
DEPARTMENT OF STATISTICS

University of Wisconsin
Madison, Wisconsin 53706

TECHNICAL REPORT NO. 457

July 1976

A SURVEY OF SOME SMOOTHING PROBLEMS AND THE METHOD OF GENERALIZED CROSS-VALIDATION FOR SOLVING THEM

by

Grace Wahba
University of Wisconsin-Madison

TYPIST: Mary E. Arthur

This research was supported by the United States Air Force under Grant No. AFOSR 72-2363-C.

A Survey of Some Smoothing Problems and the Method of Generalized Cross-Validation for Solving Them

by

Grace Wahba

Abstract

Some applications of the method of generalized cross-validation (GCV) for determining the correct degree of smoothing to minimize mean square error, are surveyed. These are ridge regression, spline smoothing, density estimation, and the approximate solution of linear operator equations when the data are noisy (Tihonov regularization).

1. Introduction. Consider the model

$$y(t) = f(t)+\epsilon(t), \qquad t \,\epsilon\, [0,1] , \tag{1.1}$$

where f is a "smooth" function, and ϵ is a noise, $E\epsilon(t)=0$, $E\epsilon(s)\epsilon(t)=\sigma^2$, $s=t$; $=0$ otherwise. Values of y are observed for $t=t_1, t_2, \ldots, t_n$, and it is desired to recover an estimate of f. Suppose $f \,\epsilon\, H_m$: $\{f: f, f', \ldots, f^{(m-1)}$ abs. cont., $f^{(m)} \epsilon L_2[0,1]\}$. Take as an estimate of f the solution, call it $f_{n,\lambda}$, to the problem: Find $f \,\epsilon\, H_m$ to minimize

$$\frac{1}{n}\sum_{j=1}^{n}(f(t_j)-y_j)^2+\lambda\int_0^1(f^{(m)}(u))^2du, \qquad y_j \equiv y(t_j). \tag{1.2}$$

If $n \geq m$, the solution $f_{n,\lambda}$ is known [20], [22] to be unique, and to be a polynomial spline of degree 2m-1, that is

1) $f_{n,\lambda}$ is a polynomial of degree 2m-1 in each interval $[t_i, t_{i+1}]$, and
2) $f_{n,\lambda}$ has 2m-2 continuous derivatives.

The parameter λ controls the tradeoff between the infidelity of $f_{n,\lambda}$ to the data, as measured by

$$\frac{1}{n}\sum_{j=1}^{n}(f_{n,\lambda}(t_j)-y_j)^2 \tag{1.3}$$

This research was supported by the United States Air Force under Grant No. AFOSR 72-2363-C.

and the "smoothness",

$$\int_0^1 [f_{n,\lambda}^{(m)}(u)]^2 \, du \, . \tag{1.4}$$

If λ is large, then (1.4) will be small while (1.3) will be large, and vice versa. As $\lambda \to \infty$, $f_{n,\lambda}$ tends (pointwise) to the polynomial of degree m best fitting the data in a least squares sense, and as $\lambda \to 0$, the solution tends to a spline function which interpolates the data. The parameter λ may be thought of as controlling the tradeoff between the squared bias and the variance of the estimate $f_{n,\lambda}(t)$ for $f(t)$. If $\lambda=0$, the bias of $f_{n,0}(t_i)$ equals 0, and the variance is σ^2. Provided f is not a polynomial of degree m or less, as λ increases, the squared bias increases and the variance decreases. For a large class of problems there is a value of λ strictly between 0 and ∞ which minimizes the squared bias plus the variance. (If f is a polynomial of degree m or less the optimum value of λ to minimize mean square error is ∞). The experimenter who wishes to smooth data by this technique must have a valid method of choosing λ. This problem, of controlling the tradeoff between squared bias and variance occurs in a variety of contexts, beginning with the problem of admissible estimates for the mean of a multivariate normal vector with quadratic loss studied by Stein and associates (See [9,17] and references cited there), ridge regression ([8, 11,16]), the correct degree of smoothing (equivalently the window width) in density and spectral density estimation ([5,7,19, 25]), the smoothing of surfaces [27], and the approximate solution of linear operator equations when the data are noisy. The operator equation problem is known in approximation theory circles as Tichonov regularization, see [26]).

In 1975 Wahba and Wold [29,30] applied an idea they learned from Stone [23], see also Geisser [10], known as cross-validation, or predictive sample reuse, to choose λ. Monte Carlo results were unbelieveably (at least to the authors!) good. Following this initial success attempts were made to establish theoretical properties of the method of cross-validation and to see if it could be applied in other contexts. It turns out that the method of "ordinary"

A Survey of Some Smoothing Problems and the Method of Generalized Cross-Validation for Solving Them

by

Grace Wahba

Abstract

Some applications of the method of generalized cross-validation (GCV) for determining the correct degree of smoothing to minimize mean square error, are surveyed. These are ridge regression, spline smoothing, density estimation, and the approximate solution of linear operator equations when the data are noisy (Tihonov regularization).

1. Introduction. Consider the model

$$y(t) = f(t)+\varepsilon(t), \qquad t \in [0,1] , \tag{1.1}$$

where f is a "smooth" function, and ε is a noise, $E\varepsilon(t)=0$, $E\varepsilon(s)\varepsilon(t)=\sigma^2$, $s=t$; $=0$ otherwise. Values of y are observed for $t=t_1,t_2,\ldots,t_n$, and it is desired to recover an estimate of f. Suppose $f \in H_m$: $\{f: f,f',\ldots,f^{(m-1)}$ abs. cont., $f^{(m)}\varepsilon L_2[0,1]\}$. Take as an estimate of f the solution, call it $f_{n,\lambda}$, to the problem: Find $f \in H_m$ to minimize

$$\frac{1}{n}\sum_{j=1}^{n}(f(t_j)-y_j)^2+\lambda\int_0^1(f^{(m)}(u))^2du, \qquad y_j \equiv y(t_j). \tag{1.2}$$

If $n\geq m$, the solution $f_{n,\lambda}$ is known [20], [22] to be unique, and to be a polynomial spline of degree 2m-1, that is

1) $f_{n,\lambda}$ is a polynomial of degree 2m-1 in each interval $[t_i,t_{i+1}]$, and

2) $f_{n,\lambda}$ has 2m-2 continuous derivatives.

The parameter λ controls the tradeoff between the infidelity of $f_{n,\lambda}$ to the data, as measured by

$$\frac{1}{n}\sum_{j=1}^{n}(f_{n,\lambda}(t_j)-y_j)^2 \tag{1.3}$$

This research was supported by the United States Air Force under Grant No. AFOSR 72-2363-C.

and the "smoothness",

$$\int_0^1 [f_{n,\lambda}^{(m)}(u)]^2 \, du \ . \tag{1.4}$$

If λ is large, then (1.4) will be small while (1.3) will be large, and vice versa. As $\lambda \to \infty$, $f_{n,\lambda}$ tends (pointwise) to the polynomial of degree m best fitting the data in a least squares sense, and as $\lambda \to 0$, the solution tends to a spline function which interpolates the data. The parameter λ may be thought of as controlling the tradeoff between the squared bias and the variance of the estimate $f_{n,\lambda}(t)$ for f(t). If $\lambda=0$, the bias of $f_{n,0}(t_i)$ equals 0, and the variance is σ^2. Provided f is not a polynomial of degree m or less, as λ increases, the squared bias increases and the variance decreases. For a large class of problems there is a value of λ strictly between 0 and ∞ which minimizes the squared bias plus the variance. (If f is a polynomial of degree m or less the optimum value of λ to minimize mean square error is ∞). The experimenter who wishes to smooth data by this technique must have a valid method of choosing λ. This problem, of controlling the tradeoff between squared bias and variance occurs in a variety of contexts, beginning with the problem of admissible estimates for the mean of a multivariate normal vector with quadratic loss studied by Stein and associates (See [9,17] and references cited there), ridge regression ([8, 11,16]),the correct degree of smoothing (equivalently the window width) in density and spectral density estimation([5,7,19, 25]), the smoothing of surfaces [27], and the approximate solution of linear operator equations when the data are noisy. The operator equation problem is known in approximation theory circles as Tichonov regularization, see [26]).

In 1975 Wahba and Wold [29,30] applied an idea they learned from Stone [23], see also Geisser [10], known as cross-validation, or predictive sample reuse, to choose λ. Monte Carlo results were unbelieveably (at least to the authors!) good. Following this initial success attempts were made to establish theoretical properties of the method of cross-validation and to see if it could be applied in other contexts. It turns out that the method of "ordinary" cross validation must be modified to have desirable properties in general. With this modification, which we call generalized cross validation (GCV), we have been able to show theoretically desirable properties in general, and pleasing Monte Carlo results when the method is used to estimate an optimum smoothing parameter in ridge regression and to estimate the correct degree of smoothing in density estimation.

In this paper we give a general formulation of the GCV method for estimating a good value of the parameter which controls the tradeoff between square bias and variance. The purpose of this paper is to provide a unified overview of some of the various contexts in which the GCV method is applicable, and survey some of the known results, published and unpublished. We will state theorems concerning the properties of the GCV estimate of λ for ridge regression, spline smoothing and density estimation, without proofs. The proofs have appeared or will appear elsewhere. Brief mention will be made of the application to smoothing of surfaces and to Tichonov regularization. In the ridge regression and density estimation cases we will present preliminary, but typical, Monte Carlo results, so that the reader may judge for himself or herself wether he or she would like to use the method on real data.

2. Ridge Regression

Consider the usual regression model,

$$y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \varepsilon_{n\times 1} \tag{2.1}$$

where subscripts denote the dimensions of the vector or matrix. Let $\varepsilon \sim N(0,\sigma^2 I_{n\times n})$. The ridge estimate $\hat{\beta}_\lambda$ of β that we consider is

$$\hat{\beta}_\lambda = (X^T X + n\lambda I)^{-1} X^T y \ . \tag{2.2}$$

It is well known that $\hat{\beta}_\lambda$ is the posterior mean of β if β has the prior $\beta \sim N(0,aI_{p\times p})$, where $\lambda=\sigma^2/na$, see for example [11]. It is also not hard to show that $\hat{\beta}_\lambda$ is the solution to the minimization problem:

Find $\hat{\beta}$ in Euclidean p-space to minimize

$$\frac{1}{n}\,||y-X\beta||_n^2 + \lambda||\beta||_p^2 \ , \tag{2.3}$$

where $||\cdot||_q$ is the Euclidean q norm. It will be convenient to think of $\frac{1}{n}||y-X\beta||_n^2$ as the "infidelity", and $||\beta||_p^2$ as the "energy" or "smoothness". For fixed β, λ controls the tradeoff between squared bias and variance. Thus the reader may take a Bayesian point of view, or minimize the infidelity plus smoothness functional, either approach can lead to estimates of the form (2.2). The "ordinary" cross validation estimate of λ for ridge regression was suggested by Allen (see [1] and the discussion in Stone [23]) and given the name PRESS. It goes as follows: Let $\hat{\beta}^{(k)}$ be the solution to:

Find $\hat{\beta} \varepsilon E_p$ to minimize

$$\frac{1}{n}\ ||y^{(k)}-X^{(k)}\beta||_{n-1}^2 + \lambda\ ||\beta||_p^2\ ,$$

where $y^{(k)}$ and $X^{(k)}$ are obtained from y and X by deleting the kth row in each case. Thus, $\beta_\lambda^{(k)}$ is a ridge estimate with the kth data point left out. The idea is, that if a particular λ is a good choice, then $[X\hat{\beta}_\lambda^{(k)}]_k$, the kth entry of $X\hat{\beta}_\lambda^{(k)}$, should be a good predictor of the missing data point y_k. This predictive ability is measured by $([X\hat{\beta}_\lambda^{(k)}]_k-y_k)^2$. The Allen's PRESS estimate for λ is then obtained by choosing λ to minimize $V_0(\lambda)$ given by

$$V_0(\lambda) = \sum_{k=1}^{n} ([X\hat{\beta}_\lambda^{(k)}]_k-y_k)^2\ . \tag{2.4}$$

This idea is intuitively appealing. However, if the components x_{ij} of X satisfy $x_{ij}=0$, $i\neq j$, that is, $y_i=x_{ii}\beta_i+\varepsilon_i$, $i=1,2,\ldots,p$, $y_i=\varepsilon_i$, $i=p+1,\ldots,n$, then $y_1,\ldots,y_{k-1}$, $y_{k+1},\ldots,y_n$ cannot be expected to provide much information about the kth component β_k of β, and in fact it can be shown that $V_0(\lambda)$ is independent of λ in that case, and insensitive to λ in nearby cases.

The generalized cross validation (GCV) estimate of λ is obtained as follows: Let the singular value decomposition (see [14]) of X be

$$X = U_{n\times n}\ D_{n\times p}\ V_{p\times p}^T$$

where U and V are orthogonal, and D is diagonal. Consider the model

$$U^Ty = DV^T\beta + U^T\varepsilon \tag{2.5}$$

This model is basically the same as the original model (2.1) with X replaced by a diagonal matrix, since if $\beta \sim N(0,aI_{p\times p})$ and $\varepsilon \sim N(0,\sigma^2 I_{n\times n})$, then $V^T\beta \sim N(0,aI_{p\times p})$ and $U^T\varepsilon \sim N(0,\sigma^2 I_{n\times n})$. Now, (2.5) is just in that form where no information about $\gamma_k = [V^T\beta]_k$ is provided from $\{[U^Ty]_j\}$, $j \neq k$. We now want to rotate the coordinate system so that when a "data point" is left out the remaining data points provide maximal information about the missing point. Let W be the n×n unitary matrix which diagonalizes the circulant matrices. (See [4,24]) In complex form, the jk^{th} entry of W is given by

$$[W]_{jk} = \frac{1}{\sqrt{n}} e^{2\pi ijk/n} .$$

(Recall that if $C_{n\times n}$ is any diagonal matrix, then WCW^* is a circulant matrix). Write

$$\tilde{y} \equiv WU^Ty = WD(V^T\beta) + WU^T\varepsilon. \tag{2.6}$$

Since we have only effected a rotation, the model is still essentially the same as (2.1), except that it is in complex form. The "new" design matrix, call it $\tilde{X}$ is now

$$\tilde{X} = WDV^T \tag{2.7}$$

and $\tilde{X}\tilde{X}^{*T}$ is a circulant matrix. The GCV estimate of λ is defined as the result of doing "ordinary" cross validation, or Allen's PRESS, on the rotated system

$$\begin{aligned}\tilde{y} = WU^Ty &= WDV^T\beta + WV^T\varepsilon \, . \\ &= \tilde{X}\beta + WU^T\varepsilon \, .\end{aligned} \tag{2.8}$$

Upon substituting $\tilde{X}$ and $\tilde{y}$ into (2.4) and after some manipulation (see [13]), it can be shown that the GCV estimate defined this way is the minimizer of $V(\lambda)$ given by

$$V(\lambda) = ||(I-A(\lambda))y||^2/[\mathrm{Trace}(I-A(\lambda))]^2 \tag{2.9}$$

where

$$A(\lambda) = X(X^TX+n\lambda I)^{-1}X^T. \qquad A(\lambda)y \equiv X\hat{\beta}_\lambda \, . \tag{2.10}$$

$V(\lambda)$ can also be written

$$V(\lambda) = \sum_{\nu=1}^{n} \left(\frac{n\lambda}{\lambda_\nu^2+n\lambda}\right)^2 z_\nu^2 \Big/ \left(\sum_{\nu=1}^{n} \frac{n\lambda}{\lambda_\nu^2+n\lambda}\right)^2 \tag{2.11}$$

where $\lambda_1,\lambda_2,\ldots,\lambda_p$ are the singular values of X, $\lambda_{p+1}=\ldots=\lambda_n=0$, and $z=(z_1,z_2,\ldots,z_n)^T$ is given by $z=U^Ty$.

We will first give some of the theoretical properties relating to the behavior of $\hat{\lambda}$, the minimizer of $V(\lambda)$. Proofs may be found in Golub, Heath, and Wahba [13]. The general idea of Theorem 1 is that $\hat{\lambda}$ is an estimate of the minimizer of the mean square prediction error.

Theorem 1. Let the mean square prediction error $T(\lambda)$ be defined by

$$T(\lambda)=\frac{1}{n}||X\beta-X\hat{\beta}_\lambda||^2$$

Let $\tilde{\lambda}$ be the minimizer of $ET(\lambda)$ and λ^* be the minimizer of $EV(\lambda)$. Let β be fixed. Suppose either

Case a) $\lim_{n\to\infty}\frac{1}{n}X^TX = B_{p\times p}$, σ^2 fixed, where B is a p×p matrix, or

Case b) X fixed, $n\geq p+1$, fixed, and $\sigma^2\to 0$

Then

$$\lambda^* = \tilde{\lambda}(1+o(1)) \qquad (2.12)$$

where $o(1)\to 0$ as $n\to\infty$ (case a) or as $\sigma^2\to 0$ (case b).

The next theorem would appeal to a Bayesian, but, practically speaking, tells us not to be surprised if $\hat{\lambda}$ is a good estimate of the minimizer of $T(\lambda)$ even if n is small and σ^2 medium-sized.

Theorem 2. Let E_p denote expectation with respect to the prior distribution $\beta \sim N(0,aI_{p\times p})$, (and E denote expectation with respect to the prior distribution $\varepsilon \sim N(0,\sigma^2I)$). Let Q be any p×p matrix. Then $E_pE||Q(\beta-\hat{\beta}_\lambda)||^2$ and $E_pEV(\lambda)$ are <u>both</u> minimized for $\lambda=\sigma^2/na$.

An unbiased estimate of the negative risk improvement $ET(\lambda)-\sigma^2$, is given by $R(\lambda)$,

$$R(\lambda) = \frac{1}{n}||(I-A(\lambda))y||^2-2\hat{\sigma}^2(\frac{1}{n}\text{Trace}(I-A(\lambda))), \qquad (2.13)$$

where $\hat{\sigma}^2$ is an unbiased estimate of σ^2. The minimizer of $R(\lambda)$ as an estimate of λ has been suggested by Hudson [17]. The GCV estimate may be considered as a form of risk improvement (RI) estimate which does not require knowledge of, or an estimate of σ^2.

For comparison we describe the maximum likelihood estimate of $\lambda=\sigma^2/na$ when $\beta \sim N(0,aI)$. By writing the prior distribution of y as

$$\begin{aligned} y &\sim N(0,aXX^T+\sigma^2I) \\ &= N(0,a(XX^T+n\lambda I)) \end{aligned} \qquad (2.14)$$

one can obtain that the maximum likelihood estimate of λ in the model (2.14) is the minimizer of

$$M(\lambda)=(y^T(I-A(\lambda))y)/[\mathrm{Det}(I-A(\lambda))]^{1/n}\equiv\Big(\sum_{\nu=1}^{n}\frac{n\lambda}{\lambda_\nu^2+n\lambda}z_\nu^2\Big)\Big/\Big[\prod_{\nu=1}^{n}\Big(\frac{\lambda}{\lambda_\nu^2+\lambda}\Big)\Big]^{1/n}. \tag{2.15}$$

It can be shown that Theorem 2 is also true for the maximum likelihood estimate, that is, $E_pE\,M(\lambda)$ is minimized for $\lambda=\sigma^2/na$. However, (2.12) of Theorem 1 does not hold for all β, for the MLE estimate.

We present a summary of the results from the first run of a Monte Carlo study evaluating the GCV estimate of λ and comparing it with several other estimates. The complete study will appear in Golub, Heath and Wahba [13]. The values for n and p were 21 and 10, and the design matrix and β come from discretizing a numerical inversion of the Laplace transform. The condition number of X, namely the ratio of the largest to the smallest singular value, was 1.54×10^5. Four values of σ^2, namely $\sigma^2=10^{-8}$, 10^{-6}, 10^{-4} and 10^{-2} were tried and for each value of σ^2 the experiment was replicated 4 times, giving a total of 16 runs. The ε_i were generated as pseudo random $N(0,\sigma^2)$ independent r.v.'s, $V(\lambda)$ was computed using the right hand side of (2.11) and the Golub-Reinsch singular value decomposition [14], and the minimizer $\hat\lambda$ of $V(\lambda)$ determined by a global search. $T(\lambda)$ was also computed, and the relative inefficiencies I_D and I_R, of $\hat\lambda$, defined by

$$I_D=||\beta-\hat\beta_{\hat\lambda}||^2/(\min_\lambda||\beta-\hat\beta_\lambda||^2);\ I_R=T(\hat\lambda)/\min_\lambda T(\lambda) \tag{2.16}$$

were computed. (D = "domain", R = "range"). Four other methods were studied for comparison. They are

1) Hoerl-Kennard-Baldwin (HKB) [16]. $n\hat\lambda=\hat\sigma^2/\frac{1}{p}||\hat\beta_0||^2$, where $\hat\sigma^2$ is the "usual" estimator of σ^2, $\hat\sigma^2=\frac{1}{n-p}||y-X\hat\beta_0||^2$.

2) Maximum Likelihood. $\hat\lambda$ is the minimizer of (2.15).

3) Allen's PRESS. $\hat\lambda$ is the minimizer of (2.4).

4) Dempster's RIDGM [8]. $\hat\lambda$ is the solution to

	Replication 1		Replication 2		Replication 3		Replication 4	
	I_D	I_R	I_D	I_R	I_D	I_R	I_D	I_R
$\sigma^2=10^{-8}$, S/N $\simeq$ 4200								
GCV	4.43*	1.06*	1.65*	1.03*	16.71*	1.10*	1.02*	1.01*
HKB	2.18E3	1.31	1.14E2	1.23	3.99E4	1.89	1.03E3	1,52
RIDGM	2.14E3	1.31	1.13E2	1.22	3.95E4	1.88	1.03E3	1.52
MLE	1.67E3	1.31	1.45E2	1.23	2.00E3	1.53	9.12E3	1.51
PRESS	2.31E3	4.8E4	6.31E2	8.6E4	3.84E3	2.1E5	2.87E3	1.2E5
Min Sol'n	1.00	1.02	1.00	1.54	1.00	2.27	1.00	1.00
Min Data	1.20	1.00	2.89	1.00	5.97	1.00	1.00	1.00
$\sigma^2=10^{-6}$, S/N $\simeq$ 420								
GCV	1.92*	1.05	1.32*	1.00*	1.51E2	1.26*	2.20*	1.02*
HKB	1.21E3	1.35	6.24E3	1.58	3.60E5	2.25	5.07E3	1.40
RIDGM	1.21E3	1.35	6.23E3	1.58	3.60E5	2.25	5.05E3	1.40
MLE	1.99E2	1.19	1.70E2	1.45	1.76E2	1.29	1.49E2	1.32
PRESS	5.80	1.01	2.41E2	1.39E4	36.37*	2.43E3	67.00	6.07E2
Min Sol'n	1.00	1.38	1.00	1.02	1.00	1.20	1.00	1.03
Min Data	3.56	1.00	1.28	1.00	7.85	1.00	41.29	1.00
$\sigma^2=10^{-4}$, S/N $\simeq$ 42								
GCV	1.27*	1.07*	1.50*	2.58*	1.00*	1.11*	1.00*	1.03*
HKB	4.34E3	1.75	1.34E4	4.30	2.03E4	1.98	2.82E6	2.66
RIDGM	4.33E3	1.75	1.31E4	4.30	2.02E4	1.97	2.81E6	2.66
MLE	1.56	1.20	12.16	3.43	1.90	1.49	2.97	1.07
PRESS	3.53	1.57	2.03	3.43	8.66	2.63	2.90	24.34
Min Sol'n	1.00	1.21	1.00	2.05	1.00	1.11	1.00	1.03
Min Data	3.26	1.00	1.16	1.00	2.39	1.00	1.16	1.00
$\sigma^2=10^{-2}$, S/N $\simeq$ 4.2								
GCV	1.40	2.47	2.01*	1.60*	1.59*	1.01	3.12	1.72*
HKB	7.05E7	13.20	3.32E5	3.35	4.58E4	2.20	5.39E5	31.0
RIDGM	7.05E7	13.20	3.31E5	3.35	4.56E4	2.20	5.26E5	31.0
MLE	2.13	3.56	3.81	1.87	2.00	1.00*	2.88	1.69
PRESS	1.04*	1.01*	2.02	2.68	1.00	1.22	2.16	2.15
Min Sol'n	1.00	1.31	1.00	1.01	1.00	1.25	1.00	1.98
Min Data	1.02	1.00	1.00	1.00	2.66	1.00	1.21	1.00

"*" indicates lowest inefficiency among the estimates

Observed Inefficiencies in Sixteen Monte Carlo Runs

Table 1

$$\frac{1}{p}\sum_{\nu=1}^{p} n\lambda\ \hat{\gamma}^2 \frac{\lambda_\nu^2}{\lambda_\nu^2+n\lambda} = \hat{\sigma}^2 ,$$

where $\hat{\gamma}=(\hat{\gamma}_1,\hat{\gamma}_2,\ldots,\hat{\gamma}_p)^T=V^T\hat{\beta}_0$.

Comparison with Hudson's risk improvement estimate is in progress.

Dempster's RIDGM was initially chosen because it was reported to be the best of a number of estimates tried in a fairly extensive Monte Carlo study [8]. It can be seen that $\hat{\lambda}$ from RIDGM will be larger than $\hat{\lambda}$ from Hoerl-Kennard and will be nearly the same if $\lambda_\nu^2/(\lambda_\nu^2+n\lambda)\simeq 1$. I_D and I_R were determined for each of these four methods as well as GCV and the results are presented in Table 1. The entries next to "Min Sol'n" and "Min Data" are the inefficiencies (2.15) with $\hat{\lambda}$ taken as the minimizer of $||\beta-\hat{\beta}_\lambda||^2$ and $T(\lambda)$ respectively. S/N, the "signal to noise ratio" is defined by S/N = $[\sigma^2/\frac{1}{n}||X\beta||^2]^{-1/2}$. Figure 1 gives a plot of $V(\lambda)$, $M(\lambda)$, $V_0(\lambda)$, $||\beta-\hat{\beta}_\lambda||^2$ and $T(\lambda)$ for Replicate 2 of the $\sigma^2=10^{-6}$ case. The $T(\lambda)$ and $V(\lambda)$ curves tend to follow each other, and this is fairly typical.

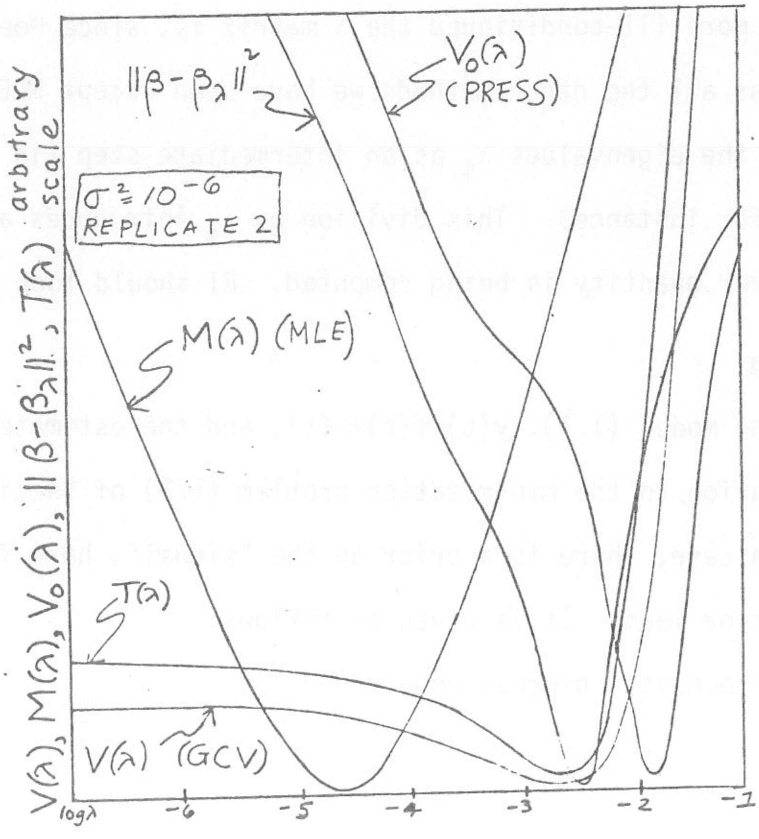

Figure 1

We make a few comments on these very preliminary results. The superiority of the GCV estimate of λ is apparent, at least in this example, compared to the other methods listed. The Gauss Markov estimate of λ was not included because it is uniformly worse than all the other estimates in this very ill-conditioned problem. A first run comparing GCV and the RI estimate of (2.13) as this goes to press indicates that the two methods give similar results for n-p=11.

We feel that GCV is generally going to be superior to PRESS, unless XX^T is circulant, in which case they will be the same. It is reasonable to think MLE will look good when β acts "as though" it came from the prior. This really means either p is small or $\{\beta_i^2\}_{i=1}^p$ have about the same spread or less as independent $\chi^2(1)$ random variables times some constant. If p is large it is possible to construct examples where MLE will look bad compared to GCV. In these examples, $\beta_i^2 \to 0$ in a manner uncharacteristic of the prior and the MLE estimate is too small. We think the Hoerl-Kennard and RIDGM estimates will look worse compared to GCV the more ill-conditioned the X matrix is, since Hoerl-Kennard and RIDGM, as well as all the other methods we have seen except MLE and RI divide somewhere by the eigenvalues λ_i as an intermediate step (in the computation of $\hat{\beta}_0$, for instance). This division by λ_i introduces a large variance for whatever quantity is being computed. RI should look good if n-p is large.

3. Spline Smoothing

We return to the model (1.1), $y(t)=f(t)+\varepsilon(t)$, and the estimate $f_{n,\lambda}$ for f. Let $f_{n,\lambda}$ be the solution to the minimization problem (1.2) of Section 1. As in the ridge regression case, there is a prior on the "signal", here f, for which $f_{n,\lambda}(t)$, is a posterior mean. It is given as follows:

Let f(t) be a stochastic process modeled by

$$f(t) = \sum_{j=0}^{m-1} \theta_j t^j + \sqrt{a}\int_0^t \frac{(t-u)_+^{m-1}}{(m-1)!} dW(u) \qquad (3.1)$$

where W(u) is the Weiner process, a is a constant, and $\{\theta_j\}$ are Gaussian random variables with a "flat" prior, i.e. infinite variance. It can be shown (by using the results in Kimeldorf and Wahba [18]) that $f_{n,\lambda}(t)$ is the posterior mean of f(t) given $y(t_i)=y_i$, with $\lambda=\sigma^2/na$.

The vector $\bar{f}_{n,\lambda}=(f_{n,\lambda}(t_1),f_{n,\lambda}(t_2),\ldots,f_{n,\lambda}(t_n))^T$ depends linearly on the data $y=(y_1,y_2,\ldots,y_n)$, and we define the matrix $A(\lambda)$ by

$$\bar{f}_{n,\lambda} = A(\lambda)y . \qquad (3.2)$$

We omit giving an explicit formula for $A(\lambda)$, it can be obtained for computational purposes in the m=2 case from Reinsch [20]; in general from Kimeldorf and Wahba, [18], and is explicitly given in Craven and Wahba [6].

The GCV estimate of λ is defined as the minimizer of

$$V(\lambda) = ||(I-A(\lambda))y||^2/[\text{Trace } (I-A(\lambda))]^2 \qquad (3.3)$$

(Compare (2.9)). The MLE estimate of λ with the prior (3.1) can be shown to be the minimizer of

$$M(\lambda) = (y^T(I-A(\lambda))y)/[\text{Det}(I-A(\lambda))]^{1/n} \qquad (3.4)$$

This MLE estimate for λ is essentially that proposed by Anderssen and Bloomfield in their pioneering papers [2] and [3], and generalizes their estimate to the non-periodic non-equally spaced data case.

In the equally spaced data case ($t_i=i/n$) for large n, GCV and "ordinary" cross validation are very similar. (In analogy to the ridge regression case, the matrix that plays the role of U is very close to the matrix that plays the role of W). Wahba and Wold [29] reported on Monte Carlo results using "ordinary" cross-validation in connection with spline smoothing. Reinsch's program [20] was used to repeatedly calculate $f_{n,\lambda}^{(k)}$, the solution to the minimization problem of (1.2) with the kth data point omitted. Favorable results were reported with the method. Cheap computational procedures for calculating $V(\lambda)$ from (3.3) are under development and will be reported in [6]. Data sets of 50 can be handled for a few dollars. The following theorem is proved in Craven and Wahba [6].

Theorem 3. Let $f \varepsilon H_m = \{f: f^{(\nu)}$ abs. cont., $\nu=0,1,\ldots,m-1$, $f^{(m)} \varepsilon L_2\}$ and suppose $\{t_i\}_{i=1}^n = \{t_{in}\}_{i=1}^n$ where $\int_0^{t_{in}} w(u)du = i/n$, $i=1,2,\ldots,n$, $n=1,2,\ldots$, and w is a strictly positive continuous weight function. Let

$$T(\lambda) = \frac{1}{n} \sum_{i=1}^n (f_{n,\lambda}(t_{in}) - f(t_{in}))^2$$

and let $\tilde{\lambda}$ be the minimizer of $ET(\lambda)$ and λ^* be the minimizer of $EV(\lambda)$.

i) If f is a polynomial of degree m-1 or less then

$$\lambda^* = \tilde{\lambda} = \infty .$$

ii) If f is not a polynomial of degree m-1 or less then

$$\lambda^* = \tilde{\lambda}(1+o(1)) \tag{3.5}$$

where $o(1) \to 0$ as $n \to \infty$.

Letting E be expectation with respect to the prior of (3.1), the following can be proved.

Theorem 4. The functions $E_pEV(\lambda)$, $E_pET(\lambda)$ and $E_pEM(\lambda)$ have the common minimizer $\lambda = \sigma^2/na$.

<u>If</u> f is a very smooth function in H_m, however it can be shown that (3.5) is false for the MLE estimate. It can be shown that as $n \to \infty$ the MLE estimate will tend to 0 more rapidly than the GCV estimate. This is not surprising, since it can be shown that sample functions from the prior (3.1) are, with probability 1, not in H_m.

The smoothing spline $f_{n,\hat{\lambda}}$ may be used to estimate $f^{(\nu)}(t)$ for any $\nu \leq m-1$. The estimate is $f_{n,\hat{\lambda}}^{(\nu)}(t)$. Pleasing numerical results for $\nu=1$, $m=2$ can be found in Wahba and Wold [29]. This problem was in fact the motivation for Anderssen and Bloomfield [23] who also had nice numerical results.

If H_m is constrained to be a space of periodic functions, the computations are vastly simplified. See section 4.

4. Density Estimation

Every reasonable non-parametric density estimate has (implicitly or explicitly) a parameter or parameters which control the tradeoff between the square bias and the variance. For example, in an estimate of kernel type, the density estimate $\hat{f}$ is given by

$$\hat{f}(t) = \frac{1}{nh} \sum_{j=1}^{n} K(\frac{t-X_j}{h})$$

where $X_1, X_2, \ldots, X_n$ are iid observations from some density f, and $K(\tau)$ is a "hill function" integrating to 1. See Parzen [19]. The parameter h controls the tradeoff between the square bias and the variance and the h which minimizes mean square error depends on the unknown f as well as the sample size n. To our knowledge, no completely automatic, practical method for choosing the smoothing parameter to minimize mean square error has been reported (See, however, Good and Gaskins [15], Woodroofe [31]). However, the method of GCV can be used to do this in conjunction with a smoothing spline density estimate. Detailed results will appear in [28]. We summarize the main idea here, for the purpose of illustrating the wide applicability of the GCV method.

We discuss the method as it applies only to densities that are in H_m of Section 3 and furthermore satisfy the periodic boundary conditions

$$f^{(\nu)}(1)-f^{(\nu)}(0)=0, \quad \nu=0,1,\ldots,m-1 \; .$$

In practice for small to medium sample sizes m=2 should be chosen. The value m=2 corresponds to visual smoothness. We restrict ourselves to the periodic case for <u>practical</u> rather than <u>theoretical</u> reasons. The computations are easy to program and are "dirt cheap" in the periodic case as can be seen below. To obtain the method we first discuss periodic smoothing splines. It can be shown (See Golomb [12], Kimeldorf and Wahba [18], for the necessary background), that the solution $f_{n,\lambda}$ to the minimization problem: Find f in H_m satisfying

$$\int_0^1 f(u)du=1, \quad f^{(\nu)}(1)-f^{(\nu)}(0)=0, \quad \nu=0,1,\ldots,m-1$$

and minimizing

$$\frac{1}{n}\sum_{j=1}^{n} (f(\tfrac{j}{n})-y_j)^2 + \lambda \int_0^1 (f^{(m)}(u))^2\, du\ , \quad y_j = y(\tfrac{j}{n})$$

satisfies

$$f_{n,\lambda}(\tfrac{j}{n}) = (1 + \sum_{\substack{\nu=-n/2 \\ \nu\neq 0}}^{n/2} \frac{\hat{f}_\nu}{(1+\lambda(2\pi\nu)^{2m})} e^{2\pi i\nu j/n})\ (1+o(1))$$

where

$$\hat{f}_\nu = \frac{1}{n}\sum_{j=1}^{n} y_j\ e^{2\pi i\nu j/n} \tag{4.1}$$

and $o(1)\to 0$ (rapidly) as $n\to\infty$. Thus, the periodic smoothing spline with equally spaced data points can be viewed as the result of passing the data through a low pass filter. The parameter λ controls the half power point of the filter, by the relationship that the frequency ν_0 at which f_ν is damped by a factor of 2 satisfies $\lambda(2\pi\nu_0)^{2m}=1$, or $\nu_0=(2\pi\lambda^{1/2m})^{-1}$. The value of λ which minimizes mean square error is estimated by GCV, as the minimizer of

$$V(\lambda) = \frac{\sum_{\nu=1}^{n} (\frac{\lambda}{(2\pi\nu)^{-2m}+\lambda})^2 \hat{f}_\nu^2}{(\sum_{\nu=1}^{n} \frac{\lambda}{(2\pi\nu)^{-2m}+\lambda})^2}$$

(Compare (2.11)). The eigenvalues of the (formal) operator $[(\frac{d}{dt})^m]^{-1}$ are $(2\pi\nu)^{-m}$ and play the role of the singular values of X. See Wahba [28] for details. In the estimation of a periodic density, (with m=2), the estimate is

$$f_{n,\lambda}(t) = 1 + \sum_{\substack{\nu=-n/2 \\ \nu\neq 0}}^{n/2} \frac{\hat{f}_\nu}{(1+\lambda(2\pi\nu)^4)} e^{2\pi i\nu t}$$

where the $\hat{f}_\nu$ of (4.1) are replaced by the sample fourier coefficients,

$$\hat{f}_\nu = \frac{1}{n}\sum_{j=1}^{n} e^{2\pi i\nu X_j} \ .$$

The estimate is related to that of Cogburn and Davis [7] for estimating spectral densities, but the method for choosing λ is new. The GCV method can be used in the Cogburn-Davis spectral density estimate to choose λ, see [30].

The following Theorem is proved in [28].

Theorem 5. Let f be a periodic density in H_m. Let $T(\lambda)=\frac{1}{n}\sum_{j=1}^{n} (f_{n,\lambda}(\frac{j}{n})-f(\frac{j}{n}))^2$

and let λ^* and $\tilde{\lambda}$ be the minimizers of $EV(\lambda)$ and $ET(\lambda)$ respectively. Then

$$\lambda^* = \tilde{\lambda}(1+o(1))$$

where $o(1)\to 0$ as $n\to\infty$.

It can be shown that if, further, $f^{(\nu)}$ abs. cont., $\nu=m, m+1,\dots,2m-1$, $f^{(m)} \in L_2$, and $f^{(\nu)}(1)-f^{(\nu)}(0) = 0$, $\nu = m,m+1,\dots,2m-1$, then

$$ET(\tilde{\lambda}) = O(n^{4m/(4m+1)}) ,$$

and hence that this density estimate shares the convergence properties of most of the well known estimates. See [25] for a discussion of the best mean square error at a point convergence rates abtainable for densities in H_m.

Figure 2 provides the results of a demonstration run of this method. One hundred and seventy-four observations from a density which is a mixture of two betas were generated by Monte Carlo methods. The density f is

$$f \sim \frac{6}{10}\,\beta(12,7) + \frac{4}{10}\,\beta(3,11) .$$

Figure 2a gives a plot of the true density and a histogram of the 174 Monte Carlo samples. The "bin size", which is the "smoothing" parameter in the histogram estimate, was chosen by eye. Figure 2b gives a plot of $T(\lambda)$ and $V(\lambda)$. It can be seen that the minimizers are quite close. The minimizer of $V(\lambda)$ is $\hat{\lambda}$ and the relative inefficiency $T(\hat{\lambda})/(\min_{\lambda} T(\lambda))$, was 1.04. Figure 2c gives a plot of $f_{n,0}(t)$ and the density estimate $f_{n,\hat{\lambda}}(t)$. $f_{n,\hat{\lambda}}(t)$ is in fact the low-pass filtered version of $f_{n,0}(t)$. Figure 2d gives a plot of f and $f_{n,\hat{\lambda}}$. We remind the reader that the computation of $f_{n,\hat{\lambda}}$ is "completely automatic", with no human intervention required to choose the optimal smoothing.

5. The approximate solution of linear operator equations when the data are noisy.

Let

$$y(t) = \int_0^1 K(t,s)f(s)ds+\varepsilon(t) \tag{5.1}$$

where K is a continuous kernel and ε is a noise satisfying $E\varepsilon(t)=0$, $E\varepsilon(s)\varepsilon(t)=\sigma^2$, $s=t$, $=0$, otherwise. y is observed for $t=t_1,t_2,\dots,t_n$, and it is desired to recover

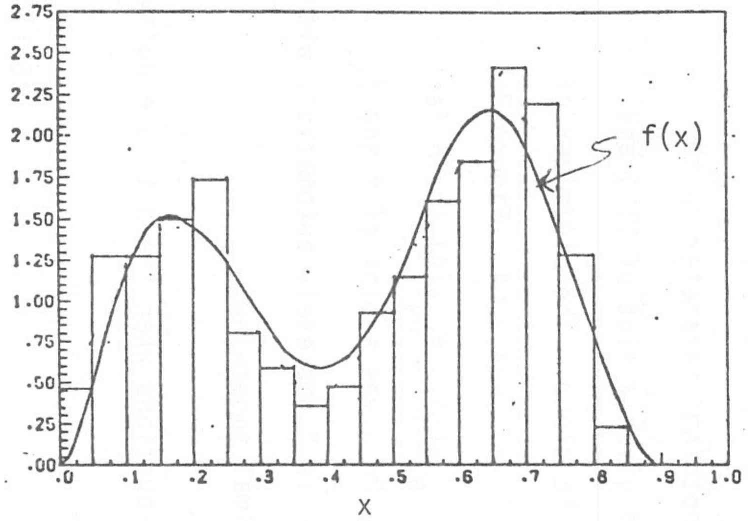

Figure 2a

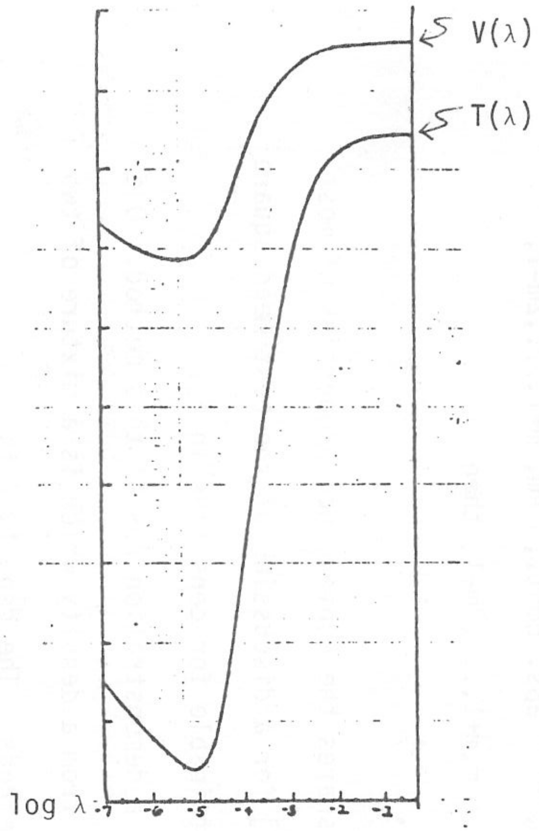

Figure 2b

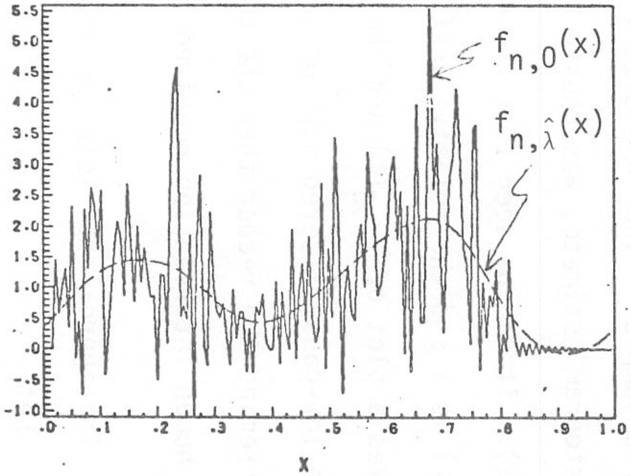

Figure 2c

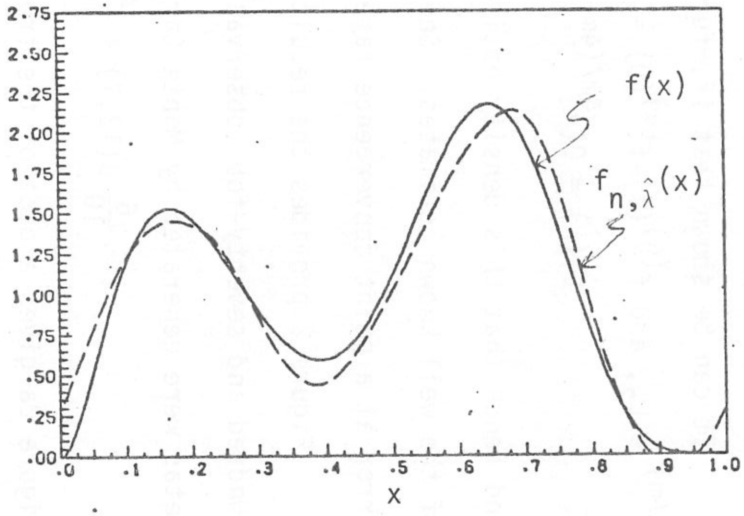

Figure 2d

f. This problem is a common one in many physical sciences. It is assumed that f is in some Hilbert space H, of smooth functions. The approximate solution $f_{n,\lambda}$ is taken as the solution to the minimization problem: Find $f \in H$ to minimize

$$\frac{1}{n} \sum_{j=1}^{n} [\int_0^1 K(t_j,s)f(s)ds-y_j]^2 + \lambda||f||_H^2 , \; y_j = y(t_j). \qquad (5.2)$$

This method for solving (5.1) approximately is known in the approximation theory literature as Tichonov regularization, and is a Hilbert space version of ridge regression. See [26].

The GCV estimate of λ is the minimizer of $V(\lambda)$ given by (3.3) where $A(\lambda)$ is defined by (3.2), with $\overline{f}_{n,\lambda}$ replaced by $\overline{g}_{n,\lambda}=(\int_0^1 K(t_1,s)f_{n,\lambda}(s)ds,\ldots, \int_0^1 K(t_n,s)f_{n,\lambda}(s)ds)^T$, where $f_{n,\lambda}$ is the solution to the minimization problem of (5.2). Analogous theorems to those of Sections 2 and 3 are found in [26].

There is nothing special about the index set $t \in [0,1]$ other than that it is bounded. The results of Section 3 and above can be shown to hold for t in an interval in Euclidean d-space, that is, for the surface smoothing problem. See [27].

REFERENCES

[1] Allen, D.M., (1974). The relationship between variable selection and data augmentation and a method for prediction, Technometrics, 16, 1, 125-127.

[2] Anderssen, R.S. and Bloomfield, P. (1974). A time series approach to numerical differentiation. Technometrics, 16, 1, 69-75.

[3] Anderssen, B. and Bloomfield, P. (1974). Numerical differentiation procedures for non-exact data, Numer. Math. 22, 157-182.

[4] Bellman, R., (1960), Introduction to Matrix Analysis, McGraw Hill, New York.

[5] Blackman, R.B., and Tukey, J.W., (1958). The measurement of power spectra, Dover, New York.

[6] Craven, P. and Wahba, G., (1976). Smoothing noisy data with spline functions; Estimating the correct degree of smoothing by the method of generalized cross-validation, in preparation.

[7] Cogburn, R. and Davis, H.T. (1974). Periodic Splines and Spectra Estimation. Ann. Statist., 2, 1108-1126.

[8] Dempster, A.P., (1973). Alternatives to least squares in multiple regression, in Multivariate Statistical Inference, Proceedings of the Research Seminar at Dalhousie University, Halifax, March 23-25, 1972. D.G. Kabe and R.P. Gupta, eds.

[9] Efron, B. and Morris, C., (1975). Data analysis using Stein's estimator and its generalizations, JASA, 70, 350, 311-319.

[10] Geisser, S., (1975). The predictive sample reuse method with applications, JASA, 70, 350, 320-328.

[11] Goldstein, M. and Smith, A.F.M., (1974). Ridge-type estimators for regression analysis, J. Roy. Stat. Soc. Ser. B. 36, 2, 284-291.

[12] Golomb, M., (1968). Approximation by periodic spline interpolants on uniform meshes, J. Approx. Thy. 1, 26-65.

[13] Golub, G., Heath, M., and Wahba, G., (1975). Cross-validation and optimum ridge regression,abstract in Abstracts of Papers To Be Presented at the SIAM-SIGNUM 1975 Fall Meeting, December 3, 4, 5, 1975, San Francisco. Technical Report to appear.

[14] Golub, G.H. and Reinsch, C., (1970). Singular value decomposition and least squares solutions, Numer. Math. 14, 403-420.

[15] Good, I.J. and Gaskins, R.A., (1971). Non-parametric roughness penalties for probability densities. Biometrika 58, 255-277.

[16] Hoerl, A.E., Kennard, R.W., and Baldwin, K.F., (1975). Ridge regression: some simulations. Comm. Statist. 4, 2, 105-123.

[17] Hudson, H.M., (1974). Empirical Bayes estimation, Technical Report #58, Stanford University Department of Statistics, Stanford, California.

[18] Kimeldorf, George and Wahba, Grace, (1971). Some results on Tchebycheffian spline functions, J. Math. Anal. Appl. 33, 82-95.

[19] Parzen, E., (1962). On the estimation of a probability density function and mode. Ann. Math. Statist. 33, 1065-1076.

[20] Reinsch, C.H., (1967). Smoothing by spline functions, Numer. Math. 10, 177-183.

[21] Reinsch, C.H., (1971). Smoothing by spline functions, II. Numer. Math. 16, 451-454.

[22] Schoenberg, I.J., (1964). Spline functions and the problem of graduation. Proc. Nat. Acad. Sci. (USA) 52, 947-950.

[23] Stone, M., (1974). Cross-validatory choice and assessment of statistical prediction, JRSS, Series B, 36, 2, 111-147.

[24] Wahba, G., (1968). On the distribution of some statistics useful in the analysis of jointly stationary time series. Ann. Math. Statist. 39, 6, 1849-1862.

[25] Wahba, Grace, (1975). Optimal convergence properties of variable knot kernel, and orthogonal series methods for density estimation. Ann. Statist. 3, 15-29.

[26] Wahba, G., (1975). Practical approximate solutions to linear operator equations when the data are noisy, University of Wisconsin-Madison, Department of Statistics, Technical Report #430. To appear, SIAM J. Num. Anal.

[27] Wahba, G., (1976). A canonical form for the problem of estimating smooth surfaces, University of Wisconsin-Madison, Department of Statistics, Technical Report #420.

[28] Wahba, G., (1976). Optimally smoothed density estimates for classification, to appear, Proceedings of the Advanced Seminar on Classification and Clustering held May 3-5, 1976 at Madison, Wisconsin. John Van Ryzin, ed.

[29] Wahba, G. and Wold, S., (1975). A completely automatic French curve: Fitting spline functions by cross-validation, Comm. Statist. 4, 1-7.

[30] Wahba, G., and Wold, S., (1975). Periodic splines for spectral density estimation: The use of cross-validation for determining the degree of smoothing, Comm. Statist. 4, 2, 125-121.

[31] Woodroofe, M., (1970). On Choosing a delta-sequence. Ann. Math. Statist., 41, 166-171.