To-do list

Contents

Haziq Jamil
Department of Statistics
London School of Economics and Political Science
February 16, 2018

Chapter 2

Vector space of functions

One of the main assumptions for regression modelling with I-priors is that the regression functions lie in some vector space of functions. At first glance, this may seem strange, that the notion of functions (as mappings from input to output space) and vector spaces are somehow equatable. Upon further thought, one realises that firstly, two functions of a similar, particular form may be added together (in some meaningful way) resulting in a function in that same form. Secondly, multiplication of a function by a scalar c can be thought of as c times the output of that function. Indeed, running through the checklist[1] of what constitutes a vector space, we find that a "space of functions" satisfies them all.

The purpose of this chapter is to provide a concise review of functional analysis leading up to the theory of reproducing kernel Hilbert and Krein spaces (RKHS/RKKS). The interest with these RKHS and RKKS 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, RKHS 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. RKHS are widely studied in the literature, but perhaps RKKS are less so. To provide an early insight, RKKS are simply an extension of RKHS when its kernel is not positive-definite.

[1] In modern linear algebra texts, this is the eight axioms of vector spaces over a field $\mathbb{F}$: The vectors forms an abelian group under addition, and this group has an $\mathbb{F}$-module structure.

The flexibility provided by RKKS will prove both useful and necessary, especially when considering scaled function spaces, as I-prior modelling does.

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 2.1–2.3. Section 2.4 describes the fundamental RKHS of interest for I-prior regression, which we refer to as the "building block" RKHS/RKKS. The reason for this is that it is possible to construct new RKKS from existing ones, and this is described in Section 2.5.

A remark on notation: Elements of the vector space $\mathcal{F}$ of real functions over a set $\mathcal{X}$ are denoted $f(\cdot)$, or 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.

2.1 Some functional analysis

The core study of functional analysis revolves around the treatment of functions as objects in vector spaces over a field[2]. 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.

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:** $\langle af_1 + bf_2, g \rangle_{\mathcal{F}} = a\langle f_1, g \rangle_{\mathcal{F}} + b\langle f_2, g \rangle_{\mathcal{F}}$, $\forall f_1, f_2, g \in \mathcal{F}$ and $\forall a, b \in \mathbb{R}$.
- **Non-degeneracy:** $\langle f, f \rangle_{\mathcal{F}} = 0 \Leftrightarrow f = 0$.

Additionally, an inner product is said to be *positive definite* if $\langle f, f \rangle_{\mathcal{F}} \geq 0$, $\forall f \in \mathcal{F}$. It is possible that this may not be the case, and we shall revisit this fact later when we cover Krein spaces. For the purposes of the discussion moving forward, we shall refer to

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

positive definite inner products, unless otherwise stated.

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

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

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:** $||\lambda f||_{\mathcal{F}} = |\lambda| \cdot ||f||_{\mathcal{F}}$, $\forall \lambda \in \mathbb{R}$, $\forall f \in \mathcal{F}$
- **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 $\|f\|_{\mathcal{F}} + \|-f\|_{\mathcal{F}} \geq \|f - f\|_{\mathcal{F}} = \|0\|_{\mathcal{F}} = 0$, thus implying non-negativity of norms. Several important relationships between norms and inner products hold in linear spaces, namely, the *Cauchy-Schwarz inequality*

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

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*

$$\|f + g\|_{\mathcal{F}}^2 + \|f + g\|_{\mathcal{F}}^2 = 4\langle f, g \rangle_{\mathcal{F}},$$

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

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 $\mathcal{F}$: $d(f, g) =$

$||f-g||_{\mathcal{F}}$, 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 be *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. If the vector space contains the limits of all Cauchy sequences (or in other words, if every Cauchy sequence converges), then it is said to be *complete*. On the other hand, a set which contains all of its limit points is said to be *closed*. Clearly, a complete set must closed, but a closed set need not necessarily be complete.

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.

Being vectors in a vector space, we can discuss mapping the vectors onto a different space, or in essence, having a function acted upon them. To establish terminology, we define linear functionals, bilinear form, and linear operators.

Definition 2.5 (Linear functional)**.** Let $\mathcal{F}$ be a Hilbert space. A *functional* L is a map from $\mathcal{F}$ to $\mathbb{R}$, and we denote its action on a function f as $L(f)$. A functional is called *linear* if it satisfies $L(f+g) = L(f) + L(g)$ and $L(\lambda f) = \lambda L(f)$, for all $f, g \in \mathcal{F}$ and $\lambda \in \mathbb{R}$.

Definition 2.6 (Bilinear form)**.** Let $\mathcal{F}$ be a Hilbert space. A *bilinear form* B takes inputs $f, g \in \mathcal{F}$ and returns a real value. It is linear in each argument separately, i.e.

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

for all $f, g, h \in \mathcal{F}$ and $\lambda_1, \lambda_2 \in \mathbb{R}$.

Definition 2.7 (Linear 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 $Af \in \mathcal{G}$. A *linear operator* satisfies $A(f+g) = A(f) + A(g)$ and $A(\lambda f) = \lambda A(f)$, for all $f, g \in \mathcal{F}$ and $\lambda \in \mathbb{R}$.

The term 'functional' is classically used in calculus of variations to denote 'a function of a function', i.e. a function having another function as its input, and outputs a real number. Really, from a function space perspective, it is simply a mapping of functions onto another vector space (the reals in this case). More generally, if the output space is another Hilbert space, then it is an operator. An interesting property of these operators to look at, besides linearity, is whether or not they are *continuous*.

Definition 2.8 (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 special case of linear operator, the evaluation functional, this means that a function in $\mathcal{F}$ is continuous if the evaluation functional is continuous—more on this later in Section 2.2. There is an even stronger notion of continuity called the *Lipschitz continuity*.

Definition 2.9 (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 2.8. 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 $\mathcal{F}^*$ 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 $L \in \mathcal{F}'$; since L is finite-dimensional, it is bounded, and therefore continuous (see Lemma 2.1) so $L \in \mathcal{F}'$ and $\mathcal{F}^* \subseteq \mathcal{F}'$; but $\mathcal{F}' \subseteq \mathcal{F}^*$ trivially, so $\mathcal{F}^* \equiv \mathcal{F}'$. 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.

Definition 2.11 (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 $\|A\| := \sup_{f\in\mathcal{F}} \frac{\|Af\|_{\mathcal{G}}}{\|f\|_{\mathcal{F}}}$.

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 a bounded if and only if it is continuous.*

Proof. Suppose that L is bounded. Then, $\forall f, f' \in \mathcal{F}$, there exists some $M > 0$ such that $\|A(f-g)\|_{\mathcal{G}} \leq M\|f-g\|_{\mathcal{G}}$. 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 $f \in \mathcal{F}$,

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

and thus A is bounded. □

The following result is an important one, which states that (continuous) linear functionals of an inner product space are nothing more than just inner products.

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

Proof. Omitted—see Rudin (1987, Theorem 4.12) for a proof. □

Corollary 2.2.1 (Riesz norm)**.** *For any $f \in \mathcal{F}$ a Hilbert space, define $L(f) = \langle f, g\rangle_{\mathcal{F}}$ for some $g \in \mathcal{F}$. Then $\|L\|_{\mathcal{F}'} = \|g\|_{\mathcal{F}}$.*

Proof. By Cauchy-Schwarz,

$$|L(f)| \leq \|f\|_{\mathcal{F}}\|g\|_{\mathcal{F}}$$

so $\|L\|_{\mathcal{F}'} \leq \|g\|_{\mathcal{F}}$. But $|L(g)| = \|g\|_{\mathcal{F}}^2$, so in fact $\|L\|_{\mathcal{F}'} = \|g\|$ □

1. Not so convinced.

The notion of isometry (transformation that preserves distance) is usually associated with metric spaces—two metric spaces being isometric means that they identical in as far as their metric properties are concerned. For Hilbert spaces (or normed spaces in general), 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 2.12 (Isometric isomorphism)**.** Two Hilbert spaces $\mathcal{F}$ and $\mathcal{G}$ are said to be *isometrically isomorphic* if there is a linear bijective map $A : \mathcal{F} \to \mathcal{G}$ which preserves the inner product, i.e.

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

In Hilbert spaces, this isometry is also known as *linear isometry*. A consequence of the Riesz representation theorem is that it gives us a canonical isometric isomorphism $A : f \mapsto \langle \cdot, f\rangle_{\mathcal{F}}$ between $\mathcal{F}$ and its continuous dual $\mathcal{F}'$, whereby $\|Af\|_{\mathcal{F}'} = \|f\|_{\mathcal{F}}$. 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 $\langle f - Pf, g\rangle_{\mathcal{G}} = 0$ for any $g \in \mathcal{G}$. The remainder $f - Pf$ belongs to the *orthogonal complement* of $\mathcal{G}$.

Definition 2.13 (Orthogonal complement)**.** Let $\mathcal{F}$ be a Hilbert space and $\mathcal{G} \subset \mathcal{F}$ be a closed subspace. The linear subspace $\mathcal{G}^{\perp} = \{f \,|\, \langle f, g\rangle_{\mathcal{G}} = 0, \forall g \in \mathcal{G}\}$ is called the orthogonal complement of $\mathcal{G}$.

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 $g = 0$.

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 $\mathcal{G}^\perp = \overline{\mathcal{G}}^\perp = \mathcal{F}^\perp = \{0\}$. □

For the last part of this introductory section on functional analysis, we discuss measures on Hilbert spaces, and in particular, a probability measure. Let $\mathcal{X}$ be a real topological space (e.g. real Hilbert spaces), and let $\mathcal{B}(\mathcal{X})$ the Borel σ-algebra of $\mathcal{X}$. A measure ν on $\big(\mathcal{X}, \mathcal{B}(\mathcal{X})\big)$ is called a *Borel measure* on $\mathcal{X}$. We shall only concern ourselves with finite Borel measures. If $\nu(\mathcal{X}) = 1$ then ν is a *(Borel) probability measure* and the measure space $\big(\mathcal{X}, \mathcal{B}(\mathcal{X}), \nu\big)$ is a *(Borel) probability space*.

Definition 2.14 (Mean vector and covariance operator)**.** Let ν be a Borel probability measure on a real topological space $\mathcal{X}$. Supposing that the function $z \mapsto \langle z, x \rangle_\mathcal{X}$ is integrable with respect to ν, the element $\mu \in \mathcal{X}$ satisfying

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

is called the *mean vector*. If, furthermore, there is a positive, symmetric linear operator C on $\mathcal{X}$ such that

$$\langle Cx, x' \rangle = \int_\mathcal{X} \langle x, z - \mu \rangle_\mathcal{X} \langle x', z - \mu \rangle_\mathcal{X} \, \mathrm{d}\nu(z), \ \forall x, x' \in \mathcal{X},$$

then C is called the *covariance operator*. The conditions requiring existence of the mean vector and covariance operator are $\int_\mathcal{X} |x| \, \mathrm{d}\nu(x) < \infty$ and $\int_\mathcal{X} |x|^2 \, \mathrm{d}\nu(x) < \infty$ respectively.

Definition 2.15 (Mean and covariance of functions)**.** Let $\big(\mathcal{X}, \mathcal{B}(\mathcal{X}), \nu\big)$ be a Borel probability space, and let $\phi : \mathcal{X} \to \mathcal{F}$ be a feature map of some Hilbert space of functions $\mathcal{F}$. The *mean element* of $\mathcal{F}$ is defined as $\mu_f \in \mathcal{F}$ satisfying

$$\mathrm{E}\langle f, f' \rangle_{\mathcal{F}} = \langle \mu_f, f' \rangle_{\mathcal{F}}$$

for all $f' \in \mathcal{F}$. The quantity $\langle \mu_f, f \rangle_{\mathcal{F}} := \mathrm{E}\langle \phi(x), f \rangle_{\mathcal{F}}$ is denoted $\mathrm{E}\, f(X)$.

Slightly confused: Do we need random functions $f \in \mathcal{F}$ or are the covariates $x \in \mathcal{X}$ assumed to be random? Later on in Section 2.4 and 2.5, we talk about $\mathrm{E}\, f(X)$ so there is some measure on $\mathcal{X}$. However, when we prove the I-prior, we talk about f itself being random.

2.2 Reproducing kernel Hilbert space theory

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 RKHS, is to convince 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 $x \in \mathcal{X}$.

Definition 2.16 (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: $\delta_x(\lambda f + g) = (\lambda f + g)(x) = \lambda f(x) + g(x) = \lambda \delta_x(f) + \delta_x(g)$. This is in fact the linearity that was implied earlier on at the beginning of Chapter 2 when introducing the notion of functions behaving like vectors. As a remark, the calculation of the (penalised) likelihood functional involves evaluations. It is therefore important for the evaluation functional to be continuous. It turns out, this is exactly what RKHS provide.

Definition 2.17 (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 $\mathcal{F}$, $\forall x \in \mathcal{X}$.

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 inspired its name.

Definition 2.18 (Reproducing kernels)**.** Let $\mathcal{F}$ be a Hilbert space of functions over a non-empty set $\mathcal{X}$. A function $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called 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 $x, x' \in \mathcal{X}$,

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

Continuity of evaluation functionals in an RKHS means that functions that are close in RKHS norm imply that they are also close pointwise. It gives some reassurance when trying to estimate f from $\mathcal{F}$—just look at the size of the norm. More formally,

3. Not so sure why this is useful?

Corollary 2.3.2 (Norm convergence implies pointwise convergence in RKHS)**.** *Let $\mathcal{F}$ be an RKHS of real functions over $\mathcal{X}$, and let f_n be a sequence of points in $\mathcal{F}$. Then, for some $f \in \mathcal{F}$,*

$$\lim_{n \to \infty} \|f_n - f\|_{\mathcal{F}} = 0 \quad \Rightarrow \quad \lim_{n \to \infty} |f_n(x) - f(x)| = 0.$$

Proof. Suppose $\mathcal{F}$ is an RKHS with reproducing kernel h. Then,

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

□

Having defined an RKHS, there are several questions we might like the answer to: What is the relationship between a reproducing kernel and an RKHS? Can we speak to its existence and uniqueness? What other properties does it have? The rest of this subsection will be dedicated to discuss the following assertion.

Theorem 2.4 (RKHS)**.** *For every reproducing kernel Hilbert space $\mathcal{F}$ of functions over a set $\mathcal{X}$, there corresponds a unique, positive-definite reproducing kernel $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Conversely, for every positive-definite function $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, there corresponds a unique reproducing kernel Hilbert space $\mathcal{F}$ that has h as its reproducing kernel.*

In essence, there is a bijection between the set of positive-definite kernels and the set of reproducing kernel Hilbert spaces. We will take take apart this theorem and inspect its constituent claims. Firstly, on the definition of kernels and its positive-definiteness.

Definition 2.19 (Kernels)**.** Let $\mathcal{F}$ be a Hilbert space (not necessarily a RKHS), $\mathcal{X}$ a non-empty set, and $\phi : \mathcal{X} \to \mathcal{F}$. A *kernel* is defined to be function $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ that satisfies

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

$\forall x, x' \in \mathcal{X}$. The map ϕ is referred to as the *feature map*, and $\mathcal{F}$ the *feature space*.

Lemma 2.5 (Positive-definiteness of kernels)**.** *The kernel as defined in Definition 2.19 is a symmetric and positive definite function, where a symmetric function $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is said to be positive definite if*

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

for all integers $n > 1$, $\forall a_1, \ldots, a_n \in \mathbb{R}$, and $\forall x_1, \ldots, x_n \in \mathcal{X}$.

Proof.

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

□

Corollary 2.5.1 (Positive-definiteness of reproducing kernels)**.** *Reproducing kernels of a RKHS are positive definite.*

Proof. Take $\phi : x \mapsto h(\cdot, x)$. By Definition 2.19, one has $h(x, x') = \langle h(\cdot, x), h(\cdot, x')\rangle_{\mathcal{F}}$, which is the reproducing property of the kernel in a RKHS, and this is positive-definite by Lemma 2.5. Incidentally, the ϕ as defined is known as the *canonical feature map.* □

We have established what a kernel is, and that reproducing kernels of an RKHS are positive-definite. But do reproducing kernels always exist, and if so, are they unique to an RKHS? Lemmas 2.6 and 2.7 answer these questions in the positive.

Lemma 2.6 (Existence of reproducing kernels)**.** *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.*

Proof. Suppose $\mathcal{F}$ is a RKHS with kernel h. Choose $\delta = \epsilon/\|h(\cdot, x)\|_{\mathcal{F}}$. Then, for any $f \in \mathcal{F}$ such that $\|f - g\|_{\mathcal{F}} < \delta$, we have

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

Thus, the evaluation functional is (uniformly) continuous on $\mathcal{F}$. To prove the reverse, follow the argument preceding Definition 2.18. □

Lemma 2.7 (Uniqueness of reproducing kernels)**.** *The reproducing kernel $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ of a RKHS $\mathcal{F}$ of functions over $\mathcal{X}$ is unique.*

Proof. Assume that $\mathcal{F}$ 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 $\|h_1(\cdot, x) - h_2(\cdot, x)\|_{\mathcal{F}}^2 = 0$ □

Naturally, having seen that every RKHS corresponds to a unique reproducing kernel, we ask whether the converse is true. That is, given a reproducing kernel, does it define a unique RKHS? Astoundingly, the answer is again positive, and this is stated by the much celebrated Moore-Aronszajn theorem below.

Theorem 2.8 (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 an RKHS. For a complete proof, see Berlinet and Thomas-Agnan (2011). 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 indeed a valid inner-product satisfying the conditions laid in Definition 2.1. At this point, the reproducing property is already had:

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

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 $x \in \mathcal{X}$,

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

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 $(w_1, \ldots, w_n) \in \mathbb{R}^n$, $n \in \mathbb{N}$. 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 (2.2).

2.3 Reproducing kernel Kreĭn space theory

Kreĭn spaces can be seen as a generalisation of Hilbert spaces, which caters for inner products not being positive definite. To motivate the need for Kreĭn spaces, we first look at several operations on reproducing kernels and the resulting vector space.

Lemma 2.9 (Scaling of kernels)**.** *If h is a kernel on $\mathcal{X}$, and $\lambda \geq 0$ a scalar, then λh is a kernel. 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.*

Proof. Multiplying a positive definite function by a positive constant results in a positive definite function still, and thus defines a unique RKHS. The scaling of functions is seen through the fact that $\mathcal{F}$ is the completion of the space spanned by the kernels, and hence $\mathcal{F}_\lambda$ by the scaled kernels. □

Lemma 2.10 (Sum of kernels)**.** *If h_1 and h_2 are kernels on $\mathcal{X}_1$ and $\mathcal{X}_2$ respectively, then $h = h_1 + h_2$ is a 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 an RKHS defined by $h = h_1 + h_2$, where*

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

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 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). □

Lemma 2.11 (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 an RKHS with reproducing kernel h.*

To prove this lemma we introduce the following definition of the *kernel matrix*.

Definition 2.20 (Kernel matrix)**.** Let $\{x_1, \ldots, x_n\}$ be a sample of observed data (covariates), where each $x_i \in \mathcal{X}$, and h a kernel over this set. Define the *kernel matrix* $\mathbf{H}$ for h as the $n \times n$ matrix with (i, j) entries equal to $h(x_i, x_j)$.

The kernel matrix is also known as the *Gram matrix*. By definition, the kernel matrix is a positive-definite matrix: $\mathbf{a}^\top \mathbf{H}\mathbf{a} = \sum_{i=1}^n \sum_{j=1}^n a_i a_j h(x_i, x_j) \geq 0$ for any choice of $a \in \mathbb{R}^n$.

Proof of Lemma 2.11. 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, we can write $\mathbf{H}_1 = \mathbf{B}^\top \mathbf{B}$ and $\mathbf{H}_1 = \mathbf{C}^\top \mathbf{C}$ for some matrices $\mathbf{B}$ and $\mathbf{C}$—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

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

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

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

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 2.9–2.11: 1) Multiplying a positive definite matrix by a positive constant results in a positive definite matrix; 2) the addition of positive definite matrices is a positive definite matrix; and 3) the *Hadamard product*[3] of two positive definite matrices is a positive definite matrix.

[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, $(\mathbf{A} \circ \mathbf{B})_{ij} = \mathbf{A}_{ij} \mathbf{B}_{ij}$.

What if we want to consider scaling or sum or products of kernels that result in a non-positive definite function? Will we lose all the nice properties of RKHS? It turns out no—we just have to switch to Kreĭn spaces. Krein spaces are inner product spaces endowed with Hilbertian topology, yet their inner products are no longer positive.

Definition 2.21 (Negative and indefinite inner products)**.** If instead of the positive-definiteness condition for inner products, as per the last item in Definition 2.1, we have that $\forall f \in \mathcal{F}$

$$\langle f, f \rangle_{\mathcal{F}} \leq 0,$$

then the inner product is said to be *negative-definite*. It is *indefinite* if is is neither positive- nor negative-definite.

Definition 2.22 (Krein space)**.** An inner product space $(\mathcal{F}, \langle \cdot, \cdot \rangle_{\mathcal{F}})$ is a *Krein space* if there exists two Hilbert spaces $\mathcal{F}_+$ and $\mathcal{F}_-$ spanning $\mathcal{F}$ such that

- All $f \in \mathcal{F}$ can be decomposed into $f = f_+ + f_-$, where $f_+ \in \mathcal{F}_+$ and $f_- \in \mathcal{F}_-$.
- $\forall f, f' \in \mathcal{F}$, $\langle f, f' \rangle_{\mathcal{F}} = \langle f_+, f'_+ \rangle_{\mathcal{F}_+} - \langle f_-, f'_- \rangle_{\mathcal{F}_-}$

This suggests that there is an "associated" Hilbert space.

Definition 2.23 (Associated Hilbert space)**.** Let $\mathcal{F}$ be a Krein space with decomposition into Hilbert spaces $\mathcal{F}_+$ and $\mathcal{F}_-$. Denote by $\overline{\mathcal{F}}$ the associated Hilbert space defined by

$$\overline{\mathcal{F}} = \mathcal{F}_+ \oplus \mathcal{F}_-,$$

and hence $\langle f, f' \rangle_{\overline{\mathcal{F}}} = \langle f_+, f'_+ \rangle_{\mathcal{F}_+} + \langle f_-, f'_- \rangle_{\mathcal{F}_-}$. Likewise,

$$\overline{\mathcal{F}} = \mathcal{F}_+ \ominus \mathcal{F}_-,$$

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

Definition 2.24 (Reproducing kernel Krein 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.

$\overline{\mathcal{F}}$ is the smallest Hilbert space majorizing the Krein space $\mathcal{F}$. The strong topology on $\mathcal{F}$ is the Hilbertian topology of $\overline{\mathcal{F}}$. The topology does not depend on the decomposition chosen. Clearly $|\langle f, f \rangle_{\mathcal{F}} \leq \|f\|^2_{\overline{\mathcal{F}}}|$ for all $f \in \mathcal{F}$.

For every reproducing kernel Krein space $\mathcal{F}$ of functions over a set $\mathcal{X}$, there corresponds a unique, positive-definite reproducing kernel $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$.

Lemma 2.12. *Let $\mathcal{F}$ be an RKKS with $\mathcal{F} = \mathcal{F}_+ \ominus \mathcal{F}_-$. Then,*

- *$\mathcal{F}_+$ and $\mathcal{F}_-$ are both RKKS with kernel h_+ and h_-.*
- *There is a unique symmetric $h(x, x')$ with $h(\cdot, x) \in \mathcal{F}$ such that for all $f \in \mathcal{F}$,*

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

- *$h = h_+ + h_-$.*

On the other hand, for every positive-definite function $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, there corresponds at least one reproducing kernel Hilbert space $\mathcal{F}$ that has h as its reproducing kernel.

Lemma 2.13. *The following are equivalent.*

- *There exists (at least) one RKKS with kernel h.*
- *h admits a positive decomposition, i.e., there exists two positive kernels h_+ and h_- such that $h = h_+ + h_-$.*
- *h is dominated by some kernel k, i.e. $h - k$ is a positive kernel.*

There is no bijection but a surjection between the set of RKKS and the set of generalized kernels defined in the vector space generated out of the cone of positive kernels.

2.4 RKHS building blocks

In what follows, each of the kernel functions will have its associated scale parameter denoted by λ. Further, to make the distinction between centred and non-centred versions of the kernels, we use the notation h to denote the uncentred version, and $\bar{h}$ to denote the centred version.

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 lemma.

Proposition 2.14 (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, $\|f\|_{\mathcal{F}} = \sqrt{\langle f, f \rangle_{\mathcal{F}}} = |c_f|$. □

In I-prior modelling, one need not consider any scale parameter on reproducing kernel, as the scale parameter would not be identified otherwise. See later chapter for details. I think the scale parameter λ would just be absorbed by the norm, which is a single value of interest and that is what is "observed", and the decomposition $\lambda \cdot c_f$ is not so interesting.

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 $f_\beta = \langle \cdot, \beta \rangle_{\mathcal{X}}$. We can show that the continuous dual space of $\mathcal{X}$ is a RKHS which consists of these linear functions.

Proposition 2.15 (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 $\mathcal{V}$ is a Hilbert space, 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).

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 2.25 (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 (i.e. $\mathrm{E}\langle x, x' \rangle_{\mathcal{X}} = \langle \mu, x' \rangle_{\mathcal{X}}$ for all $x' \in \mathcal{X}$) with respect to this probability measure. 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

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

Proof. Proof of the claim $\mathrm{E}\, f_\beta(x) = 0$:

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

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. □

Remark 2.1. In practice, the probability measure P over $\mathcal{X}$ is unknown, so we may use the empirical distribution over $\mathcal{X}$, 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 (also known as the Wiener process) has been an inquisitive subject in the mathematical sciences, and here, we describe a function space influenced 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 $t_1, \ldots, t_k$, where each $t_j \in [0, T]$, $\big(B_\gamma(t_1), \ldots, B_\gamma(t_k)\big)$ 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 or Hurst parameter. 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, then the fBm is a standard Brownian motion and its increments are independent; when $\gamma > 1/2$ ($\gamma < 1/2$) its increments are positively (negatively) correlated.

Let $\mathcal{X}$ be a Hilbert space. Defining a kernel function $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ identical to the fBm covariance kernel yields the so-called *fractional Brownian motion RKHS*.

Definition 2.26 (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 $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defined by

$$h(x, x') = \frac{1}{2}\big(\|x\|_{\mathcal{X}}^{2\gamma} + \|x'\|_{\mathcal{X}}^{2\gamma} - \|x - x'\|_{\mathcal{X}}^{2\gamma}\big),$$

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

Remark 2.2. 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 2.27 (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.

Claim 2.16. *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 $x, x' \in \mathcal{X}$,

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

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 claim. □

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.

5. This is the same for any RKHS?

Definition 2.28 (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 (i.e. $\mathrm{E}\langle x, x'\rangle_{\mathcal{X}} = \langle \mu, x'\rangle_{\mathcal{X}}$ for all $x' \in \mathcal{X}$) with respect to this probability measure. The kernel $\bar{h} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defined by

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

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 $X \in \mathcal{X}$.

6. Proof?

Remark 2.3. Again, when $\gamma = 1$, we get the reduction

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

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

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 2.29 (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 $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$

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

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 $\gamma = 1/2l^2$ is utilised. Duvenaud (2014) remarks that "exponentiated quadratic" is a more fitting 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.

Are SE smoother than fBm? Lipschitz continuous. Compact convergence. May be smoother than functions in an fBm RKHS?

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

Definition 2.30 (Universal kernel)**.** Let $\mathrm{C}(\mathcal{X})$ is the space of all continuous, complex-valued functions $f : \mathcal{X} \to \mathbb{C}$ equipped with the maximum norm $\|\cdot\|_\infty$, and denote $\mathcal{K}(\mathcal{X})$ 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 $\hat{f}$ from the SE RKHS, and these two functions can get arbitrarily close to each other in the max norm sense. This, together with some very convenient computational advantages that the SE kernel brings (more on this in a later chapter), 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 2.31 (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 $\mathcal{X}$. The *centred* squared exponential (SE) RKHS (with lengthscale l) of functions over $\mathcal{X}$ is defined by the positive definite kernel $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$

$$h(x, x') = \exp\left(-\frac{\langle x - \mu, x' - \mu\rangle}{2l^2}\right),$$

where $\mu =: \mathrm{E}\, X \in \mathcal{X}$ under P, and $\langle \cdot, \cdot \rangle$ represents the usual dot product in Euclidean space. This ensures that $\mathrm{E}\, f(X) = 0$ for any f in this RKHS.

8. Proof?

2.4.5 The Pearson RKHS

In all of the previous RKHS 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.

Definition 2.32 (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 $\mathrm{P}(X = x)$.

Claim 2.17. *Let $\mathcal{F}$ be the Pearson RKHS of functions over a finite set $\mathcal{X}$. Then,*

$$\mathcal{F} = \{f : \mathcal{X} \to \mathbb{R} | \, \mathrm{E}\, f(X) = 0\}$$

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 $\{h(\cdot, x) | x \in \mathcal{X}\}$ 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

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

and thus $h(\cdot, x)$ spans $\mathcal{F}$ so $f \in \mathcal{F}$. To provide the second part, 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 $f(X)$. □

2.5 Constructing RKKS from existing RKHS

The previous section outlined all of the basic RKHSs of functions that will form the building blocks when constructing more complex function spaces. As previously mentioned in the preliminaries, sums of kernels are kernels and products of kernels are also kernels, and thus in the context of RKHS we may construct new RKHS 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 reproducing kernel Krein spaces.

2.5.1 Scaling an RKHS

The scale of an 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, resulting in the RKHS denoted $\mathcal{F}_\lambda$ with kernel λh. The scale λ will typically need to be estimated from the data.

Restricting λ to the positive reals may be arbitrary and restrictive; in particular, we shall see when constructing new function spaces this positive restriction may turn out to be unsatisfactory. Without the positive restriction, the kernel may potentially be negative-definite. Therefore, the subsequent sections speak of RKKSs, instead of RKHSs, to account solely for the fact that λ may be negative. All other properties of RKHSs should carry over to RKKSs, so sometimes we might overlook this distinction, and make references to RKHSs when instead RKKSs would be more suited to the context.

Remark 2.4. As it turns out, for I-prior modelling, in cases where the RKHS is $\mathcal{F}_\lambda$ with kernel λh, then the sign of the single scale parameter λ is unidentified. Therefore, in such cases, we may restrict $\lambda \in \mathbb{R}^+$. More on this in Chapter 4.

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 2.33 (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

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

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 defined 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 an 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.

Claim 2.18. *The polynomial RKKS of functions over* $\mathbb{R}$*, denoted* $\mathcal{F}$*, 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 $g \in \mathcal{F}$ can therefore be written as $g = \beta_0 + \sum_{i=1}^{d} \prod_{j=1}^{i} f_j$, and $f_j(x) = b_j x$ from before, where b_j is a constant. Therefore, $g(x) = \sum_{k=0}^{d} \beta_k x^k$. □

Remark 2.5. 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 main bibliographical references for this subsection is as follows. 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 machine learning (e.g. Durrande et al., 2013), applied mathematics (e.g. Kuo et al., 2010), and sensitivity analysis (e.g. Sobol, 2001). What is interesting is that several authors who simply set out to find a suitable decomposition of a function, ended up somewhat independently recovering the ANOVA decomposition as being "optimal" in some sense. This speaks largely to this classical idea that is ANOVA.

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 $y_j = \{y_{1j}, \dots, y_{n_j j}\}$ for $j = 1, \dots, m$, 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 m *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: 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 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 the identifiability 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:

- $Af(x) = f(1) = \mu_1$. 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 $\mu_1 = 0$.
- $Af(x) = \sum_{x=1}^{m} f(x)/m =: \bar{\alpha}$. 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 [Side note: This idempotent property of the linear operator A on f speaks to the possibility of it being somewhat like an *orthogonal projection*, and indeed this is so—we shall return to this point later when we describe functional ANOVA decomposition]. We must have that $Af_t = A(I-A)f = Af - A^2 f = 0$. In other words, the choice of A is arbitrary, just like the choice of restriction, so long as it satisfies the condition that $Af_c = 0$.

The multiway ANOVA can be motivated in a similar light. Let $x = (x_1, \ldots, x_p) \in \prod_{k=1}^{p} \mathcal{X}_k$, and consider functions that map $\prod_{k=1}^{p} \mathcal{X}_j$ to $\mathbb{R}$. Let A_j be an averaging operator on $\mathcal{X}_k$ that averages the kth 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

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

where we had denoted $\mathcal{P}_p = \mathcal{P}(\{1, \ldots, p\})$ to be the power set of $\{1, \ldots, p\}$ 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 $A_k f_{\mathcal{K}} = 0, \forall k \in \mathcal{K} \in \mathcal{P}_p$.

Example 2.1 (Two-way ANOVA decomposition)**.** Let $p = 2$, $\mathcal{X}_1 = \{1, \ldots, m_1\}$, and $\mathcal{X}_2 = \{1, \ldots, m_2\}$. The power set $\mathcal{P}_2$ is $\big\{\{\}, \{1\}, \{2\}, \{1, 2\}\big\}$. The ANOVA decomposi-

tion of f is

$$f = f_0 + f_1 + f_2 + f_{12}.$$

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,

$$\begin{aligned} f_0 &= A_1 A_2 f && = f(1,1) \\ f_1 &= (I - A_1) A_2 f && = f(x_1,1) - f(1,1) \\ f_2 &= A_1 (I - A2) f && = f(1,x_2) - f(1,1) \\ f_{12} &= (I - A_1)(I - A2) f && = f(x_1,x_2) - f(x_1,1) - f(1,x_2) + f(1,1). \end{aligned}$$

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

$$\begin{aligned} f_0 &= A_1 A_2 f && = f_{\cdot\cdot} \\ f_1 &= (I - A_1) A_2 f && = f_{x_1 \cdot} - f_{\cdot\cdot} \\ f_2 &= A_1 (I - A_2) f && = f_{\cdot x_2} - f_{\cdot\cdot} \\ f_{12} &= (I - A_1)(I - A_2) f && = f - f_{x_1 \cdot} - f_{\cdot x_2} + f_{\cdot\cdot}, \end{aligned}$$

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

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 $\mathcal{F}$. Specifically, we shall consider the (Hilbert) space of square integrable functions over $\mathcal{X}$ with measure ν, $\mathrm{L}^2(\mathcal{X}, \nu) \equiv \mathcal{F}$. We shall jump straight into the multiway ANOVA analogue for functional decomposition, and to that end, consider $x = (x_1, \ldots, x_p) \in \prod_{k=1}^{p} \mathcal{X}_k =: \mathcal{X}$, where each of the spaces $\mathcal{X}_k$ has measure ν_k, and thus $\nu = \nu_1 \otimes \cdots \otimes \nu_d$. As $\mathcal{X}$ need not necessarily be a (collection of) finite set, 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

$$A_k f = \int_{\mathcal{X}_k} f(x_1, \ldots, x_p) d\nu(x_k). \tag{2.2}$$

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

$$f = \overbrace{\int_{\mathcal{X}} f(x)\,\mathrm{d}\nu(x)}^{f_0} + \overbrace{\left(f - \int_{\mathcal{X}} f(x)\,\mathrm{d}\nu(x)\right)}^{f_1} \tag{2.3}$$

and for the two-way ANOVA ($k = 2$), we have $f = f_0 + f_1 + f_2 + f_{12}$, with

$$\begin{aligned}
f_0 &= \int_{\mathcal{X}_1}\int_{\mathcal{X}_2} f(x_1, x_2)\,\mathrm{d}\nu(x_1)\,\mathrm{d}\nu(x_2) \\
f_1 &= \int_{\mathcal{X}_2}\left(f(x_1, x_2) - \int_{\mathcal{X}_1} f(x_1, x_2)\,\mathrm{d}\nu(x_1)\right)\mathrm{d}\nu(x_2) \\
f_2 &= \int_{\mathcal{X}_1}\left(f(x_1, x_2) - \int_{\mathcal{X}_2} f(x_1, x_2)\,\mathrm{d}\nu(x_2)\right)\mathrm{d}\nu(x_1) \\
f_{12} &= f(x_1, x_2) - \int_{\mathcal{X}_1} f(x_1, x_2)\,\mathrm{d}\nu(x_1) - \int_{\mathcal{X}_2} f(x_1, x_2)\,\mathrm{d}\nu(x_2) \\
&\quad + \int_{\mathcal{X}_1}\int_{\mathcal{X}_2} f(x_1, x_2)\,\mathrm{d}\nu(x_1)\,\mathrm{d}\nu(x_2).
\end{aligned}$$

As a remark, the averaging operator A_k defined in (2.2) is indeed true to its name, in that it calculates the mean function of f over the kth coordinate. For comparison, this is identical to the second type of restriction we considered in the classical ANOVA previously (i.e., setting $\sum_j \mu_j = 0$). We must also have, as before, that $A_k f_{\mathcal{K}} = 0, \forall k \in \mathcal{K} \in \mathcal{P}_p$. For the one-way functional ANOVA decomposition in (2.3), it must be that f_1 is a zero-mean function. As for the two-way ANOVA, it is the case that $\int_{\mathcal{X}_k} f_1(x_1, x_2)\,\mathrm{d}\nu(x_k) = 0, k = 1, 2$, and $\int_{\mathcal{X}_1}\int_{\mathcal{X}_2} f_{12}(x_1, x_2)\,\mathrm{d}\nu(x_1)\,\mathrm{d}\nu(x_1) = 0$.

We notice that the decomposition in (2.3) is orthogonal:

Claim 2.19. *For the ANOVA decomposition in* (2.3)*, f_0 and f_1 are orthogonal for the usual L^2 inner product.*

Proof. Note that f_0 is a constant function, and that $f_1 = f - f_0$. Thus,

$$\begin{aligned}\langle f_0, f_1 \rangle &= \int f_0 f_1 \, \mathrm{d}\nu \\ &= f_0 \int (f - f_0) \, \mathrm{d}\nu \\ &= f_0(f_0 - f_0) = 0.\end{aligned}$$

□

In fact, for $k = 1$, any $f \in \mathcal{F}$ can be decomposed as a sum of a constant plus a zero mean function, so we have the geometric decomposition of the vector space $\mathcal{F} = \mathcal{F}_0 \overset{\perp}{\oplus} \bar{\mathcal{F}}_1$, where $\mathcal{F}_0$ is a vector space of constant functions, and $\bar{\mathcal{F}}_1$ a vector space of zero-mean functions over $\mathcal{X}_1$. For $k \geq 2$ we can argue something similar. The space $\mathcal{F}$ has the tensor product structure[4] $\mathcal{F} = \mathcal{F}_1 \otimes \cdots \otimes \mathcal{F}_p$, and considered individually, each $\mathcal{F}_k$ can be decomposed orthogonally $\mathcal{F}_k = \mathcal{F}_0 \overset{\perp}{\oplus} \bar{\mathcal{F}}_k$. Note that $\mathcal{F}_k$ consists of functions $f : \mathcal{X}_k \to \mathbb{R}$. Expanding out under the distributivity rule of tensor products and rearranging slightly, we obtain

$$\begin{aligned}\mathcal{F} &= \left(\mathcal{F}_0 \overset{\perp}{\oplus} \bar{\mathcal{F}}_1\right) \otimes \cdots \otimes \left(\mathcal{F}_0 \overset{\perp}{\oplus} \bar{\mathcal{F}}_1\right) \\ &= \mathcal{F}_0^{\otimes p} \overset{\perp}{\oplus} \bigoplus_{j=1}^{p}{}^{\perp} \left(\mathcal{F}_0^{\otimes(p-1)} \otimes \bar{\mathcal{F}}_j\right) \overset{\perp}{\oplus} \bigoplus_{\substack{j,k=1 \\ j<k}}^{p}{}^{\perp} \left(\mathcal{F}_0^{\otimes(p-2)} \otimes \bar{\mathcal{F}}_j \otimes \bar{\mathcal{F}}_k\right) \\ &\quad \overset{\perp}{\oplus} \cdots \overset{\perp}{\oplus} \left(\bar{\mathcal{F}}_1 \otimes \cdots \otimes \bar{\mathcal{F}}_p\right).\end{aligned} \tag{2.4}$$

To clarify,

- $\mathcal{F}_0^{\otimes p}$ is the space of constant functions $f : \mathcal{X}_1 \times \cdots \times \mathcal{X}_p \to \mathbb{R}$.
- $\left(\mathcal{F}_0^{\otimes(p-1)} \otimes \bar{\mathcal{F}}_j\right)$ is the space of functions that are constant on all coordinates except the jth coordinate of x. Further, the functions are centred on the jth coordinate.
- $\left(\mathcal{F}_0^{\otimes(p-2)} \otimes \bar{\mathcal{F}}_j \otimes \bar{\mathcal{F}}_k\right)$ is the space of functions that are constant on all coordinates except the jth and kth coordinate of x. Further, 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 $f : \mathcal{X}_1 \times \cdots \times \mathcal{X}_p \to \mathbb{R}$.
- Similarly 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 orthogonal spaces in (2.4) leads to the *functional ANOVA representation*

$$f(x) = \mu + \sum_{j=1}^{p} f_j(x_j) + \sum_{\substack{j,k=1 \\ j<k}}^{p} f_{jk}(x_j, x_k) + \cdots + f_{1\cdots p}(x). \tag{2.5}$$

Definition 2.34 (Functional ANOVA representation)**.** Let $\mathcal{P}_d = \mathcal{P}(\{1, \ldots, d\})$, the power set of $\{1, \ldots, d\}$. For any function $f \in \mathcal{F} \equiv \mathrm{L}^2(\mathcal{X}_1 \times \cdots \times \mathcal{X}_d, \nu_1 \otimes \cdots \otimes \nu_d)$, the formula for f in (2.5) is known as the *functional ANOVA representation* of f if $\forall k \in \mathcal{K} \in \mathcal{P}_p$,

$$A_k f_{\mathcal{K}} = \int_{\mathcal{X}_{\mathcal{K}}} f_{\mathcal{K}}(x_{\mathcal{K}}) \, \mathrm{d}\nu_k(x_k) = 0, \tag{2.6}$$

where $\mathcal{X}_{\mathcal{K}} = \prod_{k \in \mathcal{K}} \mathcal{X}_k$, and $x_{\mathcal{K}} = \{x_k, k \in \mathcal{K}\}$ is an element of this space. In other words, the integral of $f_{\mathcal{K}}$ with respect to any of the variables indexed by the elements in $\mathcal{K}$ (itself an element of the power set), is zero. The requirement (2.6) ensures orthogonality of the summands in (2.5).

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

$$\begin{aligned} f_0 &= \int f \, \mathrm{d}\nu \\ f_j &= \int f \, \textstyle\prod_{i \neq j} \mathrm{d}\nu_i - f_0 \\ f_{jk} &= \int f \, \textstyle\prod_{i \neq j,k} \mathrm{d}\nu_i - f_j - f_k - f_0. \end{aligned}$$

Remark 2.6. Not all of the higher order terms need to be included. There may even be a model motivated reason for dropping certain main effects or interaction effects.

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 be not in vain. The main idea is to construct an

[4] There is an isomorphism $\mathrm{L}^2(\mathcal{X}_1 \times \cdots \times \mathcal{X}_d, \nu_1 \otimes \cdots \otimes \nu_d) \cong \mathrm{L}^2(\mathcal{X}_1, \nu_1) \otimes \cdots \otimes \mathrm{L}^2(\mathcal{X}_d, \nu_d)$. See, for example, Reed and Simon (1972) and Krée (1974).

RKKS such that the functions that lie in them will have the ANOVA representation in (2.5). 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 (2.4)). The ANOVA RKKS will be constructed by a similar manipulation of the individual kernels representing the RKHS building blocks.

Definition 2.35 (The ANOVA RKKS)**.** For $k = 1, \ldots, p$, let $\mathcal{F}_k$ be a centred RKHS of functions over the set $\mathcal{X}_k$ with kernel $h_k : \mathcal{X}_k \times \mathcal{X}_k \to \mathbb{R}$. Let $\lambda_k, k = 1, \ldots, p$ 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

$$h_\lambda(x, x') = \prod_{k=1}^{p} \big(1 + \lambda_k h_k(x_k, x'_k)\big). \tag{2.7}$$

The construction an ANOVA RKKS is very very simple in through multiplication of univariate kernels. Expanding out equations (2.7), we see that it is in fact a sum of separable 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 the expansion that the ANOVA RKKS yields functions that resemble those with the ANOVA representation in (2.5): The mean value of the function stems from the '1', i.e. it lies in an 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.

One thing to note is that restricting the λ parameters to the positive orthant might give unsatisfactory results—what if the effect of two functions are in truth opposing one another? These are handled through opposing signs of their respective scale parameters, thus the need for working in RKKSs.

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 remarked in the previous subsection, not all of the components of the ANOVA RKKS need to be included in 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. One thing's for sure—everything is ANOVA.

Remark 2.7. Unfortunately, even if centred RKHSs are used as the building blocks of the ANOVA RKKS, the properties of the function represented by (2.5) may not be preserved. In particular, any of the individual functions $f_{\mathcal{K}}$, for $\mathcal{K}$ in the power set, are not necessarily zero mean functions. Furthermore, any two terms in the summand are generally not orthogonal. Consequently, interpretation based on an ANOVA motivation may not be valid, but in spirit, they provide a conceptually strong basis for building new RKKSs from existing ones.

9. What are advantages of using ANOVA? I feel this has not been addressed properly.

2.6 Summary

Brief notes on functional analysis allows us to describe RKHS and RKKS. These are of interest because of the convenient topologies. All linear functionals are continuous in these spaces. Moreover, RKHS and RKKS can be specified through kernel functions. Although it is unique for RKHS, not so for RKKS, but need not matter greatly, because the important properties that we care about carry over anyway. Of importance is the ANOVA functional decomposition, for which we can mimic through similar manipulation of scaled positive definite kernels. Restricting to positive scalars is unsatisfactory, so RKKS is required.

Bibliography

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

Casella, G. and R. L. Berger (2002). *Statistical inference*. Vol. 2. Duxbury Pacific Grove, CA.

Cohen, S. (2002). "Champs localement auto-similaires". In: *Lois d'échelle, fractales et ondelettes*. Ed. by P. Abry, P. Gonçalves, and J. L. Véhel. Vol. 1. Hermès Sciences Publications.

Dean, A. and D. Voss (1999). *Design and analysis of experiments*. Vol. 1. Springer.

Durrande, N., D. Ginsbourger, O. Roustant, and L. Carraro (2013). "ANOVA kernels and RKHS of zero mean functions for model-based sensitivity analysis". In: *Journal of Multivariate Analysis* 115, pp. 57–67.

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

Embrechts, P. and M. Maejima (2002). *Selfsimilar Processes. Princeton series in applied mathematics*. Princeton University Press, Princeton, NJ.

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

Krée, P. (1974). "Produits tensoriels complétés d'espaces de Hilbert". In: *Séminaire Paul Krée* 1.7, pp. 1974–1975.

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

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

Micchelli, C. A., Y. Xu, and H. Zhang (2006). "Universal kernels". In: *Journal of Machine Learning Research* 7.Dec, pp. 2651–2667.

Reed, M. and B. Simon (1972). *Methods of mathematical physics I: Functional analysis*.

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

Sobol, I. 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.

Steinwart, I. and A. Christmann (2008). *Support vector machines*. Springer Science & Business Media.

Steinwart, I., D. Hush, and C. 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.

List of Figures

List of Tables

List of Theorems

List of Definitions

List of Abbreviations

ANOVA	Analysis of variance.
SE	Squared exponential (kernel).

Index