Statistics Surveys
Vol. 6 (2012) 113–141
ISSN: 1935-7516
DOI: 10.1214/12-SS101

The theory and application of penalized methods or Reproducing Kernel Hilbert Spaces made easy[*]

Nancy Heckman

Department of Statistics
The University of British Columbia
Vancouver BC Canada
e-mail: nancy@stat.ubc.ca

Abstract: The popular cubic smoothing spline estimate of a regression function arises as the minimizer of the penalized sum of squares $\sum_j (Y_j - \mu(t_j))^2 + \lambda \int_a^b [\mu''(t)]^2 \, dt$, where the data are t_j, Y_j, $j = 1, \ldots, n$. The minimization is taken over an infinite-dimensional function space, the space of all functions with square integrable second derivatives. But the calculations can be carried out in a finite-dimensional space. The reduction from minimizing over an infinite dimensional space to minimizing over a finite dimensional space occurs for more general objective functions: the data may be related to the function μ in another way, the sum of squares may be replaced by a more suitable expression, or the penalty, $\int_a^b [\mu''(t)]^2 \, dt$, might take a different form. This paper reviews the Reproducing Kernel Hilbert Space structure that provides a finite-dimensional solution for a general minimization problem. Particular attention is paid to the construction and study of the Reproducing Kernel Hilbert Space corresponding to a penalty based on a linear differential operator. In this case, one can often calculate the minimizer explicitly, using Green's functions.

AMS 2000 subject classifications: Primary 62G99, 46E22; secondary 62G08.
Keywords and phrases: Penalized likelihood, Reproducing Kernel Hilbert Space, splines.

Received November 2011.

Contents

[*] This paper was accepted by Stephen Bruce Vardeman.

1. Introduction

Data are often modelled in terms of a function: a density function provides a model for the distribution of univariate data, a regression function provides a model for dependence in bivariate data, and a logistic regression function can be used to classify an individual based on covariate information. Classic statistical methods define functions in terms of a small number of parameters such as a mean and variance, or a slope and an intercept. In contrast, many current statistical methods are less restrictive, modelling a function as smooth, lying in an infinite dimensional function space.

Modelling and computation in an infinite dimensional function space are facilitated by techniques of functional analysis, specifically, by using a Hilbert space structure. A Hilbert space is a vector space with an inner product. This inner product allows us to define projections onto subspaces, a useful tool in optimization problems. The inner product also allows us to define a norm, and through the norm, to define convergence of sequences of functions and continuity of functionals. The well-known $\mathcal{L}^2$ space of functions is a Hilbert space with the inner product of f and g simply the integral $\int f(t)g(t)\, dt$. Unfortunately, the evaluation functionals, which take f to $f(t)$, are not well-defined in $\mathcal{L}^2$, as function values are only defined "almost everywhere t". That is, we cannot with certainty state the value $f(t)$. This situation is a bit troubling for statistical analysis, for instance, for estimating a regression function. Fortunately, in a particular type of Hilbert space, namely, a Reproducing Kernel Hilbert Space (RKHS), the evaluation functionals are well-defined and, even better, they are continuous. By continuity of an evaluation functional, we mean that when two functions f and g are close in terms of the norm of the RKHS, then $f(t)$ and $g(t)$ are close, in the usual sense of closeness on the real line. These concepts of closeness and continuity are important in any principled approach to estimating a function from noisy data.

This article reviews the tools of Hilbert spaces and RKHS's for analyzing data when the parameter of interest is in an infinite dimensional function space. Specifically, consider data, $Y_1, \ldots Y_n \in \Re$ and $t_1, \ldots, t_n \in \Re^p$, where the distribution of the Y_i's depends on μ, a function of $t \in \Re^p$, which is usually assumed to be smooth. The goal is to find μ in a specified space $\mathcal{H}$ to minimize

$$\mathcal{G}(t_1, \ldots, t_n, Y_1, \ldots, Y_n, F_1(\mu), \ldots, F_n(\mu)) + \lambda P(\mu) \tag{1.1}$$

where $\mathcal{G}$ and the F_j's are known, P is a known penalty on μ, and λ serves to balance the importance between $\mathcal{G}$ and P. Typically, $F_j(\mu) = \mu(t_j)$, although we are not limited to this choice.

While some results here will concern general P and $t_j \in \Re^p$, $p \geq 1$, much of the paper considers the restricted case where $t_j \in [a, b] \subset \Re$ and P is generated from a differential operator L:

$$P(\mu) = \int_a^b [(\mathrm{L}\mu)(t)]^2 \; dt \quad \text{where} \quad (\mathrm{L}\mu)(t) = \mu^{(m)}(t) + \sum_{j=0}^{m-1} w_j(t)\mu^{(j)}(t) \tag{1.2}$$

with w_j real-valued and continuous.

The most well-known case of (1.1) occurs in regression analysis, when we seek the regression function $\hat{\mu} \in \mathcal{H}^2[a, b]$ to minimize

$$\sum_j [Y_j - \mu(t_j)]^2 + \lambda \int_a^b [\mu''(t)]^2 \; dt. \tag{1.3}$$

The minimizing μ is a cubic smoothing spline, a popular regression function estimate. The non-negative smoothing parameter λ balances the minimizing μ's fit to the data (via minimizing $\sum_j [Y_j - \mu(t_j)]^2$) with its closeness to a straight line (achieved when $\int_a^b [\mu''(t)]^2 \; dt = 0$). The value of λ is typically chosen "by eye" – by examining the resulting estimates of μ, or by some automatic data-driven method such as cross-validation. See, for instance, Wahba [30], Eubank [8] or Green and Silverman [10].

To extend (1.3) to (1.1), we can consider a first term other than a sum of squares, functionals other than $F_j(\mu) = \mu(t_j)$ and a differential operator other than the second derivative operator. Examples of these variations are given in Section 2. Section 3 provides background on Hilbert Spaces and RKHS's, and contains the reduction of (1.1) to a finite dimensional optimization problem. Section 4 relates the minimizer of (1.1) to a Bayes estimate. Sections 5 and 6 contain results and algorithms for minimizing (1.1) with P based on a differential operator as in (1.2), with Section 5 containing the "warm-up" of the cubic smoothing spline result for minimizing (1.3) and Section 6 containing the general case. The Appendix contains pertinent results from the theory of solutions of differential equations.

The material contained here is, for the most part, not original. The material is drawn from many sources: from statistical and machine learning literature, from the theory of differential equations, from numerical analysis, and from functional analysis. The purpose of this paper is to collect this diverse material in one article and to present it in an easily accessible form, to show the richness of statistical problems that involve minimizing (1.1) and to explain the theory and provide easy to follow algorithms for minimizing (1.1). A briefer review of RKHS's can be found in Wahba [31].

This article presents an approach to using the structure of a Reproducing Kernel Hilbert Space to minimize (1.1) that is different from the usual approach of, say, machine learning. A usual approach to this minimization problem is to first specify a *kernel function* K, defined on $\Re^p \times \Re^p$, and then define a set of basis functions, namely $\{K(t,\cdot), t \in \Re^p\}$, and finally to use K and the basis to construct an inner product and an RKHS. See, for instance, Hastie *et al.* [11] and Hofman *et al.* [13]. Here, the approach is different: we specify the penalty P and use P to construct the corresponding RKHS. This is a more model-based approach, since the set of functions with $P\mu \equiv 0$ typically defines a finite dimensional model space for μ. We add flexiblity to our estimation of μ by allowing departure from this finite dimensional model space: the term $\lambda P(\mu)$ in (1.1) replaces our belief that "$P(\mu) = 0$" with a belief that "$P(\mu)$ may be small". The case when $P(\mu)$ is based on derivatives of μ as in (1.2) has been recommended by Heckman and Ramsay [12] as a flexible model-based approach to smoothing. See Section 2.7. When $P(\mu)$ is as in (1.2), if one can find the solutions of $\mathrm{L}\mu = 0$, then one can explicitly calculate the kernel function K that is associated with the RKHS and thus can explicitly calculate the minimizing μ. See Section 6.

2. Examples

2.1. Penalized likelihoods with $F_j(f) = f(t_j)$

Most statistical applications that lead to minimizing (1.1) have the first term in (1.1) equal to a negative log likelihood. In these cases, the μ that minimizes (1.1) is called a penalized likelihood estimate of μ. Indeed, (1.3) yields a penalized likelihood estimator: the sum of squares arises from a likelihood by assuming that $Y_1, \ldots, Y_n$ are independent normally distributed with the mean of Y_j equal to $\mu(t_j)$ and the variance equal to σ^2. Then, apart from a constant, $-2\times$ the log likelihood is simply

$$n \log(\sigma^2) + \frac{1}{\sigma^2} \sum (Y_j - \mu(t_j))^2.$$

A penalized likelihood estimate of μ with penalty $P(\mu)$ minimizes

$$\begin{aligned} & n \log(\sigma^2) + \frac{1}{\sigma^2} \sum (Y_j - \mu(t_j))^2 + \lambda^* P(\mu) \\ & \quad = \frac{1}{\sigma^2} \left[\sigma^2 n \log(\sigma^2) + \sum (Y_j - \mu(t_j))^2 + \lambda^* \sigma^2 P(\mu) \right]. \end{aligned}$$

Thus, for a given σ^2, the penalized likelihood estimate of μ minimizes (1.1) with $\lambda = \lambda^* \sigma^2$. If the Y_j's are not independent but the vector $(Y_1, \ldots, Y_n)'$ has covariance matrix $\sigma^2 \Sigma$, then we would replace the sum of squares with $\sum_{j,k} [Y_j - \mu(t_j)]\ \Sigma^{-1}[j,k]\ [Y_k - \mu(t_k)]$.

Another likelihood, important in classification, is based on data $Y_j = 1$ or -1 with probabilities $p(t_j)$ and $1 - p(t_j)$, respectively. Thus

$$\text{the log likelihood } = \sum\nolimits_j \frac{1+Y_j}{2} \log p(t_j) + \frac{1-Y_j}{2} \log[1 - p(t_j)].$$

To avoid placing inequality constraints on the function of interest, we reparameterize by setting $\mu(t) = \log[p(t)/(1-p(t))]$ or equivalently $p(t) = \exp(\mu(t))/[1 + \exp(\mu(t))]$. This reparameterization yields

$$\text{the log likelihood} = \sum_j \frac{1+Y_j}{2} \log \frac{\exp(\mu(t_j))}{1+\exp(\mu(t_j))} + \frac{1-Y_j}{2} \log \frac{1}{1+\exp(\mu(t_j))}. \tag{2.1}$$

2.2. Regularized regression

The minimization of (1.3) is an example of *regularized regression* where the term $P(\mu)$ is called the stabilizer or regularizer. Regularized regression is used when the usual criterion, e.g. least squares, does not yield an appropriate solution due to, for instance, poorly conditioned matrices that may be difficult or impossible to invert. See Hastie *et al* [11]. The most well-known example of regularized regression is ridge regression in which $\mu(t)$ is modelled as $\sum_1^K \beta_j \phi_j(t)$ for known ϕ_j's. If K is large relative to n, the number of data points, we are in danger of over-fitting the data. To prevent this, we add a stabilizer and we minimize as a function of β

$$\sum_1^n \left[Y_i - \sum_1^K \beta_j \phi_j(t_i) \right]^2 + \lambda \sum \beta_j^2.$$

A well-known regularized regression method, the *lasso*, replaces $\sum \beta_j^2$ with $\sum |\beta_j|$. See Tibshirani [28]. This minimization problem, however, does not fit the general formulation for a RKHS.

2.3. Gaussian processes

In machine learning and in the spatial analysis technique called Kriging, the function μ is modelled as the realization of a Gaussian process with the "estimate" of $\mu(t)$ being the best linear unbiased predictor of $\mu(t)$. This estimation problem can be reformulated as an optimization problem as in (1.1) with μ lying in an RKHS that is defined in terms of the covariance of the underlying Gaussian process. Section 4 discusses this in detail, along with references. In machine learning, the covariance structure of the Gaussian process usually reflects the amount of smoothness. In Kriging, the covariance is typically taken as one of the standard spatial covariance functions. For a discussion of the connections between Kriging, spatial process covariances, penalized likelihood and RKHS's, see Furrer and Nychka [9] and Nychka [21]. The first paper also contains elegant asymptotic results that relate the estimate of $\mu(t)$ to a weighted average estimate, called the *equivalent kernel estimate*. The weights depend on the underlying reproducing kernel and spatial covariance structure via the associated Green's function.

2.4. F_j's based on integrals

While $F_j(\mu) = \mu(t_j)$ is common, $F_j(\mu)$ is sometimes chosen to involve an integral of μ, specifically, $F_j(\mu) = \int_a^b H_j(s)\mu(s)ds$, with the H_j's known. See Wahba [30].

Li [19] and Bacchetti *et al.* [5] used (1.1) to estimate $\mu(t)$, the HIV infection rate at time t, based on data, Y_j, the number of new AIDS cases diagnosed in time period $(t_{j-1}, t_j]$. The expected value of Y_j depends not only on $\mu(t_j)$, but also on $\mu(t)$ for values of $t \le t_j$. This dependence involves the distribution of the time of progress from HIV infection to AIDS diagnosis, which is estimated from cohort studies. Letting $\mathcal{F}(t|s)$ denote the probability that AIDS has developed by time t given HIV infection occurred at time s,

$$\mathrm{E}\left(\sum_1^j Y_i\right) = \int_{s=0}^{t_j} \mu(s)\mathcal{F}(t_j|s)\ ds \equiv F_j(\mu).$$

Thus we could define the first term in (1.1) as a negative log likelihood assuming the Y_j's are independent Poisson counts with $\mathrm{E}(Y_j) = F_j(\mu) - F_{j-1}(\mu)$. Or we could take the computationally simpler approach by setting the first term in (1.1) equal to

$$\sum_1^n \left\{Y_j - \left[F_j(\mu) - F_{j-1}(\mu)\right]\right\}^2.$$

Both Li [19] and Bacchetti *et al.* [5] use this simpler approach, with the former using penalty $P(\mu) = \int(\mu'')^2$ while the latter used a discretized version of $\int(\mu'')^2$.

In a density estimation setting, Nychka *et al.* [20] estimated the distribution of the volumes of tumours in livers by using data from cross-sectional slices of the livers. The authors modelled tumours as spheres and so cross-sections were circles. They estimated μ, the probability density of the spheres' radii, using an integral to relate the radius of a sphere to the radius of a random slice of the sphere. Their estimation criterion was the minimization of an expression of the form (1.1) with F_j using that integral and with $P(\mu) = \int(\mu'')^2$.

2.5. Functional linear regression

The functional $F_j(\mu) = \int_a^b H_j(s)\ d\mu(s)$ of the previous section is a key element in what is called functional linear regression with a scalar response.

See, for instance, Yuan and Cai [32], Ramsay and Silverman [23] and Horváth and Kokoszka [14] and references there-in. In functional linear regression with a scalar response, the data are $Y_1, \dots Y_n \in \Re$ and the processes $X_j(t),\ t \in [a,b],\ j = 1, \dots, n$, with the processes perhaps partially observed and/or observed with error. We suppose that Y_j depends on X_j in a way that mimics usual linear regression:

$$Y_j = \alpha_0 + \int X_j(s)\ \mu(s)\ ds + \epsilon_j$$

where ϵ_j is error, independent of X_j, and both α_0 (the "intercept") and μ (the "slope function") are unknown and to be estimated. For example, in Ramsay and Silverman [23], in a study of Canadian weather stations, Y_j is the annual precipitation at weather station j and $X_j(t)$ is the temperature at the jth station on day t. The standard estimates of α_0 and μ minimize

$$\sum_j \left[Y_j - \alpha_0 - \int X_j(s)\ \mu(s)\ ds \right]^2 + \lambda P(\mu) \tag{2.2}$$

for some penalty P. If we do not observe the process X_j exactly and only observe it for a finite number of values of t and possibly with error, then we could proceed with the minimization of (2.2) but with X_j replaced by an estimate of X_j. Ramsay and Silverman [23] follow this approach, using a least-squares estimate of X_j with a flexible basis of Bspline functions. Crambes *et al.* [7] take a slightly different approach. They consider the case with X_j observed without error at $t_1, \dots, t_n$ where $t_k - t_{k-1} = 1/n$. They use Rieman sums, finding $\hat{\mu}$ to minimize

$$\sum_j \left[Y_j - \bar{Y} - \frac{1}{n} \sum_k \left[X_j(t_k) - \bar{X}(t_k) - \mu(t_k) \right] \right]^2 + \lambda P(\mu)$$

with respect to μ.

2.6. Support vector machines

Support vector machines are a classification tool, with classification rules built from data $Y_i \in \{-1, 1\}$, $t_i \in \Re^p$ (see, for instance, Hastie *et al.* [11]). The goal is to find a function μ for classifying: classify Y_i as 1 if and only if $\mu(t_i) > 0$. We see that Y_i is misclassified by this rule if and only if $Y_i \mu(t_i)$ is negative. Thus, it is common to find μ to minimize $-\sum_i \text{sign}[Y_i \mu(t_i)]$ subject to some penalty for rough μ: that is, to find $\hat{\mu}$ to minimize

$$-\sum_i \text{sign}[Y_i \mu(t_i)] + \lambda P(\mu).$$

This can be made more general by minimizing

$$\sum_j H[Y_i \mu(t_j)] + \lambda P(\mu)$$

for a known non-increasing function H. The function $H(x) = -\operatorname{sign}(x)$ is not continuous at 0, which can make minimization challenging. To avoid this problem, Wahba [29] proposed using "softer" H functions, such as $H(x) = \ln[1 + \exp(-x)]$. This function is not only continuous, but is differentiable and convex. Wahba [29] showed that this H corresponds to a negative log likelihood. Specifically, she showed that the log likelihood in (2.1) is equal to $-\sum \log\{1 + \exp[-Y_j\mu(t_j)]\}$.

2.7. Using different differential operators in the penalty

Ansley, Kohn, and Wong [3] and Heckman and Ramsay [12] demonstrated the usefulness of appropriate choices of L in the penalty $P(\mu) = \int(\mathrm{L}\mu)^2$. For instance, Heckman and Ramsay compared two estimates of a regression function for the incidence of melanoma in males. The data, described in Andrews and Herzberg [1], are from the Connecticut Tumour Registry, for the years 1936 to 1972. The data show a roughly periodic trend superimposed on an increasing trend. A cubic smoothing spline, the minimizer of (1.3), tracks the data fairly well, but slightly dampens the periodic component. This dampening does not occur with Heckman and Ramsay's preferred estimate, the estimate that minimizes a modified version of (1.3) but with the penalty $\int[\mu''(t)]^2\,dt$ replaced by the penalty $\int[\mu^{(4)}(t) + \omega^2\mu''(t)]^2\,dt$ with $\omega = 0.58$. The differential operator $\mathrm{L}{=}\mathrm{D}^4 + \omega^2\mathrm{D}^2$ was chosen since it places no penalty on 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 satisfying $\mathrm{L}\mu \equiv 0$ 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.

The use of appropriate differential operators in the penalty has been further developed in the field of Dynamic Analysis. See, for instance, Ramsay *et al.* [22]. These authors use differential operators equal to those used by subject area researchers, who typically work in the finite dimensional space defined by solutions of $\mathrm{L}\mu \equiv 0$.

3. Results for the general minimization problem

This section contains some background on Reproducing Kernel Hilbert Spaces and shows how to use Reproducing Kernel Hilbert Space structure to reduce the minimization of (1.1) to minimization over a finite-dimensional function space (see Theorem 3.1). Whether or not the minimizer exists can be determined by studying the finite-dimensional version. While a complete review of Hilbert spaces is beyond the scope of this article, a few definitions may help the reader. Further background on Hilbert spaces can be found in any standard functional analysis textbook, such as Kolmogorov and Fomin [17] or Kreyszig [18]. For a condensed exposition of the necessary Hilbert space theory, see, for instance, Wahba [30, 31] or the appendix of Thompson and Tapia [27]. We will only consider Hilbert spaces over $\Re$.

Consider $\mathcal{H}$, a collection of functions from $\mathcal{T} \subseteq \Re^p$ to $\Re$. Suppose that $\mathcal{H}$ is a vector space over $\Re$ with inner product $< \cdot, \cdot >$. The inner product induces a norm on $\mathcal{H}$, namely $||f|| = [< f, f >]^{1/2}$. The existence of a norm allows us to define limits of sequences in $\mathcal{H}$ and continuity of functions with arguments in $\mathcal{H}$. The vector space $\mathcal{H}$ is a *Hilbert space* if it is complete with respect to this norm, that is, if any Cauchy sequence in $\mathcal{H}$ converges to an element of $\mathcal{H}$.

A *linear functional* F is a function from a Hilbert space $\mathcal{H}$ to the reals satisfying $F(\alpha f + \beta g) = \alpha F(f) + \beta F(g)$ for all $\alpha, \beta \in \Re$ and all $f, g \in \mathcal{H}$. The Riesz Representation Theorem states that a linear functional F is continuous on $\mathcal{H}$ if and only if there exists $\eta \in \mathcal{H}$ such that $< \eta, f >= F(f)$ for all $f \in \mathcal{H}$. The function η is called the representer of F.

The Hilbert space $\mathcal{H}$ is a *Reproducing Kernel Hilbert Space* if and only if, for all $t \in \mathcal{T}$, the linear functional $F_t(f) \equiv f(t)$ is continuous, that is, if and only if, for all $t \in \mathcal{T}$, there exists $R_t \in \mathcal{H}$ such that $< R_t, f >= f(t)$ for all $f \in \mathcal{H}$. Noting that the collection of R_t's, $t \in \mathcal{T}$, defines a bivariate function R, namely $R(s,t) \equiv R_t(s)$, we see that $\mathcal{H}$ is a Reproducing Kernel Hilbert Space if and only if there exists a bivariate function R defined on $\mathcal{T} \times \mathcal{T}$ such that $< R(\cdot, t), f >= f(t)$ for all $f \in \mathcal{H}$ and all $t \in \mathcal{T}$. The function R is called the reproducing kernel of $\mathcal{H}$.

One can show that the reproducing kernel is symmetric in its arguments, as follows. To aid the proof, use the notation that $R_t(s) = R(s,t)$ and $R_s(t) = R(t,s)$. By the reproducing properties of R_t and R_s, $< R_t, R_s >= R_s(t)$ and $< R_s, R_t >= R_t(s)$. But the inner product is symmetric, that is $< R_t, R_s >=< R_s, R_t >$. So $R_s(t) = R_t(s)$.

To give the form of the finite-dimensional minimizer of (1.1), we assume that the following conditions hold.

(C.1) There are $\mathcal{H}_0$ and $\mathcal{H}_1$, linear subspaces of $\mathcal{H}$, with $\mathcal{H}_1$ the orthogonal complement of $\mathcal{H}_0$.

(C.2) $\mathcal{H}_0$ is of dimension $m < \infty$, with basis $u_1, \ldots, u_m$. If $m = 0$, take $\mathcal{H}_0$ equal to the empty set and $\mathcal{H}_1 = \mathcal{H}$.

(C.3) There exists $R_0 \in \mathcal{H}_0$ and $R_1 \in \mathcal{H}_1$ such that R_i is a reproducing kernel for $\mathcal{H}_i$, in the sense that $< R_i(\cdot, t), \mu >= \mu(t)$ for all $\mu \in \mathcal{H}_i$, $i = 0, 1$.

Since $\mathcal{H}_0$ is finite dimensional, it is closed. The orthogonal complement of a subspace is always closed. Thus Condition (C.1) implies that any $\mu \in \mathcal{H}$ can be written as $\mu = \mu_0 + \mu_1$ for some $\mu_0 \in \mathcal{H}_0$ and $\mu_1 \in \mathcal{H}_1$ and that $< \mu_0, \mu_1 >= 0$. This is often written as $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$. Note that Conditions (C.1), (C.2) and (C.3) imply that $R \equiv R_0 + R_1$ is a reproducing kernel for $\mathcal{H}$.

We require one more condition, relating the penalty P to the partition of $\mathcal{H}$.

(C.4) Write $\mu = \mu_0 + \mu_1$, with $\mu_i \in \mathcal{H}_i$. Then $P(\mu) =< \mu_1, \mu_1 >$.

Theorem 3.1. *Suppose that conditions (C.1) through (C.4) hold and that* $F_1, \ldots, F_n$ *are continuous linear functionals on* $\mathcal{H}$. *Let* $\eta_{j1}(t) = F_j(R_1(\cdot, t))$, *that is,* F_j *applied to the function* R_1 *considered as a function of* s, *with* t *fixed.*

Then to minimize (1.1), it is necessary and sufficient to find

$$\mu(t) \equiv \mu_0(t) + \mu_{11}(t) \equiv \sum_1^m \hat{\alpha}_j u_j(t) + \sum_1^n \hat{\beta}_j \eta_{j1}(t)$$

where the $\hat{\alpha}_j$'s and $\hat{\beta}_j$'s minimize

$$\mathcal{G}(t_1, \ldots, t_n, Y_1, \ldots, Y_n, F_1(\mu_0 + \mu_{11}), \ldots, F_n(\mu_0 + \mu_{11})) + \lambda \boldsymbol{\beta}' K \boldsymbol{\beta}.$$

Here $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_n)'$ and the matrix K is symmetric and non-negative definite, with $K[j,k] = F_j(\eta_{k1})$. If $F_j(f) = f(t_j)$ and $F_k(f) = f(t_k)$, then $\eta_{1j}(t) = R_1(t_j, t)$, $\eta_{1k}(t) = R_1(t_k, t)$ and $K[j,k] = R_1(t_j, t_k)$.

Proof. By the Riesz Representation Theorem, there exists a representer $\eta_j \in \mathcal{H}$ such that $< \eta_j, \mu >= F_j(\mu)$ for all $\mu \in \mathcal{H}$. Applying the Riesz Representation Theorem to the subspaces $\mathcal{H}_0$ and $\mathcal{H}_1$, which can be considered as Hilbert spaces in their own rights, there exists $\eta_{j0} \in \mathcal{H}_0$ and $\eta_{j1}^* \in \mathcal{H}_1$, representers of F_j in the sense that $< \eta_{j0}, \mu >= F_j(\mu)$ for all $\mu \in \mathcal{H}_0$ and $< \eta_{j1}^*, \mu >= F_j(\mu)$ for all $\mu \in \mathcal{H}_1$. One easily shows that this η_{j1}^* is equal to η_{j1}, as defined in the statement of the Theorem: by the definition of the representer of F_j, η_{j1}^* must satisfy $F_j(R_1(\cdot, t)) =< \eta_{j1}^*, R_1(\cdot, t) >$. But, by the reproducing quality of R_1, $< \eta_{j1}^*, R_1(\cdot, t) >= \eta_{j1}^*(t)$. So $\eta_{j1}^* = \eta_{j1}$. One also easily shows that

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

We use the η_{j1}'s to partition $\mathcal{H}_1$ as follows. Let $\mathcal{H}_{11}$ be the finite dimensional subspace of $\mathcal{H}_1$ spanned by $\eta_{j1}, j = 1, \ldots, n$, and let $\mathcal{H}_{12}$ be the orthogonal complement of $\mathcal{H}_{11}$ in $\mathcal{H}_1$. Then $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_{11} \oplus \mathcal{H}_{12}$ and so any $\mu \in \mathcal{H}$ can be written as

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

We now show that any minimizer of (1.1) must have $\mu_{12} \equiv 0$. Let μ be any element of $\mathcal{H}$. Since η_j is the representer of F_j and μ_{12} is orthogonal to η_j,

$$F_j(\mu) =< \eta_j, \mu >=< \eta_j, \mu_0 + \mu_{11} + \mu_{12} >=< \eta_j, \mu_0 + \mu_{11} >= F_j(\mu_0 + \mu_{11}).$$

Therefore, μ_{12} is irrelevant in computing the first term in (1.1). To study the second term in (1.1), by (C.4) and the orthogonality of μ_{11} and μ_{12},

$$P(\mu) =< \mu_1, \mu_1 >=< \mu_{11}, \mu_{11} > + < \mu_{12}, \mu_{12} > .$$

Therefore, we want to find $\hat{\mu}_0 \in \mathcal{H}_0$, $\hat{\mu}_{11} \in \mathcal{H}_{11}$ and $\hat{\mu}_{12} \in \mathcal{H}_{12}$ to minimize

$$\mathcal{G}(t_1, \ldots, t_n, Y_1, \ldots, Y_n, F_1(\mu_0 + \mu_{11}), \ldots, F_n(\mu_0 + \mu_{11})) \\ + \lambda \left[< \mu_{11}, \mu_{11} > + < \mu_{12}, \mu_{12} >\right].$$

Clearly, we should take $\hat{\mu}_{12}$ to be the zero function and so any minimizer of (1.1) is of the form

$$\begin{aligned}\mu(t) &= \mu_0(t) + \mu_{11}(t) \\ &= \sum_1^m \alpha_j u_j(t) + \sum_1^n \beta_j \eta_{j1}(t).\end{aligned}$$

Now consider rewriting $P(\mu)$ as $\boldsymbol{\beta}' K \boldsymbol{\beta}$: $P(\mu) =< \mu_{11}, \mu_{11} >= \sum_{j,k} \beta_j \beta_k < \eta_{j1}, \eta_{k1} > \equiv \boldsymbol{\beta}' K^* \boldsymbol{\beta}$ for K^* symmetric and non-negative definite. To show that $K^*[j,k] = F_j(\eta_{k1})$, use the fact that η_{j1} is the representer of F_j in $\mathcal{H}_1$, that is, that $< \eta_{j1}, f >= F_j(f)$ for all $f \in \mathcal{H}_1$. Applying this to $f = \eta_{k1}$ yields the desired result, that $< \eta_{j1}, \eta_{k1} >= F_j(\eta_{k1})$.

Consider the case that $F_j(f) = f(t_j)$ and $F_k(f) = f(t_k)$. Then $\eta_{1j}(t) = F_j(R_1(\cdot, t)) = R_1(t_j, t)$, $\eta_{1k}(t) = R_1(t_k, t)$, and $K[j,k] = F_j(\eta_{k1}) = R_1(t_k, t_j) = R_1(t_j, t_k)$ by symmetry of R_1. □

The proof of the following Corollary is immediate, by taking $m = 0$ in (C.2).

Corollary 3.1. *Suppose that $\mathcal{H}$ is an RKHS with inner product $< \cdot, \cdot >$ and reproducing kernel R. In (1.1), suppose that $P(\mu) =< \mu, \mu >$ and assume that the F_j's are continuous linear functionals. Then the minimizer of (1.1) is of the form*

$$\mu(t) = \sum_1^n \beta_j F_j(R_1(\cdot, t)).$$

4. A Bayesian connection

Sometimes, the minimizer of (1.1) is related to a Bayes estimate of μ. In the Bayes formulation, $Y_j = \mu(t_j) + \epsilon_j$ where the ϵ_j's are independent normal random variables with zero means and variances equal to σ^2. The function μ is the realization of a stochastic process and is independent of the ϵ_j's.

The connection between $\hat{\mu}$, the minimizer of $\sum [Y_j - \mu(t_j)]^2 + \lambda \int (L\mu)^2$, and a Bayes estimate of μ was first given by Kimeldorf and Wahba [15] for the case that $\mathrm{L}\mu = \mu^{(m)}$. The result was generalized to L's as in (1.2) by Kohn and Ansley [16]. The function μ is defined on $\Re$ and is generated by the stochastic differential equation $\mathrm{L}\mu(t)\, dt = \sigma\sqrt{\lambda}\, dW(t)$ where W is a mean zero Wiener process on $[a,b]$ with $\mathrm{var}(W(t)) = t$. Assume that μ satisfies the initial conditions: $\mu(a), \mu'(a), \ldots, \mu^{(m-1)}(a)$ are independent normal random variables with zero means and variances equal to k. Let $\hat{\mu}_k(t)$ be the posterior mean of $\mu(t)$ given $Y_1, \ldots, Y_n$. Then Kimeldorf and Wahba [15] and Kohn and Ansley [16] show that $\hat{\mu}(t) = \lim_{k \to \infty} \hat{\mu}_k(t)$.

Another Bayes connection arises in Gaussian process regression, a tool of machine learning (see, for instance, Rasmussen and Williams [24]). Consider μ defined on $A \subseteq \Re^p$, with μ the realization of a mean zero stochastic process with covariance function S. Let $\hat{\mu}_B$ be the pointwise Bayes estimate of μ:

$$\hat{\mu}_B(t) = \mathrm{E}(\mu(t)|Y_1, \ldots, Y_n) = S(t, \mathbf{t}) \ \left[\sigma^2 \mathrm{I} + S(\mathbf{t}, \mathbf{t})\right]^{-1} \mathbf{Y}$$

where $S(t, \mathbf{t})'$ is an n-vector with jth entry $S(t, t_j)$, $S(\mathbf{t}, \mathbf{t})$ is the $n \times n$ matrix with jkth entry $S(t_j, t_k)$ and $\mathbf{Y} = (Y_1, \ldots, Y_n)'$. Then, as shown below, for an appropriately defined Reproducing Kernel Hilbert Space $\mathcal{H}_S$ with reproducing kernel S, the Bayes estimate of μ is equal to

$$\arg \min_{\mu \in \mathcal{H}_S} \quad \sum_{j=1}^{n} [Y_j - \mu(t_j)]^2 + \sigma^2 < \mu, \mu > . \tag{4.1}$$

The existence of the space $\mathcal{H}_{\mathcal{S}}$ with reproducing kernel S is given by the Moore-Aronszajn Theorem (Aronszajn [4]). The space is defined by constructing finite-dimensional spaces: fix $J > 0$ and $t_1, \ldots, t_J \in A$ and consider the finite dimensional linear space of functions, $\mathcal{H}_{\{t_1,\ldots,t_J\}}$, consisting of all linear combinations of $S(t_1, \cdot), S(t_2, \cdot), \ldots, S(t_J, \cdot)$. Let $\mathcal{H}^*$ be the union of these $\mathcal{H}_{\{t_1,\ldots,t_J\}}$'s over all J and all values of $t_1, \ldots, t_J$. Let $<, >$ be the inner product on $\mathcal{H}^*$ generated by $< S(t_j, \cdot), S(t_k, \cdot) >= S(t_j, t_k)$, that is, $< \sum_j a_j S(t_j, \cdot), \sum_k b_k S(x_k, \cdot) >= \sum_{j,k} a_j b_k S(t_j, x_k)$. Let $\mathcal{H}_S$ be the completion of $\mathcal{H}^*$ under the norm associated with this inner product. Then $\mathcal{H}_S$ is a Reproducing Kernel Hilbert Space with reproducing kernel S. So, by Theorem 3.1, the solution to (4.1) is of the form $\mu(t) = \sum_{l=1}^{n} \beta_l S(t_l, t) = S(t, \mathbf{t})\boldsymbol{\beta}$, with the β_j's chosen to minimize

$$\sum_{j=1}^{n} \left[Y_j - \sum_{l=1}^{n} \beta_l S(t_l, t_j) \right]^2 + \sigma^2 \sum_{l,k=1}^{n} \beta_l \beta_k S(t_l, t_k) = ||\mathbf{Y} - S(\mathbf{t}, \mathbf{t})\boldsymbol{\beta}||^{\mathbf{2}} + \boldsymbol{\sigma}^{\mathbf{2}} \boldsymbol{\beta}' \mathbf{S}(\mathbf{t}, \mathbf{t}) \boldsymbol{\beta}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_n)'$. The minimizing $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = \left[\sigma^2 \mathrm{I} + S(\mathbf{t}, \mathbf{t})\right]^{-1} \mathbf{Y}$, and so the solution to (4.1) is equal to $\hat{\mu}_B$.

5. Results for the cubic smoothing spline

Here, we minimize (1.3) using Theorem 3.1. The expressions for the reproducing kernels R_0 and R_1 are provided. The next section contains an algorithm for computing R_0 and R_1 for general L.

The first step to minimize (1.3) over $\mu \in \mathcal{H}^2[a, b]$ is to define the inner product on $\mathcal{H}^2[a, b]$:

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

Verifying that this is an inner product is straightforward, including showing that $< f, f >= 0$ if and only if $f \equiv 0$. The proof that $\mathcal{H}^2[a, b]$ is complete under this inner product uses the completeness of $\mathcal{L}^2[a, b]$.

For (C.1) and (C.2) of Section 3, we partition $\mathcal{H}^2[a, b]$ into $\mathcal{H}_0$ and $\mathcal{H}_1$:

$$\mathcal{H}_0 = \{f : f''(t) \equiv 0\} = \text{ the span of } \{1, t\}$$

and

$$\mathcal{H}_1 = \{f \in \mathcal{H}^2[a, b] : f(a) = f'(a) = 0\}.$$

$\mathcal{H}_1$ is the orthogonal complement of $\mathcal{H}_0$ and so $\mathcal{H}^2[a,b] = \mathcal{H}_0 \oplus \mathcal{H}_1$. (This is shown in Theorem 6.1 for $\mathcal{H}^m[a,b]$.)

For (C.3) let

$$R_0(s,t) = 1 + (s-a)(t-a)$$

and

$$R_1(s,t) = st\left(\min\{s,t\} - a\right) + \frac{s+t}{2}\left[(\min\{s,t\})^2 - a^2\right] + \frac{1}{3}\left[(\min\{s,t\})^3 - a^3\right].$$

Then direct calculations verify that R_0 and R_1 are the reproducing kernels of, respectively, $\mathcal{H}_0$ and $\mathcal{H}_1$, that is, that $R_i \in \mathcal{H}_i$ and that $< R_i(\cdot,t), f >= f(t)$ for all $f \in \mathcal{H}_i$, $i = 0, 1$.

To verify that condition (C.4) is satisfied, write $\mu = \mu_0 + \mu_1$, with $\mu_i \in \mathcal{H}_i$, $i = 0, 1$. Then $P(\mu) = \int (\mu'')^2 = \int (\mu_1'')^2 =< \mu_1, \mu_1 >$.

We can show that $F_j(\mu) = \mu(t_j)$ is a continuous linear functional, either by using the definition of the inner product to verify continuity of F_j or by noting that $R = R_0 + R_1$ is the reproducing kernel of $H^2[a,b]$. Thus, by Theorem 3.1, to minimize (1.3) we can restrict attention to

$$\mu(t) = \alpha_0 + \alpha_1 t + \sum_1^n \beta_j R_1(t_j, t)$$

and find $\hat{\alpha}_0$, $\hat{\alpha}_1$ and $\hat{\boldsymbol{\beta}} \equiv (\hat{\beta}_1, \ldots, \hat{\beta}_n)'$ to minimize

$$\sum_j [Y_j - \alpha_0 - \alpha_1 t_j - \sum_k \beta_k R_1(t_j, t_k)]^2 + \beta' K \beta$$

where $K[j,k] = R_1(t_j, t_k)$. In matrix/vector form, we seek $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\alpha}} = (\hat{\alpha}_0, \hat{\alpha}_1)'$ to minimize

$$||\mathbf{Y} - T\boldsymbol{\alpha} - K\boldsymbol{\beta}||^2 + \lambda \boldsymbol{\beta}' K \boldsymbol{\beta} \tag{5.1}$$

with $\mathbf{Y} = (Y_1, \ldots, Y_n)', T_{i1} = 1$ and $T_{i2} = t_i$, $i = 1, \ldots, n$. One can minimize (5.1) directly, using matrix calculus.

Unfortunately, solving the matrix equations resulting from the differentiation of (5.1) involves inverting matrices which are ill-conditioned and large. Thus, the calculations are subject to round-off errors that seriously affect 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 $R_1(t_j, \cdot)$ 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 stable basis for this problem is that made up of cubic B-splines (see, e.g., Eubank [8]). The ith B-spline basis function has support $[t_i; t_{i+2}]$ and thus the matrices involved in the minimization of (1.3) are banded, well-conditioned, and fast to invert. Another approach is that of Reinsch ([25, 26]). 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 [2]. Section 6.4 gives this technique for minimization of expressions like (5.1).

6. Results for penalties with differential operators

Now consider the problem of minimizing (1.1) with penalty P based on a differential operator L, as in (1.2), that is, of minimizing

$$\mathcal{G}(t_1,\ldots,t_n,Y_1,\ldots,Y_n,F_1(\mu),\ldots,F_n(\mu))+\lambda\int(\mathrm{L}\mu)^2. \tag{6.1}$$

We minimize over $\mu\in\mathcal{H}^m[a,b]$ where

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

Note that, for all $\mu\in\mathcal{H}^m[a,b]$, $\int_a^b[(\mathrm{L}\mu)(t)]^2\,dt$ is well defined: $\mathrm{L}\mu(t)$ exists almost everywhere t and $\mathrm{L}\mu$ is square integrable, since the ω_j's are continuous and $[a,b]$ is finite.

We can apply Theorem 3.1 using the Reproducing Kernel Hilbert Space structure for $\mathcal{H}^m[a,b]$ defined in Section 6.1 below. We can then explicitly calculate the form of $\hat{\mu}$ provided we can calculate reproducing kernels. Theorem 6.1 states a method for explicitly calculating reproducing kernels. Section 6.2 summarizes the algorithm for calculating reproducing kernels and the form of the minimizing μ, and contains three examples of calculations. Theorem 6.1 and the calculations of Section 6.2 require results from the theory of differential equations. The Appendix contains these results, including a constructive proof of the existence of $G(\cdot,\cdot)$, the Green's function associated with the differential operator L. Section 6.4 contains a fast algorithm for minimizing (6.1) when $\mathcal{G}$ is a sum of squares and $F_j(f)=f(t_j)$.

6.1. The form of the minimizer of (6.1)

Giving the form of the minimizing μ uses the result of Theorem A.1 in the Appendix, that there exist linearly independent $u_1,\ldots,u_m\in\mathcal{H}^m[a,b]$ with m derivatives and that these functions form a basis for the set of all μ with $\mathrm{L}\mu(t)=0$ almost everywhere t. Furthermore $W(t)$, the Wronskian matrix associated with $u_1,\ldots,u_m$, is invertible for all $t\in[a,b]$. The Wronskian matrix is defined as

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

The following is an inner product under which $\mathcal{H}^m[a,b]$ is a Reproducing Kernel Hilbert Space:

$$<f,g>=\sum_{j=0}^{m-1}f^{(j)}(a)g^{(j)}(a)+\int_a^b(\mathrm{L}f)(t)\;(\mathrm{L}g)(t)\;dt. \tag{6.2}$$

To show that this is, indeed, an inner product is straightforward, except to show that $<f,f>=0$ implies that $f\equiv 0$. But this follows immediately from Theorem A.4 in the Appendix.

Theorem 6.1. *Let L be as in (1.2), let $\{u_1, \ldots, u_m\}$ be a basis for the set of μ with $L\mu \equiv 0$ and let $W(t)$ be the associated Wronskian matrix. Then, under the inner product (6.2), $\mathcal{H}^m[a,b]$ is a Reproducing Kernel Hilbert Space with reproducing kernel $R(s,t) = R_0(s,t) + R_1(s,t)$ where*

$$R_0(s,t) = \sum_{i,j=1}^{m} C_{ij} u_i(s) u_j(t)$$

with

$$C_{ij} = \left[(W(a)W'(a))^{-1}\right]_{ij},$$

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

and $G(\cdot,\cdot)$ is the Green's function associated with L, as given in equations (A.1), (A.2) and (A.3) in the Appendix. Furthermore, $\mathcal{H}^m[a,b]$ can be partitioned into the direct sum of the two subspaces

$$\begin{aligned} \mathcal{H}_0 &= \textit{the set of all } f \in \mathcal{H}^m[a,b] \textit{ with } \mathrm{L}f(t) = 0 \textit{ almost everywhere } t \\ &= \textit{the span of } u_1, \ldots, u_m \end{aligned}$$

and

$$\mathcal{H}_1 = \textit{ the set of all } f \in \mathcal{H}^m[a,b] \textit{ with } f^{(j)}(a) = 0, j = 0, \ldots m-1.$$

$\mathcal{H}_1$ is the orthogonal complement of $\mathcal{H}_0$. $\mathcal{H}_0$ has reproducing kernel R_0 and $\mathcal{H}_1$ has reproducing kernel R_1.

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

(a) Any f in $\mathcal{H}^m[a,b]$ can be written as $f = f_0 + f_1$, with $f_i \in \mathcal{H}_i$ and $< f_0, f_1 >= 0$.

(b) R_0 is the reproducing kernel for $\mathcal{H}_0$ and R_1 is the reproducing kernel for $\mathcal{H}_1$.

Consider (a). Obviously, for $f_i \in \mathcal{H}_i$, $i = 0, 1$, $< f_0, f_1 >$ is equal to zero, by the definition of the inner product in (6.2). To complete the proof of (a), fix $f \in \mathcal{H}^m[a,b]$ and find $c_1, \ldots, c_m$ such that, if $f_0 = \sum c_i u_i$, then $f_1 = f - f_0 \in \mathcal{H}_1$. That is, we find $c_1, \ldots, c_m$ such that, for $j = 0, \ldots, m-1$, $f_1^{(j)}(a) = 0$, 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

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

and we can solve this for $(c_1, \ldots, c_m)$, since the Wronskian $W(a)$ is invertible.

Consider (b). To prove that R_1 is the reproducing kernel for $\mathcal{H}_1$, first simplify notation, fixing $t \in [a,b]$ and letting $r(s) = R_1(s,t)$. We must show that $r \in \mathcal{H}_1$ and that that $< r, f >= f(t)$ for all $f \in \mathcal{H}_1$. Again, to simplify notation, let

$h(u) = G(t, u)$. By definition of R_1, $r(s) = \int_a^b G(s, u)\ h(u)\ du$. By Theorems A.5 and A.6, $r \in \mathcal{H}_1$ and $\mathrm{L}r(s) = h(s) = G(t, s)$ almost everywhere s. Therefore, for $f \in \mathcal{H}_1$,

$$< r, f >= 0 + \int_a^b (\mathrm{L}r)(s)\ (\mathrm{L}f)(s)\ ds = \int_a^b G(t, s)\ (\mathrm{L}f)(s)\ ds = f(t)$$

by the definition of the Green's function. See equation (A.1).

Now consider R_0. Obviously, $R_0(\cdot, t) \in \mathcal{H}_0$, since it is a linear combination of the u_i's. To show that $< R_0(\cdot, t), f >= f(t)$, it suffices to consider $f = u_l, l = 1, \ldots, m$. Noting that $\mathrm{L}u_l \equiv 0$, write

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

□

We can now use Theorems 3.1 and 6.1 to write the form of the minimizer of (6.1). The proof of the following Theorem is straightforward.

Theorem 6.2. *Suppose that L is as in (1.2). Let $u_1, \ldots, u_m$ be a basis for the set of μ's with $L\mu \equiv 0$ and let G be the corresponding Green's function, defined in equations (A.1), (A.2) and (A.3) in the Appendix. Let*

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

and $\eta_{j1}(t) = F_j(R_1(\cdot, t))$. Then the minimizer of (6.1) must be of the form

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

where the α_j's and $\boldsymbol{\beta} \equiv (\beta_1, \ldots, \beta_n)'$ minimize

$$\mathcal{G}(t_1, \ldots, t_n, Y_1, \ldots, Y_n, F_1(\mu), \ldots, F_n(\mu)) + \lambda \boldsymbol{\beta}' K \boldsymbol{\beta}$$

with K as defined in Theorem 3.1.

6.2. Algorithm and examples for calculating R_0, R_1 and the minimizing μ

Suppose that we're given a linear differential operator L as in (1.2). The following steps summarize results so far, describing how to calculate R_0 and R_1, the required reproducing kernels associated with L, and the $\hat{\mu}$ that minimizes (6.1).

1. Find $u_1, \ldots, u_m$, a basis for the set of functions μ with $\mathrm{L}\mu \equiv 0$.
2. Calculate $W(\cdot)$, the Wronskian of the u_i's: $W_{ij}(t) = u_i^{(j-1)}(t)$.
3. Set $R_0(s,t) = \sum_{i,j} [[W(a)W'(a)]^{-1}]_{ij} u_i(s) u_j(t)$.
4. Calculate $(u_1^*(t), \ldots, u_m^*(t))$, the last row of the inverse of $W(t)$.
5. Find G, the associated Green's function: $G(t,u) = \sum u_i(t) u_i^*(u)$ for $u \leq t$, 0 else.
6. Set $R_1(s,t) = \int_a^b G(s,u)\, G(t,u)\, du$.
7. Find η_{1j}: $\eta_{1j}(t) = F_j(R_1(\cdot,t))$.
8. Calculate the symmetric matrix K: $K[j,k] = F_k(\eta_{1j})$. If $F_j(\mu) = \mu(t_j)$ and $F_k(\mu) = \mu(t_k)$ then $K[j,k] = R_1(t_j, t_k)$.
9. Set $\mu(t) = \sum \alpha_j u_j(t) + \sum_j \beta_j \eta_{1j}(t)$ and minimize $\mathcal{G}(t_1, \ldots, t_n, Y_1, \ldots, Y_n, F_1(\mu), \ldots, F_n(\mu)) + \lambda \boldsymbol{\beta}' K \boldsymbol{\beta}$ with respect to $\boldsymbol{\beta}$ and the α_j's.

The first step is the most challenging, and for some L's, it may in fact be impossible to find the u_j's in closed form. However, if L is a linear differential operator with constant coefficients, then the first step is easy, using Theorem A.2. Alternatively, if one has an approximate model in mind defined in terms of known functions $u_1, \ldots, u_m$, then one can find the corresponding L (see Example 3 below).

The reader can use these steps to derive the expressions in Section 5 for the cubic smoothing spline.

Although the calculation of the minimizing μ does not involve R_0, step 3 is included for completeness, to allow the reader to calculate the reproducing kernel, $R_0 + R_1$, for $\mathcal{H}^m[a,b]$ under the inner product (6.2).

Example 1. Suppose that $\mathrm{L}\mu = \mu'$ and that the interval $[a,b]$ is equal to $[0,1]$. In Step 1, the basis for $\mathrm{L}\mu \equiv 0$ is $u_1(t) = 1$. In Step 2, the Wronskian is the one by one matrix with element equal to 1. So in Step 3, $R_0(s,t) \equiv 1$. In Step 4, $u_1^*(s) = 1$ and so, in Step 5, $G(t,u) = 1$ if $u \leq t$, 0 else. Therefore

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

Thus, we seek μ of the form

$$\mu(t) = \alpha + \sum_{j=1}^n \beta_j F_j(R_1(\cdot,t)).$$

If $F_j(\mu) = \mu(t_j)$, $j = 1, \ldots, n$, then we seek

$$\mu(t) = \alpha + \sum_{j=1}^n \beta_j \min\{t_j, t\},$$

that is, the minimizing μ is piecewise linear with pieces defined in terms of $t_1, \ldots, t_n$. In Step 8, $K[j,k] = \min\{t_j, t_k\}$.

If, instead, $F_j(\mu) = \int_0^1 f_j \mu$ for known f_j, as in Section 2.4, then

$$\begin{aligned} F_j(R_1(\cdot,t)) &= \eta_{1j}(t) = \int_0^1 f_j(s)\ R_1(s,t)\ ds = \int_0^1 f_j(s)\ \min\{s,t\}\ ds \\ &= \int_0^t s\ f_j(s)\ ds + t\int_t^1 f_j(s)\ ds \end{aligned}$$

and, in Step 8,

$$K[j,k] = \int_{t=0}^1 f_k(t)\ \eta_{1j}(t)\ dt = \int_{s,t=0}^1 f_k(t)\ f_j(s)\ \min\{s,t\}\ ds\ dt.$$

Example 2. Suppose that $\mathrm{L}f = f'' + \gamma f'$, γ a real number.

For Step 1, we can find u_1 and u_2 via Theorem A.2 in the Appendix. We first solve $x^2 + \gamma x = 0$ for the two roots, $r_1 = 0$ and $r_2 = -\gamma$. So

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

For Step 2, we compute the Wronskian

$$W(t) = \left[\begin{array}{cc} 1 & 0 \\ \exp(-\gamma t) & -\gamma\exp(-\gamma t) \end{array}\right].$$

For Step 3 we have

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

So

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

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

For Step 4, inverting $W(t)$ we find that

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

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

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

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

$$\begin{aligned} R_1(s,t) &= \int_a^s \gamma^{-2}(1-e^{-\gamma(s-u)})\,(1-e^{-\gamma(t-u)})\,du \\ &= -\frac{1}{\gamma^3}+\frac{s^*}{\gamma^2}+\frac{1}{\gamma^3}\exp(-\gamma s^*)+\frac{1}{\gamma^3}\exp(-\gamma t^*) \\ &\quad -\frac{1}{2\gamma^3}\exp[-\gamma(t^*-s^*)]-\frac{1}{2\gamma^3}\exp[-\gamma(s^*+t^*)]. \qquad (6.3) \end{aligned}$$

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.

Therefore, to minimize (6.1) over $\mu \in \mathcal{H}^4[a,b]$, we seek μ of the form

$$\mu(t) = \alpha_1 + \alpha_2 \exp(-\gamma t) + \sum_1^n \beta_j F_j(R_1(\cdot,t)).$$

The calculations in Steps 7 and 8 for $\eta_{j1}(t) = F_j(R_1(\cdot,t))$ and K are tedious except in the case that $F_j(f) = f(t_j)$.

Example 3. Instead of specifying the operator L, one might more easily specify basis functions $u_1,\ldots,u_m$ for a preferred approximate parametric model. For instance, one might think that μ is approximately a constant plus a damped sinusoid: $\mu(t) \approx \alpha_1 + \alpha_2 \sin(t)\exp(-t)$. Given $u_1,\ldots,u_m$, one can easily find the operator L so that $\mathrm{L}u_i \equiv 0$, $i = 1,\ldots,m$, and thus one can define an estimate of μ as the minimizer of (6.1). Assume that each u_i has m continuous derivatives and that the associated Wronskian matrix $W(t)$ is invertible for all $t \in [a,b]$. To find L, we solve for the ω_j's in (1.2):

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

that is

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

This can be written in matrix/vector form as

$$W(t)\begin{bmatrix}\omega_0(t)\\ \vdots \\ \omega_{m-1}(t)\end{bmatrix} = -\begin{bmatrix}u_1^{(m)}(t)\\ \vdots \\ u_m^{(m)}(t)\end{bmatrix}$$

yielding

$$\begin{bmatrix}\omega_0(t)\\ \vdots \\ \omega_{m-1}(t)\end{bmatrix} = -W(t)^{-1}\begin{bmatrix}u_1^{(m)}(t)\\ \vdots \\ u_m^{(m)}(t)\end{bmatrix}.$$

Obviously, the ω_j's are continuous, by our assumptions concerning the u_i's and the invertibility of $W(t)$.

For the example with $u_1 \equiv 1$ and $u_2 = \sin(t)\exp(-t)$, we find that

$$W(t) = \begin{bmatrix} 1 & 0 \\ \sin(t)\exp(-t) & \exp(-t)[\cos(t) - \sin(t)] \end{bmatrix},$$

which is invertible on $[a, b]$ provided $\cos(t) \neq \sin(t)$ for $t \in [a, b]$. In this case, $\omega_0(t) \equiv 0$, $\omega_1(t) = 2\cos(t)/[\cos(t) - \sin(t)]$ and so the associated differential operator is $\mathrm{L}(\mu)(t) = \mu''(t) + 2\mu'(t)\cos(t)/[\cos(t) - \sin(t)]$. Note that we do not need L to proceed with the minimization of (6.1) – we only need $u_1, \ldots, u_m$ to calculate the required reproducing kernels. However, if we would like to cast the problem in the Bayesian model of Section 4, we require L.

6.3. Minimization of the penalized weighted sum of squares via matrix calculus

Consider minimizing a specific form of (6.1) over $\mu \in \mathcal{H}^m[a, b]$, namely minimizing

$$\sum_j d_j[Y_j - F_j(\mu)]^2 + \lambda \int (\mathrm{L}u)^2 \tag{6.4}$$

for known and positive d_j's. We can rewrite this as a minimization problem easily solved by matrix/vector calculations, provided we can find a basis $\{u_1, \ldots, u_m\}$ for the set of μ with $\mathrm{L}\mu = 0$.

Theorem 6.2 implies that, to minimize (6.4), we must find $\hat{\boldsymbol{\alpha}} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_m)'$ and $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_n)'$ to minimize

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

where $\mathbf{Y} = (Y_1, \ldots, Y_n)'$, T is $n \times m$ with $T[i, j] = u_j(t_i)$, K is $n \times n$ with $K[j, k] = F_j(\eta_{k1})$, and D is an n by n diagonal matrix with $D[i, i] = d_i$. Assume, as is typically the case, that T is of full rank and K is invertible. Taking derivatives with respect to $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ and setting equal to zero yields

$$T'D(\mathbf{Y} - K\hat{\boldsymbol{\beta}}) = T'DT\hat{\boldsymbol{\alpha}}. \tag{6.5}$$

and

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

which is equivalent to

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

Let

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

Then

$$\hat{\boldsymbol{\beta}} = M^{-1}(\mathbf{Y} - T\hat{\boldsymbol{\alpha}}). \tag{6.6}$$

Substituting this into (6.5) yields

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

that is

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

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

Therefore, provided T is of full rank,

$$\hat{\boldsymbol{\alpha}} = (T'M^{-1}T)^{-1}T'M^{-1}\mathbf{Y} \tag{6.7}$$

and

$$\hat{\boldsymbol{\beta}} = M^{-1}[\mathrm{I} - T(T'M^{-1}T)^{-1}T'M^{-1}]\mathbf{Y}. \tag{6.8}$$

Unfortunately, using equations (6.7) and (6.8) results in computational problems since typically M is an ill-conditioned matrix and thus difficult to invert. Furthermore, M is $n \times n$ and n is typically large, making inversion expensive. Fortunately, when $F_j(f) = f(t_j)$ we can transform the problem to alleviate the difficulties and to speed computation. The details are given in the next section.

6.4. Algorithm for minimizing the penalized weighted sum of squares when $F_j(f) = f(t_j)$

Assume that $F_j(f) = f(t_j)$, that $a < t_1 < \cdots < t_n < b$, that T is of full rank $n - m$ and that K is invertible. The goal is to re-write $\hat{\boldsymbol{\alpha}}$ in (6.7) and $\hat{\boldsymbol{\beta}}$ in (6.8) so that we only need to invert small or banded matrices. Meeting this goal involves defining a "good" matrix Q and showing that

$$\hat{\boldsymbol{\beta}} = Q(Q'MQ)^{-1}Q'\mathbf{Y} \tag{6.9}$$

and

$$\hat{\boldsymbol{\alpha}} = (T'T)^{-1}T'(\mathbf{Y} - M\hat{\beta}). \tag{6.10}$$

We will define Q so that $Q'MQ$ is banded and thus easy to invert. To begin, let Q be an n by $n-m$ matrix of full column rank such that $Q'T$ is an $n-m$ by m matrix of zeroes. Q isn't unique, but later, further restrictions will be placed on Q so that $Q'MQ$ is banded.

We first show that $T'\hat{\boldsymbol{\beta}} = 0$. This will imply that there exists an $n-m$ vector $\boldsymbol{\gamma}$ such that $\hat{\boldsymbol{\beta}} = Q\boldsymbol{\gamma}$. From (6.6)

$$\mathbf{Y} = M\hat{\boldsymbol{\beta}} + T\hat{\boldsymbol{\alpha}} \tag{6.11}$$

Substituting this into (6.7) yields

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

Therefore

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

and so $T'\hat{\boldsymbol{\beta}} = 0$ and $\hat{\boldsymbol{\beta}} = Q\boldsymbol{\gamma}$ for some $\boldsymbol{\gamma}$. To find $\boldsymbol{\gamma}$, use (6.6):

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

since $Q'T = 0$. So $Q'MQ\boldsymbol{\gamma} = Q'\mathbf{Y}$, yielding

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

Therefore equation (6.9) holds. Equation (6.10) follows immediately from equation (6.11).

We can also find an easy-to-compute form for $\hat{\mathbf{Y}} \equiv T\hat{\boldsymbol{\alpha}} + K\hat{\boldsymbol{\beta}}$ using (6.11):

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

and so

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

Note that we have not yet used the fact that $F_j(f) = f(t_j)$. In the special case that $F_j(f) = f(t_j)$, we can choose Q so that $Q'MQ$ is banded. Specifically, in addition to requiring that $Q'T = 0$, we also seek Q with

$$Q_{ij} = 0 \text{ unless } i = j, j+1, \ldots, j+m. \tag{6.12}$$

So we want Q with $[Q'T]_{ij} = \sum_{l=0}^{m} Q_{i+l,i}u_j(t_{i+l}) = 0$ for all $j = 1, \ldots, m, i = 1, \ldots, n-m$. That is, for each i, we seek an $(m+1)$-vector $\mathbf{q}_i \equiv (Q_{ii}, \ldots, Q_{i+m,i})'$ satisfying $\mathbf{q}_i'T_i = 0$, where T_i is the $(m+1)$ by m matrix with ljth entry equal to $u_j(t_{i+l})$. This is easily done by a QR decomposition of T_i: the matrix T_i can be written as $T_i = Q_i\, R_i$ for some Q_i, an $(m+1) \times (m+1)$ orthonormal matrix, and some R_i, $(m+1) \times m$ with last row equal to 0. Take $\mathbf{q}_i$ to be the last column of Q_i.

We now show that $Q'MQ$ is banded, specifically, that $[Q'MQ]_{kl} = 0$ whenever $|k-l| > m$. Write $Q'MQ = Q'KQ + \lambda Q'D^{-1}Q$. Since D is diagonal, one easily shows that $[QD^{-1}Q]_{kl} = 0$ for $|k-l| > m$. To show that the same is true for $Q'KQ$, write

$$\begin{aligned}
K[i,j] &= R_1(t_i, t_j) \\
&= \int G(t_i, \omega)\; G(t_j, \omega)\; d\omega \\
&= \sum_{r,s} u_r(t_i)u_s(t_j) \int_a^{\min\{t_i,t_j\}} u_r^*(\omega)\; u_s^*(\omega)\; d\omega \\
&\equiv \sum_{r,s} u_r(t_i)u_s(t_j)\; \mathcal{F}_{r,s}(\min\{t_i,t_j\}). \\
&= \sum_{r,s} T_{ir}T_{js}\; \mathcal{F}_{r,s}(\min\{t_i,t_j\}).
\end{aligned}$$

Since $Q'KQ$ is symmetric, it suffices to show that $[Q'KQ]_{kl} = 0$ for $k - l > m$. So fix k and l with $k - l > m$ and write

$$\begin{aligned}
[Q'KQ]_{kl} &= \sum_{i,j=1}^{n} Q_{ik}K_{ij}Q_{jl} = \sum_{i,j=0}^{m} Q_{k+i,k}K_{k+i,l+j}Q_{l+j,l} \\
&= \sum_{i,j=0}^{m}\sum_{r,s=1}^{m} Q_{k+i,k}\ \mathcal{F}_{r,s}(\ \min\{t_{k+i}, t_{l+j}\})\ T_{k+i,r}T_{l+j,s}Q_{l+j,l} \\
&= \sum_{j=0}^{m}\sum_{r,s=1}^{m} \mathcal{F}_{r,s}(t_{l+j})\ T_{l+j,s}Q_{l+j,l}\sum_{i=0}^{m} Q_{k+i,k}T_{k+i,r}.
\end{aligned}$$

The last equality follows since $k > l + m$ and $0 \leq i, j \leq m$ imply that $k + i > l + j$ and so $t_{l+j} < t_{k+i}$. We immediately have that $[Q'KQ]_{kl} = 0$, since $\sum_{i=0}^{m} Q_{k+i,k}T_{k+i,r} = [Q'T]_{kr} = 0$.

Thus minimizing (6.4) when $F_j(f) = f(t_j)$ is easily and quickly done through the following steps.

1. Follow steps 1 through 8 of Section 6.2 to find $u_1, \ldots, u_m$, a basis for $\mathrm{L}\mu = 0$, the reproducing kernel R_1 and the matrix K: $K[i,j] = R_1(t_i, t_j)$.
2. Calculate the matrix T: $T[i,j] = u_j(t_i)$.
3. Find Q n by $(n - m)$ of full column rank satisfying equation (6.12) and $Q'T = 0$. One can find Q directly or by the method outlined below equation (6.12).
4. Find $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\alpha}}$ using equations (6.9) and (6.10). Speed the matrix inversion by using the fact that $Q'MQ$ is banded.

Example 2 continued from Section 6.2. Suppose that we want to minimize

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

over $\mu \in \mathcal{H}^2[0,1]$. For simplicity, assume that $t_i = i/(n+1)$. Using the calculations from Section 6.2, we set $T_{i1} = 1, T_{i2} = \exp(-\gamma t_i)$, and $K[i,j] = R_1(t_i, t_j)$, with R_1 as in (6.3).

For Step 3, we find Q directly: we seek Q n by $(n-2)$ with $Q_{ij} = 0$ unless $i = j, j+1, j+2$ and

$$0 = [Q'T]_{ij} = Q_{ii}T_{ij} + Q_{i,i+1}T_{i+1,j} + Q_{i,i+2}T_{i+2,j}.$$

Thus, for $j = 1$,

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

and, for $j = 2$,

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

We 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 the fourth step to find $\hat{\alpha}$ and $\hat{\beta}$ is straightforward.

Appendix A

The Appendix contains background on the solution of linear differential equations $\mathrm{L}\mu = 0$ with L as in (1.2). Section A.2 contains results about G, the Green's function associated with L.

A.1. Differential equations

Details of results in this section can be found in Coddington [6]. The main Theorem, stated without proof, follows.

Theorem A.1. *Let L be as in (1.2). Then there exists $u_1, \ldots, u_m$ a basis for the the set of all μ with $L\mu \equiv 0$, with each u_i real-valued and having m derivatives. Furthermore, any such basis will have an invertible Wronskian matrix $W(t)$ for all $t \in [a, b]$. The Wronskian matrix is defined as*

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

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

Theorem A.2. *Suppose that L is as in (1.2), with the ω_j's real numbers. Denote the s distinct roots of the polynomial $x^m + \sum_{j=0}^{m-1} \omega_j x^j$ as $r_1, \ldots, r_s$. Let m_i denote the multiplicity of root r_i (so $m = \sum_1^s m_i$). Then the following m functions of t form a basis for the set of all μ with $L\mu \equiv 0$:*

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

The following result, stated without proof, is useful for checking that a set of functions does form a basis for the set of all μ with $\mathrm{L}\mu \equiv 0$.

Theorem A.3. *Suppose that $u_1, \ldots, u_m$ have m derivatives on $[a, b]$ and that $Lu_i \equiv 0$. If $W(t_0)$ is invertible at some $t_0 \in [a, b]$, then the u_i's are linearly independent, and thus a basis for the set of all μ with $L\mu \equiv 0$.*

The following result was useful in defining the inner product in equation (6.2), where t_0 was taken to be a.

Theorem A.4. *Suppose that L is as in (1.2) and let $t_0 \in [a, b]$. Then the only function in $\mathcal{H}^m[a, b]$ that satisfies $Lf =$ the zero function and $f^{(j)}(t_0) = 0, j = 0, \dots, m - 1$, is the zero function.*

Proof. By Theorem A.1, there exists $u_1, \dots, u_m$ a basis for the set of all μ with $\mathrm{L}\mu \equiv 0$, with $W(t)$ 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 $f^{(j)}(t_0) = 0,\ j = 0, \dots, m - 1$ can be written in matrix/vector form as $(c_1, \dots, c_m)W(t_0) = (0, \dots, 0)$. Since $W(t_0)$ is invertible, $c_i = 0, i = 1, \dots, m$. □

A.2. The Green's function associated with the differential operator L

Suppose that L is as in (1.2). The definition below gives the definition of $G(\cdot, \cdot)$, the Green's function associated with L with specified boundary conditions. Theorem A.5 gives an explicit form of G.

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

$$f(t) = \int_{u=a}^{b} G(t, u)\ (\mathrm{L}f)(u)\ du \tag{A.1}$$

for all functions f in $\mathcal{H}^m[a, b]$ satisfying the boundary conditions

$$f^{(j)}(a) = 0, j = 0, \dots, m - 1. \tag{A.2}$$

Of course, it's not immediately clear that such a function G exists. However, G exists and is easily calculated using the Wronskian matrix associated with L (see Theorem A.5). Recall from Theorem A.1 of Section A.1 that there exists a basis for the set of all μ with $\mathrm{L}\mu \equiv 0$, $u_1, \dots, u_m$, with invertible Wronskian. Furthermore, each u_i has m derivatives.

Lemma A.1. *Let $u_1^*(t), \dots, u_m^*(t)$ denote the entries in the last row of the inverse of $W(t)$. Then u_j^* is continuous, $j = 1, \dots, m$.*

Proof. The u_i^*'s are continuous, since $u_i^* = (\det W(t))^{-1}$ times an expression involving sums and products of $u_l^{(j)}, l = 1, \dots, m, j = 0, \dots, m - 1$, and the u_l's have $m - 1$ continous derivatives. □

Theorem A.5. *Let $u_1^*(t), \dots, u_m^*(t)$ denote the entries in the last row of the inverse of $W(t)$. Then*

$$G(t, u) \quad = \quad \begin{cases} \sum_{i=1}^m u_i(t) u_i^*(u) & \text{for } u \le t \\ 0 & \text{otherwise} \end{cases} \tag{A.3}$$

is a Green's function for L and, for each fixed $t \in [a, b]$, $G(t, \cdot)$ is in $L^2[a, b]$.

The following theorem will be useful in the proof of Theorem A.5.

Theorem A.6. *Let G be as in (A.3) and suppose that $h \in \mathcal{L}_2$. If*

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

Then

$$r \in \mathcal{H}^m[a,b], \tag{A.4}$$

$$(\mathrm{L}r)(t) = h(t) \quad \textit{almost everywhere } t \in [a,b] \tag{A.5}$$

and

$$r^{(j)}(a) = 0 \quad j = 0, \ldots, m-1. \tag{A.6}$$

Proof. Write

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

We'll first show that

$$r^{(j)}(t) = \sum_{i=1}^m u_i^{(j)}(t) \int_a^t u_i^*(u)\ h(u)\ du \quad j = 0, \ldots, m-1 \tag{A.7}$$

and

$$r^{(m)}(t) = h(t) + \sum_{i=1}^m u_i^{(m)}(t) \int_a^t u_i^*(u)\ h(u)\ du \quad \text{almost everywhere } t \in [a,b]. \tag{A.8}$$

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

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

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 W and the u_i^*'s, this is equal to

$$h(t) \sum_i [W(t)]_{i1} [W(t)^{-1}]_{mi} = h(t)\ [W(t)^{-1} W(t)]_{m1} = h(t)\ \mathrm{I}\{m = 1\}.$$

Therefore, for $m = 1$, (A.8) holds and for $m > 1$ (A.7) 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 (A.7) and (A.8). Clearly, the mth derivative in (A.8) is square-integrable. Therefore we've proven (A.4).

To prove (A.5), use (A.7) and (A.8) and write

$$
\begin{aligned}
(Lr)(t) &= r^{(m)}(t) + \sum_{j=0}^{m-1} \omega_j(t) r^{(j)}(t) \\
&= h(t) + \sum_{i=1}^{m} u_i^{(m)}(t) \int_a^t u_i^*(u)\ h(u)\ du \\
&\quad + \sum_{j=0}^{m-1} \sum_{i=1}^{m} \omega_j(t) u_i^{(j)}(t) \int_a^t u_i^*(u)\ h(u)\ du \\
&= h(t) + \sum_{i=1}^{m} \left[u_i^{(m)}(t) + \sum_{j=0}^{m-1} \sum_{i=1}^{m} \omega_j(t) u_i^{(j)}(t) \right] \int_a^t u_i^*(u)\ h(u)\ du \\
&= h(t) + \sum_{i=1}^{m} (\mathrm{L}u_i)(t) \int_a^t u_i^*(u)\ h(u)\ du = h(t)
\end{aligned}
$$

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

Equation (A.6) follows directly from (A.7) by taking $t = a$. □

Proof of Theorem A.5. First consider the function in equation (A.3) as a function of u with t fixed. Since the u_i's are continuous and $W(u)$ is invertible for all u, $G(t, \cdot)$ is continuous on the finite closed interval $[a, b]$. Thus it is in $L^2[a, b]$.

To show that equation (A.1) holds, let $f \in \mathcal{H}^m$ satisfy the boundary conditions (A.2). Define $r(t) = \int_a^b G(t, u)\ (\mathrm{L}f)(u)\ du$. Then, by Theorem A.6, $\mathrm{L}r = \mathrm{L}f$ almost everywhere and $r^{(j)}(a) = 0, j = 0, \ldots, m-1$. Thus $\mathrm{L}(r - f) = 0$ almost everywhere and $(r - f)^{(j)}(a) = 0, j = 0, \ldots, m-1$. By Theorem A.4, $r - f$ is the zero function, that is $r = f$. □

Acknowledgements

The author's research was supported by the Natural Sciences and Engineering Research Council of Canada through grant A7969. The author thanks the Editor and two anonymous reviewers for their helpful suggestions that improved the paper and helped to highlight the paper's contributions. Special thanks are due the Editor, for his gracious and careful handling of the submission.

References

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

[2] 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. MR0249904

[3] Ansley, C., Kohn, R., and Wong, C. (1993). Nonparametric spline regression with prior information. *Biometrika* **80**, 75–88. MR1225215

[4] Aronszain, N. (1950). Theory of reproducing kernels. *Transactions of the American Mathematical Society* **68**, 337–404. MR0051437

[5] 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.

[6] Coddington, E.A. (1961). *An Introduction to Ordinary Differential Equations*. New Jersey: Prentice-Hall. MR0126573

[7] Crambes, C., Kneip, A., and Sarda, P. (2009). Smoothing splines estimators for functional linear regression. *Annals of Statistics* **37**, 35–72. MR2488344

[8] Eubank, R.L. (1999). *Spline Smoothing and Nonparametric Regression, Second Edition*. New York: Marcel Dekker. MR1680784

[9] Furrer, E.M. and Nychka, D. A framework to understand the asymptotic properties of Kriging and splines. URL: `http://www.image.ucar.edu/~nychka/man.html`

[10] Green, P.J. and Silverman, B.W. (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach.* London: Chapman and Hall. MR1270012

[11] Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning Data Mining, Inference, and Prediction, Second Edition*. Springer Series in Statistics, Springer. MR2722294

[12] Heckman, N. and Ramsay, J.O. (2000). Penalized regression with model based penalties. *Canadian Journal of Statistics* **28**, 241–258. MR1792049

[13] Hofmann, T., Schölkopf, B., and Smola, A. (2008). Kernel methods in machine learning. *Annals of Statistics* **36**, 1171–1220. MR2418654

[14] Horváth, L. and Kokoszka, P. (2012). *Inference for Functional Data with Applications*. Springer Series in Statistics, Volume 200. MR2920735

[15] Kimeldorf, G. and Wahba, G. (1971). Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications* **33**, 82–95. MR0290013

[16] Kohn, R. and Ansley, C.F. (1988). Equivalence between Bayesian smoothness priors and optimal smoothing for function estimation. *Bayesian Analysis of Time Series and Dynamic Models* **1**, 393–430.

[17] Kolmogorov, A.N. and Fomin, S.V. (1999). *Elements of the Theory of Functions and Functional Analysis*. Dover Publications.

[18] Kreyszig, E. (1989). *Introductory Functional Analysis with Applications.* Wiley. MR0992618

[19] Li, Xiaochun. (1996). *Local Linear Regression versus Backcalculation in Forecasting.* Ph.D. thesis, Statistics Department, University of British Columbia. MR2695370

[20] 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. MR0770276

[21] Nychka, D. (2000). Spatial Process Estimates as Smoothers. *Smoothing and Regression. Approaches, Computation and Application*, ed. Schimek, M. G., Wiley, New York. MR1795148

[22] Ramsay, J.O., Hooker, G., Campbell, D., and Cao, J. (2007). Parameter estimation for differential equations: a generalized smoothing approach. *Journal of the Royal Statistical Society, Series B* **69**, 741-796. MR2368570

[23] Ramsay, J.O. and Silverman, B.W. (2005). *Functional Data Analysis, Second Edition*. Springer. MR2168993

[24] Rasmussen, C.E. and Williams, C.K.I. (2006). *Gaussian Processes for Machine Learning*. The MIT Press. MR2514435

[25] Reinsch, C. (1967). Smoothing by spline functions. *Numerische Mathematik* **10**, 177-183. MR0295532

[26] Reinsch, C. (1970). Smoothing by spline functions II. *Numerische Mathematik* **16**, 451-454. MR1553981

[27] Thompson, J.R. and Tapia, R.A. (1990). *Nonparametric Function Estimation, Modeling, and Simulation*. Society for Industrial Mathematics.

[28] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B* **58**, 267–288. MR1379242

[29] Wahba, G. (1999). Support vector machines, Reproducing Kernel Hilbert Spaces, and randomized GCV. *Advances in Kernel Methods: Support Vector Learning.* Bernhard Schölkopf, Christopher J. C. Burges and Alexander J. Smola, Editors. MIT Press, Cambridge, MA, 69–88.

[30] Wahba, G. (1990). *Spline Models for Observational Data.* Philadelpha: Society for Industrial and Applied Mathematics. MR1045442

[31] Wahba, G. (2003). An introduction to Reproducing Kernel Hilbert Spaces and why they are so useful. *Proceedings Volume from the 13th IFAC Symposium on System Identification*, 27–29. IPV-IFAC Proceedings Volume. Paul M.J. Van Den Hof, Bo Wahlberg and Siep Weiland, Editors.

[32] Yuan, M. and Cai, T. (2010). A Reproducing Kernel Hilbert Space approach to functional linear regression. *Annals of Statistics* **38** 3412–3444. MR2766857