An Introduction to Reproducing Kernel Hilbert Spaces and Why They are So Useful

Grace Wahba
Department of Statistics
University of Wisconsin-Madison
`http://www.stat.wisc.edu/~wahba`
$\longrightarrow$ *TRLIST*

SYSID 2003
Rotterdam
August 27, 2003

We review some of the basic facts about reproducing kernel Hilbert spaces (RKHS), and the solution of various optimization problems of interest in them.

OUTLINE

1. What is an RKHS?

2. The Moore Aronszajn Theorem.

3. Gaussian Processes.

4. More RKHS.

5. The representer theorem.

6. Varieties of cost functions. (Univariate case).

7. The bias-variance tradeoff and adaptive tuning.

8. Methods for choosing λ from the data.

9. Concluding remarks.

References

Akhiezer, N. & Glazman, I. (1963), *Theory of Linear Operators in Hilbert Space*, Ungar, New York.

Aronszajn, N. (1950), ‘Theory of reproducing kernels’, *Trans. Am. Math. Soc.* **68**, 337–404.

Craven, P. & Wahba, G. (1979), ‘Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation’, *Numer. Math.* **31**, 377–403.

Gao, F., Wahba, G., Klein, R. & Klein, B. (2001), ‘Smoothing spline ANOVA for multivariate Bernoulli observations, with applications to ophthalmology data, with discussion’, *J. Amer. Statist. Assoc.* **96**, 127–160.

Golub, G., Heath, M. & Wahba, G. (1979), ‘Generalized cross validation as a method for choosing a good ridge parameter’, *Technometrics* **21**, 215–224.

Gu, C. & Wahba, G. (1991), ‘Minimizing GCV/GML scores with multiple smoothing parameters via the Newton method’, *SIAM J. Sci. Statist. Comput.* **12**, 383–398.

Halmos, P. (1957), *Introduction to Hilbert Space and the Theory of Spectral Multiplicity*, Chelsea, New York.

Kimeldorf, G. & Wahba, G. (1971), ‘Some results on Tchebycheffian spline functions’, *J. Math. Anal. Applic.* **33**, 82–95.

Lee, Y., Lin, Y. & Wahba, G. (2002), Multicategory support vector machines, theory, and application to the classification of microarray data and satellite radiance data, Technical Report 1064, Department of Statistics, University of Wisconsin, Madison WI.

Lin, X. (1998), Smoothing spline analysis of variance for polychotomous response data, Technical Report 1003, PhD thesis, Department of Statistics, University

of Wisconsin, Madison WI.
Available via G. Wahba's website.

Lin, X., Wahba, G., Xiang, D., Gao, F., Klein, R. & Klein, B. (2000), 'Smoothing spline ANOVA models for large data sets with Bernoulli observations and the randomized GACV', *Ann. Statist.* **28**, 1570–1600.

Micchelli, C. (1986), 'Interpolation of scattered data: distance matrices and conditionally positive definite functions', *Constructive Approximation* **2**, 11–22.

Nychka, D. (1988), 'Bayesian confidence intervals for smoothing splines', *J. Amer. Statist. Assoc.* **83**, 1134–1143.

Nychka, D., Wahba, G., Goldfarb, S. & Pugh, T. (1984), '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.

O’Sullivan, F. & Wahba, G. (1985), ‘A cross validated Bayesian retrieval algorithm for non-linear remote sensing’, *J. Comput. Physics* **59**, 441–455.

Parzen, E. (1970), Statistical inference on time series by rkhs methods, *in* R. Pyke, ed., ‘Proceedings 12th Biennial Seminar’, Canadian Mathematical Congress, Montreal. 1-37.

Riesz, F. & Nagy, B. S. (1955), *Functional Analysis*, Ungar, New York.

Schoenberg, I. (1964a), ‘Spline functions and the problem of graduation’, *Proc. Nat. Acad. Sci. U.S.A.* **52**, 947–950.

Tikhonov, A. (1963), ‘Solution of incorrectly formulated problems and the regularization method’, *Soviet Math. Dokl.* **4**, 1035–1038.

Wahba, G. (1977a), ‘Practical approximate solutions to linear operator equations when the data are noisy’, *SIAM J. Numer. Anal.* **14**, 651–667.

Wahba, G. (1981), ‘Spline interpolation and smoothing on the sphere’, *SIAM J. Sci. Stat. Comput.* **2**, 5–16.

Wahba, G. (1982), ‘Erratum: Spline interpolation and smoothing on the sphere’, *SIAM J. Sci. Stat. Comput.* **3**, 385–386.

Wahba, G. (1982b), Vector splines on the sphere, with application to the estimation of vorticity and divergence from discrete, noisy data, *in* W. Schempp & K. Zeller, eds, ‘Multivariate Approximation Theory, Vol.2’, Birkhauser Verlag, pp. 407–429.

Wahba, G. (1983), ‘Bayesian “confidence intervals” for the cross-validated smoothing spline’, *J. Roy. Stat. Soc. Ser. B* **45**, 133–150.

Wahba, G. (1990), *Spline Models for Observational Data*, SIAM. CBMS-NSF Regional Conference Series in Applied Mathematics, v. 59.

Wahba, G. (1992), Multivariate function and operator estimation, based on smoothing splines and reproducing kernels, *in* M. Casdagli & S. Eubank, eds, 'Nonlinear Modeling and Forecasting, SFI Studies in the Sciences of Complexity, Proc. Vol XII', Addison-Wesley, pp. 95–112.

Wahba, G. (1996), 'NIPS 1996 model complexity workshop notes', lecture overheads. Available via `http://www.st`
goto TALKKS goto 1996.

Wahba, G. (1999), Support vector machines, reproducing kernel Hilbert spaces and the randomized GACV, *in* B. Scholkopf, C. Burges & A. Smola, eds, 'Advances in Kernel Methods-Support Vector Learning', MIT Press, pp. 69–88.

Wahba, G. (2002), ‘Soft and hard classification by reproducing kernel Hilbert space methods’, *Proc.National Academy of Sciences* **99**, 16524–16530.

Wahba, G. & Wang, Y. (1990), ‘When is the optimal regularization parameter insensitive to the choice of the loss function?’, *Commun. Statist.-Theory Meth.* **19**, 1685–1700.

Wahba, G. & Wendelberger, J. (1980), ‘Some new mathematical methods for variational objective analysis using splines and cross-validation’, *Monthly Weather Review* **108**, 1122–1145.

Wahba, G., Wang, Y., Gu, C., Klein, R. & Klein, B. (1995), ‘Smoothing spline ANOVA for exponential families, with application to the Wisconsin Epidemiological Study of Diabetic Retinopathy’, *Ann. Statist.* **23**, 1865–1895.

Wang, Y. & Wahba, G. (1994), Bootstrap confidence intervals for smoothing splines and their comparison to Bayesian ‘confidence intervals’, Technical Report 913, Dept. of Statistics, University of Wisconsin, Madison, WI, to appear, *J. Stat. Comp. Sim.*

Xiang, D. & Wahba, G. (1996), ‘A generalized approximate cross validation for smoothing splines with non-Gaussian data’, *Statistica Sinica* **6**, 675–692.

♣♣ 1. What is an RKHS?

An RKHS is a Hilbert space (Akhiezer and Glazman:1963) in which all the point evaluations are bounded linear functionals. (Unlike $\mathcal{L}_2$.) Letting $\mathcal{H}$ be a Hilbert space of functions on some domain $\mathcal{T}$, this means, that for every $t \in \mathcal{T}$ there exists an element $\eta_t \in \mathcal{H}$, such that

$$f(t) = < \eta_t, f >, \quad \forall f \in \mathcal{H},$$

where $<,>$ is the inner product in $\mathcal{H}$. Let $< \eta_s, \eta_t > = K(s,t)$. Then $K(s,t)$ is positive definite on $\mathcal{T} \otimes \mathcal{T}$, that is, for $\forall t_1, \cdots, t_n \in \mathcal{T}$, $\sum_{i,j} a_i a_j K(t_i, t_j) \geq 0$. K is called the reproducing kernel (RK) for $\mathcal{H}$, and η_t is the "representer of evaluation" at t. Since $\eta_t \equiv K(t, \cdot)$, then $< K(t,\cdot), K(s,\cdot) > \equiv K(s,t)$, this being the origin of the term "reproducing kernel".

♣♣ 2. The Moore-Aronszajn Theorem

The Moore-Aronszajn theorem (Aronszajn:1950) theorem states that for every positive definite function $K(\cdot,\cdot)$ on $\mathcal{T}\otimes\mathcal{T}$, there exists a unique RKHS and vice versa. The Hilbert space associated with K can be constructed as containing all finite linear combinations of the form $\sum a_j K(t_j,\cdot)$, and their limits under the norm induced by the inner product $< K(s,\cdot), K(t,\cdot >= K(s,t)$. Norm convergence implies pointwise convergence in a RKHS, as can be seen by observing that

$$|f_n(t) - f_m(t)| = | < K(t,\cdot), f_n - f_m > |$$

$$\leq K(t,t)\|f_n - f_m\|.$$

Thus, these limit functions are well defined pointwise. Nothing has been said about $\mathcal{T}$. The discussion above applies to any domain on which it is possible to define a positive definite function, a matrix being a special case when $\mathcal{T}$ has only a countable or finite number of points.

♣♣ 3. Gaussian Processes.

Note that, for every positive definite $K(\cdot,\cdot)$ on $\mathcal{T} \otimes \mathcal{T}$ there exists a zero mean Gaussian process with K as its covariance. Thus, there is a relation between Bayes estimates, Gaussian processes and optimization problems in RKHS. See Parzen:1970, Kimeldorf and Wahba:1971, Wahba:1990 and elsewhere.

♣♣ 4. More RKHS

Tensor sums and products of RK's are RK's, which allow construction of all sorts of spaces (Smoothing Spline ANOVA spaces as an example Wahba:1990). Letting $s_1, t_1 \in \mathcal{T}^{(1)}$, $s_2, t_2 \in \mathcal{T}^{(2)}$, and letting $s = (s_1, s_2), t = (t_1, t_2)$, then

$$K(s,t) = K_1(s_1,t_1)K(s_2,t_2)$$

is an RK on $\mathcal{T} = \mathcal{T}^{(1)} \otimes \mathcal{T}^{(2)}$ whenever K_1 and K_2 are RK's on their respective domains. Subspaces of RKHS are also RKHS, and the RK for a subspace can be obtained by e. g. projecting the representers of evaluation in $\mathcal{H}$ onto the subspace.

♣♣ 5. The Representer Theorem.

A special but important case of the representer theorem (Kimeldorf:Wahba:1971) is:
The solution to the problem: Find $f \in \mathcal{H}$ to minimize

$$\sum_{i=1}^{n} \mathcal{C}(y_i, f(t_i)) + \lambda \|f\|^2 \qquad (1)$$

where $\mathcal{C}$ is convex in f, has a representation as

$$f_\lambda(\cdot) = \sum_{i=1}^{n} c_i K(t_i, \cdot). \qquad (2)$$

Then (2) is substituted in (1) and the c_i's are found numerically. When $\mathcal{C}$ is quadratic, it is only necessary to solve a linear system, but otherwise a descent algorithm is used. The general form includes unpenalized (low-dimensional) subspaces, different λ's applied to different subspaces, and other generalizations.

♣♣ 5. The Representer Theorem (continued).

If we replace $f(t_i)$ by $L_i f$, where L_i is some bounded linear functional in the RKHS in

$$\sum_{i=1}^{n} \mathcal{C}(y_i, f(t_i)) + \lambda \|f\|^2$$

then the minimizer has a representation of the form

$$f_\lambda(\cdot) = \sum_{i=1}^{n} c_i \eta_i(\cdot)$$

where η_i is the representer of L_i. An important example is: let

$$y_i = \int H(t_i, u) f(u) du + \epsilon_i$$

where the ϵ_i are i.i.d Gaussian random variables. In this case $\mathcal{C}$ would correspond to least squares. Under appropriate regularity conditions,

$$L_i f = \int H(t_i, u) f(u) du,$$

$$\eta_i(s) = \int H(t_i, u) K(u, s) du.$$

and

$$\|f\|^2 = \sum_{i,j} c_i c_j < \eta_i, \eta_j >$$

where

$$< \eta_i, \eta_j > = \int\int H(t_i, u) H(t_j, v) K(u, v) du dv.$$

This setup is a generalized version of Tikhonov regularization (Tikhonov:1963, Wahba:1977a, O'Sullivan:Wahba:1985, Nychka:Wahba:Goldfarb:Pugh:1984)

♣♣ 6. Varieties of Cost Functions (Univariate Case).

	$\mathcal{C}(y, f)$
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}, $\log(1 + e^{-yf})$, $(1 - yf)^2$ and numerous other functions of (yf)
....	
(MV) Density estimation: $y \equiv 1$	$-yf + \int e^f$

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

♣♣ 7. The bias-variance tradeoff and adaptive tuning.

The parameter λ controls the tradeoff between the size of $\sum_{i=1}^{n} \mathcal{C}(y_i, f(t_i))$ and the size of $\|f\|^2$ in

$$\sum_{i=1}^{n} \mathcal{C}(y_i, f(t_i)) + \lambda \|f\|^2.$$

More generally there may be other so-called tuning parameters (such as σ in the Gaussian reproducing kernel), or, different λ's penalizing components in different subspaces differently.

Choosing λ reasonably well is usually important.

♣♣ 8. 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” (see Wahba:1990). ”choose” ”smoothing parameter” gave 2850 hits)

- Bernoulli Data: Generalized Approximate Cross Validation (GACV) (Xiang:Wahba:96),other earlier related

- Support Vector Machines: GACV for SVM’s (Wahba:Lin:Zhang:00) other related, esp. Joachim’s $\xi\alpha$ method.

- Multivariate Density Estimation: GACV for density estimation. (Wahba:Lin:Leng:02)

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

♣♣ 9. Concluding remarks.

Methods for model building, regression and classification by solving optimization problems in RKHS are an important tool for the Engineer, Computer Scientist and Statistician.