The University of British Columbia
Department of Statistics
Technical Report # 216

The Theory and Application of Penalized Least Squares Methods or Reproducing Kernel Hilbert Spaces Made Easy

by Nancy E. Heckman

The Theory and Application of Penalized Least Squares Methods or Reproducing Kernel Hilbert Spaces Made Easy

Nancy E. Heckman
Department of Statistics
University of British Columbia
Vancouver, British Columbia V6T 1Z2 Canada

August 22, 1997

The author wishes to acknowledge the support of the Natural Sciences and Engineering Research Council of Canada through grant A7969. The inspiration for this work came from J.O. Ramsay, Psychology Department, McGill University. Requests for this paper should be directed to the author, electronic mail address: nancy@stat.ubc.ca. A ps file of the manuscript can also be retrieved via anonymous ftp to newton@stat.ubc.ca, in directory pub/nancy, file name PLS.ps, or via the web: http://www.stat.ubc.ca/nhpubl.html.

1 Abstract

The popular cubic smoothing spline estimate of a regression function is the minimizer of

$$\sum_j d_j(Y_j - \mu(t_j))^2 + \lambda \int_a^b \left[\mu''(t)\right]^2 dt,$$

where (Y_j, t_j) are the data and the d_j's are positive weights. However, sometimes the data are related to the function of interest μ in another way, i.e., $\mathrm{E}(Y_i) = \mathrm{F}_i(\mu)$ for some known F_i's. And sometimes, one may wish to replace

$\int(\mu'')^2$ with another expression. This paper discusses the solution for these generalizations, that is, the minimization of

$$\sum_j d_j(Y_j - \mathrm{L}_j(\mu))^2 + \lambda \int_a^b \left[(\mathrm{L}\mu)(t)\right]^2 \, dt.$$

Here, L is a linear differential operator of order $m \geq 1$: $(\mathrm{L}\mu)(t) = \mu^{(m)}(t) + \sum_{j=0}^{m-1} w_j(t)\mu^{(j)}(t)$. This paper outlines basic theory for this general minimization problem, and provides explicit directions for calculating the minimizer. The minimizer depends on the easily calculated reproducing kernel associated with L.

2 Introduction

The cubic smoothing spline, a popular regression function estimate, is the minimizer of

$$\sum_j d_j(Y_j - \mu(t_j))^2 + \lambda \int_a^b (\mu''(t))^2 \, dt. \tag{1}$$

Here, the regression data are (t_j, Y_j), $j = 1, \cdots, n$, $t_j \in [a,b]$ a finite interval, and the d_j's are positive weights. The non-negative smoothing parameter λ balances μ's fit to the data (via minimizing $\sum d_j(Y_j - \mu(t_j))^2$) with μ's closeness to a straight line (via forcing $\mu''(t)$ to be zero). The minimization is performed over the function space

$$\mathcal{H}^2[a,b] = \{\mu : [a,b] \to \Re : \mu \text{ and } \mu' \text{ are absolutely continuous}$$
$$\text{and } \int_a^b (\mu''(t))^2 \, dt < \infty\}.$$

The purpose of this paper is to explain to the average statistician the theory and techniques for minimizing (1) and for minimizing penalized least squares expressions which are more general than (1). The material contained here is drawn from many sources: from statistical literature, from the theory of differential equations, from numerical analysis, and from functional analysis.

The general penalized least squares problem is to minimize

$$\sum_j d_j(Y_j - \mathrm{F}_j(\mu))^2 + \lambda \int_a^b \left[(\mathrm{L}\mu)(t)\right]^2 \, dt \tag{2}$$

where $\lambda \geq 0$, the d_j's are positive weights, the F_j's are continuous linear functionals and L is a linear differential operator of order $m \geq 1$:

$$(\mathrm{L}\mu)(t) \;=\; \mu^{(m)}(t) + \sum_{j=0}^{m-1} w_j(t)\mu^{(j)}(t) \qquad (3)$$
$$\text{with } w_j(\cdot) \text{ real} - \text{valued and continuous.}$$

The minimization is over all μ in the Sobolev space

$$\mathcal{H}^m[a,b] = \{\mu : [a,b] \to \Re : \mu^{(j)}, j = 0, \cdots, m-1 \text{ are absolutely continuous}$$
$$\text{and } \int_a^b (\mu^{(m)}(t))^2 \, dt < \infty\}.$$

To simplify notation, $\mathcal{H}^m$ will be used instead of $\mathcal{H}^m[a,b]$.

F_j's have been studied other than $\mathrm{F}_j(\mu) = \mu(t_j)$ of equation (1). For instance, to estimate $\mu(t)$, the HIV infection rate at time t, the data are Y_j, the number of new AIDS cases diagnosed in time period j. The expected value of Y_j depends on $\mu(t)$ for values of t up to and including period j, and $\mathrm{E}(Y_j)$ can be written as a continuous linear functional of μ. The functional depends on the distribution of the time of progress from HIV infection to AIDS diagnosis. Li, 1996, has estimated μ by minimizing (2) with $\mathrm{L}\mu = \mu''$. Bacchetti et al, 1993, have considered minimizing a discretized version of (2) with $\mathrm{L}\mu = \mu''$. This technique is known as backcalculation.

In a non-regression setting, Nychka et al (1984) minimize (2) to estimate μ, the distribution of the volumes of tumours found in livers of experimental animals. The data are the areas of cross-sections of tumours, gotten from cross-sectional slices of liver. The authors model tumours as spheres and, via a continuous linear functional, relate μ to the distribution of the area of a randomly chosen cross-section of a sphere. Thus the observed data are directly related to a linear functional of μ. These authors use $\mathrm{L}\mu = \mu''$.

Wahba (1990) considers F_j's based on Fredholm integral equations of the first kind, that is, $\mathrm{E}(Y_j) = g(t_j)$ where $g(t_j) = \int_a^b H(s,t_j)\mu(s)\, ds \equiv \mathrm{F}_j(\mu)$, with H known. Such data can arise in tomography. For other applications, see the references in Wahba. Wahba also takes $\mathrm{L}\mu = \mu''$.

Ansley, Kohn, and Wong (1993) and Heckman and Ramsay (1996) demonstrate the usefulness of using L's other than $\mathrm{L}\mu = \mu''$. Figure 1, taken from the Heckman and Ramsay paper, shows two estimates of a regression function for the incidence of melanoma in males. The data, described in Andrews

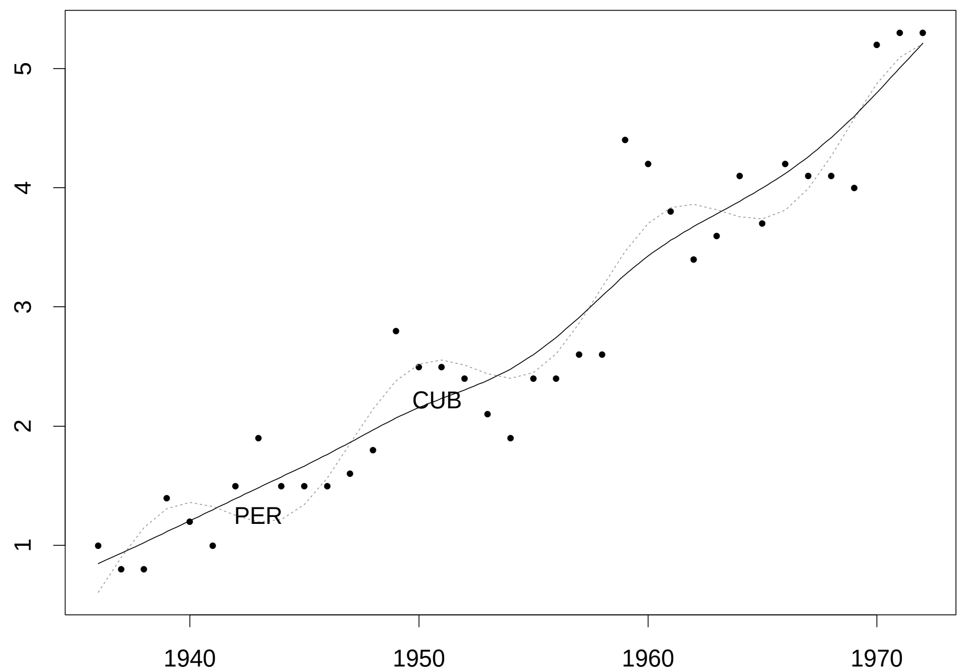

Figure 1: Male Melanoma Data. Estimates are the solid line, which is CUB and uses $\mathrm{L} = \mu''$, and the dashed line, which is PER, which uses $\mathrm{L} = \mu^{(4)} + (.58)^2\mu''$. Each estimate uses 5.6 parameters.

and Herzberg (1985), are from the Connecticut Tumour Registry and can be found in Statlib, whose WEB address is http://www.stat.cmu.edu/. The estimate labelled CUB is the minimizer of (1). The estimate labelled PER is the minimizer of (2) with $(\mathrm{L}\mu)(t) = \mu^{(4)}(t) + \omega^2\mu''(t)$ with $\omega = 0.58$. In both estimates, the smoothing parameter λ was chosen so that the "number of parameters used" was equal to 5.6. The differential operator L was chosen because we didn't want to penalize functions of the form $\mu(t) = \alpha_1 + \alpha_2 t + \alpha_3 \cos\omega t + \alpha_4 \sin\omega t$. Such functions are exactly the functions $\mathrm{L}\mu \equiv 0$ and form a popular parametric model for fitting melanoma data. The value of ω was chosen by a nonlinear least squares fit to this parametric model.

3 Results for the Cubic Smoothing Spline

Here, standard results for the minimization of (1) are stated without proof. For details, see Eubank (1988), Wahba (1990), or Green and Silverman (1994). Later sections contain the analogous results for the minimizer of (2). In those sections, some proofs will be given, along with references.

The minimization of (1) over $\mu \in \mathcal{H}^2[a,b]$ is easily done by considering the Hilbert space structure of $\mathcal{H}^2[a,b]$. The inner product is given by

$$\langle f, g\rangle = f(a)g(a) + f'(a)g'(a) + \int_a^b f''(t)\ g''(t)\ dt.$$

With this inner product, the linear functional $\mathrm{F}_t(f) = f(t)$ is continuous and so, by the Riesz representation theorem, there exists $R_t \in \mathcal{H}^2[a,b]$ such that $\langle R_t, f\rangle = f(t)$ for all $f \in \mathcal{H}$. One easily verifies that

$$R_t(s) \ = \ 1 \ + \ (s-a)(t-a) \ + \ R_{1t}(s)$$

where

$$R_{1t}(s) \ = \ st\left(\min\{s,t\} - a\right) + \frac{s+t}{2}\left((\min\{s,t\})^2 - a^2\right) + \frac{1}{3}\left((\min\{s,t\})^3 - a^3\right).$$

We call the bivariate function R with $R(s,t) = R_t(s)$ the reproducing kernel for $\mathcal{H}^2[a,b]$ and we say that $\mathcal{H}^2[a,b]$ is a reproducing kernel Hilbert space.

One can show that the minimizer of (1) is of the form

$$\mu(t) = \alpha_0 + \alpha_1 t + \sum_1^n \beta_j R_{1t_j}(t).$$

Direct calculation yields that, for μ of this form, (1) becomes

$$(\mathbf{Y} - \mathbf{T}\boldsymbol{\alpha} - \mathbf{K}\boldsymbol{\beta})'\mathbf{D}(\mathbf{Y} - \mathbf{T}\boldsymbol{\alpha} - \mathbf{K}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}'\mathbf{K}\boldsymbol{\beta} \tag{4}$$

where $\boldsymbol{\alpha} = (\alpha_0, \alpha_1)'$, $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_n)'$, $\mathbf{Y} = (Y_1, \cdots, Y_n)$, $\mathbf{T}_{i1} = 1$, $\mathbf{T}_{i2} = t_i$, $i = 1, \cdots, n$, $\mathbf{K}_{ij} = R_{1t_j}(t_i)$, $i, j = 1, \cdots, n$, and $\mathbf{D}$ is an n by n diagonal matrix with $\mathbf{D}_{ii} = d_i$. Thus one can minimize (4) directly, using matrix calculus.

Unfortunately, solving the matrix equations resulting from the differentiation of (4) involves inverting matrices which are very ill-conditioned. Thus, the calculations are subject to round-off errors that seriously effect the accuracy of the solution. In addition, the matrices to be inverted are not sparse, so that $O(n^3)$ operations are required. This can be a formidable task for, say, $n = 1000$. The problem is due to the fact that the bases functions 1, t, and $R_{1t_j}(t)$ are almost dependent with supports equal to the entire interval $[a, b]$. There are two ways around this problem. One way is to replace this inconvenient basis with a more stable one, one in which the elements have close to non-overlapping support. The most popular basis for this problem is that made up of B-splines (see, e.g., Eubank, 1988). The ith B-spline basis function has support $[t_i, t_{i+2}]$ and thus the matrices involved in the minimization of (1) are banded, well-conditioned, and fast to invert. Another approach is that of Reinsch (1967, 1970). The Reinsch algorithm yields a minimizer in $O(n)$ calculations. The approach for the Reinsch algorithm is based on a paper of Anselone and Laurent (1968). The application of this technique to the minimization of (2) with $\mathrm{F}_j(\mu) = \mu(t_j)$ is given in Section 5.

4 Hilbert Space Structure for the General Problem

We would like to set up a Hilbert space structure on $\mathcal{H}^m$, similar to the structure on $\mathcal{H}^2$ of Section 3, so that the minimization of (2) is easy. In particular, we would like to define a useful inner product on $\mathcal{H}^m$ so that it is a reproducing kernel Hilbert space.

Definition. $\mathcal{H}$ with inner product $\langle \cdot, \cdot \rangle$ is a *reproducing kernel Hilbert space* of functions on $[a, b]$ if and only if $\mathcal{H}$ is a Hilbert space (that is, a complete vector space with inner product $\langle \cdot, \cdot \rangle$) where, for all $t \in [a, b]$, the linear operator $\mathrm{F}_t(f) \equiv f(t)$ is continuous. By the Riesz representation theorem, the continuity of F_t is equivalent to the existence of a bivariate function R defined on $[a, b] \times [a, b]$ such that $R(\cdot, t) \in \mathcal{H}$ for all t and $\langle R(\cdot, t), f \rangle = f(t)$ for all $f \in \mathcal{H}$ and all $t \in [a, b]$. The function R is called the *reproducing kernel* of $\mathcal{H}$.

Results from the theory of differential equations are important in calculating the reproducing kernel of a Hilbert space. These results are given in Sections 6 and 7, and involve $G(\cdot, \cdot)$, the Green's function associated with the differential operator L.

Assume throughout that L is as in (3).

First note the following:

(4.1) For all $\mu \in \mathcal{H}^m$, $\mathrm{L}\mu(t)$ exists almost everywhere t and $\mathrm{L}\mu$ is square integrable, since the w_j's are continuous and $[a, b]$ is finite.

(4.2) By Theorem 4 of Section 6, there exist $u_1, \cdots, u_m \in \mathcal{H}^m$ with m derivatives that are linearly independent and form a basis for the set of all μ with $\mathrm{L}\mu(t) = 0$ almost everywhere t. Furthermore $\mathbf{W}(t)$, the Wronskian matrix associated with $u_1, \cdots, u_m$, is invertible for all $t \in [a, b]$. The Wronskian matrix is defined as $[\mathbf{W}(t)]_{ij} = u_i^{(j-1)}(t)$, $i, j = 1, \cdots, m$.

(4.3) By Theorem 7 of Section 6, if $f \in \mathcal{H}^m$ with $\mathrm{L}f(t) = 0$ almost everywhere t and with $f^{(j)}(a) = 0$, $j = 0, \cdots, m-1$, then $f \equiv 0$.

We can define an inner product on $\mathcal{H}^m$ under which $\mathcal{H}^m$ is a reproducing kernel Hilbert space. Let

$$\langle f, g \rangle = \sum_{j=0}^{m-1} f^{(j)}(a) g^{(j)}(a) + \int_a^b (\mathrm{L}f)(t)\ (\mathrm{L}g)(t)\ dt. \tag{5}$$

Theorem 1 *Let* $\{u_1, \cdots, u_m\}$ *be a basis for the set of* μ *with* $\mathrm{L}\mu \equiv 0$ *and let* $\mathbf{W}(t)$ *be the associated Wronskian matrix. Then, under the inner product (5),* $\mathcal{H}^m$ *is a reproducing kernel Hilbert space with reproducing kernel* $R(s,t) = R_0(s,t) + R_1(s,t)$ *where*

$$R_0(s,t) = \sum_{i,j=1}^{m} C_{ij} u_i(s) u_j(t) \quad \text{where } C_{ij} = \left[\left(\mathbf{W}(a) \mathbf{W}'(a) \right)^{-1} \right]_{ij},$$

$$R_1(s,t) = \int_{u=a}^{b} G(s,u)\ G(t,u)\ du$$

and $G(\cdot,\cdot)$ is the Green's function associated with L*, as defined in Section 7. Furthermore $\mathcal{H}^m$ can be partitioned into the direct sum of the two subspaces*

$$\mathcal{H}_0^m = \text{the set of all } f \in \mathcal{H}^m \text{ with } \mathrm{L}f(t) = 0 \text{ almost everywhere } t.$$

and

$$\mathcal{H}_1^m = \text{the set of all } f \in \mathcal{H}^m \text{ with } f^{(j)}(a) = 0, j = 0, \cdots, m-1.$$

$\mathcal{H}_0^m$ has reproducing kernel R_0 and $\mathcal{H}_1^m$ has reproducing kernel R_1.

Proof. To prove the theorem, it suffices to show the following.

(a) Any f in $\mathcal{H}^m$ can be written as $f = f_0 + f_1$, with $f_i \in \mathcal{H}_i^m$.

(b) $\mathcal{H}_0^m$ is orthogonal to $\mathcal{H}_1^m$ under the inner product (5).

(c) $\mathcal{H}_0^m$ is a reproducing kernel Hilbert space with reproducing kernel R_0 under the inner product $\langle f,g\rangle_0 = \sum_{j=0}^{m-1} f^{(j)}(a)g^{(j)}(a)$.

(d) $\mathcal{H}_1^m$ is a reproducing kernel Hilbert space with reproducing kernel R_1 under the inner product $\langle f,g\rangle_1 = \int_a^b (\mathrm{L}f)(t)\ (\mathrm{L}g)(t)\ dt$.

To prove (a), we find $c_1, \cdots, c_m$ such that, if $f_0 = \sum c_i u_i$, then $f_1 = f - f_0 \in \mathcal{H}_1^m$. That is, we require that, for $j = 0, \cdots, m-1$, $f_1^{(j)}(a) = 0$, that is $f^{(j)}(a) - \sum_i c_i u_i^{(j)}(a) = 0$. Writing this in matrix notation and using the Wronskian matrix yields

$$(f(a), f'(a), \cdots, f^{(m-1)}(a)) \;=\; (c_1, \cdots, c_m)\ \mathbf{W}(a)$$

and we can solve this for $(c_1, \cdots, c_m)$, since the Wronskian is invertible, by comment 4.2.

To prove statement (b), suppose that $f_i \in \mathcal{H}_i^m$, $i = 0, 1$. Then $\langle f_0, f_1\rangle$ is obviously equal to zero, by the definition (5).

To prove statements (c) and (d), we must show that, for $i = 0, 1$, $\mathcal{H}_i^m$ is a vector space (this is obvious), that $\langle\cdot,\cdot\rangle_i$ is an inner product, that $\mathcal{H}_i^m$ is complete, that $R_i(\cdot, t) \in \mathcal{H}_i^m$, and that $\langle R_i(\cdot,t), f\rangle_i = f(t)$ for all $f \in \mathcal{H}_i^m$.

The only difficulty in verifying that $\langle \cdot, \cdot \rangle_i$ is an inner product lies in showing that $\langle f, f \rangle_i = 0$ implies that $f \equiv 0$, $i = 0, 1$. But this follows immediately from comment 4.3.

To prove that $\mathcal{H}_1^m$ is complete, suppose that f_n is a Cauchy sequence in $\mathcal{H}_1^m$. Let $\mathcal{L}_2$ be the set of all functions on $[a, b]$ that are square integrable, with usual inner product $\langle f, g \rangle = \int_a^b f(t)\ g(t)\ dt$. Then, by the definition of the inner product on $\mathcal{H}_1^m$, $\mathrm{L}f_n$ is a Cauchy sequence in $\mathcal{L}_2$ and, by completeness of $\mathcal{L}_2$, there exists $h \in \mathcal{L}_2$ such that $\int_a^b ((\mathrm{L}f_n)(t) - h(t))^2\ dt$ converges to zero. Let G be the Green's function associated with L and let $f(t) = \int_a^b G(t, u)\ h(u)\ du$. Then, by Theorem 9 of Section 7, $f \in \mathcal{H}_1^m$ and $\mathrm{L}f(t) = h(t)$ almost everywhere t. Therefore, f_n converges to f in $\mathcal{H}_1^m$.

To prove that R_1 is the reproducing kernel for $\mathcal{H}_1^m$, first simplify notation, fixing $t \in [a, b]$ and letting $r(s) = R_1(s, t)$. We must show that $r \in \mathcal{H}_1^m$ and that $\langle r, f \rangle_1 = f(t)$ for all $f \in \mathcal{H}_1^m$. But $r(s) = \int_a^b G(s, u)\ h(u)\ du$ for $h(u) = G(t, u)$, which is in $\mathcal{L}^2$, so by Theorem 9 of Section 7, $r \in \mathcal{H}_1^m$ and $\mathrm{L}r(s) = h(s) = G(t, s)$ almost everywhere s. Therefore, for $f \in \mathcal{H}_1^m$,

$$\langle r, f \rangle_1 = \int_a^b (\mathrm{L}r)(s)\ (\mathrm{L}f)(s)\ ds = \int_a^b G(t, s)\ (\mathrm{L}f)(s)\ ds = f(t)$$

since G is the Green's function. (See the definition in Section 7.)

To finish the proof of (d), we first note that $\mathcal{H}_0^m$ is complete since it is finite dimensional, having as a basis $u_1, \cdots, u_m$. Obviously, $R_0(\cdot, t) \in \mathcal{H}_0^m$, since it is a linear combination of the u_i's. To show that $\langle R_0(\cdot, t), f \rangle_0 = f(t)$, it suffices to consider $f = u_l$, $l = 1, \cdots, m$. Then

$$\begin{aligned}
\langle R_0(\cdot, t), u_l \rangle_0 &= \sum_{i,j=1}^{m} C_{ij}\ u_j(t)\ \langle u_i, u_l \rangle_0 \\
&= \sum_{i,j=1}^{m} C_{ij}\ u_j(t) \sum_{k=0}^{m-1} u_i^{(k)}(a)\ u_l^{(k)}(a) \\
&= \sum_{i,j=1}^{m} C_{ij}\ u_j(t) \sum_{k=0}^{m-1} [\mathbf{W}(a)]_{i,k+1} [\mathbf{W}(a)]_{l,k+1} \\
&= \sum_{i,j=1}^{m} C_{ij}\ u_j(t)\ [\mathbf{W}(a)\mathbf{W}'(a)]_{li} \\
&= \sum_{j=1}^{m} u_j(t)\ [\mathbf{W}(a)\mathbf{W}'(a)\mathbf{C}]_{lj}
\end{aligned}$$

$$= \quad u_l(t).$$

Algorithm for calculating R_0, R_1 and R.

Suppose that we're given a linear differential operator L as in equation (3). The following steps describe how to calculate R_0, R_1, and R, the associated reproducing kernels.

1. Find $u_1, \cdots, u_m$, a basis for the set of functions μ with $\mathrm{L}\mu \equiv 0$. (If L is a linear differential operator with constant coefficients, this is easy to do. See Theorem 5 of Section 6.)
2. Calculate $\mathbf{W}(\cdot)$, the Wronskian of the u_i's: $\mathbf{W}_{ij}(t) = u_i^{(j-1)}(t)$.
3. $R_0(s,t) = \sum_{i,j} [\, [\mathbf{W}(a)\mathbf{W}'(a)]^{-1}]_{ij}\, u_i(s)u_j(t)$.
4. Calculate $(u_1^*(t), \cdots, u_m^*(t))$, the last row of the inverse of $\mathbf{W}$.
5. Find G, the associated Green's function: $G(t,u) = \sum u_i(t)u_i^*(u)$ for $u \le t$, 0 else.
6. $R_1(s,t) = \int_a^b G(s,u)G(t,u)du$.
7. $R = R_0 + R_1$.

Example. Suppose that $\mathrm{L}f = f'' + \gamma f'$, γ a real number.

For 1, we can find u_1 and u_2 via Theorem 5 of Section 6. We first solve $x^2 + \gamma x = 0$ for the two roots, $r_1 = 0$ and $r_2 = -\gamma$. Then

$$u_1(t) = 1 \quad \text{and} \quad u_2(t) = \exp(-\gamma t).$$

For 2, we compute the Wronskian

$$\mathbf{W}(t) = \begin{bmatrix} 1 & 0 \\ \exp(-\gamma t) & -\gamma \exp(-\gamma t) \end{bmatrix}.$$

For 3 we have

$$[\mathbf{W}(a)\mathbf{W}'(a)]^{-1} = \begin{bmatrix} 1 + \frac{1}{\gamma^2} & -\frac{1}{\gamma^2}\exp(\gamma a) \\ -\frac{1}{\gamma^2}\exp(\gamma a) & \frac{1}{\gamma^2}\exp(2\gamma a) \end{bmatrix}.$$

So

$$
\begin{aligned}
R_0(s,t) &= C_{11}u_1(s)u_1(t) + C_{12}u_1(s)u_2(t) + C_{21}u_2(s)u_1(t) + C_{22}u_2(s)u_2(t) \\
&= 1 + \frac{1}{\gamma^2} - \frac{1}{\gamma^2}\exp(-\gamma t^*) - \frac{1}{\gamma^2}\exp(-\gamma s^*) + \frac{1}{\gamma^2}\exp\big(-\gamma(s^*+t^*)\big)
\end{aligned}
$$

with $s^* = s - a$ and $t^* = t - a$.

For 4, inverting $\mathbf{W}(t)$ we find that

$$
u_1^*(t) = \frac{1}{\gamma} \quad \text{and} \quad u_2^*(t) = -\frac{1}{\gamma}\exp(\gamma t)
$$

and so, in 5, the Green's function is given by

$$
G(t,u) = \begin{cases} \frac{1}{\gamma}\Big(1 - \exp(-\gamma(t-u))\Big) & \text{for } u \le t \\ 0 & \text{else.} \end{cases}
$$

To find $R_1(s,t)$ in 6, first suppose that $s \le t$. Then

$$
\begin{aligned}
R_1(s,t) &= \int_a^b \gamma^{-2}\left(1 - e^{-\gamma(s-u)}\right)\left(1 - e^{-\gamma(t-u)}\right)\ I\{u \le s\}\ I\{u \le t\}\ du \\
&= \int_a^s \gamma^{-2}\left(1 - e^{-\gamma(s-u)}\right)\left(1 - e^{-\gamma(t-u)}\right)\ du \\
&= -\frac{1}{\gamma^3} + \frac{s^*}{\gamma^2} + \frac{1}{\gamma}\exp(-\gamma s^*) + \frac{1}{\gamma^3}\exp(-\gamma t^*) \\
&\quad -\frac{1}{2\gamma^3}\exp(\gamma(s^*-t^*)) - \frac{1}{2\gamma^3}\exp(-\gamma(s^*+t^*)). \qquad (6)
\end{aligned}
$$

Since $R_1(s,t) = R_1(t,s)$, if $t < s$, then $R_1(s,t)$ is gotten by interchanging s^* and t^* in the above.

5 Minimization of the Penalized Sum of Squares

We're now ready to minimize (2) over $\mu \in \mathcal{H}^m$, where $\mathcal{H}^m$ has inner product defined in (5) and L is as in (3). Most of the material here can be found in Wahba (1990). We assume that the F_j's are continuous linear functionals in the inner product (5) defined on $\mathcal{H}^m$.

Definition. F is a *linear* functional if $\mathrm{F} : \mathcal{H}^m \to \Re$ and $\mathrm{F}(\alpha f + \beta g) = \alpha \mathrm{F}(f) + \beta \mathrm{F}(g)$ for all $f, g \in \mathcal{H}^m$ and all reals α and β. A linear functional F is *continuous* if and only if there exists a constant C such that, for all $\mu \in \mathcal{H}^m$, $|\mathrm{F}(\mu)| \leq C\|\mu\|$, where $\|\mu\|^2 = \langle \mu, \mu \rangle$. By the Riesz representation theorem, this is equivalent to the existence of $\eta \in \mathcal{H}^m$ such that $\langle \eta, \mu \rangle = \mathrm{F}(\mu)$ for all $\mu \in \mathcal{H}^m$. The function η is called the representer of F, and it is unique, since if $\mathrm{F}(\mu) = \langle \eta^*, \mu \rangle = \langle \eta, \mu \rangle$ for all $\mu \in \mathcal{H}^m$ then $\langle \eta^* - \eta, \mu \rangle = 0$ for all $\mu \in \mathcal{H}^m$ and so $\eta^* - \eta = 0$.

Note that $\mathrm{F}(\mu) = \mu(t)$ is linear. It is also continuous, since it has representer $\eta(\cdot) = R(\cdot, t)$, as defined in Theorem 1 of Section 4. The following theorem is useful for calculating representers of continuous linear functionals.

Theorem 2 *Suppose that* F *is a continuous linear functional on* $\mathcal{H}^m$ *with inner product as in (5). Let* η *be* F*'s representer. Using the notation and results of Theorem 1,*

$$\eta(t) = \mathrm{F}(R(\cdot, t)),$$

that is, we apply F *to* $R(s,t)$ *as a function of* s*, keeping* t *fixed. Furthermore, the representer of* F *in* $\mathcal{H}_i^m$*,* $i = 0, 1$*, is given by*

$$\eta_i(t) = \mathrm{F}(R_i(\cdot, t)), \quad i = 0, 1$$

and $\eta = \eta_0 + \eta_1$.

Proof. First, since η is the representer of F, η must satisfy $\mathrm{F}(R(\cdot, t)) = \langle \eta, R(\cdot, t) \rangle$. But, by the reproducing quality of R, this is equal to $\eta(t)$.

By the same argument $\eta_i(t) = \mathrm{F}(R_i(\cdot, t))$, $i = 0, 1$, since, by Theorem 1 of Section 4, R_i is the reproducing kernel of $\mathcal{H}_i^m$. To see that $\eta = \eta_0 + \eta_1$, write $\eta = \eta_0^* + \eta_1^*$ with $\eta_i^* \in \mathcal{H}_i^m$, $i = 0, 1$. (This is possible since $\mathcal{H}^m$ is the direct sum of $\mathcal{H}_0^m$ and $\mathcal{H}_1^m$.) Then, using the facts that R and R_i are reproducing kernels, that $\mathcal{H}_0^m$ and $\mathcal{H}_1^m$ are orthogonal, and that η is the representer of F,

$$\eta_i^*(t) = \langle \eta_i^*, R(\cdot, t) \rangle = \langle \eta, R_i(\cdot, t) \rangle = \mathrm{F}(R_i(\cdot, t)) = \eta_i(t).$$

So $\eta_i^* = \eta_i$.

Theorem 3 *Suppose that* L *is as in (3). Let* $u_1, \cdots, u_m$ *be a basis for the kernel of* L *and* G *the corresponding Green's function. Let* η_j *be the representer of* F_j *and write* $\eta_j = \eta_{j0} + \eta_{j1}$ *with* $\eta_{ji} \in \mathcal{H}_i^m$, $i = 0, 1$. *Then the minimizer of (2) exists and is of the form*

$$\mu(t) = \sum_{j=1}^{m} \alpha_j u_j(t) \ + \ \sum_{j=1}^{n} \beta_j \eta_{j1}(t). \tag{7}$$

Furthermore, η_{j1} *can be calculated via*

$$\eta_{j1}(t) = \mathrm{F}_j(R_1(\cdot, t)),$$

that is, we apply F_j *to* $R_1(s,t)$ *as a function of* s *with* t *held fixed, where*

$$R_1(s,t) = \int_a^b G(s,u)\ G(t,u)\ dt.$$

For μ *of the form (7), (2) becomes*

$$(\mathbf{Y} - \mathbf{T}\boldsymbol{\alpha} - \mathbf{K}\boldsymbol{\beta})'\mathbf{D}(\mathbf{Y} - \mathbf{T}\boldsymbol{\alpha} - \mathbf{K}\boldsymbol{\beta}) \ + \ \lambda\boldsymbol{\beta}'\mathbf{K}\boldsymbol{\beta}$$

where $\mathbf{Y} = (Y_1, \cdots, Y_n)'$, $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_m)'$, $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_n)'$, $\mathbf{T}_{ij} = \mathrm{F}_i(u_j)$, $i = 1, \cdots, n$, $j = 1, \cdots, m$,

$$\mathbf{K}_{ij} = \mathrm{F}_j(\eta_{i1}) = \mathrm{F}_i(\eta_{j1}), \quad i, j = 1, \cdots, n,$$

and $\mathbf{D}$ *is an* n *by* n *diagonal matrix with* $\mathbf{D}_{ii} = d_i$. *If* $\mathbf{T}$ *is of full column rank, the minimizer is unique.*

Proof. Using the notation and results of Theorem 1 of Section 4, we know that we can write $\mathcal{H}^m$ as the sum of two orthogonal subspaces $\mathcal{H}^m = \mathcal{H}_0^m \oplus \mathcal{H}_1^m$. We further partition $\mathcal{H}^m$ as follows. Let η_j be the representer of F_j and write

$$\eta_j = \eta_{j0} + \eta_{j1}$$

with $\eta_{ji} \in \mathcal{H}_i^m, i = 0, 1$. Then

$$\mathcal{H}^m = \mathcal{H}_0^m \oplus \mathcal{H}_{11}^m \oplus \mathcal{H}_{12}^m,$$

where $\mathcal{H}_{11}^m$ is the finite dimensional space spanned by η_{j1}, $j = 1, \cdots, n$, and $\mathcal{H}_{12}^m$ is the orthogonal complement of $\mathcal{H}_{11}^m$ in $\mathcal{H}_1^m$. Therefore, any $\mu \in \mathcal{H}^m$ can be written as

$$\mu = \mu_0 + \mu_{11} + \mu_{12} \quad \mu_{1j} \in \mathcal{H}_{1j}^m, j = 1, 2, \quad \text{and } \mu_0 \in \mathcal{H}_0^m.$$

Statement (7) will follow if we show that any minimizer of (2) must have $\mu_{12} \equiv 0$. Let $\mu \in \mathcal{H}^m$. First it's shown that $\mathrm{F}_j(\mu) = \mathrm{F}_j(\mu_0 + \mu_{11})$. Since η_j is the representer of F_j and μ_{12} is perpendicular to η_j,

$$\mathrm{F}_j(\mu) = \langle \eta_j, \mu \rangle = \langle \eta_j, \mu_0 + \mu_{11} + \mu_{12} \rangle = \langle \eta_j, \mu_0 + \mu_{11} \rangle = \mathrm{F}_j(\mu_0 + \mu_{11}).$$

To study the second term in (2), use the fact that $\mathrm{L}\mu_0 \equiv 0$ and write

$$\begin{aligned} \int_a^b (\mathrm{L}\mu(t))^2 \, dt &= \int_a^b (\mathrm{L}\mu_1(t))^2 \, dt \\ = \langle \mu_1, \mu_1 \rangle &= \langle \mu_{11}, \mu_{11} \rangle + \langle \mu_{12}, \mu_{12} \rangle. \end{aligned}$$

Therefore, to minimize

$$\begin{aligned} &\sum_{j=1}^n d_j \Big(Y_j - \mathrm{F}_j(\mu) \Big)^2 + \lambda \int_a^b ((\mathrm{L}\mu)(t))^2 \, dt \\ = &\sum_{j=1}^n d_j \Big(Y_j - \mathrm{F}_j(\mu_0 + \mu_{11}) \Big)^2 + \lambda \Big[\langle \mu_{11}, \mu_{11} \rangle + \langle \mu_{12}, \mu_{12} \rangle \Big] \end{aligned}$$

we should take μ_{12} to be the zero function.

Therefore, the minimizing μ must be of the form (7). For a μ of this form, we see that

$$\mathrm{F}_i \mu = \sum_{j=1}^m \alpha_j T_{ij} + \sum_{j=1}^n \beta_j K_{ij}$$

and

$$\int_a^b [\mathrm{L}\mu(t)]^2 = \sum_{i,j=1}^n \beta_i \; \beta_j \int_a^b (\mathrm{L}\eta_{i1})(t) \; (\mathrm{L}\eta_{j1})(t) \; dt = \sum_{ij} \beta_i \beta_j \langle \eta_{i1}, \eta_{j1} \rangle.$$

By Theorem 2 of Section 5, η_{i1} is the representer of F_i in $\mathcal{H}_1^m$ and so $\langle \eta_{i1}, \eta_{j1} \rangle = \mathrm{F}_i(\eta_{j1})$. Also by Theorem 2, η_{j1} is the representer of F_j in $\mathcal{H}_1^m$ and so $\langle \eta_{i1}, \eta_{j1} \rangle$ is also equal to $\mathrm{F}_j(\eta_{i1})$. Therefore $\int_a^b [\mathrm{L}\mu(t)]^2 \; dt = \boldsymbol{\beta}' \mathbf{K} \boldsymbol{\beta}$.

Computing $\hat{\mu}$

From Theorem 3 we see that we must minimize

$$(\mathbf{Y} - \mathbf{T}\boldsymbol{\alpha} - \mathbf{K}\boldsymbol{\beta})'\mathbf{D}(\mathbf{Y} - \mathbf{T}\boldsymbol{\alpha} - \mathbf{K}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}'\mathbf{K}\boldsymbol{\beta}. \tag{8}$$

Taking the derivative with respect to $\boldsymbol{\alpha}$ yields

$$-2\mathbf{T}'\mathbf{D}(\mathbf{Y} - \mathbf{T}\hat{\boldsymbol{\alpha}} - \mathbf{K}\hat{\boldsymbol{\beta}}) = 0,$$

that is

$$\mathbf{T}'\mathbf{D}(\mathbf{Y} - \mathbf{K}\hat{\boldsymbol{\beta}}) = \mathbf{T}'\mathbf{D}\mathbf{T}\hat{\boldsymbol{\alpha}}. \tag{9}$$

Taking the derivative of (8) with respect to $\boldsymbol{\beta}$ yields

$$-2\mathbf{K}'\mathbf{D}(\mathbf{Y} - \mathbf{T}\hat{\boldsymbol{\alpha}} - \mathbf{K}\hat{\boldsymbol{\beta}}) + 2\lambda\mathbf{K}\hat{\boldsymbol{\beta}} = 0.$$

Since $\mathbf{K}$ and $\mathbf{D}$ are invertible and symmetric, this last equation is equivalent to

$$\mathbf{Y} - \mathbf{T}\hat{\boldsymbol{\alpha}} - (\mathbf{K} + \lambda\mathbf{D}^{-1})\hat{\boldsymbol{\beta}} = 0$$

Let

$$\mathbf{M} = \mathbf{K} + \lambda\mathbf{D}^{-1}.$$

Then

$$\hat{\boldsymbol{\beta}} = \mathbf{M}^{-1}(\mathbf{Y} - \mathbf{T}\hat{\boldsymbol{\alpha}}). \tag{10}$$

Substituting this into (9) yields

$$\mathbf{T}'\mathbf{D}[I - \mathbf{K}\mathbf{M}^{-1}]\mathbf{Y} = \mathbf{T}'\mathbf{D}[I - \mathbf{K}\mathbf{M}^{-1}]\mathbf{T}\hat{\boldsymbol{\alpha}},$$

that is

$$\mathbf{T}'\mathbf{D}[\mathbf{M} - \mathbf{K}]\mathbf{M}^{-1}\mathbf{Y} = \mathbf{T}'\mathbf{D}[\mathbf{M} - \mathbf{K}]\mathbf{M}^{-1}\mathbf{T}\hat{\boldsymbol{\alpha}},$$

or

$$\lambda\mathbf{T}'\mathbf{D}\mathbf{D}^{-1}\mathbf{M}^{-1}\mathbf{Y} = \lambda\mathbf{T}'\mathbf{D}\mathbf{D}^{-1}\mathbf{M}^{-1}\mathbf{T}\hat{\boldsymbol{\alpha}}.$$

Therefore

$$\hat{\boldsymbol{\alpha}} = (\mathbf{T}'\mathbf{M}^{-1}\mathbf{T})^{-1}\mathbf{T}'\mathbf{M}^{-1}\mathbf{Y} \tag{11}$$

and

$$\hat{\boldsymbol{\beta}} = \mathbf{M}^{-1}[I - \mathbf{T}(\mathbf{T}'\mathbf{M}^{-1}\mathbf{T})^{-1}\mathbf{T}'\mathbf{M}^{-1}]\mathbf{Y}. \tag{12}$$

Algorithm for Minimizing (2) for General F_j's

We now have an algorithm for finding the minimizer of (2).

A. Follow steps 1-6 of Section 4 to find $u_1, \cdots, u_m$, a basis for $\mathrm{L}\mu = 0$, and the reproducing kernel R_1.

B. Find $\eta_{j1}(t) = \mathrm{F}_j(R_1(\cdot, t))$, $j = 1, \cdots, n$.

C. Let $T_{ij} = \mathrm{F}_i(u_j)$, $K_{ij} = \mathrm{F}_i(\eta_{j1})$.

D. Find $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$ using (11) and (12).

E. $\hat{\mu}(t) = \sum_{j=1}^{m} \hat{\alpha}_j u_j(t) + \sum_{j=1}^{n} \hat{\beta}_j \eta_{j1}(t)$.

Example

Suppose we want to find $\mu \in \mathcal{H}^1[0,1]$ to minimize

$$\sum_{j=1}^{n} [Y_j - \int_0^1 f_j(t)\mu(t)\ dt]^2 + \lambda \int_0^1 [\mu'(t)]^2\ dt$$

where the f_j's are known. Thus $\mathrm{F}_j(\mu) = \int_0^1 f_j(t)\mu(t)\ dt$ and $\mathrm{L}\mu = \mu'$.

For A, our basis for $\mathrm{L}\mu \equiv 0$ is $u_1(t) = 1$. The Wronskian is a one by one matrix [1]. So $u_1^*(s) = 1$ and $G(t, u) = 1$ if $u \leq t$, 0 else. Therefore

$$R_1(s,t) = \int_0^{\min\{s,t\}} 1\ du = \min\{s,t\}.$$

For B

$$\eta_{j1}(t) = \int_0^1 f_j(s) R_1(s,t)\ ds = \int_0^t s f_j(s)\ ds\ + t \int_t^1 f_j(s)\ ds.$$

For C, $\mathbf{T}$ is n by 1 with

$$T_{i1} = \mathrm{F}_i(u_1) = \int_0^1 f_i(t)\ dt$$

and

$$\begin{aligned} K_{ij} = \mathrm{F}_i(\eta_{j1}) &= \int_{t=1}^{1} f_i(t)\eta_{j1}(t)\ dt \\ &= \int_{t=0}^{1} f_i(t) \left[\int_{s=0}^{t} s f_j(s)\ ds\ + t \int_{s=t}^{1} f_j(s)\ ds \right]\ dt \\ &= \int_{t=0}^{1} \int_{s=0}^{t} s f_i(s) f_j(t)\ ds\ dt + \int_{t=0}^{1} \int_{s=0}^{t} s f_j(s) f_i(t)\ ds\ dt \end{aligned}$$

Continue with D and E.

Minimizing (2) when $\mathrm{F}_j(\mu) = \mu(t_j)$

Unfortunately, equations (11) and (12) result in computational problems since $\mathbf{M}$ is an ill-conditioned matrix and thus difficult to invert. Fortunately, when $\mathrm{F}_j(\mu) = \mu(t_j)$ we can transform the problem to alleviate the difficulties. Assume that $a \le t_1 \le \cdots \le t_n$.

Let $\mathbf{Q}$ be an n by $n-m$ matrix of full column rank such that $\mathbf{Q}'\mathbf{T}$ is an $n-m$ by m matrix of zeroes. ($\mathbf{Q}$ isn't unique. Later, a "good" $\mathbf{Q}$ is described.) The goal here is to show that

$$\hat{\boldsymbol{\beta}} = \mathbf{Q}(\mathbf{Q}'\mathbf{M}\mathbf{Q})^{-1}\mathbf{Q}'\mathbf{Y} \tag{13}$$

and

$$\hat{\mathbf{Y}} = \mathbf{Y} - \lambda\mathbf{D}^{-1}\hat{\boldsymbol{\beta}}.$$

We then seek $\mathbf{Q}$ so that $\mathbf{Q}'\mathbf{M}\mathbf{Q}$ is easy to invert.

We first show that $\mathbf{T}'\hat{\boldsymbol{\beta}} = 0$. (This will imply that there exists an $n-m$ vector $\boldsymbol{\gamma}$ such that $\hat{\boldsymbol{\beta}} = \mathbf{Q}\boldsymbol{\gamma}$.) Multiplying both sides of (10) by $\mathbf{M}$ yields

$$\mathbf{Y} = \mathbf{M}\hat{\boldsymbol{\beta}} + \mathbf{T}\hat{\boldsymbol{\alpha}}. \tag{14}$$

Substituting this into (11) yields

$$\hat{\boldsymbol{\alpha}} = (\mathbf{T}'\mathbf{M}^{-1}\mathbf{T})^{-1}\mathbf{T}'\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\alpha}}.$$

Therefore

$$(\mathbf{T}'\mathbf{M}^{-1}\mathbf{T})^{-1}\mathbf{T}'\hat{\boldsymbol{\beta}} = 0$$

and so $\mathbf{T}'\hat{\boldsymbol{\beta}} = 0$ and $\hat{\boldsymbol{\beta}} = \mathbf{Q}\boldsymbol{\gamma}$ for some $\boldsymbol{\gamma}$. To find $\boldsymbol{\gamma}$, use (10):

$$\mathbf{Q}'\mathbf{M}\hat{\boldsymbol{\beta}} = \mathbf{Q}'(\mathbf{Y} - \mathbf{T}\hat{\boldsymbol{\alpha}}) = \mathbf{Q}'\mathbf{Y}$$

since $\mathbf{Q}'\mathbf{T} = 0$, and so $\mathbf{Q}'\mathbf{M}\mathbf{Q}\boldsymbol{\gamma} = \mathbf{Q}'\mathbf{Y}$, yielding

$$\boldsymbol{\gamma} = (\mathbf{Q}'\mathbf{M}\mathbf{Q})^{-1}\mathbf{Q}'\mathbf{Y}.$$

Therefore (13) holds.

We easily solve for $\hat{\mathbf{Y}} = \mathbf{T}\hat{\boldsymbol{\alpha}} + \mathbf{K}\hat{\boldsymbol{\beta}}$ using (14):

$$\mathbf{Y} = (\mathbf{K} + \lambda\mathbf{D}^{-1})\hat{\boldsymbol{\beta}} + \mathbf{T}\hat{\boldsymbol{\alpha}} = \hat{\mathbf{Y}} + \lambda\mathbf{D}^{-1}\hat{\boldsymbol{\beta}}$$

and so

$$\hat{\mathbf{Y}} = \mathbf{Y} - \lambda \mathbf{D}^{-1}\hat{\boldsymbol{\beta}}.$$

Note that we have not yet used the fact that $\mathrm{F}_j(\mu) = \mu(t_j)$.

If $\mathrm{F}_j(\mu) = \mu(t_j)$, we can choose $\mathbf{Q}$ so that $\mathbf{Q}'\mathbf{M}\mathbf{Q}$ is banded and thus very easy to invert. In addition to requiring that $\mathbf{Q}'\mathbf{T} = 0$ we also seek $\mathbf{Q}$ with

$$Q_{ij} = 0 \quad \text{unless} \quad i = j, j+1, \cdots, j+m.$$

So we want $\mathbf{Q}$ with $[\mathbf{Q}'\mathbf{T}]_{ij} = \sum_{\ell=0}^{m} Q_{i+\ell,i} u_j(t_{i+\ell}) = 0$ for all $j = 1, \cdots, m$, $i = 1, \cdots, n-m$. That is, for each i, we seek an $m+1$ vector $\mathbf{q}_i = (Q_{ii}, \cdots, Q_{i+m,i})'$ with $\mathbf{q}_i'\mathbf{T}_i = 0$ where $\mathbf{T}_i$ is the $m+1$ by m matrix with ℓjth entry equal to $u_j(t_{i+\ell})$. This is easily done by a QR-decomposition of $\mathbf{T}_i$. Write the decomposition as $\mathbf{T}_i = \mathbf{Q}_i \mathbf{R}_i$. Then the required vector $\mathbf{q}_i$ is the $(m+1)$st column of $\mathbf{Q}_i$.

We now show that $\mathbf{Q}'\mathbf{M}\mathbf{Q}$ is banded, specifically, that $[\mathbf{Q}'\mathbf{M}\mathbf{Q}]_{k\ell} = 0$ whenever $|k-\ell| > m$. Write $\mathbf{Q}'\mathbf{M}\mathbf{Q} = \mathbf{Q}'\mathbf{K}\mathbf{Q} + \lambda\mathbf{Q}'\mathbf{D}^{-1}\mathbf{Q}$. Obviously, since $\mathbf{D}$ is diagonal, $[\mathbf{Q}\mathbf{D}^{-1}\mathbf{Q}]_{k\ell} = 0$ for $|k-\ell| > m$. We'll show that the same is true for $\mathbf{Q}'\mathbf{K}\mathbf{Q}$. Write

$$\begin{aligned} K_{ij} &= R_1(t_i, t_j) \\ &= \int G(t_i, w) G(t_j, w)\, dw \\ &= \sum_{r,s} u_r(t_i) u_s(t_j) \int_a^{\min\{t_i,t_j\}} u_r^*(w) u_s^*(w)\, dw \\ &\equiv \sum_{r,s} u_r(t_i) u_s(t_j) F_{r,s}(\min\{t_i, t_j\}) \end{aligned} \tag{15}$$

Since $\mathbf{Q}'\mathbf{K}\mathbf{Q}$ is symmetric, it suffices to show that $[\mathbf{Q}'\mathbf{K}\mathbf{Q}]_{k\ell} = 0$ for $k-\ell > m$.

$$\begin{aligned} [\mathbf{Q}'\mathbf{K}\mathbf{Q}]_{k\ell} &= \sum_{i,j=1}^{n} Q_{ik} K_{ij} Q_{j\ell} \\ &= \sum_{i,j=0}^{m} Q_{k+i,k} K_{k+i,\ell+j} Q_{\ell+j,\ell}. \end{aligned}$$

Since $k - \ell > m \geq j - i$ whenever $0 \leq i, j \leq m$, in the above summation we have $k + i > \ell + j$. So, using (15),

$$[\mathbf{Q}'\mathbf{K}\mathbf{Q}]_{k\ell} = \sum_{i,j=0}^{m} Q_{k+i,k} \sum_{r,s=1}^{m} F_{r,s}(t_{\ell+j}) T_{k+i,r} T_{\ell+j,s} Q_{\ell+j,\ell}$$

$$= \sum_{j=0}^{m} \sum_{r,s=1}^{m} F_{r,s}(t_{\ell+j}) T_{\ell+j,s} Q_{\ell+j,j} \sum_{i=0}^{m} Q_{k+i,k} T_{k+i,r}.$$

But $\sum_{i=0}^{m} Q_{k+i,k} T_{k+i,r} = [\mathbf{Q}'\mathbf{T}]_{kr} = 0$.

Algorithm for Minimizing (2) when $\mathrm{F}_j(\mu) = \mu_j(t)$

A. Follow steps 1-6 of Section 4 to find $u_1, \cdots, u_m$, a basis for $\mathrm{L}\mu = 0$, and the reproducing kernel R_1.

B. Let $T_{ij} = u_j(t_i)$, $K_{ij} = R_1(t_i, t_j)$.

C. Find $\mathbf{Q}$ n by $n - m$ of full column rank with $\mathbf{Q}'\mathbf{T} = 0$ and $Q_{ij} = 0$ unless $i = j, j+1, \cdots, j+m$.

D. Let
$$\hat{\boldsymbol{\beta}} = \mathbf{Q}[\mathbf{Q}'(\mathbf{K} + \lambda \mathbf{D}^{-1})\mathbf{Q}]^{-1}\mathbf{Q}'\mathbf{Y},$$
speeding up the matrix inversion by using the fact that $\mathbf{Q}'(\mathbf{K}+\lambda\mathbf{D}^{-1})\mathbf{Q}$ is banded. Then
$$\hat{\mathbf{Y}} = \mathbf{Y} - \lambda \mathbf{D}^{-1}\hat{\boldsymbol{\beta}}.$$

Example

Suppose that we want to minimize

$$\sum_{j=1}^{n} d_j (Y_j - \mu(t_j))^2 + \lambda \int_0^1 (\mu''(t) + \gamma\mu'(t))^2 \; dt$$

over $\mu \in \mathcal{H}^2[0,1]$. So $\mathrm{F}_j(\mu) = \mu(t_j)$ and $\mathrm{L}\mu = \mu'' + \gamma\mu'$. For simplicity, assume that $t_i = i/(n+1)$.

For A, by the example in Section 4,

$$u_1(t) = 1 \quad \text{and} \quad u_2(t) = e^{-\gamma t}$$

and $R_1(s,t) = R_1(t,s)$ with, for $s \leq t$ and $s^* = s - a$ and $t^* = t - a$,

$$\begin{aligned} R_1(s,t) &= -\frac{1}{\gamma^3} + \frac{s^*}{\gamma^2} + \frac{1}{\gamma}\exp(-\gamma s^*) + \frac{1}{\gamma^3}\exp(-\gamma t^*) \\ &\quad -\frac{1}{2\gamma^3}\exp(\gamma(s^* - t^*)) - \frac{1}{2\gamma^3}\exp(-\gamma(s^* + t^*)). \end{aligned}$$

For B, $T_{i1} = 1$, $T_{i2} = \exp(-\gamma t_i)$, and $K_{ij} = R_1(t_i, t_j)$.

For C, we seek $\mathbf{Q}$ n by $n-2$ with $Q_{ij} = 0$ unless $i = j, j+1, j+2$ and

$$0 = [\mathbf{Q}'\mathbf{T}]_{ij} = \sum_{k=1}^{n} Q_{ki}T_{kj} = Q_{ii}T_{ij} + Q_{i,i+1}T_{i+1,j} + Q_{i,i+2}T_{i+2,j}.$$

Thus

$$0 = Q_{ii} + Q_{i,i+1} + Q_{i,i+2}$$

and

$$0 = Q_{ii}\exp(-\gamma t_i) + Q_{i,i+1}\exp(-\gamma t_{i+1}) + Q_{i,i+2}\exp(-\gamma t_{i+2}).$$

There are many solutions. For instance, we can take

$$Q_{ii} = 1 - \exp\left(-\frac{\gamma}{n+1}\right) \quad Q_{i,i+1} = \exp\left(\frac{\gamma}{n+1}\right) - \exp\left(-\frac{\gamma}{n+1}\right)$$

and

$$Q_{i,i+2} = \exp\left(\frac{\gamma}{n+1}\right) - 1.$$

Continuing with D is straightforward.

6 Differential Equations

This section contains the results from differential equations that were used in the definition of our reproducing kernel Hilbert space. Details can be found in Coddington(1961). The main theorem, stated without proof, is

Theorem 4 *Let* L *be as in (3). Then there exists $u_1, \cdots, u_m$ a basis for the kernel of* L*, with each u_i real-valued and having m derivatives. Furthermore, any basis for the kernel of* L *will have an invertible Wronskian matrix $\mathbf{W}(t)$. The Wronskian matrix is defined as*

$$[\mathbf{W}(t)]_{ij} = u_i^{(j-1)} \quad i, j = 1, \cdots, m.$$

The following theorem, stated without proof, is useful for calculating the basis functions in the case that the w_j's are constants.

Theorem 5 *Suppose that* L *is as in (3), with the* w_j*'s real numbers. Denote the* s *distinct roots of the polynomial* $x^m + \sum_{j=0}^{m-1} w_j x^j$ *as* $r_1, \cdots, r_s$*. Let* m_i *denote the multiplicity of root* r_i *(so* $m = \sum_1^s r_i$*). Then the following* m *functions of* t *form a basis for the kernel of* L*:*

$$\exp(r_i t), \quad t\exp(r_i t), \quad \cdots, \quad t^{m_i-1}\exp(r_i t) \quad i = 1, \cdots, s.$$

The following result, stated without proof, is useful for checking that a set of functions does form a basis for the kernel of L.

Theorem 6 *Suppose that* $u_1, \cdots, u_m$ *have* m *derivatives on* $[a, b]$ *and that* $\mathrm{L}u_i \equiv 0$*. If* $\mathbf{W}(t_0)$ *is invertible at some* $t_0 \in [a, b]$*, then the* u_i*'s are linearly independent, and thus a basis for the kernel of* L*.*

The following result was useful in defining the inner product in Section 4, where t_0 was taken to be a.

Theorem 7 *Suppose that* L *is as in (3) and let* $t_0 \in [a, b]$*. Then the only function in* $\mathcal{H}^m$ *that satisfies* $\mathrm{L}f =$ *the zero function and* $f^{(j)}(t_0) = 0$*,* $j = 0, \cdots, m-1$ *is the zero function.*

Proof. By Theorem 4, there exists $u_1, \cdots, u_m$ a basis for the kernel of L with $\mathbf{W}(t)$ invertible for all $t \in [a, b]$. Suppose $\mathrm{L}f \equiv 0$. Then $f = \sum_i c_i u_i$ for some c_i's. We see that the conditions $f^{(j)}(t_0) = 0$, $j = 0, \cdots, m-1$ can be written in matrix/vector form as $(c_1, \cdots, c_m)\mathbf{W}(t_0) = (0, \cdots, 0)$. Since $\mathbf{W}(t_0)$ is invertible, $c_i = 0$, $i = 1, \cdots, m$.

Before using the smoothing technique in (2), one must first, of course, decide on the operator L. Often, the easiest way to do this is by specifying basis functions $u_1, \cdots, u_m$ for a preferred parametric model. One must then calculate the w_j's in (3) so that $\mathrm{L}u_i \equiv 0$, $i = 1, \cdots, m$ for these w_j's. This is simple to do, assuming that each u_i has m continuous derivatives and that the associated Wronskian matrix $\mathbf{W}(t)$ is invertible for all $t \in [a, b]$. Suppose this is true for a specified set of u_i with m continous derivatives. To solve for the w_j's, write

$$0 = (\mathrm{L}u_i)(t) = u_i^{(m)}(t) + \sum_{j=0}^{m-1} w_j(t) u_i^{(j)}(t),$$

that is

$$u_i^{(m)}(t) = -\sum_{j=0}^{m-1} w_j(t) u_i^{(j)}(t).$$

This can be written in matrix/vector form as

$$\mathbf{W}(t) \begin{pmatrix} w_1(t) \\ \vdots \\ w_m(t) \end{pmatrix} = - \begin{pmatrix} u_1^{(m)}(t) \\ \vdots \\ u_m^{(m)}(t) \end{pmatrix}$$

yielding

$$\begin{pmatrix} w_1(t) \\ \vdots \\ w_m(t) \end{pmatrix} = -\mathbf{W}(t)^{-1} \begin{pmatrix} u_1^{(m)}(t) \\ \vdots \\ u_m^{(m)}(t) \end{pmatrix}.$$

Obviously, the w_j's are continuous, by our assumptions concerning the u_i's.

7 The Green's Function Associated with the Differential Operator L

The definition below gives the definition of $G(\cdot,\cdot)$, the Green's function associated with the differential operator L with specified boundary conditions. Theorem 8 gives an easily calculated form of G. The Green's function is used in Section 4 to calculate the reproducing kernel.

Let L be the linear differential operator (3) defined on $\mathcal{H}^m$.

Definition G is a *Green's function* for L if and only if

$$f(t) = \int_{u=a}^{b} G(t,u)\ (\mathrm{L}f)(u)\ du$$

for all functions f in $\mathcal{H}^m$ satisfying the boundary conditions

$$f^{(j)}(a) = 0, \quad j = 0, \cdots, m-1. \tag{16}$$

Of course, it's not immediately clear that such a function G exists. However, G exists and is easily calculated by using the Wronskian matrix associated with L. Recall from Theorem 4 of Section 6 that there exists a basis for the kernel of L, $u_1, \cdots, u_m$, with invertible Wronskian. Furthermore, each u_i has m derivatives. Theorem 8 shows how to calculate G.

Theorem 8 *Let $u_1^*(t), \cdots, u_m^*(t)$ denote the entries in the last row of the inverse of $\mathbf{W}(t)$. Then*

$$G(t,u) = \begin{cases} \sum_{i=1}^m u_i(t)\ u_i^*(u) & \text{for } u \le t \\ 0 & \text{else} \end{cases} \tag{17}$$

is a Green's function for L*.*

The following theorem will be useful in the proof of Theorem 8.

Theorem 9 *Let G be as in (17) and suppose that $h \in \mathcal{L}^2$. If*

$$r(t) = \int_a^b G(t,u)\ h(u)\ du$$

Then

$$r \in \mathcal{H}^m, \tag{18}$$

$$(\mathrm{L}r)(t) = h(t) \quad \text{almost everywhere } t \in [a,b] \tag{19}$$

and

$$r^{(j)}(a) = 0 \quad j = 0, \cdots, m-1. \tag{20}$$

Proof. Write

$$r(t) = \int_a^b G(t,u)\ h(u)\ du = \sum_{i=1}^m u_i(t) \int_a^t u_i^*(u)\ h(u)\ du.$$

Note that the u_i^*'s are continuous, since $u_i^* = (\det \mathbf{W}(t))^{-1}$ times an expression involving sums and products of $u_l^{(j)}$, $l = 1, \cdots, m$, $j = 0, \cdots, m-1$, and the u_l's have $m-1$ continous derivatives. We'll first show that

$$r^{(j)}(t) = \sum_{i=1}^m u_i^{(j)}(t) \int_a^t u_i^*(u)\ h(u)\ du \quad j = 0, \cdots, m-1 \tag{21}$$

and

$$r^{(m)}(t) = h(t) \; + \; \sum_{i=1}^{m} u_i^{(m)}(t) \int_a^t u_i^*(u) \; h(u) \; du \quad \text{almost everywhere } t \in [a,b]. \tag{22}$$

These equations follow easily by induction on j. We only present the case $j = 1$. Then

$$r'(t) = \sum_{i=1}^{m} u_i'(t) \int_a^t u_i^*(u) \; h(u) \; du \quad + \quad \sum_{i=1}^{m} u_i(t) \; \frac{d}{dt}\left[\int_a^t u_i^*(u) \; h(u) \; du \right].$$

Since u_i^* and h are in $\mathcal{L}_2$,

$$\sum_{i=1}^{m} u_i(t) \; \frac{d}{dt}\left[\int_a^t u_i^*(u) \; h(u) \; du \right] = \sum_{i=1}^{m} u_i(t) \; u_i^*(t) \; h(t)$$

almost everywhere t. But, by definition of $\mathbf{W}$ and the u_i^*'s, this is equal to

$$h(t) \; \sum_i [\mathbf{W}(t)]_{i1} \; [\mathbf{W}(t)^{-1}]_{mi} = h(t) \; [\mathbf{W}(t)^{-1} \; \mathbf{W}(t)]_{m1}$$

which equals zero for $m > 1$ and equals $h(t)$ for $m = 1$. Therefore, for $m = 1$, (22) holds and for $m > 1$ (21) holds when $j = 1$. For $m > 1$ and $j > 1$, we can calculate derivatives of r up to order $m - 1$, and can calculate the mth derivative almost everywhere to prove (21) and (22). Clearly, the mth derivative in (22) is square-integrable. Therefore we've proven (18).

To prove (19), use (21) and (22) and write

$$(\mathrm{L}r)(t) = r^{(m)}(t) + \sum_{j=0}^{m-1} w_j(t) \; r^{(j)}(t)$$

$$= h(t) + \sum_{i=1}^{m} u_i^{(m)}(t) \int_a^t u_i^*(u) \; h(u) \; du + \sum_{j=0}^{m-1} \sum_{i=1}^{m} w_j(t) u_i^{(j)}(t) \int_a^t u_i^*(u) \; h(u) \; du$$

$$= h(t) + \sum_{i=1}^{m} \left[u_i^{(m)}(t) + \sum_{j=0}^{m-1} \sum_{i=1}^{m} w_j(t) u_i^{(j)}(t) \right] \int_a^t u_i^*(u) \; h(u) \; du$$

$$= h(t) \; + \; \sum_{i=1}^{m} (\mathrm{L}u_i)(t) \int_a^t u_i^*(u) \; h(u) \; du = h(t)$$

since $\mathrm{L}u_i \equiv 0$.

Equation (20) follows directly from (21) by taking $t = a$.

Proof of Theorem 8. Let $f \in \mathcal{H}^m$ satisfy (16). Define $r(t) = \int_a^b G(t,u)\,(\mathrm{L}f)(u)\,du$. Then, by Theorem 9 of section 6, $\mathrm{L}r = \mathrm{L}f$ almost everywhere and $r^{(j)}(a) = 0$, $j = 0, \cdots, m-1$. Thus $\mathrm{L}(r-f) = 0$ almost everywhere and $(r-f)^{(j)}(a) = 0, j = 0, \cdots, m-1$. By Theorem 7, $r - f$ is the zero function, that is $r = f$.

8 Bibliography

Andrews, D. F. and Herzberg, A. M. (1985) *Data: A Colleciton of Problems from Many Fields for the Student and Research Worker*, New York: Springer Verlag.

Anselone, P. M. and Laurent, P. J. (1967) A general method for the construction of interpolating or smoothing spline-functions. *Numerische Mathematik*, **12**, 66-82.

Ansley, C., Kohn, R., and Wong, C. (1993) Nonparametric spline regression with prior information. *Biometrika* **80**, 75-88.

Bacchetti, P., Segal, M.R., Hessol, N.A., and Jewell, N.P. (1993) Different AIDS Incubation Periods and Their Impacts on Reconstructing Human Immunodeficiency Virsu Epidemics and Projecting AIDS Incidence, *Proceeding of the National Academy of Sciences, USA*, **90**, 2194-2196.

Coddington, E. A. (1961) *An Introduction to Ordinary Differential Equations.* New Jersey: Prentice-Hall.

Eubank, R. L. (1988) *Spline Smoothing and Nonparametric Regression.* New York: Marcel Dekker.

Green, P. J. and Silverman, B. W. (1994) *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach.* London: Chapman and Hall.

Heckman, N. and Ramsay, J.O. (1996) Penalized Regression with Model-Based Penalties.

Li, Xiaochun. (1996) *Local Linear Regression versus Backcalculation in Forecasting* . Ph.D. thesis, Statistics Department, University of British Columbia.

Nychka, D., Wahba, G., Goldfarb, S. and Pugh, T. (1984) Cross-validated Spline Methods for the Estimation of Three-dimensional Tumor Size Distributions from Observations on Two-dimensional Cross Sections, *Journal of the American Statistical Association* **78**, 832-846.

Ramsay, J.O. and Heckman, N. (1996) Some theory for L-spline smoothing. Technical Report 165, Statistics Department University of British Columbia.

Reinsch, C. (1967) Smoothing by spline functions. *Numerische Mathematik,* **10,** 177-183.

Reinsch, C. (1970) Smoothing by spline functions II. *Numerische Mathematik,* **16**, 451-454.

Wahba, G. (1990) *Spline Models for Observational Data.* Philadelpha: Society for Industrial and Applied Mathematics.