To-do list

Contents

Haziq Jamil
Department of Statistics
London School of Economics and Political Science
February 11, 2018

Chapter 2

Reproducing kernel Krein spaces

This chapter provides a concise review of functional analysis, especially on topic of reproducing kernel Hilbert and Krein spaces. In addition, this chapter also describes several reproducing kernel Hilbert space (RKHSs) of interest for the purpose of I-prior modelling. Choosing the appropriate RKHS allows us to fit various models of interest. In I-prior modelling, the kernel defining the RKHS turn out to be negative. In such a case, it is necessary to consider *Krein spaces*, in order to give us the required mathematical platform for I-prior modelling. Krein spaces are simply a generalisation of Hilbert spaces for which the kernels allowed to be non-positive definite it its reproducing kernel space. It is emphasised that a deep knowledge of functional analysis is not necessary for I-prior modelling; the advanced reader may wish to skip Sections 2.1–2.3. Section 2.4 describes the RKHSs and RKKSs of interest.

2.1 Preliminaries

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 it is known, are sets for which its elements adhere to a set of rules (axioms) relating to additivity and multiplication by a constant. Additionally, vector spaces are 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.

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

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:** $\langle f, f \rangle_{\mathcal{F}} = 0 \Leftrightarrow f = 0$
- **Positive-definiteness:** $\langle f, f \rangle_{\mathcal{F}} \geq 0$, $\forall f \in \mathcal{F}$

We can always define a *norm* on $\mathcal{F}$ using the inner product as $\|f\|_{\mathcal{F}} = \sqrt{\langle f, f \rangle_{\mathcal{F}}}$. Norms are another form of structure that specifically describes 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:** $||\lambda f||_{\mathcal{F}} = |\lambda| \, ||f||_{\mathcal{F}}$, $\forall \lambda \in \mathbb{R}$, $\forall f \in \mathcal{F}$
- **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 norm $|| \cdot ||_{\mathcal{F}}$ induces a metric (a notion of distance) on $\mathcal{F}$: $d(f, g) = ||f - g||_{\mathcal{F}}$. 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 $\|f\|_{\mathcal{F}} + \|-f\|_{\mathcal{F}} \geq \|f - f\|_{\mathcal{F}} = \|0\|_{\mathcal{F}} = 0$, which implies non-negativity of norms.

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, but not always the other way around. With these notions of distances we can then define *Cauchy sequences.* A sequence is said to be Cauchy if the elements of the sequence become arbitrarily close to one another as the sequence progresses.

Definition 2.3 (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 $\forall n, m > N$, $||f_n - f_m||_{\mathcal{F}} < \epsilon$.

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.*

A vector space equipped with a (positive definite) inner product that is also complete is known as a *Hilbert space*. Out of interest, an incomplete inner product space is known as a *pre-Hilbert space*, since its completion with respect to the norm induced by the inner product is a Hilbert space. A complete normed space is called a *Banach space*.

The next few definitions are introduced as a necessary precursor to defining a reproducing kernel Hilbert space. Firstly,

For a space of functions $\mathcal{F}$ on $\mathcal{X}$, we define the evaluation functional that assigns a value to $f \in \mathcal{F}$ for each $x \in \mathcal{X}$.

Definition 2.4 (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 function $\delta_x : \mathcal{F} \to \mathbb{R}$ as defined by $\delta_x(f) = f(x)$ is called the (Dirac) evaluation functional at x. Evaluation functionals are always linear.

There are two more concepts that we need to cover before defining a reproducing kernel Hilbert/Krein space.

Definition 2.5 (Linear operator)**.** A function $A : \mathcal{F} \to \mathcal{G}$, where $\mathcal{F}$ and $\mathcal{G}$ are both normed vector spaces over $\mathbb{R}$, is called a linear operator if and only if it satisfies the following properties:

- **Homogeneity**: $A(af) = aA(f)$, $\forall a \in \mathbb{R}$, $\forall f \in \mathcal{F}$
- **Additivity**: $A(f + g) = A(f) + A(g)$, $\forall f \in \mathcal{F}, g \in \mathcal{G}$.

Definition 2.6 (Bounder operator)**.** The linear operator $A : \mathcal{F} \to \mathcal{G}$ between two normed spaces $(\mathcal{F}, ||\cdot||_\mathcal{F})$ and $(\mathcal{G}, ||\cdot||_\mathcal{G})$ is said to be a bounded operator if $\exists \lambda \in [0, \infty)$ such that

$$||A(f)||_\mathcal{G} < \lambda ||f||_\mathcal{F}.$$

Now we define a reproducing kernel Hilbert space.

Definition 2.7 (Reproducing kernel Hilbert space)**.** A Hilbert space 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 bounded (equivalently, continuous[2]), i.e. $\exists \lambda_x \geq 0$ such that $\forall f \in \mathcal{F}$,

$$|f(x)| = |\delta_x(f)| \leq \lambda_x ||f||_\mathcal{F}.$$

[2] For any two function $f, g \in \mathcal{F}$, $|f(x) - g(x)| = |\delta_x(f) - \delta_x(g)| = |\delta_x(f - g)| \leq \lambda_x ||f - g||_\mathcal{F}$ for some

Theorem 2.1 (Representation theorem)**.** *Every continuous linear functional f on a Hilbert space $\mathcal{H}$ has the form*

$$f(x) = \langle x, y \rangle$$

with a unique $y \in \mathcal{M}$ and $\|f\| = \|y\|_{\mathcal{H}}$.

Theorem 2.2 (Orthogonal decomposition)**.** *Let $\mathcal{H}$ be a Hilbert space and $\mathcal{M} \subset \mathcal{H}$ be a closed subspace. For every $x \in \mathcal{H}$, we can write*

$$x = y + z$$

where $y \in \mathcal{M}$ and $z \in \mathcal{M}^\perp$, and y and z are uniquely determined by x.

Corollary 2.2.1. *Let $\mathcal{M}$ be a subspace of a Hilbert space $\mathcal{H}$. Then, $\mathcal{M}^\perp = \{0\}$ if and only if $\mathcal{M}$ is dense in $\mathcal{H}$.*

`https://en.wikibooks.org/wiki/Functional_Analysis/Hilbert_spaces`

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 $\mathcal{H}$ is the closure of $\mathcal{H}$, i.e. $(\mathcal{H}^\perp)^\perp = \overline{\mathcal{H}}$. If $\mathcal{M}$ is a closed linear subspace of $\mathcal{H}$, then $\mathcal{H} = \mathcal{M} \oplus \mathcal{M}^\perp$.

2.2 Reproducing kernel Hilbert spaces

2.3 Reproducing kernel Krein spaces

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.

$\lambda_x \geq 0$, thus is said to be Lipschitz continuous, which implies uniform continuity. This property implies pointwise convergence from norm convergence in $\mathcal{F}$.

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 2.3 (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 2.4 (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 2.8 (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 $\operatorname{E} f_\beta(x) = 0$. The reproducing kernel of $(\mathcal{X} - \mu)'$ is

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

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

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

and since $\operatorname{E}\langle x, \beta \rangle_\mathcal{X} = \langle \mu, \beta \rangle_\mathcal{X}$ 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 $t_1, \dots, t_k$, 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 $\operatorname{E} B_\gamma(t) = 0$ for all $t \in [0, T]$ and

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

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 2.9 (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 2.10 (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 2.5. *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 $x, x' \in \mathcal{X}$,

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

since

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

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.

2. This is the same for any RKHS?

Definition 2.11 (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 $X \in \mathcal{X}$.

3. 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 2.12 (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 2.13 (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* $\overline{\operatorname{span}}\{h(\cdot, x) | x \in \mathcal{X}\}$, 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 2.14 (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.

5. 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 2.15 (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 2.6. *Let $\mathcal{F}$ be the Pearson RKHS of functions over a finite set $\mathcal{X}$. Then,*

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

with

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

Proof. Write $p_x = \operatorname{P}(X = x)$. 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 $\operatorname{E} h(X, x') = \operatorname{E}[\delta_{Xx'}]/p_{x'} - 1 = p_{x'}/p_{x'} - 1 = 0$, and thus $\operatorname{E} f(X) = 0$. Conversely, suppose $f : \mathcal{X} \to \mathbb{R}$ is such that $\operatorname{E} f(X) = 0$. Taking $w_x = p_x f(x)$, we see that

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

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

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

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 2.16 (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 2.7. *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, $AAf = A^2f = Af$, 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 $Af_t = A(I-A)f = Af - A^2f = 0$. 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 $x = (x_1, \dots, x_p) \in \prod_{k=1}^p \mathcal{X}_k$, 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_kf 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 $\mathcal{P}_p = \mathcal{P}(\{1, \dots, p\})$ 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 $A_kf_{\mathcal{K}} = 0, \forall k \in \mathcal{K} \in \mathcal{P}_p$.

Example 2.1 (Two-way ANOVA decomposition)**.** Let $p = 2$, $\mathcal{X}_1 = \{1, \dots, m_1\}$, 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 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 $x = (x_1, \dots, x_p) \in \prod_{k=1}^p \mathcal{X}_k =: \mathcal{X}$, 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

$$A_k f = \int_{\mathcal{X}_k} f(x_1, \dots, x_p) d\nu(x_k). \tag{2.1}$$

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

$$f = \overbrace{\int_{\mathcal{X}} f(x)\,\mathrm{d}\nu(x)}^{f_0} + \overbrace{\left(f - \int_{\mathcal{X}} f(x)\,\mathrm{d}\nu(x)\right)}^{f_1} \tag{2.2}$$

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 (2.1) 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 (2.2), 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 (2.2) is orthogonal:

Claim 2.8. *For the ANOVA decomposition in* (2.2)*,* 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[3] $\mathcal{F} = \mathcal{F}_1 \otimes \cdots \otimes \mathcal{F}_p$, 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

$$\begin{aligned}
\mathcal{F} &= \big(\mathcal{F}_0 \overset{\perp}{\oplus} \bar{\mathcal{F}}_1\big) \otimes \cdots \otimes \big(\mathcal{F}_0 \overset{\perp}{\oplus} \bar{\mathcal{F}}_1\big) \\
&= \mathcal{F}_0^{\otimes p} \ \overset{\perp}{\oplus} \ \bigoplus_{j=1}^{p}{}^{\perp} \Big(\mathcal{F}_0^{\otimes (p-1)} \otimes \bar{\mathcal{F}}_j\Big) \ \overset{\perp}{\oplus} \ \bigoplus_{\substack{j,k=1 \\ j<k}}^{p}{}^{\perp} \Big(\mathcal{F}_0^{\otimes (p-2)} \otimes \bar{\mathcal{F}}_j \otimes \bar{\mathcal{F}}_k\Big) \qquad (2.3) \\
&\quad \overset{\perp}{\oplus} \ \cdots \ \overset{\perp}{\oplus} \ \Big(\bar{\mathcal{F}}_1 \otimes \cdots \otimes \bar{\mathcal{F}}_p\Big).
\end{aligned}$$

To clarify,

- $\mathcal{F}_0^{\otimes p}$ is the space of constant functions $f : \mathcal{X}_1 \times \cdots \times \mathcal{X}_p \to \mathbb{R}$.
- $\Big(\mathcal{F}_0^{\otimes (p-1)} \otimes \bar{\mathcal{F}}_j\Big)$ 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.
- $\Big(\mathcal{F}_0^{\otimes (p-2)} \otimes \bar{\mathcal{F}}_j \otimes \bar{\mathcal{F}}_k\Big)$ is the space of functions that are constant on all coordinates except the jth and kth coordinate of x. 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 (2.3) leads to the *functional ANOVA representation*

$$f(x) = \mu + \sum_{j=1}^{p} f_j(x_j) + \sum_{\substack{j,k=1 \\ j<k}}^{p} f_{jk}(x_j, x_k) + \cdots + f_{1\cdots p}(x). \tag{2.4}$$

Definition 2.17 (Functional ANOVA representation)**.** Let $\mathcal{P}_d = \mathcal{P}(\{1,\dots,d\})$, the power set of $\{1,\dots,d\}$. 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 (2.4) is known as the *functional ANOVA representation* of f if $\forall k \in \mathcal{K} \in \mathcal{P}_p$,

$$A_k f_{\mathcal{K}} = \int_{\mathcal{X}_{\mathcal{K}}} f_{\mathcal{K}}(x_{\mathcal{K}}) \, \mathrm{d}\nu_k(x_k) = 0, \tag{2.5}$$

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 (2.5) ensures orthogonality of the summands in (2.4).

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

[3] 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 (2.4). 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 (2.3)). The ANOVA RKKS will be constructed by a similar manipulation of the individual kernels representing the RKHS building blocks.

Definition 2.18 (The ANOVA RKKS)**.** For $k = 1, \dots, p$, 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 $\lambda_k, k = 1, \dots, p$ 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

$$h_\lambda(x, x') = \prod_{k=1}^{p} \big(1 + \lambda_k h_k(x_k, x'_k)\big). \tag{2.6}$$

The construction an ANOVA RKKS is very very simple in through multiplication of univariate kernels. Expanding out equations (2.6), we see that it is in fact a sum of separable kernels with increasing orders of interaction:

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

It is now clear from the expansion that the ANOVA RKKS yields functions that resemble those with the ANOVA representation in (2.4): 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 (2.4) 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.

2.6 The Sobolev-Hilbert inner product

2.7 Discussion

Resolving the uncentred polynomial and ANOVA RKKS.

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.*

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