that a function $\nu : \Sigma \to [0, \infty]$ is called a *measure* if it satisfies

- **Non-negativity:** $\nu(S) \geq 0$ for all S in Σ;
- **Null empty set:** $\nu(\emptyset) = 0$; and
- **σ-additivity:** for all countable, mutually disjoint sets $\{S_i\}_{i=1}^\infty$,

$$\nu\left(\bigcup_{i=1}^\infty S_i\right) = \sum_{i=1}^\infty \nu(S_i).$$

A measure ν on $\big(\mathcal{H}, \mathcal{B}(\mathcal{H})\big)$ is called a *Borel measure* on $\mathcal{H}$. We shall only concern ourselves with finite Borel measures. In addition, if $\nu(\mathcal{H}) = 1$ then ν is a *(Borel) probability measure* and the measure space $\big(\mathcal{H}, \mathcal{B}(\mathcal{H}), \nu\big)$ is a *(Borel) probability space.*

Let $(\Omega, \mathcal{E}, \mathrm{P})$ be a probability space. We say that a mapping $X : \Omega \to \mathcal{H}$ is a *random element* in $\mathcal{H}$ if $X^{-1}(B) \in \mathcal{E}$ for every Borel set, i.e., X is a function such that for every $B \in \mathcal{B}(\mathcal{H})$, its preimage $X^{-1}(B) = \{\omega \in \Omega \,|\, X(\omega) \in B\}$ lies in $\mathcal{E}$. This is simply a generalisation of the definition of random variables in regular Euclidean space. From this definition, we can also properly define random functions f in a Hilbert space of functions $\mathcal{F}$. In any case, every random element X induces a probability measure on $\mathcal{H}$ defined by

$$\nu(B) = \mathrm{P}\left(X^{-1}(B)\right) = \mathrm{P}\left(\omega \in \Omega | X(\omega) \in B\right) = \mathrm{P}(X \in B).$$

The measure ν is called the *distribution* of X. The *density* p of X is a measurable function with the property that

$$\mathrm{P}(X \in B) = \int_{X^{-1}(B)} \omega \,\mathrm{dP}(\omega) = \int_B p(x) \,\mathrm{d}\nu(x).$$

Definition 2.15 (Mean vector)**.** Let ν be a Borel probability measure on a real Hilbert space $\mathcal{H}$. Supposing that a random element X of $\mathcal{H}$ is *integrable*, that is to say

$$\mathrm{E}\|X\|_{\mathcal{H}} = \int_{\mathcal{H}} \|z\|_{\mathcal{H}} \,\mathrm{d}\nu(z) < \infty,$$

then the unique element $\mu \in \mathcal{H}$ satisfying

$$\langle \mu, x \rangle = \int_{\mathcal{X}} \langle z, x \rangle_{\mathcal{X}} \,\mathrm{d}\nu(z) = \mathrm{E}\langle X, x \rangle_{\mathcal{H}}$$

for all $x \in \mathcal{H}$ is called the *mean vector*.

def:covoper

Definition 2.16 (Covariance operator)**.** Let ν be a Borel probability measure on a real Hilbert space $\mathcal{H}$. Suppose that a random element X of $\mathcal{H}$ is *square integrable*, i.e., $\mathrm{E}\|X\|_{\mathcal{H}}^2 < \infty$, and let μ be the mean vector of X. Then the *covariance operator* C is defined by the mapping

$$\begin{aligned} C : \mathcal{H} &\to \mathcal{H} \\ x &\mapsto \mathrm{E}\left[\langle X - \mu, x \rangle_{\mathcal{H}} (X - \mu)\right]. \end{aligned}$$

The covariance operator C is also an element of $\mathcal{H} \otimes \mathcal{H}$ that satisfies

$$\begin{aligned} \langle C, x \otimes x' \rangle_{\mathcal{H} \otimes \mathcal{H}} &= \int_{\mathcal{H}} \langle z - \mu, x \rangle_{\mathcal{H}} \langle z - \mu, x' \rangle_{\mathcal{H}} \,\mathrm{d}\nu(z) \\ &= \mathrm{E}\left[\langle X - \mu, x \rangle_{\mathcal{H}} \langle X - \mu, x' \rangle_{\mathcal{H}}\right] \end{aligned}$$

for all $x, x' \in \mathcal{H}$.

From the definition of the covariance operator, we see that it induces a symmetric, bilinear form, which we shall denote by $\mathrm{Cov} : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$, through

$$\begin{aligned} \langle Cx, x' \rangle_{\mathcal{H}} &= \left\langle \mathrm{E}\left[\langle X - \mu, x \rangle_{\mathcal{H}} (X - \mu)\right], x' \right\rangle_{\mathcal{H}} \\ &= \mathrm{E}\left[\langle X - \mu, x \rangle_{\mathcal{H}} \langle X - \mu, x' \rangle_{\mathcal{H}}\right] \\ &=: \mathrm{Cov}[x, x']. \end{aligned}$$

Definition 2.17 (Gaussian vectors)**.** A random element X is called *Gaussian* if $\langle X, x \rangle_{\mathcal{H}}$ has a normal distribution for all fixed $x \in \mathcal{H}$. A Gaussian vector X is characterised by its mean element $\mu \in \mathcal{H}$ and its covariance $C \in \mathcal{H} \otimes \mathcal{H}$.

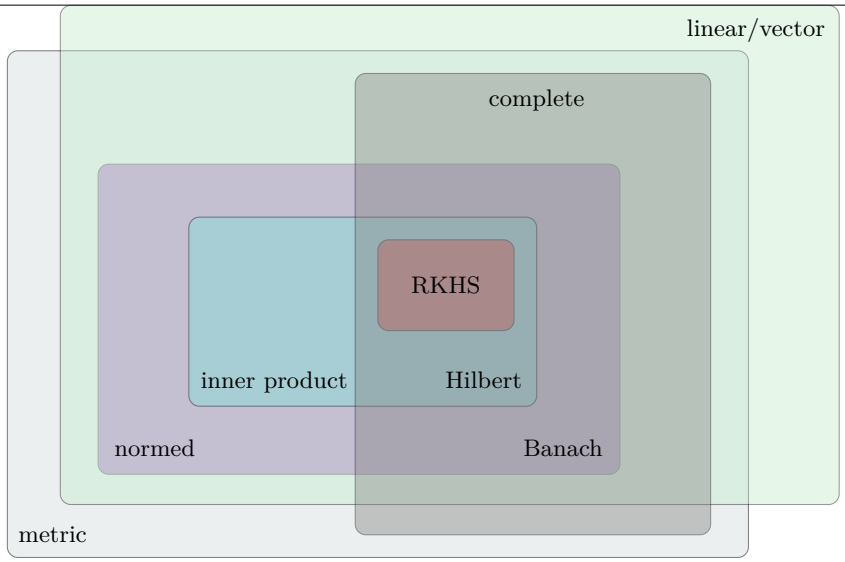

Figure 2.1: A hierarchy of vector spaces[2].

2.2 Reproducing kernel Hilbert space theory

sec:rkhstheory

The introductory section sets us up nicely to discuss the coveted reproducing kernel Hilbert space. This is a subset of Hilbert spaces for which its evaluation functionals are continuous (by definition, in fact). The majority of this section, apart from defining RKHSs, is an exercise in convincing ourselves that each and every RKHS of functions can be specified solely through its reproducing kernel. To begin, we consider a fundamental linear functional on a Hilbert space of functions $\mathcal{F}$, that assigns a value to $f \in \mathcal{F}$ for each $x \in \mathcal{X}$, called the *evaluation functional*.

Definition 2.18 (Evaluation functional)**.** Let $\mathcal{F}$ be a vector space of functions $f : \mathcal{X} \to \mathbb{R}$, defined on a non-empty set $\mathcal{X}$. For a fixed $x \in \mathcal{X}$, the functional $\delta_x : \mathcal{F} \to \mathbb{R}$ as defined by $\delta_x(f) = f(x)$ is called the (Dirac) evaluation functional at x.

It is easy to see that evaluation functionals are always linear: $\delta_x(\lambda f + g) = (\lambda f + g)(x) = \lambda f(x) + g(x) = \lambda\delta_x(f) + \delta_x(g)$ for $\lambda \in \mathbb{R}$, $f, g \in \mathcal{F}$ real functions over $\mathcal{X}$. Humble as they may seem, the entirety of the evaluation functionals over the domain $\mathcal{X}$ determines f uniquely, and thus are of great importance in understanding the space $\mathcal{F}$. Core topological properties like convergence are hinged on continuity, and it is therefore important that evaluation functionals are continuous. As it turns out, RKHSs by definition provide exactly this.

[2]Reproduced from the lecture slides of Dino Sejdinovic and Arthur Gretton entitled 'Foundations of Reproducing Kernel Hilbert Spaces: Advanced Topics in Machine Learning', 2014. URL: http://www.stats.ox.ac.uk/~sejdinov/teaching/atml14/Theory_slides2_2014.pdf.

def:rkhs

Definition 2.19 (Reproducing kernel Hilbert space)**.** A Hilbert space $\mathcal{F}$ of real-valued functions $f : \mathcal{X} \to \mathbb{R}$ on a non-empty set $\mathcal{X}$ is called a *reproducing kernel Hilbert space* if the evaluation functional $\delta_x : f \mapsto f(x)$ is continuous (equivalently, bounded) on $\mathcal{F}$, $\forall x \in \mathcal{X}$.

Continuity (boundedness) of evaluation functionals in an RKHS means that functions that are close in RKHS norm imply that they are also close pointwise, since $|\delta_x(f) - \delta_x(g)| = |\delta_x(f-g)| \leq M\|f-g\|_{\mathcal{F}}$ for some real $M > 0$. Note that the converse is not necessarily true. RKHSs are particularly well behaved in this respect, compared to other Hilbert spaces, and this property in particular has desirable consequences for a wide variety of applications, including nonparametric curve estimation, learning and decision theory, and many more.

While the continuity condition by definition is what makes an RKHS, it is neither easy to check this condition in practice, nor is it intuitive as to the meaning of its name. In fact, there isn't even any mention of what a reproducing kernel actually is. In order to benefit from the desirable continuity property of RKHS, we should look at this from another, more intuitive, perspective. By invoking the Riesz representation theorem, we see that for all $x \in \mathcal{X}$, there exists a unique element $h_x \in \mathcal{F}$ such that

$$f(x) = \delta_x(f) = \langle f, h_x \rangle_{\mathcal{F}}, \forall f \in \mathcal{F}$$

holds. Since h_x itself is a function in $\mathcal{F}$, it holds that for every $x' \in \mathcal{X}$ there exists a $h_{x'} \in \mathcal{F}$ such that

$$h_x(x') = \delta_{x'}(h_x) = \langle h_x, h_{x'} \rangle_{\mathcal{F}}.$$

This leads us to the definition of a *reproducing kernel* of an RKHS—the very notion that inspires its name.

def:repkern

Definition 2.20 (Reproducing kernels)**.** Let $\mathcal{F}$ be a Hilbert space of functions over a non-empty set $\mathcal{X}$. A symmetric, bivariate function $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a *kernel*, and it is a *reproducing kernel* of $\mathcal{F}$ if h satisfies

- $\forall x \in \mathcal{X}$, $h(\cdot, x) \in \mathcal{F}$; and
- $\forall x \in \mathcal{X}$, $f \in \mathcal{F}$, $\langle f, h(\cdot, x) \rangle_{\mathcal{F}} = f(x)$ (the reproducing property).

In particular, for any $x, x' \in \mathcal{X}$,

$$h(x, x') = \langle h(\cdot, x), h(\cdot, x') \rangle_{\mathcal{F}}.$$

An important property for reproducing kernels of a RKHS is that they are positive definite functions. That is, $\forall a_1, \dots, a_n \in \mathbb{R}$ and $\forall x_1, \dots, x_n \in \mathcal{X}$,

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j h(x_i, x_j) \geq 0.$$

thm:posdef

Proposition 2.4 (Reproducing kernels of RKHS are positive-definite)**.** *Let $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a reproducing kernel for a Hilbert space $\mathcal{F}$. Then h is a symmetric and positive definite function.*

Proof.

$$\begin{aligned}
\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j h(x_i, x_j) &= \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j \langle h(\cdot, x_i), h(\cdot, x_j) \rangle_{\mathcal{F}} \\
&= \left\langle \sum_{i=1}^{n} a_i h(\cdot, x_i), \sum_{j=1}^{n} a_j h(\cdot, x_j) \right\rangle_{\mathcal{F}} \\
&= \left\| \sum_{i=1}^{n} a_i h(\cdot, x_i) \right\|_{\mathcal{F}}^2 \\
&\geq 0
\end{aligned}$$

■

Remark 2.3. In the kernel method literature, a *kernel* $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is usually defined as the inner product between inputs in feature space. That is, take $\phi : \mathcal{X} \to \mathcal{V}$, $x \mapsto \phi(x)$, where $\mathcal{V}$ is a Hilbert space. Then the kernel is defined as $h(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{V}}$, for any $x, x' \in \mathcal{X}$. The space $\mathcal{V}$ is known as the *feature space* and the mapping ϕ the *feature map*. In many mathematical models involving feature space mappings, elucidation of the feature map and feature space is not necessary, and thus computation is made simpler by the use of kernels (known as the *kernel trick*—Hofmann et al., 2008). Note that kernels defined in this manner are positive definite, while in this thesis, we opt for a more general definition allowing kernels to not necessarily be positive. The relevance of this generality will be appreciated when we discuss reproducing kernel Kreĭn spaces in Section 2.3.

Introducing the following definition of the *kernel matrix* (also known as the *Gram matrix*) is useful at this point.

Definition 2.21 (Kernel matrix)**.** Let $\{x_1, \dots, x_n\}$ be a sample of points, where each $x_i \in \mathcal{X}$, and h a kernel over $\mathcal{X}$. Define the *kernel matrix* $\mathbf{H}$ for h as the $n \times n$ matrix with (i, j) entries equal to $h(x_i, x_j)$.

Obviously, $\mathbf{H}$ is a positive definite matrix if the kernel that defines it is positive definite: $\mathbf{a}^\top \mathbf{H} \mathbf{a} = \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j h(x_i, x_j) \geq 0$ for any choice of $a_1, \dots, a_n \in \mathbb{R}$ and $x_1, \dots, x_n \in \mathcal{X}$.

So far, we have seen that reproducing kernels of a RKHS are positive-definite functions, and that RKHSs are Hilbert spaces with continuous evaluation functionals, but one might wonder what exactly the relationship between a reproducing kernel and a RKHS is. We assert the following:

- For every RKHS $\mathcal{F}$ of functions over a set $\mathcal{X}$, there corresponds a unique, positive-definite reproducing kernel $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, and vice-versa. That is, a Hilbert space is a RKHS if it possesses a unique, reproducing kernel.
- For every positive-definite function $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, there corresponds a unique RKHS $\mathcal{F}$ that has h as its reproducing kernel.

Pictorially, the following relationships are established:

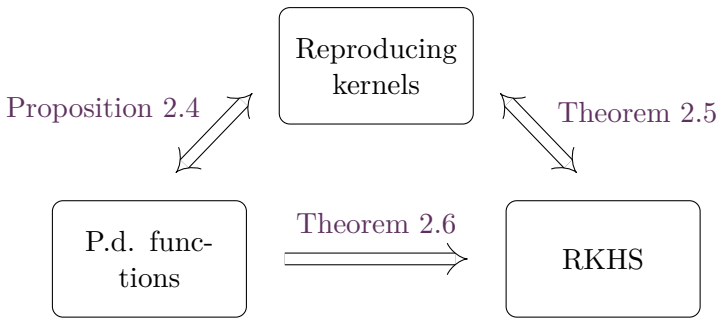

Figure 2.2: Relationships between positive definite functions, reproducing kernels, and RKHSs.

In essence, the notion of positive-definite functions and reproducing kernels of RKHSs are equivalent, and that there is a bijection between the set of positive-definite kernels and the set of RKHSs. The rest of this section is a consideration of these assertions, addressed by the two theorems that follow.

thm:rkhsunique

Theorem 2.5 (RKHS uniqueness)**.** *Let $\mathcal{F}$ be a Hilbert space of functions over $\mathcal{X}$. $\mathcal{F}$ is a RKHS if and only if $\mathcal{F}$ has a reproducing kernel $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, and that h is unique to $\mathcal{F}$.*

Proof. First we tackle existence, i.e. we prove that $\mathcal{F}$ is a RKHS if and only if $\mathcal{F}$ has a reproducing kernel. Suppose $\mathcal{F}$ is a Hilbert space of functions, and $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a reproducing kernel for $\mathcal{F}$. Then, choosing $\delta = \epsilon / \|h(\cdot, x)\|_{\mathcal{F}}$, for any $f \in \mathcal{F}$ such that $\|f - g\|_{\mathcal{F}} < \delta$, we have

$$\begin{aligned}
|\delta_x(f) - \delta_x(g)| &= |(f - g)(x)| \\
&= |\langle f - g, h(\cdot, x)\rangle_{\mathcal{F}}| \quad \text{(reproducing property)} \\
&\leq \|h(\cdot, x)\|_{\mathcal{F}} \, \|f - g\|_{\mathcal{F}} \quad \text{(Cauchy-Schwarz)} \\
&= \epsilon.
\end{aligned}$$

Thus, the evaluation functional is (uniformly) continuous on $\mathcal{F}$, and by definition, $\mathcal{F}$ is a RKHS. Now suppose that $\mathcal{F}$ is a RKHS, and h is a kernel function over $\mathcal{X} \times \mathcal{X}$. The reproducing property of h is had by following the argument preceding Definition 2.20.

As for uniqueness, assume that the RKHS $\mathcal{F}$ has two reproducing kernels h_1 and h_2. Then, $\forall f \in \mathcal{F}$ and $\forall x \in \mathcal{X}$,

$$\langle f, h_1(\cdot, x) - h_2(\cdot, x)\rangle_{\mathcal{F}} = f(x) - f(x) = 0.$$

In particular, if we take $f = h_1(\cdot, x) - h_2(\cdot, x)$, we obtain $\|h_1(\cdot, x) - h_2(\cdot, x)\|_{\mathcal{F}}^2 = 0$. Thus, $h_1 = h_2$. ■

thm:moorea

Theorem 2.6 (Moore-Aronszajn)**.** *If $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a positive-definite function then there exists a unique RKHS whose reproducing kernel is h.*

Sketch proof. Most of the details here have been omitted, except for the parts which we feel are revealing as to the properties of a RKHS. For a complete proof, see Gu (2013, Theorem 2.3). Start with the linear space

$$\mathcal{F}_0 = \left\{ f_n : \mathcal{X} \to \mathbb{R} \,\middle|\, f_n = \sum_{i=1}^n w_i h(\cdot, x_i), x_i \in \mathcal{X}, w_i \in \mathbb{R}, n \in \mathbb{N} \right\}$$

and endow this linear space with the following inner product:

$$\left\langle \sum_{i=1}^n w_i h(\cdot, x_i), \sum_{j=1}^m w_j' h(\cdot, x_j') \right\rangle_{\mathcal{F}_0} = \sum_{i=1}^n \sum_{j=1}^m w_i w_j' h(x_i, x_j').$$

It may be shown that this is indeed a valid inner product satisfying the conditions laid in Definition 2.1. At this point, the reproducing property is already had:

$$\begin{aligned} \langle f_n, h(\cdot, x)\rangle_{\mathcal{F}_0} &= \left\langle \sum_{i=1}^n w_i h(\cdot, x_i), h(\cdot, x) \right\rangle_{\mathcal{F}_0} \\ &= \sum_{i=1}^n w_i h(x, x_i) \\ &= f_n(x), \end{aligned}$$

for any $f_n \in \mathcal{F}_0$.

Let $\mathcal{F}$ be the completion of $\mathcal{F}_0$ with respect to this inner product. In other words, define $\mathcal{F}$ to be the set of functions $f : \mathcal{X} \to \mathbb{R}$ for which there exists a Cauchy sequence $\{f_n\}_{n=1}^\infty$ in $\mathcal{F}_0$ converging pointwise to $f \in \mathcal{F}$. The inner product for $\mathcal{F}$ is defined to be

$$\langle f, f'\rangle_{\mathcal{F}} = \lim_{n\to\infty} \langle f_n, f_n'\rangle_{\mathcal{F}_0}.$$

The sequence $\{\langle f_n, f_n' \rangle_{\mathcal{F}_0}\}_{n=1}^{\infty}$ is convergent and does not depend on the sequence chosen, but only on the limits f and f' (Berlinet and Thomas-Agnan, 2011, Lemma 5). We may check that this indeeds defines a valid inner product. The reproducing property carries over to the completion:

$$\begin{aligned}\langle f, h(\cdot, x)\rangle_{\mathcal{F}} &= \lim_{n\to\infty} \langle f_n, h(\cdot, x)\rangle_{\mathcal{F}_0} \\ &= \lim_{n\to\infty} f_n(x) \\ &= f(x).\end{aligned}$$

To prove uniqueness, let $\mathcal{G}$ be another RKHS with reproducing kernel h. $\mathcal{F}$ has to be a closed subspace of $\mathcal{G}$, since $h(\cdot, x) \in \mathcal{G}$ for all $x \in \mathcal{X}$, and because $\mathcal{G}$ is complete and contains $\mathcal{F}_0$ and hence its completion. Using the orthogonal decomposition theorem, we have $\mathcal{G} = \mathcal{F} \oplus \mathcal{F}^{\perp}$, i.e. any $g \in \mathcal{G}$ can be decomposed as $g = f + f^c$, $f \in \mathcal{F}$ and $f^c \in \mathcal{F}^{\perp}$. For each element $g \in \mathcal{G}$ we have that, for all $x \in \mathcal{X}$,

$$\begin{aligned}g(x) &= \langle g, h(\cdot, x)\rangle_{\mathcal{G}} \\ &= \big\langle f + f^c, h(\cdot, x)\big\rangle_{\mathcal{G}} \\ &= \big\langle f, h(\cdot, x)\big\rangle_{\mathcal{G}} + \cancelto{0}{\big\langle f^c, h(\cdot, x)\big\rangle_{\mathcal{G}}} \\ &= f(x)\end{aligned}$$

so therefore $g \in \mathcal{F}$ too. It must be that $\mathcal{F} \equiv \mathcal{G}$. ■

A consequence of the above proof is that we can show that any function f in a RKHS $\mathcal{F}$ with kernel h can be written in the form $f(x) = \sum_{i=1}^{n} h(x, x_i) w_i$, with some $(w_1, \dots, w_n) \in \mathbb{R}^n$, $n \in \mathbb{N}$. More precisely, $\mathcal{F}$ is the completion of the space $\mathcal{G} = \operatorname{span}\{h(\cdot, x) \,|\, x \in \mathcal{X}\}$ endowed with the inner product as stated in Section 2.2.

2.3 Reproducing kernel Kreĭn space theory

sec:rkkstheory

In this section, we review elementary Kreĭn and reproducing kernel Kreĭn space theory, and comment on the similarity and differences between it and RKHSs. Kreĭn spaces are linear spaces endowed with a Hilbertian topology, characterised by an inner product which is non-positive.

Definition 2.22 (Negative and indefinite inner products)**.** Let $\langle \cdot, \cdot \rangle_{\mathcal{F}}$ be an inner product of a vector space $\mathcal{F}$, as per Definition 2.1. An inner product is said to be *negative-definite* if for all $f \in \mathcal{F}$, $\langle f, f \rangle_{\mathcal{F}} \leq 0$. It is *indefinite* if it is neither positive- nor negative-definite.

def:krein

Definition 2.23 (Kreĭn space)**.** An inner product space $\big(\mathcal{F}, \langle\cdot,\cdot\rangle_{\mathcal{F}}\big)$ is a *Krein space* if there exists two Hilbert spaces $\big(\mathcal{F}_+, \langle\cdot,\cdot\rangle_{\mathcal{F}_+}\big)$ and $\big(\mathcal{F}_-, \langle\cdot,\cdot\rangle_{\mathcal{F}_-}\big)$ spanning $\mathcal{F}$ such that

- All $f \in \mathcal{F}$ can be decomposed into $f = f_+ + f_-$, where $f_+ \in \mathcal{F}_+$ and $f_- \in \mathcal{F}_-$.
- This decomposition is orthogonal, i.e. $\mathcal{F}_+ \cup \mathcal{F}_- = \{0\}$, and $\langle f_+, f_-\rangle_{\mathcal{F}} = 0$ for all $f_+ \in \mathcal{F}_+$ and $f_- \in \mathcal{F}_-$, with the inner product on $\mathcal{F}$ defined below.
- $\forall f, f' \in \mathcal{F}$, $\langle f, f'\rangle_{\mathcal{F}} = \langle f_+, f'_+\rangle_{\mathcal{F}_+} - \langle f_-, f'_-\rangle_{\mathcal{F}_-}$.

Remark 2.4. Any Hilbert space is also a Kreĭn space, which is seen by taking $\mathcal{F}_- = \{0\}$ in the above Definition 2.23.

Let P be the projection of the Kreĭn space $\mathcal{F}$ onto $\mathcal{F}_+$, and $Q = I - P$ the projection onto $\mathcal{F}_-$, where I is the identity map. These are caleld the *fundamental projections* of $\mathcal{F}$. We shall refer to $\mathcal{F}_+$ as the *positive subspace*, and $\mathcal{F}_-$ as the *negative subspace*. These monikers stem from the fact that for all $f, f' \in \mathcal{F}$, $\langle Pf, Pf'\rangle_{\mathcal{F}_+} \geq 0$ while $\langle Qf, Qf'\rangle_{\mathcal{F}_-} \leq 0$. We introduce the notation $\ominus$ to refer to the Kreĭn space decomposition: $\mathcal{F} = \mathcal{F}_+ \ominus \mathcal{F}_-$. There is then a notion of an *associated Hilbert space*.

Definition 2.24 (Associated Hilbert space)**.** Let $\mathcal{F}$ be a Kreĭn space with decomposition into Hilbert spaces $\mathcal{F}_+$ and $\mathcal{F}_-$. Denote by $\mathcal{F}_{\mathcal{H}}$ the associated Hilbert space defined by $\mathcal{F}_{\mathcal{H}} = \mathcal{F}_+ \oplus \mathcal{F}_-$, with inner product

$$\langle f, f'\rangle_{\mathcal{F}_{\mathcal{H}}} = \langle f_+, f'_+\rangle_{\mathcal{F}_+} + \langle f_-, f'_-\rangle_{\mathcal{F}_-},$$

for all $f, f' \in \mathcal{F}$.

The associated Hilbert space can be found via the linear operator $J = P - Q$ called the *fundamental symmetry*. That is, a Kreĭn space $\mathcal{F}$ can be turned into its associated Hilbert space by using the positive-definite inner product of the associated Hilbert space as $\langle f, f'\rangle_{\mathcal{F}_{\mathcal{H}}} = \langle f, Jf'\rangle_{\mathcal{F}}$, for all $f, f' \in \mathcal{F}$. The converse is true too: starting from a Hilbert space $\mathcal{F}_{\mathcal{H}}$ and an operator J, the vector space endowed with the inner product $\langle f, f'\rangle_{\mathcal{F}} = \langle f, Jf'\rangle_{\mathcal{F}_{\mathcal{H}}}$, for all $f, f' \in \mathcal{F}$, is a Kreĭn space.

We realise that for a Kreĭn space $\mathcal{F}$, $|\langle f, f\rangle_{\mathcal{F}}| \leq \|f\|^2_{\mathcal{F}_{\mathcal{H}}}|$ for all $f \in \mathcal{F}$ (we say that $\mathcal{F}_{\mathcal{H}}$ majorises the $\mathcal{F}$), and in fact it is the smallest Hilbert space to do so. The strong topology on $\mathcal{F}$ is defined to be the topology arising from the norm of $\mathcal{F}_{\mathcal{H}}$, and this does not depend on the decomposition chosen (Ong et al., 2004). Now, we define a RKKS.

Definition 2.25 (Reproducing kernel Kreĭn space)**.** A Krein space $\mathcal{F}$ of real-valued functions $f : \mathcal{X} \to \mathbb{R}$ on a non-empty set $\mathcal{X}$ is called a *reproducing kernel Krein space* if the evaluation functional $\delta_x : f \mapsto f(x)$ is continuous on $\mathcal{F}$, $\forall x \in \mathcal{X}$, endowed with its strong topology (i.e. the topology of its associated Hilbert space $\mathcal{F}_{\mathcal{H}}$).

One might wonder whether the uniqueness theorem (Theorem 2.5) holds for RKKS. Indeed, for every RKKS $\mathcal{F}$ of functions over a set $\mathcal{X}$, there corresponds a unique reproducing kernel $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$.

Lemma 2.7 (Uniqueness of kernel for RKKS)**.** *Let $\mathcal{F}$ be a RKKS of functions over a set $\mathcal{X}$, with $\mathcal{F} = \mathcal{F}_+ \ominus \mathcal{F}_-$. Then, $\mathcal{F}_+$ and $\mathcal{F}_-$ are both RKHS with kernel h_+ and h_-, and the kernel $h = h_+ - h_-$ is a unique, symmetric, reproducing kernel for $\mathcal{F}$.*

Proof. Since $\mathcal{F}$ is a RKKS, evaluation functionals are continuous on $\mathcal{F}$ with respect to topology of the associated Hilbert space $\mathcal{F}_\mathcal{H} = \mathcal{F}_+ \oplus \mathcal{F}_-$. Therefore, $\mathcal{F}_\mathcal{H}$ is a RKHS, and so too are $\mathcal{F}_+$ and $\mathcal{F}_-$ with respective kernels h_+ and h_-.

Furthermore, $h(\cdot, x) \in \mathcal{F}$ since $h_+(\cdot, x) \in \mathcal{F}_+$ and $h_-(\cdot, x) \in \mathcal{F}_-$ for some $x \in \mathcal{X}$. Then, for any $f \in \mathcal{F}$,

$$\begin{aligned}
\langle f, h(\cdot, x)\rangle_\mathcal{F} &= \langle f, h_+(\cdot, x)\rangle_\mathcal{F} - \langle f, h_-(\cdot, x)\rangle_\mathcal{F} \\
&= \langle f_+, h_+(\cdot, x)\rangle_{\mathcal{F}_+} - \cancelto{0}{\langle f_-, h_+(\cdot, x)\rangle_{\mathcal{F}_-}} \\
&\quad - \cancelto{0}{\langle f_+, h_-(\cdot, x)\rangle_{\mathcal{F}_+}} + \langle f_-, h_-(\cdot, x)\rangle_{\mathcal{F}_-} \\
&= f_+(x) + f_-(x) \\
&= f(x)
\end{aligned}$$

The last two lines are achieved by linearity of evaluation functionals ($\delta_x(f_+) + \delta_x(f_-) = \delta_x(f_+ + f_-)$), and the fact that $f = f_+ + f_-$ (by the Kreĭn space decomposition). We have that $h = h_+ - h_-$ is a reproducing kernel for $\mathcal{F}$. Uniqueness follows as a consequence of the non-degeneracy condition of the respective inner products for $\mathcal{F}_+$ and $\mathcal{F}_-$. ■

Remark 2.5. Unlike reproducing kernels of RKHSs, reproducing kernels of RKKSs may not be positive definite.

The analogue of the Moore-Aronszajn theorem holds partially for RKKS, up to uniqueness. That is, there is *at least* one associated RKKS with kernel $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ if and only if h can be decomposed as the difference between two positive kernels h_+ and h_- over $\mathcal{X}$, i.e. $h = h_+ - h_-$. The proof of this statement is rather involved, so is omitted in the interest of maintaining coherence to the discussion at hand. This subject has been studied by various authors; one may refer to works by Alpay (1991, Theorem 2 & Example in Section 4), and Mary (2003, Theorem 2.28).

The take-away message as we close this section is that there is no bijection, but a surjection, between the set of RKKS and the set of bivariate, symmetric functions over $\mathcal{X} \times \mathcal{X}$. In any case, Hilbertian topology applies to Kreĭn spaces via the associated Hilbert space, and in particular, RKKS provide a functional space for which evaluation

functionals are continuous. The motivation for the use of Kreĭn spaces will become clear when constructing function spaces out of (scaled) building block RKHS later in Section 2.5.

2.4 RKHS building blocks

sec:rkhsbuild

This section describes what we refer to as the "building block" RKHSs of functions. In the context of regression modelling using I-priors, we may assume that the regression function lies in any one of these single RKHSs, although it may be more appropriate to consider function spaces built upon these RKHSs for more complex models. Construction of new function spaces from these building block RKHSs will be discussed in the next section.

2.4.1 The RKHS of constant functions

The vector space of constant functions $\mathcal{F}$ over a set $\mathcal{X}$ contains the functions $f : \mathcal{X} \to \mathbb{R}$ such that $f(x) = c_f \in \mathbb{R}$, $\forall x \in \mathcal{X}$. These functions would be useful to model an overall average, i.e. an "intercept effect". The space $\mathcal{F}$ can be equipped with a norm to form an RKHS, as shown in the following proposition.

Proposition 2.8 (RKHS of constant functions)**.** *The space $\mathcal{F}$ as described above endowed with the norm $\|f\|_{\mathcal{F}} = |c_f|$ forms an RKHS with the reproducing kernel $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ as defined, rather simply, by*

$$h(x, x') = 1,$$

known as the constant kernel.

Proof. If $\mathcal{F}$ is an RKHS with kernel h as described, then $\mathcal{F}$ is spanned by the functions $h(\cdot, x) = 1$, so it is clear that $\mathcal{F}$ consists of constant functions over $\mathcal{X}$. On the other hand, if the space $\mathcal{F}$ is equipped with the inner product $\langle f, f' \rangle_{\mathcal{F}} = c_f c_{f'}$, then the reproducing property follows, since $\langle f, h(\cdot, x) \rangle_{\mathcal{F}} = c_f = f(x)$. Hence, $\|f\|_{\mathcal{F}} = \sqrt{\langle f, f \rangle_{\mathcal{F}}} = |c_f|$. ■

Remark 2.6. In I-prior modelling, it is simpler to consider the intercept of a regression model as a parameter to be estimated, rather than a separate function within an RKHS of constant functions for which its posterior is to be estimated. See Section 4.2.1 in Chapter 4 for further details.

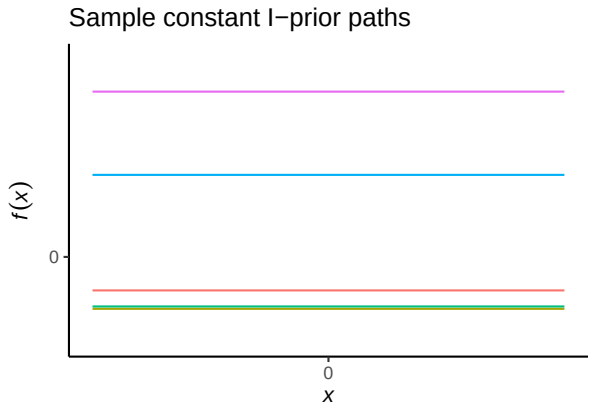

Figure 2.3: Sample I-prior paths from the RKHS of constant functions.

2.4.2 The canonical (linear) RKHS

Consider a function space $\mathcal{F}$ over $\mathcal{X}$ which consists of functions of the form $f_\beta : \mathcal{X} \to \mathbb{R}$, $f_\beta : x \mapsto \langle x, \beta \rangle_\mathcal{X}$ for some $\beta \in \mathbb{R}$. Suppose that $\mathcal{X} \equiv \mathbb{R}^p$, then $\mathcal{F}$ consists of the linear functions $f_\beta(x) = x^\top \beta$. More generally, if $\mathcal{X}$ is a Hilbert space, then its continuous dual consists of elements of the form $f_\beta = \langle \cdot, \beta \rangle_\mathcal{X}$ by the Riesz representation theorem. We can show that the continuous dual space of $\mathcal{X}$ is a RKHS which consists of these linear functions.

Proposition 2.9 (The canonical RKHS)**.** *The continuous dual space a Hilbert space $\mathcal{X}$, denoted by $\mathcal{X}^*$, is a RKHS of linear functions over $\mathcal{X}$ of the form $\langle \cdot, \beta \rangle_\mathcal{X}$, $\beta \in \mathcal{X}$. Its reproducing kernel $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is defined by*

$$h(x, x') = \langle x, x' \rangle_\mathcal{X}.$$

Proof. Define $f_\beta := \langle \cdot, \beta \rangle_\mathcal{X}$ for some $\beta \in \mathcal{X}$. Clearly this is linear and continuous, so $f_\beta \in \mathcal{X}^*$, and so $\mathcal{X}^*$ is a Hilbert space containing functions $f : \mathcal{X} \to \mathbb{R}$ of the form $f_\beta(x) = \langle x, \beta \rangle_\mathcal{X}$. By the Riesz representation theorem, every element of $\mathcal{X}^*$ has the form f_β. It also gives us a natural isometric isomorphism such that the following is true:

$$\langle \beta, \beta' \rangle_\mathcal{X} = \langle f_\beta, f_{\beta'} \rangle_{\mathcal{X}^*}.$$

Hence, for any $f_\beta \in \mathcal{X}^*$,

$$\begin{aligned} f_\beta(x) &= \langle x, \beta \rangle_\mathcal{X} \\ &= \langle f_x, f_\beta \rangle_{\mathcal{X}^*} \\ &= \big\langle \langle \cdot, x \rangle_\mathcal{X}, f_\beta \big\rangle_{\mathcal{X}^*}. \end{aligned}$$

Thus, $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ as defined by $h(x, x') = \langle x, x' \rangle_\mathcal{X}$ is the reproducing kernel of $\mathcal{X}^*$. ∎

In many other literature, the kernel $h(x, x') = \langle x, x' \rangle_{\mathcal{X}}$ is also known as the *linear kernel*. The use of the term 'canonical' is fitting not just due to the relation between a Hilbert space and its continuous dual space. Let $\phi : \mathcal{X} \to \mathcal{V}$ be the feature map from the space of covariates (inputs) to some feature space $\mathcal{V}$. Suppose both $\mathcal{X}$ and $\mathcal{V}$ are Hilbert spaces, then a kernel is defined as

$$h(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{V}}.$$

Taking the feature map to be $\phi(x) = \langle \cdot, x \rangle_{\mathcal{X}}$, we can prove the reproducing property to obtain $h(x, x') = \langle x, x' \rangle_{\mathcal{X}}$, which implies $\phi(x) = h(\cdot, x)$, and thus ϕ is the *canonical feature map* (Steinwart and Christmann, 2008, Lemma 4.19).

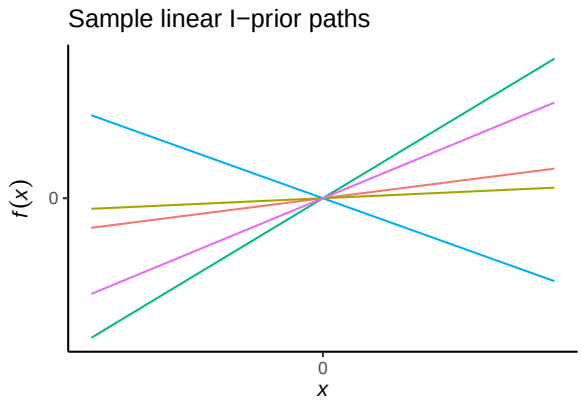

Figure 2.4: Sample paths from the canonical RKHS.

The origin of a Hilbert space may be arbitrary, in which case a centring may be appropriate. We define the centred canonical RKHS as follows.

Definition 2.26 (Centred canonical RKHS)**.** Let $\mathcal{X}$ be a Hilbert space, P be a probability measure over $\mathcal{X}$, and $\mu \in \mathcal{X}$ be the mean of a random element $X \in \mathcal{X}$. Define $(\mathcal{X} - \mu)'$, the continuous dual space of $\mathcal{X} - \mu$, to be the *centred canonical RKHS*. $(\mathcal{X} - \mu)'$ consists of the centred linear functions $f_\beta(x) = \langle x - \mu, \beta \rangle_{\mathcal{X}}$, for $\beta \in \mathcal{X}$, such that $\mathrm{E}\, f_\beta(X) = 0$. The reproducing kernel of $(\mathcal{X} - \mu)'$ is

$$h(x, x') = \langle x - \mu, x' - \mu \rangle_{\mathcal{X}}.$$

That the centred canonical RKHS consists of zero mean function, $\mathrm{E}[f_\beta(X)] = 0$, consider the following argument:

$$\begin{aligned} \mathrm{E}[f_\beta(X)] &= \mathrm{E}\langle X - \mu, \beta \rangle_{\mathcal{X}} \\ &= \mathrm{E}\langle X, \beta \rangle_{\mathcal{X}} - \langle \mu, \beta \rangle_{\mathcal{X}}, \end{aligned}$$

and since $\mathrm{E}\langle X, \beta \rangle_{\mathcal{X}} = \langle \mu, \beta \rangle_{\mathcal{X}}$ for any $\beta \in \mathcal{X}$, the results follows.

rem:empircent

Remark 2.7. In practice, the probability measure P over $\mathcal{X}$ is unknown, so we find it useful to use the empirical distribution over $\mathcal{X}$ instead, so that $\mathcal{X}$ is centred by the sample mean $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

2.4.3 The fractional Brownian motion RKHS

Brownian motion, which also goes by the name Wiener process, has been an inquisitive subject in the mathematical sciences, and here, we describe a function space motivated by a generalised version of Brownian motion paths.

Suppose $B_\gamma(t)$ is a continuous-time Gaussian process on $[0, T]$, i.e. for any finite set of indices $t_1, \dots, t_k$, where each $t_j \in [0, T]$, $\big(B_\gamma(t_1), \dots, B_\gamma(t_k)\big)$ is a multivariate normal random variable. $B_\gamma(t)$ is said to be a *fractional Brownian motion* (fBm) if $\mathrm{E}[B_\gamma(t)] = 0$ for all $t \in [0, T]$ and

$$\mathrm{Cov}\left[B_\gamma(t), B_\gamma(s)\right] = \frac{1}{2}\big(|t|^{2\gamma} + |s|^{2\gamma} - |t-s|^{2\gamma}\big) \qquad \forall t, s \in [0, T],$$

where $\gamma \in (0, 1)$ is called the *Hurst index*, *Hurst parameter* or even *Hurst coefficient*. Introduced by Mandelbrot and Van Ness (1968), fBms are a generalisation of Brownian motion. The Hurst parameter plays two roles: 1) it describes the raggedness of the resultant motion, with higher values leading to smoother motion; and 2) it determines the type of process the fBm is, as past increments of $B_\gamma(t)$ are weighted by $(t-s)^{\gamma-1/2}$. When $\gamma = 1/2$ exactly, the fBm is a standard Brownian motion and its increments are independent; when $\gamma > 1/2$ (resp. $\gamma < 1/2$) its increments are positively (resp. negatively) correlated.

Now, let $\mathcal{X}$ be a Hilbert space. Schoenberg (1937, Theorem 3) has shown that, for $0 < \gamma \le 1$, there exists a Hilbert space $\mathcal{V}$ and a function $\phi_\gamma : \mathcal{X} \to \mathcal{V}$ such that $\forall x, x' \in \mathcal{X}$,

$$\big\|\phi_\gamma(x) - \phi_\gamma(x')\big\|_{\mathcal{V}} = \|x - x'\|_{\mathcal{X}}^{\gamma}.$$

Using the polarisation identity, we find that the kernel of the RKHS with feature space $\mathcal{V}$ and feature map ϕ_γ defines a kernel function $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ identical to the fBm covariance kernel.

def:fbmrkhs

Definition 2.27 (Fractional Brownian motion RKHS)**.** The fractional Brownian motion (fBm) RKHS $\mathcal{F}$ is the space of functions on the Hilbert space $\mathcal{X}$ possessing the reproducing kernel $h_\gamma : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defined by

$$h_\gamma(x, x') = \big\langle \phi_\gamma(x), \phi_\gamma(x') \big\rangle_{\mathcal{V}} = \frac{1}{2}\big(\|x\|_{\mathcal{X}}^{2\gamma} + \|x'\|_{\mathcal{X}}^{2\gamma} - \|x - x'\|_{\mathcal{X}}^{2\gamma}\big),$$

which depends on the Hurst coefficient $\gamma \in (0, 1)$. We shall reference this space as the fBm-γ RKHS.

Remark 2.8. When $\gamma = 1$, by the polarisation identity we get $h(x, x') = \langle x, x' \rangle_{\mathcal{X}}$, which is the (reproducing) kernel of the canonical RKHS.

From its construction, it is clear that the fBm kernel is positive definite, and thus defines an RKHS. That the fBm RKHS describes a space of functions is proved in Cohen (2002), who studied this space in depth. It is also noted in the collection of examples of Berlinet and Thomas-Agnan (2011, pp.71 & 319).

The Hurst coefficient γ controls the "smoothness" of the functions in the RKHS. We can talk about smoothness in the context of Hölder continuity of functions.

Definition 2.28 (Hölder condition)**.** A function f over a set $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$ is said to be *Hölder continuous* of order $0 < \gamma \leq 1$ if there exists a $C > 0$ such that $\forall x, x' \in \mathcal{X}$,

$$|f(x) - f(x')| \leq C\|x - x'\|^{\gamma}.$$

Functions in the Hölder space $\mathrm{C}^{k,\gamma}(\mathcal{X})$, where $k \geq 0$ is an integer, consists of those functions over $\mathcal{X}$ having continuous derivatives up to order k and such that the kth partial derivatives are Hölder continuous of order γ. Unlike realisations of actual fBm paths with Hurst index γ, which are well-known to be almost surely Hölder continuous of order less than γ (Embrechts and Maejima, 2002, Theorem 4.1.1), functions in its namesake RKHS are strictly smoother.

Proposition 2.10. *The fBm-γ RKHS $\mathcal{F}$ of functions over $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$ are Hölder continuous of order γ.*

Proof. For some $f \in \mathcal{F}$ we have $f(x) = \langle f, h(\cdot, x) \rangle_{\mathcal{F}}$ by the reproducing property of the kernel h of $\mathcal{F}$. It follows from the Cauchy-Schwarz inequality that for any $x, x' \in \mathcal{X}$,

$$\begin{aligned} |f(x) - f(x')| &= |\langle f, h(\cdot, x) - h(\cdot, x') \rangle_{\mathcal{F}}| \\ &\leq \|f\|_{\mathcal{F}} \left\| h(\cdot, x) - h(\cdot, x') \right\|_{\mathcal{F}} \\ &= \|f\|_{\mathcal{F}} \|x - x'\|_{\mathcal{X}}^{\gamma}, \end{aligned}$$

since

$$\begin{aligned} \left\| h(\cdot, x) - h(\cdot, x') \right\|_{\mathcal{F}}^2 &= \left\| h(\cdot, x) \right\|_{\mathcal{F}}^2 + \left\| h(\cdot, x') \right\|_{\mathcal{F}}^2 - 2\langle h(\cdot, x), h(\cdot, x') \rangle_{\mathcal{F}} \\ &= h(x, x) + h(x', x') - 2h(x, x') \\ &= \|x - x'\|_{\mathcal{X}}^{2\gamma}, \end{aligned}$$

and thus proving the proposition. ■

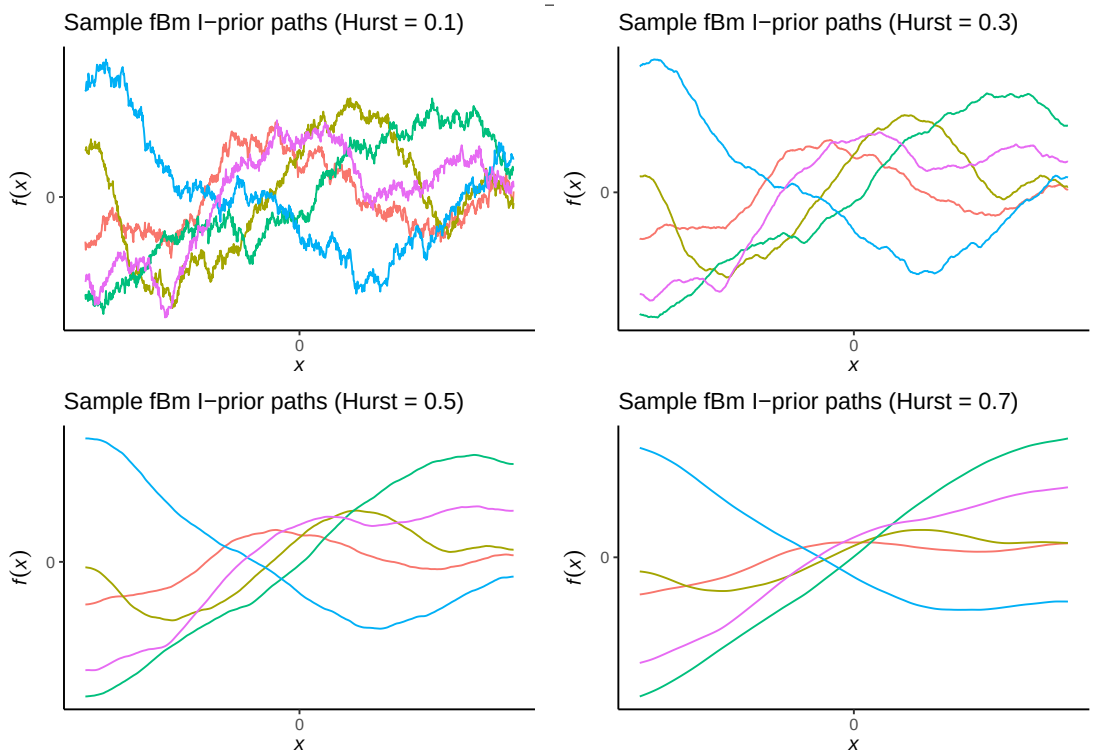

Figure 2.5: Sample I-prior paths from the fBm RKHS with varying Hurst coefficients. Note that the fBm-γ RKHS contains functions that are rougher than these I-prior paths.

The fBm-γ RKHS is spanned by the functions $h(\cdot, x)$, which means that $f(0) = 0$ for all $f \in \mathcal{F}$, which may be undesirable. We define the centred fBm RKHS as follows.

Definition 2.29 (Centred fBm RKHS)**.** Let $\mathcal{X}$ be a Hilbert space, P be a probability measure over $\mathcal{X}$, and $\mu \in \mathcal{X}$ be the mean with respect to this probability measure. The kernel $\bar{h}_\gamma : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defined by

$$\bar{h}_\gamma(x, x') = \frac{1}{2} \mathrm{E}\left[\|x - X\|_{\mathcal{X}}^{2\gamma} + \|x' - X'\|_{\mathcal{X}}^{2\gamma} - \|x - x'\|_{\mathcal{X}}^{2\gamma} - \|X - X'\|_{\mathcal{X}}^{2\gamma}\right]$$

is the reproducing kernel of the *centred* fBm-γ RKHS, which consists of functions f in the fBm-γ RKHS such that $\mathrm{E}[f(X)] = 0$. In the above definition, $X, X' \sim \mathrm{P}$ are two independent copies of a random vector $X \in \mathcal{X}$.

Remark 2.9. Again, when $\gamma = 1$, we get the reduction

$$\begin{aligned}
\bar{h}_{\gamma=1}(x, x') &= \frac{1}{2} \mathrm{E}\left[\|x - X\|_{\mathcal{X}}^{2} + \|x' - X'\|_{\mathcal{X}}^{2} - \|x - x'\|_{\mathcal{X}}^{2} - \|X - X'\|_{\mathcal{X}}^{2}\right] \\
&= \frac{1}{2} \mathrm{E}\left[\langle X, X\rangle_{\mathcal{X}} + \langle X', X'\rangle_{\mathcal{X}} + 2\langle x, x'\rangle_{\mathcal{X}} - 2\langle x, X\rangle_{\mathcal{X}} - 2\langle x', X'\rangle_{\mathcal{X}}\right] \\
&= \langle \mu, \mu\rangle_{\mathcal{X}} + \langle x, x'\rangle_{\mathcal{X}} - \langle x, \mu\rangle_{\mathcal{X}} - \langle \mu, x'\rangle_{\mathcal{X}} \\
&= \langle x - \mu, x' - \mu\rangle_{\mathcal{X}},
\end{aligned}$$

which is the (reproducing) kernel of the centred canonical RKHS.

Remark 2.10. For posterity, a general centring of any (positive-definite) kernel $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ can be achieved via

$$\bar{h}(x, x') = h(x, x') - \mathrm{E}[h(x, X')] - \mathrm{E}[h(X, x')] + \mathrm{E}[h(X, X')],$$

where expectations are taken for the random elements $X, X' \overset{\text{iid}}{\sim} \mathrm{P}$, a probability measure over $\mathcal{X}$. This centred kernel gives rise to the centred RKHS $\bar{\mathcal{F}}$ of centred functions $\mathrm{E}[f(X)]$, $f \in \bar{\mathcal{F}}$. As per Remark 2.7, the empirical distribution of P can be used to approximate the unknown, true P.

2.4.4 The squared exponential RKHS

The squared exponential (SE) kernel function is indeed known to be the default kernel used for Gaussian process regression in machine learning. It is a positive definite function, and hence defines an RKHS. The definition of the SE RKHS is as follows.

Definition 2.30 (Squared exponential RKHS)**.** The squared exponential (SE) RKHS $\mathcal{F}$ of functions over some set $\mathcal{X} \subseteq \mathbb{R}^p$ equipped with the 2-norm $\|\cdot\|_2$ is defined by the positive definite kernel $h_l : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$

$$h_l(x, x') = \exp\left(-\frac{\|x - x'\|_2^2}{2l^2}\right).$$

The real-valued parameter $l > 0$ is called the *lengthscale* parameter, and is a smoothing parameter for the functions in the RKHS.

It is known by many other names, including the Gaussian kernel, due to its semblance to the kernel of the Gaussian pdf. Especially in the machine learning literature, the term Gaussian radial basis functions (RBF) is used, and commonly the simpler parameterisation $\gamma = (2l^2)^{-1}$ is utilised. Duvenaud (2014) remarks that "exponentiated quadratic" is a more aptly descriptive name for this kernel.

Despite being used extensively for learning algorithms using kernels, an explicit study of the RKHS defined by the SE kernel was not done until recently by Steinwart, Hush, et al. (2006). In that work, the authors describe the nature of real-valued functions in the SE RKHS by considering a a real restriction on the SE RKHS of functions over complex values. Their derivation of an orthonormal basis of such an RKHS proved the SE kernel to be the reproducing kernel for the SE RKHS.

SE kernels are known to be "universal". That is, it satisfies the following definition of universal kernels due to Micchelli et al. (2006).

Definition 2.31 (Universal kernel)**.** Let $\mathrm{C}(\mathcal{X})$ be the space of all continuous, complex-valued functions $f : \mathcal{X} \to \mathbb{C}$ equipped with the maximum norm $\|\cdot\|_\infty$, and denote $\mathcal{K}(\mathcal{X})$

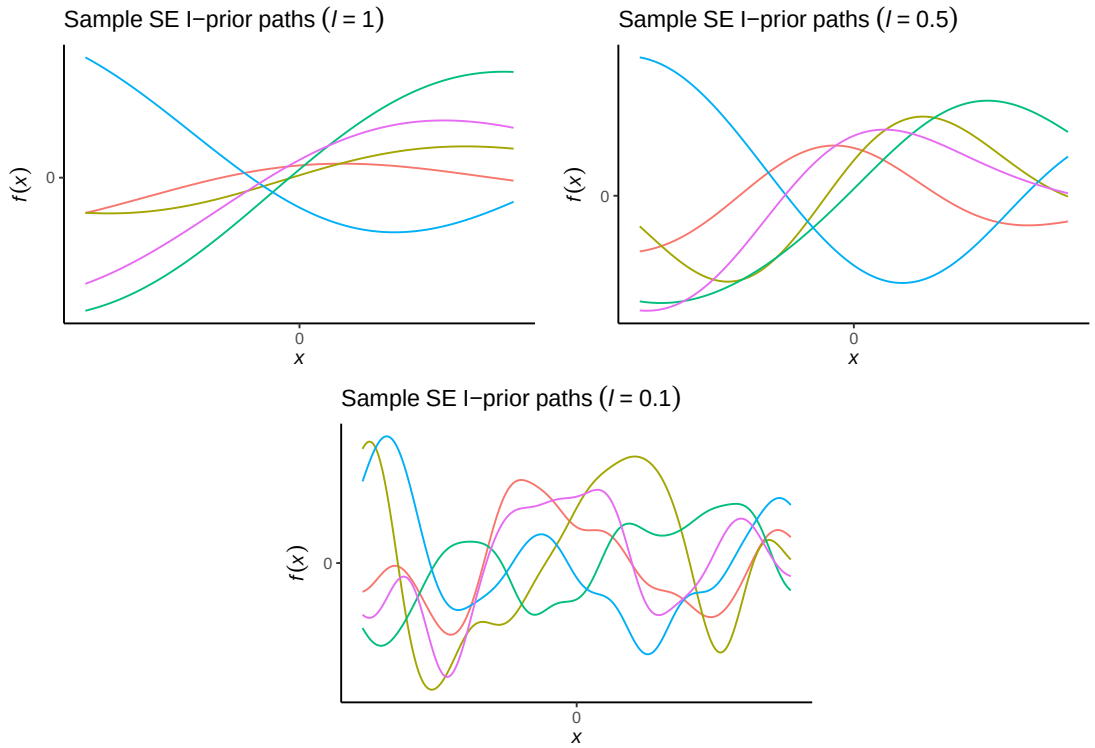

Figure 2.6: Sample paths from the SE RKHS with varying values for the lengthscale.

as the space of *kernel sections* $\overline{\text{span}}\{h(\cdot, x) | x \in \mathcal{X}\}$, where here, h is a complex-valued kernel function. A kernel h is said to be *universal* if given any compact subset $\mathcal{Z} \subset \mathcal{X}$, any positive number ϵ and any function $f \in \mathrm{C}(\mathcal{Z})$, there is a function $g \in \mathcal{K}(\mathcal{Z})$ such that $\|f - g\|_{\mathcal{Z}} \leq \epsilon$.

The consequence of this universal property vis-à-vis regression modelling is that any (continuous) regression function f may be approximated very well by a function $\hat{f}$ belonging to the SE RKHS, and these two functions can get arbitrarily close to each other in the max norm sense. This, together with the convenient computational advantages that the SE kernel brings (Raykar and Duraiswami, 2007), is a testament to the popularity of SE kernels.

In a similar manner to the two previous subsections, we may also derive the *centred* SE RKHS.

Definition 2.32 (Centred SE RKHS)**.** Let $\mathcal{X} \subseteq \mathbb{R}^p$ be equipped with the 2-norm $\|\cdot\|_2$, and let P denote the distribution over $\mathcal{X}$. Assuming integrability of $h(x, X)$, for any $x \in \mathcal{X}$ and a random element $X \in \mathcal{X}$, the *centred* squared exponential (SE) RKHS (with lengthscale l) of functions over $\mathcal{X}$ is defined by the positive definite kernel $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$

$$h(x, x') = e^{-\frac{\|x-x'\|_2^2}{2l^2}} - \mathrm{E}\left[e^{-\frac{\|x-X'\|_2^2}{2l^2}}\right] - \mathrm{E}\left[e^{-\frac{\|X-x'\|_2^2}{2l^2}}\right] + \mathrm{E}\left[e^{-\frac{\|X-X'\|_2^2}{2l^2}}\right],$$

where $X, X' \sim \mathrm{P}$ are two independent random elements of $\mathcal{X}$. This ensures that $\mathrm{E}[f(X)] = 0$ for any f in this RKHS.

2.4.5 The Pearson RKHS

In all of the previous RKHSs of functions, the domain $\mathcal{X}$ was taken to be some Euclidean space. The Pearson RKHS is a vector space of functions whose domain $\mathcal{X}$ is a finite set. Let P be a probability measure over the finite set $\mathcal{X}$. The Pearson RKHS is defined as follows.

def:pearson

Definition 2.33 (Pearson RKHS)**.** The *Pearson RKHS* is the RKHS of functions over a finite set $\mathcal{X}$ defined by the reproducing kernel

$$h(x, x') = \frac{\delta_{xx'}}{\mathrm{P}(X = x)} - 1,$$

where $X \sim \mathrm{P}$ and δ is the Kronecker delta.

The Pearson RKHS contains functions which are centred, and has the desirable property that the contribution of $[f(x)]^2$ to the squared norm of f is proportional to $\mathrm{P}(X = x)$.

Proposition 2.11. *Let $\mathcal{F}$ be the Pearson RKHS of functions over a finite set $\mathcal{X}$. Then,*

$$\mathcal{F} = \{f : \mathcal{X} \to \mathbb{R} \mid \mathrm{E}[f(X)] = 0\}$$

with

$$\|f\|_{\mathcal{F}}^2 = \mathrm{Var}[f(X)] = \sum_{x \in \mathcal{X}} \mathrm{P}(X = x)[f(x)]^2, \ \forall f \in \mathcal{F}.$$

Proof. Write $p_x = \mathrm{P}(X = x)$. The set of functions $\{h(\cdot, x) \mid x \in \mathcal{X}\}$ form a basis for $\mathcal{F}$, and thus each $f \in \mathcal{F}$ can be written as $f(x) = \sum_{x' \in \mathcal{X}} w_{x'} h(x, x')$ for some scalars $w_i \in \mathbb{R}$, $i \in \mathcal{X}$. But $\mathrm{E}[h(X, x')] = \mathrm{E}[\delta_{Xx'}]/p_{x'} - 1 = p_{x'}/p_{x'} - 1 = 0$, and thus $\mathrm{E}[f(X)] = 0$. Conversely, suppose $f : \mathcal{X} \to \mathbb{R}$ is such that $\mathrm{E}[f(X)] = 0$. Taking $w_x = p_x f(x)$, we see that

$$\begin{aligned}
\sum_{x' \in \mathcal{X}} w_{x'} h(x, x') &= \frac{w_x}{p_x} - \sum_{x' \in \mathcal{X}} w_{x'} \\
&= \frac{f(x) \cancel{p_x}}{\cancel{p_x}} - \cancel{\sum_{x' \in \mathcal{X}} p_{x'} f(x')}^{\mathrm{E}[f(X)] = 0} \qquad = f(x)
\end{aligned}$$

and thus $h(\cdot, x)$ spans $\mathcal{F}$ so $f \in \mathcal{F}$.

The second part is proved as follows. Noting that with the choice $w_x = p_x f(x)$ and due to the reproducing property of h for the RKHS $\mathcal{F}$, the squared norm is

$$\begin{aligned}
\langle f, f \rangle_{\mathcal{F}} &= \left\langle \sum_{x \in \mathcal{X}} w_x h(\cdot, x), \sum_{x' \in \mathcal{X}} w_{x'} h(\cdot, x') \right\rangle_{\mathcal{F}} \\
&= \sum_{x \in \mathcal{X}} \sum_{x' \in \mathcal{X}} w_x w_{x'} \left\langle h(\cdot, x), h(\cdot, x') \right\rangle_{\mathcal{F}} \\
&= \sum_{x \in \mathcal{X}} \sum_{x' \in \mathcal{X}} w_x w_{x'} h(x, x') \\
&= \sum_{x \in \mathcal{X}} w_x f(x) \\
&= \sum_{x \in \mathcal{X}} \mathrm{P}(X = x) [f(x)]^2,
\end{aligned}$$

which is also the variance of $f(X)$. ■

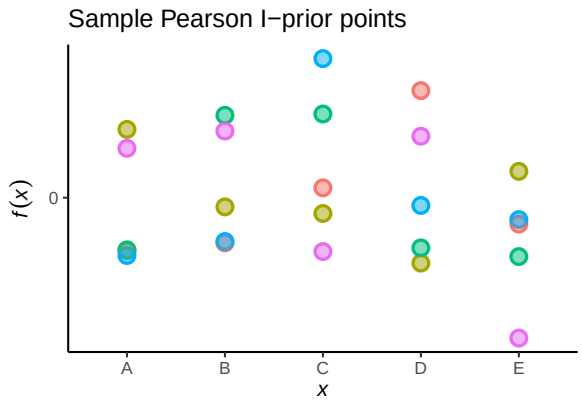

Figure 2.7: Sample I-prior "paths" from the Pearson RKHS. These are represented as points over a finite set. Similarly coloured points are from the same "path", and since they are zero-mean functions, they sum to zero.

2.5 Constructing RKKSs from existing RKHSs

sec:construc
trkks

The previous section outlined all of the basic RKHSs of functions that will form the building blocks when constructing more complex function spaces. We will see, at the outset, that sums of kernels are kernels and products of kernels are also kernels. This provides us a platform for constructing new function spaces from existing ones. To be more flexible in the specification of these new function spaces, we do not restrict ourselves to positive-definite kernels only, thereby necessitating us to use the theory of RKKSs.

2.5.1 Sums, products and scaling of RKHS

Sums of positive definite kernels are also positive definite kernels, and the product of positive definite kernel is a positive definite kernel. They each, in turn, are associated with a RKHS that is defined by the sum of kernels and product of kernels, respectively. The two lemmas below formalise these two facts.

thm:sumkernels

Lemma 2.12 (Sum of kernels)**.** *If h_1 and h_2 are positive-definite kernels on $\mathcal{X}_1$ and $\mathcal{X}_2$ respectively, then $h = h_1 + h_2$ is a positive-definite kernel on $\mathcal{X}_1 \times \mathcal{X}_2$. Moreover, denote $\mathcal{F}_1$ and $\mathcal{F}_2$ the RKHS defined by h_1 and h_2 respectively. Then $\mathcal{F} = \mathcal{F}_1 \oplus \mathcal{F}_2$ is a RKHS defined by $h = h_1 + h_2$, where*

$$\mathcal{F}_1 \oplus \mathcal{F}_2 = \{f : \mathcal{X}_1 \times \mathcal{X}_2 \to \mathbb{R} \,|\, f = f_1 + f_2, f_1 \in \mathcal{F}_1 \text{ and } f_2 \in \mathcal{F}_2\}.$$

For all $f \in \mathcal{F}$,

$$\|f\|_{\mathcal{F}}^2 = \min_{f_1+f_2=f} \left\{ \|f_1\|_{\mathcal{F}_1}^2 + \|f_2\|_{\mathcal{F}_2}^2 \right\}.$$

Proof. That h_1+h_2 is a positive-definite kernel should be obvious, as the sum of two positive definite functions is also positive definite. For a proof of the remaining statements, see Berlinet and Thomas-Agnan (2011, Theorem 5). ■

thm:prodkernels

Lemma 2.13 (Products of kernels)**.** *Let $\mathcal{F}_1$ and $\mathcal{F}_2$ be two RKHS of functions over $\mathcal{X}_1$ and $\mathcal{X}_2$, with respective reproducing kernels h_1 and h_2. Then, $h = h_1 h_2$ is a kernel on $\mathcal{X}_1 \times \mathcal{X}_2$. Moreover, the tensor product space $\mathcal{F}_1 \otimes \mathcal{F}_2$ is a RKHS with reproducing kernel h.*

Proof. Fix $n \in \mathbb{N}$, and let $\mathbf{H}_1$ and $\mathbf{H}_2$ be the kernel matrices for h_1 and h_2 respectively. Since these kernel matrices are symmetric and positive definite by virtue of h_1 and h_2 being symmetric and positive-definite functions, we can write $\mathbf{H}_1 = \mathbf{A}^\top \mathbf{A}$ and $\mathbf{H}_1 = \mathbf{B}^\top \mathbf{B}$ for some matrices $\mathbf{A}$ and $\mathbf{B}$: perform an (orthogonal) eigendecomposition of each of the kernel matrices, and take square roots of the eigenvalues. Let $\mathbf{H}$ be the kernel matrix for $h = h_1 h_2$. With $x_i = (x_{i1}, x_{i2})$, its (i, j) entries are

$$\begin{aligned} h(x_i, x_j) &= h_1(x_{i1}, x_{i2})\, h_2(x_{j1}, x_{j2}) \\ &= (\mathbf{A}^\top \mathbf{A})_{ij}\, (\mathbf{B}^\top \mathbf{B})_{ij} \\ &= \sum_{k=1}^n a_{ik} a_{jk} \sum_{l=1}^n b_{il} b_{jl}, \end{aligned}$$

where we have denoted b_{ij} and c_{ij} to be the (i,j)'th entries of $\mathbf{B}$ and $\mathbf{C}$ respectively Then,

$$\begin{aligned}
\sum_{i=1}^{n}\sum_{j=1}^{n} h(x_i, x_j) &= \sum_{k=1}^{n}\sum_{l=1}^{n}\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j a_{ik} a_{jk} b_{il} b_{jl} \\
&= \sum_{k=1}^{n}\sum_{l=1}^{n} \left(\sum_{i=1}^{n} \lambda_i a_{ik} b_{il}\right)\left(\sum_{j=1}^{n} \lambda_j a_{jk} b_{jl}\right) \\
&= \sum_{k=1}^{n}\sum_{l=1}^{n} \left(\sum_{i=1}^{n} \lambda_i a_{ik} b_{il}\right)^2 \\
&\geq 0
\end{aligned}$$

Again, for the remainder of the statement in the lemma, we refer to Berlinet and Thomas-Agnan (2011, Theorem 13). ■

A familiar fact from linear algebra is realised here from Lemmas 2.12 and 2.13: 1) the addition of positive (semi-)definite matrices is a positive-definite matrix; and 2) the *Hadamard product*[3] of two positive (semi-)definite matrices is a positive (semi-)definite matrix.

The scale of a RKHS of functions $\mathcal{F}$ over a set $\mathcal{X}$ with kernel h may be arbitrary. To resolve this issue, a scale parameter $\lambda \in \mathbb{R}$ for the kernel h may be introduced, which will typically need to be estimated from the data. If h is a positive definite-kernel on $\mathcal{X} \times \mathcal{X}$, and $\lambda \geq 0$ a scalar, then this yields a scaled RKHS $\mathcal{F}_\lambda = \{\lambda f \mid f \in \mathcal{F}\}$ with reproducing kernel λh, where $\mathcal{F}$ is the RKHS defined by h.

Restricting λ to the positive reals is arbitrary and unnecessarily restrictive. Especially when considering sums and products of scaled RKHSs, having negative scale parameters also give additional flexibility. The resulting kernels from summation and/or multiplication with negative kernels may no longer be positive definite, and in such cases, they give rise to RKKSs instead.

Remark 2.11. Recall that a RKKS $\mathcal{F}$ of functions over $\mathcal{X}$ can be uniquely decomposed as the difference between two RKHSs $\mathcal{F}_+$ and $\mathcal{F}_-$, and its associated Hilbert space $\mathcal{F}_\mathcal{H}$ is the RKHS $\mathcal{F}_+ \oplus \mathcal{F}_-$. It is important to note that both $\mathcal{F}$ and $\mathcal{F}_\mathcal{H}$ contain identical functions over $\mathcal{X}$, but different topologies. That is to say, functions that are close with respect to the norm of $\mathcal{F}$ may not be close to each other in the norm of $\mathcal{F}_\mathcal{H}$.

[3] The Hadamard product is an element-wise multiplication of two matrices $\mathbf{A}$ and $\mathbf{B}$ of identical dimensions, denoted $\mathbf{A} \circ \mathbf{B}$. That is, $(\mathbf{A} \circ \mathbf{B})_{ij} = \mathbf{A}_{ij}\mathbf{B}_{ij}$.

2.5.2 The polynomial RKKS

A polynomial construction based on a particular RKHS building block is considered here. For example, using the canonical RKHS in the polynomial construction would allow us to easily add higher order effects of the covariates $x \in \mathcal{X}$. In particular, we only require a single scale parameter in polynomial kernel construction.

Definition 2.34 (The polynomial RKKS)**.** Let $\mathcal{X}$ be a Hilbert space. The kernel function $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ obtained through the d-degree polynomial construction of linear kernels is

$$h_\lambda(x, x') = \big(\lambda \langle x, x' \rangle_{\mathcal{X}} + c\big)^d,$$

where $\lambda \in \mathbb{R}$ is a scale parameter for the linear kernel, and $c \in \mathbb{R}$ is a real constant called the *offset*. This kernel defines the *polynomial RKKS* of degree d.

Write

$$h_\lambda(x, x')_{\mathcal{F}} = \sum_{k=0}^{d} \frac{d!}{k!(d-k)!} c^{k-d} \lambda^k \langle x, x' \rangle_{\mathcal{X}}^k.$$

Evidently, as the name suggests, this is a polynomial involving the canonical kernel. In particular, each of the k-powered kernels (i.e., $\langle x, x' \rangle_{\mathcal{X}}^k$) defines a RKHS of their own (since these are merely products of kernels), and therefore the sum of these k-powered kernels define the polynomial RKHS.

The offset parameter influences trade-off between the higher-order versus lower-order terms in the polynomial. It is sometimes known as the bias term.

Proposition 2.14. *The polynomial RKKS $\mathcal{F}$ of real functions over $\mathcal{X}$ contains polynomial functions of the form $f(x) = \sum_{k=0}^{d} \beta_k x^k$.*

Proof. By construction, $\mathcal{F} = \mathcal{F}_0 \oplus \bigoplus_{i=1}^{d} \bigotimes_{j=1}^{i} \mathcal{F}_j$, where each $\mathcal{F}_j, j \neq 0$ is the canonical RKHS, and $\mathcal{F}_0$ is the RKHS of constant functions. Each $f \in \mathcal{F}$ can therefore be written as $f = \beta_0 + \sum_{i=1}^{d} \prod_{j=1}^{i} f_j$, and $f_j(x) = b_j x$ as they are functions from the canonical RKHS, where b_j is a constant. Therefore, $f(x) = \sum_{k=0}^{d} \beta_k x^k$. ■

Remark 2.12. We may opt to use other RKHSs as the building blocks of the polynomial RKHS. In particular, using the centred canonical kernel seems natural, so that each of the functions in the constituents of the direct sum of spaces is centred. However, the polynomial RKKS itself will not be centred.

2.5.3 The ANOVA RKKS

We find it useful to begin this subsection by spending some time to elaborate on the classical analysis of variance (ANOVA) decomposition, and the associated notions of main effects and interactions. This will go a long way in understanding the thinking behind constructing an ANOVA-like RKKS of functions.

The classical ANOVA decomposition

The standard one-way ANOVA is essentially a linear regression model which allows comparison of means from two or more samples. Given sets of observations $y_j = \{y_{1j}, \dots, y_{n_j j}\}$, $j = 1, \dots, m$, we consider the linear model $y_{ij} = \mu_j + \epsilon_{ij}$, where ϵ_{ij} are independent, univariate, normal random variables with a common variance. This covariate-less model is used to make inferences about the *treatment means* μ_j. Often, the model is written in the *overparameterised* form by substituting $\mu_j = \mu + \tau_j$. This gives a different, arguably better, interpretability to the model: the τ_j's, referred to as the *treatment effects*, now represent the amount of deviation from the grand, *overall mean* μ. Estimating all τ_j's and μ separately is not possible because there is one degree of freedom that needs to be addressed in the model: there are $p+1$ mean parameters to estimate but only information from p means. A common fix to this identification issue is to set one of the μ_j's, say the first one μ_1, to zero, or impose the restriction $\sum_{j=1}^m \mu_j = 0$. The former treats one of the m levels as the control, while the latter treats all treatment effects symmetrically.

Now write the ANOVA model slightly differently, as $y_i = f(x_i) + \epsilon_i$, where f is defined on the discrete domain $\mathcal{X} = \{1, \dots, m\}$, and i indexes all of the $n := \sum_{j=1}^m n_j$ observations. Here, f represents the group-level mean, returning μ_j for some $j \in \mathcal{X}$. In a similar manner, we can perform the ANOVA decomposition on f as

$$f = Af + (I - A)f = f_o + f_t,$$

where A is an averaging operator that "averages out" its argument x and returns a constant, and I is the identity operator. $f_o = Af$ is a constant function representing the *overall mean*, whereas $f_t = (I - A)f$ is a function representing the *treatment effects* τ_j. Here are two choices of A:

- $\boldsymbol{Af(x) = f(1) = \mu_1}$. This implies $f(x) = f(1) + \big(f(x) - f(1)\big)$. The overall mean μ is the group mean μ_1, which corresponds to setting the restriction $\mu_1 = 0$.
- $\boldsymbol{Af(x) = \sum_{x=1}^m f(x)/m =: \bar{\alpha}}$. This implies $f(x) = \bar{\alpha} + \big(f(x) - \bar{\alpha}\big)$. The overall mean is $\mu = \sum_{j=1}^m \alpha_j / m$, which corresponds to the restriction $\sum_{j=1}^m \mu_j = 0$.

By definition, $AAf = A^2 f = Af$, because averaging a constant returns that constant. We must have that $Af_t = A(I - A)f = Af - A^2 f = 0$. The choice of A is arbitrary, as is the choice of restriction, so long as it satisfies the condition that $Af_t = 0$.

The multiway ANOVA can be motivated in a similar fashion. Let $x = (x_1, \dots, x_p) \in \prod_{k=1}^p \mathcal{X}_k$, and consider functions that map $\prod_{k=1}^p \mathcal{X}_k$ to $\mathbb{R}$. Let A_j be an averaging operator on $\mathcal{X}_k$ that averages the k'th component of x from the active argument list, i.e. $A_k f$ is constant on the $\mathcal{X}_k$ axis but not necessarily an overall constant function. An ANOVA decomposition of f is

$$f = \left(\prod_{k=1}^p (A_k + I - A_k) \right) f = \sum_{\mathcal{K} \in \mathcal{P}_p} \left(\prod_{k \in \mathcal{K}} (I - A_k) \prod_{k \notin \mathcal{K}} A_k \right) f = \sum_{\mathcal{K} \in \mathcal{P}_p} f_{\mathcal{K}} \tag{2.2}$$

where we had denoted $\mathcal{P}_p = \mathcal{P}(\{1, \dots, p\})$ to be the power set of $\{1, \dots, p\}$ whose cardinality is 2^p. The summands $f_{\mathcal{K}}$ will compose of the overall effect, main effects, two-way interaction terms, and so on. Each of the terms will satisfy the condition $A_k f_{\mathcal{K}} = 0, \forall k \in \mathcal{K} \in \mathcal{P}_p \backslash \{\}$.

Example 2.1 (Two-way ANOVA decomposition)**.** Let $p = 2$, $\mathcal{X}_1 = \{1, \dots, m_1\}$, and $\mathcal{X}_2 = \{1, \dots, m_2\}$. The power set $\mathcal{P}_2$ is $\big\{\{\}, \{1\}, \{2\}, \{1, 2\}\big\}$. The ANOVA decomposition of f (with indices derived trivially from the power set) is

$$f = f_\emptyset + f_1 + f_2 + f_{12}.$$

Here are two choices for the averaging operator A_k analogous to the previous illustration in the one-way ANOVA.

- Let $A_1 f(x) = f(1, x_2)$ and $A_2 f(x) = f(x_1, 1)$. Then,

$$\begin{aligned}
f_\emptyset(x) &= A_1 A_2 f &&= f(1,1) \\
f_1(x) &= (I - A_1) A_2 f &&= f(x_1, 1) - f(1,1) \\
f_2(x) &= A_1 (I - A2) f &&= f(1, x_2) - f(1,1) \\
f_{12}(x) &= (I - A_1)(I - A2) f &&= f(x_1, x_2) - f(x_1, 1) - f(1, x_2) + f(1,1).
\end{aligned}$$

- Let $A_k f(x) = \sum_{x_k=1}^{m_k} f(x_1, x_2)/m_k, k = 1, 2$. Then,

$$\begin{aligned}
f_\emptyset(x) &= A_1 A_2 f &&= f_{\cdot\cdot} \\
f_1(x) &= (I - A_1) A_2 f &&= f_{x_1 \cdot} - f_{\cdot\cdot} \\
f_2(x) &= A_1 (I - A_2) f &&= f_{\cdot x_2} - f_{\cdot\cdot} \\
f_{12}(x) &= (I - A_1)(I - A_2) f &&= f - f_{x_1 \cdot} - f_{\cdot x_2} + f_{\cdot\cdot},
\end{aligned}$$

  where $f_{\cdot\cdot} = \sum_{x_1, x_2} f(x_1, x_2)/m_1 m_2$, $f_{x_1 \cdot} = \sum_{x_2} f(x_1, x_2)/m_2$, and $f_{\cdot x_1} = \sum_{x_1} f(x_1, x_2)/m_1$.

It is also easy to convince ourselves that $A_1 f_1 = A_2 f_2 = A_1 f_{12} = A_2 f_{12} = 0$ in either choice of the averaging operator A_k.

Functional ANOVA decomposition

Let us now extend the ANOVA decomposition idea to a general function $f : \mathcal{X} \to \mathbb{R}$ in some vector space $\mathcal{F}$. We shall jump straight into the multiway ANOVA analogue for functional decomposition, and to that end, consider $x = (x_1, \dots, x_p) \in \prod_{k=1}^p \mathcal{X}_k =: \mathcal{X}$ a measurable space, where each of the spaces $\mathcal{X}_k$ has measure ν_k, and $\nu = \nu_1 \times \cdots \times \nu_p$ is the product measure on $\mathcal{X}$. In the following, denote by $\mathcal{F}_k$ the vector space of functions over the set $\mathcal{X}_k$, $k = 1, \dots, p$, and $\mathcal{F}_\emptyset$ the vector space of constant functions.

As $\mathcal{X}$ need not necessarily be a collection of finite sets, we need to figure out a suitable linear operator that performs an "averaging" of some sort. Consider the linear operator $A_k : \mathcal{F} \to \mathcal{F}_{-k}$, where $\mathcal{F}_{-k}$ is a vector space of functions for which the kth component is constant over $\mathcal{X}$, defined by

$$A_k f(x) = \int_{\mathcal{X}_k} f(x_1, \dots, x_p) \, \mathrm{d}\nu_k(x_k). \tag{2.3}$$

{eq:avgoper}

Thus, for the one-way ANOVA ($p = 1$), we get

$$f(x) = \overbrace{\int_{\mathcal{X}} f(x) \, \mathrm{d}\nu(x)}^{f_\emptyset(x)} + \overbrace{\left(f - \int_{\mathcal{X}} f(x) \, \mathrm{d}\nu(x) \right)}^{f_1(x)} \tag{2.4}$$

{eq:functionalanova1}

and for the two-way ANOVA ($p = 2$), we have $f = f_\emptyset + f_1 + f_2 + f_{12}$, with

$$\begin{aligned}
f_\emptyset(x) &= \int_{\mathcal{X}_1} \int_{\mathcal{X}_2} f(x_1, x_2) \, \mathrm{d}\nu_1(x_1) \, \mathrm{d}\nu_2(x_2) \\
f_1(x) &= \int_{\mathcal{X}_2} \left(f(x_1, x_2) - \int_{\mathcal{X}_1} f(x_1, x_2) \, \mathrm{d}\nu_1(x_1) \right) \mathrm{d}\nu_2(x_2) \\
f_2(x) &= \int_{\mathcal{X}_1} \left(f(x_1, x_2) - \int_{\mathcal{X}_2} f(x_1, x_2) \, \mathrm{d}\nu_2(x_2) \right) \mathrm{d}\nu_1(x_1) \\
f_{12}(x) &= f(x_1, x_2) - \int_{\mathcal{X}_1} f(x_1, x_2) \, \mathrm{d}\nu_1(x_1) - \int_{\mathcal{X}_2} f(x_1, x_2) \, \mathrm{d}\nu_2(x_2) \\
&\quad + \int_{\mathcal{X}_1} \int_{\mathcal{X}_2} f(x_1, x_2) \, \mathrm{d}\nu_1(x_1) \, \mathrm{d}\nu_2(x_2).
\end{aligned}$$

The averaging operator A_k defined in (2.3) generalises the concept of the previous subsection's averaging operator. We must then also have, as before, that $A_k f_{\mathcal{K}} = 0, \forall k \in \mathcal{K} \in \mathcal{P}_p \backslash \{\}$. For the one-way functional ANOVA decomposition in (2.4), it must be that f_1 is a zero-mean function. As for the two-way ANOVA, it is the case that $\int_{\mathcal{X}_k} f_{\mathcal{K}}(x_1, x_2) \, \mathrm{d}\nu_k(x_k) = 0, k = 1, 2$, and $\mathcal{K} \in \{\{1\}, \{2\}, \{1, 2\}\}$ (Durrande et al., 2013).

This is highly suggestive as to what the ANOVA decomposition of the space $\mathcal{F}$ should look like in general. Starting with $p = 1$, any $f \in \mathcal{F}$ can be decomposed as a sum of a constant plus a zero mean function, so we have the decomposition of the vector space $\mathcal{F} = \mathcal{F}_\emptyset \oplus \bar{\mathcal{F}}_1$, where a bar over $\mathcal{F}_k$, $k = 1, \dots, p$ will be used to denote the vector space of zero mean functions over $\mathcal{X}_k$. For $p \geq 2$ we can argue something similar. Take the vector space space $\mathcal{F}$ of functions over $\prod_{k=1}^p \mathcal{X}_k$ to be the tensor product space $\mathcal{F} = \mathcal{F}_1 \otimes \cdots \otimes \mathcal{F}_p$ whose elements are identified as being tensor product functions $f_1 \otimes \cdots \otimes f_p$, where each $f_k : \mathcal{X}_k \to \mathbb{R}$ belongs to $\mathcal{F}_k$. This is constructed by repeatedly taking the completion of linear combinations of the tensor product $f_k \otimes f_j$, $k, j \in \{1, \dots, p\}$ as per Definition 2.14. Considered individually, each $\mathcal{F}_k$ can then be decomposed as $\mathcal{F}_k = \mathcal{F}_{\emptyset k} \oplus \bar{\mathcal{F}}_k$, where $\mathcal{F}_{\emptyset k}$ is the space of functions constant along the k'th axis. Expanding out under the distributivity rule of tensor products and rearranging slightly, we obtain

$$\begin{aligned}
\mathcal{F} &= \left(\mathcal{F}_{\emptyset 1} \oplus \bar{\mathcal{F}}_1\right) \otimes \cdots \otimes \left(\mathcal{F}_{\emptyset p} \oplus \bar{\mathcal{F}}_p\right) \\
&= \mathcal{F}_\emptyset \ \oplus \ \bigoplus_{j=1}^p \Big(\bigotimes_{i \neq j} \mathcal{F}_{\emptyset i} \otimes \bar{\mathcal{F}}_j \Big) \ \oplus \ \bigoplus_{\substack{j,k=1 \\ j<k}}^p \Big(\bigotimes_{i \neq j,k} \mathcal{F}_{\emptyset i} \otimes \bar{\mathcal{F}}_j \otimes \bar{\mathcal{F}}_k \Big) \\
&\quad \oplus \ \cdots \ \oplus \ \Big(\bar{\mathcal{F}}_1 \otimes \cdots \otimes \bar{\mathcal{F}}_p \Big),
\end{aligned} \tag{2.5}$$

{eq:funcanovaspace}

where '$\bigoplus$' and '$\bigotimes$' represent the summation and product operator for direct/tensor sums and products, respectively. To clarify,

- $\mathcal{F}_\emptyset$ is the space of constant functions $f_\emptyset : \mathcal{X}_1 \times \cdots \times \mathcal{X}_p \to \mathbb{R}$;
- $\Big(\bigotimes_{i \neq j} \mathcal{F}_{\emptyset i} \otimes \bar{\mathcal{F}}_j \Big)$ is the space of functions that are constant on all coordinates except the j'th coordinate of x, and the functions are centred on the j'th coordinate;
- $\Big(\bigotimes_{i \neq j,k} \mathcal{F}_{\emptyset i} \otimes \bar{\mathcal{F}}_j \otimes \bar{\mathcal{F}}_k \Big)$ is the space of functions that are constant on all coordinates except the jth and kth coordinate of x, and the functions are centred on these two coordinates;
- $\bar{\mathcal{F}}_1 \otimes \cdots \otimes \bar{\mathcal{F}}_p$ is the space of zero-mean functions $f : \mathcal{X}_1 \times \cdots \times \mathcal{X}_p \to \mathbb{R}$;

and so on for for the rest of the spaces in the summand, of which there are 2^p members all together. Therefore, given an arbitrary function $f \in \mathcal{F}$, the projection of f onto the above respective spaces in (2.5) leads to the *functional ANOVA representation*

$$f(x) = \alpha + \sum_{j=1}^p f_j(x_j) + \sum_{\substack{j,k=1 \\ j<k}}^p f_{jk}(x_j, x_k) + \cdots + f_{1 \cdots p}(x), \tag{2.6}$$

{eq:functionalanova2}

where α is the grand intercept (a constant).

Definition 2.35 (Functional ANOVA representation)**.** Let $\mathcal{P}_p = \mathcal{P}(\{1,\dots,p\})$, the power set of $\{1,\dots,p\}$. For any function $f \in \mathcal{F} \equiv \mathcal{F}_1 \otimes \cdots \otimes \mathcal{F}_p$, with each $\mathcal{F}_k$ a space of functions over $\mathcal{X}_k$, $k = 1,\dots,p$, the formula for f in (2.6) is known as the *functional ANOVA representation* of f if $\forall k \in \mathcal{K} \in \mathcal{P}_p \backslash \{\}$,

$$A_k f_{\mathcal{K}}(x) = \int_{\mathcal{X}_k} f_{\mathcal{K}}(x)\,\mathrm{d}\nu_k(x_k) = 0. \tag{2.7}$$

{eq:funcanovaorth}

In other words, the integral of $f_{\mathcal{K}}$ with respect to any of the variables indexed by the elements in $\mathcal{K}$, is zero.

For the constant term, main effects, and two-way interaction terms, the familiar classical expressions are obtained:

$$\begin{aligned} f_\emptyset &= \int f\,\mathrm{d}\nu; \\ f_j &= \int f \textstyle\prod_{i\neq j} \mathrm{d}\nu_i - f_\emptyset; \\ f_{jk} &= \int f \textstyle\prod_{i\neq j,k} \mathrm{d}\nu_i - f_j - f_k - f_\emptyset. \end{aligned}$$

The ANOVA kernel

At last, we come to the section of deriving the ANOVA RKKS, and, rest assured, the preceding long build-up will prove to not be in vain. The main idea is to construct a RKKS such that the functions that lie in them will have the ANOVA representation in (2.6). The bulk of the work has been done, and in fact we know exactly how this ANOVA RKKS should be structured—it is the space as specified in (2.5). The ANOVA RKKS will be constructed by a similar manipulation of the individual kernels representing the RKHS building blocks.

def:anovarkks

Definition 2.36 (The ANOVA RKKS)**.** For $k = 1,\dots,p$, let $\mathcal{F}_k$ be centred RKHSs of functions over the set $\mathcal{X}_k$ with kernel $h_k : \mathcal{X}_k \times \mathcal{X}_k \to \mathbb{R}$. Let $\lambda_k, k = 1,\dots,p$ be real-valued scale parameters. The ANOVA RKKS of functions $f : \mathcal{X}_1 \times \cdots \times \mathcal{X}_p \to \mathbb{R}$ is specified by the ANOVA kernel, defined by

$$h_\lambda(x,x') = \prod_{k=1}^{p} \big(1 + \lambda_k h_k(x_k, x_k')\big). \tag{2.8}$$

{eq:anovarkks}

It is interesting to note that an ANOVA RKKS is constructed very simply through multiplication of univariate kernels. Expanding out equations (2.8), we see that it is in

fact a sum of products of kernels with increasing orders of interaction:

$$h_\lambda(x, x') = 1 + \sum_{j=1}^{p} \lambda_j h_j(x_j, x'_j) + \sum_{\substack{j,k=1 \\ j<k}}^{p} \lambda_j \lambda_k h_j(x_j, x'_j) h_k(x_k, x'_k) + \cdots + \prod_{j=1}^{p} \lambda_j h_j(x_j, x'_j).$$

It is now clear from this expansion that the ANOVA RKKS yields functions that resemble those with the ANOVA representation in (2.6): the mean value of the function stems from the '1', i.e. it lies in a RKHS of constant functions; the main effects are represented by the sum of the individual kernels; the two-way interaction terms are represented by the second-order kernel interactions; and so on.

Example 2.2. Consider two RKKSs $\mathcal{F}_k$ with kernel $\lambda_k h_k$, $k = 1, 2$. The ANOVA kernel defining the ANOVA RKKS $\mathcal{F}$ is

$$h_\lambda\big((x_1, x_2), (x'_1, x'_2)\big) = 1 + \lambda_1 h_1(x_1, x'_1) + \lambda_2 h_2(x_2, x'_2) + \lambda_1 \lambda_2 h_1(x_1, x'_1) h_2(x_2, x'_2).$$

Suppose that $\mathcal{F}_1$ and $\mathcal{F}_2$ are the centred canonical RKKS of functions over $\mathbb{R}$. Then, functions in $\mathcal{F} = \mathcal{F}_0 \oplus \mathcal{F}_1 \oplus \mathcal{F}_2 \oplus (\mathcal{F}_1 \otimes \mathcal{F}_2)$ are of the form

$$f(x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2.$$

As a remark, not all of the components of the ANOVA RKKS need to be included in the construction. The selective exclusion of certain interactions characterises many interesting statistical models. Excluding certain terms of the ANOVA RKKS is equivalent to setting the scale parameter for those relevant components to be zero, i.e. they play no role in the decomposition of the function. With this in mind, the ANOVA RKKS then gives us an objective way of model-building, from linear regression, to multilevel models, longitudinal models, and so on.

Finally, we note that the functional ANOVA decomposition of a RKKS is orthogonal. Without loss of generality, assume that all scale parameters are positive. For $p = 1$, we have that $\mathcal{F} = \mathcal{F}_\emptyset \oplus \mathcal{F}_1$, where each of $\mathcal{F}_\emptyset$ and $\mathcal{F}_1$ is a RKHS. From Lemma 2.12, the squared norm of $\mathcal{F}$ is given by

$$\|f\|_{\mathcal{F}}^2 = \|f_\emptyset\|_{\mathcal{F}_\emptyset}^2 + \|f_1\|_{\mathcal{F}_1}^2,$$

if the decomposition $f = f_\emptyset + f_1$ is minimal. Hence, the decomposition is orthogonal. An inductive argument can be used to extend and generalise to any $p \geq 2$.

sec:summarychapter2

2.6 Summary

The review of functional analysis allows us to describe the theory of RKHSs and RKKSs, which are of interest to us because the topology endowed on such spaces gives appreciable assurances—in particular, all evaluation functionals are continuous in these spaces. Moreover, RKHSs and RKKSs can be specified completely through kernel functions, with new and complex function spaces built simply by manipulation of these kernel functions. Of particular importance is the ANOVA functional decomposition, for which we realise provides an objective way of constructing various function spaces for regression and modelling. Such models will be described later on in detail in Chapter 4.

An annotated collection of bibliographical references used for this chapter is as follows.

- **Functional analysis**. On the introductory material relating to functional analysis in Section 2.1, the lecture notes by Sejdinovic and Gretton (2012) is recommended, and forms the basis for most of our material. Additionally, Kokoszka and Reimherr (2017), Rudin (1987), and Yamamoto (2012) provides a complementary reading.
- **RKHS theory**. There are certainly no shortages of introductory texts relating to the theory of RKHS: Steinwart and Christmann (2008), Berlinet and Thomas-Agnan (2011), and Gu (2013) to name a few. The concise sketch proof for the Moore-Aronszajn theorem was mostly inspired by Hein and Bousquet (2004, Theorem 4).
- **Kreĭn space and RKKS theory**. The innovation of indefinite inner product spaces perhaps started in mathematical physics literature, for which the theory of special relativity depends. Four-dimensional space-time is an often cited example. In any case, we referred to mainly Ong et al. (2004), which gives an overview in the context of learning using indefinite kernels. Alpay (1991) and Zafeiriou (2012) were also useful for understanding the fundamental concepts of RKKSs.
- **RKHS building blocks**. The main building block RKHSs, i.e. the canonical RKHS, the fBm RKHS and the Pearson RKHS, are described in the manuscript of Bergsma (2017).
- **ANOVA and functional ANOVA**. Classical ANOVA is pretty much existent in every fundamental statistical textbook. These texts have extremely well written introductions to this very important concept: Casella and Berger (2002, Ch. 11), Dean and Voss (1999, Ch. 3). On the relation between classical ANOVA and functional ANOVA decomposition, Gu (2013) offers novel insights. There is diverse literature concerning functional ANOVA, namely from the fields of statistical learning (e.g. Wahba, 1990), applied mathematics (e.g. Kuo et al., 2010), and sensitivity analysis (e.g. Durrande et al., 2013; Sobol, 2001).

Bibliography

alpay1991some — Alpay, Daniel (1991). "Some remarks on reproducing kernel Krein spaces". In: *The Rocky Mountain Journal of Mathematics*, pp. 1189–1205.

bergsma2017 — Bergsma, Wicher (2017). "Regression with I-priors". In: *Unpublished manuscript*.

berlinet2011reproducing — Berlinet, Alain and Christine Thomas-Agnan (2011). *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer-Verlag. DOI: 10.1007/978-1-4419-9096-9.

casella2002statistical — Casella, George and Roger L Berger (2002). *Statistical inference*. Vol. 2. Duxbury Pacific Grove, CA.

cohen2002 — Cohen, S (2002). "Champs localement auto-similaires". In: *Lois d'échelle, fractales et ondelettes*. Ed. by Patrice Abry, Paulo Gonçalves, and Jacques Lévy Véhel. Vol. 1. Hermès Sciences Publications.

dean1999design — Dean, Angela and Daniel Voss (1999). *Design and analysis of experiments*. Vol. 1. Springer.

durrande2013anova — Durrande, Nicolas, David Ginsbourger, Olivier Roustant, and Laurent Carraro (2013). "ANOVA kernels and RKHS of zero mean functions for model-based sensitivity analysis". In: *Journal of Multivariate Analysis* 115, pp. 57–67.

duvenaud2014automatic — Duvenaud, David (2014). "Automatic model construction with Gaussian processes". PhD thesis. University of Cambridge.

embrechts2002selfsimilar — Embrechts, Paul and Makoto Maejima (2002). *Selfsimilar Processes. Princeton series in applied mathematics*. Princeton University Press, Princeton, NJ.

gu2013smoothing — Gu, Chong (2013). *Smoothing spline ANOVA models*. Vol. 297. Springer Science & Business Media.

hein2004kernels — Hein, Matthias and Olivier Bousquet (2004). "Kernels, associated structures and generalizations". In: *Max-Planck-Institut fuer biologische Kybernetik, Technical Report*.

hofmann2008kernel — Hofmann, Thomas, Bernhard Schölkopf, and Alexander J Smola (2008). "Kernel methods in machine learning". In: *The annals of statistics*, pp. 1171–1220.

kokoszka2017introduction — Kokoszka, Piotr and Matthew Reimherr (2017). *Introduction to functional data analysis*. CRC Press.

kuo2010decompositions — Kuo, F, I Sloan, G Wasilkowski, and Henryk Woźniakowski (2010). "On decompositions of multivariate functions". In: *Mathematics of computation* 79.270, pp. 953–966.

mandelbrot1968fractional — Mandelbrot, Benoit B and John W Van Ness (1968). "Fractional Brownian motions, fractional noises and applications". In: *SIAM review* 10.4, pp. 422–437.

mary2003hilbertian — Mary, Xavier (2003). "Hilbertian subspaces, subdualities and applications". PhD thesis. INSA de Rouen.

micchelli2006universal — Micchelli, Charles A, Yuesheng Xu, and Haizhang Zhang (2006). "Universal kernels". In: *Journal of Machine Learning Research* 7.Dec, pp. 2651–2667.

ong2004learning — Ong, Cheng Soon, Xavier Mary, Stéphane Canu, and Alexander J Smola (2004). "Learning with non-positive kernels". In: *Proceedings of the twenty-first international conference on Machine learning*. ACM, p. 81.

raykar2007fast — Raykar, Vikas C and Ramani Duraiswami (2007). "Fast large scale Gaussian process regression using approximate matrix-vector products". In:

rudin1987real — Rudin, Walter (1987). *Real and complex analysis*. Tata McGraw-Hill Education.

schoenberg1937 — Schoenberg, Isaac J (1937). "On certain metric spaces arising from Euclidean spaces by a change of metric and their imbedding in Hilbert space". In: *Annals of mathematics*, pp. 787–793.

sejdinovic2012 — Sejdinovic, Dino and Arthur Gretton (2012). "Lecture notes: What is an RKHS?" In: *COMPGI13 Advanced Topics in Machine Learning. Lecture conducted at University College London*, pp. 1–24. URL: http://www.gatsby.ucl.ac.uk/%7B~%7Dgretton/coursefiles/RKHS%7B%5C_%7DNotes1.pdf.

sobol2001global — Sobol, Ilya M (2001). "Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates". In: *Mathematics and computers in simulation* 55.1-3, pp. 271–280.

steinwart2008support — Steinwart, Ingo and Andreas Christmann (2008). *Support vector machines*. Springer Science & Business Media.

steinwart2006explicit — Steinwart, Ingo, Don Hush, and Clint Scovel (2006). "An explicit description of the reproducing kernel Hilbert spaces of Gaussian RBF kernels". In: *IEEE Transactions on Information Theory* 52.10, pp. 4635–4643.

wahba1990spline — Wahba, Grace (1990). *Spline models for observational data*. Vol. 59. Siam.

yamamoto2012vector — Yamamoto, Yutaka (2012). *From vector spaces to function spaces: introduction to functional analysis with applications*. Vol. 127. SIAM.

zafeiriou2012subspace

Zafeiriou, Stefanos (2012). "Subspace learning in krein spaces: Complete kernel fisher discriminant analysis with indefinite kernels". In: *European Conference on Computer Vision*. Springer, pp. 488–501.

Figures

2.1 A hierarchy of vector spaces 12
2.2 Relationship between p.d. functions, rep. kernels, and RKHSs 15
2.3 Sample I-prior paths from the RKHS of constant functions. 21
2.4 Sample paths from the canonical RKHS. 22
2.5 Sample I-prior paths from the fBm RKHS with varying Hurst coefficients. . . 25
2.6 Sample paths from the SE RKHS with varying values for the lengthscale. . . 27
2.7 Sample I-prior "paths" from the Pearson RKHS 29

Tables

Theorems

2.1 Lemma (Equivalence of boundedness and continuity) 6
2.2 Theorem (Riesz-Fréchet) 7
2.3 Theorem (Orthogonal decomposition) 8
2.5 Theorem (RKHS uniqueness) 15
2.6 Theorem (Moore-Aronszajn) 16
2.7 Lemma (Uniqueness of kernel for RKKS) 19
2.12 Lemma (Sum of kernels) 30
2.13 Lemma (Products of kernels) 30

Definitions

2.1 Definition (Inner products) . . . 3
2.2 Definition (Norms) . . . 3
2.3 Definition (Convergent sequence) . . . 4
2.4 Definition (Cauchy sequence) . . . 4
2.5 Definition (Linear map/operator) . . . 5
2.6 Definition (Bilinear map/operator) . . . 5
2.7 Definition (Continuity) . . . 5
2.8 Definition (Lipschitz continuity) . . . 6
2.9 Definition (Bounded operator) . . . 6
2.10 Definition (Dual spaces) . . . 6
2.11 Definition (Isometric isomorphism) . . . 7
2.12 Definition (Orthogonal complement) . . . 8
2.13 Definition (Tensor products) . . . 8
2.14 Definition (Tensor product space) . . . 9
2.15 Definition (Mean vector) . . . 11
2.16 Definition (Covariance operator) . . . 11
2.17 Definition (Gaussian vectors) . . . 11
2.18 Definition (Evaluation functional) . . . 12
2.19 Definition (Reproducing kernel Hilbert space) . . . 12
2.20 Definition (Reproducing kernels) . . . 13
2.21 Definition (Kernel matrix) . . . 14
2.22 Definition (Negative and indefinite inner products) . . . 17
2.23 Definition (Kreĭn space) . . . 17
2.24 Definition (Associated Hilbert space) . . . 18
2.25 Definition (Reproducing kernel Kreĭn space) . . . 18
2.26 Definition (Centred canonical RKHS) . . . 22
2.27 Definition (Fractional Brownian motion RKHS) . . . 23
2.28 Definition (Hölder condition) . . . 24
2.29 Definition (Centred fBm RKHS) . . . 25
2.30 Definition (Squared exponential RKHS) . . . 26
2.31 Definition (Universal kernel) . . . 26
2.32 Definition (Centred SE RKHS) . . . 27

2.33 Definition (Pearson RKHS) 28
2.34 Definition (The polynomial RKKS) 32
2.35 Definition (Functional ANOVA representation) 36
2.36 Definition (The ANOVA RKKS) 37

Nomenclature

As much as possible, and unless otherwise stated, the following conventions are used throughout this thesis.

Conventions

$\mathbf{a}, \mathbf{b}, \mathbf{c}, \ldots$	Boldface lower case letters denote real vectors
$\mathbf{A}, \mathbf{B}, \mathbf{C}, \ldots$	Boldface upper case letters denote real matrices
$\mathcal{A}, \mathcal{B}, \mathcal{C}, \ldots$	Calligraphic upper case letters denote sets

Indexing

$\mathbf{A}_{ij}$, A_{ij}, a_{ij}	The (i, j)'th element of the matrix $\mathbf{A}$
$\mathbf{A}_{i\cdot}$	The i'th row of the matrix $\mathbf{A}$ as a tall vector (transposed row vector)
$\mathbf{A}_{\cdot j}$	The j'th column vector of the matrix $\mathbf{A}$

Symbols

$\mathbb{N}$	The set of natural numbers (excluding zero)
$\mathbb{Z}$	The set of integers
$\mathbb{R}$	The set of real numbers
$\mathbb{R}^d$	The d-dimensional Euclidean space
x'	Primes are used to distinguish elements, rather than to denote derivatives
$\hat{\theta}$	Hats are used to denote estimators of a parameter θ
$\mathcal{A}^c$	The complement of a set $\mathcal{A}$
$\mathcal{P}(\mathcal{A})$	The power set of the set $\mathcal{A}$
$\{\}, \emptyset$	The empty set
$\mathbf{0}$	A vector of zeroes
$\mathbf{1}_n$	A length n vector of ones
$\mathbf{I}_n$	The $n \times n$ identity matrix
$\exists$	(short hand) There exists
$\forall$	(short hand) For all
$\lim_{n\to\infty}$	The limit as n tends to infinity
$\xrightarrow{\text{dist.}}$	Convergence in distribution
$O(n)$	Computational complexity (time or storage)
Δx	A quantity representing a change in x

Relations

$a \approx b$	a is approximately or almost equal to b
$a \propto b$	a is equivalent to b up to a constant of proportionality
$a \equiv b$	a is identical to b
$A \Rightarrow B$	The statement B being true is predicated on A being true
$A \Leftrightarrow B$	The statement A is true if and only if B is true
$a \in \mathcal{A}$	a is an element of the set $\mathcal{A}$
$\mathcal{A} \subseteq \mathcal{B}$	$\mathcal{A}$ is a subset of $\mathcal{B}$ which may include itself
$\mathcal{A} \subset \mathcal{B}$	$\mathcal{A}$ is a subset of $\mathcal{B}$ which does not include itself
$a := b$, $a \leftarrow b$	a is assigned the value b
$X \sim p(X)$	The random variable X is distributed according to the pdf $p(X)$
$X \sim D$	The random variable X is distributed according to the pdf specified by the distribution D, e.g. $D \equiv \mathrm{N}(0,1)$
$X_1,\ldots,X_n \overset{\text{iid}}{\sim} D$	Each random variable X_i, $i = 1, \ldots, n$ is independently and identically distributed according to the pdf specified by the distribution D
$X \vert Y$	The (random) variable X given/conditional on Y

Functions

$\inf \mathcal{A}$	The infimum of a set $\mathcal{A}$
$\sup \mathcal{A}$	The supremum of a set $\mathcal{A}$
$\min \mathcal{A}$	The minimum value of a set $\mathcal{A}$
$\max \mathcal{A}$	The maximum value of a set $\mathcal{A}$
$\arg\min_x f(x)$	The value of x which minimises the function $f(x)$
$\arg\max_x f(x)$	The value of x which maximises the function $f(x)$
$\vert a \vert$ with $a \in \mathbb{R}$	The absolute value of a; $\vert a \vert = a$ if a is positive, and $-a$ if a is negative, and $\vert 0 \vert = 0$
$\delta_{xx'}$	The Kronecker delta; $\delta_{xx'} = 1$ if $x = x'$, and 0 otherwise
$[A]$	The Iverson bracket; $[A] = 1$ if the logical proposition A is true, and 0 otherwise
$\mathbb{1}_{\mathcal{A}}(x)$	The indicator function; $\mathbb{1}_{\mathcal{A}}(x) = 1$ if $x \in \mathcal{A}$, and 0 otherwise
e^x, $\exp(x)$	The natural exponential function
$\log(x)$	The natural logarithmic function
$\frac{\mathrm{d}}{\mathrm{d}x} f(x)$, $\dot{f}(x)$	The derivative of f with respect to x

Abstract vector space operations and notations

$\mathcal{V}^\perp$	The orthogonal complement of the space $\mathcal{V}$
$\mathcal{V}^\vee$	The algebraic dual space of $\mathcal{V}$
$\mathcal{V}^*$	The continuous dual space of $\mathcal{V}$
$\overline{\mathcal{V}}$	The closure of the space $\mathcal{V}$
$\mathcal{B}(\mathcal{V})$	The Borel σ-algebra of $\mathcal{V}$
$\mathrm{L}^p(\mathcal{X}, \nu)$	The set of p-integrable functions over the measure space $\mathcal{X}$ with measure ν
$\mathrm{L}(\mathcal{V}; \mathcal{W})$	The set of bounded, linear operators from $\mathcal{V}$ to $\mathcal{W}$
$\dim(\mathcal{V})$	The dimensions of the vector space $\mathcal{V}$
$\langle x, y \rangle_{\mathcal{V}}$	The inner product between x and y in the vector space $\mathcal{V}$

$\|x\|_{\mathcal{V}}$	The norm of x in the vector space $\mathcal{V}$
$D(x, y)$	The distance between x and y
$x \otimes y$	The tensor product of x and y which are elements of a vector space
$\mathcal{F} \otimes \mathcal{G}$	The tensor product space of two vector spaces
$\mathcal{F} \oplus \mathcal{G}$	The direct sum (or tensor sum) of two vector spaces
$\mathrm{d}f(x)$, $\mathrm{d}^2 f(x)$	The first and second Fréchet differentials of f at x
$\partial_v f(x)$, $\partial_v^2 f(x)$	The first and second Gâteaux differentials of f at x in the direction v
$\nabla f(x)$, $\nabla^2 f(x)$	The gradient and Hessian of f at x in the direction v (f is a mapping of a Hilbert space)

Matrix and vector operations

$\mathbf{a}^\top$, $\mathbf{A}^\top$	The transpose of a vector $\mathbf{a}$ or matrix $\mathbf{A}$
$\mathbf{A}^{-1}$	The inverse of a square matrix $\mathbf{A}$
$\|\mathbf{a}\|^2$	The squared 2-norm the vector $\mathbf{a}$, equivalent to $\mathbf{a}^\top\mathbf{a}$
$\|\mathbf{A}\|$	The determinant of a matrix $\mathbf{A}$
$\mathrm{tr}(\mathbf{A})$	The trace of a square matrix $\mathbf{A}$
$\mathrm{diag}(\mathbf{A})$	The diagonal elements of a square matrix $\mathbf{A}$
$\mathrm{rank}(\mathbf{A})$	The rank of a matrix $\mathbf{A}$
$\mathrm{vec}(\mathbf{A})$	The column-wise vectorisation of a matrix $\mathbf{A}$
$\mathbf{a} \otimes \mathbf{b}$	The outer product of two vectors $\mathbf{a}$ and $\mathbf{b}$
$\mathbf{A} \otimes \mathbf{B}$	The Kronecker product of matrix $\mathbf{A}$ with matrix $\mathbf{B}$
$\mathbf{A} \circ \mathbf{B}$	The Hadamard product two matrices $\mathbf{A}$ and $\mathbf{B}$

Statistical functions

$\mathrm{P}(A)$	The probability of event A occurring
$p(X\|\theta)$	The probability density function of X given parameters θ
$L(\theta\|X)$	The log-likelihood of θ given data X, sometimes simply $L(\theta)$ or $L(\theta\|M_k)$, the (marginal) log-likelihood under model assumptions M_k
$\mathrm{BF}(M, M')$	Bayes factor for comparing two models M and M'
$I(\theta)$	The Fisher information for θ
$\mathrm{E}[X]$	The expectation[4] of the random element X
$\mathrm{Var}[X]$	The variance[4] of the random element X
$\mathrm{Cov}[X, Y]$	The covariance[4] between two random elements X and Y
$H(p)$	The entropy of the distribution $p(X)$
$\mathrm{KL}[q(x)\|p(x)]$	The Kullbeck-Leibler divergence from $p(x)$ to $q(x)$, denoted also by $\mathrm{KL}(q\|p)$

Statistical distributions

$\mathrm{N}(\mu, \sigma^2)$	Univariate normal distribution with mean μ and variance σ^2
$\mathrm{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	d-dimensional multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$
$\phi(z)$	The standard normal pdf

foot:exp

[4] When there is ambiguity as to which random element the expectation or variance is taken under or what its distribution is, this is explicated by means of subscripting, e.g. $\mathrm{E}_{X\sim\mathrm{N}(0,1)}[X]$ to denote the expectation of a standard normal random variable.

$\Phi(z)$	The standard normal cdf
$\phi(x\|\mu,\sigma^2)$	The pdf of $\mathrm{N}(\mu,\sigma^2)$
$\phi(\mathbf{x}\|\boldsymbol{\mu},\boldsymbol{\Sigma})$	The pdf of $\mathrm{N}_d(\boldsymbol{\mu},\boldsymbol{\Sigma})$
$\mathrm{MN}_{n,m}(\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\Psi})$	Matrix normal distribution with mean $\boldsymbol{\mu}$ and row variances $\boldsymbol{\Sigma}\in\mathbb{R}^{n\times n}$ and column variances $\boldsymbol{\Psi}\in\mathbb{R}^{m\times m}$
${}^{\mathrm{t}}\mathrm{N}(\mu,\sigma^2,a,b)$	Truncated univariate normal distribution with mean μ and variance σ^2 restricted to the interval (a,b)
$\mathrm{N}_+(\mu,\sigma^2)$	The half-normal distribution with mean μ and variance σ^2
${}^{\mathrm{t}}\mathrm{N}_d(\boldsymbol{\mu},\boldsymbol{\Sigma},\mathcal{A})$	Truncated d-dimensional multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ restricted to the set $\mathcal{A}$
$\Gamma(s,r)$	Gamma distribution with shape s and rate r parameters
$\Gamma^{-1}(s,\sigma)$	Inverse gamma distribution with shape s and scale σ parameters
χ^2_d	Chi-squared distribution with d degrees of freedom
$\mathrm{Bern}(p)$	Bernoulli distribution with probability of success p
$\mathrm{Cat}(p_1,\ldots,p_m)$	Categorical distribution with m categories, and each category has probability of success p_j

Abbreviations

ANOVA	Analysis of variance
CRAN	Comprehensive R Archive Network
EM	expectation-maximisation
fBm	Fractional Brownian motion
GPR	Gaussian process regression
Lasso	Least absolute shrinkage and selection operator
MCMC	Markov chain Monte Carlo
ML	maximum likelihood
RKHS	Reproducing kernel Hilbert space
RKKS	Reproducing kernel Kreĭn space
SE	Squared exponential (kernel)