To-do list

Contents

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

Chapter 2

Vector space of functions

chapter2

One of the main assumptions for regression modelling with I-priors is that the regression functions lie in some vector space of functions. The purpose of this chapter is to provide a concise review of functional analysis leading up to the theory of reproducing kernel Hilbert and Kreĭn spaces (RKHS/RKKS). The interest with these 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 applied statistical and machine learning literature, but perhaps RKKS are less so. To provide an early insight, RKKS are simply a generalisation of RKHS, and are defined as the difference between two RKHSs. The flexibility provided by RKKS will prove both useful and necessary, especially when considering the sums and products of scaled function spaces, as is done in I-prior modelling.

It is emphasised that a deep knowledge of functional analysis, including RKHS and RKKS theory, is not at all necessary for I-prior modelling, so perhaps the advanced reader may wish to skip Sections 2.1–2.3. 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.

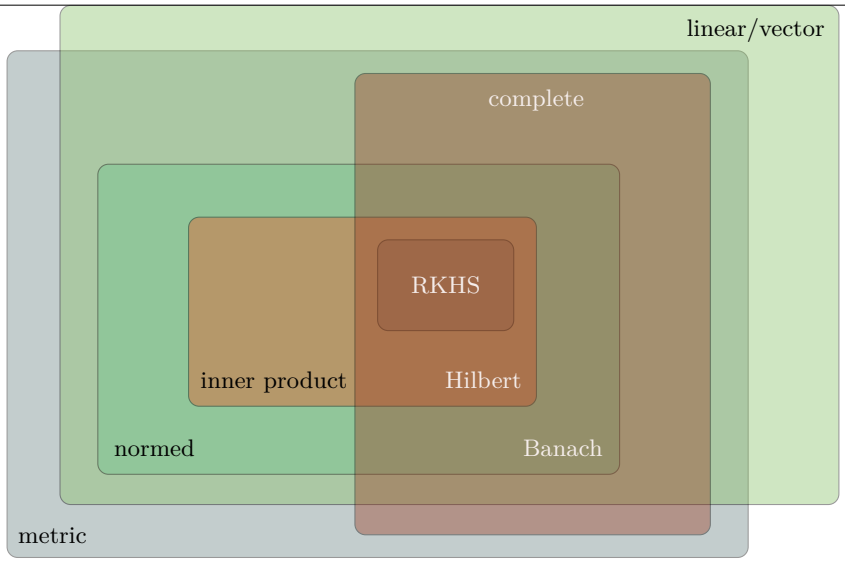

Figure 2.1: A hierarchy of vector spaces[1].

A remark on notation: Sets and vector spaces are denoted by calligraphic letters, and as much as possible, we shall stick to the convention that $\mathcal{F}$ denotes function spaces, and $\mathcal{X}$ denotes set of covariates or function inputs. Occasionally, we will describe a generic Hilbert space denoted by $\mathcal{H}$. Elements of the vector space of real functions over a set $\mathcal{X}$ are denoted $f(\cdot)$, 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 Reproducing kernel Hilbert space theory

sec:rkhstheory

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

Definition 2.1 (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.

def:rkhs

Definition 2.2 (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 inspires its name.

def:repkern

Definition 2.3 (Reproducing kernels)**.** Let $\mathcal{F}$ be a Hilbert space of functions over a non-empty set $\mathcal{X}$. A symmetric, bivariate function $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a *kernel*, and it is a *reproducing kernel* of $\mathcal{F}$ if h satisfies

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

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

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

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

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

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

thm:posdef

Claim 2.1 (Reproducing kernels of RKHS are positive-definite)**.** *Let $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a reproducing kernel for a Hilbert space $\mathcal{F}$. Then h is a symmetric and positive definite function.*

Proof.

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

□

Remark 2.1. In the kernel method literature, a *kernel* $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is usually defined as the inner product between inputs in feature space. That is, take $\phi : \mathcal{X} \to \mathcal{V}$, $x \mapsto \phi(x)$, where $\mathcal{V}$ is a Hilbert space. Then the kernel is defined as $h(x, x') = \langle \phi(x), \phi(x')\rangle_{\mathcal{V}}$, for any $x, x' \in \mathcal{X}$. The space $\mathcal{V}$ is known as the *feature space* and the mapping ϕ the *feature map*. In many mathematical models involving feature space mappings, elucidation of the feature map and feature space is not necessary, and thus computation is made simpler by the use of kernels (known as the *kernel trick*—Hofmann et al., 2008). Note that kernels defined in this manner are positive definite, while in this thesis, we opt for a more general definition allowing kernels to not necessarily be positive. The relevance of

this generality will be appreciated when we discuss reproducing kernel Kreĭn spaces in **??**.

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

Definition 2.4 (Kernel matrix)**.** Let $\{x_1, \dots, x_n\}$ be a sample of points, where each $x_i \in \mathcal{X}$, and h a kernel over $\mathcal{X}$. Define the *kernel matrix* $\mathbf{H}$ for h as the $n \times n$ matrix with (i, j) entries equal to $h(x_i, x_j)$.

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

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

- **RKHS $\Leftrightarrow$ reproducing kernel** (Theorem 2.2). For every RKHS $\mathcal{F}$ of functions over a set $\mathcal{X}$, there corresponds a unique, positive-definite reproducing kernel $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, and vice-versa. That is, a Hilbert space is a RKHS if it possesses a unique, reproducing kernel.
- **P.d. function $\Rightarrow$ RKHS** (Theorem 2.3). For every positive-definite function $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, there corresponds a unique RKHS $\mathcal{F}$ that has h as its reproducing kernel.

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

thm:rkhsuni que

Theorem 2.2 (RKHS uniqueness)**.** *Let $\mathcal{F}$ be a Hilbert space of functions over $\mathcal{X}$. $\mathcal{F}$ is a RKHS if and only if $\mathcal{F}$ has a reproducing kernel $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, and that h is unique to $\mathcal{F}$.*

Proof. Omitted—see Sejdinovic and Gretton (2012). □

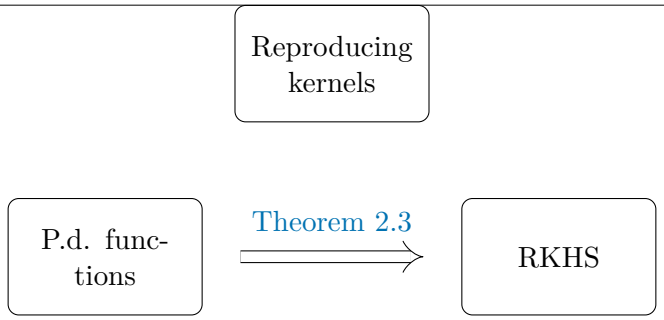

Figure 2.2: Test

thm:moorea

Theorem 2.3 (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.*

Proof. Omitted—see Berlinet and Thomas-Agnan (2011). □

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 $\mathcal{G} = \text{span}\{h(\cdot, x) \,|\, x \in \mathcal{X}\}$ endowed with the inner product as stated in **??**.

2.2 Summary

The brief notes on functional analysis allow us to describe the theory of reproducing kernel Hilbert and Kreĭn spaces. These are of great interest to us because the topology endowed on such spaces gives great assurances—in particular, all evaluation functionals are continuous in these spaces. Moreover, RKHS and RKKS can be specified completely through kernel functions, with new and complex function spaces built simply by manipulation of these kernel functions. Of particular importance is the ANOVA functional decomposition, for which we realise provides an objective way of constructing various statistical models (such models will be described later on in detail in Chapter 4).

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

- **Functional analysis**. On the introductory material relating to functional analysis in Section 2.1, the lecture notes by Sejdinovic and Gretton (2012) is recommended, and forms the basis for most of the material described. Additionally, Rudin (1987) provides a complementary reading.

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

Bibliography

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

zafeiriou2012subspace
Zafeiriou, Stefanos (2012). "Subspace learning in krein spaces: Complete kernel fisher discriminant analysis with indefinite kernels". In: *European Conference on Computer Vision*. Springer, pp. 488–501.