DEPARTMENT OF STATISTICS

TECHNICAL REPORT NO. 491

GENERALIZED CROSS-VALIDATION AS A METHOD FOR CHOOSING A GOOD RIDGE PARAMETER

by

Gene H. Golub
Stanford University

Michael Heath
Stanford University

Grace Wahba
University of Wisconsin

UNIVERSITY OF WISCONSIN-MADISON

DEPARTMENT OF STATISTICS

University of Wisconsin
Madison, Wisconsin 53706

TECHNICAL REPORT NO. 491

May 1977

GENERALIZED CROSS-VALIDATION AS A
METHOD FOR CHOOSING A GOOD RIDGE PARAMETER

by

Gene H. Golub
Department of Computer Sciences
Stanford University

Michael Heath
Department of Computer Sciences
Stanford University

Grace Wahba
Department of Statistics
University of Wisconsin

TYPIST: Mary E. Arthur

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

This report is also being issued as a Stanford University Department of Computer Sciences Technical Report.

GENERALIZED CROSS-VALIDATION AS A
METHOD FOR CHOOSING A GOOD RIDGE PARAMETER

by

Gene H. Golub*, Michael Heath+, and Grace Wahba#

Consider the ridge estimate $\hat{\beta}(\lambda)$ for β in the model $y = X\beta + \varepsilon$, $\varepsilon \sim \mathcal{N}(0,\sigma^2 I)$, σ^2 unknown, $\hat{\beta}(\lambda) = (X^T X + n\lambda I)^{-1} X^T y$. We study the method of generalized cross-validation (GCV) for choosing a good value $\hat{\lambda}$ for λ, from the data. The estimate $\hat{\lambda}$ is the minimizer of $V(\lambda)$ given by

$$V(\lambda) = \frac{1}{n}||(I - A(\lambda))y||^2 / [\frac{1}{n} \text{Trace}(I - A(\lambda))]^2 ,$$

where $A(\lambda) = X(X^T X + n\lambda I)^{-1} X^T$. This estimate is a rotation-invariant version of Allen's PRESS, or ordinary cross-validation. This estimate behaves like a risk improvement estimator, but does not require an estimate of σ^2, so can be used when n-p is small, or even if $p \geq n$ in certain cases. The GCV method can also be used in subset selection and singular value truncation methods for regression, and even to choose from among mixtures of these methods.

*Research supported in part under Energy Research and Development Administration Grant E(04-3) 326 PA #30 and in part under U.S. Army Grant DAHC04-75-G-0185.

+Research supported in part under Energy Research and Development Administration Grant E(04-3) 326 PA #30.

#Research supported in part under U.S. Air Force Grant AF-AFOSR-2363-C, and by the Science Research Council (GB).

1. Introduction

Consider the standard regression model

$$y = X\beta + \varepsilon \tag{1.1}$$

where y and ε are column n-vectors, β is a p-vector and X is an $n\times p$ matrix; ε is random with $E\varepsilon = 0$, $E\varepsilon\varepsilon^T = \sigma^2 I$, where I is the $n\times n$ identity.

For $p \geq 3$, it is known that there exist estimates of β with smaller mean square error than the minimum variance unbiased, or Gauss-Markov, estimate $\hat{\beta}(0) = (X^TX)^{-1}X^Ty$. (See Berger [8], Thisted [37], for recent results and references to the earlier literature.) Allowing a bias may reduce the variance tremendously.

In this paper we primarily consider the (one parameter) family of ridge estimates $\hat{\beta}(\lambda)$ given by

$$\hat{\beta}(\lambda) = (X^TX + n\lambda I)^{-1}X^Ty\ . \tag{1.2}$$

The estimate $\hat{\beta}(\lambda)$ is the posterior mean of β if β has the prior $\beta \sim \mathcal{N}(0, aI)$, and $\lambda = \sigma^2/na$. $\hat{\beta}(\lambda)$ is also the solution to the problem:

Find β which satisfies the constraint

$$||\beta|| = \gamma$$

and for which

$$\frac{1}{n}||y - X\beta|| = \min.$$

Here $||\cdot||$ indicates the Euclidean norm and we use this norm throughout the paper. Introducing the Lagrangian we find that the

above problem is equivalent to finding the minimum of

$$\frac{1}{n}||y - X\beta||^2 + \lambda||\beta||^2 \tag{1.3}$$

where λ is a Lagrange multiplier. Methods for computing λ given γ are given in [16]. See [28] for discussion of (1.3). The method of minimizing equation (1.3), or its Hilbert space generalizations, is called the method of regularization in the approximation theory literature, see [20,40] for further references.

It is known that for any problem there is a $\lambda > 0$ for which the expected mean square error $E||\beta - \hat{\beta}(\lambda)||^2$ is less than the Gauss-Markov estimate, however the λ which minimizes, say, $E||\beta - \hat{\beta}(\lambda)||^2$, or any other given non-trivial quadratic loss function depends on σ^2 and the unknown β.

There has been a substantial amount of interest in estimating a good value of λ from the data. See [10,11,12,14,19,21,22,24,29,30,31 34,36,37]. A conservative guess might put the number of published estimates for λ at several dozen.

In this paper we examine the properties of the method of generalized cross-validation (GCV) for obtaining a good estimate of λ from the data. The GCV estimate of λ in the ridge estimate (1.2) is the minimizer of $V(\lambda)$ given by

$$V(\lambda) = \frac{1}{n}||(I - A(\lambda)y||^2/[\frac{1}{n}\text{Trace}\ (I - A(\lambda))]^2\ , \tag{1.4}$$

where

$$A(\lambda) = X(X^TX + n\lambda I)^{-1}X^T\ . \tag{1.5}$$

A discussion of the source of $V(\lambda)$ will be given in Section 2. This estimate is a rotation-invariant version of Allen's PRESS or ordinary cross-validation, as described in Hocking's discussion to Stone's paper [35], see also Allen [3], and Geisser [13].

Let $T(\lambda)$ be the "true" predictive mean square error, defined by

$$T(\lambda) = \frac{1}{n}||X\beta - X\hat{\beta}(\lambda)||^2 . \tag{1.6}$$

It is straight-forward to show that

$$ET(\lambda) = \frac{1}{n}||(I - A(\lambda))g||^2 + \frac{\sigma^2}{n} \operatorname{Tr} A^2(\lambda) \tag{1.7}$$

where

$$g = X\beta .$$

An unbiased estimator $\hat{T}(\lambda)$ of $ET(\lambda)$ is, for $n > p$, given by

$$\hat{T}(\lambda) = \frac{1}{n}||(I - A(\lambda))y||^2 - \frac{2\hat{\sigma}^2}{n} \operatorname{Tr}(I - A(\lambda)) + \hat{\sigma}^2 , \tag{1.8}$$

where

$$\hat{\sigma}^2 = \frac{1}{n-p}||(I - X(X^TX)^{-1}X^T)y||^2 .$$

Mallows [27,p.672] has suggested choosing λ to minimize Mallows' C_L, which is equivalent to minimizing $n\hat{T}(\lambda)/\hat{\sigma}^2$. (This follows from [27] upon noting that $||(I-A(\lambda))y||^2$ is the "residual sum of squares".) The minimizer of $\hat{T}$ was also suggested by Hudson [24]. We shall call an estimate formed by minimizing $\hat{T}$ an RR ("range risk") estimate.

We shall show that the GCV estimate is, for large n an estimate for the λ which approximately minimizes $ET(\lambda)$ of (1.7), without the necessity of estimating σ^2. As a consequence of not needing an estimate of σ^2, GCV can be used on problems where n-p is small, or (in certain circumstances), where the "real" model may be

$$y_i = \sum_{j=1}^{\infty} x_{ij}\beta_j + \epsilon_i, \qquad i = 1,2,\ldots,n . \tag{1.9}$$

It is also natural for solving regression-like problems that come from an attempt to solve ill-posed linear operator equations numerically.

In these problems there is typically no way of estimating σ^2 from the data. See Hanson [18], Hilgers [20], Varah [38] for descriptions of these problems. See Wahba [40] for the use of GCV in estimating λ in the context of ridge-type approximate solutions for ill-posed linear operator equations, and for further references to the numerical analysis literature. See Wahba and Wahba and Wold [41,42,43,44] for the use of GCV for curve smoothing, numerical differentiation, and the optimal smoothing of density and spectral density estimation. At the time of this writing, the only other methods we know of for estimating λ from the data without either knowledge of or an estimate of σ^2, are PRESS and maximum likelihood, to be described. We shall indicate why GCV can be expected to be generally better than either. (PRESS and GCV will coincide if XX^T is a circulant matrix).

A fundamental tool in our analysis and in our computations is the singular value decomposition. Given any $n \times p$ matrix X, we may write

$$X = UDV^T$$

where U is an $n \times n$ orthogonal matrix, V is a $p \times p$ orthogonal matrix, and D is a diagonal matrix whose entries are the square roots of the eigenvalues of X^TX. The number of non-zero entries in D is equal to the rank of X. The singular value decomposition arises in a number of statistical applications [17]. Good numerical procedures are given in [15].

In Section 2 we derive the GCV estimate as a rotation-invariant version of Allen's PRESS and discuss why it should be generally superior to PRESS. In Section 3 we give some theorems concerning its properties. In Section 4 we show how GCV can be used in other regression procedures, namely, subset selection, and eigenvalue truncation, or principal components. Indeed GCV can be used to compare between the best of the three different methods, or mixtures, of them, if you will. In Section 5 we present the results of a Monte Carlo example.

2. The Generalized Cross-Validation Estimate of λ as an Invariant Version of Allen's PRESS

The Allen's PRESS, or ordinary cross-validation estimate of λ goes as follows. Let $\beta^{(k)}(\lambda)$ be the ridge estimate (1.2) of β with the kth data point y_k, omitted. The argument is, that if λ is a good choice, then the kth component $[X\beta^{(k)}(\lambda)]_k$ of $X\beta^{(k)}(\lambda)$ should be a good predictor of y_k. Therefore, the Allen's PRESS estimate of λ is the minimizer of

$$P(\lambda) = \frac{1}{n}\sum_{k=1}^{n}([X\beta^{(k)}(\lambda)]_k - y_k)^2 . \tag{2.1}$$

It can be shown, by use of the Sherman-Morrison-Woodbury formula (see [23]), that

$$P(\lambda) = \frac{1}{n}||B(\lambda)(I - A(\lambda))y||^2 , \tag{2.2}$$

where $B(\lambda)$ is the diagonal matrix with jjth entry $1/(1-a_{jj}(\lambda))$, $a_{jj}(\lambda)$ being the jjth entry of $A(\lambda) = X(X^TX+n\lambda I)^{-1}X^T$.

Although the idea of PRESS is intuitively appealing, it can be seen that in the extreme case where the entries of X are 0 except for x_{ii}, $i = 1,2,\ldots,p$, then $[X\beta^{(k)}(\lambda)]_k$ cannot be expected to be much of a

predictor of y_k. In fact, in this case $A(\lambda)$ is diagonal, $P(\lambda) = \frac{1}{n}\sum_{k=1}^{n} y_k^2$, and so $P(\lambda)$ does not have a unique minimizer. It is reasonable to conclude that PRESS would not do very well in the near diagonal case. If β and ε both have spherical normal priors, then various arguments can be brought to bear that any good estimate of λ should be invariant under rotations of the (measurement) coordinate system. The GCV estimate is a rotation-invariant form of ordinary cross-validation. It may be derived as follows: Let the singular value decomposition [15] of X be

$$X = UDV^T .$$

Let W be the unitary matrix which diagonalizes the circulants. (See Bellman [6], Wahba [39].) In complex form the jkth entry $[W]_{jk}$ of W is

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

The GCV estimate for λ can be defined as the result of using Allen's PRESS on the transformed model

$$\tilde{y} = WU^Ty = WDV^T\beta + WU^T\varepsilon$$

$$\equiv \tilde{X}\beta + WU^T\varepsilon .$$

The new "data vector" is $\tilde{y} = (\tilde{y}_1,\ldots,\tilde{y}_n)^T$, and the new "design matrix" is $\tilde{X} = WDV^T$. $\tilde{X}\tilde{X}^*$ is a circulant matrix (see [6,39]). Thus intuitively,

$[\tilde{X}\beta^{(k)}(\lambda)]_k$ should contain a "maximal" amount of information about $\tilde{y}_k$, on the average. By substituting $\tilde{X}$ and $\tilde{y}$ into (2.2), and observing that $\tilde{X}(\tilde{X}^*\tilde{X} + n\lambda I)^{-1}\tilde{X}^*$ is a circulant matrix and hence constant down the diagonals, it is seen that $P(\lambda)$ becomes $V(\lambda)$ given by

$$V(\lambda) = \frac{1}{n}||(I - A(\lambda)y||^2/[\frac{1}{n}\,\mathrm{Tr}(I - A(\lambda))]^2 \qquad (1.4)$$

$$\equiv \frac{1}{n}\sum_{\nu=1}^{n}\left(\frac{n\lambda}{\lambda_\nu^{(n)}+n\lambda}\right)^2 z_\nu^2 \Big/ \left[\frac{1}{n}\sum_{\nu=1}^{p}\frac{n\lambda}{\lambda_\nu^{(n)}+n\lambda} + n - p\right]^2 \qquad (2.3)$$

where $z = (z_1,\ldots,z_n)^T = U^Ty$ and $\lambda_\nu^{(n)}$, $\nu = 1,2,\ldots,n$, are the eigenvalues of XX^T, $\lambda_\nu^{(n)} = 0$, $\nu > p$.

We define the GCV estimate of λ as the minimizer of (1.4), equivalently (2.3), and proceed to an investigation of its properties.

3. Properties of the GCV Estimate of λ.

Theorem 1. (The GCV Theorem)

Let $\mu_1 = \frac{1}{n}\,\mathrm{Tr}\,A(\lambda)$, $\mu_2 = \frac{1}{n}\,\mathrm{Tr}\,A^2(\lambda)$, $b^2 = \frac{1}{n}||(I - A(\lambda))g||^2$.

Then

$$\frac{ET(\lambda) - EV(\lambda) + \sigma^2}{ET(\lambda)} = \frac{-\mu_1(2-\mu_1)}{(1-\mu_1)^2} + \frac{\sigma^2}{b^2+\sigma^2\mu_2}\,\frac{\mu_1^2}{(1-\mu_1)^2} \qquad (3.1)$$

and so

$$\frac{|ET(\lambda) - EV(\lambda) + \sigma^2|}{ET(\lambda)} < (2\mu_1 + \frac{\mu_1^2}{\mu_2})\,\frac{1}{(1-\mu_1)^2}\,.$$

Proof.

Since $ET = b^2 + \sigma^2\mu_2$ $EV = [b^2+\sigma^2(1-2\mu_1+\mu_2)]/(1-\mu_1)^2$, the result follows from

$$ET - EV = (b^2+\sigma^2\mu_2)(1 - \frac{1}{(1-\mu_1)^2}) - \sigma^2\,\frac{(1-2\mu_1)}{(1-\mu_1)^2}$$

$$ET + \sigma^2 - EV = ET(1 - \frac{1}{(1-\mu_1)^2}) + \sigma^2\,\frac{\mu_1^2}{(1-\mu_1)^2}$$

Remark: This theorem implies that if

$$\frac{1}{n}\,\mathrm{Tr}\,A(\lambda) = \mu_1 \to 0 \qquad \text{as } n \to \infty$$

and

$$(\frac{1}{n}\,\mathrm{Tr}\,A(\lambda))^2/(\frac{1}{n}\,\mathrm{Tr}\,A^2(\lambda)) = \frac{\mu_1^2}{\mu_2} \to 0 \qquad \text{as } n \to \infty$$

then the difference between $ET(\lambda) + \sigma^2$ and $EV(\lambda)$ is small compared to $ET(\lambda)$.

Corollary: Let

$$h = (2\mu_1 + \frac{\mu_1^2}{\mu_2})\,\frac{1}{(1-\mu_1)^2}$$

Let λ^0 be the minimizer of $ET(\lambda)$. Then $EV(\lambda)$ always has a (possibly local) minimum $\tilde{\lambda}$ so that the "expectation inefficiency" I^0 defined by

$$I^0 = \frac{ET(\tilde{\lambda})}{ET(\lambda^0)}$$

satisfies

$$I^0 \leq \frac{1+h(\lambda^0)}{1-h(\tilde{\lambda})}\,.$$

Remark: This Corollary says that if $h(\lambda^0)$ and $h(\tilde{\lambda})$ are small then the mean square error at the minimizer of $EV(\lambda)$ is not much bigger than the minimum possible mean square error $\min_{\lambda} ET(\lambda)$

Proof: Let $\Lambda = \{\lambda: \ 0 \le \lambda \le \infty, \ EV(\lambda) - \sigma^2 \le T(\lambda^0)(1+h(\lambda^0))\}$.

Since

$$ET(\lambda)(1-h(\lambda)) < EV(\lambda) < ET(\lambda)(1+h(\lambda)), \quad 0 \le \lambda < \infty ,$$

and ET, EV and h are continuous functions of λ, then Λ is a non-empty closed set. If 0 is not a boundary point of Λ, then $EV(\lambda)-\sigma^2$ has at least one minimum in the interior of Λ, call it $\tilde{\lambda}$. See Fig. 1. Now by the Theorem

$$ET(\tilde{\lambda})(1-h(\tilde{\lambda})) < EV(\tilde{\lambda})-\sigma^2 < ET(\lambda^0)(1+h(\lambda^0))$$

and so

$$I^0 = \frac{T(\tilde{\lambda})}{T(\lambda^0)} \le \frac{1+h(\lambda^0)}{1-h(\tilde{\lambda})} .$$

If Λ includes 0, then $\tilde{\lambda}$ may be on the boundary of Λ, i.e., $\tilde{\lambda} = 0$, but the above bound on I^0 still holds.

Example 1. Note that

$$\mu_1 = \frac{1}{n} \operatorname{Tr} A = \frac{1}{n} \sum_{\nu=1}^{p} \frac{\lambda_{\nu n}}{\lambda_{\nu n}+n\lambda} \le \frac{p}{n}$$

$$\frac{\mu_1^2}{\mu_2} = \frac{(\frac{1}{n} \operatorname{Tr} A)^2}{\frac{1}{n} \operatorname{Tr} A^2} = \frac{1}{n} \frac{(\sum_{\nu=1}^{p} \frac{\lambda_{\nu n}}{\lambda_{\nu n}+\lambda})^2}{\sum_{\nu=1}^{p} (\frac{\lambda_{\nu n}}{\lambda_{\nu n}+\lambda})^2} \le \frac{p}{n}$$

Then $h \le 3 \frac{p}{n} \frac{1}{(1-\frac{p}{n})^2}$. Hence for p fixed and $n \to \infty$, it follows

that $I^0 \le 1 + \cdot \frac{p}{n} + o(\frac{p}{n})$

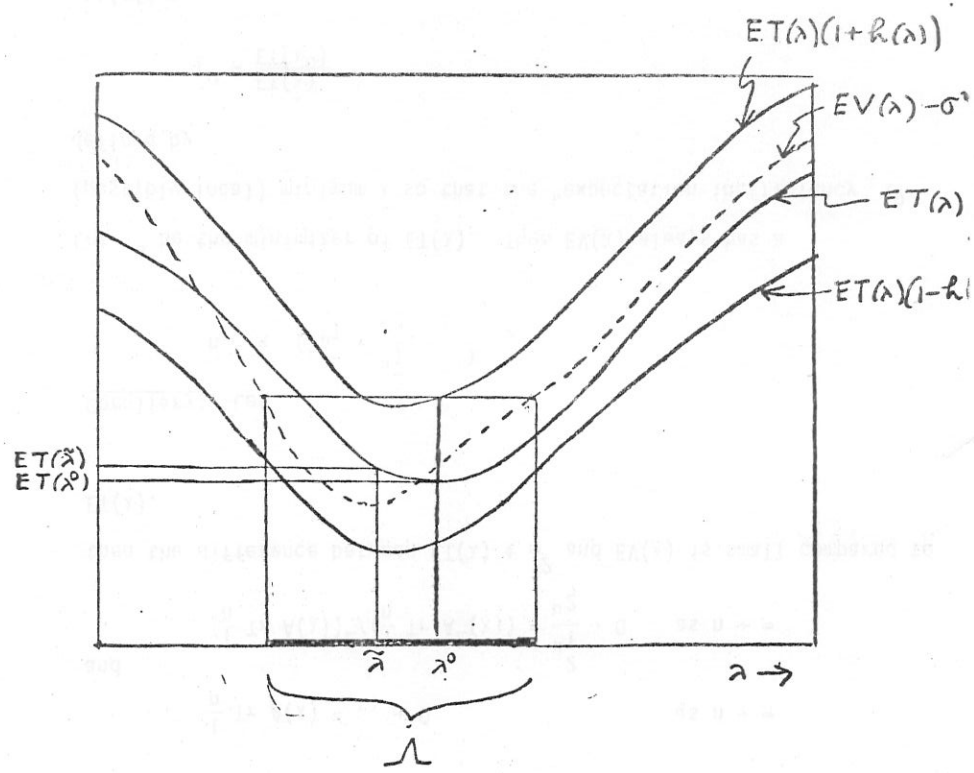

Figure 1.

Graphical Suggestion of the Proof of the Corollary to the GCV Theorem.

Example 2. $p > n$.

It is not necessary that $p << n$ for I^0 to tend to 1, as this example suggests. What is required is that XX^T become ill conditioned for n large.

Let

$$y_i = \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i, \qquad i = 1,2,\ldots, \qquad p > n \tag{3.2}$$

with

$$\sum_{j=1}^{\infty} x_{ij}^2 \le k_1 < \infty, \text{ all } i, \quad \sum_{j=1}^{\infty} \beta_j^2 \le k_2 < \infty .$$

Suppose

$$\lim_{n\to\infty} \frac{1}{n} \operatorname{Tr} XX^T = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{\infty} x_{ij}^2 = k_3 < \infty$$

and suppose the eigenvalues $\{\lambda_\nu^{(n)}, \nu = 1,2,\ldots,n\}$ of XX^T satisfy

$$\lambda_\nu^{(n)} \simeq n\, \nu^{-m} ,$$

say, for some $m > 1$. $(k_3 = \sum_{\nu=1}^{\infty} \nu^{-m})$.

Then

$$\mu_1 = \frac{1}{n} \sum_{\nu=1}^{n} \frac{\lambda_\nu^{(n)}}{\lambda_\nu^{(n)}+n\lambda} \simeq \frac{1}{n} \sum_{\nu=1}^{n} \frac{1}{1+\lambda\nu^m} \simeq \frac{1}{n} \int_0^\infty \frac{dx}{(1+\lambda x^m)} = \frac{1}{n\lambda^{1/m}} \int_0^\infty \frac{dx}{(1+x^m)}$$

$$\mu_2 = \frac{1}{n} \sum_{\nu=1}^{n} \left(\frac{\lambda_\nu^{(n)}}{\lambda_\nu^{(n)}+n\lambda}\right)^2 \simeq \frac{1}{n} \sum_{\nu=1}^{n} \frac{1}{(1+\lambda\nu^m)^2} \simeq \frac{1}{n} \int_0^\infty \frac{dx}{(1+\lambda x^m)^2} = \frac{1}{n\lambda^{1/m}} \int_0^\infty \frac{dx}{(1+x^m)^2} .$$

and $\mu_1 \to 0$, $\mu_1^2/\mu_2 \to 0$ if $n\lambda^{1/m} \to \infty$.

Now

$$b^2(\lambda) = \lambda\, \beta^T(X^TX+n\lambda I)^{-1}(n\lambda)X^TX(X^TX+n\lambda I)^{-1}\beta$$

$$\le \frac{\lambda}{2} ||\beta||^2 \le \frac{\lambda}{2} k_2 ,$$

since the largest eigenvalue of $(X^TX+n\lambda I)^{-1}(n\lambda)X^TX(X^TX+n\lambda I)^{-1}$ is

$$= \max_\nu \frac{(\lambda_\nu^{(n)})(n\lambda)}{(\lambda_\nu^{(n)})^2+(n\lambda)^2} \le \frac{1}{2} .$$

As $n \to \infty$, the minimizing sequence $\lambda^0 = \lambda^0(n)$ of $ET(\lambda) = b^2(\lambda)+\sigma^2\mu_2(\lambda)$ clearly must satisfy $\lambda^0 \to 0$, $n(\lambda^0)^{1/m} \to \infty$, so that the GCV lemma may be applied. It is proved in [9,40] in a different context that $\tilde\lambda$ as well as λ^0 satisfies $(n\lambda^{1/m}) \to \infty$ so that $h(\tilde\lambda) \to 0$, $h(\lambda^0) \to 0$ and $I^0 \to 1$ as $n \to \infty$.

Instead of viewing β as fixed but unknown, suppose that β has the prior $\beta \sim \mathcal{N}(0,aI)$. Let E_{prior} be expectation with respect to the prior. (We reserve E for expectation with respect to ϵ.) Then

Theorem 2.

The minimizer of $E_{prior}\, EV(\lambda)$ is the same as the minimizer of $E_{prior}\, ET(\lambda)$ and is $\lambda = \sigma^2/na$.

Proof:

Since $E\, g\, g^T = E\, X\beta\beta^TX^T = a\, XX^T$,

$$E_{prior}\, ET(\lambda) = \frac{a}{n} \operatorname{Tr} (I-A)^2\, XX^T + \frac{\sigma^2}{n} \operatorname{Tr} A^2$$

$$E_{prior}\, EV(\lambda) = [\tfrac{a}{n} \operatorname{Tr} (I-A)^2XX^T + \tfrac{\sigma^2}{n} \operatorname{Tr} (I-A)^2]/[\tfrac{1}{n} \operatorname{Tr} (I-A)]^2 . \tag{3.3}$$

The proof proceeds by differentiating (3.3) with respect to λ and setting the remainder equal to 0. This calculation has appeared elsewhere [42, p.8], and will be omitted.

4. GCV in Subset Selection and General Linear Model Building

Let $y = g+\varepsilon$, where g is a fixed (unknown) n-vector and $\varepsilon \sim \mathcal{N}(0,\sigma^2 I)$, σ^2 unknown. Let $A(\nu)$, ν in some index set, be a family of symmetric non-negative definite n×n matrices and let

$$\mu_1(\nu) = \frac{1}{n} \operatorname{Tr} A(\nu)$$
$$\mu_2(\nu) = \frac{1}{n} \operatorname{Tr} A^2(\nu) .$$

Letting $T(\nu) = \frac{1}{n}||g - A(\nu)y||^2$ and $V(\nu)$ as before with $A(\lambda)$ replaced by $A(\nu)$, then (3.1) clearly holds irrespective of the nature of A.

A different way of dealing with ill conditioning in the design matrix is to reduce the number of predictor variables by choosing a subset $\beta_{j_1}, \beta_{j_2}, \ldots, \beta_{j_k}$ of the β_i's. Let ν be an index on the 2^p possible subsets of $\beta_1,\ldots,\beta_p$, let $X^{(\nu)}$ be the n×k(ν) design matrix corresponding to the ν^{th} subset, and let

$$\hat{\beta}(\nu) = (X^{(\nu)^T} X^{(\nu)})^{-1} X^{(\nu)} y$$
$$A(\nu) = X^{(\nu)}(X^{(\nu)^T} X^{(\nu)})^{-1} X^{(\nu)} y .$$

Then

$$\mu_1 = k/n, \quad \mu_1^2/\mu_2 = k/n .$$

Mallows [27] suggestion to choose the subset minimizing C_p becomes, in our notation, the equivalent of minimizing $\hat{T}(\cdot)$ of (1.8) with

$A(\lambda)$ replaced by $A(\nu)$, see also Allen [2]. This assumes that an estimate of σ^2 is available. Parzen [32] has observed that, if one prefers to choose a subset without estimating σ^2, (because one believed in the model (3.2), say), GCV can be used. The subset of size $\leq k_{max}$ with smallest V can be chosen, knowing that

$$\left|\frac{ET(\nu)-EV(\nu)-\sigma^2}{ET(\nu)}\right| \leq \frac{k_{max}}{n} ,$$

even if the model (3.2) is nontrivially true.

In the subset selection case, GCV asymptotically coincides with the use of Akaike's information criteria AIC [1] since

$$AIC = (-2) \log \text{maximum likelihood} + 2k$$
$$= n \log \frac{1}{n} ||(I-A)y||^2 + 2k$$

and so

$$e^{AIC/n} = \frac{\frac{1}{n}||(I-A)y||^2}{(e^{-\frac{k}{n}})^2} \approx \frac{\frac{1}{n}||(I-A)y||^2}{(1-\frac{k}{n})^2} = V$$

as $\frac{k}{n} \to 0$.

We thank E. Parzen for pointing this out. M. Stone, in an unpublished manuscript kindly brought to the attention of one of the authors (G.W.) by H. Tong, has also investigated relations between AIC and (ordinary) cross-validation.

Another approach, the principal components approach, is also popular in solving ill posed linear operator equations, see Baker et al [7], Hanson [18], Varah [38]). The method is to replace X by X(ν) defined by $X(\nu) = U D(\nu) V^T$, where D(ν) is the diagonal matrix of singular values of V with all but the ν^{th} subset of singular values set equal to 0. Then

$$A(\nu) = U\, D(\nu)\, (D(\nu)D(\nu)^T)^{\dagger}\, D(\nu)\, U^T$$

$$= U \begin{pmatrix} 1 & & & 0 \\ & \ddots & & \\ & & 0 & \\ 0 & & & 0 \end{pmatrix} U^T$$

where the ones are located at positions of the ν^{th} subset of singular values, and, again $\mu_1 \leq p/n$, $\mu_1^2/\mu_2 \leq p/n$, where p can be replaced by the number of singular values in the largest subset considered.

In fact, it is reasonable to select from among any family $\{A(\nu)\}$ of matrices for which the corresponding μ_1 and μ_1^2/μ_2 are uniformly small, by choosing that member for which $V(\nu)$ is smallest. Mixtures of the above methods, e.g. a ridge method on a subset can be handled this way. Note that the conditions μ_1 small, μ_1^2/μ_2 small are just those conditions which make it plausible that the "signal" g can be separated from the noise. These conditions say that the A matrix essentially maps the data vector (roughly) into some much smaller subspace than the whole space. Parzen [33] has also indicated how GCV can be used to choose the order of an autoregressive model to fit a stationary time series.

5. A Numerical Example

We choose a discretization of the Laplace transform as given in Varah, [38,p.262] as an example, to see how the method would work on a very ill conditioned problem. The values for n and p were 21 and 10 and the condition number of X, namely the ratio of the largest to the smallest (non-zero) 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 four times, giving a total of 16 runs. The ε_i were generated as pseudo-random $\mathcal{N}(0,\sigma^2)$ independent r.v.'s, $V(\lambda)$ was computed using the right hand side of (2.3) and the Golub-Reinsch singular value decomposition [15]. The minimizer $\hat{\lambda}$ of $V(\lambda)$ was determined by a global search. $T(\lambda)$ was also computed and the relative inefficiencies $\hat{I}_D$ and $\hat{I}_R$ of $\hat{\lambda}$ defined by

$$I_D = ||\beta - \hat{\beta}_{\hat{\lambda}}||^2 / (\min_{\lambda} ||\beta - \hat{\beta}_{\lambda}||^2)$$
$$I_R = T(\hat{\lambda}) / \min_{\lambda} T(\lambda) \qquad (5.1)$$

were computed. (D = "domain", R = "range")

The results of a comparison with three other methods are also presented. The methods are, respectively

1. PRESS, the minimizer of $P(\lambda)$
2. Range risk, (RR) the minimizer of $\hat{T}(\lambda)$
3. Maximum likelihood (MLE)

The maximum likelihood estimate is obtained from the model

$$y = X\beta + \varepsilon$$

with $\varepsilon \sim \mathcal{N}(0,\sigma^2 I)$ and β having the prior distribution $\beta \sim \mathcal{N}(0,aI)$. Then the posterior distribution of y is

$$y \sim \mathcal{N}(0,\ a(XX^T + n\lambda I)) \qquad (5.2)$$

where $\lambda = \sigma^2/na$. The ML estimate for λ from the model (5.2) is then the minimizer of $M(\lambda)$ given by

$$M(\lambda) = \frac{1}{n} \frac{y^T(I-A(\lambda))y}{[\mathrm{Det}(I-A(\lambda))]^{1/n}} \qquad (5.3)$$

This estimate is the general form of the maximum likelihood estimate suggested by Anderssen and Bloomfield in the context of numerical differentiation [4,5]. It can be shown that the minimizer of $E_{prior}\, E\, M(\lambda)$ is σ^2/na. However, it can also be shown that if β behaves as though it did not come from the prior (e.g. as in the model (1.9), $\sum_{i=1}^{\infty} \beta_i^2 < \infty$), then the minimizer of $E\, M(\lambda)$ may not be a good estimate of the minimizer of $ET(\lambda)$.

I_D and I_R of (5.1) were determined for each of these three 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 (5.1) with $\hat{\lambda}$ replaced by the minimizers 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 2 gives a plot of $V(\lambda)$, $\hat{T}(\lambda)$, $M(\lambda)$, $P(\lambda)$, $||\beta-\hat{\beta}_\lambda||^2$ and $T(\lambda)$ for Replicate 2 of the $\sigma^2 = 10^{-6}$ case. The $V(\lambda)$, $\hat{T}(\lambda)$ and $T(\lambda)$ curves tend to follow each other as predicted.

	Replication 1		Replication 2		Replication 3		Replication
	I_D	I_R	I_D	I_R	I_D	I_R	I_D I_R
			$\sigma^2=10^{-8}$, S/N = 4200				
GCV	4.43	1.06	1.65	1.03	16.71	1.10	1.02 1.
RR	1.46	1.00	1.66	1.03	8.69	1.01	1.22 1.
MLE	1.67E3	1.31	1.45E2	1.23	2.00E3	1.53	9.12E3 1.
PRESS	2.31E3	4.8E4	6.31E2	8.6E4	3.84E3	2.1E5	2.87E3 1.
Min Sol'n	1.00	1.02	1.00	1.54	1.00	2.27	1.00 1.
Min Data	1.20	1.00	2.89	1.00	5.97	1.00	1.00 1.
			$\sigma^2=10^{-6}$, S/N = 420				
GCV	1.92	1.05	1.32	1.00	1.51E2	1.26	2.20 1.
RR	1.83	1.06	1.90	1.01	7.03E1	1.10	1.18 1.
MLE	1.99E2	1.19	1.70E2	1.45	1.76E2	1.29	1.49E2 1.
PRESS	5.80	1.01	2.41E2	1.39E4	36.37	2.43E3	67.00 6.
Min Sol'n	1.00	1.38	1.00	1.02	1.00	1.20	1.00 1.
Min Data	3.56	1.00	1.28	1.00	7.85	1.00	41.29 1.
			$\sigma^2=10^{-4}$, S/N = 42				
GCV	1.27	1.07	1.50	2.58	1.00	1.11	1.00 1.
RR	1.18	1.08	1.03	2.27	1.07	1.13	1.00 1.
MLE	1.56	1.20	12.16	3.43	1.90	1.49	2.97 1.
PRESS	3.53	1.57	2.03	3.43	8.66	2.63	2.90 24.
Min Sol'n	1.00	1.21	1.00	2.05	1.00	1.11	1.00 1.
Min Data	3.26	1.00	1.16	1.00	2.39	1.00	1.16 1.
			$\sigma^2=10^{-2}$, S/N = 4.2				
GCV	1.40	2.47	2.01	1.60	1.59	1.01	31.20 17.
RR	1.38	2.39	2.41	1.79	1.41	1.02	10.8 10.
MLE	2.13	3.56	3.81	1.87	2.00	1.00	28.8 16.
PRESS	1.04	1.01	2.02	2.68	1.00	1.22	2.16 21.
Min Sol'n	1.00	1.31	1.00	1.01	1.00	1.25	1.00 1.
Min Data	1.02	1.00	1.00	1.00	2.66	1.00	1.21 1.

Observed Inefficiencies in Sixteen Monte Carlo Runs

Table 1

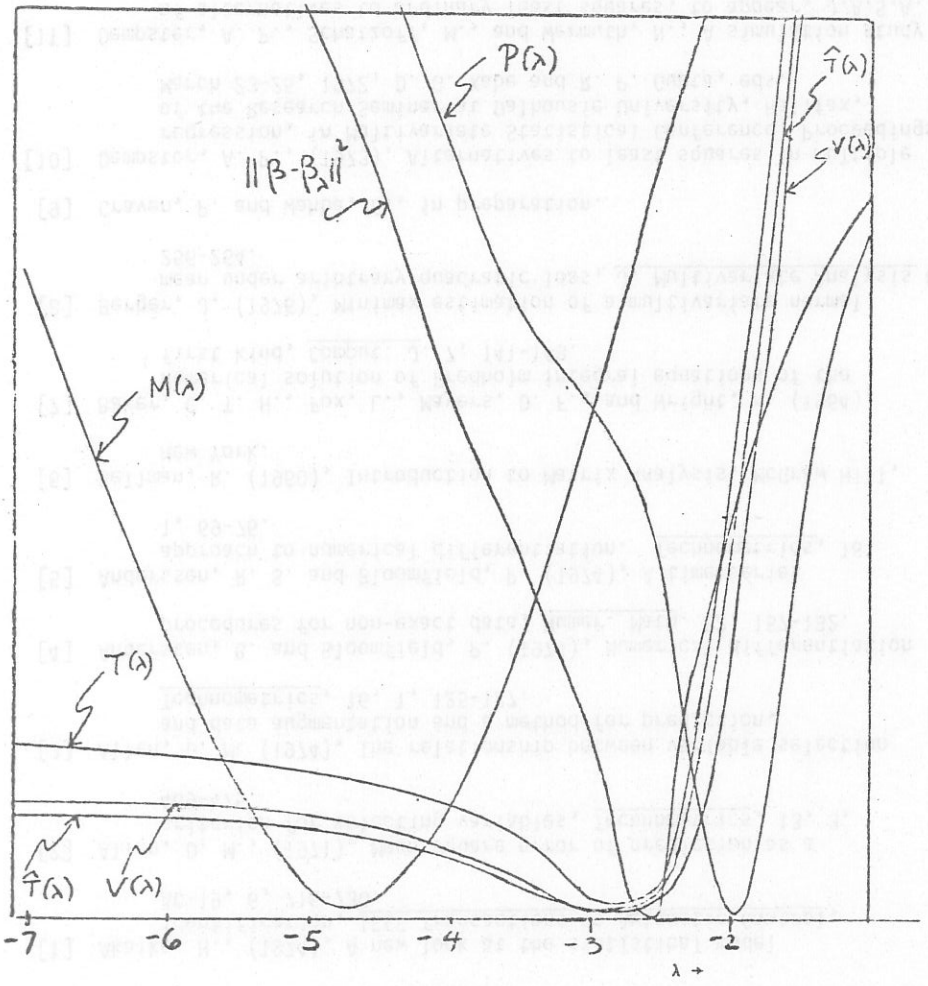

Figure 2.

ACKNOWLEDGMENT

The work of Gene H. Golub was initiated while a guest of the Eidgenossische Technische Hochschule. He is very pleased to acknowledge the gracious hospitality and stimulating environment provided by Professors Peter Henrici and Peter Huber.

The work of Grace Wahba was initiated while a visitor at the Oxford University Mathematical Institute at the invitation of Professor J. F. C. Kingman. The hospitality of Professor Kingman, the Mathematical Institute, and St. Cross College, Oxford is gratefully acknowledged.

REFERENCES

[1] Akaike, H., (1974), A new look at the statistical model identification, IEEE Transactions on Automatic Control, AC-19, 6, 716-730.

[2] Allen, D. M., (1971), Mean square error of prediction as a criterion for selecting variables, Technometrics, 13, 3, 469-475.

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

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

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

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

[7] Baker, C. T. H., Fox, L., Mayers, D. F., and Wright, K. (1964), Numerical solution of Fredholm integral equations of the first kind, Comput. J. 7, 141-148.

[8] Berger, J. (1976), Minimax estimation of a multivariate normal mean under aribtrary quadratic loss, J. Multivariate Analysis 6, 256-264.

[9] Craven, P. and Wahba, G., in preparation.

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

[11] Dempster, A. P., Schatzoff, M., and Wermuth, N., A simulation study of alternatives to ordinary least squares, to appear, J.A.S.A.

[12] Farebrother, R. W. (1975), The minimum mean square error linear estimator and ridge regression, Technometrics, 17, 1, 127-128.

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

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

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

[16] Golub, G. H. (1973), Some modified matrix eigenvalue problems, SIAM Review, 15, pp. 318-334.

[17] Golub, G. H. and Luk, F. T. (1977), Singular value decomposition: applications and computations, Transactions of the Twenty-Second Conference of Army Mathematicians, pp. 577-605.

[18] Hanson, R. J. (1971), A numerical method for solving Fredholm integral equations of the first kind using singular values, SIAM J. Num. Anal., 8, 616-622.

[19] Hemmerle, W. J. (1975), An explicit solution for generalized ridge regression, Technometrics, 17, 3, 309-313.

[20] Hilgers, J. W. (1976), On the equivalence of regularization and certain reproducing kernel Hilbert space approaches for solving first kind problems, SIAM J. Num. Anal., 13, 2, 172-184.

[21] Hoerl, A. E. and Kennard, R. W. (1976), Ridge regression; iterative estimation of the biasing parameter, Commun. Statis. A5(1), 77-88.

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

[23] Householder, A., (1964), The Theory of Matrices in Numerical Analysis, Blaisdell, New York.

[24] Hudson, H. M. (1974), Empirical Bayes Estimation, Technical Report #58, Stanford University, Department of Statistics, Stanford, CA.

[25] Lawless, J. F. and Wang, P. (1976), A simulation study of ridge and other regression estimators, Commun. Statist. A5(4), 307-324.

[26] Lindley, D. V. and Smith, A. F. M. (1972), Bayes estimates for the linear model (with discussion), J. Roy. Stat. Soc. B, 34, part 1 1-41.

[27] Mallows, C. L. (1973). Some comments on C_p. Technometrics 15, 661-675.

[28] Marquardt, D. W. (1970), Generalized inverses, ridge regression, biased linear estimation and nonlinear estimation, Technometrics 12, 591-64.

[29] Marquardt, D. W. and Snee, R. D. (1975), Ridge regression in practice, The American Statistician, 29, 1, 3-20.

[30] McDonald, C. and Galarneau, D. (1975), A Monte-Carlo evaluation of some ridge-type estimators. JASA, 70, 407-416.

[31] Obenchain, R. L. (1975), Ridge analysis following a preliminary test of the shrunken hypothesis, Technometrics, 17, 4

[32] Parzen, E. (1976), Time series theoretic nonparametric statistical methods, Preliminary Report, Statistical Science Division, SUNY, Buffalo, New York.

[33] Parzen, E. (1977), Forecasting and whitening filter estimation, manuscript.

[34] Rolph, J. E. (1976), Choosing shrinkage estimators for regression problems. Commun. Statist. - Theor. Meth., A5(9), 789-802.

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

[36] Swindel, B. F. (1976), Good ridge estimators based on prior information. Commun. Statist. - Theor. Meth., A5(1) 985-997.

[37] Thisted, R. A. (1976), Ridge regression, minimax estimation, and empirical Bayes methods. Division of Biostatistics, Stanford University TR#28.

[38] Varah, J. M. (1973), On the numerical solution of ill-conditioned linear systems with applications to ill posed problems, SIAM J. Num. Anal. 10, 2, 257-267.

[39] 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.

[40] Wahba, G. (1977), The approximate solution of linear operator equations when the data are noisy. SIAM J. Num. Anal., 14, 4, to appear.

[41] Wahba, G. (1976), A survey of some smoothing problems and the method of generalized cross-validation for solving them. Department of Statistics, University of Wisconsin-Madison TR#457. To appear, Conference on the Applications of Statistics, held at Dayton, Ohio, June 14-17, 1976, P. R. Krishnaiah, ed.

[42] Wahba, G. (1976), Optimal smoothing of density estimates. Department of Statistics, University of Wisconsin-Madison, TR#469. To appear, Proceedings of the Conference on Classification and Clustering, held at Madison, WI., May 3-5, 1976, J. Van Ryzin, ed.

[43] Wahba, G. and Wold, S. (1975), Periodic splines for spectral density estimation: The use of cross-validation for determining the correct degree of smoothing, Commun. Statist. 4, 2, 125-141.

[44] Wahba, G. and Wold, S. (1975), A completely automatic French curve: Fitting spline functions by cross-validation, Commun. Statist. 4, 1, 1-17.