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

NIPS 96 MODEL COMPLEXITY WORKSHOP NOTES

December 2 1996

RBF's, SBF's, TreeBF's, Smoothing Spline ANOVA: Representers and pseudo-representers for a dictionary of basis functions for penalized likelihood estimates [1]

by

Grace Wahba

[1]These draft notes were prepared on December 2 as handouts/overheads for G. Wahba's talk at the Neural Information Processing Society Workshop on Model Complexity, December 6, 1996, Snowmass CO, organized by Chris Williams and Joachim Utans. Permission to quote with attribution granted. Some minor typos and additions based on the actual talk have been added on December 9.

ABSTRACT

This work in progress represents an attempt to combine radial basis functions (RBF's), sigmoidal basis functions (SBF's) and basis functions that may be useful in conjunction with tree-structured methods (TreeBF's) under a single 'umbrella' of a reproducing kernel Hilbert space. Once this is done, several ways of generating a 'list' of basis functions in which to solve a penalized likelihood problem suggest themselves. Support vector methods may be used to refine the list. Given such a list, regularized forward selection methods generalizing those suggested by Orr and by Luo and Wahba may be used to fit the model.

Large to very large data sets are assumed ($n > 1000$). It is envisioned that the approach could prove useful in building models where more than three or four but less than, say ten or fifteen predictor variables are involved, and that the umbrella provides some intuition concerning how the basis functions are related and what they are doing, so as to give some interpretability to models built from them. Also, some intuition may be provided as to how to parametrize the basis functions so that the optimization problems to be solved numerically to obtain the fit are well conditioned in some sense. A 'super-umbrella' which also includes smoothing spline ANOVA models, can also be constructed. Although we are not discussing context here we remind the listener that the scientific context in which model building takes place should not be ignored.

OUTLINE

- Describe RBF's in a reproducing kernel Hilbert space (RKHS) setting. The Bayes-RKHS duality. Examples.
- Put TreeBF's and Sigmoidal BF's (SBF's) in an RKHS setting. Representers and pseudo-representers. Scaling.
- Creating the basis list for RBF's, TreeBF's and SBF's.
- Regularized forward basis function selection.
- Smoothing Spline ANOVA, 0-1 data.

RBF's

$$
\begin{aligned}
s, t &\in E^d \\
R(s,t) &\equiv \text{isotropic covariance} \\
&\equiv \text{symmetric positive definite function on} \\
&\quad E^d \times E^d \text{ which only depends on } \|s-t\|. \\
&\quad \text{(special case of positive definite functions)}
\end{aligned}
$$

$R_k(t) \equiv R(t, t(k))$ is an RBF with center $t(k)$.

Example: $R(s,t) = e^{-\|s-t\|^\beta}$, $0 < \beta \leq 2$

$$
\begin{aligned}
\beta &= 1 - \text{negative exponential} \\
\beta &= 2 - \text{gaussian}
\end{aligned}
$$

Example: $R(s,t) = e^{-\|s-t\|} P_q(\|s-t\|), q = 0, 1, 2, ..$

$$
\begin{aligned}
P_0(\tau) &= 1 \quad - \text{negative exponential} \\
P_1(\tau) &= 1 + \tau \\
P_2(\tau) &= 3 + 3\tau + \tau^2 \\
\cdots &\quad \cdots
\end{aligned}
$$

Remark: Scale factors: $R(s,t) \rightarrow R_\alpha(s,t) = R(\alpha\|s-t\|)$.

THE BAYES-RKHS DUALITY

BAYES: $Z(t), t \in E^t$, zero-mean gaussian, $EZ(s)Z(t) = R(s,t)$.

RKHS: $\mathcal{H}_R$, RKHS with reproducing kernel R.

(Duality: Parzen 1962, Kimeldorf & Wahba 1971, Wahba, 1990)

BAYES MODEL

$$
\begin{aligned}
y_i &= f(t(i)) + \epsilon_i, i = 1, \cdots, n, \epsilon' \sim \mathcal{N}(0, \sigma^2 I). \\
\text{Let } \; Ef(s)f(t) &= bR(s,t) \\
E\{f(t)|y_1, \cdots y_n\} &= f_\lambda(t)
\end{aligned}
$$

where

$$
\begin{aligned}
f_\lambda(t) &= \sum_{i=1}^{n} c_i R(t, t(i)) \\
c &= (\Sigma_n + \lambda I)^{-1} \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \lambda = \sigma^2/nb \\
\text{and} & \\
\{\Sigma_n\}_{i,j} &= \{R(t(i), t(j))\}
\end{aligned}
$$

Important **Remark:** If R is isotropic then f_λ is a linear combination of RBF's with centers at the data points.

PENALIZED LIKELIHOOD, OR, THE VARIATIONAL MODEL, OR, REGULARIZATION

Theorem (Special case of Kimeldorf & Wahba 1971, see Wahba, 1990). Let $\|f\|_R$ be the norm in $\mathcal{H}_R$. Then:

The solution to the problem: find $f \in \mathcal{H}_R$ to minimize

$$\sum_{i=1}^{n}(y_i - f(t(i)))^2 + \lambda\|f\|_R^2 \qquad (*)$$

is f_λ of the previous slide, a linear combination of RBF's with centers at the data points.

Important **Remark**: for large n a good approximation to the minimizer of $(*)$ can be obtained by finding f in $span\{R_{i_1}(t), \cdots R_{i_K}(t)\}$, a subset of the RBF's, to minimize $(*)$. (Wahba 1980)

$$i_1, \cdots, i_K = \text{"centers"}.$$

Methods to choose centers:

random subsets, stratified subsets (Hutchinson 1984 ...)
stratified subsets (O'Sullivan 1990 ....)
clustering (Xiang 1995 ...)
.....
forward selection with GCV stopping (Friedman 1991)
two stage forward selection (Luo and Wahba 1996)
first order regularized forward selection (Orr 1993)
general regularized forward selection (Diaz 1995)

The methods above the dotted line do not use the responses y_i whereas those below do.

WHAT DOES $\|f\|_R^2$ LOOK LIKE?

Example: Let

$$\begin{aligned} R(s,t) &= \frac{1}{\alpha^5} e^{-\alpha\|s-t\|}(3 + 3\alpha\|s-t\| + \alpha^2\|s-t\|^2) \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{i(s-t)\cdot\omega}(\|\omega\|^2 + \alpha^2)^{-\frac{d+5}{2}} d\omega_1 \cdots d\omega_d \end{aligned}$$

Then, it can be shown that

$$\|f\|_R^2 = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} (\|\omega\|^2 + \alpha^2)^{\frac{d+5}{2}} |\tilde{f}(\omega)|^2 d\omega_1 \cdots d\omega_d$$

where $\tilde{f}$ is the Fourier transform of f. To understand this formula note that, if $\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \|\omega\|^{2m} |\tilde{f}(\omega)|^2 d\omega_1 \cdots d\omega_d$ is finite, then it is equal to to the thin plate spline penalty functional $J_m(f)$:

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \|\omega\|^{2m} |\tilde{f}(\omega)|^2 d\omega_1 \cdots d\omega_d =$$

$$\sum_{i_1=1}^{d} \cdots \sum_{i_m=1}^{d} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left(\frac{\partial^m f}{\partial x_{i_1} \cdots \partial x_{i_m}}\right)^2 d\omega_1 \cdots d\omega_d = J_m(f).$$

The right hand side is the square integral of the total derivative of order m. Moody and Rognvaldsson 1996 discuss the use of and approximations to this kind of penalty functional.

REPRESENTERS OF INTEGRATION OVER Ω_k

Let Ω_k be a bounded region in E^d and suppose

$$z_k = \int \cdots \int_{\Omega_k} f(u)du + \epsilon_k.$$

(KW 1971) The minimizer $f_\lambda \in \mathcal{H}_R$ of

$$\sum_{k=1}^{K} (z_k - \int \cdots \int_{\Omega_k} f(u)du)^2 + \lambda \|f\|_R^2$$

is

$$f_\lambda(t) = \sum_{k=1}^{K} c_k \eta_k(t)$$

where η_k is the **representer** of integration over Ω_k in $\mathcal{H}_R$, that is,

$$< \eta_k, f >= \int \cdots \int_{\Omega_k} f(u)du, \quad \text{all } f \in \mathcal{H}_R.$$

η_k is obtained from the RK as

$$\eta_k(t) = \int \cdots \int_{\Omega_k} R(t,u)du$$

and $c = (\Sigma_K + \lambda I)^{-1} \begin{pmatrix} z_1 \\ \vdots \\ z_K \end{pmatrix}$, where

$$\{\Sigma_K\}_{j,k} = \{< \eta_j, \eta_k >\} = \int \cdots \int_{\Omega_j} du \int \cdots \int_{\Omega_k} dv R(u,v).$$

The message here is the role of **representers** in an RKHS, how they are obtained from the RK and how the inner product between two of them is obtained from the RK. Note that $\|f\|_R^2 = c'\Sigma_K c$.

TreeBF's

Assume the predictor variables $t(i), i = 1, \cdots n$ have been scaled to sit in $[0,1]^d$. Generate a tree. This results in a partition of $[0,1]^d$ into $\{\Omega_1, \cdots, \Omega_K\}$ where the boundaries of Ω_k are parallel to the coordinate axes. A common way of estimating via a tree is to let

$$\begin{aligned} I_k(s) &= 1, s \in \Omega_k \\ I_k(s) &= 0, s \notin \Omega_k \end{aligned}$$

and let

$$f(t) = \sum_{k=1}^{K} c_k I_k(t).$$

Then f is obtained by finding c to minimize

$$\sum_{i=1}^{K} (y_i - \sum_{k=1}^{K} c_k I_k(t(i)))^2.$$

(giving $c_k = \frac{1}{n_k} \Sigma_{y_i \in \Omega_k} y_i$). A **regularized tree** may be found by firstly, overgrowing the tree somewhat, since it will be smoothed, and then, finding $f_\lambda = \Sigma_{k=1}^{K} c_k \eta_k$, where the η_k are as before, to minimize

$$\sum_{i=1}^{n} (y_i - \sum_{k=1}^{K} c_k \eta_k(t(i))^2 + \lambda \| f \|_R^2.$$

If Ω_k is a rectangle with boundaries parallel to the coordinate axes, we will call η_k a TreeBF. TreeBF's and RBF's may be included in the same list of basis functions by considering

$$f_\lambda = \sum_k c_k^{TreeBF} \eta_k + \sum_\ell c_\ell^{RBF} R_{i_\ell}. \qquad (R_{i_\ell}(t) = R(t, t_{i_\ell}))$$

$\| f \|_R^2$ is readily found by using the fact that $< \eta_k, R_{i_\ell} >= \eta_k(t(i_\ell)$.

SBF's AND THE PSEUDO-PUNCH LINE

Let $\gamma_j \in E^d, \|\gamma_j\| = 1$.
Consider

$$L_j f = \int \cdots \int_{\gamma_j' u \le b_j} f(u)du \quad \text{Integration over a half-space.}$$

and

$$\sigma_j(s) = \int \cdots \int_{\gamma_j' u \le b_j} R(s,u)du \quad \text{Looks like the representer of } L_j.$$

If $R(s,t) = r(\|s-t\|)$, let

$$\sigma(\tau) = \int_{-\infty}^{\tau} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} r(\sum_{\nu=1}^{d} v_\nu^2) dv_1 \cdots dv_d.$$

Then

$$\sigma_j(s) = \sigma(b_j - \gamma_j' s). \qquad \text{AN SBF}$$

THE PSEUDO-PUNCH LINE: Consider

$$f_\lambda = \sum_j c_j^{SBF} \sigma_j + \sum_k c_k^{TreeBF} \eta_k + \sum_\ell c_\ell^{RBF} R_{i_\ell},$$

carry on, computing $\|f_\lambda\|_R^2$ as before, given all the inner products. Unfortunately since the region of integration for the σ_j is infinite, $< \sigma_j, \sigma_j >$ is not finite. For the Bayesians among us, this is the mathematical equivalent of observing that (since $Z(t)$ is stationary), $E[\int \cdots \int_{\gamma_j' u \le b_j} Z(u)du]^2$ is infinite. On the other hand all the cross-inner products between a σ_j and an η_k or an R_{i_ℓ} are well defined, corresponding to, for example, the finiteness of

$$E[\int \cdots \int_{\gamma_j' u \le b_j} Z(u)du][\int \cdots \int_{\Omega_k} Z(v)dv] \equiv < \sigma_j, \eta_k > .$$

For this reason we call the σ_j's **pseudo-representers**.

PATCHING THINGS UP

Since we don't really care about about what happens at infinity, we should be able to modify things to result in a set of inner products or similar objects for the pseudo-representers which will result in reasonable penalty functionals. Two mathematically 'kosher' ways of patching things up but may be messy in practice are:

- Let $w(t)$ be a positive real valued function on E^d which is 1 on a region $\Omega \subset E^d$ containing all the t of interest, and satisfies $\int \cdots \int_{E^d} \int \cdots \int_{E^d} R(u,v)w(u)w(v)dudu < \infty$. Then replace $R(u,v)$ by $R(u,v)w(u)w(v)$ in the calculation of the inner products.
- Replace the penalty $\|f\|_R^2$ by $\|Pf\|_R^2$, where P is the orthogonal projection operator onto some subspace of $\mathcal{H}_R$ contained in $span\{R_t, t \in \Omega\}$. This will entail that the penalty only involves the values of f in Ω. One such space is $span\{R_{t(i)}, i = 1, \cdots, n\}$.

A simple and appealing way of determining a penalty functional is to replace each representer of integration over a region by the representer of averaging over the data points (or some subset of them) in the region, when calculating the penalty. Thus if there are n_j observations in the half space $\gamma_j' t \leq b_j$, then σ_j is replaced by $\frac{1}{n_j} \Sigma_{\{t(i):\gamma_j' t(i) \leq bj\}} R_{t(i)}$ in the definition of the penalty functional.

Remark: The constraint $\|\gamma_j\| = 1$ is meaningful. Usually sigmoidal functions do not have this restriction. $\sigma(b_j - \gamma_j' s)$ may be replaced by $\sigma(\alpha \times (b_j - \gamma_j' s))$, but α should be treated as a scale factor.

CREATING THE BASIS LIST

With regard to RBF's the list of data points $\{t(i), i = 1, \cdots, n\}$ is an appropriate upper bound to the list of centers to be considered. Running a tree program with a slightly generous stopping criteria can provide an upper bound for the list of TreeBF's to be considered. Representers of integration over quarter spaces, slabs, and so forth are also possible. With respect to the pseudo-representers σ_j, which can be characterized by their boundaries $\{\gamma_j' t = b_j\}$, there is a (in general) many-one map between boundaries and partitions of the data sets, at most one boundary per partition could be allowed in the basis list, furthermore the boundaries can be required to pass through at least one data point. A boundary is a better candidate if knowing which side of the boundary a data point is on provides more useful predictive information about its response value. Clever methods needed. [2]

Support vector methods (Vapnik 1995, KW 1971) could be used to screen out basis functions. Letting $\{h_1, \cdots, h_N\}$ be a list of candidate basis functions, let $f = \Sigma_k c_k h_k$. Use a quadratic optimization program to find the c_k to minimize $\|f\|_R^2 = c'\Sigma_N c$, say, subject to $|y_i - \Sigma_{k=1}^N c_k h_k(t(i))| \leq \delta, i = 1, \cdots, n$. Such programs run fast, and in practice (Vapnik 1995, Villalobos and Wahba 1987) it has been found that, even with fairly small δ, many of the c_k will be 0, suggesting that the basis functions associated with them can be discarded.

[2] Added December 9: In the talk in the Error Surfaces Workshop, I suggested generating a very large number of direction cosines γ_j were chosen by generating random numbers on the d dimensional sphere and then screened. Reactions both pro and con were registered by the audience.

REGULARIZED FORWARD BASIS FUNCTION SELECTION

Let $\{h_1, \cdots, h_N\}$ be a list of candidate basis functions, as established previously. Given that $h_1, \cdots, h_{k-1}$ have been selected and λ is (temporarily) fixed, the regularized forward selection problem is to choose h_k from the remaining elements in the list to minimize

$$\inf_c \sum_{i=1}^{n} (y_i - \sum_{k=1}^{k} c_k h_k(t(i)))^2 + \lambda c' \Sigma_k c$$

where Σ_k is given, for example, $c'\Sigma_k c = \Sigma_{i,j}\, c_i c_j < h_i, h_j >$. A general regularized forward selection method would, after selecting h_k, then update λ to minimize the GCV function

$$V_k(\lambda) = \|(I - A_k(\lambda))y\|^2 / (trace(I - A_k(\lambda))^2,$$

where

$$A_k(\lambda) = H_k(H_k' H_k + \lambda \Sigma_K)^{-1} H_k'.$$

Here H_k is the design matrix for $\{h_1, \cdots, h_k\}$. One would continue to increase k and update λ until no useful decrease in $V_k(\lambda)$ obtains.

A similar approach was taken in Diaz 1995 thesis in the context of small density estimation problems. Orr 1966 develops fast methods for doing this in large data sets, via rank-one updating formulae in the zero-th order regularization case $\Sigma = I$ and provides Matlab routines. The general case can be done with matrix decompositions for relatively small k, but it would be nice to find a fast procedure for large k and very large n, possibly via the randomized trace technique, see Wahba, Johnson, Gao and Gong 1995 and references there. Luo and Wahba 1996 used the $\lambda = 0$ updating formula with a modified GCV stopping criteria followed by regularization.

SMOOTHING SPLINE ANOVA, NON-GAUSSIAN DATA

Smoothing spline ANOVA models (Wahba, Wang, Gu, Klein and Klein 1995 and references cited there) can be incorporated into a super-Hilbert space with $\mathcal{H}_R$ as a subspace. In smoothing spline ANOVA models $f(u) \equiv f(u_1, \cdots, u_d)$ is represented as linear combinations of functions of u_α (main effects), functions of u_α, u_β (two factor interactions), etc. These models are based on RK's for d variables which are tensor products of d single variable RK's, and are relatively easy to interpret. They may have a modest number of different smoothing parameters for the different components. See also Hastie and Tibshirani 1990. The penalty functional(s) typically have a non-trivial null space and then the fits shrink towards the nul space as the smoothing parameter(s) become large.

For Bernoulli (0-1) data, the residual sum of squares can be replaced by the log likelihood, with $p(t) \equiv prob\ y_i = 1$ replaced by the logit $f(t) = log[p(t)/(1-p(t)]$. Then $p(t) = \frac{e^{f(t)}}{(1+e^{f(t)})}$, where f is a spline or a spline ANOVA model. See Wahba, Wang, Gu, Klein and Klein (1995). Approximate unbiased risk methods for smoothing parameter selection in the Bernoulli and other cases are discussed there. For the penalized likelihood method, it is assumed that f is 'smooth', as might be expected in some demographic and environmental data sets. The GACV estimate for the Bernoulli case (related to Moody's GPE 1992) is discussed in Xiang and Wahba 1996.

References

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

Chambers, J. & Hastie, T. (1992), *Statistical Models in S*, Wadsworth and Brooks.

Girosi, F., Jones, M. & Poggio, T. (1995), 'Regularization theory and neural networks architectures', *Neural Computation* **7**, 219–269.

Gu, C. & Wahba, G. (1993*a*), 'Semiparametric analysis of variance with tensor product thin plate splines', *J. Royal Statistical Soc. Ser. B* **55**, 353–368.

Gu, C. & Wahba, G. (1993*b*), 'Smoothing spline ANOVA with component-wise Bayesian "confidence intervals"', *J. Computational and Graphical Statistics* **2**, 97–117.

Hastie, T. & Tibshirani, R. (1990), *Generalized Additive Models*, Chapman and Hall.

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

Luo, Z. & Wahba, G. (1995), Hybrid adaptive splines, Technical Report 947, Dept. of Statistics, University of Wisconsin, Madison WI, to appear, J.A.S.A.

Moody, J. (1992), The effective number of parameters: An analysis of generalization and regularization in nonlinear learning systems, *in* J. Moody, S. Hanson & R. Lippman, eds, 'Advances in Neural Information Processing Systems 4', Kaufmann, San Mateo, pp. 847–854.

Moody, J. & Rognvaldsson, T. (1996), Smoothing regularizers for projective basis function networks, Technical Report 96-006, Oregon Graduate Institute of Science and Technology, Portland OR.

Orr, M. (1995), 'Regularization in the selection of radial basis function centers', *Neural Computation* **7**, 606–623.

Orr, M. (1996), 'Introduction to radial basis function networks', Manuscript. in `http://www.cns.ed.ac.uk/people/mark/neuro.html`.

O'Sullivan, F. (1990), 'An iterative approach to two-dimensional laplacian smoothing with application to image restoration', *J. Amer. Statist. Assoc.* **85**, 213–219.

Sin, S. & DeFigueiredo, R. (1993), 'Efficient learning procedures for optimal interpolative nets', *Neural Networks* **6**, 99–113.

Vapnik, V. (1995), *The Nature of Statistical Learning Theory*, Springer.

Villalobos, M. & Wahba, G. (1987), 'Inequality constrained multivariate smoothing splines with application to the estimation of posterior probabilities', *J. Am. Statist. Assoc.* **82**, 239–248.

Wahba, G. (1980), Spline bases, regularization, and generalized cross validation for solving approximation problems with large quantities of noisy data, *in* W. Cheney, ed., 'Approximation Theory III', Academic Press, pp. 905–912.

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

Wahba, G. (1995), Generalization and regularization in nonlinear learning systems, *in* M. Arbib, ed., 'Handbook of Brain Theory and Neural Networks', MIT Press, pp. 426–430.

Wahba, G., Johnson, D., Gao, F. & Gong, J. (1995*a*), 'Adaptive tuning of nummerical weather prediction models: randomized GCV in three and four dimensional data assimilation', *Mon. Wea. Rev.* **123**, 3358–3369.

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

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