IPAM Graduate Summer School: Intelligent Extraction of Information From Graphs and High Dimensional Data

Grace Wahba

Tutorial I: Reproducing Kernel Hilbert Spaces, and Why they Are Important

Overheads will be up via the TALKS link on my website: `http://www.stat.wisc.edu/~wahba`.
References in TALKS: JSM Wald Lectures 2003, Interface Short Course 2000
Tech reports up via the TRLIST link

July 14, 2005

I Reproducing Kernel Hilbert Spaces and Why They are Important

Abstract

We assume no previous knowledge of reproducing kernel Hilbert spaces (rkhs). We discuss the Moore-Aronszajn Theorem which gives a 1:1 correspondence between positive definite functions and rkhs, and note that through this relationship a positive definite function defines a distance measure with an inner product. In addition we identify the relationship with Bayes estimates. After describing several popular classes of rkhs we describe the related optimization problems which estimate functions or vectors via (Tihonov) regularization, where the regularization term is a norm or semi-norm in rkhs. We note a variety of cost functions, including those resulting in penalized likelihood estimates and support vector machines, and briefly mention the problem of tuning, or, the bias-variance tradeoff in function estimation, which balances fit to the data against complexity of the solution.

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 (up to a point.)

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

9. Are the foundation for so-called ”kernel methods” which have become immensely popular as part of the popularity of support vector machines for classification.

Outline Part I

1. Positive Definite Functions

2. Bayes Estimates,Variational Problems and Positive Definite Functions

3. Reproducing Kernel Hilbert Spaces

4. The Moore-Aronszajn Theorem and Inner Products in RKHS, Radial Basis Functions.

5. The Representer Theorem (simple case)

6. The Polynomial Smoothing Splines, Tuning

7. Varieties of Cost Functions and Tuning Methods

♣♣ Positive definite matrices and functions.

The concept of positive definite functions is key, so we begin by reviewing it.

Let $\mathcal{T}$ be an index set. A symmetric function of two variables, $K(s,t), s,t \in \mathcal{T}$ is said to be positive definite (pd) if, for every n and $t_1, \cdots, t_n \in \mathcal{T}$, and every $a_1, \cdots, a_n$,

$$\sum_{i,j=1}^{n} a_i a_j K(t_i, t_j) \geq 0.$$

In the case $\mathcal{T} = \{1, 2, \cdots, N\}$ K reduces to an $N \times N$ matrix. But we will be interested in a (limitless) variety of other index sets-anything on which you can construct a positive definite function:

Some useful domains for positive definite functions:

$\mathcal{T} = (\ldots, -1, 0, 1, \ldots)$

$\mathcal{T} = [0, 1]$

$\mathcal{T} = E^d$ (Euclidean *d*-space)

$\mathcal{T} = \mathcal{S}$ (the unit sphere)

$\mathcal{T} =$ the atmosphere

$\mathcal{T} = \{\diamondsuit, \triangle, \heartsuit\}$ (unordered set)

$\mathcal{T} =$ A Riemannian manifold

$\mathcal{T} =$ A collection of trees

$\mathcal{T} =$ A collection of graphs

$\mathcal{T} =$ A collection of proteins

$\mathcal{T} =$ A collection of gene microarray chips

... A positive definite function (kernel) defined on $\mathcal{T} \otimes \mathcal{T}$ defines a metric on a certain class of functions defined on $\mathcal{T}$ possessing an inner product. Thus allowing solutions of optimization problems, clustering and classification.

♣♣ Bayes Estimates, Variational Problems and Positive Definite Functions

(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

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

- 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

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

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

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

- 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

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

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:

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

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

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 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) general case:

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

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

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

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

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 $\mathcal{H}_K$.

♣♣ Radial Basis functions.

Radial basis functions (rbf's) are obtained from positive definite functions defined on $E^d \otimes E^d$ which depend only on the Euclidean destance between s and t.

For $s, t \in E^d$ let

$$K(s,t) = \mathcal{F}(\Lambda) = \int_{E^d} \Lambda(d\omega) e^{i\omega \cdot s} e^{-i\omega \cdot t}$$

If Λ is symmetric and non-negative, K will be positive definite. If Λ depends only on $\|\omega\|$ then $K(s,t)$ will depend only on $\|s-t\|$ and will generate rbf's.

The most commonly used kernel of this form is the Gaussian: $K(s,t) = e^{-\frac{1}{\sigma^2}\|s-t\|^2}$. Letting $K(s,t) = q(\|s-t\|)$, for the Gaussian kernel positive definiteness follows from

$$q(\tau) = e^{-\tau^2/\sigma^2} = \mathcal{F}(ce^{-k\|\omega\|^2})$$

In general, the RKHS norm of $f \in \mathcal{H}$ is given by

$$\|f\|_{\mathcal{H}}^2 = \int_{E^d} \Lambda(\omega)^{-1} |\tilde{f}(\omega)| d\omega_1, \cdots, d\omega_d$$

where $\tilde{f}$ is the Fourier transform of f. The Matern class is also useful in some applications:

$q(\tau) = \mathcal{F}(\|\omega\|^2 + \alpha^2)^{-(d+1)/2+m)}), d, m = 1, 2, ..$

Some members of the Matern Class:

$$\begin{aligned}
q(\tau) &\sim \Lambda(\omega) \\
\frac{1}{\alpha} e^{-\alpha\tau} &\sim (\|\omega\|^2 + \alpha^2)^{-(\frac{d+1}{2})}, m = 0 \\
\frac{1}{\alpha^3} e^{-\alpha\tau}[1 + \alpha\tau] &\sim (\|\omega\|^2 + \alpha^2)^{-(\frac{d+3}{2})}, m = 1 \\
\frac{1}{\alpha^5} e^{-\alpha\tau}[3 + 3\alpha\tau + \alpha^2\tau^2] &\sim (\|\omega\|^2 + \alpha^2)^{-(\frac{d+5}{2})}, m = 2 \\
\cdots &\sim \cdots
\end{aligned}$$

m indexes the number of derivatives/continuity properties of the rbf's $K(t*, \cdot)$ at the origin.

♣♣ The Representer Theorem (simple case)

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

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

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 $f \in \mathcal{H}_K$, writing $K_{t(i)}(\cdot)$ for $K(t(i), \cdot)$, then we can always write

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

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 $\|\rho\|^2_{\mathcal{H}_K} = 0$.

♣♣ 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

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

Letting $k_\nu(t) = B_\nu(t)/m!$, (B_ν are Bernoulli polynomials) it can be shown that the solution f_λ satisfies

$$f_\lambda = \sum_{\nu=0}^{m-1} d_\nu k_\nu + f_1$$

where $f_1 \in \mathcal{H}_K$ with the reproducing kernel

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

and square norm $\|f\|^2_{\mathcal{H}_K} = \int_0^1 (f^{(m)}(u))^2 du$.

By an argument generalizing the representer theorem, and upon observing that $\{k_\nu\}_{\nu=0}^{m-1}$ are not penalized, it follows that the minimizer f_λ has the form

$$f_\lambda(t) = \sum_{\nu=1}^{m} d_\nu k_{\nu-1}(t) + \sum_{i=1}^{n} c_i K(t(i), t), \quad (3)$$

and that

$$\int_0^1 (f^{(m)}(u))^2 du = \sum_{i,j=1}^{n} c_i c_j K(t(i), t(j)). \quad (4)$$

Upon substituting (3) and (4) into the original variational problem, the solution is obtained by minimizing a quadratic form in $d = (d_1, \cdots, d_m)'$ and $c = (c_1, \cdots, c_n)'$. (There are easier ways to get the polynomial spline.)

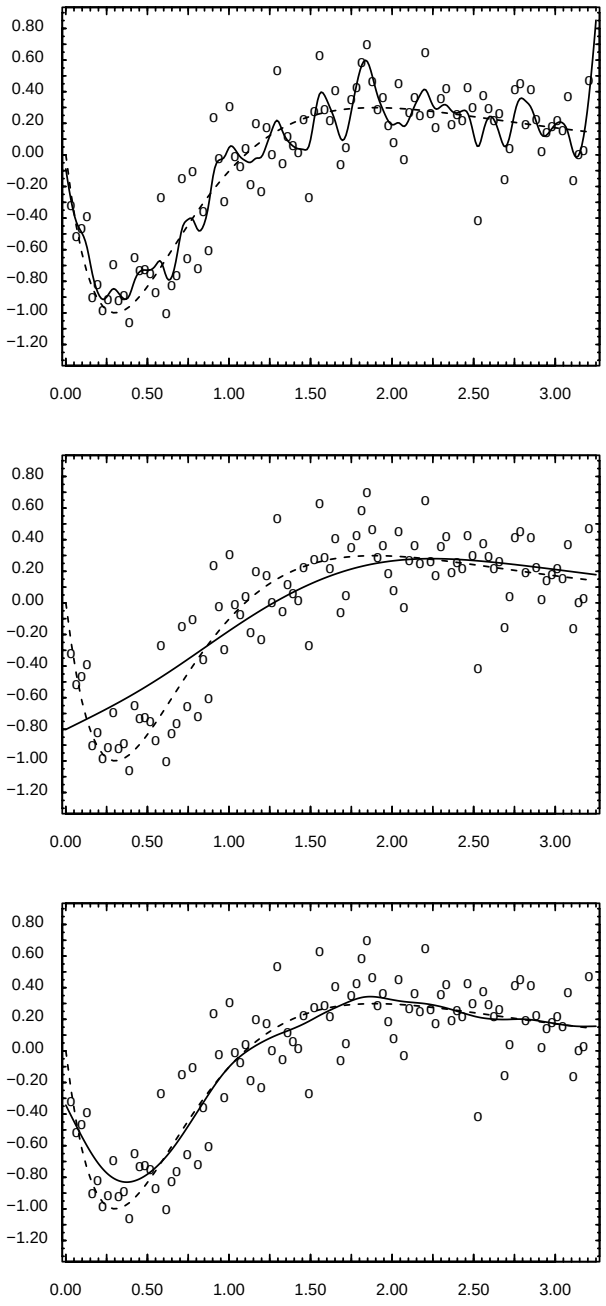

Figure 1: Dashed Lines are Smoothing Splines with λ too small, λ too large, and λ estimated via GCV, from the top. Solid line is[1]'truth'.

♣♣ Varieties of Cost Functions and Tuning Methods.

	Cost Function $C(y, f)$
Regression	
.....	
Gaussian data	$(y-f)^2$
Bernoulli, $f = log[p/(1-p)]$	$-yf + log(1+e^f)$
Other exponential families	other log likelihoods
Data with outliers	robust functionals
Quantile functionals	$\rho_q(y-f)$
.....	
Classification: $y \in \{-1, 1\}$	
.....	
Support vector machines	$(1-yf)_+$
Other "large margin classifiers"	e^{-yf} and other functions of (yf) including $log(1+e^{-yf})$
.....	

(Here $(\tau)_+ = \tau, \tau \geq 0, = 0$ otherwise.
$\rho_q(\tau) = \tau(q - I(\tau \leq 0))$.

♣♣ Varieties of cost functions and tuning methods (continued).

Tuning methods for choosing λ from the data:

- Gaussian Data: Generalized Cross Validation (GCV), Generalized Maximum Likelihood (GML)(aka REML), Unbiassed risk (UBR), others (`google` "methods" "choose" "smoothing parameter" gave 15,000 hits)

- Bernoulli Data: Generalized Approximate Cross Validation (GACV), other earlier related

- Support Vector Machines: GACV for SVM's other related, esp. Joachim's $\xi\alpha$ method.

- All problems: Leaving-out-one, k-fold cross validation

♣♣ 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.