

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.

© Grace Wahba 2000

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> → 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, 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 $\mathcal{T} = 1, 2, \dots, N$. $K(s, t), (s, t) \in \mathcal{T} \otimes \mathcal{T}$ 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 $a = (a_1, \dots, a_N)$ 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 $t_1, \dots, t_n \in \mathcal{T}$, and every $a = (a_1, \dots, a_n)$, $\sum_{i,j=1}^n a_i a_j K(t_i, t_j) \geq 0$.

♣♣ 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

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

- 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 process with $E f(s) f(t) = bK(s, t), (s, t) \in \mathcal{T} \otimes \mathcal{T}$.

Let

$$y_i = f(t(i)) + \epsilon_i, \quad i = 1, \dots, n$$

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

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

where K is the $n \times n$ matrix with ij th entry $K(t(i), t(j))$.

Note that $E f(t) | y$ is defined for all $t \in \mathcal{T}$. However, evaluating \hat{f} at $t(1), \dots, t(n)$ results in the familiar looking formula:

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

- 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

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

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

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

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 $\|f\|_{\mathcal{H}_K}^2$ 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

$$\langle f, g \rangle = f' \Sigma^{-1} g.$$

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

$$\boxed{\langle \sigma_i, \sigma_j \rangle = \sigma_{ij}},$$

where σ_{ij} is the ij th entry of Σ .

- The N dimensional case continued:

Given the inner product

$$\langle f, g \rangle = f' \Sigma^{-1} g,$$

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

$$\boxed{\langle \sigma_i, \sigma_j \rangle = \sigma_{ij}} \quad ??$$

Because

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

since $\Sigma^{-1} \begin{pmatrix} | & | & \cdots & | \\ \sigma_1 & \sigma_2 & \cdots & \sigma_N \\ | & | & \cdots & | \end{pmatrix} = I$. More gener-

ally, let $f = (f(t(1)), \dots, f(N))'$, then $\boxed{\langle \sigma_i, f \rangle = f(i)}$.

Taking the inner product of f with the i th row of Σ^{-1} 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 $\sigma_i, i = 1, \dots, N$ 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 t th ‘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:

$$K_t \in \mathcal{H}_K \quad \text{for each } t \in \mathcal{T},$$

$$\sum_{\ell=1}^L a_\ell K_{t_\ell} \in \mathcal{H}_K \quad \text{for any finite } L \text{ and } \{a_\ell\}. \quad (*)$$

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

$$\langle K_s, K_t \rangle = K(s, t)$$

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

$$\boxed{\langle K_t, f \rangle = f(t)}$$

since $\sum_{\ell} a_{\ell} K_{t_{\ell}}(t) \equiv \langle K_t, \sum_{\ell} a_{\ell} K_{t_{\ell}} \rangle = \langle K_t, f \rangle$

Let $f_n, f_m \in \mathcal{H}_K$. Then

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

by the Cauchy-Schwartz Inequality ($(u, v) \leq \|u\| \|v\|$).
 Therefore, if $f_n, f_{n+1} \dots$ is a Cauchy sequence (this means $\|f_n - f_m\| \rightarrow 0$ as $n, m \rightarrow \infty$) then $|f_n(t) - f_m(t)| \rightarrow 0$. (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 **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 $\langle \cdot, \cdot \rangle_{\mathcal{H}_K}$, we have for every $f \in \mathcal{H}_K$, and every $t \in \mathcal{T}$, $\langle K_t, f \rangle_{\mathcal{H}_K} = f(t)$, 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 $\langle \xi_t, f \rangle_{\mathcal{H}_K} = f(t)$. From the above, $\xi_t = K_t$, furthermore, $\langle K_s, K_t \rangle = K(s, t)$, 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 ij th entry of Σ as

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

We have

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

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 $\int_{\mathcal{T}} \int_{\mathcal{T}} K^2(s, t) ds dt = C \leq \infty$. Then \exists an orthonormal set on \mathcal{T} , $\{\Phi_\nu\}_{\nu=1}^\infty$

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

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

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

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

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

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

$$\frac{d^m}{dt^m} \begin{cases} \cos 2\pi\nu t \\ \sin 2\pi\nu t \end{cases} = (2\pi\nu)^m \times \begin{matrix} \pm \sin 2\pi\nu t \\ \pm \cos 2\pi\nu t, \end{matrix}$$

then if (??) holds, we have

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

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

$$\int_0^1 f(u) du = 0$$

$$\int_0^1 f^{(k)}(u) du = f^{(k-1)}(1) - f^{(k-1)}(0) = 0,$$

$$k = 1, \dots, m.$$

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:

$$B_0(t) = 1$$

$$B_1(t) = t - 1/2$$

$$B_2(t) = t^2 - t + 1/6$$

$$B_3(t) = t^3 - 3t^2/2 + t/2$$

$$B_4(t) = t^4 - 2t^3 + t^2 - 1/30$$

Let $k_m(t) = B_m(t)/m!$. $K(s, t)$ is given (Abramowitz 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

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

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

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

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

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

where $\rho \perp K_{t(i)}$. (This means that $\langle K_{t(i)}, \rho \rangle \equiv \rho(t(i)) = 0!$). 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.

Part II*

1. The polynomial smoothing spline.
2. Leaving-out-one, GCV and other smoothing parameter estimates.
3. The thin plate smoothing spline.
4. Generalizations: Different kinds of observations: Non-gaussian, indirect, constrained.
5. Examples: The histospline, convolution equations with positivity constraints. GCV with inequality constraints.

*Part II of 'An Introduction to Model Building With Reproducing Kernel Hilbert Spaces', by Grace Wahba, Univ. of Wisconsin Statistics Department TR 1020, Overheads for Interface 2000 Short Course. © Grace Wahba, 2000

♣♣ The Polynomial Smoothing Spline

- The polynomial smoothing spline is the forerunner of much more general RKHS models.

Let W_m be the collection of functions on $[0, 1]$ with $\int_0^1 (f^{(m)}(u))^2 du \leq \infty$. The polynomial smoothing spline is the solution to the problem: Find $f \in W_m$ to min

$$\frac{1}{n} \sum_{i=1}^n (y_i - f(t(i)))^2 + \lambda \int_0^1 (f^{(m)}(u))^2 du.$$

It can be shown that $W_m = W_m^0 \oplus \pi_m$, where π_m is the span of the polynomials of degree m or less. We rearrange things so that

$$W_m = \mathcal{H}_0 \oplus \mathcal{H}_1$$

where $\mathcal{H}_0 = \pi_{m-1}$, the polynomials of degree $m-1$ or less, on which there will be no penalty, and $\mathcal{H}_1 = W_m^0 \oplus \{k_m\}$. It can be shown that the RK for \mathcal{H}_1 with square norm $\|f\|^2 = \int_0^1 (f^{(m)}(u))^2 du$ is

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

By an argument generalizing the representer theorem, and upon observing that $\{k_\nu\}_{\nu=0}^{m-1}$ span \mathcal{H}_0 , it follows that the minimizer f_λ has the form

$$f_\lambda(t) = \sum_{\nu=1}^m d_\nu k_{\nu-1}(t) + \sum_{i=1}^n c_i K(t(i), t), \quad (1)$$

and that

$$\int_0^1 (f^{(m)}(u))^2 du = \sum_{i,j=1}^n c_i c_j K(t(i), t(j)). \quad (2)$$

Upon substituting (1) and (2) into the original variational problem, the solution is obtained by minimizing a quadratic form in $d = (d_1, \dots, d_m)'$ and $c = (c_1, \dots, c_n)'$. (There are easier ways to get the polynomial spline, but the present way of going about it is the one which we see will generalize in many ways.)

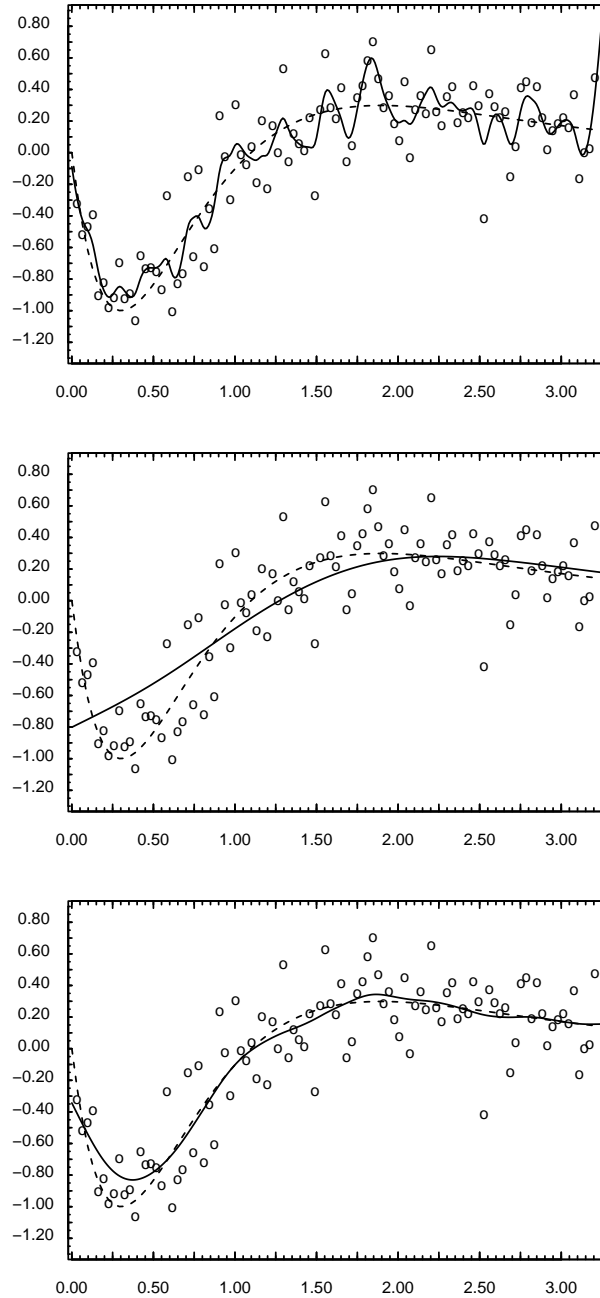


Figure 1: Dashed Lines are Smoothing Splines with λ too small, λ too large, and λ estimated via GCV, from the top. Solid line is ¹'truth'.

♣♣ Choosing λ .

- Leaving-out-one.

Let $f_\lambda^{[k]}(\cdot)$ be the minimizer of

$$\frac{1}{n} \sum_{\substack{i=1 \\ i \neq k}}^n (y_i - f(t(i)))^2 + \lambda \int_0^1 (f^{(m)}(u))^2 du.$$

The leaving-out-one estimate of λ is the minimizer of

$$o(\lambda) = \frac{1}{n} \sum_{k=1}^n (y_k - f_\lambda^{[k]}(t(k)))^2.$$

- GCV (Generalized Cross Validation).

The influence matrix $A(\lambda)$ with kk th entry $a_{kk}(\lambda)$ plays an important role. The influence matrix relates the data to the predicted data:

$$\begin{pmatrix} f_\lambda(t(1)) \\ f_\lambda(t(2)) \\ \vdots \\ f_\lambda(t(n)) \end{pmatrix} \equiv A(\lambda)y.$$

- GCV (continued)

We have the Lemma:

$$o(\lambda) \equiv \frac{1}{n} \sum_{k=1}^n \frac{(y_k - f_\lambda(t(k)))^2}{(1 - a_{kk}(\lambda))}$$

The GCV estimate of λ :

$$\begin{aligned} \min (\lambda) &= \frac{\frac{1}{n} \sum_{k=1}^n (y_k - f_\lambda(t(k)))^2}{(1 - \frac{1}{n} \sum_{\ell=1}^n a_{\ell\ell}(\lambda))^2} \\ &\equiv \frac{\|(I - A(\lambda))y\|^2}{\frac{1}{n}(I - \text{tr} A(\lambda))^2} \end{aligned}$$

- Unbiased Risk (if you know σ^2).

$$\min (\lambda) = \|(I - A(\lambda))y\|^2 + 2\sigma^2 \text{tr} A(\lambda)$$

- Generalized Max Likelihood (GML, aka REML).

$$\min M(\lambda) = \frac{y'(I - A(\lambda))y}{[\text{+}(I - A(\lambda))]^{1/(n-M)}}$$

+ = product of the $n - M$ non-zero eigenvalues.

- Cubic smoothing spline with GCV, SOFTWARE.

Codes for cubic (or higher order) smoothing splines with GCV to choose the smoothing parameter. Approximately reverse chronological order.

...

Code- Author- Where Found (* = *freeware*)

— — —

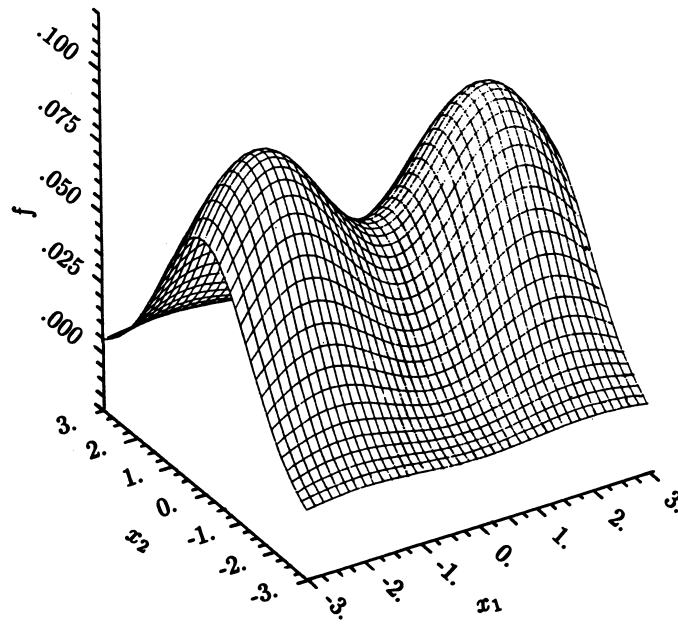
- * pspline-Jim Ramsay-<http://www.r-project.org>
- smooth.spline()-Trevor Hastie-Splus
- * sbart-Finbarr O'Sullivan-<http://www.netlib.org/gcv>
- * gcvspl-H. J. Woltring-<http://www.netlib.org/gcv>

♣♣♣ The Thin Plate Spline

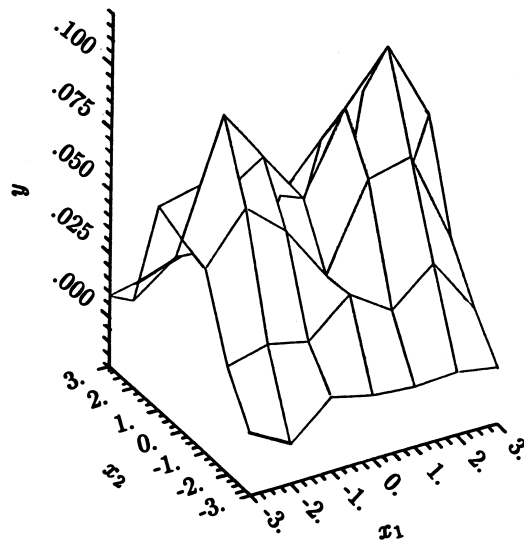
- The thin plate spline is one of the two-dimensional generalizations of the univariate spline.

Letting $t = (t_1, t_2)$, the penalty $\int (f^{(2)})^2$ is replaced by

$$(f) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} [f_{t_1 t_1}^2 + 2f_{t_1 t_2}^2 + f_{t_2 t_2}^2] dt_1 dt_2.$$

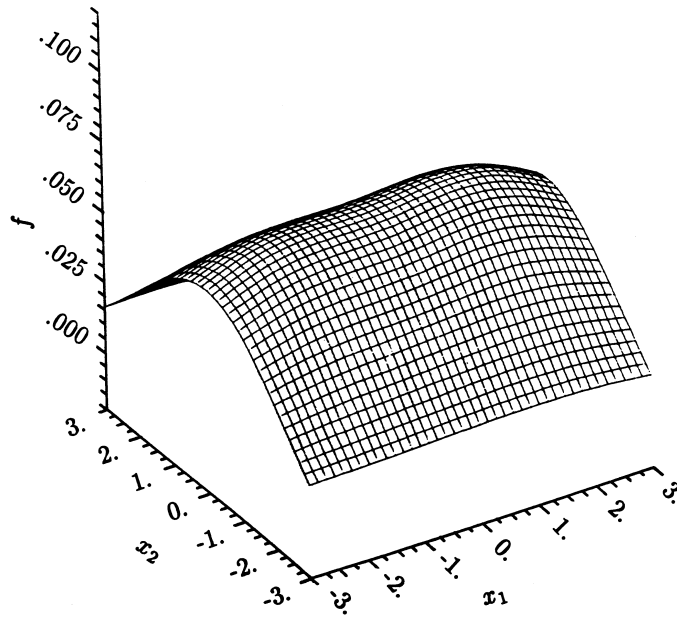


The actual surface.

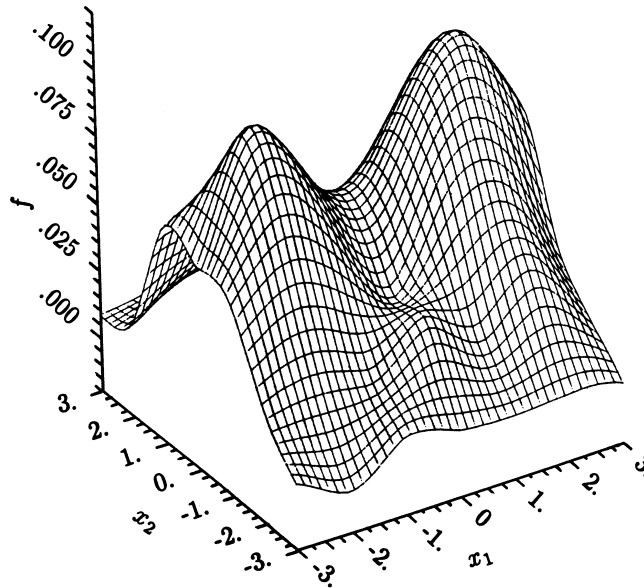


The data.

Figure 2a: Thin plate spline demo. Top: True surface. Bottom: The observations.

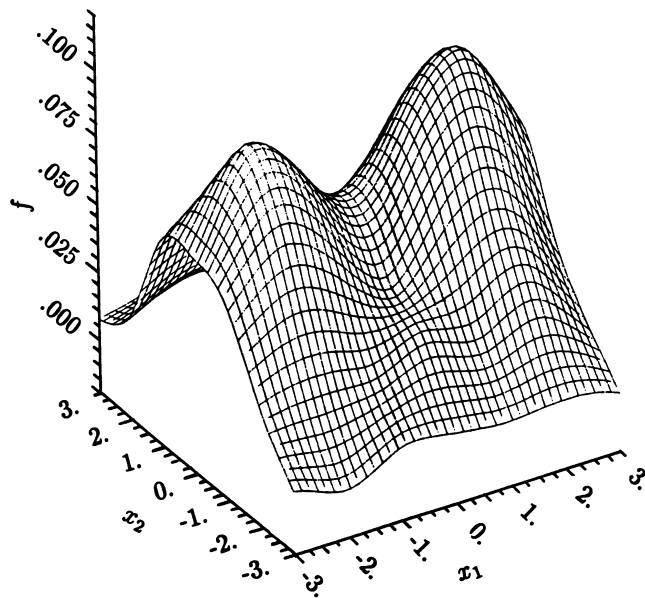


f_λ with λ too large, $\lambda = 100\hat{\lambda}$.



f_λ with λ too small, $\lambda = .01\hat{\lambda}$.

Figure 2b. Top: f_λ with λ too large. Bottom: f_λ with λ too small.



f_λ with λ estimated by GCV.

Figure 2c. $f_{\hat{\lambda}}$ with λ estimated by GCV.

- Thin plate spline with GCV, SOFTWARE.

Codes for thin plate smoothing splines with GCV to choose the smoothing parameter. Approximately reverse chronological order.

...

Code- Author- Where Found (* = *freeware*)

--- --- ---

tpspline-Dong Xiang-SAS

* funfits-Doug Nychka-[http://www.cdg.ucar.edu/
/stats/software.shtml](http://www.cdg.ucar.edu/stats/software.shtml)

ANUSPLIN-M. Hutchinson-[http://cres20.anu.edu/
/au/software/anusplin.html](http://cres20.anu.edu/au/software/anusplin.html)

* GCVPACK-Bates et al-<http://www.netlib.org/gcv>

♣♣ The Representer Theorem (more general case)

Let $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$, where \mathcal{H}_0 is a finite dimensional space spanned by $\phi_\nu, \nu = 1, \dots, M$, and \mathcal{H}_1 is an RKHS with square norm $\|f\|_{\mathcal{H}_K}^2$. Find $f = f_0 + f_1$ with $f_0 \in \mathcal{H}_0$ and $f_1 \in \mathcal{H}_1$ to min

$$I_\lambda(y, f) = \frac{1}{n} \sum_{i=1}^n g_i(y_i, f(t(i))) + \lambda \|f_1\|_{\mathcal{H}_K}^2.$$

Suppose g is convex in f and the minimizer of $\sum_{i=1}^n g_i(y_i, f(t(i)))$ in \mathcal{H}_0 is unique. Then the minimizer f_λ of I_λ is unique and has a representation

$$f(\cdot) = \sum_{\nu=1}^M d_\nu \phi_\nu(\cdot) + \sum_{i=1}^n c_i K_{t(i)}(\cdot).$$

where (d, c) minimize

$$\sum_{i=1}^n g_i(y_i, (d + Kc)_i) + \lambda c' K c.$$

Here, ${}_{n \times M} = \{\phi_\nu(t(i))\}$, $K_{n \times n} = \{K(t(i), t(j))\}$ and $(\)_i$ means the i th component of $\$.

♣♣♣ Quick List of Generalizations

- In the 'distance' of f from observations.
 1. $g(y, f) = \log$ likelihood
 2. $g(y, f) =$ robust functional
 3. $g(y, f) =$ support vector machine (SVM) functional
 4. $g(y, f) =$ indicator functionals, e. g. $g(y, f) = 0$ or ∞ according as $f \in [y + u, y -]$
- In the kinds of observations.
- In the imposition of constraints.
- In the domain of the model, $\mathcal{T} \rightarrow \mathcal{T}^{(1)} \otimes \dots \otimes \mathcal{T}^{(d)}$

- In the kinds of observations: Integrals:

Replace $f(t)$ by $L_t f$ where the $L_t f$ are bounded linear functionals in \mathcal{H} : Example: Tomography.

$$L_t f = \int_{\mathcal{I}} H(t, u) f(u) du.$$

Then K_t is replaced by

$$\xi_t(\cdot) = \int_{\mathcal{U}} H(t, u) K(u, \cdot) du$$

and $\langle K_s, K_t \rangle$ is replaced by

$$\langle \xi_s, \xi_t \rangle = \int_{\mathcal{U}} \int_{\mathcal{U}} H(s, u) K(u, v) H(t, u) du dv.$$

ξ_t is called the representer of L_t in \mathcal{H} ,

$$L_t f \equiv \langle \xi_t, f \rangle = \int_{\mathcal{I}} H(t, u) f(u) du.$$

Where does this come from?

- In the kinds of observations: The Eta Theorem:

Theorem: Let L be a bounded linear functional in an RKHS \mathcal{H}_K with RK K . By the Riesz representation theorem there exists an η in \mathcal{H}_K such that

$$Lf = \langle \eta, f \rangle, \quad \forall f \in \mathcal{H}_K.$$

We may find η by observing that

$$(Lf)(s) = \langle K_s, \eta \rangle.$$

Since

$$\langle K_s, \eta \rangle \equiv LK_s,$$

the representer of any bounded linear functional in \mathcal{H}_K may be found by applying the bounded linear functional to K_s , and then looking at the result as a function of s .

This allows us to estimate f based on observations on integrals and even derivatives if \mathcal{H}_K is chosen so that these are bounded linear functionals. Derivatives up to the $m - 1$ st are bounded linear functionals in W_m .

- In the kinds of observations: The Representer Theorem (even more general case)

Let $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$, where \mathcal{H}_0 is a finite dimensional space spanned by $\phi_\nu, \nu = 1, \dots, M$, and \mathcal{H}_1 is an RKHS with square norm $\|f\|_{\mathcal{H}_K}^2$. Find $f = f_0 + f_1$ with $f_0 \in \mathcal{H}_0$ and $f_1 \in \mathcal{H}_1$ to min

$$I_\lambda(y, f) = \frac{1}{n} \sum_{i=1}^n g_i(y_i, L_i f) + \lambda \|f_1\|_{\mathcal{H}_K}^2,$$

where the $\{L_i\}$ are bounded linear functionals on \mathcal{H} . Suppose g is convex in f and the minimizer of $\sum_{i=1}^n g_i(y_i, L_i f)$ over f in \mathcal{H}_0 is unique. Then the minimizer f_λ of I_λ is unique and has a representation

$$f(\cdot) = \sum_{\nu=1}^M d_\nu \phi_\nu(\cdot) + \sum_{i=1}^n c_i \xi_i(\cdot). \quad (3)$$

where ξ_i is the representer for L_i in \mathcal{H}_1 , $L_i f \equiv \langle \xi_i, f \rangle$ and (d, c) minimize

$$\frac{1}{n} \sum_{i=1}^n g_i(y_i, (d + Kc)_i) + \lambda c' K c.$$

Here, $_{n \times M} = \{L_i \phi_\nu\}$, $K_{n \times n} = \{\langle \xi_i, \xi_j \rangle\}$.

- The histospline. Given area integrals.

Female lung cancer rates in Wisconsin, by county.

$$y_i = \frac{1}{|\Omega_i|} \int_{\Omega_i} f(A) dA + \epsilon_i.$$

The thin plate penalty is used for the histospline.

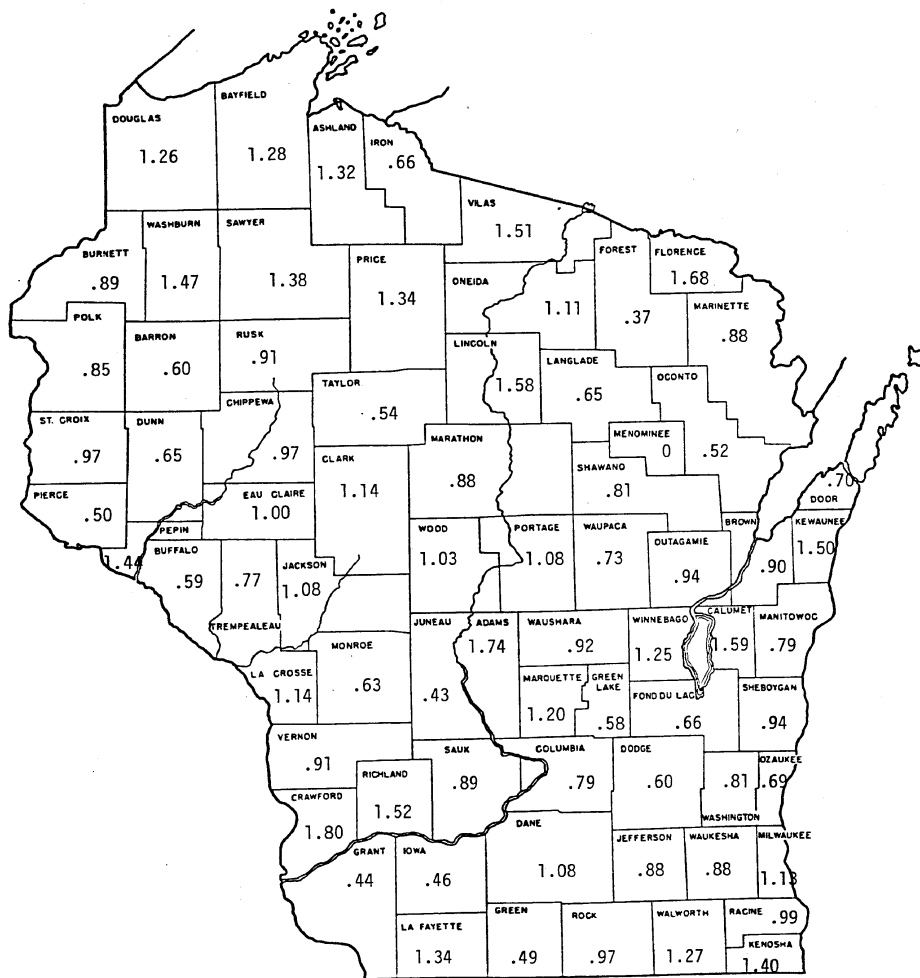


Fig. 3.13. 1970-1975 Female Lung Cancer, Revised SMR's by County

Figure 3a. Female lung cancer rates, by county.

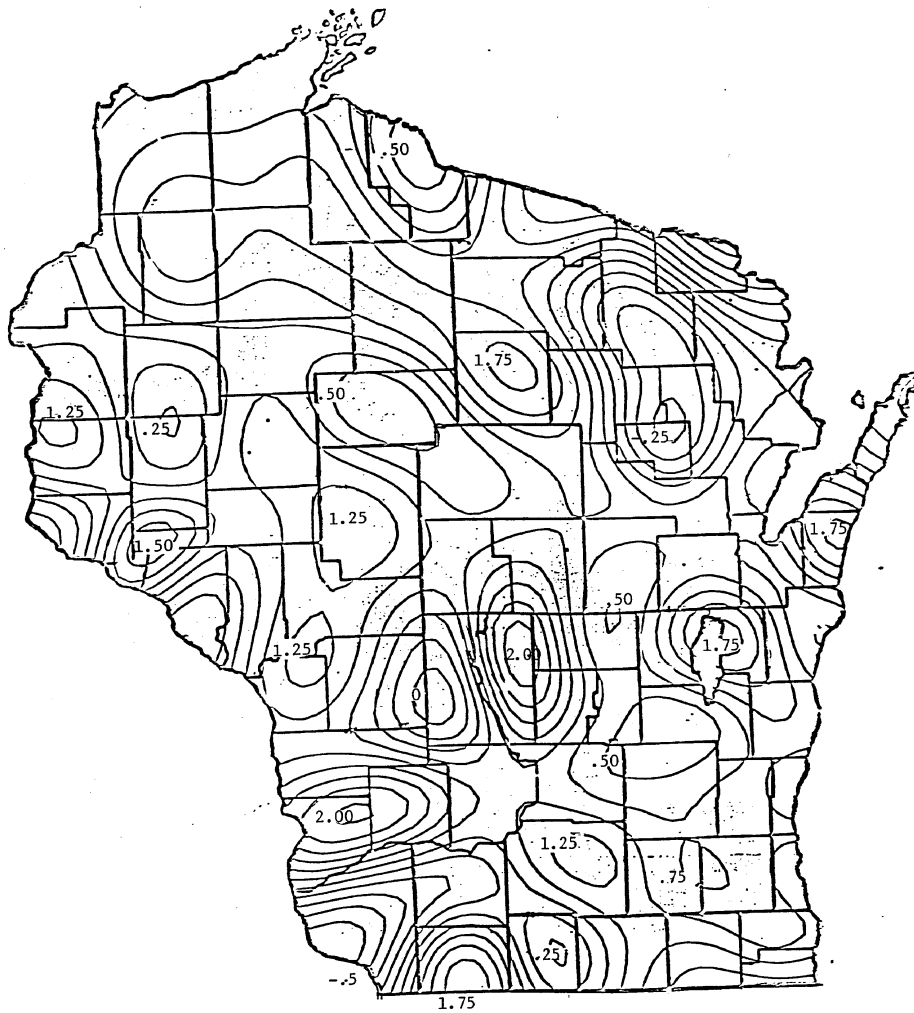


Figure 3.14. Female Lung Cancer Revised SMR's. Volume Matching Histospline. Contour Interval: 0.25.

Figure 3b. Volume matching. Min (f) subject to $\frac{1}{|\Omega_i|} \int_{\Omega_i} \hat{f}(A) dA = y_i$

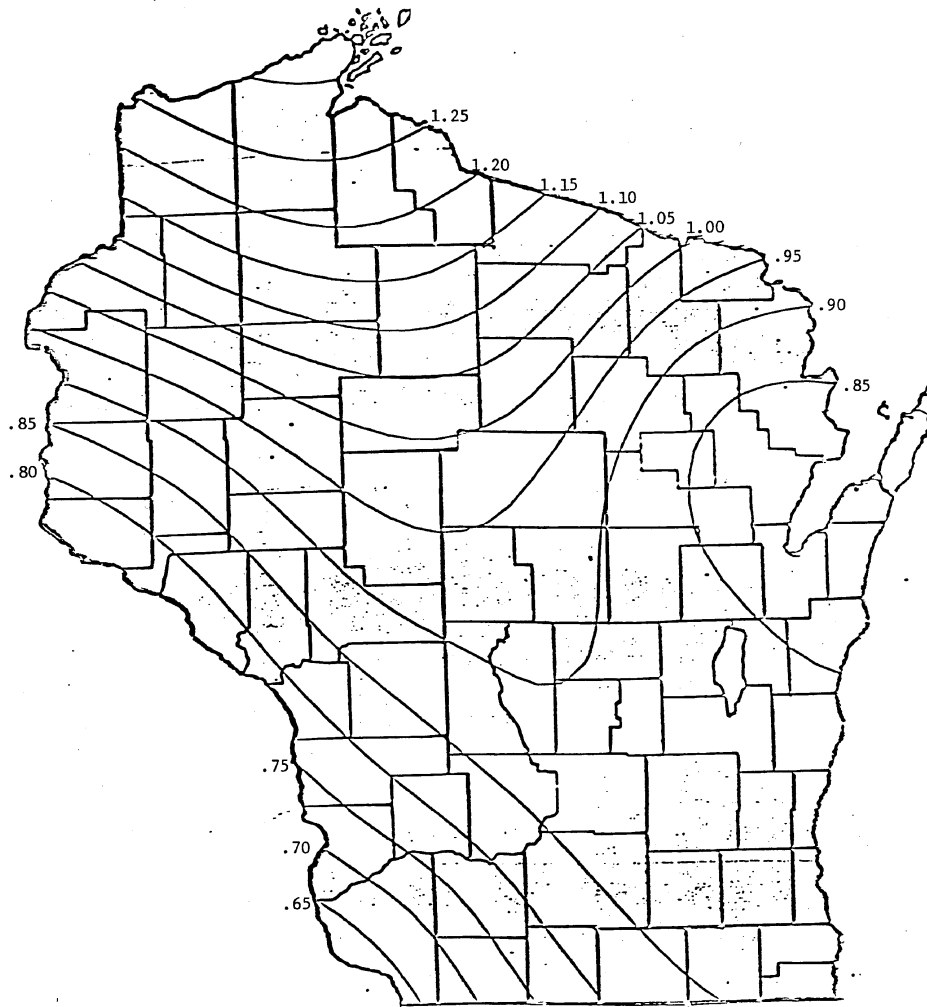


Figure 3.15. 1979-1975 Female Lung Cancer, Histospline Smoothed by GCV. Contour Interval: 0.05.

Figure 3c. Volume smoothing. Find $f_\lambda \in \mathcal{H}$ to min $\sum_i (y_i - \frac{1}{|\Omega_i|} \int_{\Omega_i} f(A) dA)^2 + \lambda (f)$

- In the kinds of observations: Constraints

Let \mathcal{H} as before. Find $f = f_0 + f_1$ with $f_0 \in \mathcal{H}_0$ and $f_1 \in \mathcal{H}_1$ to min

$$I_\lambda(y, f) = \frac{1}{n} \sum_{i=1}^n g_i(y_i, L_i f) + \lambda \|f_1\|_{\mathcal{H}_K}^2$$

subject to

1. Positivity: $f(t) \geq 0$
2. Linear inequality constraints: $N_t f \geq a_t$
3. Constraints via solutions to PDE's:
 $t = (\text{time}, \text{space}), H(\mathcal{L}f) \leq C$

To compute, the constraints are discretized. The representers of the constraints are incorporated in the representation of the solution. For inequality constraints, the coefficients are obtained by solving a mathematical programming problem. (MINOS) In typical cases where the family of constraints is 'smooth' the addition of a few constraints will lead to the constraints actually being satisfied everywhere.

Let $\{N_j\}$ be a finite set of discretized constraints, and let $\{\xi_j\}$ be their representers, $\langle \xi_j, f \rangle = N_j f$. The problem then becomes Find $f = f_0 + f_1$ to min

$$I_\lambda(y, f) = \frac{1}{n} \sum_{i=1}^n g_i(y_i, L_i f) + \lambda \|f_1\|_{\mathcal{H}_K}^2$$

subject to

$$\langle \xi_j, f \rangle \geq a_j, \quad j = 1, \dots, m,$$

and the solution has a representation

$$f(\cdot) = \sum_{\nu=1}^M d_\nu \phi_\nu(\cdot) + \sum_{i=1}^n c_i \xi_i(\cdot) + \sum_j c_j \xi_j(\cdot).$$

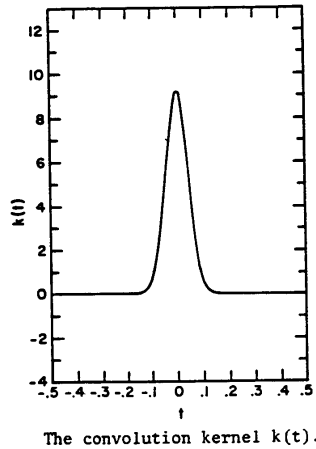


Figure 4a. The convolution kernel

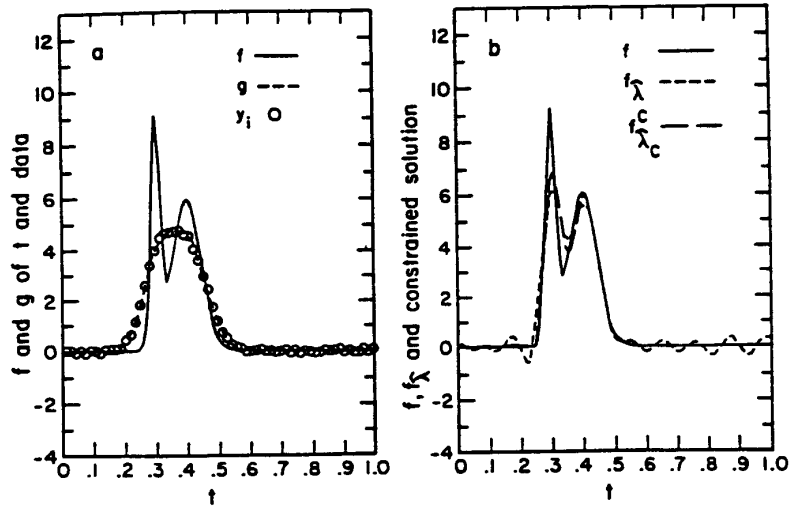
- Positivity constraints in a convolution equation.

$$y_i = \int k(v_i, u) f(u) du + \epsilon_i.$$

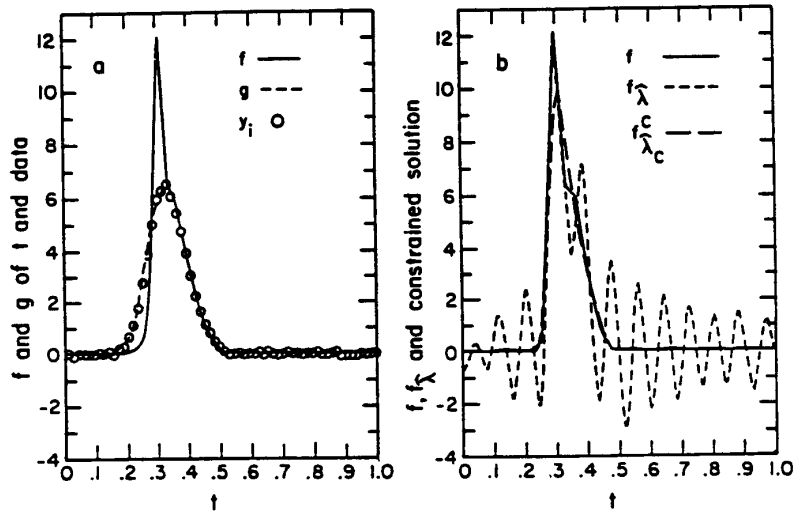
Find f_λ to min

$$\frac{1}{n} \sum_i (y_i - \int k(v_i, u) f(u) du)^2 + \int (f^{(2)})^2,$$

subject to $f_\lambda(u_j) \geq 0$. The GCV for constrained problems: For fixed λ , solve the quadratic programming problem, and find the active constraints. At the solution, the same answer will be obtained by throwing away the inactive constraints and putting in the active inequality constraints as equality constraints. This problem is linear - compute the GCV for it.



$f, g, \text{data}, \hat{f}_\lambda$ and $\hat{f}_{\lambda^c}^c$
 peak separation = .10.



f, g, \hat{f}_λ and $\hat{f}_{\lambda^c}^c$
 peak separation = .05.

Figure 4b. Two examples: Left panels-true f and observations y_i . Right panels-true f , unconstrained solution (wiggly) and constrained solution \hat{f}_λ .

Part III*

1. SS-ANOVA Spaces on General Domains
2. Averaging Operators and ANOVA Decompositions
3. Reproducing Kernel Spaces for ANOVA Decompositions
4. Building Blocks for SS-ANOVA Spaces, General and Particular
5. Representation of SS-ANOVA Fits
6. Example: Risk of Progression of Diabetic Retinopathy in the WESDR Study. Bernoulli data.

*Part III of 'An Introduction to Model Building With Reproducing Kernel Hilbert Spaces', by Grace Wahba, Univ. of Wisconsin Statistics Department TR 1020, Overheads for Interface 2000 Short Course. © Grace Wahba, 2000

7. GACV for smoothing parameters in the Bernoulli case.

- General Model Domains: $\mathcal{T} = \mathcal{T}^{(1)} \otimes \dots \otimes \mathcal{T}^{(d)}$.
SS-ANOVA spaces.

Let $t = (t_1, \dots, t_d)$, $t \in \mathcal{T}^{(\alpha)}$, $\alpha = 1, \dots, d$. Some examples are:

$$\mathcal{T}^{(\alpha)} = [0, 1] \quad \text{unit interval}$$

$$\mathcal{T}^{(\alpha)} = E^r \quad \text{Euclidean } r - \text{space}$$

$$\mathcal{T}^{(\alpha)} = \mathcal{S} \quad \text{the sphere}$$

$$\mathcal{T}^{(\alpha)} = \{1, \dots, N\} \quad \text{ordered categorical}$$

$$\mathcal{T}^{(\alpha)} = \{\diamond, \triangle, \heartsuit\} \quad \text{unordered categorical}$$

... ..

We let $t \in \mathcal{T} \equiv \mathcal{T}^{(1)} \otimes \dots \otimes \mathcal{T}^{(d)}$. Let \mathcal{E} be an averaging operator on $\mathcal{T}^{(\alpha)}$, defined by

$$\mathcal{E} f = \int_{\mathcal{T}^{(\alpha)}} f d\mu$$

where $d\mu$ is some given probability distribution on $\mathcal{T}^{(\alpha)}$, for example if $\mathcal{T}^{(\alpha)} = [0, 1]$ the uniform distribution is convenient. Given the \mathcal{E} , any real valued function $f(t) = f(t_1, \dots, t_d)$ on \mathcal{T} has an ANOVA decomposition as follows:

Here's the ANOVA decomposition:

$$\begin{aligned}
 f(t) &= \mu + \sum f(t) \\
 &+ \sum_{\leq} f(t, t) \\
 &+ \cdots + f_{1, \dots, d}(t_1, \dots, t_d)
 \end{aligned}$$

where the components are generated by the decomposition of the identity:

$$\begin{aligned}
 f &= [\mathcal{E} + (I - \mathcal{E})]f \\
 f &= \mathcal{E} f + \sum (I - \mathcal{E}) \mathcal{E} f \\
 &+ \sum_{<} (I - \mathcal{E})(I - \mathcal{E}) \mathcal{E} f \\
 &+ \cdots + \sum_{=1}^d (I - \mathcal{E})f.
 \end{aligned}$$

(from the previous slide)

$$\begin{aligned}
 f &= \mathcal{E} f + \sum (I - \mathcal{E}) \mathcal{E} f \\
 &\quad \neq \\
 &+ \sum_{<} (I - \mathcal{E})(I - \mathcal{E}) \mathcal{E} f \\
 &\quad \neq , \\
 &+ \dots + \sum_{=1}^d (I - \mathcal{E}) f.
 \end{aligned}$$

Therefore:

$$\begin{aligned}
 \mu &= \mathcal{E} f, \quad f = (I - \mathcal{E}) \mathcal{E} f \\
 &\quad \neq \\
 f_{,} &= (I - \mathcal{E})(I - \mathcal{E}) \mathcal{E} f \\
 &\quad \neq , \\
 \dots &\quad \dots \\
 f_{1,2,\dots,d} &= \sum_{=1}^d (I - \mathcal{E}) f,
 \end{aligned}$$

and satisfy the ANOVA SIDE CONDITIONS

$$\begin{aligned}
 \mathcal{E} f &= 0 \\
 \mathcal{E} f &= \mathcal{E} f = 0 \\
 \mathcal{E} f &= \mathcal{E} f = \mathcal{E} f = 0 \\
 &\vdots
 \end{aligned}$$

Now, let $\mathcal{H}^{(\alpha)}$ be an RKHS of functions on $\mathcal{T}^{(\alpha)}$ with $\int_{\mathcal{T}^{(\alpha)}} f(t) d\mu = 0$ for all $f(t) \in \mathcal{H}^{(\alpha)}$, and let $[1^{(\alpha)}]$ be the one dimensional space of constant functions on $\mathcal{T}^{(\alpha)}$. We can construct an RKHS \mathcal{H} as the direct sum of subspaces which correspond to this decomposition:

$$\begin{aligned} \mathcal{H} &= \bigoplus_{\alpha=1}^d [\{[1^{(\alpha)}]\} \oplus \{\mathcal{H}^{(\alpha)}\}] \\ \mathcal{H} &= [1] \oplus \sum_{\alpha} \mathcal{H}^{(\alpha)} \\ &\quad \oplus \sum_{\alpha < \beta} [\mathcal{H}^{(\alpha)} \otimes \mathcal{H}^{(\beta)}] \\ &\quad \oplus \dots \oplus \bigoplus_{\alpha=1}^d \mathcal{H}^{(\alpha)}. \end{aligned}$$

($[1]^{(\alpha)}$ are omitted wherever they occur).

Let $R(s, t)$ be the RK for $\mathcal{H}^{(1)}$. A (Smoothing Spline) ANOVA space \mathcal{H}_K of functions on $\mathcal{T} = \mathcal{T}^{(1)} \otimes \dots \otimes \mathcal{T}^{(d)}$ is given by:

$$\mathcal{H}_K = \sum_{i=1}^d [[1^{(i)}] \oplus \mathcal{H}^{(i)}]$$

which then has the RK

$$\begin{aligned} K(s, t) &= \sum_{i=1}^d [1 + R(s, t)] \\ &= 1 + \sum_{i=1}^d R(s, t) \\ &\quad + \sum_{\substack{1 \leq i < j \leq d}} R(s, t) R(s, t) \\ &\quad + \dots + \sum_{i=1}^d \otimes R(s, t). \end{aligned}$$

- SS-ANOVA spaces, continued.

$\mathcal{H}(\cdot)$ may be further decomposed into a low dimensional parametric part $\mathcal{H}_\pi(\cdot)$, and a ‘smooth’ part $\mathcal{H}_s(\cdot)$, $\mathcal{H}(\cdot) = \mathcal{H}_\pi(\cdot) \oplus \mathcal{H}_s(\cdot)$. This can be done in many ways, depending on the $\mathcal{T}(\cdot)$ and what part of the model it is desired not to penalize. A useful example when $\mathcal{T}(\cdot) = [0, 1]$, which will be employed later is:

$$\begin{aligned} \mathcal{H}(\cdot) &= \{k_1\} \oplus [\{k_2\} \oplus W_2^0] \\ R(s, t) &= r_\pi(s, t) + r_s(s, t) \text{ say} \end{aligned}$$

where

$$\begin{aligned} r_\pi(s, t) &= k_1(s)k_1(t), \\ r_s(s, t) &= k_2(s)(k_2(t)) - k_4([s - t]) \end{aligned}$$

We have encountered r_s before, the square norm in its associated RKHS is $\int_0^1 (f'')^2$. In this example the $\mathcal{T}(\cdot)$ and r_π and r_s will be the same for each component of t , but this not necessary.

- SS-ANOVA spaces, continued.

Now $K(s, t)$ can be seen to be expandable in the tensor sums and products of $r_\pi(s, t)$ and $r_s(s, t)$, $\alpha = 1, \dots, d$. The expansion is carried out and truncated (Model selection!), in our experience, interactions higher than two-factor can generally be deleted, and frequently only a few two factor interactions are important. Finally, terms containing only r_π 's will not be penalized, and are collected into \mathcal{H}_0 , and the spanning set for \mathcal{H}_0 will be relabeled as $\{\phi_1, \dots, \phi_M\}$, The terms with one or more r_s are collected into \mathcal{H}_1 , and relabeled as $\mathcal{H}_1 = \sum \mathcal{H}_i$, with the RK's Q_i for the \mathcal{H}_i weighted and relabeled as

$$Q(s, t) = \sum \theta_i Q_i(s, t).$$

Note that the Q_i generally depend only on a subset of the components of (s, t) . The θ_i allow for different smoothing parameters for the different components.

- SS-ANOVA spaces, continued.

We have, finally, reduced an arbitrary ANOVA model to the case established via the representer theorem: Find $f = f_0 + f_1$ with $f_0 \in \mathcal{H}_0$ and $f_1 \in \mathcal{H}_1$ to min

$$I_\lambda(y, f) = \frac{1}{n} \sum_{i=1}^n g_i(y_i, f(t(i))) + \lambda \sum \theta^{-1} \|P f\|_{\mathcal{H}_{Q_\beta}}^2.$$

where $P f$ is the component of f in \mathcal{H}_β . Then the minimizer f_λ of I_λ is unique and has, by the representer theorem, the representation

$$f(\cdot) = \sum_{\nu=1}^M d_\nu \phi_\nu(\cdot) + \sum_{i=1}^n c_i Q(t(i), \cdot).$$

- SS-ANOVA Example: Risk of Progression of Diabetic Retinopathy in the Younger Onset Population in the Wisconsin Epidemiologic Study of Diabetic Retinopathy.

Data:

$$\{y_i, t(i)\}, t = (\text{dur}, \text{gly}, \text{bmi}), i = 1, \dots, n = 66 .$$

where

$$\begin{aligned} y_i &= 1, \text{ progression yes} \\ &= 0, \text{ progression no} \\ \text{dur} &= \text{duration of diabetes at baseline} \\ \text{gly} &= \text{glycosylated hemoglobin} \\ \text{bmi} &= \text{body mass index} \end{aligned}$$

Goal: Estimate $p(t)$, the probability of progression given t . Let $f(t) = \log[p(t)/(1-p(t))]$. The $-\loglik(y, f)$ is

$$-\log[p^y(1-p)^{1-y}] \equiv -yf + \log(1 + e^f) \equiv g(y, f)$$

- SS-ANOVA Example: Diabetic Retinopathy (con't).

We selected the model

$$f(\text{dur}, \text{gly}, \text{bmi}) = \mu + f_1(\text{dur}) + a_2 \cdot \text{gly} \\ + f_3(\text{bmi}) + f_{13}(\text{dur}, \text{bmi})$$

\mathcal{H}_0

$$\{\phi_\nu(\text{dur}, \text{gly}, \text{bmi})\} = \{1, k_1(\text{dur}), k_1(\text{gly}), k_1(\text{bmi})\}$$

\mathcal{H}_1

$$\beta \quad Q(\text{dur}, \text{bmi}; \text{dur}', \text{bmi}')$$

$$1 \quad r_s(\text{dur}, \text{dur}')$$

$$2 \quad r_s(\text{bmi}, \text{bmi}')$$

$$3 \quad r_\pi(\text{dur}, \text{dur}')r_s(\text{bmi}, \text{bmi}')$$

$$4 \quad r_s(\text{dur}, \text{dur}')r_\pi(\text{bmi}, \text{bmi}')$$

$$r_s(\text{dur}, \text{dur}')r_s(\text{bmi}, \text{bmi}')$$

$$Q(\text{dur}, \text{bmi}; \cdot) = \sum_{=1}^5 \theta Q(\text{dur}, \text{bmi}; \cdot).$$

•SS-ANOVA EXAMPLE: Diabetic Retinopathy
(con't).

We will minimize

$$I_\lambda(y, f) = \frac{1}{n} \sum_{i=1}^n [-\loglik(y_i, f_i) + \lambda \sum \theta^{-1} \|P f\|_{\mathcal{H}_{Q\beta}}^2]$$

where $f_i = f(t(i))$, $P f$ is the component of f in $\mathcal{H}_{Q\beta}$. The minimizer f_λ of I_λ has the representation

$$f(\cdot) = \sum_{\nu=1}^M d_\nu \phi_\nu(\cdot) + \sum_{i=1}^n c_i Q(t(i), \cdot),$$

and we need to compute (d, c) to min

$$\frac{1}{n} \sum_{i=1}^n (-y_i f_i + \log(1 + e^{f_i})) + \lambda c' K c$$

where $f_i = (Td + Kc)_i$, $T_{n \times 4} = \{\phi_\nu(t(i))\}$, $K_{n \times n} = Q(t(i), t(j))$; and we need to estimate $\lambda = \lambda \theta^{-1}$, $\beta = 1, \dots, \dots$

♣♣♣ Choosing $\lambda = (\lambda_1, \dots, \lambda_n)$, Bernoulli data.

Notation: Let $f_\lambda^{[k]}(\cdot)$ be the minimizer of $I_\lambda(y, f)$ with the k th data point omitted. Let $f_{\lambda k}^{[k]} = f_\lambda^{[k]}(t(k))$, $f_{\lambda k} = f_\lambda(t(k))$. Let $b(f) = \log(1 + e^f)$, thus $g(y, f) = -yf + b(f)$.

• Leaving-out-one.

Choose λ to min

$$V_0(\lambda) = \frac{1}{n} \sum_{k=1}^n [-y_k f_{\lambda k}^{[k]} + b(f_{\lambda k})].$$

Generally not practical.

- GACV (Generalized Approximate Cross Validation).

The inverse Hessian $H(\lambda)$ of I_λ with respect to $(f_{\lambda 1}, \dots, f_{\lambda n})$ at the minimizer plays an important role. It is an interesting fact that the influence matrix $A(\lambda)$ in the Gaussian case is also the inverse Hessian of I_λ in the Gaussian case, and this is true to first order in the general exponential family case, and in some other situations. H can be thought of as the (local) influence matrix, since in the nonquadratic case it depends on f_λ . It can be shown that

$$V_0(\lambda) \approx ACV(\lambda) = \frac{1}{n} \sum_{k=1}^n [-y_k f_{\lambda k} + b(f_{\lambda k})] + D_0(\lambda).$$

where

$$D_0 = \frac{1}{n} \sum_{i=1}^n \frac{h_{ii} y_i (y_i - p_{\lambda i})}{[1 - h_{ii} \sigma_{ii}]},$$

h_{ii} is the ii th entry of $H(\lambda)$, $p_{\lambda i} = e^{f_{\lambda i}} / (1 + e^{f_{\lambda i}})$, $\sigma_{ii} = p_{\lambda i} (1 - p_{\lambda i})$. The GACV is obtained from the ACV by replacing h_{ii} and $h_{ii} \sigma_{ii}$ by their averages, as follows:

In the expression for D_0 h_{ii} is replaced by $\frac{1}{n} \sum_{i=1}^n h_{ii} \equiv \frac{1}{n} \text{tr}(H)$ and $1 - h_{ii}\sigma_{ii}$ is replaced by $\frac{1}{n} \text{tr}[I - (W^{1/2}HW^{1/2})]$, where $W = \text{diag}\{\sigma_{ii}\}$, giving

$$GACV(\lambda) = \frac{1}{n} \sum_{i=1}^n [-y_i f_{\lambda i} + b(f_{\lambda i})] + \frac{\text{tr}(H)}{n} \frac{\sum_{i=1}^n y_i (y_i - p_{\lambda i})}{\text{tr}[I - (W^{1/2}HW^{1/2})]}.$$

The randomized trace technique may be used to evaluate $GACV$:

$$\text{ran}GACV(\lambda) = \frac{1}{n} \sum_{i=1}^n [-y_i f_{\lambda i} + b(f_{\lambda i})] + \frac{\delta' (f_{\lambda}^{y+} - f_{\lambda}^y)}{n} \frac{\sum_{i=1}^n y_i (y_i - p_{\lambda i})}{[\delta' \delta - \delta' W (f_{\lambda}^{y+} - f_{\lambda}^y)]}.$$

δ is a random white noise perturbation n -vector and f_{λ}^{y+} is the n -vector of values of the fit at the observation points **based on estimating f with perturbed data $y + \delta$** . We show next that

$$\delta' (f_{\lambda}^{y+} - f_{\lambda}^y)$$

provides a randomized estimate of $\text{trace}H(\lambda)$.

- Randomized Trace Estimates.

δ is a (small) random perturbation with $E\delta = 0$ and $cov\delta = \sigma I$. For any matrix H , $E\delta'H\delta = \sigma \text{trace}H$. Now, let $H[\cdot]$ be the operator which maps a data vector z into the vector of values of f_λ at the observation points, that is, $H[z] = f_\lambda$. In the Gaussian case H is linear and we just have $H[z] = Hz$. We have, to first order

$$f_\lambda^{y^+} - f_\lambda^y \approx H[y + \delta] - H[y] \approx H[y^*]\delta$$

where y^* is some intermediate value between $y + \delta$ and y . Thus, we have the approximation

$$E\delta'(f_\lambda^{y^+} - f_\lambda^y) \sim \delta'H[y^*]\delta \approx \sigma \text{tr}H[y]$$

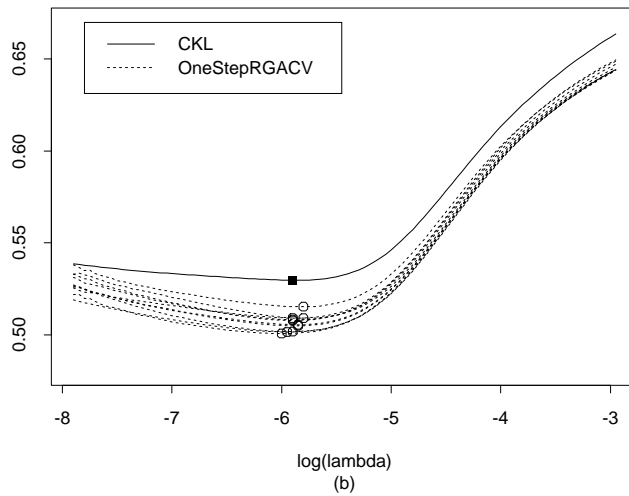
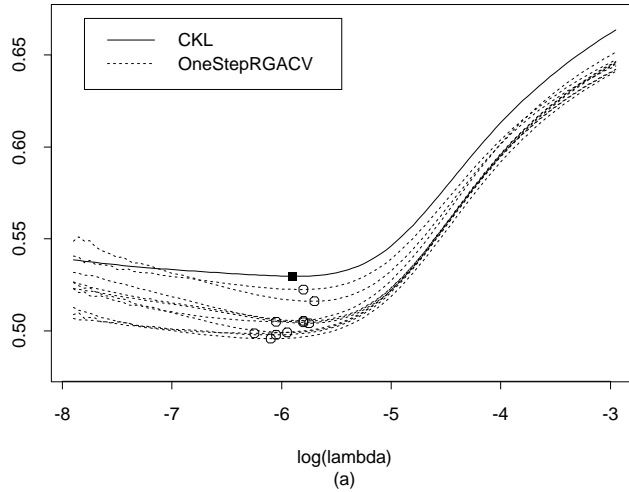
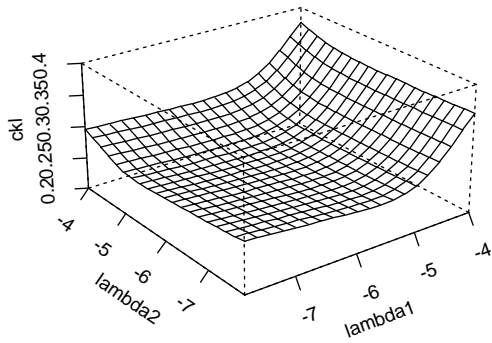
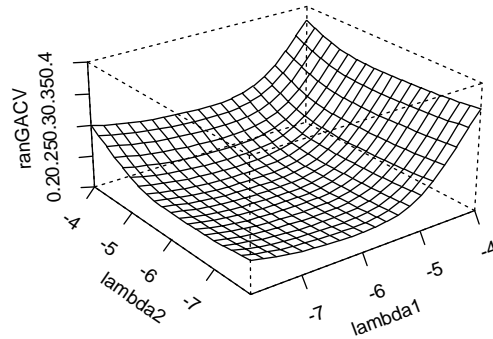


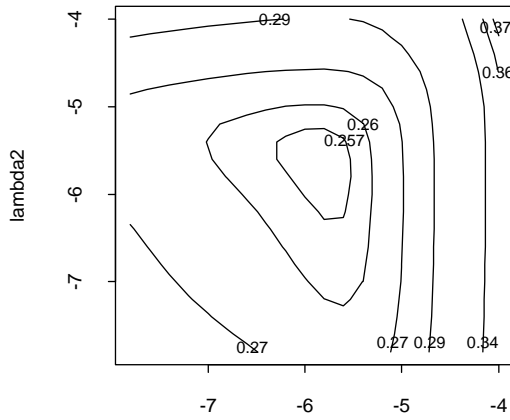
Figure 5a. 10 replicates of $ranGACV(\lambda)$, compared to $CKL(\lambda)$, where the Comparative Kullback-Liebler distance (CKL) is given by $CKL(\lambda)[p_\lambda, p_{R E}] = \frac{1}{n} \sum_{i=1}^n [-p_{R E} f_{\lambda i} + b(f_{\lambda i})]$



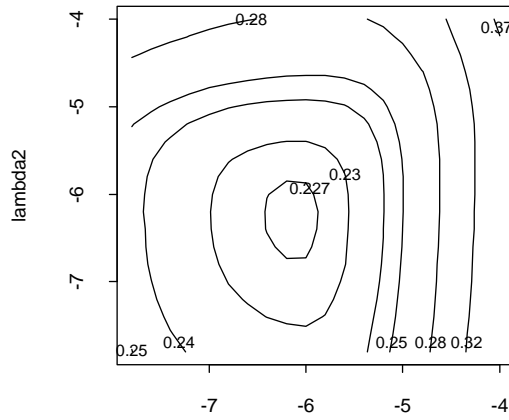
(a1) True CKL Surface



(a2) ranGACV Surface



(b1) Contour of CKL



(b2) Contour of ranGACV

Figure 5b. *ranGACV*, compared to the true *CKL*, $\lambda = (\lambda_1, \lambda_2)$. Left: *CKL*. Right: *ranGACV*.

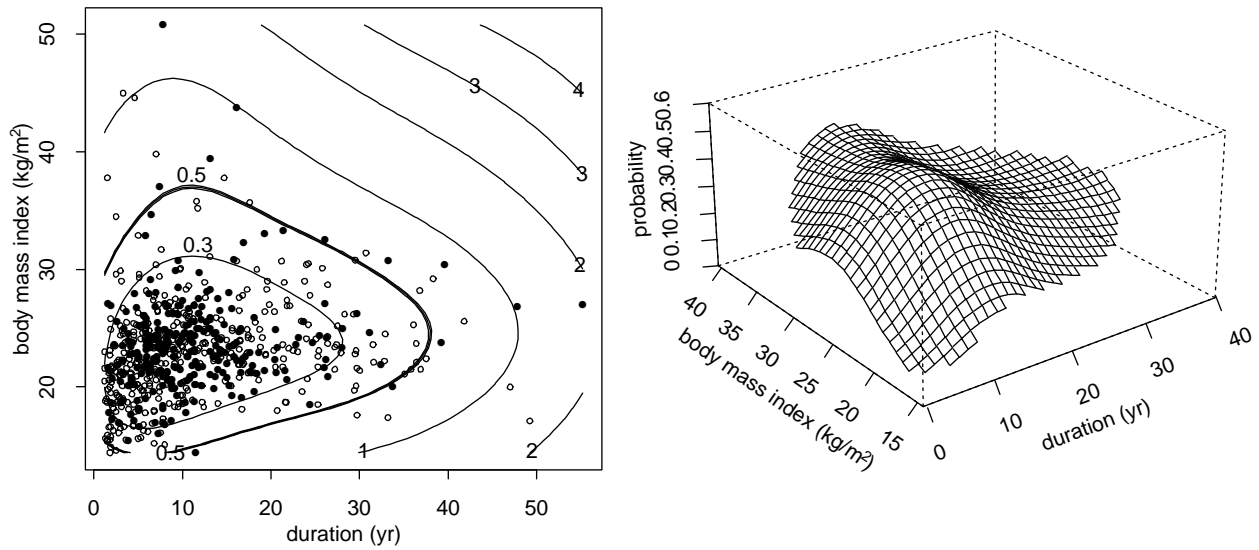


Figure 6a. Left: Data and contours of constant posterior standard deviation. Right: Estimated probability of progression as a function of duration and body mass index for glycosylated hemoglobin fixed at its median.

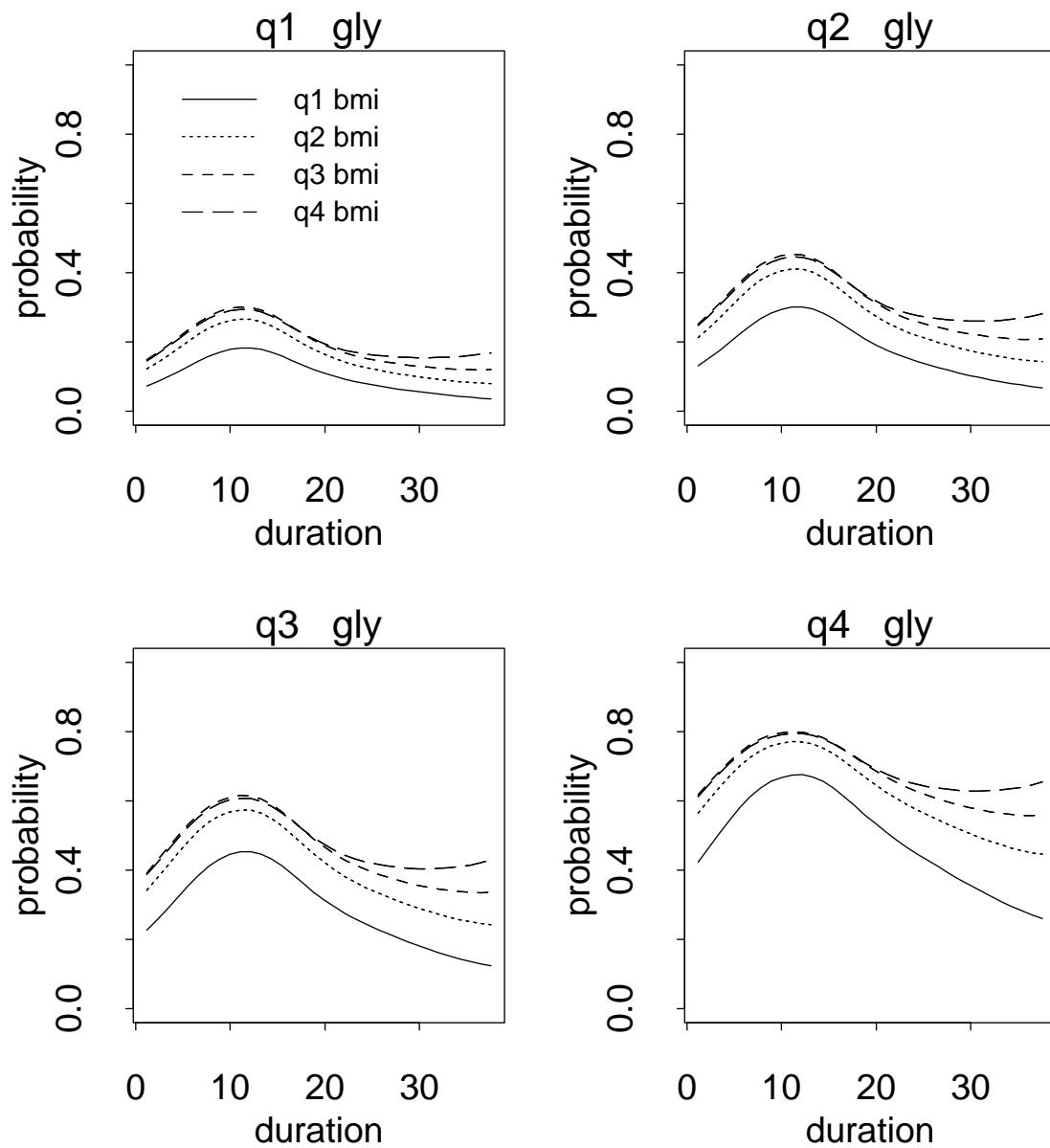


Figure 6b. Estimated probability of progression as a function of `dur` for four levels of `bmi` by four levels of `gly`.

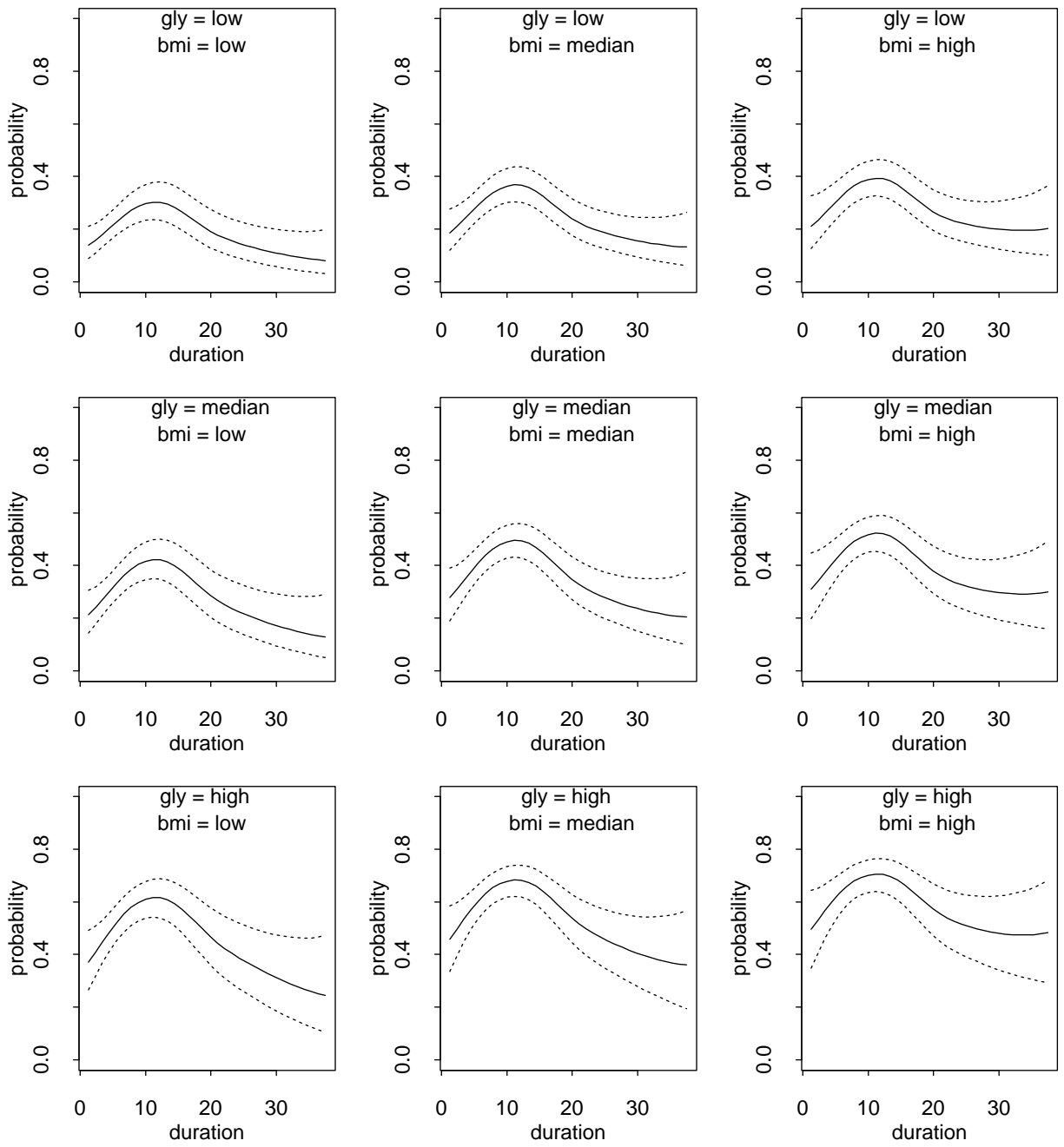


Figure 6c. Bayesian 'Confidence Intervals'

- SS-ANOVA spaces, SOFTWARE.

Codes for SS-ANOVA models, reverse chronological order: Use a Newton-Raphson algorithm for (d, c) given λ . Use an iterative unbiased risk estimate for λ in the Bernoulli case.

...

Code- Author- Where Found (* = *freeware*)

--- --- ---

- * gss- Chong Gu- <http://www.r-project.org>
- * GRKPACK- Yuedong Wang- <http://www.netlib.org/gcv>
- * RKPACK- Chong Gu- <http://www.netlib.org/gcv>

recap **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

Part II

1. The polynomial smoothing spline.
2. Leaving-out-one, GCV and other smoothing parameter estimates.
3. The thin plate smoothing spline.
4. Generalizations: Different kinds of observations: Non-gaussian, indirect, constrained.
5. Examples: The histospline, convolution equations with positivity constraints. GCV with inequality constraints.

Part III

1. SS-ANOVA Spaces on General Domains
2. Averaging Operators and ANOVA Decompositions
3. Reproducing Kernel Spaces for ANOVA Decompositions
4. Building Blocks for SS-ANOVA Spaces, General and Particular
5. Representation of SS-ANOVA Fits
6. Example: Risk of Progression of Diabetic Retinopathy in the WESDR Study. Bernoulli data.
7. GACV for smoothing parameters in the Bernoulli case.

♣♣♣ Ending Comments

Reproducing Kernel Hilbert Spaces apparently first appeared in the Statistics Literature in the work of Parzen in the late 60's, and although there was theoretical work in the early 70's there were several things that were necessary to make models based on them useful to the data analyst: (i) high speed computers that could handle the solution of large linear systems, (ii) method(s) for choosing the smoothing parameter(s), (iii) user friendly software, since the models are generally non-trivial to code from scratch. These things have come to pass for some models, but for some of the more recent methods, user-friendly software is not (yet) available. There are still many interesting open theoretical and practical problems for the research-minded - particularly related to variable and model selection in very large, complex data sets, and efficient code development. However, we hope we have shown that model building with RKHS has the flexibility and generality to handle a very wide variety of statistical data analysis problems, and have given the interested user ideas on how to begin doing this.

References for G. Wahba Short Course

PART I

Reproducing Kernel Hilbert Spaces

- [Amo97] L. Amodei. Reproducing kernels of vector-valued function spaces. In A. LeMehaute, C. Rabut, and L. Schumaker, editors, *Surface Fitting and Multiresolution Methods*, pages 17–26, Nashville TN, 1997. Vanderbilt University Press.
- [Aro50] N. Aronszajn. Theory of reproducing kernels. *Trans. Am. Math. Soc.*, 68:337–404, 1950.
- [Par63] E. Parzen. Probability density functionals and reproducing kernel Hilbert spaces. In M. Rosenblatt, editor, *Proceedings of the Symposium on Time Series Analysis*, pages 155–169. Wiley, 1963.
- [Wah90] G. Wahba. *Spline Models for Observational Data*. SIAM, 1990. CBMS-NSF Regional Conference Series in Applied Mathematics, v. 59.
- [Wah92] G. Wahba. Multivariate function and operator estimation, based on smoothing splines and reproducing kernels. In M. Casdagli and S. Eubank, editors, *Nonlinear Modeling and Forecasting, SFI Studies in the Sci of Complexity, Proc. Vol XII*, pages 95–112. Addison-Wesley, 1992.
- [Wei82] H. Weinert, editor. *Reproducing kernel Hilbert spaces: Application in signal processing*. Hutchinson Ross, Stroudsburg, PA, 1982.

Moore-Aronszajn Theorem, Mercer-Hilbert-Schmidt Theorem

- [Aro50] N. Aronszajn. Theory of reproducing kernels. *Trans. Am. Math. Soc.*, 68:337–404, 1950.
- [RN55] F. Riesz and B. Sz. Nagy. *Functional Analysis*. Ungar, New York, 1955.

Bernoulli Polynomials

- [AS65] M. Abramowitz and I. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables*. U. S. Gov't. Printing Office, Washington, D.C., 1965.

The Representer Theorem

- [KW71] G. Kimeldorf and G. Wahba. Some results on Tchebycheffian spline functions. *J. Math. Anal. Applic.*, 33:82–95, 1971.

PART II

GCV, GML and Unbiased Risk for estimating smoothing parameters

- [CW79] P. Craven and G. Wahba. Smoothing noisy data with spline functions: estimating the correct degree of smoothing by

the method of generalized cross-validation. *Numer. Math.*, 31:377–403, 1979.

- [GW91] C. Gu and G. Wahba. Minimizing GCV/GML scores with multiple smoothing parameters via the Newton method. *SIAM J. Sci. Statist. Comput.*, 12:383–398, 1991.
- [Li85] K. C. Li. From Stein’s unbiased risk estimates to the method of generalized cross-validation. *Ann. Statist.*, 13:1352–1377, 1985.
- [Li86] K. C. Li. Asymptotic optimality of C_L and generalized cross validation in ridge regression with application to spline smoothing. *Ann. Statist.*, 14:1101–1112, 1986.
- [Li7b] K. C. Li. Asymptotic optimality for $C_{sub p}$, $C_{sub L}$, cross-validation and generalized cross validation: discrete index set. *Ann. Math. Statist.*, 15:958–975, 1987b.
- [Mal73] C. Mallows. Some comments on C_p . *Technometrics*, 15:661–675, 1973.
- [Wah85] G. Wahba. A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. *Ann. Statist.*, 13:1378–1402, 1985.

The Thin Plate Spline

- [Duc77] J. Duchon. Splines minimizing rotation-invariant seminorms in Sobolev spaces. In *Constructive Theory of Functions of Several Variables*, pages 85–100. Springer-Verlag, Berlin, 1977.

- [HG94] M. Hutchinson and P. Gessler. Splines - more than just a smooth interpolator. *Geoderma*, 62:45–67, 1994.
- [WW80] G. Wahba and J. Wendelberger. Some new mathematical methods for variational objective analysis using splines and cross-validation. *Monthly Weather Review*, 108:1122–1145, 1980.

‘Distance’ $g(y, f)$

- [HNP98] Xuming He, Pin Ng, and Stephen Portnoy. Bivariate quantile smoothing splines. *Journal of the Royal Statistical Society, Series B, Methodological*, 60:537–550, 1998.
- [Len77] Russell V. Lenth. Robust splines. *Communications in Statistics, Part A – Theory and Methods*, 6:847–854, 1977.
- [LLW00] Y. Lin, Y. Lee, and G. Wahba. Support vector machines for classification in nonstandard situations. Technical Report 1016, Department of Statistics, University of Wisconsin, Madison WI, 2000.
- [MN89] P. McCullagh and J. Nelder. *Generalized Linear Models, Second Edition*. Chapman and Hall, 1989.
- [O’S83] F. O’Sullivan. *The analysis of some penalized likelihood estimation schemes*. PhD thesis, Dept. of Statistics, University of Wisconsin, Madison, WI, 1983. Technical Report 726.
- [OYR86] F. O’Sullivan, B. Yandell, and W. Raynor. Automatic smoothing of regression functions in generalized linear models. *J. Amer. Statist. Assoc.*, 81:96–103, 1986.

- [Utr81] Florencio I. Utreras. On computing robust splines and applications. *SIAM Journal on Scientific and Statistical Computing*, 2:153–163, 1981.
- [Wah69] G. Wahba. Estimating derivatives from outer space. Technical Report 989, Mathematics Research Center, 1969.
- [WLZ99] G. Wahba, Y. Lin, and H. Zhang. Generalized approximate cross validation for support vector machines, or, another way to look at margin-like quantities. Technical Report 1006, Department of Statistics, University of Wisconsin, Madison WI, 1999. to appear, *Advances in Large Margin Classifiers*, A. Smola, P. Bartlett, B. Scholkopf and D. Schurmans, eds, MIT Press.

Integrals

- [Gir7c] D. Girard. Optimal regularized reconstruction in computerized tomography. *SIAM J. Sci. Statist. Comput.*, 8:934–950, 1987c.
- [NWGP84] D. Nychka, G. Wahba, S. Goldfarb, and T. Pugh. Cross-validated spline methods for the estimation of three dimensional tumor size distributions from observations on two dimensional cross sections. *J. Am. Stat. Assoc.*, 79:832–846, 1984.
- [O’S6a] F. O’Sullivan. A statistical perspective on ill-posed inverse problems. *Statistical Science*, 1:502–527, 1986a.

- [OW85] F. O’Sullivan and G. Wahba. A cross validated Bayesian retrieval algorithm for non-linear remote sensing. *J. Comput. Physics*, 59:441–455, 1985.
- [Wah85] G. Wahba. Design criteria and eigensequence plots for satellite- computed tomography. *J. Atmos. Ocean. Tech.*, 2:125–132, 1985.

The Eta Theorem

- [Par70] E. Parzen. Statistical inference on time series by rkhs methods. In R. Pyke, editor, *Proceedings 12th Biennial Seminar*, Montreal, 1970. Canadian Mathematical Congress. 1-37.

The Histospline

- [DW82] N. Dyn and G. Wahba. On the estimation of functions of several variables from aggregated data. *SIAM J. Math. Anal.*, 13:134–152, 1982.
- [DWW79] N. Dyn, G. Wahba, and W. Wong. Comment on “Smooth pchnophylactic interpolation for geographical regions by W. Tobler. *J. Am. Statist. Assoc.*, 74(367):530–535, 1979.
- [Wah81] G. Wahba. Numerical experiments with the thin plate histospline. *Commun. Statist.-Theor. Meth.*, A10:2475–2514, 1981.

- [Amo97] L. Amodei. Reproducing kernels of vector-valued function spaces. In A. LeMehaute, C. Rabut, and L. Schumaker, editors, *Surface Fitting and Multiresolution Methods*, pages 17–26, Nashville TN, 1997. Vanderbilt University Press.
- [BCL96] A. Bennett, B. Chua, and L. Leslie. Generalized inversion of a global weather prediction model. *Meteorology and Atmospheric Physics*, 60:165–178, 1996.
- [KS85] C. Kravaris and J.H. Seinfeld. Identification of parameters in distributed parameter systems by regularization. *SIAM J. Control Opt.*, 23:217–241, 1985.
- [Utr85] F. Utreras. Smoothing noisy data under monotonicity constraints, existence, characterization and convergence rates. *Numer. Math.*, 47:611–625, 1985.
- [VW87] M. Villalobos and G. Wahba. Inequality constrained multivariate smoothing splines with application to the estimation of posterior probabilities. *J. Am. Statist. Assoc.*, 82:239–248, 1987.
- [Wah82] G. Wahba. Constrained regularization for ill posed linear operator equations, with applications in meteorology and medicine. In S. Gupta and J. Berger, editors, *Statistical Decision Theory and Related Topics, III, Vol.2*, pages 383–418. Academic Press, 1982.
- [Wah99] G. Wahba. Adaptive tuning, four dimensional variational data assimilation and representers in rkhs. In ECMWF,

editor, *Diagnosis of Data Assimilation Systems*, pages 45–52, Reading England, 1999. European Center for Medium Range Weather Prediction.

PART III

SS-ANOVA

- [BR98] B. Brumback and J. Rice. Smoothing spline models for the analysis of nested and crossed samples of curves. *J. Amer. Statist. Assoc.*, 93:961–991, 1998.
- [CGW89] Z. Chen, C. Gu, and G. Wahba. Comments to ‘Linear Smoothers and Additive Models’, by Buja, Hastie and Tibshirani. *Ann. Statist.*, 17:515–521, 1989.
- [GW91] C. Gu and G. Wahba. Comments to ‘Multivariate Adaptive Regression Splines’, by J. Friedman. *Ann. Statist.*, 19:115–123, 1991.
- [GW93a] C. Gu and G. Wahba. Semiparametric analysis of variance with tensor product thin plate splines. *J. Royal Statistical Soc. Ser. B*, 55:353–368, 1993.
- [GW93b] C. Gu and G. Wahba. Smoothing spline ANOVA with component-wise Bayesian “confidence intervals”. *J. Computational and Graphical Statistics*, 2:97–117, 1993.
- [GWKK99] F. Gao, G. Wahba, R. Klein, and B. Klein. Smoothing spline ANOVA for multivariate Bernoulli observations,

with applications to ophthalmology data. Technical Report 1009, Department of Statistics, University of Wisconsin, Madison WI, 1999.

- [VCKW99] A. Verbyla, B. Cullis, M. Kenward, and S. Welham. The analysis of designed experiments and longitudinal data using smoothing splines. *J. Roy. Stat. Soc. C*, 48:269–311, 1999.
- [Wan97] Y. Wang. GRKPACK: Fitting smoothing spline analysis of variance models to data from exponential families. *Commun. Statist. Simulation and Computation*, 26:765–782, 1997.
- [Wan98a] Y. Wang. Mixed-effects smoothing spline ANOVA. *J. Roy. Stat. Soc. B*, 60:159–174, 1998.
- [Wan98b] Y. Wang. Smoothing spline models with correlated random errors. *J. Amer. Statist. Assoc.*, 93:34–348, 1998.
- [Wan98c] Y. Wang. Smoothing spline models with correlated random errors. *J. Amer. Statist. Assoc.*, 93:341–348, 1998.
- [WB96] Y. Wang and M. Brown. A flexible model for human circadian rhythms. *Biometrics*, 52:588–596, 1996.
- [WW98] Y. Wang and G. Wahba. Comments to ‘Smoothing spline models for the analysis of nested and crossed samples of curves’ by B. Brumback and J. Rice. *J. Amer. Statist. Assoc.*, 93:976–980, 1998.
- [WWG⁺94] G. Wahba, Y. Wang, C. Gu, R. Klein, and B. Klein. Structured machine learning for ‘soft’ classification with smoothing spline ANOVA and stacked tuning, testing and evaluation. In J. Cowan, G. Tesauero, and

J. Alspector, editors, *Advances in Neural Information Processing Systems 6*, pages 415–422. Morgan Kaufman, 1994.

- [WWG⁺95] G. Wahba, Y. Wang, C. Gu, R. Klein, and B. Klein. Smoothing spline ANOVA for exponential families, with application to the Wisconsin Epidemiological Study of Diabetic Retinopathy. *Ann. Statist.*, 23:1865–1895, 1995. Neyman Lecture.
- [WWG⁺97] Y. Wang, G. Wahba, C. Gu, R. Klein, and B. Klein. Using smoothing spline ANOVA to examine the relation of risk factors to the incidence and progression of diabetic retinopathy. *Statistics in Medicine*, 16:1357–1376, 1997.

Tuning Non-Gaussian Models

- [Gu92] C. Gu. Cross-validating non-Gaussian data. *J. Comput. Graph. Stats.*, 1:169–179, 1992.
- [GWJT98] J. Gong, G. Wahba, D. Johnson, and J. Tribbia. Adaptive tuning of numerical weather prediction models: simultaneous estimation of weighting, smoothing and physical parameters. *Monthly Weather Review*, 125:210–231, 1998.
- [LWX⁺98] X. Lin, G. Wahba, D. Xiang, F. Gao, R. Klein, and B. Klein. Smoothing spline ANOVA models for large data sets with Bernoulli observations and the randomized GACV. Technical Report 998, Department of

Statistics, University of Wisconsin, Madison WI, tent. acc. *Ann. Statist.*, 1998.

- [WJGG95] G. Wahba, D. Johnson, F. Gao, and J. Gong. Adaptive tuning of numerical weather prediction models: randomized GCV in three and four dimensional data assimilation. *Mon. Wea. Rev.*, 123:3358–3369, 1995.
- [WLG⁺99] G. Wahba, X. Lin, F. Gao, D. Xiang, R. Klein, and B. Klein. The bias-variance tradeoff and the randomized GACV. In M. Kearns, S. Solla, and D. Cohn, editors, *Advances in Information Processing Systems 11*, pages 620–626. MIT Press, 1999.
- [XW96] D. Xiang and G. Wahba. A generalized approximate cross validation for smoothing splines with non-Gaussian data. *Statistica Sinica*, 6:675–692, 1996.

Randomized Trace Techniques

- [Gao99] F. Gao. Iterated ranGACV: a computational proxy for the comparative Kullback-Leibler distance. Technical Report 1011, Department of Statistics, University of Wisconsin, Madison WI, 1999.
- [Gir89] D. Girard. A fast ‘Monte-Carlo cross-validation’ procedure for large least squares problems with noisy data. *Numer. Math.*, 56:1–23, 1989.
- [Gir91] D. Girard. Asymptotic optimality of the fast randomized versions of GCV and C_L in ridge regression and regularization. *Ann. Statist.*, 19:1950–1963, 1991.

- [Gir98] D. Girard. Asymptotic comparison of (partial) cross-validation, GCV and randomized GCV in nonparametric regression. *Ann. Statist.*, 126:315–334, 1998.
- [Hut89] M. Hutchinson. A stochastic estimator for the trace of the influence matrix for Laplacian smoothing splines. *Commun. Statist.-Simula.*, 18:1059–1076, 1989.

Bayesian ‘Confidence Intervals’

- [GW93] C. Gu and G. Wahba. Smoothing spline ANOVA with component-wise Bayesian “confidence intervals”. *J. Computational and Graphical Statistics*, 2:97–117, 1993.
- [Wah83] G. Wahba. Bayesian “confidence intervals” for the cross-validated smoothing spline. *J. Roy. Stat. Soc. Ser. B*, 45:133–150, 1983.
- [WW95] Y. Wang and G. Wahba. Bootstrap confidence intervals for smoothing splines and their comparison to Bayesian ‘confidence intervals’. *J. Statist. Comput. Simul.*, 51:263–279, 1995.