DEPARTMENT OF STATISTICS
University of Wisconsin
1210 West Dayton St.
Madison, WI 53706

TECHNICAL REPORT NO. 1020
April 18, 2000

An Introduction to Model Building with Reproducing Kernel Hilbert Spaces * †

Grace Wahba

*This document contains the overhead slides as given in the Short Course of the same name at Interface 2000, minor typos fixed.

†Research on which part of this course is based has been supported by NSF Grant DMS9704758, NIH Grant EY09946, and NASA Grant NAG5 3769.

An Introduction to Model Building with Reproducing Kernel Hilbert Spaces (With Applications)

Grace Wahba

Interface 2000 Short Course

New Orleans, April 5, 2000

All TR's since late 93 up in
`http://www.stat.wisc.edu/~wahba` $\rightarrow$ TRLIST
Home directory for this talk
`ftp://ftp.stat.wisc.edu/pub/wahba/talks/interface.00/`

Abstract

We assume no knowledge of reproducing kernel Hilbert spaces, but review some basic concepts, with a view towards demonstrating how this setting allows the building of interesting statistical models that allow the simultaneous analysis of heterogeneous, scattered observations, and other information. The abstract ideas will be illustrated with several specific data analyses, including modeling risk factors for eye diseases.

What you should get out of this shortcourse:

1. An understanding of what reproducing kernel Hilbert spaces are and the advantages they provide in multivariate function estimation and statistical model building.

2. Ideas for using old models or developing new ones for your particular application.

3. Where to go for software and further information.

Why should we be interested in RKHS?

1. Provide a framework for flexible function estimation and statistical model building with scattered, noisy, direct and indirect data on very general domains.

2. Models based on RKHS are the foundation for penalized likelihood estimation and regularization methods and can handle a wide variety of data distributions and problems - Gaussian, general exponential families, robust estimation, interval observations, ..

3. Constraints such as positivity, convexity, other linear inequality constraints can be incorporated in the models.

4. Can deal with noisy observations on derivatives, integrals, and other bounded linear functionals, provides a framework for merging different kinds of information - e. g. observations averaged over irregular and inconsistent areas or time intervals.

5. Can estimate model integrals and derivatives as well as function values. Can estimate meaningful projections or components of the model.

6. Methods for model tuning to optimize the bias-variance tradeoff are readily available.

7. Have a dual interpretation as Bayes estimates, $\boxed{\text{prior to bias-variance or generalization-error tuning}}$.

8. Bayesian ‘confidence intervals’ with frequentist properties are available.

9. Can incorporate dynamical systems equations and other physical models into the empirical model.

Part I

1. Positive Definite Functions
2. Bayes Estimates and Variational Problems
3. Reproducing Kernel Hilbert Spaces
4. The Moore-Aronszajn Theorem and Inner Products in RKHS
5. Example: Periodic Splines
6. The Representer Theorem (simple case)
7. Sums and Products of Positive Definite Functions

♣♣ What is a positive definite function?

This concept is key, so we begin by reviewing it.

- The N dimensional case:

Let $\mathcal{T} = 1, 2, \cdots, N$. $K(s,t), (s,t) \in \mathcal{T} \otimes \mathcal{T}$ is said to be a positive definite function on $\mathcal{T} \otimes \mathcal{T}$ (`which means it is actually an` $N \times N$ `matrix`) if, for every $a = (a_1, \cdots, a_N)$ we have that $\sum_{i,j=1}^{N} a_i a_j K(i,j) \geq 0$.

- The general case:

Let $\mathcal{T}$ be a (possibly continuous) index set, for example, the unit interval, the unit cube, the surface of the unit sphere, the real line, the plane, Euclidean D space, etc. $K(s,t), (s,t) \in \mathcal{T} \otimes \mathcal{T}$ is said to be a positive definite function on $\mathcal{T} \otimes \mathcal{T}$ if, for every n, and every $t_1, \cdots, t_n \in \mathcal{T}$, and every $a = (a_1, \cdots, a_n)$, $\sum_{i,j=1}^{n} a_i a_j K(t_i, t_j) \geq 0$.

♣♣ Bayes Estimates and Variational Problems

(Certain) Bayes estimates are solutions to variational problems, and vice versa.

- The N dimensional case: The Bayes estimate:

Let $y, f, \epsilon \in E^N$, with $f \sim \mathcal{N}(0, b\Sigma), \epsilon \sim \mathcal{N}(0, \sigma^2 I)$, f, ϵ independent, and let

$$y = f + \epsilon.$$

Here b is a fixed constant whose role will become apparent shortly. Σ is a given (strictly) positive definite matrix. We want to estimate f. Standard calculations give

$$\begin{aligned} \hat{f} = E(f|y) &= \Sigma(\Sigma + (\sigma^2/b)I)^{-1}y \\ &= A(\lambda)y, \;\; say, with \;\; \lambda = (\sigma^2/b). \end{aligned}$$

- The N dimensional case: The variational problem:

Consider the ridge regression estimate: Find f in E^N to minimize

$$\|y - f\|^2 + \lambda f' \Sigma^{-1} f.$$

The minimizer, f_λ is easily seen to satisfy

$$(I + \lambda \Sigma^{-1}) f = y,$$

or,

$$\begin{aligned} f_\lambda &= \Sigma(\Sigma + \lambda I)^{-1} y \\ &= A(\lambda) y \end{aligned}$$

MORAL: Given the prior $f \sim \mathcal{N}(0, b\Sigma)$ and $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$, the posterior mean for f given y is the ridge regression estimate for f with penalty $f'\Sigma^{-1} f$ and penalty parameter $\lambda = \sigma^2/b$. $A(\lambda)$ is known as the influence matrix and will play an important role later.

- The general case: The Bayes estimate:

Let $f(t), t \in \mathcal{T}$ be a zero mean Gaussian stochastic prosess with $Ef(s)f(t) = bK(s,t), (s,t) \in \mathcal{T} \otimes \mathcal{T}$. Let

$$y_i = f(t(i)) + \epsilon_i, \;\; i = 1, ..., n$$

with $\epsilon = (\epsilon_1, \cdots \epsilon_n)' \sim \mathcal{N}(0, \sigma^2 I)$. Then

$$\begin{aligned} \hat{f}(t) &= Ef(t)|y \\ &= (K(t,t(1)), \cdots, K(t,t(n))(K + (\sigma^2/b)I)^{-1}y, \\ &\quad t \in \mathcal{T} \end{aligned}$$

where K is the $n \times n$ matrix with ij th entry $K(t(i), t(j))$. Note that $Ef(t)|y$ $\boxed{\textit{is defined for all } t \in \mathcal{T}}$. However, evaluating $\hat{f}$ at $t(1), \cdots, t(n)$ results in the familiar looking formula:

$$\begin{aligned} E(\begin{pmatrix} f(t(1)) \\ f(t(2)) \\ \vdots \\ f(t(n)) \end{pmatrix} |y) &= K(K + (\sigma^2/b)I)^{-1}y \\ &\equiv A(\lambda)y, say, with \lambda = (\sigma^2/b). \end{aligned}$$

• The general case. The variational problem:

What is the variational problem corresponding to $\min \|y - f\|^2 + \lambda f'\Sigma^{-1} f$? Let $\mathcal{H}_K$ be the RKHS with reproducing kernel $K(s,t)$. *I AM NOT TELLING YOU WHAT THAT OBJECT IS, YET*, other than it is a collection of functions defined on $\mathcal{T}$. Let f_λ in $\mathcal{H}_K$ minimize

$$\sum_{i=1}^{n} (y_i - f(t(i)))^2 + \lambda \|f\|^2_{\mathcal{H}_K}$$

where $\|f\|^2$ is the squared norm in $\mathcal{H}_K$. Then

$$\begin{aligned} \hat{f}_\lambda(t) &= Ef(t)|y \\ &= (K(t,t(1)), \cdots, K(t,t(n))(K+\lambda I)^{-1} y, \\ &\quad\; t \in \mathcal{T}. \end{aligned}$$

MORAL: Given the prior $f(t), t \in \mathcal{T}$ a 0 mean Gaussian stochastic process with $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$, the posterior mean for $f|y$ is the solution to a variational problem in an RKHS. *I STILL HAVENT TOLD YOU WHAT AN RKHS is*, but you should suspect that $\|f\|^2_{\mathcal{H}_K}$ somehow generalizes the square norm $f'\Sigma^{-1} f$ on E^d.

♣♣ Reproducing Kernel Hilbert Spaces

We describe N dimensional and infinite dimensional RKHS and their inner products.

• The N dimensional case:

Let Σ be strictly positive definite. Then Σ defines a perfectly good inner product on E^N by

$$< f, g >= f' \Sigma^{-1} g.$$

Let $(\sigma_1, \sigma_2, \cdots, \sigma_N)$ be the columns of Σ. Then

$$\boxed{< \sigma_i, \sigma_j >= \sigma_{ij}},$$

where σ_{ij} is the ijth entry of Σ.

- The N dimensional case continued:

Given the inner product

$$< f, g >= f'\Sigma^{-1}g,$$

letting $(\sigma_1, \sigma_2, \cdots, \sigma_N)$ be the columns of Σ. Why do we have

$$\boxed{< \sigma_i, \sigma_j >= \sigma_{ij}} \quad ??$$

Because

$$\begin{aligned}
< \sigma_i, \sigma_j > &= \sigma_i'\Sigma^{-1}\sigma_j \\
&= \sigma_i' \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix} = \sigma_{ij}
\end{aligned}$$

since $\Sigma^{-1} \begin{pmatrix} | & | & \cdots & | \\ \sigma_1 & \sigma_2 & \cdots & \sigma_N \\ | & | & \cdots & | \end{pmatrix} = I$. More generally, let $f = f(t(1)), \cdots, f(N))'$, then $\boxed{< \sigma_i, f >= f(i).}$ Taking the inner product of f with the ith row of Σ picks out the value of f at $t(i)$.

• The general case: Construction of an RKHS from a positive definite function.

Recall that the columns $\sigma_i, i = 1, \cdots N$ span E^N. We are now going to construct a general RKHS from the 'columns' of an arbitrary positive definite function. Let $K(\cdot, \cdot)$ be a positive definite function on $\mathcal{T} \otimes \mathcal{T}$. Define the tth 'column' of K as

$$K_t(\cdot) = K(t, \cdot).$$

By this we mean that t is fixed and K_t is a function of $(\cdot)$. K_t is a function on $\mathcal{T}$. With $K(\cdot, \cdot)$ we can associate a (unique!) collection of functions, to be called $\mathcal{H}_K$, as follows:

$$\begin{aligned} K_t &\in \mathcal{H}_K \quad for\, each\, t \in \mathcal{T}, \\ \sum_{\ell=1}^{L} a_\ell K_{t_\ell} &\in \mathcal{H}_K \;\; for\, any\, finite\, L\, and\, \{a_\ell\}. \;\; (*) \end{aligned}$$

The inner product in $\mathcal{H}_K$ is defined by

$$< K_s, K_t >= K(s, t)$$

and extened by linearity to functions of the form $(*)$. Note that for $f \in \mathcal{H}_K$,

$$\boxed{< K_t, f >= f(t)}$$

since $\sum_\ell a_\ell K_{t_\ell}(t) \equiv < K_t, \sum_\ell a_\ell K_{t_\ell} >=< K_t, f >$ Let $f_n, f_m \in \mathcal{H}_K$. Then

$$\begin{aligned} |f_n(t) - f_m(t)| &= | < K_t, f_n - f_m > | \\ &\leq \|K_t\| \|f_n - f_m\| \end{aligned}$$

by the Cauchy-Schwartz Inequality ($(u, v) \leq \|u\| \|v\|$). Therefore, if $f_n, f_{n+1} \cdots$ is a Cauchy sequence (this means $\|f_n - f_m\| \to 0$ as $n, m \to \infty$) then $|f_n(t) - f_m(t)| \to 0$. (In words, strong convergence implies pointwise convergence here). We add the pointwise limits of all these functions to $\mathcal{H}_K$ and we have a $\boxed{\text{REPRODUCING KERNEL HILBERT SPACE}}$. K is called the reproducing kernel for $\mathcal{H}_K$.

♣♣ The Moore-Aronszajn Theorem: (Aronszajn 1950).

Let $\mathcal{T}$ be an index set. To every positive definite function K on $\mathcal{T} \times \mathcal{T}$ there corresponds a unique RKHS $\mathcal{H}_K$ of real valued functions on $\mathcal{T}$ and vice versa. Letting $< \cdot, \cdot >_{\mathcal{H}_K}$, we have for every $f \in \mathcal{H}_K$, and every $t \in \mathcal{T}, < K_t, f >_{\mathcal{H}_K} = f(t)$, where $K_t(\cdot) = K(t, \cdot)$.

Remark: The formal definition of an $RKHS$ is: A Hilbert space where all the evaluation functionals are bounded. What this means is, that, if $\mathcal{H}_K$ is a Hilbert space, it is an RKHS if and only if, for $f \in \mathcal{H}_K$, and each $t \in \mathcal{T}$, there exists M_t, not depending on f such that $|f(t)| \leq M_t \|f\|$. As a consequence, by the Riesz representation theorem there exists a representer, ξ_t, with the property that $< \xi_t, f >_{\mathcal{H}_K} = f(t)$. From the above, $\xi_t = K_t$, furthermore, $< K_s, K_t >= K(s, t)$, which is the source of the name 'Reproducing Kernel'.

♣♣ More on Inner Products in RKHS

We will describe the inner product in the N dimensional case and see (one of the) generalizations to infinite dimensional spaces.

• The N dimensional case:

Let $\Sigma = \Gamma D \Gamma$, where $\Gamma = \{\Phi_\nu(i)\}$ is orthogonal and D is diagonal, with diagonal entries λ_ν. Then we can write the ijth entry of Σ as

$$\sigma_{ij} = \sum_{\nu=1}^{N} \lambda_\nu \Phi_\nu(i) \Phi_\nu(j).$$

We have

$$< f, g >= f' \Sigma^{-1} g \equiv \sum_{\nu=1}^{N} \frac{(f, \Phi_\nu)(g, \Phi_\nu)}{\lambda_\nu}$$

where (u, v) is the Euclidean inner product.

• The (almost most) general case:

The Mercer-Hilbert-Schmidt Theorem: Let $K(s,t)$ be a positive definite function with $\int_{\mathcal{T}}\int_{\mathcal{T}} K^2(s,t)dsdt = C \le \infty$. Then $\exists$ an orthonormal set on $\mathcal{T}$, $\{\Phi_\nu\}_{\nu=1}^{\infty}$

$$\int_{\mathcal{T}}\int_{\mathcal{T}} \Phi_\mu(s)\Phi_\nu(s)ds \;=\; 1, \mu = \nu; = 0 \;\; otherwise$$

and nonnegative eigenvalues λ_ν with $\sum_{\nu=1}^{\infty} \lambda_\nu^2 = C$ such that

$$K(s,t) = \sum_{\nu=1}^{\infty} \lambda_\nu \Phi_\nu(s)\Phi_\nu(t). \qquad \diamond$$

The inner product in $\mathcal{H}_K$ will have a representation

$$< f, g >= \sum_{\nu=1}^{\infty} \frac{(f,\Phi_\nu)(g,\Phi_\nu)}{\lambda_\nu}$$

where $(u,v) = \int_{\mathcal{T}} u(s)v(s)ds$. In practice we need only to be given $K(\cdot,\cdot)$ but not $\{\phi_\nu, \lambda_\nu\}$ to solve problems in $\mathcal{H}_K$. However in the next slide we know the eigenfunctions Φ_ν and eigenvalues λ_ν along with a closed form expression for $K(\cdot,\cdot)$ in the case of periodic polynomial splines.

♣♣ Examples

• Periodic Splines

Let W_m^o (per) be the collection of all functions on $[0,1]$ of the form

$$f(t) \sim \sqrt{2} \sum_{\nu=1}^{\infty} a_\nu \cos 2\pi\nu t + \sqrt{2} \sum_{\nu=1}^{\infty} b_\nu \sin 2\pi\nu t$$

with

$$\sum_{\nu=1}^{\infty} (a_\nu^2 + b_\nu^2)(2\pi\nu)^{2m} < \infty.$$

Since

$$\frac{d^m}{dt^m} \left\{ \begin{array}{c} \cos 2\pi\nu t \\ \sin 2\pi\nu t \end{array} \right\} = (2\pi\nu)^m \times \begin{array}{c} \pm \sin 2\pi\nu t \\ \pm \cos 2\pi\nu t, \end{array}$$

then if (**??**) holds, we have

$$\sum_{v=1}^{\infty} (a_\nu^2 + b_\nu^2)(2\pi\nu)^{2m} = \int_0^1 (f^{(m)}(u))^2 du.$$

Elements in W_m^o (per) satisfy the periodic boundary conditions

$$\begin{aligned}
\int_0^1 f(u)du &= 0 \\
\int_0^1 f^{(k)}(u)du &= f^{(k-1)}(1) - f^{(k-1)}(0) = 0, \\
& \quad k = 1, \ldots, m.
\end{aligned}$$

It can be shown that that the RK for W_m^o (per) is

$$\begin{aligned}
K(s,t) &= \sum_{\nu=1}^{\infty} \lambda_\nu \Phi_\nu(s)\Phi_\nu(t) \\
&= \sum_{\nu=1}^{\infty} \frac{2}{(2\pi\nu)^{2m}} [\cos 2\pi\nu s \cos 2\pi\nu t \\
&\quad + \sin 2\pi\nu s \sin 2\pi\nu t] \\
&= \sum_{\nu=1}^{\infty} \frac{2}{(2\pi\nu)^{2m}} \cos 2\pi\nu(s-t).
\end{aligned}$$

A closed form expression for $K(s,t)$ using Bernoulli polynomials is available:

The first few Bernoulli polynomials are:

$$\begin{array}{rcl} B_0(t) & = & 1 \\ B_1(t) & = & t - 1/2 \\ B_2(t) & = & t^2 - t + 1/6 \\ B_3(t) & = & t^3 - 3t^2/2 + t/2 \\ B_4(t) & + & t^4 - 2t^3 + t^2 - 1/30 \end{array}$$

Let $k_m(t) = B_m(t)/m!$. $K(s,t)$ is given (Abromowitz and Stegun, 1965) by

$$K(s,t) = (-1)^{m-1} k_{2m}([s-t])$$

where $[s-t]$ is the fractional part of $s-t$. (For example, [1.2] = .2.) The inner product in $\mathcal{H}_K$ is

$$\begin{array}{rcl} < f, g > & = & \sum_{\nu=1}^{\infty} [a_\nu(f) a_\nu(g) + b_\nu(f) b_\nu(g)](2\pi\nu)^{2m} \\ & \equiv & \int_0^1 f^{(m)}(u) g^{(m)}(u) du. \end{array}$$

where $a_\nu(h), b_\nu(h)$ are the Fourier cosine and sine coefficients for h.

♣♣ The Representer Theorem (simple case)

Let $g_i(y_i, \tau)$ be convex in τ for each i, y_i. Then `Any solution to the problem: find` $f \in \mathcal{H}_K$ `to minimize`

$$\frac{1}{n}\sum_{i=1}^{n} g_i(y_i, f(t(i))) + \lambda\|f\|^2_{\mathcal{H}_K} \tag{1}$$

`has a representation of the form`

$$f(\cdot) = \sum_{i=1}^{n} c_i K(t(i), \cdot).$$

The proof goes back to Kimeldorf and Wahba(1971), and we only sketch it here. If $f \in \mathcal{H}_K$, then we can always write

$$f(\cdot) = \sum_{i=1}^{n} c_i K_{t(i)}(\cdot) + \rho \tag{2}$$

where $\rho \perp K_{t(i)}$. (This means that $< K_{t(i)}, \rho > \equiv \rho(t(i)) = 0$!). Substituting (2) into (1) will show that $\|\rho\|^2 = 0$.

♣♣ Sums and Products of Positive Definite Functions

There are many ways to obtain positive definite functions, for example $K(s,t) = \int_{\mathcal{U}} G(s,u)G(t,u)du$ will be positive definite for any G. Tensor sums and products of positive definite functions are positive definite functions. For example let $s = (s_1, s_2), t = (t_1, t_2)$ in $[0,1]^2$, the unit square. Let $r_1(s_1,t_1)$ and $r_2(s_2,t_2)$ be positive definite functions on $[0,1] \otimes [0,1]$ Then, for example $K(s,t) = r_1(s_1,t_1) + r_2(s_2,t_2)$ $+ r_1(s_1,t_1)r_2(s_2,t_2)$ is a positive definite function on $[0,1]^2 \otimes [0,1]^2$. Furthermore, with some care r_1 and r_2 can be chosen so that $\mathcal{H}_K$ is the direct sum of three orthogonal subspaces corresponding to the three positive definite functions in the sum. This allows us to build up useful models with various combinations of reproducing kernels as building blocks. We will return to this later.