Chapter 11
Reproducing Kernel Hilbert Spaces

The ideas used in analyzing linear models extend naturally to much more general spaces than $\mathbf{R}^n$. One theoretical extension was introduced in PA Subsection 6.3.5. These extensions lead to interesting tools for examining penalized estimation in regression (Pearce and Wand, 2006) and support vector machines in classification/discrimination problems (Moguerza and Muñoz, 2006 and Zhu, 2008), They are commonly used in areas such as functional data analysis (Ramsay and Silverman, 2005), computer model analysis (Storlie et al. 2009), image processing (Berman, 1994), and various applications of spatial statistics (Bivand, Pebesma, and Gómez-Rubio, 2008 or Storlie, Bondell, and Reich, 2010), to name a few. The flexibility and elegance of RKHS methods are remarkable. Texts and survey articles on related subjects include Wahba (1990), Eubank (1999), Hastie, Tibshirani, and Friedman (2001), Gu (2002), Berlinet and Thomas-Agnan (2004), Bühlmann and van de Geer (2011), Heckman (2012), Bühlmann, Kalisch, and Meier (2014), and Wainwright (2014). These include a wealth of additional references as do the papers mentioned earlier.

In particular, the key ideas of linear models extend completely to *finite dimensional Hilbert spaces* whereas much of the newer theory has been developed for infinite dimensional *Reproducing Kernel Hilbert Spaces* (*RKHS*s). We provide an introduction to the mathematical ideas behind this work emphasizing its connections to linear model theory and two applications to problems that we have previously solved using linear model theory: ridge regression from Chapter 10 and (without a penalty function) spline regression from Chapter 9. We also provide an illustration of using reproducing kernels to test lack of fit in a linear model. Our development follows closely that of Nosedal-Sanchez, Storlie, Lee, and Christensen (2012).

Ridge regression and smoothing splines can both be viewed as solutions to minimization problems in a function space. If such a minimization problem is posed on an RKHS, the solution is guaranteed to exist and has a very simple form. We begin by solving some simple problems that relate to our broader goals. Section 2 introduces Banach spaces and Hilbert spaces. Section 3 provides basic ideas of RKHSs. Section 4 discusses the two distinct ways of using RKHS results and exploits the one not primarily used here to look at testing lack of fit in a linear model. Section 5

discusses penalized regression with RKHS's and the two specific examples of ridge regression and smoothing splines.

11.1 Introduction

For a known $n \times p$ matrix

$$X = [X_1, \cdots, X_p] = \begin{bmatrix} x_1' \\ \vdots \\ x_n' \end{bmatrix}$$

and vector Y, consider solving the system of n equations in p unknowns

$$X\beta = Y. \tag{1}$$

For a solution to exist, we must have $Y \in C(X)$. In standard linear model problems, we typically have $n > r(X)$, which typically causes $Y \notin C(X)$ and precludes us from finding a solution. That is precisely why we go to the trouble of finding (generalized) least squares estimates for β. More commonly in the current era of "big data," people have become interested in solving such problems when $r(X) \geq n$ in which case solutions often exist.

We wish to find the solution that minimizes the norm $\|\beta\| = \sqrt{\beta'\beta}$ but we solve the problem using concepts that extend to RKHSs. As mentioned, a solution $\tilde{\beta}$ (not necessarily a minimum norm solution) exists whenever $Y \in C(X)$. Given one solution $\tilde{\beta}$, all solutions β must satisfy

$$X\tilde{\beta} = X\beta$$

or

$$X(\tilde{\beta} - \beta) = 0.$$

Write $\tilde{\beta}$ uniquely as $\tilde{\beta} = \beta_0 + \beta_1$ with $\beta_0 \in C(X') = \text{span}\{x_1, \ldots, x_n\}$ and $\beta_1 \in C(X')^{\perp}$, then β_0 is a solution because $X(\tilde{\beta} - \beta_0) = X\beta_1 = 0$.

In fact, β_0 is both the unique solution in $C(X')$ and the minimum norm solution. If β is any other solution in $C(X')$ then $X(\beta - \beta_0) = 0$ so we have both $(\beta - \beta_0) \in C(X')^{\perp}$ and $(\beta - \beta_0) \in C(X')$, two sets whose intersection is only the 0 vector. Thus $\beta - \beta_0 = 0$ and $\beta = \beta_0$. In other words, every solution $\tilde{\beta}$ has the same β_0 vector. Finally, β_0 is also the minimum norm solution because the arbitrary solution $\tilde{\beta}$ has

$$\beta_0'\beta_0 \leq \beta_0'\beta_0 + \beta_1'\beta_1 = \tilde{\beta}'\tilde{\beta}.$$

We have established the existence of a unique, minimum norm solution in $C(X')$ that can be written as

$$\beta_0 = X'\xi = \sum_{i=1}^{n} \xi_i x_i,$$

for some ξ_i, $i = 1,\ldots,n$. To find β_0 explicitly, write $\beta_0 = X'\xi$ and the defining equation (1) becomes

$$XX'\xi = Y, \tag{2}$$

which is again a system of linear equations. Even if there exist multiple solutions ξ, $X'\xi$ is unique.

EXAMPLE 11.1.1. We use this framework to illustrate the "smallest" solution to the system of equations $\beta_1 + \beta_3 = 0$ and $\beta_2 = 1$. In the general framework (1), these become

$$\begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix},$$

whereas (2) becomes

$$\begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

The unique solution is $(\xi_1, \xi_2)' = (0,1)'$ which implies that the solution to the original problem is $\beta_0 = 0x_1 + 1x_2 = (0,1,0)'$.

Virtually the same methods can be used to solve similar problems in any inner-product space $\mathscr{M}$. As discussed in PA Subsection 6.3.5, an inner product $\langle \cdot, \cdot \rangle$ assigns real numbers to pairs of vectors. There the notion was used to treat random variables as vectors whereas in most of this book and PA, vectors are elements of $\mathbf{R}^n$ or $\mathbf{R}^p$. In this chapter we often use functions from $\mathscr{E}$ into $\mathbf{R}$ as vectors. Frequently we take $\mathscr{E} \subset \mathbf{R}$ but $\mathscr{E}$ can be a subset of a more general vector space. Note that an element of $\mathbf{R}^n$ can be thought of as a function from the integers $\{1,2,\ldots,n\}$ into $\mathbf{R}$.

We generalize the problem of solving a system of linear equations as follows. For n given vectors $x_i \in \mathscr{M}$ and numbers $y_i \in \mathbf{R}$, we might want to find the vector $\beta \in \mathscr{M}$ such that

$$\langle x_i, \beta \rangle = y_i, \qquad i = 1,2,...,n \tag{3}$$

for which the norm $\|\beta\| \equiv \sqrt{\langle \beta, \beta \rangle}$ is minimal. The solution (if one exists) has the form

$$\beta_0 = \sum_{i=1}^{n} \xi_i x_i, \tag{4}$$

with ξ_i satisfying the linear equations

$$\sum_{k=1}^{n} \langle x_i, x_k \rangle \xi_k = y_i, \qquad i = 1,\ldots,n \tag{5}$$

or, equivalently, we can solve the matrix equation

$$\tilde{R}\xi = Y$$

where the $n \times n$ matrix $\tilde{R}$ has elements $\tilde{r}_{ij}$ with

$$\tilde{r}_{ij} = \langle x_i, x_j \rangle.$$

For a formal proof see Máté (1989). From the matrix equation it is easy to check whether a solution exists. After a brief digression, we apply this result to the interpolating spline problem.

The illustrations in this chapter focus on three vector spaces, $\mathbf{R}^n$, because it is familiar,

$$\mathscr{L}^2[0,1] = \left\{ f : \int_0^1 [f(t)]^2 dt < \infty \right\},$$

because it relates to fitting splines, and the vector space of all functions from $\mathbf{R}^{p-1}$ into $\mathbf{R}$

$$\mathscr{F} = \left\{ f : f(x) \in \mathbf{R},\ x \in \mathbf{R}^{p-1} \right\},$$

because it relates to fitting multiple regression models that include an intercept. We will look at two different inner products on $\mathbf{R}^n$. We will look at the standard inner product on $\mathscr{L}^2[0,1]$ and at two subspaces with different inner products on each. For $\mathscr{F}$ we focus on linear multiple regression by focusing on subspaces of constant, linear, and affine functions and the orthogonality properties that they display.

The first illustration looks at a subset of $\mathscr{L}^2[0,1]$. Throughout, $f^{(m)}(t)$ denotes the m-th derivative of f with $\dot{f} \equiv f^{(1)}$ and $\ddot{f} \equiv f^{(2)}$. Recall from Chapter 1 that we also have $\dot{f}(t) \equiv d_t f(t)$ and $\ddot{f}(t) \equiv d_{tt}^2 f(t)$.

11.1.1 Interpolating Splines

Suppose we want to find a function $f(t)$ that interpolates between the points (t_i, y_i), $i = 0, 1, \ldots, n$, where $y_0 \equiv 0$ and $0 = t_0 < t_1 < t_2 < \cdots < t_n \leq 1$. We restrict attention to functions $f \in \mathscr{W}_0^1$ where

$$\mathscr{W}_0^1 = \left\{ f \in \mathscr{L}^2[0,1] : f(0) = 0, \int_0^1 [\dot{f}(t)]^2 dt < \infty \right\}$$

and it is understood that the derivatives exist because they are assumed to be square integrable. The restriction that $y_0 = f(0) = 0$ is not really necessary, but simplifies the presentation.

We want to find the smoothest function $f(t)$ that satisfies $f(t_i) = y_i$, $i = 1, \ldots, n$. Defining an inner product on $\mathscr{W}_0^1$ by

$$\langle f, g \rangle = \int_0^1 \dot{f}(x)\dot{g}(x)dx \tag{6}$$

determines a norm (length) $\|f\| \equiv \sqrt{\langle f, f \rangle}$ for $\mathscr{W}_0^1$ that is small for "smooth" functions. To address the interpolation problem, define the indicator function $I_A(t)$ for a

set A to be 1 if $t \in A$ and 0 if $t \notin A$. Note that the functions

$$R_i(s) \equiv \min(s,t_i) = sI_{[0,t_i]}(s) + t_i I_{(t_i,1]}(s),$$

$i = 1,2,...,n$, have $R_i(0) = 0$, $\dot{R}_i(s) = I_{[0,t_i]}(s)$, so $R_i(\cdot) \in \mathscr{W}_0^1$ and have the property that $\langle R_i, f\rangle = f(t_i)$ because

$$\begin{aligned}\langle f, R_i\rangle &= \int_0^1 \dot{f}(s)\dot{R}_i(s)ds \\ &= \int_0^{t_i} \dot{f}(s)ds = f(t_i) - f(0) = f(t_i).\end{aligned}$$

Thus, any interpolator f satisfies a system of equations like (3), namely

$$f(t_i) = \langle R_i, f\rangle = y_i, \qquad i = 1,\ldots,n. \tag{7}$$

and by (4), the smoothest function f (minimum norm) that satisfies the requirements has the form

$$\hat{f}(t) = \sum_{i=1}^{n} \xi_i R_i(t).$$

The ξ_j's are the solutions to the system of real linear equations obtained by substituting $\hat{f}$ into (7), that is

$$\sum_{j=1}^{n} \langle R_i, R_j\rangle \xi_j = y_i, \qquad i = 1,2,\ldots,n$$

or

$$\tilde{R}\xi = Y.$$

Note that

$$\langle R_i, R_j\rangle = R_j(t_i) = R_i(t_j) = \min(t_i,t_j)$$

and we define the function

$$R(s,t) = \min(s,t)$$

that turns out to be a reproducing kernel.

EXAMPLE 11.1.2. Given points $f(t_i) = y_i$, say, $f(0) = 0$, $f(0.1) = 0.1$, $f(0.25) = 1$, $f(0.5) = 2$, $f(0.75) = 1.5$, and $f(1) = 1.75$, we can now find $f \in \mathscr{W}_0^1$ that satisfies these six conditions and minimizes the norm associated with (1). The vector $\xi \in \mathbf{R}^5$ that satisfies the system of equations

$$\begin{bmatrix} 0.1 & 0.1 & 0.1 & 0.1 & 0.1 \\ 0.1 & 0.25 & 0.25 & 0.25 & 0.25 \\ 0.1 & 0.25 & 0.5 & 0.5 & 0.5 \\ 0.1 & 0.25 & 0.5 & 0.75 & 0.75 \\ 0.1 & 0.25 & 0.5 & 0.75 & 1 \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \\ \xi_4 \\ \xi_5 \end{bmatrix} = \begin{bmatrix} 0.1 \\ 1 \\ 2 \\ 1.5 \\ 1.75 \end{bmatrix}$$

is $\xi = (-5, 2, 6, -3, 1)'$, which implies that the smoothest interpolating function is

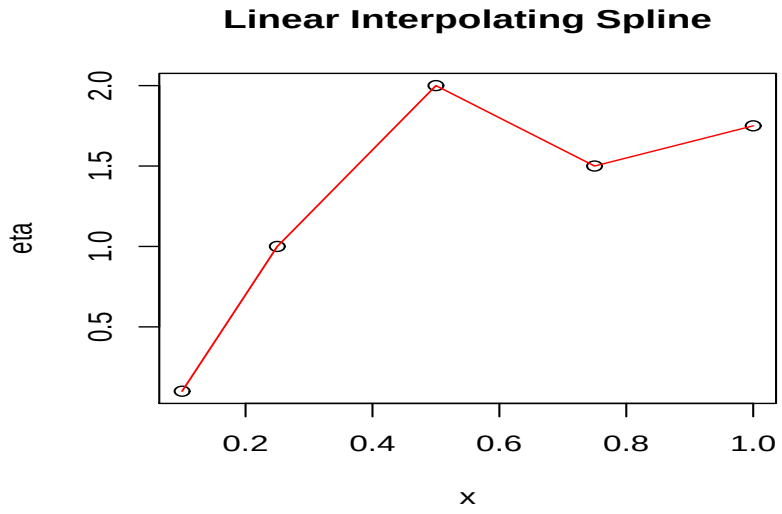
$$\begin{aligned} \hat{f}(t) &= -5R_1(t) + 2R_2(t) + 6R_3(t) - 3R_4(t) + 1R_5(t) \\ &= -5R(t,t_1) + 2R(t,t_2) + 6R(t,t_3) - 3R(t,t_4) + 1R(t,t_5) \end{aligned} \tag{8}$$

or, adding together the slopes for $t > t_i$ and finding the intercepts,

$$\hat{f}(t) = \begin{cases} t & 0 \le t \le 0.1 \\ 6t - 0.5 & 0.1 \le t \le 0.25 \\ 4t & 0.25 \le t \le 0.5 \\ -2t + 3 & 0.5 \le t \le 0.75 \\ t + 0.75 & 0.75 \le t \le 1. \end{cases}$$

This is the linear interpolating spline as can be seen graphically in Figure 11.1, although for reasons discussed later, the plot does not go below $t = 0.1$.

Fig. 11.1 Linear Interpolating Spline



For this illustration we restricted f so that $f(0) = 0$ but only for convenience of presentation. It can be shown that the form of the solution remains the same with any shift to the function, so that in general the solution takes the form $\hat{f}(t) = \beta_0 + \sum_{j=1}^{n} \xi_j R_j(t)$ where $\beta_0 = y_0$.

The two key points are (a) that the functions $R_j(t)$ allow us to express a function evaluated at a point as an inner-product constraint and (b) the restriction to functions in $\mathscr{W}_0^1$. $\mathscr{W}_0^1$ is a very special function space, a reproducing kernel Hilbert space, and R_j is determined by a reproducing kernel function R.

Ultimately, our goal is to address more complicated regression problems like

EXAMPLE 11.1.3. *Smoothing Splines.*
Consider simple regression data (x_i, y_i), $i = 1, \ldots, n$ and finding the function that minimizes

$$\frac{1}{n}\sum_{i=1}^{n}[y_i - f(x_i)]^2 + \lambda \int_0^1 \left[f^{(m)}(x)\right]^2 dx. \tag{9}$$

If $f(x)$ is restricted to be in an appropriate class of functions, minimizing only the first term gives least squares estimation within the class. If the class contains functions with $f(x_i) = y_i$ for all i, such functions minimize the first term but are typically very "unsmooth," i.e., have large second term. For example, an $n-1$ order polynomial will always fit the data perfectly but is typically very unsmooth. The second "penalty" term is minimized whenever the mth derivative is 0 everywhere, but (at least for small m) that rarely has a small first term. For $m = 1$, $0 \le x \le 1$, and a suitable class of functions, as we will see later, the minimizer takes the form

$$\hat{f}(x) = \beta_0 + \sum_{i=1}^{n} \xi_i R_i(x),$$

where the R_is were given earlier and the coefficients are found by solving a slightly different system of linear equations. Choosing $m = 1, 2$ determines linear and cubic smoothing splines, respectively.

If our goal is only to derive the solution to the linear smoothing spline problem with one predictor variable, RKHS theory is overkill. The value of RKHS theory lies in its generality. For example, the spline penalty can be replaced by many other penalties having associated inner products, and the x_i's can be vectors. Using RKHS results, we can solve the general problem of finding the minimizer of $\frac{1}{n}\sum_{i=1}^{n}[y_i - f(x_i)]^2 + \lambda Q(f)$ for quite general functions Q that correspond to a squared norm in a Hilbert subspace. See Wahba (1990) or Gu (2002) for a full treatment.

11.2 Banach and Hilbert Spaces

As discussed in PA's Appendix A, a vector space is a set $\mathscr{M}$ that contains elements called "vectors" and supports two kinds of operations: addition of vectors and multiplication by scalars. The scalars are real numbers in this book and in PA but in general they can be from any "field." We rely on context to distinguish between the vector $0 \in \mathscr{M}$ and the scalar $0 \in \mathbf{R}$. For vectors $u, v \in \mathscr{M}$, we also write $u + (-1 \times v)$ as $u - v$. Any subset of $\mathscr{M}$ that is closed under vector addition and scalar multiplication is a subspace of $\mathscr{M}$. The classic book on finite dimensional vector spaces is Halmos (1958). Harville (1997) also contains a wealth of information. For more on

vector spaces and the other topics in this section, see, for example, Naylor and Sell (1982), Young (1988), Máté (1989) or Rustagi (1994).

11.2.1 Banach Spaces

A Banach space is a vector space that has some additional structure. First, a Banach space has a length measure, called a *norm*, associated with it and, second, a Banach space is complete under that norm. Banach spaces provide a convenient introduction to Hilbert spaces, who have another bit of structure, namely an inner product.

Definition 11.2.1. A *norm* of a vector space $\mathscr{M}$, denoted $||\cdot||$, is a nonnegative real valued function satisfying the following properties for all $u, v \in \mathscr{M}$ and all $a \in \mathbf{R}$.

1. Non-negative: $||u|| \geq 0$.
2. Strictly positive: $||u|| = 0$ implies $u = 0$.
3. Homogeneous: $||au|| = |a|\ ||u||$.
4. Triangle inequality: $||u+v|| \leq ||u|| + ||v||$.

A vector space is called a *normed vector space* when a norm is defined on the space. The norm of a vector is also called its *length*. For vectors $u, v \in \mathscr{M}$, the *distance* between u and v is defined as $||u-v||$.

Definition 11.2.2. A sequence $\{v_n\}$ in a normed vector space $\mathscr{M}$ is said to *converge* to $v_0 \in \mathscr{M}$ if

$$\lim_{n\to\infty} ||v_n - v_0|| = 0.$$

Definition 11.2.3. A sequence $\{v_n\} \subset \mathscr{M}$ is called a *Cauchy sequence* if for any given $\varepsilon > 0$, there exists an integer N such that

$$||v_m - v_n|| < \varepsilon, \qquad \text{whenever} \quad m, n \geq N.$$

Convergence of sequences in normed vector spaces follows the same general idea as sequences of real numbers except that the distance between two vectors of the space is measured by the norm of the difference between the two vectors.

Definition 11.2.4. (Banach Space). A normed vector space $\mathscr{M}$ is *complete* if every Cauchy sequence in $\mathscr{M}$ converges to an element of $\mathscr{M}$. A complete normed vector space is a *Banach Space*.

EXAMPLE 11.2.5. $\mathbf{R}$ with the absolute value norm $\|x\| \equiv |x|$ is a complete, normed vector space over $\mathbf{R}$, and is thus a Banach space.

EXAMPLE 11.2.6. For $\mathbf{R}^n$, let $x = (x_1, ..., x_n)'$ be a vector. The l_p norm on $\mathbf{R}^n$ is defined by

$$||x||_p \equiv \left[\sum_{i=1}^{n} |x_i|^p\right]^{1/p} \qquad \text{for } 1 \leq p < \infty.$$

One can verify properties 1-4 of Definition 11.2.1 for each p, validating that $||x||_p$ is a norm on $\mathbf{R}^n$. Under the l_p norm, $\mathbf{R}^n$ is complete and thus a Banach space. Euclidean distance on $\mathbf{R}^n$ corresponds to choosing $p = 2$.

Alternatively, if A is a positive definite matrix,

$$||x||_A \equiv \sqrt{x'Ax}$$

defines a norm and a Banach space on $\mathbf{R}^n$. Euclidean distance on $\mathbf{R}^n$ corresponds to choosing I for A.

11.2.2 Hilbert Spaces

A Hilbert Space is a Banach space in which the norm is defined by an inner-product (also called a dot-product) that maps any two vectors into a real number. Banach spaces incorporate concepts of length and distance; Hilbert spaces add the concept of orthogonality (perpendicularity).

We typically denote Hilbert spaces by $\mathscr{H}$. For elements $u, v \in \mathscr{H}$, write the inner product of u and v as either $\langle u, v\rangle_{\mathscr{H}}$ or, when it is clear from the context that the inner product is taking place in $\mathscr{H}$, as $\langle u, v\rangle$. An *inner product* must satisfy four properties for all $u, v, w \in \mathscr{H}$ and all $a \in \mathbf{R}$.

1. Associative: $\langle au, v\rangle = a\langle u, v\rangle$.
2. Commutative: $\langle u, v\rangle = \langle v, u\rangle$.
3. Distributive: $\langle u, v + w\rangle = \langle u, v\rangle + \langle u, w\rangle$.
4. Positive Definite: $\langle u, u\rangle \geq 0$ with equality holding only if $u = 0$.

Definition 11.2.7. A vector space $\mathscr{M}$ with an inner product defined on it is called an *inner product space*. Vectors u and v are *orthogonal*, written $u \perp v$, if $\langle u, v\rangle = 0$. Two sets of vectors are said to be orthogonal if every vector in one set is orthogonal to every vector in the other. The set of all vectors orthogonal to a subspace $\mathscr{N}$ of $\mathscr{M}$ is called the *orthogonal complement* of $\mathscr{N}$ with respect to $\mathscr{M}$ and is written $\mathscr{N}^{\perp}_{\mathscr{M}}$, or just $\mathscr{N}^{\perp}$. The *norm* of u in an inner product space is $\|u\| \equiv \sqrt{\langle u, u\rangle}$. The angle θ between two vectors u and v is defined by

$$\cos(\theta) \equiv \frac{\langle x, y\rangle}{\|u\|\,\|v\|}.$$

A complete inner-product space is called a *Hilbert space*.

Most of the ideas related to orthogonality that we have exploited here and in PA extend immediately to finite dimensional Hilbert spaces. It is easy to see that an orthogonal complement is a subspace because it is closed under vector addition and scalar multiplication. Moreover, because only the 0 vector can be orthogonal to itself, $\mathscr{N} \cap \mathscr{N}_{\mathscr{M}}^{\perp} = \{0\}$, which means that for any vector x that can be written as $x = x_0 + x_1$ with $x_0 \in \mathscr{N}$ and $x_1 \in \mathscr{N}_{\mathscr{M}}^{\perp}$, the representation is unique. If $\mathscr{M} = \text{span}\{x_1, \cdots, x_n\}$, Gram-Schmidt applies so that we can find an orthonormal spanning set $\{o_1, \cdots, o_n\}$ in which all vectors are orthgonal to each other and each $\|o_j\|$ is 0 or 1 and $\text{span}\{x_1, \cdots, x_r\} = \text{span}\{o_1, \cdots, o_r\}$, $r = 1, \ldots, n$. Eliminating the 0 vectors from $\{o_1, \cdots, o_n\}$ gives an orthonormal basis for $\mathscr{M}$. The key idea in the inductive proof of Gram-Schmidt is to set

$$w_{s+1} = x_{s+1} - \sum_{j=1}^{s} o_j \langle o_j, x_{s+1} \rangle \quad \text{and} \quad o_{s+1} = \begin{cases} \frac{1}{\|w_{s+1}\|} w_{s+1} & \text{if } \|w_{s+1}\| > 0 \\ 0 & \text{if } \|w_{s+1}\| = 0 \, . \end{cases}$$

By taking spanning sets $\{x_1, \cdots, x_n\}$ for $\mathscr{M}$ and $\{v_1, \cdots, v_s\}$ for a subspace $\mathscr{N}$, we can Gram-Schmidt the spanning set $\{v_1, \cdots, v_s, x_1, \cdots, x_n\}$ of $\mathscr{M}$ to get orthonormal bases for $\mathscr{N}$ and $\mathscr{N}_{\mathscr{M}}^{\perp}$ that combine to give a basis for $\mathscr{M}$ thus establishing that any vector $x \in \mathscr{M}$ can be written uniquely as $x = x_0 + x_1$ with $x_0 \in \mathscr{N}$ and $x_1 \in \mathscr{N}_{\mathscr{M}}^{\perp}$ and allowing us to define x_0 as the perpendicular projection of x onto $\mathscr{N}$ Sometimes the perpendicular projection of x into a subspace $\mathscr{N}$ of $\mathscr{M}$ is defined as the unique vector $x_0 \in \mathscr{N}$ with the property that $\langle x - x_0, u \rangle = 0$ for any $u \in \mathscr{N}$.

EXAMPLE 11.2.8. For $\mathbf{R}^n$ we can define a Hilbert space with the inner product

$$\langle u, v \rangle \equiv u'v = \sum_{i=1}^{n} u_i v_i$$

that conforms with Euclidean geometry. More generally, for any positive definite matrix A, we can define a Hilbert space with the inner product $\langle u, v \rangle \equiv u'Av$.

EXAMPLE 11.2.9. For

$$\mathscr{L}^2[0,1] = \left\{ f : \int_0^1 [f(t)]^2 dt < \infty \right\},$$

we can define a Hilbert space with the inner product

$$\langle f, g \rangle_{\mathscr{L}^2[0,1]} \equiv \int_0^1 f(t) g(t) dt.$$

The space is well-known to be complete, see de Barra (1981).

For the subspace

$$\mathscr{W}_0^1 = \left\{ f \in \mathscr{L}^2[0,1] : f(0) = 0, \int_0^1 [\dot{f}(t)]^2 dt < \infty \right\},$$

define the inner product

$$\langle f,g\rangle_{\mathscr{W}_0^1}=\int_0^1 \dot{f}(t)\dot{g}(t)dt. \tag{1}$$

Note that $\langle f,f\rangle_{\mathscr{W}_0^1}=0$ if and only if $f=0$ because if $\langle f,f\rangle_{\mathscr{W}_0^1}=0$, $\dot{f}(t)=0$ for all t, so $f(t)$ must be a constant, however $f(0)=0$, so $f(t)=0$ for all t.

For the subspace

$$\mathscr{W}_0^2=\left\{f\in\mathscr{L}^2[0,1]:f(0)=\dot{f}(0)=0,\int_0^1[\ddot{f}(t)]^2dt<\infty\right\},$$

define an inner product

$$\langle f,g\rangle_{\mathscr{W}_0^2}=\int_0^1 \ddot{f}(t)\ddot{g}(t)dt. \tag{2}$$

Note that $\langle f,f\rangle_{\mathscr{W}_0^2}=0$ if and only if $f=0$ because if $\langle f,f\rangle_{\mathscr{W}_0^2}=0$, $\ddot{f}(t)=0$ for all t, so $\dot{f}(t)$ must be a constant, however $\dot{f}(0)=0$, so $\dot{f}(t)=0$ for all t, hence $f(t)$ must be a constant, however $f(0)=0$, so $f(t)=0$ for all t.

EXAMPLE 11.2.10. Consider the vector space of all functions from $\mathbf{R}^{p-1}$ to $\mathbf{R}$,

$$\mathscr{F}=\left\{f:f(x)\in\mathbf{R},\ x\in\mathbf{R}^{p-1}\right\}.$$

The subspace of all constant functions on $\mathbf{R}^{p-1}$ is

$$\mathscr{F}_0=\{f_a\in\mathscr{F}:f_a(x)=a,\ a\in\mathbf{R},\ x\in\mathbf{R}^{p-1}\}$$

and define the inner product

$$\langle f_a,f_b\rangle_{\mathscr{F}_0}=ab.$$

Since $\mathbf{R}^{p-1}$ is a Hilbert Space, so is $\mathscr{F}_0$.

The subspace of all linear functions on $\mathbf{R}^{p-1}$ passing through the origin is

$$\mathscr{F}_1=\{f_\gamma\in\mathscr{F}:f_\gamma(x)=x'\gamma,\ \gamma\in\mathbf{R}^{p-1},\ x\in\mathbf{R}^{p-1}\}$$

and define the inner product

$$\langle f_\eta,f_\gamma\rangle_{\mathscr{F}_1}=\eta'\gamma=\eta_1\gamma_1+\eta_2\gamma_2+\cdots+\eta_{p-1}\gamma_{p-1}.$$

Again, since $\mathbf{R}^{p-1}$ is a Hilbert Space, so is $\mathscr{F}_1$

Now consider the subspace of all affine (i.e., linear plus a constant) functions on $\mathbf{R}^{p-1}$,

$$\mathscr{F}_*=\{f_\beta\in\mathscr{F}:f_\beta(x)=\beta_0+\beta_1x_1+\ldots+\beta_{p-1}x_{p-1},\ \beta\in\mathbf{R}^p,\ x\in\mathbf{R}^{p-1}\},$$

with the inner product

$$\langle f_\beta, f_\eta \rangle_{\mathscr{F}_*} = \beta_0\eta_0 + \beta_1\eta_1 + \ldots + \beta_{p-1}\eta_{p-1}.$$

This too is a Hilbert space.

The subspace of $\mathscr{F}_*$ that contains constant functions,

$$\mathscr{F}_0 = \{f_\beta \in \mathscr{F}_* : f_\beta(x) = \beta_0,\ 0 = \beta_1 = \cdots = \beta_p\}$$

has an orthogonal complement with respect to $\mathscr{F}_*$ of

$$\mathscr{F}_0^\perp = \{f_\beta \in \mathscr{F}_* : f_\beta(x) = \beta_1 x_1 + \ldots + \beta_{p-1}x_{p-1},\ 0 = \beta_0\} = \mathscr{F}_1.$$

For any $f_\beta, f_\eta \in \mathscr{F}_*$, write $\beta = [\beta_0, \beta_*']$ and $\eta = [\eta_0, \eta_*']$. We have the unique decompositions $f_\beta = f_{\beta_0} + f_{\beta_*}$ and $f_\eta = f_{\eta_0} + f_{\eta_*}$, and

$$\langle f_\beta, f_\eta \rangle_{\mathscr{F}_*} = \langle f_{\beta_0}, f_{\eta_0} \rangle_{\mathscr{F}_0} + \langle f_{\gamma_*}, f_{\eta_*} \rangle_{\mathscr{F}_1}.$$

11.3 Reproducing Kernel Hilbert Spaces

Hilbert spaces that display certain properties on certain linear operators are called reproducing kernel Hilbert spaces.

Definition 11.3.1. A function (operator) T mapping a vector space $\mathscr{X}$ into another vector space $\mathscr{Y}$ is called *linear* if $T(\lambda_1 x_1 + \lambda_2 x_2) = \lambda_1 T(x_1) + \lambda_2 T(x_2)$ for any $x_1, x_2 \in \mathscr{X}$ and any $\lambda_1, \lambda_2 \in \mathbf{R}$.

Any $p \times n$ matrix A maps vectors in $\mathbf{R}^n$ into vectors in $\mathbf{R}^p$ via $T_A(x) \equiv Ax$ and is linear.

Exercise 11.1 Consider a finite dimensional Hilbert space $\mathscr{H}$ (one that contains a finite basis) and a subspace $\mathscr{H}_0$. The operator M is a *perpendicular projection operator* onto $\mathscr{H}_0$ if $M(x) = x$ for any $x \in \mathscr{H}_0$ and $M(w) = 0$ for any $w \in \mathscr{H}_0^\perp$. Show that M must be both unique and linear. Let $\{o_1, \cdots, o_r\}$ be an orthonormal basis for $\mathscr{H}_0$ and show that

$$M(x) = \sum_{j=1}^{r} o_j \langle o_j, x \rangle.$$

Do these results hold for subspaces with infinite dimensions? In particular, if $\mathscr{H}$ has a countable basis and a subspace $\mathscr{H}_0$ is finite dimensional, can any $x \in \mathscr{H}$ be decomposed into $x = x_0 + x_1$ with $x_0 \in \mathscr{H}_0$ and $x_1 \in \mathscr{H}_0^\perp$?

Definition 11.3.2. The operator $T : \mathscr{X} \to \mathscr{Y}$ mapping a Banach space into a

Banach space is *continuous* at $x_0 \in \mathscr{X}$ if and only if for every $\varepsilon > 0$ there exists $\delta = \delta(\varepsilon) > 0$ such that for every x with $||x - x_0||_{\mathscr{X}} < \delta$ we have $||T(x) - T(x_0)||_{\mathscr{Y}} < \varepsilon$.

Linear operators are continuous everywhere if they are continuous at 0.

We can write $x_m \to x_0$, if $\|x_m - x_0\|_{\mathscr{X}} \to 0$. Continuity occurs if $x_m \to x_0$ implies $T(x_m) \to T(x_0)$ in the sense that $\|T(x_m) - T(x_0)\|_{\mathscr{Y}} \to 0$.

Definition 11.3.3. A real valued function defined on a vector space is called a *functional*.

Any vector $a \in \mathbf{R}^n$ defines a linear functional on $\mathbf{R}^n$ via $\phi_a(x) \equiv a'x$.

EXAMPLE 11.3.4. Let $\mathscr{S}$ be the set of bounded real valued differentiable functions $\{f(x)\}$ defined on the real line. Then $\mathscr{S}$ is a vector space with the usual $+$ and $\times$ operations for functions. Some linear functionals on $\mathscr{S}$ are $\phi(f) = \int_a^b f(x)dx$ and $\phi_a(f) = \dot{f}(a)$ for some fixed a. A nonlinear functional is $\phi_{a,b}(f) = \sup_{x \in [a,b]} f(x)$.

A linear functional of particular importance is the evaluation functional.

Definition 11.3.5. Let $\mathscr{M}$ be a vector space of functions defined from $\mathscr{E}$ into $\mathbf{R}$. For any $t \in \mathscr{E}$, denote by e_t the *evaluation functional* at the point t, i.e., for $g \in \mathscr{M}$, the mapping is $e_t(g) = g(t)$.

For $\mathscr{M} = \mathbf{R}^n$, vectors can be viewed as functions from the set $\mathscr{E} = \{1, 2, \ldots, n\}$ into $\mathbf{R}$. An evaluation functional is $e_i(x) = x_i$.

While it is simplest to take $\mathscr{E} \subset \mathbf{R}$ in Definition 11.3.5, we will need to consider $\mathscr{E} \subset \mathbf{R}^p$, and there is no reason not to use even more general vector spaces to define $\mathscr{E}$.

In a Hilbert space (or any normed vector space) of functions, the notion of pointwise convergence is related to the continuity of the evaluation functionals. The following are equivalent for a normed vector space $\mathscr{H}$ of real valued functions defined on $\mathscr{E}$.

(i) The evaluation functionals are continuous for all $t \in \mathscr{E}$.
(ii) If $f, f_1, f_2, \ldots \in \mathscr{H}$ and $||f_n - f|| \to 0$ then $f_n(t) \to f(t)$ for every $t \in \mathscr{E}$.
(iii) For every $t \in \mathscr{E}$ there exists $K_t > 0$ such that $|f(t)| \leq K_t ||f||$ for all $f \in \mathscr{H}$.

Here (ii) is the definition of (i). See Máté (1989) for a proof of (iii).

To define a reproducing kernel, we need the famous *Riesz Representation Theorem*.

Theorem 11.3.6. Let $\mathscr{H}$ be a Hilbert space and let ϕ be a continuous linear functional on $\mathscr{H}$. Then there is one and only one vector $g \in \mathscr{H}$ such that

$$\phi(f) = \langle g, f \rangle, \qquad \text{for all } f \in \mathscr{H}.$$

The vector g is sometimes called the *representation* of ϕ. Nonetheless, ϕ and g are different objects: ϕ is a linear functional on $\mathscr{H}$ and g is a vector in $\mathscr{H}$. For a proof of this theorem, see Naylor and Sell (1982) or Máté (1989).

For $\mathscr{H} = \mathbf{R}^n$ with the Euclidean inner product, the representation theorem is well known because for $\phi(x)$ to be a linear functional there must exist a vector g such that

$$\phi(x) = g'x.$$

In particular, an evaluation functional is $e_i(x) = x_i$. The representation of this linear functional is the indicator vector $R_i \in \mathbf{R}^n$ that is 0 everywhere except has a 1 in the ith place because

$$e_i(x) = x_i = R_i'x.$$

In the future we will use e_i to denote both the indicator vector in $\mathbf{R}^n$ that is 0 everywhere except has a 1 in the ith place and the evaluation functional that the indicator vector represents.

An element of a set of functions, say f, from $\mathscr{E}$ into $\mathbf{R}$, is sometimes denoted $f(\cdot)$ to be explicit that the elements are functions, whereas $f(t)$ is the value of $f(\cdot)$ evaluated at $t \in \mathscr{E}$.

Applying the Riesz Representation Theorem to a Hilbert space $\mathscr{H}$ of real valued functions in which all evaluation functionals are continuous, for every $t \in \mathscr{E}$ there is a unique symmetric function $R : \mathscr{E} \times \mathscr{E} \to \mathbf{R}$ with $R(\cdot,t) \in \mathscr{H}$ the representation of the evaluation functional e_t, so that

$$f(t) = e_t(f) = \langle R(\cdot,t), f(\cdot)\rangle, \qquad f \in \mathscr{H}.$$

The function R is called a *reproducing kernel* (r.k.) and $f(t) = \langle R(\cdot,t), f(\cdot)\rangle$ is called the *reproducing property* of R. In particular, by the reproducing property

$$R(t,s) = \langle R(\cdot,t), R(\cdot,s)\rangle.$$

The fact that $\langle R(\cdot,t), R(\cdot,s)\rangle = \langle R(\cdot,s), R(\cdot,t)\rangle$ is why R must be a symmetric function.

Again, our use of t is suggestive of it being a real number but in general it can be a vector.

Definition 11.3.7. A Hilbert space $\mathscr{H}$ of functions defined on $\mathscr{E}$ into $\mathbf{R}$ is called a *reproducing kernel Hilbert space* if all evaluation functionals are continuous.

EXAMPLE 11.3.8. For $\mathbf{R}^n$ with inner product $\langle u, v\rangle \equiv u'Av$ where the $n \times n$ matrix A is positive definite, the r.k. $R(\cdot,\cdot)$ maps $s,t = 1,\ldots,n$ into $\mathbf{R}$, so it is really just a matrix. To see that $[R(s,t)] = A^{-1}$, note that $R(\cdot,t)$ is the tth column of $[R(s,t)]$ and we must have thar $R(\cdot,t) = A^{-1}e_t$ because for any x

$$\langle R(\cdot,t), x\rangle \equiv x_t = e_t'x = e_t'A^{-1}Ax = \langle A^{-1}e_t, x\rangle.$$

As earlier, e_t is the vector with 0s everywhere except a 1 in the tth place. For Euclidean $\mathbf{R}^n$, the r.k. is I.

EXAMPLE 11.3.9. For

$$\mathscr{L}^2[0,1] = \left\{ f : \int_0^1 [f(t)]^2 dt < \infty \right\}$$

with the inner product

$$\langle f, g \rangle_{\mathscr{L}} \equiv \int_0^1 f(t) g(t) dt,$$

the evaluation functionals are not continuous, so no r.k. exists. For example, if $t_n = t_0 - (1/n)$ and we define the $\mathscr{L}^2[0,1]$ functions $f(t) = I_{[0,t_0]}(t)$ and $f_n(t) = I_{[0,t_n)}(t)$ we have

$$||f - f_n|| = \sqrt{\int_0^1 I_{[t_n,t_0]}(t) dt} = \frac{1}{\sqrt{n}} \to 0$$

but

$$f_n(t_0) = 0 \not\to 1 = f(t_0).$$

EXAMPLE 11.3.10. Consider the Hilbert space

$$\mathscr{W}_0^1 = \left\{ f \in \mathscr{L}^2[0,1] : f(0) = 0, \int_0^1 [\dot{f}(t)]^2 dt < \infty \right\}$$

with inner product

$$\langle f, g \rangle_{\mathscr{W}_0^1} = \int_0^1 \dot{f}(t) \dot{g}(t) dt. \tag{1}$$

In Subsection 11.1.1, we found the reproducing kernel to be $R(s,t) = \min(s,t)$. For fixed t, $R(\cdot,t)$ is an element of the function space $\mathscr{W}_0^1$, since $R(0,t) = 0$ and $\int_0^1 [d_s R(s,t)]^2 ds < \infty$. Also, as shown earlier, $R(\cdot,\cdot)$ has the reproducing property.

EXAMPLE 11.3.11. Consider the Hilbert space

$$\mathscr{W}_0^2 = \left\{ f \in \mathscr{L}^2[0,1] : f(0) = \dot{f}(0) = 0, \int_0^1 [\ddot{f}(t)]^2 dt < \infty \right\}$$

with inner product

$$\langle f, g \rangle_{\mathscr{W}_0^2} = \int_0^1 \ddot{f}(t) \ddot{g}(t) dt.$$

We begin by finding the second derivative of the representation of the evaluation functional. In particular, we show that

$$f(s) = \int_0^1 (s-u)_+ \ddot{f}(u) du\,, \tag{2}$$

where $(a)_+$ is a for $a > 0$ and 0 for $a \leq 0$.

Given any arbitrary and fixed $s \in [0,1]$,

$$\int_0^1 (s-u)_+ \ddot{f}(u)du = \int_0^s (s-u)\ddot{f}(u)du.$$

Integrating by parts

$$\int_0^s (s-u)\ddot{f}(u)du = (s-s)\dot{f}(s) - (s-0)\dot{f}(0) + \int_0^s \dot{f}(u)du = \int_0^s \dot{f}(u)du$$

and applying the Fundamental Theorem of Calculus to the last term,

$$\int_0^s (s-u)\ddot{f}(u)du = f(s) - f(0) = f(s).$$

Since the r.k. of the space $\mathscr{W}_2^0$ must satisfy $f(s) = \langle f(\cdot), R(\cdot,s)\rangle$, we see that $R(\cdot,s)$ is a function such that

$$d_{uu}^2 R(u,s) = (s-u)_+.$$

We also know that $R(\cdot,s) \in \mathscr{W}_2^0$, so using $R(s,t) = \langle R(\cdot,t), R(\cdot,s)\rangle$

$$R(s,t) = \int_0^1 (t-u)_+(s-u)_+ du = \frac{\max(s,t)\min^2(s,t)}{2} - \frac{\min^3(s,t)}{6}. \qquad (3)$$

Exercise 11.2 Do the calculus to establish equation (3).

EXAMPLE 11.3.12. Consider all constant functionals on $\mathbf{R}^{p-1}$,

$$\mathscr{F}_0 = \{f_a \in \mathscr{F} : f_a(x) = a,\ a \in \mathbf{R},\ x \in \mathbf{R}^{p-1}\},$$

with inner product

$$\langle f_a, f_b\rangle_{\mathscr{F}_0} = ab.$$

$\mathscr{F}_0$ has continuous evaluation functionals $e_x(f) = f(x)$, so it is an RKHS and has a unique reproducing kernel. To find the r.k., observe that $R(\cdot,x) \in \mathscr{F}_0$, so it is a constant for any x. Write $R(x) \equiv R(\cdot,x)$. By the representation theorem and the defined inner product

$$a = f_a(x) = \langle f_a(\cdot), R(\cdot,x)\rangle_{\mathscr{F}_0} = aR(x)$$

for any x and a. This implies that $R(x) \equiv 1$ so that $R(\cdot,x) \equiv 1$ and $R(\cdot,\cdot) \equiv 1$.

EXAMPLE 11.3.13. Consider all linear functionals on $\mathbf{R}^{p-1}$ passing through the origin,

$$\mathscr{F}_1 = \{f_\gamma \in \mathscr{F} : f_\gamma(x) = x'\gamma,\ \gamma \in \mathbf{R}^{p-1},\ x \in \mathbf{R}^{p-1}\},$$

with inner product

$$\langle f_\eta, f_\gamma \rangle_{\mathscr{F}_1} = \eta'\gamma = \eta_1\gamma_1 + \eta_2\gamma_2 + \cdots + \eta_{p-1}\gamma_{p-1}.$$

The r.k. R must satisfy

$$f_\gamma(x) = \langle f_\gamma(\cdot), R(\cdot,x) \rangle_{\mathscr{F}_1}$$

for all γ and any x. Since $R(\cdot,x) \in \mathscr{F}_1$, $R(v,x) = v'u$ for some u that depends on x, i.e., $u(x)$ is the vector in $\mathbf{R}^{p-1}$ that determines the functional $R(\cdot,x) \in \mathscr{F}_1$, so $R(\cdot,x) = f_{u(x)}(\cdot)$ and $R(v,x) = v'u(x)$. By our definition of $\mathscr{F}_1$ we have

$$x'\gamma = f_\gamma(x) = \langle f_\gamma(\cdot), R(\cdot,x) \rangle_{\mathscr{F}_1} = \langle f_\gamma(\cdot), f_{u(x)}(\cdot) \rangle_{\mathscr{F}_1} = \gamma' u(x),$$

so we need $u(x)$ such that for any γ and x we have

$$x'\gamma = u(x)'\gamma.$$

It follows that $x = u(x)$. For example, taking γ to be the indicator vector e_i implies that $x_i = u_i(x)$ for every $i = 1, \ldots, p-1$. We now have $R(\cdot,x) = f_x(\cdot)$ so that

$$R(\tilde{x},x) = \langle R(\cdot,\tilde{x}), R\cdot,x) \rangle_{\mathscr{F}_1} = \langle f_{\tilde{x}}, f_x \rangle_{\mathscr{F}_1} = \tilde{x}'x = x_1\tilde{x}_1 + x_2\tilde{x}_2 + \ldots + x_{p-1}\tilde{x}_{p-1}.$$

One point of these examples is that if you can characterize the evaluation functional $e_t(\cdot)$, then frequently you can find $R(s,t)$. For further examples of RKHSs with various inner products, see Berlinet and Thomas-Agnan (2004).

One further concept is useful in RKHS approaches to regression problems.

11.3.1 The projection principle for an RKHS

Consider the connection between the reproducing kernel R of the RKHS $\mathscr{H}$ and the reproducing kernel R_0 for a subspace $\mathscr{H}_0 \subset \mathscr{H}$. When any vector $f \in \mathscr{H}$ can be written uniquely as $f = f_0 + f_1$ with $f_0 \in \mathscr{H}_0$ and $f_1 \in \mathscr{H}_0^\perp$, more particularly, $R(\cdot,t) = R_0(\cdot,t) + R_1(\cdot,t)$ with $R_0(\cdot,t) \in \mathscr{H}_0$ and $R_1(\cdot,t) \in \mathscr{H}_0^\perp$ if and only if R_0 is the r.k. of $\mathscr{H}_0$ and R_1 is the r.k. of $\mathscr{H}_0^\perp$. For a proof see Gu (2002).

EXAMPLE 11.3.14. Consider the affine functionals on $\mathbf{R}^{p-1}$,

$$\mathscr{F}_* = \{f_\beta \in \mathscr{F} : f_\beta(x) = \beta_0 + \beta_1 x_1 + \ldots + \beta_{p-1}x_{p-1},\ \beta \in \mathbf{R}^p,\ x \in \mathbf{R}^{p-1}\}$$

with inner product

$$\langle f_\beta, f_\eta \rangle_{\mathscr{F}_*} = \beta'\eta = \beta_0\eta_0 + \beta_1\eta_1 + \ldots + \beta_{p-1}\eta_{p-1}.$$

We have already derived the r.k.'s for the constant functionals $\mathscr{F}_0$ and linear functionals $\mathscr{F}_1 \equiv \mathscr{F}_0^\perp$ (call them R_0 and R_1, respectively) in Examples 11.3.12 and

11.3.13. Applying the projection principle, the r.k. for $\mathscr{F}_*$ is the sum of R_0 and R_1, i.e.,

$$R(\tilde{x},x) = 1 + \tilde{x}'x.$$

For two subspaces $\mathscr{A}$ and $\mathscr{B}$ contained in a vector space $\mathscr{H}$, the *direct sum* is the space $\mathscr{D} = \{a+b : a \in \mathscr{A}, b \in \mathscr{B}\}$, written $\mathscr{D} = \mathscr{A} \oplus \mathscr{B}$. Any elements $d_1, d_2 \in \mathscr{D}$ can be written as $d_1 = a_1 + b_1$ and $d_2 = a_2 + b_2$, respectively, for some $a_1, a_2 \in \mathscr{A}$ and $b_1, b_2 \in \mathscr{B}$. When the two subspaces have $\mathscr{A} \cap \mathscr{B} = \{0\}$, those decompositions are unique. If the vector space $\mathscr{H}$ is also a Hilbert space and $\mathscr{A} \perp \mathscr{B}$, the decomposition is unique and the inner product between d_1 and d_2 is $\langle d_1, d_2 \rangle = \langle a_1, a_2 \rangle + \langle b_1, b_2 \rangle$. In particular, for any finite dimensional subspace $\mathscr{H}_0$ of a Hilbert space $\mathscr{H}$, we have $\mathscr{H} = \mathscr{H}_0 \oplus \mathscr{H}_0^\perp$ and $\mathscr{H}_0 \perp \mathscr{H}_0^\perp$.

In something of a converse, suppose $\mathscr{A}$ and $\mathscr{B}$ are Hilbert space subspaces of a vector space $\mathscr{M}$ that has no norm, and suppose $\mathscr{A} \cap \mathscr{B} = \{0\}$. Then we can define a Hilbert space on $\mathscr{D} = \mathscr{A} \oplus \mathscr{B}$ by defining $\langle d_1, d_2 \rangle_{\mathscr{D}} \equiv \langle a_1, a_2 \rangle_{\mathscr{A}} + \langle b_1, b_2 \rangle_{\mathscr{B}}$. With respect to the Hilbert space $\mathscr{D}$, $\mathscr{A} \perp \mathscr{B}$. For more information about direct sum decompositions see, for example, Berlinet and Thomas-Agnan (2004) or Gu (2002).

11.4 Two Approaches

There seems to be at least two ways to use RKHS results. One is to define a penalty function for estimation that relates to an RKHS and figure out (or at least demonstrate) what the appropriate reproducing kernel is. Our applications up to this point and in the next section focus on that approach. The other approach is to choose an appropriate "kernel" function for vectors in $\mathscr{M}$, say, $R(u,v)$ and to build an RKHS around the kernel. The idea then is to work within this RKHS without really ever leaving $\mathscr{M}$. We now show how to construct such an RKHS. The next subsection illustrates the use of such an RKHS.

For an appropriate kernel function $R(\cdot,\cdot)$, one can find (but the point is that one need not find) a positive definite eigen-decomposition of the function,

$$R(u,v) = \sum_{j=1}^{\infty} \eta_j \phi_j(u) \phi_j(v). \tag{1}$$

The existence of such an eigen-decomposition implies that

$$R(\cdot,v) = \sum_{j=1}^{\infty} \eta_j \phi_j(\cdot) \phi_j(v).$$

Associated with the eigen-decomposition is a Hilbert space

$$\mathscr{H} = \left\{ f : f = \sum_{j=1}^{\infty} \alpha_j \phi_j \quad \text{with} \quad \sum_{j=1}^{\infty} \alpha_j^2/\eta_j < \infty \right\}$$

having inner product

$$\left\langle \sum_{j=1}^{\infty} \alpha_j \phi_j, \sum_{j=1}^{\infty} \beta_j \phi_j \right\rangle \equiv \left\langle \sum_{j=1}^{\infty} \alpha_j \phi_j(\cdot), \sum_{j=1}^{\infty} \beta_j \phi_j(\cdot) \right\rangle \equiv \sum_{j=1}^{\infty} \alpha_j \beta_j / \eta_j.$$

Although the eigen-decomposition in (1) need not be unique because eigenvalues η_j may have multiplicities greater than 1, nonetheless every decomposition involves a set of basis functions ϕ_j that define a unique representation and all span the same space $\mathscr{H}$.

The chosen kernal function is the r.k. on $\mathscr{H}$ because $R(\cdot, v)$ represents the evaluation functional, i.e.,

$$\begin{aligned} \left\langle R(\cdot, v), \sum_{j=1}^{\infty} \alpha_j \phi_j(\cdot) \right\rangle &= \left\langle \sum_{j=1}^{\infty} \eta_j \phi_j(\cdot) \phi_j(v), \sum_{j=1}^{\infty} \alpha_j \phi_j(\cdot) \right\rangle \\ &= \sum_{j=1}^{\infty} \eta_j \phi_j(v) \alpha_j / \eta_j \\ &= \sum_{j=1}^{\infty} \alpha_j \phi_j(v). \end{aligned}$$

In this context, some commonly used kernels are, when $\mathscr{M}$ is a Hilbert space, the polynomial kernels for positive integers r

$$R(u, v) \equiv (1 + \langle u, v \rangle_{\mathscr{M}})^r$$

and, when $\mathscr{M}$ is a Banach space, the radial basis kernel

$$R(u, v) \equiv \exp(-\gamma \|u - v\|_{\mathscr{M}})$$

for some positive scalar γ. Most often the space $\mathscr{M}$ is Euclidean $\mathbf{R}^p$.

We now briefly apply these ideas, and the idea of never actually using the space $\mathscr{H}$, to the problem of testing lack of fit in a linear model.

11.4.1 Testing Lack of Fit

To translate these ideas back into linear regression theory, consider a linear model $Y = X\beta + e$ with X written in terms of its p dimensional row vectors as

$$X = \begin{bmatrix} x_1' \\ \vdots \\ x_n' \end{bmatrix}.$$

For a kernel function defined on $\mathbf{R}^p$, whenever the eigen-decomposition is finite, say,

$$R(u,v) = \sum_{j=1}^{s} \eta_j \phi_j(u)\phi_j(v)$$

we can define

$$\Phi_k = [\phi_k(x_1), \cdots, \phi_k(x_n)]', \quad \Phi = [\Phi_1, \cdots, \Phi_s],$$

and fit a new regression model

$$Y = \Phi\alpha + e, \quad \mathrm{E}(e) = 0. \tag{2}$$

This new model is appropriate for testing lack of fit because either we get $C(X) \subset C(\Phi)$ or, if not, we could use the model matrix $[X, \Phi]$ instead, with little change to our discussion.

Because the η_ks are positive, using PA Proposition B.51 we see that

$$C(\Phi) = C(\Phi D(\sqrt{\eta_i})) = C(\Phi D(\eta_i)\Phi').$$

The key fact is that

$$\Phi D(\eta_i)\Phi' = [R(x_h, x_i)]_{n\times n} \equiv \tilde{R},$$

so instead of fitting the linear model (2) we can fit the equivalent model

$$Y = \tilde{R}\xi + e$$

to obtain $\hat{\xi}$, fitted values, and residuals. Moreover, we can make a prediction for a new observed vector x_0 by using

$$\hat{y}_0 = [R(x_0, x_1), \cdots, R(x_0, x_n)]\hat{\xi}.$$

Thus we can execute the analysis without ever specifying the ϕ_ks or η_ks, but it would be useful to know s in order to find the MSE. The fact that the matrix Φ can be replaced by the matrix $\tilde{R}$ is known as the *kernel trick*.

Typically, one would also want to modify this discussion to deal with an intercept in the model and thus use vectors in $\mathbf{R}^{p-1}$ rather than $\mathbf{R}^p$. Moreover, there is nothing in this discussion that precludes the eigen-decomposition from being infinite. The only problem with an infinite decomposition is the likelihood of getting $C(\tilde{R}) = \mathbf{R}^n$ and a not very interesting regression model. In fact, even for finite s we need to worry about overfitting the data.

EXAMPLE 11.4.1. Consider a linear model $y_i = \beta_0 + x_i'\beta_* + \varepsilon_i$ or, in matrix form,

$$Y = X\beta + e = [J, Z]\begin{bmatrix}\beta_0 \\ \beta_*\end{bmatrix} + e$$

with

$$Z = \begin{bmatrix} x_1' \\ \vdots \\ x_n' \end{bmatrix}.$$

This is the model used in Chapter 10 for ridge regression and in PA Section 6.2.

We illustrate the use of

$$R(x_i, x_h) \equiv (1 + x_h' x_i)^2$$

when $p - 1 = 2$ so that $x_i' = (x_{i1}, x_{i2})$. Write

$$Z = [X_1, X_2], \quad X_1 = [x_{i1}], \quad X_2 = [x_{i2}]$$

and define the n dimensional vectors

$$X_1^2 \equiv [x_{i1}^2], \quad X_2^2 \equiv [x_{i2}^2], \quad X_1 X_2 \equiv [x_{i1} x_{i2}].$$

Note that the linear model

$$Y = W\gamma + e \equiv \gamma_{00} + \gamma_{10} X_1 + \gamma_{01} X_2 + \gamma_{20} X_1^2 + \gamma_{02} X_2^2 + \gamma_{11} X_1 X_2 + e$$

is a quadratic model in the two predictor variables.

It is not difficult to see that in our example

$$R(x_i, x_h) \equiv (1 + x_h' x_i)^2 = 1 + 2x_{h1} x_{i1} + 2x_{h2} x_{i2} + x_{h1}^2 x_{i1}^2 + x_{h2}^2 x_{i2}^2 + 2x_{h1} x_{i1} x_{h2} x_{i2}$$

which leads to the hth column of $\tilde{R}$ being

$$\tilde{R}_h = J + 2x_{h1} X_1 + 2x_{h2} X_2 + x_{h1}^2 X_1^2 + x_{h2}^2 X_2^2 + 2x_{h1} x_{h2} X_1 X_2.$$

Clearly, $\tilde{R}_h \in C(J, X_1, X_2, X_1^2, X_2^2, X_1 X_2) = C(W)$, so $C(\tilde{R}) \subset C(W)$. Moreover, for regression data it is almost inconceivable that no 6 of the columns of the $n \times n$ matrix $\tilde{R}$ would be linearly independent, so in all likelihood $C(\tilde{R}) = C(W)$. In this case, fitting the model using the kernel trick is equivalent to fitting a full quadratic model to the data.

In this illustration, the point of the kernel trick is that instead of having to go to the trouble of defining W, one can use $R(x_i, x_h)$ to define $\tilde{R}$. Doesn't seem like much of an advantage in this problem, does it? In fact, it is hard to see where using the kernel trick with the polynomial kernels would ever provide much of an advantage. On the other hand, we saw in Chapter 9 that fitting cubic splines involved fitting a linear model with some pretty horrible linear constraints. In the next section, an appropriate r.k. enables us to fit the cubic spline model without linear constraints.

Exercise 11.3. In terms of the notation for this section, Example 11.4.1 uses $\mathscr{M} = \mathbf{R}^2$. What is $\mathscr{H}$?

11.5 Penalized Regression with RKHSs

As mentioned in the introduction, nonparametric regression is a powerful approach for solving many modern problems. The nonparametric regression model is given by

$$y_i = f(x_i) + \varepsilon_i, \quad i = 1, 2, \ldots, n,$$

where f is an unknown regression function and the ε_i are independent, mean 0 error terms. We start this section with three common examples of penalized regression: ridge regression, lasso regression, and smoothing splines.

11.5.1 Ridge and Lasso Regression

The classical linear regression setting is notationally identical to Example 11.4.1, i.e., $y_i = \beta_0 + x_i'\beta_* + \varepsilon_i$ with $\beta' = (\beta_0, \beta_*')$. The classical ridge regression estimator $\hat{\beta}_R$ proposed by Hoerl and Kennard (1970) minimizes

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p-1} \beta_j x_{ij} \right)^2 + \lambda n \sum_{j=1}^{p-1} \beta_j^2 \tag{1}$$

where x_{ij} is the ith observation on the jth predictor variable. The resulting estimate is biased but can reduce the variance relative to least squares estimates. The tuning parameter $\lambda \geq 0$ is a constant that controls the trade-off between bias and variance in $\hat{\beta}_R$, and is often selected by some form of cross validation, see Section 6.

The lasso regression estimator $\hat{\beta}_L$ proposed by Tibshirani (1996) minimizes

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p-1} \beta_j x_{ij} \right)^2 + \lambda n \sum_{j=1}^{p-1} |\beta_j|$$

The l_1 norm for the penalty defines a Banach space but does not lend itself to an inner product, so RKHS results do not readily apply to lasso problems. However, Lin and Zhang (2006) and Storlie et al. (2010) used the RKHS framework to develop smoothing spline versions of the lasso and the adaptive lasso (cf. Zou, 2006), respectively.

11.5.2 Smoothing Splines

Smoothing splines are among the most popular methods for the estimation of f due to their good empirical performance and sound theoretical support. We assume that the domain of f is $[0,1]$. With $f^{(m)}$ the mth derivative of f, a smoothing spline

estimate $\hat{f}$ is a minimizer of

$$\sum_{i=1}^{n} [y_i - f(x_i)]^2 + \lambda n \int \left[f^{(m)}(x) \right]^2 dx \tag{2}$$

The minimization of (2) is implicitly over functions with square integrable m-th derivatives. The first term of (2) encourages the fitted f to be close to the data, while the second term penalizes the roughness of f. The smoothing parameter λ controls the trade-off between the two conflicting goals. The special case of $m = 1$ is the linear smoothing spline problem.

In practice it is common to choose $m = 2$ in which case the minimizer $\hat{f}$ of (2) is called a cubic smoothing spline. As $\lambda \to \infty$, $\hat{f}$ approaches the least squares simple linear regression line, while as $\lambda \to 0$, $\hat{f}$ approaches the minimum curvature interpolant.

11.5.3 Solving the General Penalized Regression Problem

We now review a general framework that allows us to minimize (1), (2) and many other similar problems, cf. O'Sullivan (1986), Lin (2006), Storlie (2009), Storlie et al. (2010), Gu (1993). The data model is

$$y_i = f(x_i) + \varepsilon_i, \quad \mathrm{E}(\varepsilon_i) = 0, \quad i = 1, 2, \ldots, n, \tag{3}$$

where the ε_i are error terms and $f \in \mathscr{M}$, a given vector space of functions on $\mathscr{E}$.

Let Q be a nonnegative penalty functional on $\mathscr{M}$ with restrictions that are discussed later. An estimate of f is obtained by minimizing

$$\frac{1}{n} \sum_{i=1}^{n} [y_i - f(x_i)]^2 + \lambda Q(f), \tag{4}$$

over $f \in \mathscr{S} \subset \mathscr{M}$ where $\mathscr{S}$ is chosen in a specific way.

We require that Q, in addition to being nonnegative, has a null set $\mathscr{N} = \{f \in \mathscr{M} : Q(f) = 0\}$ that is a subspace, that for $f_N \in \mathscr{N}$ and $f \in \mathscr{M}$, we have $Q(f_N + f) = Q(f)$, and that there exists an RKHS $\mathscr{H} \subset \mathscr{M}$ for which the inner product satisfies $\langle f, f \rangle = Q(f)$. This condition forces $\mathscr{N} \cap \mathscr{H} = \{0\}$. Finally, we consider a finite dimensional subspace of $\mathscr{N}$, say $\mathscr{N}_0$, with basis functions $\{\psi_1, \ldots, \psi_s\}$ and define

$$\mathscr{S} \equiv \mathscr{N}_0 \oplus \mathscr{H}.$$

In many applications $\mathscr{N}_0 = \mathscr{N}$.

EXAMPLE 11.5.1. *Linear Interpolating Splines.* Using Example 11.3.10 we can take $\mathscr{M}$ as the subspace of $\mathscr{L}^2[0,1]$ with finite values of

$$Q(f) \equiv \int_0^1 \left[\dot{f}(t)\right]^2 dt.$$

This $Q(f)$ satisfies our four conditions with $\mathscr{H} = \mathscr{W}_0^1$. The constant functions $f(t) = a$ comprise $\mathscr{N} = \mathscr{N}_0$, so we take $\psi_1(x) \equiv 1$. Note that equation (11.3.1) defines an inner product on $\mathscr{W}_0^1$ but does not define an inner product on all of $\mathscr{M}$ because nonzero functions can have a zero inner product with themselves, hence the nontrivial nature of $\mathscr{N}$.

Some nice things happen if $\mathscr{M}$ is a Hilbert space. In particular, for any two subspaces $\mathscr{N}$ and $\mathscr{H}$ with $\mathscr{N} \perp \mathscr{H}$, we have for any f in their direct sum a unique decomposition $f = f_0 + f_1$ with $f_0 \in \mathscr{N}$ and $f_1 \in \mathscr{H}$. This allows us to define $Q(f) \equiv \langle f_1, f_1 \rangle$ and have all our assumptions met. (Technically, we should define Q for all functions $f \in \mathscr{M}$ but how we extend it beyond the direct sum is irrelevant.) The orthogonal decomposition of $\mathscr{S}$ is closely related to generalized additive models, cf. Wood (2006), and more generally to smoothing spline ANOVA models, cf. Gu (2002), which also include tensor product splines as a special case. Thin plate splines also fall nicely into the general RKHS framework, cf. Wahba (1990).

The key result, Wahba's Representation Theorem, also known as the "dual form" or "kernel trick", cf. Pearce and Wand (2006), is that the minimizer of (4) is a finite linear combination of known basis functions and functions involving the reproducing kernel on $\mathscr{H}$. This allows us to find the coefficients of the linear combination by solving a quadratic minimization problem similar to those in standard linear models.

Theorem 11.5.2. *Representation Theorem.* Any minimizer $\hat{f} \in \mathscr{S}$ of equation (4) has the form

$$\hat{f}(x) = \sum_{j=1}^{s} \beta_j \psi_j(x) + \sum_{i=1}^{n} \xi_i R(x_i, x), \tag{5}$$

where $R(\cdot,\cdot)$ is the r.k. for $\mathscr{H}$.

PROOF: An informal proof is given. See Wahba (1990) or Gu (2002) for a formal proof.

Since we are working in $\mathscr{S}$, clearly, any minimizer $\hat{f}$ must have $\hat{f} = \hat{f}_0 + \hat{f}_1$ with $\hat{f}_0 \in \mathscr{N}_0$ and $\hat{f}_1 \in \mathscr{H}$. We want to establish that $\hat{f}_1(\cdot) = \sum_{i=1}^n \xi_i R(x_i, \cdot)$ for some ξ_is. Decompose $\mathscr{H}$ as $\mathscr{H} = \mathscr{H}_0 \oplus \mathscr{H}_0^{\perp}$ where $\mathscr{H}_0 = \text{span}\{R(x_i,\cdot), i = 1,\ldots,n\}$ so that

$$\hat{f}_1(\cdot) = \hat{f}_R(\cdot) + \eta(\cdot),$$

with

$$\hat{f}_R(\cdot) \equiv \sum_{i=1}^{n} \xi_i R(x_i, \cdot) \tag{6}$$

and $\eta(\cdot) \in \mathscr{H}_0^{\perp}$. By orthogonality and the reproducing property of the r.k.,

$$0 = \langle R(x_i,\cdot), \eta(\cdot) \rangle = \eta(x_i), \quad i = 1,\ldots,n.$$

We now establish the representation theorem. Using our assumptions about Q,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}[y_i-\hat{f}(x_i)]^2+\lambda Q(\hat{f}) &= \frac{1}{n}\sum_{i=1}^{n}[y_i-\hat{f}_0(x_i)-\hat{f}_1(x_i)]^2+\lambda Q(\hat{f}_0+\hat{f}_1)\\ &= \frac{1}{n}\sum_{i=1}^{n}[y_i-\hat{f}_0(x_i)-\hat{f}_1(x_i)]^2+\lambda Q(\hat{f}_1)\\ &= \frac{1}{n}\sum_{i=1}^{n}[y_i-\hat{f}_0(x_i)-\hat{f}_R(x_i)-\eta(x_i)]^2+\lambda Q(\hat{f}_R+\eta).\end{aligned}$$

Because $\eta(x_i)=0$ and using orthogonality within $\mathscr{H}$

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}[y_i-\hat{f}(x_i)]^2+\lambda Q(\hat{f}) &= \frac{1}{n}\sum_{i=1}^{n}[y_i-\hat{f}_0(x_i)-\hat{f}_R(x_i)]^2+\lambda\left[Q(\hat{f}_R)+Q(\eta)\right]\\ &\geq \frac{1}{n}\sum_{i=1}^{n}[y_i-\hat{f}_0(x_i)-\hat{f}_R(x_i)]^2+\lambda Q(\hat{f}_R).\end{aligned}$$

Clearly, any $\eta\neq 0$ makes the inequality strict, so minimizers have $\eta=0$ and $\hat{f}=\hat{f}_0+\hat{f}_R$ with the last inequality an equality. □

Corollary 11.5.3. Among $f\in\mathscr{S}$, the least squares estimate with minimum penalty satisfies the relation (5).

PROOF: As in the proof of the Theorem 11.5.2, any $f\in\mathscr{S}$ can be written as $f_0+f_R+\eta$ and

$$\frac{1}{n}\sum_{i=1}^{n}[y_i-f_0(x_i)-f_R(x_i)-\eta(x_i)]^2=\frac{1}{n}\sum_{i=1}^{n}[y_i-f_0(x_i)-f_R(x_i)]^2.$$

However,

$$Q(f)=Q(f_0+f_R+\eta)=Q(f_R+\eta)=Q(f_R)+Q(\eta)\geq Q(f_R)=Q(f_0+f_R).$$

Thus for any $f\in\mathscr{S}$, there is a function of the form (5) that has the same squared error and at least as small a penalty, so there exists a least squares estimate that minimizes the penalty and has this form. □

A remarkable feature of the result in (5) is that the form of the minimizer is represented by a finite dimensional basis, regardless of the dimension of $\mathscr{H}$. For example, $\mathscr{H}=\mathscr{W}_0^1$ requires an infinite expansion of basis functions to represent all functions in the space, yet the *solution* of the minimization can be represented by a finite basis!

Once we know that the minimizer takes the form (5), we can find the coefficients of the linear combination by solving a quadratic minimization problem similar to those in standard linear models. This occurs because we can write $Q(\hat{f})=Q(\hat{f}_R)$

as a quadratic form in $\xi = (\xi_1, \ldots, \xi_n)'$. Define $\tilde{R}$ as the $n \times n$ matrix where the i, j entry is $\tilde{r}_{ij} = R(x_i, x_j)$. The matrix $\tilde{R}$ is commonly referred to as the *Gram* matrix. Now, using the reproducing property of R, write

$$Q(\hat{f}_R) = \left\langle \sum_{i=1}^{n} \xi_i R(x_i, \cdot), \sum_{j=1}^{n} \xi_j R(x_j, \cdot) \right\rangle = \sum_{i=1}^{n} \sum_{j=1}^{n} \xi_i \xi_j R(x_i, x_j) = \xi' \tilde{R} \xi.$$

Define the observation vector $Y = [y_1, \ldots, y_n]'$, and let W be the $n \times s$ matrix defined by $w_{ij} = \psi_j(x_i)$. With the usual norm for Euclidean $\mathbf{R}^n$, the minimization of (4) takes the form

$$\min_{\beta, \xi} \left\{ \frac{1}{n} \| Y - W\beta - \tilde{R}\xi \|^2 + \lambda \xi' \tilde{R} \xi \right\}. \tag{7}$$

The minimization in (7) is a special case of the generalized ridge regression minimization problem (10.2.1) and has the solutions of (10.2.3),

$$\hat{\xi} = [\tilde{R}(I - M_W)\tilde{R} + \lambda n \tilde{R}]^{-} \tilde{R}(I - M_W)Y \qquad \hat{\beta} = [W'W]^{-} W'(Y - \tilde{R}\hat{\xi}). \tag{8}$$

Alternatively, the solution to the minimization of (4) can be found from the normal equations associated with the linear model (10.2.2) used for solving generalized ridge regression. In this application the model becomes

$$\begin{bmatrix} Y \\ 0 \end{bmatrix} = \begin{bmatrix} W & \tilde{R} \\ 0 & \sqrt{\lambda n} \tilde{S} \end{bmatrix} \begin{bmatrix} \beta \\ \xi \end{bmatrix} + \begin{bmatrix} e \\ \tilde{e} \end{bmatrix} \tag{9}$$

where $\tilde{R} = \tilde{S}'\tilde{S}$. The normal equations are

$$\begin{bmatrix} W'W & W'\tilde{R} \\ \tilde{R}W & \tilde{R}\tilde{R} + \lambda n \tilde{R} \end{bmatrix} \begin{bmatrix} \beta \\ \xi \end{bmatrix} = \begin{bmatrix} W'Y \\ \tilde{R}Y \end{bmatrix} \tag{10}$$

or they can be reduced to

$$\begin{bmatrix} W'W & 0 \\ 0 & \tilde{R}(I - M_W)\tilde{R} + \lambda n \tilde{R} \end{bmatrix} \begin{bmatrix} \delta \\ \xi \end{bmatrix} = \begin{bmatrix} W'Y \\ \tilde{R}(I - M_W)Y \end{bmatrix} \tag{11}$$

with $W\delta \equiv W\beta + M_W \tilde{R}\xi$. Both of these require solving a system of $s + n$ linear equations to find the estimates.

The simplest way to *program* these results may be to find

$$\begin{bmatrix} \hat{\beta} \\ \hat{\xi} \end{bmatrix} = \begin{bmatrix} W'W & W'\tilde{R} \\ \tilde{R}W & \tilde{R}\tilde{R} + \lambda n \tilde{R} \end{bmatrix}^{-} \begin{bmatrix} W'Y \\ \tilde{R}Y \end{bmatrix}$$

where it is easy to find the Moore-Penrose generalized inverse of a nonnegative definite matrix by using its eigen-decomposition as demonstrated in the proof of PA Theorem B.38. See the supplemental material to Nosedal-Sanchez et al. (2012) for examples of programs in the R language. However, it may be more efficient to

find the two generalized inverses of dimensions s and n associated with solving (11) rather than one generalized inverse of dimension $s+n$ to solve (10).

For clarity, we have restricted our attention to minimizing (4), which incorporates squared error loss between the observations and the unknown function evaluations. The representation theorem holds for more general loss functions, e.g., those from logistic or Poisson regression, see Gu (2002).

EXAMPLE 11.5.4. Reconsider fitting linear splines to the data of Example 11.1.2 but without the requirement that $f(0) = 0 = y_0$. The observed data are

$$Y \equiv \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 0.1 \\ 1 \\ 2 \\ 1.5 \\ 1.75 \end{bmatrix}, \quad X \equiv \begin{bmatrix} x_1' \\ \vdots \\ x_n' \end{bmatrix} \equiv \begin{bmatrix} t_1 \\ \vdots \\ t_n \end{bmatrix} = \begin{bmatrix} 0.1 \\ 0.25 \\ 0.5 \\ 0.75 \\ 1 \end{bmatrix}$$

We have

$$Q(f) \equiv \int_0^1 [\dot{f}(t)]^2 dt,$$

$\mathscr{N} = \mathscr{N}_0$, the one dimensional space spanned by $\psi(x) = 1$, and $\mathscr{H} = \mathscr{W}_0^1$, so $R(s,t) = \min(s,t)$. It follows that $W = J_n$ and for these data

$$\tilde{R} = \begin{bmatrix} 0.1 & 0.1 & 0.1 & 0.1 & 0.1 \\ 0.1 & 0.25 & 0.25 & 0.25 & 0.25 \\ 0.1 & 0.25 & 0.5 & 0.5 & 0.5 \\ 0.1 & 0.25 & 0.5 & 0.75 & 0.75 \\ 0.1 & 0.25 & 0.5 & 0.75 & 1 \end{bmatrix}.$$

A similar structure holds whenever $t_1 \le t_2 \le \cdots \le t_n$, which we henceforth assume.

If we set $\lambda = 0$ to get linear interpolating splines, we are just solving a least squares problem and the least squares problem is minimized by any function with $f(x_i) = y_i$, $i = 1, \ldots, n$. In particular, with five distinct x_i values, a quartic (fourth degree) polynomial will fit the data perfectly as would appropriate sine and cosine models as discussed in Chapter 9. One problem with these solutions is that they will not be particularly smooth as judged by having small values of $Q(f)$. In any case, Theorem 11.5.2 tells us that there is also a solution that can be found by fitting the model

$$Y = J\beta_0 + \tilde{R}\xi + e.$$

The first column of $\tilde{R}$ is $t_1 J$, so the estimates are not unique. We could fit this model using a standard regression package deleting the first column of $\tilde{R}$, which is equivalent to imposing a side condition of $\xi_1 = 0$. If we do so, we get $\hat{\beta}_0 = -0.5$ and $\hat{\xi} = (0, 2, 6, -3, 1)'$. A predicted value for a new t is

$$\begin{aligned} \hat{f}(t) &= \hat{\beta}_0 + [R(t,t_1), \cdots, R(t,t_n)]\hat{\xi} \\ &= -0.5 + 0R(t,t_1) + 2R(t,t_2) + 6R(t,t_3) - 3R(t,t_4) + 1R(t,t_5). \end{aligned}$$

Or we could drop the "intercept" by imposing the side condition $\beta_0 = 0$ and fit only $Y = \tilde{R}\xi + e$. Now a predicted value is

$$\begin{aligned}\hat{f}(t) &= \hat{\beta}_0 + [R(t,t_1),\cdots,R(t,t_n)]\hat{\xi} \\ &= 0 - 5R(t,t_1) + 2R(t,t_2) + 6R(t,t_3) - 3R(t,t_4) + 1R(t,t_5).\end{aligned}$$

which is the solution we found in Example 11.1.2. These two solutions are different for $0 \le t < t_1$ but agree for $t_1 \le t \le 1$. There is no unique solution. Figure 1 reflects that fact by not including the plot below $t = 0.1$ even though in Example 11.1.2, $\hat{f}$ is defined over the entire interval. In fact, for these data we need only have $\hat{\beta}_0$ and $\hat{\xi}_1$ satisfying

$$-0.5 = \hat{\beta}_0 + \hat{\xi}_1 \times 0.1 = \hat{\beta}_0 + \hat{\xi}_1 R(t_1,t_1).$$

Moreover all solutions give a horizontal line from t_n to 1, an issue hidden in this example by the fact that $t_n = 1$. Actually, to fit interpolating splines, you do not need to do regression, you merely need to solve the system of linear equations $Y = J\beta_0 + \tilde{R}\xi$.

We can find an interpolating function that uses the form of (5) but what if we wanted an interpolator that minimized $Q(f)$? Corollary 11.5.3 tells us the the minimizer has this form, but not how to find the minimizer itself. Intuitively, the answer is quite clear. We need an interpolator that is a flat a possible. Which of the possible answers from the linear model will minimize the roughness penalty? Clearly, pick $\hat{\xi}_1$ to give a flat function on $[0,t_1]$, so the derivative is 0 and its contribution to Q is minimized. Moreover, this shows that having a horizontal line from t_n to 1 is actually a feature and not a bug. We return to this question after the next example.

EXAMPLE 11.5.5. Reconsider Example 11.5.1 but now including the point at $(t_0,y_0) = (0,0)$. The observed data are

$$Y \equiv \begin{bmatrix} y_0 \\ \vdots \\ y_{n-1} \end{bmatrix} = \begin{bmatrix} 0 \\ 0.1 \\ 1 \\ 2 \\ 1.5 \\ 1.75 \end{bmatrix}, \quad \begin{bmatrix} t_0 \\ \vdots \\ t_{n-1} \end{bmatrix} = \begin{bmatrix} 0 \\ 0.1 \\ 0.25 \\ 0.5 \\ 0.75 \\ 1 \end{bmatrix}$$

As before $W = J_n$ (but n is 6 rather than 5) and for these data

$$\tilde{R} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 \\ 0 & 0.1 & 0.25 & 0.25 & 0.25 & 0.25 \\ 0 & 0.1 & 0.25 & 0.5 & 0.5 & 0.5 \\ 0 & 0.1 & 0.25 & 0.5 & 0.75 & 0.75 \\ 0 & 0.1 & 0.25 & 0.5 & 0.75 & 1 \end{bmatrix}.$$

For $\lambda = 0$, as with the previous example, any interpolation function minimizes the squared errors. Similar to the previous example, fitting a fifth degree polynomal provides an interpolator, but not one that is necessarily very smooth. As before, fitting the linear model $Y = J\beta_0 + \tilde{R}\xi + e$ with $\xi = (\xi_0, \ldots, \xi_5)$ is one way to produce an interpolator. The estimate of ξ_0 can be anything. ξ_0 is irrelevant to fitting the data because it corresponds to a column of 0s. A predicted value is

$$\begin{aligned}\hat{f}(t) = \hat{\beta}_0 + [R(t,t_0), \cdots, R(t,t_n)]\hat{\xi} \\ = 0 - 5R(t,t_1) + 2R(t,t_2) + 6R(t,t_3) - 3R(t,t_4) + 1R(t,t_5)\end{aligned}$$

because $R(t,t_0) = 0$, making the estimate of ξ_0 irrelevant. This is the solution we found in Example 11.1.2.

Now suppose $y_0 \neq 0$. We get a similar solution but $\hat{\beta}_0 = y_0$ with a different $\hat{\xi}_1$ (the other $\hat{\xi}_k$s remain the same). In particular, if $y_0 = 2$, we get $\hat{\xi}_1 = -25$ so that

$$\begin{aligned}y_1 = 0.1 = 2 - 25R(t_1,t_1) + 2R(t_1,t_2) + 6R(t_1,t_3) - 3R(t_1,t_4) + 1R(t_1,t_5) \\ = 2 + (-25 + 2 + 6 - 3 + 1)(0.1).\end{aligned}$$

Typically, for bivariate data (x_i, y_i), $i = 1, \ldots, n$, the predictor variable x_i is not between 0 and 1, although our procedure requires that it be. If we standardize the predictor so that

$$t_i = \frac{x_i - \min_k(x_k)}{\max_k(x_k) - \min_k(x_k)}$$

neither the nonuniqueness at the beginning of $\hat{f}(t)$ or the flatness of the end of $\hat{f}(t)$ remain issues in linear interpolation. In that case, as in Example 11.5.3, $t_1 = 0$, so the first column of $\tilde{R}$ is 0, the estimates of all parameters are unique except for ξ_1, but $R(t,t_1) = 0$ so ξ_1 is irrelevant to predictions. Of course, this standardization disallows any possibility of extrapolating the results beyond the observed data.

Proposition 11.5.6. For simple regression data (t_i, y_i), $i = 1, \ldots, n$ with strictly increasing t_i, there exists a minimum "linear spline" penalty interpolator for $f \in \mathscr{S}$ that has the form $\hat{f}(t) = \hat{\beta}_0 + \sum_{i=1}^n \hat{\xi}_i \min(t_i, t)$ with $\sum_{i=1}^n \hat{\xi}_i = 0$ and thus has $\hat{f}(t) = y_1$ for $0 \leq t \leq t_1$ and $\hat{f}(t) = y_n$ for $t_n \leq t \leq 1$. This minimizer is unique.

PROOF: We obtain interpolating coefficients (that minimize the sum of squared errors), say, $\hat{\beta}_0$ and $\hat{\xi}_i$s from solving $Y = J\beta_0 + \tilde{R}\xi$. Using the structure of $\tilde{R}$, if we set $\hat{\xi}_1 = 0$, there is a unique solution with, say, $\hat{\beta}_{00}$, but in general we merely have $\hat{\beta}_{00} = \hat{\beta}_0 + \hat{\xi}_1 t_1$ with no other restrictions on the choices of $\hat{\beta}_0$ and $\hat{\xi}_1$. Choose $\hat{\xi}_1$ to minimize the penalty by writing

$$Q(\hat{f}) = \hat{\xi}'\tilde{R}\hat{\xi} = \begin{bmatrix}\hat{\xi}_1 & \hat{\xi}_*'\end{bmatrix}\begin{bmatrix}t_1 & t_1 J_{n-1}' \\ t_1 J_{n-1} & \tilde{R}_2\end{bmatrix}\begin{bmatrix}\hat{\xi}_1 \\ \hat{\xi}_*\end{bmatrix}$$

or

$$Q(\hat{f}) = t_1\hat{\xi}_1^2 + 2t_1\hat{\xi}_1\hat{\xi}_*'J_{n-1} + \hat{\xi}_*'\tilde{R}_2\hat{\xi}_*.$$

For $t_1 = 0$, $\hat{\xi}_1$ can be anything, so it can be chosen so that the proposition holds, noting that $\hat{f}$ is unique regardless of how $\hat{\xi}_1$ is chosen. For $t_1 > 0$, to minimize the penalty as a function of $\hat{\xi}_1$ set the derivative equal to 0 yielding $0 = \hat{\xi}_1 + \hat{\xi}_*'J_{n-1} = \sum_{i=1}^n \hat{\xi}_i$. Since $\hat{f}$ is an interpolator,

$$y_1 = \hat{f}(t_1) = \hat{\beta}_0 + \sum_{i=1}^n \hat{\xi}_i \min(t_i, t_1) = \hat{\beta}_0 + \sum_{i=1}^n \hat{\xi}_i t_1 = \hat{\beta}_0.$$

and for $0 \le t \le t_1$,

$$\hat{f}(t) = \hat{\beta}_0 + \sum_{i=1}^n \hat{\xi}_i \min(t_i, t) = \hat{\beta}_0 + \sum_{i=1}^n \hat{\xi}_i t = \hat{\beta}_0.$$

Finally, any interpolator of this form based on $\min(t_i,t)$ has, for $t_n \le t \le 1$, $y_n = \hat{f}(t_n) = \hat{f}(t)$. □

It is pretty clear that we do not want to use this procedure to extrapolate the data!

In general, to fit linear smoothing splines with $\lambda > 0$ using a regression program, we need to incorporate

$$\tilde{S} = \begin{bmatrix} \sqrt{t_1} & \sqrt{t_1} & \sqrt{t_1} & \cdots & \sqrt{t_1} & \sqrt{t_1} \\ 0 & \sqrt{t_2 - t_1} & \sqrt{t_2 - t_1} & \cdots & \sqrt{t_2 - t_1} & \sqrt{t_2 - t_1} \\ 0 & 0 & \sqrt{t_3 - t_2} & \cdots & \sqrt{t_3 - t_2} & \sqrt{t_3 - t_2} \\ \vdots & \vdots & & \ddots & & \vdots \\ 0 & 0 & 0 & \cdots & 0 & \sqrt{t_n - t_{n-1}} \end{bmatrix}$$

into model (9).

Exercise 11.4. Fit linear smoothing splines to the data of Example 11.5.2 using $\lambda = 0.001, 0.01, 0.1, 0.5, 1, 2$ and compare the results. Graph the results.

11.5.4 General Solution Applied to Ridge Regression

In Chapter 10, we used standard linear model ideas to solve the generalized ridge regression problem. In this section we used the generalized ridge regression results of Chapter 10, along with RKHS theory, to solve the general penalized estimation problem. Now, somewhat circularly, we demonstrate how the general penalized estimation results apply to the classical ridge regression problem.

To solve the classical ridge problem we use the same notation as in Example 11.4.1, which is that of Chapter 10 and PA Section 6.2,

$$Y = X\beta + e = [J, Z] \begin{bmatrix} \beta_0 \\ \beta_* \end{bmatrix} + e$$

where the rows of Z are the predictor vectors x_i, $i = 1, \ldots, n$. In particular, the classical ridge regression problem estimates β_0 and β_* by minimizing

$$[Y - J\beta_0 - Z\beta_*]'[Y - J\beta_0 - Z\beta_*] + \lambda n \beta_*' \beta_*$$

We will see that RKHS theory essentially rewrites the model as $Y = [J, ZZ'][\beta_0, \xi] + e$ before solving the generalized ridge regression problem of minimizing

$$[Y - J\beta_0 - ZZ'\xi]'[Y - J\beta_0 - ZZ'\xi] + \lambda n \xi' ZZ' \xi$$

in which, quite clearly, $\beta_* = Z'\xi$.

To put the ridge regression problem in the framework of Theorem 11.5.2, reconsider Example 11.3.14. Take

$$\mathscr{M} = \mathscr{F}_* = \left\{ f : f(x) = \beta_0 + \sum_{j=1}^{p-1} \beta_j x_j \right\}. \tag{12}$$

with the penalty function

$$Q(f) = \sum_{j=1}^{p-1} \beta_j^2.$$

Take $\mathscr{N}$ as $\mathscr{F}_0$ and $\mathscr{H}$ as $\mathscr{F}_0^{\perp}$ (which is $\mathscr{F}_1$ from Example 11.3.13). The dimension of $\mathscr{N} = \mathscr{N}_0$ is $s = 1$ with $\psi_1(x) = 1$. The r.k. on $\mathscr{H}$ comes from Example 11.3.13 and is

$$R(\tilde{x}, x) = x'\tilde{x}.$$

All of this leads us to

$$W = J, \qquad \tilde{R} = ZZ', \qquad \tilde{S} = Z'$$

in the minimization (7) and the model (9)

The estimates from (8) reduce to

$$\hat{\xi} = \left[ZZ' \left(I - \frac{1}{n} JJ' \right) ZZ' + \lambda n ZZ' \right]^{-} ZZ' \left(I - \frac{1}{n} JJ' \right) Y, \qquad \hat{\beta}_0 = (J'J)^{-1} J'(Y - ZZ'\hat{\xi})$$

or

$$\hat{\xi} = \left\{ Z \left[Z' \left(I - \frac{1}{n} JJ' \right) Z + \lambda n I \right] Z' \right\}^{-} ZZ' \left(I - \frac{1}{n} JJ' \right) Y, \qquad \hat{\beta}_0 = \frac{1}{n} J'(Y - ZZ'\hat{\xi}).$$

It is easy to see by checking the definition of a generalized inverse that we can take

$$\left\{ Z \left[Z' \left(I - \frac{1}{n} JJ' \right) Z + \lambda n I \right] Z' \right\}^{-} = Z(Z'Z)^{-1} \left[Z' \left(I - \frac{1}{n} JJ' \right) Z + \lambda n I \right]^{-1} (Z'Z)^{-1} Z'$$

so

$$\begin{aligned}\hat{\xi} &= Z(Z'Z)^{-1}\left[Z'\left(I-\frac{1}{n}JJ'\right)Z+\lambda nI\right]^{-1}(Z'Z)^{-1}Z'ZZ'\left(I-\frac{1}{n}JJ'\right)Y\\ &= Z(Z'Z)^{-1}\left[Z'\left(I-\frac{1}{n}JJ'\right)Z+\lambda nI\right]^{-1}Z'\left(I-\frac{1}{n}JJ'\right)Y\\ &= Z(Z'Z)^{-1}\hat{\gamma}\end{aligned}$$

where $\hat{\gamma}$ is the classical ridge estimator from (10.2.3) in which "classical" means that $Q = I$ in that formula. With $\beta_* = Z'\xi$,

$$\hat{\beta}_{*R} = Z'\hat{\xi} = Z'Z(Z'Z)^{-1}\hat{\gamma} = \hat{\gamma}.$$

This version of $\hat{\xi}$ leads to

$$\hat{\beta}_{0R} = \frac{1}{n}J'[Y - ZZ'Z(Z'Z)^{-1}\hat{\gamma}] = \frac{1}{n}J'(Y - Z\hat{\beta}_{*R}).$$

which also agrees with the classical ridge estimates from (10.2.3). Finally we have

$$\hat{Y} = \hat{\beta}_{0R}J + ZZ'\hat{\xi} = \hat{\beta}_{0R}J + ZZ'\left[Z(Z'Z)^{-1}\hat{\gamma}\right] = \hat{\beta}_{0R}J + Z\hat{\beta}_{*R}.$$

In particular, for prediction at a new vector x_0 we have

$$\hat{f}(x_0) = \hat{\beta}_0 + [R(x_0,x_1),\cdots,R(x_0,x_1)]\hat{\xi} = \hat{\beta}_{0R} + x_0'Z'\hat{\xi} = \hat{\beta}_{0R} + x_0'\hat{\beta}_{*R}.$$

11.5.5 General Solution Applied to Cubic Smoothing Splines

Consider again the bivariate regression problem $y_i = f(t_i) + \varepsilon_i$, $i = 1,2,\ldots,n$ where $t_i \in [0,1]$ and $\mathrm{E}(\varepsilon_i) = 0$. Now consider the cubic smoothing spline estimates obtained by finding a function that minimizes

$$\sum_{i=1}^{n}[y_i - f(t_i)]^2 + \lambda n\int \ddot{f}(u)^2 du.$$

For the general solution we take $\mathscr{M}$ as the subset of $\mathscr{L}^2[0,1]$ with square integrable second derivatives. We take $Q(f)$ as in (11.3.2) which is also the inner product on $\mathscr{H} = \mathscr{W}_0^2$ from Example 11.3.11. The r.k. was given in Example 11.3.11 as equation (11.3.3) and is also given later. With this Q, the space $\mathscr{N}$ is all functions with $\ddot{f} = 0$. We take $\mathscr{N}_0$ to be linear functions $f_0(t) = \beta_0 + \beta_1 t$, so basis functions are $\psi_1(t) = 1$ and $\psi_2(t) = t$.

This structure leads to

$$W = \begin{bmatrix} 1 & t_1 \\ \vdots & \vdots \\ 1 & t_n \end{bmatrix}$$

and an $\tilde{R}$ that is ugly, but easily computed, in the minimization (7) and the model (9). Using

$$R(s,t) = \int_0^1 (t-u)_+(s-u)_+ du = \frac{\max(s,t)\min^2(s,t)}{2} - \frac{\min^3(s,t)}{6},$$

it is well known that the basis functions $R(t_h,t)$, that determine the columns $\tilde{R}_h$ of $\tilde{R}$ when evaluated at the data points t_i, form a natural cubic spline with knots at the distinct values of t_h. See Wahba (1990) for a justification, which just involves some algebra. The $\max(t_h,t)$ and $\min(t_h,t)$ in $R(t_h,t)$ combine in a way to produce knots at the t_h, while the degree of the polynomial spline is clearly three, since it is the highest power present in $R(s,t)$. This is the reason that the minimization problem in this section has been given the name *cubic smoothing spline*.

Nosedal-Sanchez et al. (2012) demonstrate with R code the fitting of cubic smoothing splines on some real data. The demonstration also includes searching for the best value of the tuning parameter λ which is briefly discussed in the next section.

11.6 Choosing the Degree of Smoothness

With the penalized regression procedures described above, the choice of the smoothing parameter λ is an important issue. There are many methods available for this task, e.g., visual inspection of the fit; m-fold cross-validation, Kohavi (1995); AIC/unbiased risk estimation; generalized maximum likelihood, Wahba (1990); generalized cross-validation (GCV), Craven and Wahba (1979); among others. Nosedal-Sanchez et al. (2012) used the GCV approach which works as follows. Suppose that for fixed λ the estimates determine fitted values $\hat{Y} = A(\lambda)Y$. The GCV choice of λ is the minimizer of

$$V(\lambda) = \frac{\frac{1}{n}\|I - A(\lambda)Y\|^2}{\left\{\frac{1}{n}\mathrm{tr}[I - A(\lambda)]\right\}^2}.$$

For more details about GCV and other methods of finding λ see Golub, Heath, and Wahba (1979), Allen (1974), Wecker and Ansley (1983) and Wahba (1990).