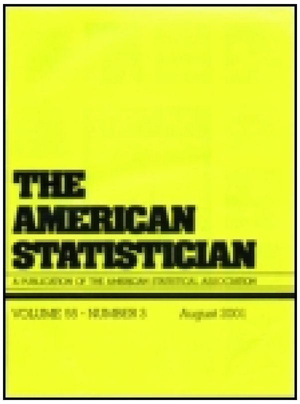

The American Statistician
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/utas20

Reproducing Kernel Hilbert Spaces for Penalized Regression: A Tutorial

Alvaro Nosedal-Sanchez [a] , Curtis B. Storlie [b] , Thomas C.M. Lee [c] & Ronald Christensen [d]

[a] Mathematics Department, Indiana University of Pennsylvania, Indiana, PA, 15705

[b] Statistical Sciences Group, Los Alamos National Laboratory, Los Alamos, NM, 87545

[c] Department of Statistics, University of California, Davis, Davis, CA, 95616

[d] Department of Mathematics and Statistics, University of New Mexico, Albuquerque, NM, 87131

Accepted author version posted online: 23 Mar 2012.Version of record first published: 12 Jun 2012.

To cite this article: Alvaro Nosedal-Sanchez , Curtis B. Storlie , Thomas C.M. Lee & Ronald Christensen (2012): Reproducing Kernel Hilbert Spaces for Penalized Regression: A Tutorial, The American Statistician, 66:1, 50-60

To link to this article: http://dx.doi.org/10.1080/00031305.2012.678196

Reproducing Kernel Hilbert Spaces for Penalized Regression: A Tutorial

Alvaro NOSEDAL-SANCHEZ, Curtis B. STORLIE, Thomas C.M. LEE, and Ronald CHRISTENSEN

Penalized regression procedures have become very popular ways to estimate complicated functions. The smoothing spline, for example, is the solution of a minimization problem in a functional space. If such a minimization problem is posed on a reproducing kernel Hilbert space (RKHS), the solution is guaranteed to exist, is unique, and has a very simple form. There are excellent books and articles about RKHS and their applications in statistics; however, this existing literature is very dense. This article provides a friendly reference for a reader approaching this subject for the first time. It begins with a simple problem, a system of linear equations, and then gives an intuitive motivation for reproducing kernels. Armed with the intuition gained from our first examples, we take the reader from vector spaces to Banach spaces and to RKHS. Finally, we present some statistical estimation problems that can be solved using the mathematical machinery discussed. After reading this tutorial, the reader will be ready to study more advanced texts and articles about the subject, such as those by Wahba or Gu. Online supplements are available for this article.

KEY WORDS: Projection principle; Regularization; Representation Theorem; Ridge Regression; Smoothing Splines.

1. INTRODUCTION

Penalized regression procedures have become a very popular approach to estimating complex functions (Wahba 1990; Eubank 1999; Hastie, Tibshirani, and Friedman 2001). 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), to name a few. Penalized regression procedures use an estimator that is defined as the solution to a minimization problem. In any minimization problem, there are the following questions: Does the solution exist? If yes, is the solution unique? How can we find it? If the problem is posed in the reproducing kernel Hilbert space (RKHS) framework that we discuss below, then the solution is guaranteed to exist, is unique, and takes a particularly simple form.

Reproducing kernel Hilbert spaces (RKHS) and reproducing kernels (RK) play a central role in penalized regression. The purpose of this article is to provide a constructive tutorial for statisticians interested in learning about RKHS methods in regression before studying more advanced texts and articles about the subject, such as Wahba (1990), Gu (2002), or Pearce and Wand (2006). Pearce and Wand (2006) provided a review of the connection between penalized splines and support vector machines (SVMs) using the RKHS framework. This article is intended to complement that review by providing the reader with the necessary background about RKHS to fully understand how RKHS results are used in penalized regression problems. Much of the associated literature begins with picking an RK and goes from there. Reproducing kernels may be the beginning of an application but they are the end of a body of theory. This article explicates that body of theory in an effort to make its application to penalized regression (and hence SVMs) more lucid. We describe how to construct a kernel with the properties needed for a given application and how to use the properties of that kernel for penalized regression. We provide several examples to help motivate and solidify the concepts as well as a transparent justification for the so-called “kernel trick.”

In the first section of this tutorial, we present, and solve, simple problems while gently introducing key concepts. The remainder of the article is organized as follows. Section 2 takes the reader from a basic understanding of fields through Banach spaces and Hilbert spaces. In Section 3, we provide elementary theory of RKHS along with some examples. Section 4 discusses penalized regression with RKHS. Two specific examples involving ridge regression and smoothing splines are given, with code written in the R language (R Development Core Team 2005) to solidify the concepts. Some methods for

Alvaro Nosedal-Sanchez, Assistant Professor, Mathematics Department, Indiana University of Pennsylvania, Indiana, PA 15705 (E-mail: *anosedal@iup.edu*). Curtis B. Storlie, Technical Staff, Statistical Sciences Group, Los Alamos National Laboratory, Los Alamos, NM 87545 (E-mail: *storlie@lanl.gov*). Thomas C.M. Lee, Professor of Statistics, Department of Statistics, University of California, Davis, Davis, CA 95616 (E-mail: *tcmlee@ucdavis.edu*). Ronald Christensen, Professor of Statistics, Department of Mathematics and Statistics, University of New Mexico, Albuquerque, NM 87131 (E-mail: *fletcher@stat.unm.edu*). The authors are grateful to the referees and the associate editor for their comments, which led to a much improved version of the article. Raymond Wong also provided helpful comments on an earlier version of the article. Nosedal was partially supported by the National Council of Science and Technology of Mexico (CONACYT). Lee was partially supported by the National Science Foundation under grant DMS 1007520. Storlie’s work was funded by Los Alamos National Security (LANS), LLC, operator of the Los Alamos National Laboratory under contract no. DE-AC52-06NA25396 with the U.S. Department of Energy. This article is published under LA-UR-12-10169.

DOI: 10.1080/00031305.2012.678196

smoothing parameter selection are briefly mentioned. Section 5 contains some closing remarks.

1.1 Why We Care About RKHS

Before introducing new concepts, we present some simple illustrations of the tools used to solve problems in the RKHS framework. Consider solving the following system of linear equations:

$$x_1 + x_3 = 0 \tag{1}$$

$$x_2 = 1. \tag{2}$$

Clearly, the real-valued solutions to this system are the vectors $\mathbf{x}_*^t = (-\alpha, 1, \alpha)$ for $\alpha \in \Re$. Suppose we want to find the "smallest" solution. Under the usual squared norm $\|\mathbf{x}\|^2 = x_1^2 + x_2^2 + x_3^2$, the smallest solution is $\mathbf{x}_s^t = (0, 1, 0)$.

Now consider a more general problem. For a given $p \times n$ matrix $\mathbf{R}$ and $n \times 1$ matrix $\boldsymbol{\eta}$, solve

$$\mathbf{R}^t\mathbf{x} = \boldsymbol{\eta}, \tag{3}$$

where $\mathbf{R}^t$ is the transpose of $\mathbf{R}$, $\mathbf{x}$ and the columns of $\mathbf{R}$, say $\mathbf{R}_k$, $k = 1, 2, \ldots, n$, are all in $\Re^p$, and $\boldsymbol{\eta} \in \Re^n$. We wish to find the solution $\mathbf{x}_s$ that minimizes the norm $\|\mathbf{x}\| = \sqrt{\mathbf{x}^t\mathbf{x}}$. We solve the problem using concepts that extend to RKHS.

A solution $\mathbf{x}_*$ (not necessarily a minimum norm solution) exists whenever $\boldsymbol{\eta} \in C(\mathbf{R}^t)$. Here, $C(\mathbf{R}^t)$ denotes the column space of $\mathbf{R}^t$. Given one solution $\mathbf{x}_*$, all solutions $\mathbf{x}$ must satisfy

$$\mathbf{R}^t\mathbf{x} = \mathbf{R}^t\mathbf{x}_*$$

or

$$\mathbf{R}^t(\mathbf{x}_* - \mathbf{x}) = \mathbf{0}.$$

The vector $\mathbf{x}_*$ can be written uniquely as $\mathbf{x}_* = \mathbf{x}_0 + \mathbf{x}_1$, with $\mathbf{x}_0 \in C(\mathbf{R})$ and $\mathbf{x}_1 \in C(\mathbf{R})^\perp$, where $C(\mathbf{R})^\perp$ is the orthogonal complement of $C(\mathbf{R})$ (i.e., $\mathbf{x}_0^t\mathbf{x}_1 = 0$, with orthogonality defined more formally in Definition 2.6). Clearly, $\mathbf{x}_0$ is a solution because $\mathbf{R}^t(\mathbf{x}_* - \mathbf{x}_0) = \mathbf{R}^t\mathbf{x}_1 = \mathbf{0}$.

In fact, $\mathbf{x}_0$ is both the unique solution in $C(\mathbf{R})$ and the minimum norm solution. If $\mathbf{x}$ is any other solution in $C(\mathbf{R})$, then $\mathbf{R}^t(\mathbf{x} - \mathbf{x}_0) = \mathbf{0}$ so we have both $(\mathbf{x} - \mathbf{x}_0) \in C(\mathbf{R})^\perp$ and $(\mathbf{x} - \mathbf{x}_0) \in C(\mathbf{R})$, two sets whose intersection is only the $\mathbf{0}$ vector. Thus, $\mathbf{x} - \mathbf{x}_0 = \mathbf{0}$ and $\mathbf{x} = \mathbf{x}_0$. In other words, every solution $\mathbf{x}_*$ has the same $\mathbf{x}_0$ vector. Finally, $\mathbf{x}_0$ is also the minimum norm solution because the arbitrary solution $\mathbf{x}_*$ has

$$\mathbf{x}_0^t\mathbf{x}_0 \le \mathbf{x}_0^t\mathbf{x}_0 + \mathbf{x}_1^t\mathbf{x}_1 = \mathbf{x}_*^t\mathbf{x}_*.$$

We have established the existence of a unique, minimum norm solution in $C(\mathbf{R})$ that can be written as

$$\mathbf{x}_s \equiv \mathbf{x}_0 = \mathbf{R}\xi = \sum_{k=1}^{n} \xi_k \mathbf{R}_k, \tag{4}$$

for some ξ_k, $k = 1, \ldots, n$. To find $\mathbf{x}_s$ explicitly, write $\mathbf{x}_s = \mathbf{R}\xi$ and the defining Equation (3) becomes

$$\mathbf{R}^t\mathbf{R}\boldsymbol{\xi} = \boldsymbol{\eta}, \tag{5}$$

which is just a system of linear equations. Even if there exist multiple solutions $\boldsymbol{\xi}$, $\mathbf{R}\boldsymbol{\xi}$ is unique.

Now we use this framework to find the "smallest" solution to the system of Equations (1) and (2). In the general framework, we have

$$\begin{aligned} \mathbf{x}^t &= (x_1, x_2, x_3), \\ \boldsymbol{\eta}^t &= (0, 1), \\ \mathbf{R}_1^t &= (1, 0, 1), \\ \mathbf{R}_2^t &= (0, 1, 0). \end{aligned}$$

We know that the solution has the form (4) and we also know that we have to solve a system of equations given by (5). In this case, the system of equations is

$$\begin{aligned} 2\xi_1 + 0\xi_2 &= 0, \\ 0\xi_1 + 1\xi_2 &= 1. \end{aligned}$$

The solution to the system is $(\xi_1, \xi_2) = (0, 1)$, which implies that our solution to the original problem is $\mathbf{x}_s = 0\mathbf{R}_1 + 1\mathbf{R}_2 = (0, 1, 0)^t$, as expected.

Virtually, the same methods can be used to solve a similar problem in any inner-product space Ω. As discussed later, an inner product $\langle \cdot, \cdot \rangle$ assigns real numbers to pairs of "vectors." For given vectors $\mathbf{R}_k \in \Omega$ and numbers $\eta_k \in \Re$, find $\mathbf{x} \in \Omega$ such that

$$\langle \mathbf{R}_k, \mathbf{x} \rangle = \eta_k, \qquad k = 1, 2, \ldots, n, \tag{6}$$

for which the norm of $\|\mathbf{x}\| \equiv \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$ is minimal. The solution has the form

$$\mathbf{x}_s = \sum_{k=1}^{n} \xi_k \mathbf{R}_k, \tag{7}$$

with ξ_k satisfying the linear equations

$$\sum_{k=1}^{n} \langle \mathbf{R}_i, \mathbf{R}_k \rangle \xi_i = \eta_i, \qquad i = 1, \ldots, n.$$

For a formal proof, see Máté (1990, p. 70). In RKHS applications, vectors are typically functions. We now apply this result to the interpolating spline problem.

1.2 Interpolating Splines

Suppose we want to find a function $f(t)$ that interpolates between the points (t_k, η_k), $k = 0, 1, 2, \ldots, n$, where $\eta_0 \equiv 0$ and $0 = t_0 < t_1 < \cdots < t_n = 1$. We restrict attention to functions $f \in F$ where F=$\{f$: f is absolutely continuous on $[0, 1]$, $f(0) = 0$, $\int_0^1 [f'(t)]^2 dt < \infty$ $\}$. Throughout, $f^{(m)}$ denotes the mth derivative of f, with $f' \equiv f^{(1)}$ and $f'' \equiv f^{(2)}$. The restriction that $\eta_0 = f(0) = 0$ is not really necessary, but simplifies the presentation.

We want to find the smoothest function $f(t)$ that satisfies $f(t_k) = \eta_k$, $k = 1, \ldots, n$. Defining an inner product on F by

$$\langle f, g \rangle = \int_0^1 f'(x)g'(x)dx$$

implies a norm over the space F that is small for "smooth" functions. To address the interpolation problem, note that the functions $R_k(s) \equiv \min(s, t_k)$, $k = 1, 2, \ldots, n$ have $R_k(0) = 0$ and the property that $\langle R_k, f\rangle = f(t_k)$ because

$$\begin{aligned}\langle f, R_k\rangle &= \int_0^1 f'(s)R_k'(s)ds\\ &= \int_0^{t_k} f'(s)1ds + \int_{t_k}^1 f'(s)0ds\\ &= \int_0^{t_k} f'(s)ds = f(t_k) - f(0) = f(t_k).\end{aligned}$$

Thus, an interpolator f satisfies a system of equations such as (6), namely

$$f(t_k) = \langle R_k, f\rangle = \eta_k, \quad k = 1, \dots, n, \tag{8}$$

and by (7), the smoothest function f (minimum norm) that satisfies the requirements has the form

$$\hat{f}(t) = \sum_{k=1}^{n} \xi_k R_k(t).$$

The ξ_j's are the solutions to the system of real linear equations obtained by substituting $\hat{f}$ into (8),

$$\sum_{j=1}^{n} \langle R_k, R_j\rangle \xi_j = \eta_k, \quad k = 1, 2, \dots, n.$$

Note that

$$\langle R_k, R_j\rangle = R_j(t_k) = R_k(t_j) = \min(t_k, t_j)$$

and define the function

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

which turns out to be an RK.

1.2.1 Numerical Example

Given points $f(t_i) = \eta_i$, 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 find

$$\arg\min_{f\in F} \|f\|^2 = \int_0^1 f'(x)^2 dx.$$

The system of equations is

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

The solution is $\xi = (-5, 2, 6, -3, 1)^t$, which implies that our function is

$$\hat{f}(t) = -5R_1(t) + 2R_2(t) + 6R_3(t) - 3R_4(t) + 1R_5(t) \tag{9}$$

$$\begin{aligned}&= -5R(t, t_1) + 2R(t, t_2) + 6R(t, t_3) - 3R(t, t_4)\\ &\quad + 1R(t, t_5)\end{aligned} \tag{10}$$

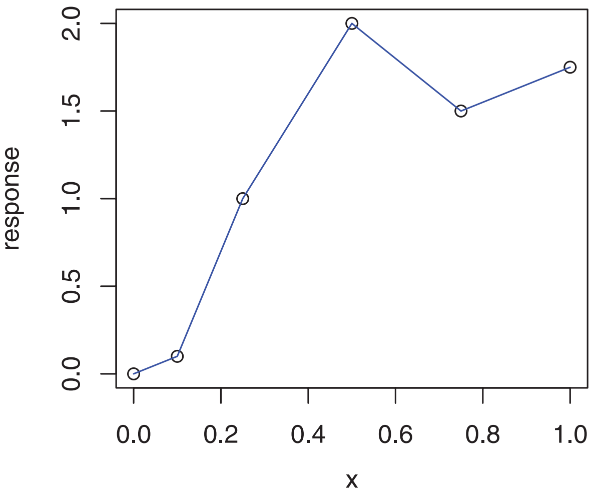

Figure 1. Linear interpolating spline. The online version of this figure is in color.

or, adding 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 1.

For this illustration, we restricted f so that $f(0) = 0$. This was 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 in general, the solution takes the form $\hat{f}(t) = \xi_0 + \sum_{j=1}^{n} \xi_j R_j(t)$, where $\xi_0 = \eta_0$.

The key points are (a) the elements R_i that allow us to express a function evaluated at a point as an inner-product constraint, and (b) the restriction to functions in F. F is a very special function space, an RKHS, and R_i is determined by an RK R.

Ultimately, our goal is to address more complicated regression problems like the linear smoothing spline problem.

1.2.2 Linear Smoothing Spline Problem

Consider simple regression data (x_i, y_i) with $0 \le x_i \le 1$, $i = 1, \dots, n$, and finding the function that minimizes

$$\frac{1}{n}\sum_{i=1}^{n}\{y_i - f(x_i)\}^2 + \lambda \int_0^1 f'(x)^2 dx. \tag{11}$$

If $f(x)$ is restricted to be in some class of functions F, minimizing only the first term gives least-squares estimation within F. If F contains functions with $f(x_i) = y_i$ for all i, such functions minimize the first term but are typically very "unsmooth," that is, have a large second term. The second "penalty" term is minimized by having a horizontal line, but that rarely has a small first term. As we will see in Section 4, for suitable F, the minimizer takes the form

$$\hat{f}(x) = \xi_0 + \sum_{i=1}^{n} \xi_i R_i(x),$$

where the R_i's are known functions and the ξ_i's are coefficients found by solving a system of linear equations. This produces a linear smoothing spline.

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. The linear spline penalty can be replaced by any other penalty with an associated inner product, and the x_i's can be vectors in $\Re^p$. Using RKHS results, we can solve the general problem of finding the minimizer of $\frac{1}{n}\sum_{i=1}^{n}(y_i - f(\mathbf{x}_i))^2 + \lambda J(f)$ for a general functional J that corresponds to a squared norm in a subspace. See Wahba (1990) or Gu (2002) for a full treatment of this approach. We now present an introduction to this theory.

2. VECTOR, BANACH, AND HILBERT SPACES

This section presents background material required for the formal development of the RKHS framework.

2.1 Vector Spaces

A vector space is a set that contains elements called "vectors" and supports two kinds of operations: addition of vectors and multiplication by scalars. The scalars are drawn from some field (the real numbers in the rest of this article) and the vector space is said to be a vector space over that field. Formally, a set V is a vector space over a field F if there exists a structure $\{V, F, +, \times, 0_v\}$ consisting of V, F, a vector addition operation $+$, a scalar multiplication $\times$, and an identity element $0_v \in V$. This structure must obey the following axioms for any $\mathbf{u}, \mathbf{v}, \mathbf{w} \in V$ and $a, b \in F$:

- Associative law: $(\mathbf{u} + \mathbf{v}) + \mathbf{w} = \mathbf{u} + (\mathbf{v} + \mathbf{w})$.
- Commutative law: $\mathbf{u} + \mathbf{v} = \mathbf{v} + \mathbf{u}$.
- Inverse law: $\exists \mathbf{s} \in V$ s.t. $\mathbf{u} + \mathbf{s} = 0_v$. (Write $-\mathbf{u} \equiv \mathbf{s}$.)
- Identity laws:
  - ○ $0_v + \mathbf{u} = \mathbf{u}$.
  - ○ $1 \times \mathbf{u} = \mathbf{u}$.
- Distributive laws:
  - ○ $a \times (b \times \mathbf{u}) = (a \times b) \times \mathbf{u}$.
  - ○ $(a + b) \times \mathbf{u} = a \times \mathbf{u} + b \times \mathbf{u}$.
  - ○ $a \times (\mathbf{u} + \mathbf{v}) = a \times \mathbf{u} + a \times \mathbf{v}$.

We will write $\mathbf{0}$ for $0_v \in V$ and $\mathbf{u} + (-\mathbf{v})$ as $\mathbf{u} - \mathbf{v}$. Any subset of a vector space that is closed under vector addition and scalar multiplication is called a subspace.

The simplest example of a vector space is just $\Re$ itself, which is a vector space over $\Re$. Vector addition and scalar multiplication are just addition and multiplication on $\Re$. For more on vector spaces and the other topics to follow in this section, see Naylor and Sell (1982), Young (1988), Máté (1990), and Rustagi (1994).

2.2 Banach Spaces

A Banach space has a level of additional structure over that required to be a vector space. It is a vector space that also has a distance measure called a "norm" and is "complete" under that norm. Defining a Banach space sets the stage for defining a Hilbert space, which involves an additional bit of structure (an "inner product") beyond that required to be a Banach space.

Definition 2.1. A norm of a vector space V, denoted by $||\cdot||$, is a nonnegative real-valued function satisfying the following properties for all $\mathbf{u}, \mathbf{v} \in V$ and all $a \in \Re$:

(1) Nonnegative: $||\mathbf{u}|| \geq 0$
(2) Strictly positive: $||\mathbf{u}|| = 0$ implies $\mathbf{u} = 0$
(3) Homogeneous: $||a\mathbf{u}|| = |a|\ ||\mathbf{u}||$
(4) Triangle inequality: $||\mathbf{u} + \mathbf{v}|| \leq ||\mathbf{u}|| + ||\mathbf{v}||$

Definition 2.2. A vector space is called a normed vector space when a norm is defined on the space.

Definition 2.3. A sequence $\{\mathbf{v}_n\}$ in a normed vector space V is said to converge to $\mathbf{v}_0 \in V$ if

$$\lim_{n\to\infty} ||\mathbf{v}_n - \mathbf{v}_0|| = 0.$$

Definition 2.4. A sequence $\{\mathbf{v}_n\} \subset V$ is called a Cauchy sequence if for any given $\epsilon > 0$, there exists an integer N such that

$$||\mathbf{v}_m - \mathbf{v}_n|| < \epsilon, \quad \text{whenever} \quad m, n \geq N.$$

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

Definition 2.5 (Banach space). A normed vector space V is called complete if every Cauchy sequence in V converges to an element of V. A complete normed vector space is called a Banach space.

Example 2.1. $\Re$ with the absolute value norm $\|x\| \equiv |x|$ is a complete, normed vector space over $\Re$, and is thus a Banach space.

Example 2.2. Let $\mathbf{x} = (x_1, \ldots, x_n)^t$ be a point in $\Re^n$. The l_p norm on $\Re^n$ is defined by

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

One can verify properties (1)–(4) at the beginning of this section for each p, validating that $||\mathbf{x}||_p$ is a norm on $\Re^n$. Under the l_p norm, $\Re^n$ is complete and thus a Banach space.

2.3 Hilbert Spaces

A Hilbert space is a Banach space in which the norm is defined by an inner product (also called dot product), which we define next. We typically denote Hilbert spaces by H. For elements $\mathbf{u}, \mathbf{v} \in H$, write the inner product of $\mathbf{u}$ and $\mathbf{v}$ either as $\langle \mathbf{u}, \mathbf{v} \rangle_H$ or, when it is clear by context that the inner product is taking place in H, as $\langle \mathbf{u}, \mathbf{v} \rangle$. If H is a vector space over F, the result of the inner product is an element in F. We have $F = \Re$, so the result of an inner product will be a real number. The inner-product operation must satisfy four properties for all $\mathbf{u}, \mathbf{v}, \mathbf{w} \in H$ and all $a \in F$.

(1) Associative: $\langle a\mathbf{u}, \mathbf{v}\rangle = a\langle \mathbf{u}, \mathbf{v}\rangle$.
(2) Commutative: $\langle \mathbf{u}, \mathbf{v}\rangle = \langle \mathbf{v}, \mathbf{u}\rangle$.
(3) Distributive: $\langle \mathbf{u}, \mathbf{v} + \mathbf{w}\rangle = \langle \mathbf{u}, \mathbf{v}\rangle + \langle \mathbf{u}, \mathbf{w}\rangle$.
(4) Positive definite: $\langle \mathbf{u}, \mathbf{u}\rangle \geq 0$, with equality holding only if $\mathbf{u} = \mathbf{0}$.

Definition 2.6. A vector space with an inner product defined on it is called an inner-product space. The norm of an element u in an inner-product space is taken as $||\mathbf{u}|| = \langle \mathbf{u}, \mathbf{u}\rangle^{1/2}$. Two vectors are said to be orthogonal if their inner product is 0, and two sets of vectors are said to be orthogonal if every vector in one is orthogonal to every vector in the other. The set of all vectors orthogonal to a subspace is called the orthogonal complement of the subspace. A complete inner-product space is called a Hilbert space.

Example 2.3. $\Re^n$ with inner product defined by

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

is a Hilbert space. For any positive definite matrix $\mathbf{A}$, $\langle \mathbf{u}, \mathbf{v}\rangle \equiv \mathbf{u}^t\mathbf{A}\mathbf{v}$ also defines a valid inner product.

Example 2.4. Let $L^2(a, b)$ be the vector space of all real-valued functions defined on the interval (a, b) that are square integrable and define the inner product

$$\langle f, g\rangle \equiv \int_a^b f(x)g(x)dx.$$

The inner-product space $L^2(a, b)$ is well known to be complete (see de Barra 1981); thus, $L^2(a, b)$ is a Hilbert space.

3. REPRODUCING KERNEL HILBERT SPACES (RKHS)

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

Definition 3.1. A function T mapping a vector space X into another vector space Y is called a linear operator if $T(\lambda_1\mathbf{x}_1 + \lambda_2\mathbf{x}_2) = \lambda_1 T(\mathbf{x}_1) + \lambda_2 T(\mathbf{x}_2)$ for any $\mathbf{x}_1, \mathbf{x}_2 \in X$ and any $\lambda_1, \lambda_2 \in \Re$.

Any $m \times n$ matrix $\mathbf{A}$ maps vectors in $\Re^n$ into vectors in $\Re^m$ via $\mathbf{A}\mathbf{x} = \mathbf{y}$ and is linear.

Definition 3.2. The operator $T : X \to Y$ mapping a Banach space into a Banach space is continuous at $\mathbf{x}_0 \in X$ if and only if for every $\epsilon > 0$, there exists $\delta = \delta(\epsilon) > 0$ such that for every $\mathbf{x}$ with $||\mathbf{x} - \mathbf{x}_0|| < \delta$, we have $||T\mathbf{x} - T\mathbf{x}_0|| < \epsilon$.

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

Definition 3.3. A real-valued function defined on a vector space is called a functional.

A $1 \times n$ matrix defines a linear functional on $\Re^n$.

Example 3.1. Let S be the set of bounded real-valued continuous functions $\{f(x)\}$ defined on the real line. Then, S is a vector space with the usual + and × operations for functions. Some functionals on S are $\phi(f) = \int_a^b f(x)dx$ and $\phi_a(f) = f'(a)$ for some fixed a and b. A functional of particular importance is the evaluation functional.

Definition 3.4. Let V be a vector space of functions defined from E into $\Re$. For any $t \in E$, denote by e_t the evaluation functional at the point t; that is, for $g \in V$, the mapping is $e_t(g) = g(t)$.

For $V = \Re^p$, vectors can be viewed as functions from the set $E = \{1, 2, \ldots, p\}$ into $\Re$. An evaluation functional is $e_i(\mathbf{x}) = x_i$. Clearly, evaluation functionals are linear operators.

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 H of real-valued functions:

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

Here, (ii) is the definition of (i). See Máté (1990, p. 123) for a proof of (iii).

To define an RK, we need the famous *Riesz representation theorem.*

Theorem 3.1. Let H be a Hilbert space and let ϕ be a continuous linear functional on H. Then, there is one and only one vector $g \in H$ such that

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

The vector g is sometimes called the *representation* of ϕ. However, ϕ and g are different objects: ϕ is a linear functional on H and g is a vector in H. For a proof of this theorem, see Naylor and Sell (1982) or Máté (1990, p. 84).

Recall, for $H = \Re^p$, an evaluation functional is $e_i(\mathbf{x}) = x_i$. The representation of this linear functional is the indicator vector $\mathbf{e}_i$ that is 0 everywhere except has a 1 in the ith place. Then,

$$x_i = e_i(\mathbf{x}) = \mathbf{x}^t\mathbf{e}_i.$$

In fact, the entire representation theorem is well known in $\Re^p$ because for $\phi(\mathbf{x})$ to be a linear functional, there must exist a vector $\boldsymbol{\phi}$ such that

$$\phi(\mathbf{x}) = \boldsymbol{\phi}^t\mathbf{x}.$$

An element of a set of functions, say f, is sometimes denoted $f(\cdot)$ to be explicit that the elements are functions, whereas $f(t)$ is the value of $f(\cdot)$ evaluated at $t \in E$. Applying the Riesz representation theorem to a Hilbert space H of real-valued functions in which evaluation functionals are continuous, for every $t \in E$, there is a unique symmetric function $R : E \times E \to \Re$, with $R(\cdot, t) \in H$ the representation of e_t, so that

$$f(t) = e_t(f) = \langle f(\cdot), R(\cdot, t)\rangle_H, \quad f \in H.$$

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

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

In Section 1.2, we found the RK for the linear interpolating spline problem, $R(s, t) = \min(s, t)$. For any fixed t, $R(\cdot, t)$ is a part of the space F defined there, since $R(0, t) = 0$ and $\int_0^1 [\partial R(s, t)/\partial s]^2 ds < \infty$. Also, R has the reproducing property since

$$\begin{aligned}\langle f, R(\cdot, t)\rangle &= \int_0^1 [f'(s)\partial R(s, t)/\partial s]ds \\ &= \int_0^t f'(s)1ds + \int_t^1 f'(s)0ds \\ &= f(t) - f(0) = f(t).\end{aligned}$$

Many other, more detailed, examples involving RKHS and their RK's follow, but first, the formal definition of an RKHS is presented next.

Definition 3.5. A Hilbert space H of functions defined on E is called a reproducing kernel Hilbert space if all evaluation functionals are continuous.

We now present several RKHS examples that we will then use to solve some familiar penalized regression problems in Section 4.

3.1 Examples of RKHS

Example 3.2. Consider the space of all constant functionals over $\mathbf{x} = (x_1, x_2, \ldots, x_p)^t \in \Re^p$,

$$H = \{f_\theta : f_\theta(\mathbf{x}) = \theta, \theta \in \Re\},$$

with $\langle f_\theta, f_\lambda\rangle = \theta\lambda$. (For simplicity, think of $p = 1$.) Since $\Re^p$ is a Hilbert space, so is H. H has continuous evaluation functionals, so it is an RKHS and has a unique RK. To find the RK, observe that $R(\cdot, \mathbf{x}) \in H$, so it is a constant for any $\mathbf{x}$. Write $R(\mathbf{x}) \equiv R(\cdot, \mathbf{x})$. By the representation theorem and the defined inner product

$$\theta = f_\theta(\mathbf{x}) = \langle f_\theta(\cdot), R(\cdot, \mathbf{x})\rangle = \theta R(\mathbf{x})$$

for any $\mathbf{x}$ and θ. This implies that $R(\mathbf{x}) \equiv 1$ so that $R(\cdot, \mathbf{x}) = R(\mathbf{x}) \equiv 1$ and $R(\cdot, \cdot) \equiv 1$.

Example 3.3. Consider all linear functionals over $\mathbf{x} \in \Re^p$ passing through the origin

$$H = \{f_\theta : f_\theta(\mathbf{x}) = \boldsymbol{\theta}^t\mathbf{x},\ \boldsymbol{\theta} \in \Re^p\}.$$

Define $\langle f_\theta, f_\lambda\rangle = \boldsymbol{\theta}^t\boldsymbol{\lambda} = \theta_1\lambda_1 + \theta_2\lambda_2 + \cdots + \theta_p\lambda_p$. The kernel R must satisfy

$$f_\theta(\mathbf{x}) = \langle f_\theta(\cdot), R(\cdot, \mathbf{x})\rangle$$

for all $\boldsymbol{\theta}$ and any $\mathbf{x}$. Since $R(\cdot, \mathbf{x}) \in H$, $R(\mathbf{v}, \mathbf{x}) = \mathbf{u}^t\mathbf{v}$ for some $\mathbf{u}$ that depends on $\mathbf{x}$, that is, $R(\cdot, \mathbf{x}) = f_{u(x)}(\cdot)$, so $R(\mathbf{v}, \mathbf{x}) = \mathbf{u}(\mathbf{x})^t\mathbf{v}$. By our definition of H, we have

$$\boldsymbol{\theta}^t\mathbf{x} = f_\theta(\mathbf{x}) = \langle f_\theta(\cdot), R(\cdot, \mathbf{x})\rangle = \langle f_\theta(\cdot), f_{u(x)}(\cdot)\rangle = \boldsymbol{\theta}^t\mathbf{u}(\mathbf{x}),$$

so we need $\mathbf{u}(\mathbf{x})$ such that for any $\boldsymbol{\theta}$ and $\mathbf{x}$, we have

$$\boldsymbol{\theta}^t\mathbf{x} = \boldsymbol{\theta}^t\mathbf{u}(\mathbf{x}).$$

It follows that $\mathbf{u}(\mathbf{x}) = \mathbf{x}$. For example, taking $\boldsymbol{\theta}$ to be the indicator vector $\mathbf{e}_i$ implies that $u_i(\mathbf{x}) = x_i$ for every $i = 1, \ldots, p$. We now have $R(\cdot, \mathbf{x}) = f\mathbf{x}(\cdot)$ so that

$$R(\tilde{\mathbf{x}}, \mathbf{x}) = \mathbf{x}^t\tilde{\mathbf{x}} = x_1\tilde{x}_1 + x_2\tilde{x}_2 + \cdots + x_p\tilde{x}_p.$$

Before moving on to the next example, we present one more concept that is useful in RKHS approaches to regression problems.

3.1.1 *The Projection Principle for an RKHS*

Consider the connection between the RK R of the RKHS H and the RK R_0 for a subspace $H_0 \subset H$. Let $H_0^\perp$ be the orthogonal complement of H_0. Then, any vector $f \in H$ can be written uniquely as $f = f_0 + f_1$, with $f_0 \in H_0$ and $f_1 \in H_0^\perp$. More particularly, $R(\cdot, t) = R_0(\cdot, t) + R_1(\cdot, t)$, with $R_0(\cdot, t) \in H_0$ and $R_1(\cdot, t) \in H_0^\perp$ if and only if R_0 is the RK of H_0 and R_1 is the RK of $H_0^\perp$. For a proof, see Gu (2002).

Example 3.4. Now consider all affine (i.e., linear plus a constant) functionals in $\Re^p$,

$$H = \{f_\theta : f_\theta(\mathbf{x}) = \theta_0 + \theta_1 x_1 + \cdots + \theta_p x_p,\ \boldsymbol{\theta} \in \Re^{p+1}\},$$

with $\langle f_\theta, f_\lambda\rangle = \theta_0\lambda_0 + \theta_1\lambda_1 + \cdots + \theta_p\lambda_p$. The subspace $H_0 = \{f_\theta \in H : \theta_0 \in \Re, 0 = \theta_1 = \cdots = \theta_p\}$ has the orthogonal complement $H_0^\perp = \{f_\theta \in H : 0 = \theta_0\}$. For practical purposes, H_0 is the space of constant functionals from Example 3.2 and $H_0^\perp$ is the space of linear functionals from Example 3.3. Note that the inner product on H when applied to vectors in H_0 and $H_0^\perp$, respectively, reduces to the inner products used in Examples 3.2 and 3.3.

Write H as $H = H_0 \oplus H_0^\perp$, where $\oplus$ denotes the *direct sum* of two vector spaces. For two subspaces A and B contained in a vector space C, the direct sum is the space $D = \{a + b : a \in A, b \in B\}$. Any elements $d_1, d_2 \in D$ can be written as $a_1 + b_1$ and $a_2 + b_2$, respectively, for some $a_1, a_2 \in A$ and $b_1, b_2 \in B$. When the two subspaces are orthogonal, as in our example, those decompositions are unique and the inner product between d_1 and d_2 is $\langle d_1, d_2\rangle = \langle a_1, a_2\rangle + \langle b_1, b_2\rangle$. For more information about direct sum decomposition see, for example, Berlinet and Thomas-Agnan (2004) or Gu (2002).

We have already derived the RK's for H_0 and $H_0^\perp$ (call them R_0 and R_1, respectively) in Examples 3.2 and 3.3. Applying the projection principle, the RK for H is the sum of R_0 and R_1, that is,

$$R(\tilde{\mathbf{x}}, \mathbf{x}) = 1 + \mathbf{x}^t\tilde{\mathbf{x}}.$$

Example 3.5. Denote by V the collection of functions f with $f'' \in L^2[0, 1]$ and consider the subspace

$$W_2^0 = \{f(x) \in V : f, f' \text{ absolutely continuous and } f(0) = f'(0) = 0\}.$$

Define an inner product on W_2^0 as

$$\langle f, g\rangle = \int_0^1 f''(t)g''(t)dt. \tag{12}$$

Below, we demonstrate that for $f \in W_2^0$ and any s, $f(s)$ can be written as

$$f(s) = \int_0^1 (s - u)_+ f''(u)du, \tag{13}$$

where $(a)_+$ is a for $a > 0$ and 0 for $a \le 0$. Given any arbitrary and fixed $s \in [0, 1]$,

$$\int_0^1 (s-u)_+ f''(u)du = \int_0^s (s-u) f''(u)du \, .$$

Integrating by parts

$$\int_0^s (s-u) f''(u)du = (s-s)f'(s) - (s-0)f'(0) + \int_0^s f'(u)du = \int_0^s f'(u)du$$

and applying the fundamental theorem of calculus to the last term,

$$\int_0^s (s-u) f''(u)du = f(s) - f(0) = f(s) \, .$$

Since the RK of the space W_2^0 must satisfy $f(s) = \langle f(\cdot), R(\cdot, s)\rangle$ from (12) and (13), we observe that $R(\cdot, s)$ is a function such that

$$\frac{d^2 R(u,s)}{du^2} = (s-u)_+ .$$

We also know that $R(\cdot, s) \in W_2^0$, so using $R(s,t) = \langle R(\cdot, t), R(\cdot, s)\rangle$

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

For more examples of RKHS with various inner products, see Berlinet and Thomas-Agnan (2004).

4. PENALIZED REGRESSION WITH RKHS

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

$$y_i = f(\mathbf{x}_i) + \epsilon_i, \quad i = 1, 2, \ldots, n,$$

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

Ridge regression: In the classical linear regression setting $y_i = \mathbf{x}_i^t \boldsymbol{\beta} + \epsilon_i$, the ridge regression estimator $\hat{\boldsymbol{\beta}}_R$ proposed by Hoerl and Kennard (1970) minimizes

$$\frac{1}{n}\sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^p \beta_j^2, \qquad (14)$$

where x_{ij} is the ith observation of the jth component. The resulting estimate is biased but can reduce the variance relative to least-squares estimates. The tuning parameter $\lambda \ge 0$ is a constant that controls the tradeoff between bias and variance in $\hat{\boldsymbol{\beta}}_R$, and is often selected by some form of cross-validation; see Section 4.4.

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. It is often assumed, without loss of generality, that the domain of f is $[0, 1]$. With $f^{(m)}$ the mth derivative of f, a smoothing spline estimate $\hat{f}$ is the unique minimizer of

$$\frac{1}{n}\sum_{i=1}^n \{y_i - f(x_i)\}^2 + \lambda \int [f^{(m)}(x)]^2 dx. \qquad (15)$$

The minimization of (15) is implicitly over functions with square integrable mth derivatives. The first term of (15) encourages the fitted f to be close to the data, while the second term penalizes the roughness of f. The smoothing parameter λ, usually prespecified, controls the tradeoff between the two conflicting goals. The special case of $m = 1$ reduces to the linear smoothing spline problem from (11). In practice, it is common to choose $m = 2$, in which case the minimizer f_λ of (15) is called a cubic smoothing spline. As $\lambda \to \infty$, $\hat{f}_\lambda$ approaches the least-squares simple linear regression line, while as $\lambda \to 0$, $\hat{f}_\lambda$ approaches the minimum curvature interpolant.

4.1 Solving the General Penalized Regression Problem

We now review a general framework to minimize (14), (15), and many other similar criteria (cf. O'Sullivan, Yandell, and Raynor 1986; Lin and Zhang 2006; Storlie, Bondell, and Reich 2010; Storlie et al. 2010; Gu and Qiu 1993). The data model is

$$y_i = f(\mathbf{x}_i) + \epsilon_i, \quad i = 1, 2, \ldots, n, \qquad (16)$$

where the ϵ_i are error terms and $f \in V$, a given vector space of functions on a set E.

An estimate of f is obtained by minimizing

$$\frac{1}{n}\sum_{i=1}^n \{y_i - f(\mathbf{x}_i)\}^2 + \lambda J(f), \qquad (17)$$

over $f \in V$, where J is a penalty functional that must satisfy several restrictions and that helps to define the vector space V. We require (a) that $J(f) \ge 0$ for any f in some vector space $\tilde{V}$; (b) that the null set $N = \{f \in \tilde{V} : J(f) = 0\}$ be a subspace, that is, that it be closed under vector addition and scalar multiplication; (c) that for $f_N \in N$ and $f \in \tilde{V}$, we have $J(f_N + f) = J(f)$; and (d) that there exists an RKHS H contained in $\tilde{V}$ for which the inner product satisfies $\langle f, f\rangle = J(f)$. This condition forces the intersection of N and H to contain only the zero vector.

Define a finite-dimensional subspace of N, say N_0, with a basis of known functions, say $\{\phi_1, \ldots, \phi_M\}$, and $M \le n$. In applications, N is often finite-dimensional, and we can simply take $N_0 = N$. In the minimization problem, we restrict attention to $f \in V$, where

$$V \equiv N_0 \oplus H.$$

For Example 3.5, $\tilde{V}$ consists of functions in $L^2[0, 1]$ with finite values of

$$J(f) \equiv \int_0^1 \left[f''(t)\right]^2 dt.$$

$J(f)$ satisfies our four conditions with $H = W_2^0$. The linear functions $f(x) = a + bx$ are in N, so we can take $\phi_1(x) \equiv 1$ and $\phi_2(x) = x$. Note that Equation (12) defines an inner product on

W_2^0 but does not define an inner product on all of $\tilde{V}$ because nonzero functions could have a zero inner product with themselves, hence the nontrivial nature of N.

The key result [Wahba's representation theorem, also known as the "dual form" or "kernel trick" (Pearce and Wand 2006)] is that the minimizer of (17) is a linear combination of known functions involving the RK on H. This fact will allow us to find the coefficients of the linear combination by solving a quadratic minimization problem similar to those in standard linear models.

Representation theorem: The minimizer $\hat{f}_\lambda$ of Equation (17) has the form

$$\hat{f}_\lambda(\mathbf{x}) = \sum_{j=1}^{M} d_j \phi_j(\mathbf{x}) + \sum_{i=1}^{n} c_i R(\mathbf{x}_i, \mathbf{x}), \qquad (18)$$

where $R(\mathbf{s}, \mathbf{t})$ is the RK for H. An informal proof is given next, see Wahba (1990) or Gu (2002) for a formal proof.

Since we are working in V, clearly, a minimizer $\hat{f}$ must have $\hat{f} = \hat{f}_0 + \hat{f}_1$, with $\hat{f}_0 \in N_0$ and $\hat{f}_1 \in H$. We want to show that $\hat{f}_1(\cdot) = \sum_{i=1}^n c_i R(\mathbf{x}_i, \cdot)$. To simplify notation, write

$$\hat{f}_R(\cdot) \equiv \sum_{i=1}^{n} c_i R(\mathbf{x}_i, \cdot). \qquad (19)$$

Decompose H as $H = H_0 \oplus H_0^\perp$, where $H_0 = \text{span}\{R(\mathbf{x}_i, \cdot), i = 1, \ldots, n\}$ so that

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

with $\eta(\cdot) \in H_0^\perp$. By orthogonality and the reproducing property of the RK,

$$0 = \langle R(\mathbf{x}_i, \cdot), \eta(\cdot)\rangle = \eta(\mathbf{x}_i).$$

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

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\{y_i - \hat{f}(\mathbf{x}_i)\}^2 + \lambda J(\hat{f}) \\
&\quad = \frac{1}{n}\sum_{i=1}^{n}\{y_i - \hat{f}_0(\mathbf{x}_i) - \hat{f}_1(\mathbf{x}_i)\}^2 + \lambda J(\hat{f}_0 + \hat{f}_1) \\
&\quad = \frac{1}{n}\sum_{i=1}^{n}\{y_i - \hat{f}_0(\mathbf{x}_i) - \hat{f}_1(\mathbf{x}_i)\}^2 + \lambda J(\hat{f}_1) \\
&\quad = \frac{1}{n}\sum_{i=1}^{n}\{y_i - \hat{f}_0(\mathbf{x}_i) - \hat{f}_R(\mathbf{x}_i) - \eta(\mathbf{x}_i)\}^2 + \lambda J(\hat{f}_R + \eta).
\end{aligned}$$

Because $\eta(\mathbf{x}_i) = 0$ and using orthogonality within H,

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

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

A remarkable feature of the result in (18) is that the form of the minimizer is represented by a finite-dimensional basis, regardless of the dimension of H. For example, H could be all functions with second derivative in L^2, such as in the cubic smoothing spline problem. This H would require 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! So once we know that the minimizer takes the form (18), 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 $J(\hat{f}) = J(\hat{f}_R)$ as a quadratic form in $\mathbf{c} = (c_1, \ldots, c_n)^t$. Define Σ as the $n \times n$ matrix where the i, j entry is $\Sigma_{ij} = R(\mathbf{x}_i, \mathbf{x}_j)$. The matrix Σ is commonly referred to as the *Gram* matrix (Wahba 1990; Gu 2002). Now, using the reproducing property of R, write

$$\begin{aligned}
J(\hat{f}_R) &= \left\langle \sum_{i=1}^{n} c_i R(\mathbf{x}_i, \cdot), \sum_{j=1}^{n} c_j R(\mathbf{x}_j, \cdot) \right\rangle \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j R(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{c}^t \Sigma \mathbf{c}.
\end{aligned}$$

Define the observation vector $\mathbf{y} = [y_1, \ldots, y_n]^t$, and let $\mathbf{T}$ be the $n \times M$ matrix with the ijth entry defined by $\mathbf{T}_{ij} = \{\phi_j(\mathbf{x}_i)\}$. The minimization of (17) then takes the form

$$\min_{c,d} \frac{1}{n}||\mathbf{y} - (\mathbf{Td} + \Sigma\mathbf{c})||^2 + \lambda \mathbf{c}^t \Sigma \mathbf{c}. \qquad (20)$$

To solve (20), we define the following matrices:

$$\mathbf{Q}_{n\times(n+M)} = \begin{bmatrix} \mathbf{T}_{n\times M} & \Sigma_{n\times n} \end{bmatrix},$$

$$\boldsymbol{\gamma}_{(n+M)\times 1} = \begin{bmatrix} \mathbf{d}_{M\times 1} \\ \mathbf{c}_{n\times 1} \end{bmatrix},$$

and

$$\mathbf{S}_{(n+M)\times(n+M)} = \begin{bmatrix} \mathbf{0}_{M\times M} & \mathbf{0}_{M\times n} \\ \mathbf{0}_{n\times 1} & \Sigma_{n\times n} \end{bmatrix}.$$

Now, Equation (20) becomes

$$\min_{\gamma} \frac{1}{n}||\mathbf{y} - \mathbf{Q}\boldsymbol{\gamma}||^2 + \lambda \boldsymbol{\gamma}^{\mathbf{t}}\mathbf{S}\boldsymbol{\gamma}. \qquad (21)$$

The minimization in (21) is the same as a generalized ridge regression. Taking derivatives with respect to $\boldsymbol{\gamma}$, we have

$$(\mathbf{Q}^{\mathbf{t}}\mathbf{Q} + \lambda\mathbf{S})\hat{\boldsymbol{\gamma}} = \mathbf{Q}^{\mathbf{t}}\mathbf{y},$$

which requires solving a system of $n + M$ equations to find $\hat{\boldsymbol{\gamma}}$. For analytical purposes, we can write $\hat{\boldsymbol{\gamma}}$ as

$$\hat{\boldsymbol{\gamma}} = (\mathbf{Q}^{\mathbf{t}}\mathbf{Q} + \lambda\mathbf{S})^{-}\mathbf{Q}^{\mathbf{t}}\mathbf{y}. \qquad (22)$$

As long as $\mathbf{T}$ is of full-column rank, $\hat{f}_\lambda$ is unique. If the $\mathbf{x}_i$ are not unique, then $\hat{\boldsymbol{\gamma}}$ is not unique as defined, but $\hat{f}_\lambda$ is still unique. One could simply use only the unique $\mathbf{x}_i$ in the definition of f_R in (19) to ensure that $\hat{\boldsymbol{\gamma}}$ is unique in that case as well.

Alternatively, $\hat{\gamma}$ can be obtained as the generalized least-squares estimate from fitting the linear model

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{T} & \Sigma \\ \mathbf{0}_{n\times M} & \mathbf{I}_{n\times n} \end{bmatrix} \begin{bmatrix} \mathbf{d} \\ \mathbf{c} \end{bmatrix} + e,$$
$$\operatorname{cov}(e) \propto \begin{bmatrix} \mathbf{I}_{n\times n} & \mathbf{0}_{n\times n} \\ \mathbf{0}_{n\times n} & (1/\lambda)\Sigma^{-1}_{n\times n} \end{bmatrix}.$$

For clarity, we have restricted our attention to minimizing (17), 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).

4.2 General Solution Applied to Cubic Smoothing Spline

Consider again the regression problem $y_i = f(x_i) + \epsilon_i, i = 1, 2, \ldots, n$, where $x_i \in [0, 1]$ and $\epsilon_i \sim N(0, \sigma^2)$. We focus on the cubic smoothing spline solution to this problem. That is, we find a function that minimizes

$$\sum_{i=1}^{n} \{y_i - f(x_i)\}^2 + \lambda \int f''(x)^2 dx.$$

As discussed in Section 4.1, N is the space of linear functions from Example 3.4 with $p = 1$ [see also Equation (24)] and a basis of $\phi_1(x) = 1$, $\phi_2(x) = x$. In this case, N happens to be an RKHS, but in general, it is not even necessary to define an inner product on N. $H = W_2^0$ comes from Example 3.5. We know that the RK for H is

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

so the solution has the form

$$\hat{f}(x) = \hat{d}_0(1) + \hat{d}_1 x_i + \sum_{i=i}^{n} \hat{c}_i R(x_i, x).$$

From (22), we have

$$(\mathbf{Q}^t\mathbf{Q} + \lambda\mathbf{S})^{-1}\mathbf{Q}^t\mathbf{y} = \begin{bmatrix} \hat{\mathbf{d}} \\ \hat{\mathbf{c}} \end{bmatrix}. \tag{23}$$

The online supplementary material provides code written in the R language (R Development Core Team 2005) and plots for fitting the cubic smoothing spline solution in (23) to some motorcycle accident data. The demonstration also includes searching for the best value of the tuning parameter λ, which is briefly discussed in Section 4.4.

As an aside, it is well known that the basis functions $R(x_i, x)$ (that form the columns of $\mathbf{Q}$ when evaluated at the data points) form a natural cubic spline with knots at the distinct values of x_i; see Wahba (1990) for a justification, which just involves some algebra. The $\max(x_i, x)$ and $\min(x_i, x)$ in $R(x_i, x)$ combine in a way to produce knots at the x_i, while the degree of the polynomial spline would clearly be 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."

4.3 General Solution Applied to Ridge Regression

We now solve the linear ridge regression problem of minimizing (14) with the RKHS approach detailed earlier. Although the RKHS framework is not necessary to solve the ridge regression problem, it serves as a good illustration of the RKHS machinery.

To put the ridge regression problem in the framework of (17), consider

$$\tilde{V} = \left\{ f(\mathbf{x}) = \beta_0 + \sum_{j=1}^{p} \beta_j x_j \right\}, \tag{24}$$

with the penalty function

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

Note that by letting $\tilde{V} = V_0 \oplus H$, where V_0 is the RKHS from Example 3.2 and H is from Example 3.3, we have $J(f) = \langle f, f \rangle_H$. The dimension of $V_0 = N = N_0$ is $M = 1$ and $\phi_1(x) = 1$.

From (18), the solution takes the form

$$\hat{f}_\lambda(\mathbf{x}) = \hat{d}_1 + \sum_{i=1}^{n} \hat{c}_i R(\mathbf{x}_i, \mathbf{x}), \tag{25}$$

with

$$(\mathbf{Q}^t\mathbf{Q} + \lambda\mathbf{S})^{-1}\mathbf{Q}^t\mathbf{y} = \begin{bmatrix} \hat{d}_1 \\ \hat{\mathbf{c}} \end{bmatrix}, \tag{26}$$

as given by (22). In this case, it is more familiar to write the solution in what Pearce and Wand (2006) referred to as the "primal" form,

$$\hat{f}_\lambda(\mathbf{x}) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p.$$

This can be done by recalling from Example 3.3 that

$$R(\mathbf{x}_i, \mathbf{x}) = x_{i1}x_1 + x_{i2}x_2 + \cdots + x_{ip}x_p.$$

Substituting into (25) gives

$$\hat{f}(\mathbf{x}) = \hat{d}_1 + \sum_{i=1}^{n} \hat{c}_i x_{i1} x_1 + \sum_{i=1}^{n} \hat{c}_i x_{i2} x_2 + \cdots + \sum_{i=1}^{n} \hat{c}_i x_{ip} x_p,$$

which implies that

$$\hat{\beta}_0 = \hat{d}_1 \quad \text{and} \quad \hat{\beta}_j = \sum_{i=1}^{n} \hat{c}_i x_{ij} \tag{27}$$

for $j = 1, 2, \ldots, p$. In the online supplementary material, we demonstrate this solution on Longley's (1967) employment data using the statistical software R. The RKHS solution applied to ridge regression is mostly for illustrative (not practical) purposes, as it requires an $(n+1) \times (n+1)$ matrix solve. This is, of course, less efficient than the standard approach unless $p > n$.

Both of the previous examples illustrate a typical case in which $\tilde{V}$ is constructed as the direct sum of a vector space V_0 and an RKHS H whose intersection is the zero vector,

$$\tilde{V} = V_0 \oplus H.$$

Because the intersection is zero, any $f \in \tilde{V}$ can be written uniquely as $f = f_0 + f_1$, with $f_0 \in V_0$ and $f_1 \in H$. The importance of this is that if the penalty functional J happens to satisfy

$$J(f) = \langle f_1, f_1 \rangle_H,$$

then all of our assumptions about J hold immediately with $N = V_0$ and the solution in (22) can be applied.

As a further aside, when V_0 is itself an RKHS, then $\tilde{V}$ is an RKHS under the inner product

$$\langle f, g \rangle_{\tilde{V}} \equiv \langle f_0, g_0 \rangle_{V_0} + \langle f_1, g_1 \rangle_H,$$

with the RK $R_{\tilde{V}} = R_{V_0} + R_H$. Note that V_0 and H are orthogonal under this inner product. This orthogonal decomposition of $\tilde{V}$ is also closely related to additive models (Wood 2006) and more generally to smoothing spline ANOVA (analysis of variance) models (Gu 2002), which also include tensor product splines as a special case. Thin-plate splines (Wahba 1990) also fall nicely into the general RKHS framework.

4.4 Choosing the Degree of Smoothness

With the penalized regression procedures described earlier, the choice of the smoothing parameter λ is an important issue. There are many methods available for this task, for example, visual inspection of the fit, m-fold cross-validation (Kohavi 1995), Akaike information criterion (AIC)/unbiased risk estimation, generalized maximum likelihood (Wahba 1990), and generalized cross-validation (GCV) (Craven and Wahba 1979). For those examples given in the online supplementary material, we use the GCV approach, which works as follows for any generalized ridge regression solution, such as in (26). Suppose that an estimate admits the following closed-form expression:

$$\hat{\boldsymbol{\beta}} = (\mathbf{Q}^t\mathbf{Q} + \lambda\mathbf{S})^{-1}\mathbf{Q}^t\mathbf{y}.$$

The GCV choice of λ for this generalized ridge estimate is the minimizer of

$$V(\lambda) = \frac{1}{n}||(\mathbf{I} - \mathbf{A}(\lambda)\mathbf{y}||^2 \Big/ \left[\frac{1}{n}\text{trace}\{\mathbf{I} - \mathbf{A}(\lambda)\}\right]^2,$$

where

$$\mathbf{A}(\lambda) = \mathbf{Q}(\mathbf{Q}^t\mathbf{Q} + \lambda\mathbf{S})^{-1}\mathbf{Q}^t.$$

The goal of GCV is to find the optimal λ so that the resulting $\hat{\boldsymbol{\beta}}$ has the smallest mean squared error. 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).

5. CONCLUDING REMARKS

We have given several examples illustrating the utility of the RKHS approach to solve penalized regression problems. We reviewed the building blocks (spaces) necessary to define an RKHS and presented several key results about these spaces. Finally, we used the results to illustrate application to cubic smoothing spline problems and ridge regression, providing transparent R code to enhance understanding.

The reader is now encouraged to explore some more advanced articles and texts that they now have the tools to access and unlock the full potential of RKHS methods. Wahba (1990) and Gu (2002) discussed smoothing spline ANOVA, which is a flexible modeling framework ranging from additive modeling on one extreme to full tensor product splines on the other. Gu (2002) also covered generalized models in the smoothing spline ANOVA framework. Pearce and Wand (2006) explored the intimate connection between RKHS in penalized regression and the much used SVM. Lin and Zhang (2006) and Storlie et al. (2010) used the RKHS framework to develop a smoothing spline version of the popular Lasso (Tibshirani 1996) and adaptive Lasso (Zou 2006), respectively. The RKHS framework can also be used to do spatially adaptive smoothing (Storlie, Bondell, and Reich 2010). The flexibility and elegance of RKHS methods are remarkable.

SUPPLEMENTAL MATERIALS

Examples using R code to illustrate application of RKHS to cubic smoothing spline and ridge regression.

[Received August 2011. Revised March 2012.]

REFERENCES

Allen, D. (1974), "The Relationship Between Variable Selection and Data Augmentation and a Method for Prediction," *Technometrics*, 16, 125–127. [59]

Berlinet, A., and Thomas-Agnan, C. (2004), *Reproducing Kernel Hilbert Spaces in Probability and Statistics*, Norwell, MA: Kluwer Academic Publishers. [55,56]

Berman, M. (1994), "Automated Smoothing of Image and Other Regularly Spaced Data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16, 460–468. [50]

Bivand, R., Pebesma, E., and Gómez-Rubio, V. (2008), *Applied Spatial Data Analysis With R: Use R*, New York: Springer. [50]

Craven, P., and Wahba, G. (1979), "Smoothing Noisy Data With Spline Functions: Estimating the Correct Degree of Smoothing by the Method of Generalized Cross-Validation," *Numerical Mathematics*, 31, 377–403. [59]

de Barra, G. (1981), *Measure Theory and Integration*, West Sussex: Horwood Publishing. [54]

Eubank, R. (1999), *Nonparametric Regression and Spline Smoothing*, New York: Marcel Dekker. [50]

Golub, G., Heath, M., and Wahba, G. (1979), "Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter," *Technometrics*, 21, 215–223. [59]

Gu, C. (2002), *Smoothing Spline ANOVA Models*, New York: Springer-Verlag. [50,53,55,57,58,59]

Gu, C., and Qiu, C. (1993), "Smoothing Spline Density Estimation: Theory," *The Annals of Statistics*, 21, 217–234. [56]

Hastie, T., Tibshirani, R., and Friedman, J. (2001), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, New York: Springer-Verlag. [50]

Hoerl, A., and Kennard, R. (1970), "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics*, 12, 55–67. [56]

Kohavi, R. (1995), "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection," *Proceedings of the 14th International Joint Conference on Artificial Intelligence* (Vol. 2), ed. Chris S. Mellish, Waltham, MA: Morgan Kaufmann, pp. 1137–1143. [59]

Lin, Y., and Zhang, H. (2006), "Component Selection and Smoothing in Smoothing Spline Analysis of Variance Models," *The Annals of Statistics*, 34, 2272–2297. [56,59]

Longley, J. (1967), "An Appraisal of Least-Squares Programs From the Point of View of the User," *Journal of the American Statistical Association*, 62, 819–841. [58]

Máté, L. (1990), *Hilbert Space Methods in Science and Engineering*, Oxford: Taylor & Francis. [51,53,54]

Naylor, A., and Sell, G. (1982), *Linear Operator Theory in Engineering and Science*, New York: Springer. [53,54]

O'Sullivan, F., Yandell, B., and Raynor, W. (1986), "Automatic Smoothing of Regression Functions in Generalized Linear Models," *Journal of the American Statistical Association*, 81, 96–103. [56]

Pearce, N., and Wand, M. (2006), "Penalised Splines and Reproducing Kernel Methods," *The American Statistician*, 60, 233–240. [50,57,58,59]

R Development Core Team. (2005), *R: A Language and Environment for Statistical Computing*, Vienna: R Foundation for Statistical Computing. Available at: *http://www.R-project.org* [50,58]

Ramsay, J., and Silverman, B. (2005), *Functional Data Analysis*, New York: Springer-Verlag. [50]

Rustagi, J. (1994), *Optimization Techniques in Statistics*, London: Academic Press. [53]

Storlie, C., Bondell, H., and Reich, B. (2010), "A Locally Adaptive Penalty for Estimation of Functions With Varying Roughness," *Journal of Computational and Graphical Statistics*, 19, 569–589. [56,59]

Storlie, C., Bondell, H., Reich, B., and Zhang, H. (2010), "Surface Estimation, Variable Selection, and the Nonparametric Oracle Property," *Statistica Sinica*, 21, 679–705. [56,59]

Storlie, C., Swiler, L., Helton, J., and Sallaberry, C. (2009), "Implementation and Evaluation of Nonparametric Regression Procedures for Sensitivity Analysis of Computationally Demanding Models," *Reliability Engineering and System Safety*, 94, 1735–1763. [50]

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society,* Series B, 58, 267–288. [59]

Wahba, G. (1990), *Spline Models for Observational Data* (Vol. 59: CBMS-NSF Regional Conference Series in Applied Mathematics), Philadelphia, PA: SIAM. [50,53,57,58,59]

Wecker, W., and Ansley, C. (1983), "The Signal Extraction Approach to Nonlinear Regression and Spline Smoothing," *Journal of the American Statistical Association*, 78, 81–89. [59]

Wood, S. (2006), *Generalized Additive Models: An Introduction With R*, Boca Raton, FL: CRC Press. [59]

Young, N. (1988), *An Introduction to Hilbert Space*, Cambridge: Cambridge University Press. [53]

Zou, H. (2006), "The Adaptive Lasso and Its Oracle Properties," *Journal of the American Statistical Association*, 101(476), 1418–1429. [59]