Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter

Gene H. Golub
Department of Computer Science
Stanford University
Stanford, CA 94303

Michael Heath
Computer Sciences Division
Oak Ridge National Laboratory
Oak Ridge, TN 37830

Grace Wahba
Department of Statistics
University of Wisconsin
Madison, WI 53705

Consider the ridge estimate $\hat{\beta}(\lambda)$ for β in the model $y = X\beta + \epsilon$, $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$, 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

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

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.

KEY WORDS

Ridge regression
Cross-validation
Ridge parameter

1. INTRODUCTION

Consider the standard regression model

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

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 $E\epsilon = 0$, $E\epsilon\epsilon^T = \sigma^2 I$, 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 $\hat{\beta}(0) = (X^T X)^{-1} X^T y$. (See Berger [8], Thisted [39], 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

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

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

Received June 1977; revised April 1978

Lagrangian we find that the above problem is equivalent to finding the minimum over β 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 [17]. See [29] 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 [21, 44] 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 nontrivial 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, 15, 20, 22, 23, 25, 26, 27, 30, 31, 32, 35, 38, 39]. 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

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

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 [36] (see also Allen [3], and Geisser [13]).

Let $T(\lambda)$ be the mean square error in estimating $X\beta$, that is,

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

It is straightforward 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 $ET(\lambda)$, for $n > p$, is 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 [28, 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 [28] upon noting that $\|(I - A(\lambda))y\|^2$ is the "residual sum of squares.") The minimizer of $\hat{T}$ was also suggested by Hudson [25]. 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

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

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 [19], Hilgers [21], Varah [40] for descriptions of these problems. See Wahba [44] 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, Wahba and Wold, and Craven and Wahba [9, 42, 43, 45, 46] for the use of GCV for curve smoothing, numerical differentiation, and the optimal smoothing of density and spectral density estimates. 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 an $n \times p$ 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 [18]. Good numerical procedures are given in [16].

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

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

It has been observed by one of the referees that $P(\lambda)$ may be viewed as a direct sample estimate of $\frac{1}{n}E_{y^*}\|y^* - X\hat{\beta}(\lambda)\|^2 \equiv T(\lambda) + \sigma^2$, where here $\hat{\beta}(\lambda)$ is supposed fixed, y^* is a future hypothetical observation vector, and E_{y^*} denotes expectation over the distribution of y^*.

It can be shown, by use of the Sherman-Morrison-Woodbury formula (see [24]), that

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

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}, $i = 1, 2, \cdots, p$, then $[X\beta^{(k)}(\lambda)]_k$ cannot be expected to be a good 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 [16] of X be

$$X = UDV^T.$$

Let W be the unitary matrix which diagonalizes the circulants. (See Bellman [7], Wahba [41].) In complex form the jkth entry $[W]_{jk}$ of W is

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

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

$$\tilde{y} = WU^Ty = WDV^T\beta + WU^T\epsilon \equiv \tilde{X}\beta + WU^T\epsilon.$$

The new "data vector" is $\tilde{y} = (\tilde{y}_1, \cdots, \tilde{y}_n)^T$, and the new "design matrix" is $\tilde{X} = WDV^T$. $\tilde{X}\tilde{X}^*$ ("*" means complex conjugate transpose) is a circulant matrix (see [6,41]). 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 $\tilde{A}(\lambda) \equiv \tilde{X}(\tilde{X}^*\tilde{X} + n\lambda I)^{-1}\tilde{X}^*$ is a circulant matrix and hence constant down the diagonals, and $\tilde{A}(\lambda)$ and $A(\lambda)$ have the same eigenvalues, it is seen that $P(\lambda)$ becomes $V(\lambda)$ (see (1.4)) given by

$$V(\lambda) = \frac{1}{n}\|(I - \tilde{A}(\lambda))\tilde{y}\|^2 \Big/ \left[\frac{1}{n}\mathrm{Tr}(I - \tilde{A}(\lambda))\right]^2$$
$$\equiv \frac{1}{n}\sum_{\nu=1}^{n}\left(\frac{n\lambda}{\lambda_{\nu n} + n\lambda}\right)^2 z_\nu^2 \Big/ \left[\frac{1}{n}\sum_{\nu=1}^{p}\frac{n\lambda}{\lambda_{\nu n} + n\lambda} + n - p\right]^2 \qquad (2.3)$$

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

It can also be shown that $V(\lambda)$ is a weighted version of $P(\lambda)$, namely

$$V(\lambda) \equiv \frac{1}{n}\sum_{k=1}^{n} ([X\beta^{(k)}(\lambda)]_k - y_k)^2 w_k(\lambda)$$

where

$$w_k(\lambda) = \frac{1 - a_{kk}(\lambda)}{1 - \frac{1}{n}\mathrm{Tr}\,A(\lambda)}.$$

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

Then

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

and so

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

whenever $0 < \mu_1 < 1$.

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

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

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

Remark This theorem implies that if

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

and

$$\left(\frac{1}{n}\operatorname{Tr} A(\lambda)\right)^2 \Big/ \left(\frac{1}{n}\operatorname{Tr} A^2(\lambda)\right) = \frac{\mu_1^2}{\mu_2} \to 0 \quad \text{as } n \to \infty$$

then the difference between $ET(\lambda) + \sigma^2$ and $EV(\lambda)$ is small compared to $ET(\lambda)$. This result and the fact that in the extreme diagonal case $P(\lambda)$ does not have a unique minimum suggests that the minimizer of $V(\lambda)$ is preferable to the minimizer of $P(\lambda)$ if one wants to choose λ to minimize

$$\frac{1}{n}E_{y^*}\|y^* - X\beta(\lambda)\|^2.$$

Corollary Let

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

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

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

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

Since

$$ET(\lambda)(1 - h(\lambda)) < EV(\lambda) - \sigma^2 < ET(\lambda)(1 + h(\lambda)), \quad 0 \le \lambda < \infty,$$

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 Figure 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

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

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

Then

$$h \le 3\frac{p}{n}\frac{1}{\left(1 - \frac{p}{n}\right)^2}$$

Hence for p fixed and $n \to \infty$, it follows that

$$I^0 \le 1 + 6\frac{p}{n} + 0\left(\frac{p}{n}\right)$$

Example 2 $p > n$.

It is not necessary that $p \ll n$ 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

$$y_i = \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i, \qquad i = 1, 2, \cdots,$$

$$p > n \tag{3.2}$$

with

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

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 $\{\lambda_{\nu n}, \nu = 1, 2, \cdots, n\}$ 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

$$(X^TX + n\lambda I)^{-1}(n\lambda)X^TX(X^TX + n\lambda I)^{-1}$$

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

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 Theorem may be applied. It is proved in [9, 44] 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 $I^0 \mid I$ 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_β be expectation with respect to the prior. (We reserve E for expectation with respect to ϵ.) Then

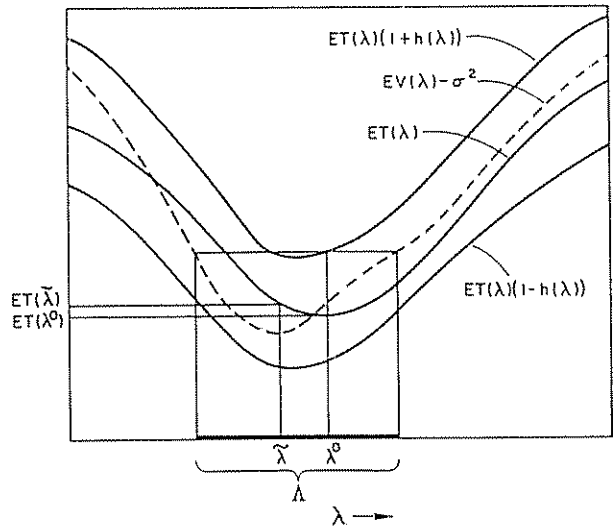

FIGURE 1 Graphical suggestion of the proof of the corollary to the GCV theorem

Theorem 2

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

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

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

$$E_\beta EV(\lambda) = \left[\frac{a}{n} \operatorname{Tr} (I - A)^2XX^T + \frac{\sigma^2}{n} \operatorname{Tr} (I - A)^2\right] \Big/ \left[\frac{1}{n} \operatorname{Tr} (I - A)\right]^2 \tag{3.3}$$

The proof proceeds by differentiating (3.3) with respect to λ and setting the remainder equal to 0. This calculation has appeared elsewhere [43 p. 8], and will be omitted.

4 GCV IN SUBSET SELECTION AND GENERAL LINEAR MODEL BUILDING

Let $y = g + \epsilon$, where g is a fixed (unknown) n-vector and $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$, σ^2 unknown. Let $A(\nu)$, ν in some index set, be a family of symmetric nonnegative definite $n \times n$ 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 $\beta_{j_1}, \beta_{j_2}, \cdots, \beta_{j_k}$ of the β_i's. Let ν be an index on the 2^p possible subsets of $\beta_1, \cdots, \beta_p$, let $X^{(\nu)}$ be the $n \times k(\nu)$ design matrix corresponding to the ν^{th} subset, and let

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

$$A(\nu) = X^{(\nu)} (X^{(\nu)T} X^{(\nu)}) X^{(\nu)}$$

Then

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

Mallows [28] suggestion to choose the subset minimizing C_p becomes, in our notation, the equivalent of minimizing $\tilde{T}(\cdot)$ 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 [33] 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 criterion AIC [1] since

$$\text{AIC} = (-2) \log \text{maximum likelihood} + 2k$$

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

and so

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

as

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

We thank E. Parzen for pointing this out. M. Stone, [37] has investigated the 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. [6], Hanson [19], Varah [40]). The method is to replace X by $X(\nu)$ defined by $X(\nu) = U D(\nu) V^T$, where $D(\nu)$ is the diagonal matrix of singular values of X with all but the ν^{th} subset of singular values set equal to 0. Then

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

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

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 [34] 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, [40, p. 262] as an example in which $X^T X$ is very ill conditioned.

We emphasize that the following is nothing more than a single example, with a single X and β. It does not indicate what may happen as X and β are varied. It is intended as an indication of the type of Monte Carlo evaluation study that an experimenter might perform with the particular X that he has at hand, and perhaps one or several β that represent the class of β's he believes he is likely to encounter. We suggest that an experimenter with particular design matrix at hand evaluate candidate methods (at least crudely), perhaps including subset selection and/or principal components, as well as ridge methods against his X and against a realistic set of β, before final selection of a method. The values for n and p in the experiment presented here were 21 and 10 and the condition number of X, namely the ratio of the largest to the smallest (non-zero) singular value, was 1.54×10^3. The value of $\|X\beta\|^2$ was 370.84.

Four values of σ^2, namely $\sigma^2 = 10^{-4}, 10^{-6}, 10^{-8}$ 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 [16]. The minimizer $\hat{\lambda}$ of $V(\lambda)$ was determined by a global search. $T(\lambda)$ was also computed and the relative inefficiencies I_D and I_R of $\hat{\lambda}$ defined by

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

$$I_R = T(\hat{\lambda}) / \min_{\lambda} T(\lambda) \tag{5.1}$$

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

TABLE 1—*Observed inefficiencies in sixteen Monte Carlo runs*

	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 = 4200				
GCV	4.43	1.06	1.65	1.03	16.71	1.10	1.02	1.01
RR	1.46	1.00	1.66	1.03	8.69	1.01	1.22	1.03
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 = 420				
GCV	1.92	1.05	1.32	1.00	1.51E2	1.26	2.20	1.02
RR	1.83	1.06	1.90	1.01	7.03E1	1.10	1.18	1.00
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 = 42				
GCV	1.27	1.07	1.50	2.58	1.00	1.11	1.00	1.03
RR	1.18	1.08	1.03	2.27	1.07	1.13	1.00	1.03
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 = 4.2				
GCV	1.40	2.47	2.01	1.60	1.59	1.01	31.20	17.2
RR	1.38	2.39	2.41	1.70	1.41	1.02	10.8	10.6
MLE	2.13	3.56	3.81	1.87	2.00	1.00	28.8	16.8
PRESS	1.04	1.01	2.02	2.68	1.00	1.22	2.16	21.5
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

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

$$y = X\beta + \epsilon$$

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

$$y \sim N(0, a(XX^t + n\lambda I)) \qquad (5.2)$$

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 $E_\beta E M(\lambda)$ is σ^2/na. However, it can also be shown that if β behaves as though it did not come from the prior

$$\left(\text{e.g. as in the model (1.9).} \quad \sum_{i=1}^{\infty} \beta_i^2 < \infty\right),$$

then the minimizer of $E M(\lambda)$ may not be a good estimate of the minimizer of $ER(\lambda)$.

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 = $[(1/n)\|\beta\|^2/\sigma^2]^{1/2}$. Figure 2 gives a

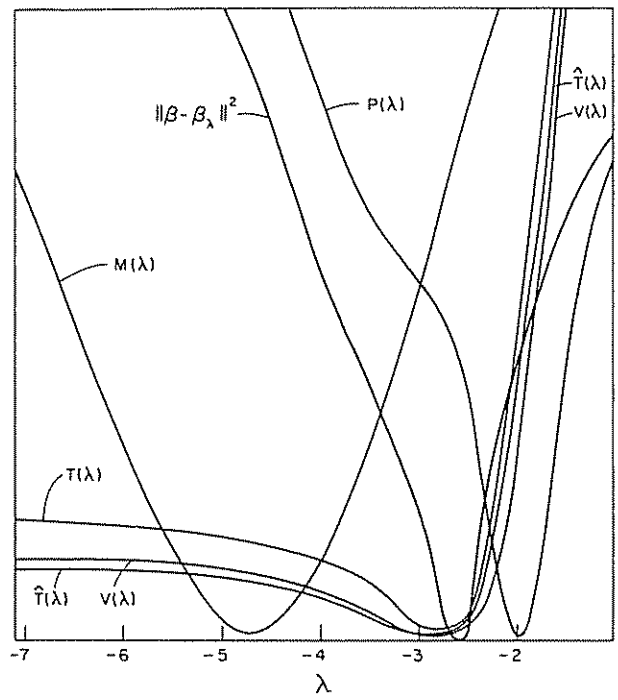

FIGURE 2 $V(\lambda)$, $T(\lambda)$, $\hat{T}(\lambda)$, $M(\lambda)$, $P(\lambda)$ and $\|\beta-\hat{\beta}\lambda\|^2$

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

D. I. Gibbons [14] has recently completed a Monte Carlo comparison of 10 methods of choosing k. Three estimators, GCV, HKB (described in [23]), and RIDGM (described in [10,11]) were identified as the best performers in the examples studied. HKB and RIDGM use estimates of σ^2.

6 CONCLUSIONS

The generalized cross-validation method for estimating the ridge parameter in ridge regression has been given. This estimate does not require an estimate of σ^2, and thus may be used when the number of degrees of freedom for estimating σ^2 is small or even; in some cases, when the "real" model actually involves more than n parameters. The method may also be used to do subset selection or selection of principal components instead of ridge regression, or even to choose between various combinations of ridge, subset selection or principal components methods. A numerical example, briefly suggestive of the behavior of the method, has been carried out. It illustrates what an experimenter might wish to do to examine the properties of the method with respect to his/her design matrix.

7 ACKNOWLEDGMENTS

The work of Gene H. Golub was initiated while a guest of the Eidgenössische Technische Hochschule. He is very pleased to acknowledge the gracious hospitality and stimulating environment provided by Professors Peter Henrici and Peter Huber. His research was supported in part under Energy Research and Development Administration Grant E(04-3) PA # 30, and in part under U.S. Army Grant DAHC04-75-G-0185.

Michael Heath's research was supported in part under Energy Research and Development Administration Grant E(04-3) 326 PA # 30.

The work of Grace Wahba was initiated while she was 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. Her research was supported by the Science Research Council (GB), and by U. S. Air Force Grant AF-AFOSR-2363-C.

REFERENCES

[1] AKAIKE, H. (1974) A new look at the statistical model identification *IEEE Transaction 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.* 469–475

[3] ALLEN, D. M. (1974) The relationship between variable selection and data augmentation and a method for prediction *Technometrics. 16.* 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.* 69–75.

[6] BAKER, C. T. H., FOX, L., MAYERS, D. F., and WRIGHT, K. (1964) Numerical solution of Fredholm integral equations of the first kind *Comp J. 7.* 141–148

[7] BELLMAN, R. (1960) *Introduction to Matrix Analysis* New York: McGraw-Hill

[8] BERGER, J. (1976) Minimax estimation of a multivariate normal mean under arbitrary quadratric loss *J. Multivariate Analysis. 6.* 256–264

[9] CRAVEN, P. and WAHBA, G. (1979) Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation *Numer Math., 31.* 377–403

[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.* ed. by D. G. Kabe and R. P. Gupta.

[11] DEMPSTER, A. P., SCHATZOFF, M., and WERMUTH, N. (1975) A simulation study of alternatives to ordinary least squares *J. Amer. Statist. Assoc., 70.* 77–106.

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

[13] GEISSER, S. (1975) The predictive sample reuse method with applications *J. Amer. Statist. Assoc., 70.* 320–328

[14] GIBBONS, D. I. (1978) A simulation study of some ridge estimators General Motors Research Laboratories, Research Publication GMR-2659, Warren, Michigan

[15] GOLDSTEIN, M., and SMITH, A. F. M. (1974) Ridge type estimators for regression analysis *J. Roy Statist Soc., Ser B. 36.* 284–291.

[16] GOLUB, G , and REINSCH, C (1970) Singular value decomposition and least squares solutions *Numer Math . 14*. 403–420

[17] GOLUB, G H (1973) Some modified matrix eigenvalue problems *SIAM Review, 15*. 318–334

[18] GOLUB, G H and LUK, F T (1977) Singular value decomposition: applications and computations *Transactions of the Twenty-Second Conference of Army Mathematicians*. 577–605

[19] 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

[20] HEMMERLE, W J (1975) An explicit solution for generalized ridge regression *Technometrics, 17*. 309–313

[21] 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*. 172–184

[22] HOERL, A E , and KENNARD, R W (1976) Ridge regression: iterative estimation of the biasing parameter *Comm in Statist . A5*, 77–88

[23] HOERL, A E . KENNARD, R W , and BALDWIN, K F (1975) Ridge regression: some simulations *Comm in Statist . 4*. 105–123

[24] HOUSEHOLDER, A (1964) *The Theory of Matrices in Numerical Analysis* New York: Blaisdell

[25] HUDSON, H M (1974) Empirical Bayes estimation Technical Report No 58. Stanford University, Department of Statistics, Stanford, CA

[26] LAWLESS, J F and WANG, P (1976) A simulation study of ridge and other regression estimators *Comm in Statist . A5*. 307–324

[27] LINDLEY, D V , and SMITH, A F M (1972) Bayes estimate for the linear model (with discussion), part I *J Roy Statist Soc . B, 34*. 1–41

[28] MALLOWS, C L (1973) Some comments on C_p *Technometrics, 15*, 661–675

[29] MARQUARDT, D W (1970) Generalized inverses, ridge regression, biased linear estimation and nonlinear estimation *Technometrics, 12*. 591–64

[30] MARQUARDT, D W , and SNEE, R D (1975) Ridge regression in practice *The American Statistician. 29*. 3–20

[31] MCDONALD, G and GALARNEAU, D (1975) A Monte Carlo evaluation of some ridge-type estimators *J Amer Statist Assoc . 70*. 407–416

[32] OBENCHAIN, R L (1975) Ridge Analysis following a preliminary test of the shrunken hypothesis *Technometrics. 17*. 431–446

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

[34] PARZEN, E (1977) Forecasting and whitening filter estimation Manuscript

[35] ROLPH, J E (1976) Choosing shrinkage estimators for regression problems *Comm in Statist . A5*. 789–802

[36] STONE, M (1974) Cross-validatory choice and assessment of statistical prediction *J Roy Statist Soc , B, 36*. 111–147

[37] STONE, M (1977) An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion *J Roy Statist Soc . B . 39*. 44–47

[38] SWINDEL, B F (1976) Good ridge estimators based on prior information *Comm in Statist . A5*. 985–997

[39] THISTED, R A (1976) Ridge regression, minimax estimation, and empirical Bayes methods Division of Biostatistics, Stanford University, Tech Report No 28

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

[41] WAHBA, G (1968) On the distribution of some statistics useful in the analysis of jointly stationary time series *Ann Math Statist . 39*, 1849–1862

[42] WAHBA, G (1976) A survey of some smoothing problems and the method of generalized cross-validation for solving them In *Proceedings* of the Conference on the Applications of Statistics, held at Dayton, Ohio, June 14–17, 1976, ed by P R Krishnaiah

[43] WAHBA, G (1976) Optimal smoothing of density estimates *Classification and Clustering*, pp 423–458, ed by J Van Ryzin New York: Academic Press

[44] WAHBA, G (1977) The approximate solution of linear operator equations when the data are noisy *SIAM J Num Anal . 14*. 651–667

[45] WAHBA, G , and WOLD, S (1975) Periodic splines for spectral density estimation: the use of cross-validation for determining the correct degree of smoothing *Comm in Statist . 4*. 125–141

[46] WAHBA, G , and WOLD, S (1975) A completely automatic French curve: fitting spline functions by cross-validation *Comm in Statist . 4*. 1–17