J. Austral. Math. Soc. (Series B) 22 (1981), 501–514

THE EXTENSION OF GENERALIZED CROSS-VALIDATION TO A MULTI-PARAMETER CLASS OF ESTIMATORS

D. M. O'BRIEN and J. N. HOLT

(Received 3 September 1980)

(Revised 26 November 1980)

Abstract

The method of generalized cross-validation (GCV) provides a good value for the "ridge" regularization parameter for an ill-conditioned linear system, such as the system produced by discretization of a Fredholm integral equation of the first kind. In this note we apply GCV to a wider class of estimators than the one parameter ridge estimators. We observe that the expected values of the parameter mean-square error, the predictive mean-square error, and the GCV function are simultaneously minimized over this new class, so we accept the minimizer of the GCV function as the best computable estimator. We present a simple algorithm for computing this estimator from the data, so that a numerical search is not needed.

1. Introduction

In this note, we consider the problem of estimating the solution of the matrix equation

$$y = Ax \tag{1}$$

under the assumptions that (i) A is an $m \times n$ matrix with $m > n$, (ii) A is ill-conditioned, and (iii) the measurement of y is subject to error. We will assume that the data is

$$z = y + u, \tag{2}$$

where the error u is random, with mean zero and covariance Σ^2,

$$\langle u \rangle = 0, \qquad \langle uu^* \rangle = \Sigma^2, \tag{3}$$

https://doi.org/10.1017/S0334270000002836 Published online by Cambridge University Press

and where $\langle\ \rangle$ denotes the expected value, and * the complex conjugate transpose.

An example of an ill-conditioned system such as (1) is a discrete approximation to a linear integral equation of the first kind,

$$y(s) = \int_0^1 k(s, t)x(t)\, dt, \tag{4}$$

whose kernel defines a compact operator K on $L^2(0, 1)$. The singular values of K converge to zero and the inverse of K is unbounded, so the solution of (4) (if indeed one exists) may not depend continuously on the data and may contain large errors. These properties are reflected in the matrix A, whose condition will be poor and will worsen as its size is increased.

The method of regularization (introduced by Tihonov [7] and surveyed by de Hoog [4] and Lukas [5]) is one attempt to control the instability inherent in (1). Its strategy is to regard the unstable problem as a limit of a family of stable problems, depending upon a number of regularization parameters, and to solve instead one of the stable problems for a suitable choice of the parameters. For example, we could replace (1) by the family of problems,

$$\underset{x}{\text{minimize}}\ \|z - Ax\|^2 + \lambda\|x\|^2, \tag{5}$$

depending upon the parameter λ. However, the method faces an obvious difficulty, the choice of the parameter λ: if λ is too small, the problem remains numerically unstable, but if λ is too large the solution of the regularized problem may bear little relation to the true solution.

What is a general program for choosing the "best" regularization parameters? We allow the estimate $\hat{x}$ of x to be a non-linear function of the data, restricted by the following conditions.

(1) $\hat{x}$ should vanish identically whenever z vanishes, so the most general relationship between $\hat{x}$ and z is

$$\hat{x} = Bz, \tag{6}$$

in which B is an $n \times m$ matrix, which may itself be a function of z.

(2) The range of B should be contained within the orthogonal complement of the kernel of A. This is a natural condition, because the component of x in the kernel of A cannot contribute to y.

(3) B must be the "best" estimator, in the sense that B must minimize a stated objective function over a stated class of $n \times m$ matrices. For example, we might consider as possible objective functions

$$X = \|x - \hat{x}\|^2, \tag{7}$$

$$Y = \|y - \hat{y}\|^2, \quad \text{where } \hat{y} = A\hat{x}, \tag{8}$$

https://doi.org/10.1017/S0334270000002836 Published online by Cambridge University Press

$$Z = \|z - \hat{y}\|^2 \left[m^{-1} \operatorname{tr} \Sigma^2 / \operatorname{tr}(1 - AB)\Sigma^2 \right]^2, \tag{9}$$

or the expected values of these functions.

(4) B must be stable, in the sense that any matrix norm of B must be a bounded function of the singular values of A. This guarantees that small errors in the data cannot produce aribitrarily large errors in the estimate.

The contentious components in this program are the objective function and the class of matrices to which the estimators are restricted: different choices lead to conflicting definitions of "best" and a deluge of papers. (A good survey of the literature is given by Golub, Heath and Wahba [2].) The difficulty is that the "natural" objective functions, such as $\langle X \rangle$ and $\langle Y \rangle$, contain unknown quantities, such as x and y, and so are uncomputable. Only when $\hat{x}$ is required to be linear in z, unbiased, and the minimizer of the variance

$$(x - \hat{x})(x - \hat{x})^*$$

does this not pose a problem. In this case, the best estimator is the Gauss-Markov estimator,

$$B = (A^* \Sigma^{-2} A)^+ A^* \Sigma^{-2},$$

where $^+$ denotes the Moore–Penrose inverse, which is computable either if Σ^2 is known or if Σ^2 has the form

$$(\Sigma^2)_{ij} = \sigma^2 \delta_{ij} \tag{10}$$

with σ^2 possibly unknown. If we relax the requirements that the estimate should be unbiased and should depend linearly upon the data, then we must find a *computable* function Q with the property that the minimizer of $\langle Q \rangle$ is close to the minimizer of one of the natural loss functions, such as $\langle X \rangle$. Then we can assert that the *ideal estimator* is the minimizer $B^\square$ of $\langle X \rangle$, but the *best computable estimator* is the minimizer $B^\sharp$ of Q, for $B^\sharp$ should be close to $B^\square$ whenever Q is close to $\langle Q \rangle$.

When the covariance is known, it is not difficult to find a function Q with this property. For example, suppose that the covariance Σ^2 has the form (10) in which σ^2 is known, and consider the class of ridge estimators,

$$B_\lambda = (A^* A + \lambda)^{-1} A^*,$$

and corresponding ridge estimates,

$$\hat{x}_\lambda = B_\lambda z.$$

It is well known that $\hat{x}_\lambda$ is the solution of (5), so λ is a regularization parameter and its value represents a compromise between fitting the data and controlling the noise in the solution. If the ideal $\lambda^\square$ is taken to be the minimizer of $\langle Y \rangle$, then Craven and Wahba [1], referring to the work of Mallows [6], have suggested

https://doi.org/10.1017/S0334270000002836 Published online by Cambridge University Press

that a *good* computable $\lambda^{\#}$ is the minimizer of

$$Q = \|(1 - AB)z\|^2 - \sigma^2 \operatorname{tr}(1 - AB)(1 - AB)^* + \sigma^2 \operatorname{tr} ABB^*A^*,$$

for this function has the property that

$$\langle Q \rangle = \langle Y \rangle.$$

When the covariance is unknown, the problem is much more difficult. However, an attractive solution was proposed by Wahba [8] and further developed by Golub, Heath and Wahba [2] and Craven and Wahba [1] in papers on the solution of integral equations with noisy data, regression, and curve fitting. Their method, known as generalized cross-validation (GCV), allows a good value for the ridge parameter to be chosen from the data, when the coveriance has the form (10) but the value of σ^2 is unknown. They suggest that the ideal $\lambda^{\square}$ is the minimizer of $\langle Y \rangle$, and that the best computable $\lambda^{\#}$ should be taken to be the minimizer of the GCV function, which reduces to

$$Z = \|(1 - AB)z\|^2 / \left[\operatorname{tr}(1 - AB)\right]^2$$

under the present assumptions on Σ^2. In support of this suggestion, they offer strong intuitive arguments and then prove that $\lambda^{\square}$ will be close to the minimizer $\lambda^{\sim}$ of $\langle Z \rangle$ whenever AA^* becomes ill-conditioned for large m. For example, when A is a discrete approximation to a compact integral operator K, Wahba [8] has estimated the difference $\lambda^{\square} - \lambda^{\sim}$ in terms of the asymptotic rate of decay of the singular values of K. So they argue that $\lambda^{\#}$ should provide a good choice for λ, provided Z is close to $\langle Z \rangle$. The best computable $\lambda^{\#}$ can be found by a global parameter search.

In this note we consider a natural extension of the method of generalized cross-validation in which we retain the GCV objective function Z, but enlarge the class of allowed estimators. When restricted to the new class, which we denote by $\mathcal{F}$, each of the objective functions $\langle X \rangle$, $\langle Y \rangle$ and $\langle Z \rangle$ has only one extremum, a global minimum, and these are attained simultaneously at $B^{\square}$ in $\mathcal{F}$. Furthermore, $B^{\square}$ is stable. Thus, we assert that the best computable estimator is the minimizer $B^{\#}$ of Z over $\mathcal{F}$. We will present an algorithm which allows $B^{\#}$ to be computed trivially, without the need for a parameter search.

2. Class of estimators

Let us suppose that A has the singular value decomposition,

$$A = USV^*,$$

https://doi.org/10.1017/S0334270000002836 Published online by Cambridge University Press

in which:

(1) U is an $m \times m$ unitary matrix whose columns are orthonormal eigenvectors of AA^*;

(2) S is an $m \times n$ diagonal matrix

$$S = \left[\begin{array}{ccc} s_1 & & \\ & \ddots & \\ & & s_n \\ \hline & 0 & \end{array}\right]$$

in which $s_1, s_2, \ldots, s_n$ are the singular values of A, the positive square roots of the eigenvalues of A^*A;

(3) V is an $n \times n$ unitary matrix whose columns are orthonormal eigenvectors of A^*A.

We can assume that the singular values are ordered so that

$$s_1 \geqslant s_2 \geqslant \cdots \geqslant s_n \geqslant 0.$$

The rank of A, denoted by r, is the number of non-zero singular values, and the Moore–Penrose generalized inverse of A is given by

$$A^+ = VS^+U^*,$$

where

$$S^+ = \left[\begin{array}{ccc|c} s_1 & & & \\ & \ddots & & 0 \\ & & s_n^+ & \end{array}\right]$$

and

$$s_i^+ = \begin{cases} s_i^{-1} & \text{if } s_i > 0, \\ 0 & \text{if } s_i = 0. \end{cases}$$

We introduce the family $\mathcal{F}$ of $n \times m$ matrices with the form

$$B = VFS^+U^*, \tag{11}$$

where

$$0 \leqslant F \leqslant 1 \tag{12}$$

and

$$F = \left[\begin{array}{ccc|c} f_1 & & & \\ & \ddots & & 0 \\ & & f_r & \\ \hline & 0 & & 0 \end{array}\right], \qquad r = \operatorname{rank} A. \tag{13}$$

https://doi.org/10.1017/S0334270000002836 Published online by Cambridge University Press

We call F a filter, because it clearly filters the spectrum of singular values of A, and B the corresponding filtered estimator. Such matrices satisfy the following conditions:

(1) both AB and BA are hermitian;

(2) $0 \leqslant AB \leqslant 1$ and $0 \leqslant BA \leqslant 1$;

(3) the range of B is contained in the orthogonal complement of the kernel of A.

Conversely, any matrix B which satisfies conditions (1) to (3) must have the form shown in equation (11) with $0 \leqslant F \leqslant 1$. Furthermore, F will be block diagonal with the dimension of each block equal to the multiplicity of the corresponding singular value of A. In particular, if the singular values of A are distinct, then F must have the form (13). In order to prove these assertions, let

$$B = VTU^*$$

and partition S and T

$$S = \begin{pmatrix} S_{11} & 0 \\ 0 & 0 \end{pmatrix}, \qquad T = \begin{pmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{pmatrix},$$

so that S_{11} contains the non-zero singular values of A. The condition that the range of B should be in the orthogonal complement of the kernel of A forces $T_{21} = T_{22} = 0$. Next, the conditions that AB and BA should be hermitian show that $T_{12} = 0$ and

$$S_{11}T_{11} = T_{11}^* S_{11},$$
$$S_{11}T_{11}^* = T_{11}S_{11}.$$

Hence, S_{11} (anti) commutes with the (skew) hermitian part of T_{11}, so T_{11} must be hermitian and block diagonal with the dimensionality of each block equal to the multiplicity of the corresponding singular value in S_{11}. In particular, if the singular values are distinct, then T_{11} must be diagonal. Lastly, the conditions

$$0 \leqslant AB \leqslant 1 \quad \text{and} \quad 0 \leqslant BA \leqslant 1$$

show that T can be factorized, $T = FS^+$, with $0 \leqslant F \leqslant 1$.

The ridge estimators lie in this class. Indeed, it is not hard to show that the corresponding filter has

$$f_i = \frac{s_i^2}{s_i^2 + \lambda}, \qquad i = 1, 2, \ldots, r.$$

The elements $f_1, f_2, \ldots, f_r$ are the regularization parameters in our approach. When all f_i are equal to one, the matrix B coincides with A^+, whose norm will be large if any of the singular values is small. However, if the elements of F are matched to the singular values of A so that the product FS^+ remains bounded as a function of the singular values, then the estimator will be stable, but may

https://doi.org/10.1017/S0334270000002836 Published online by Cambridge University Press

not be accurate if the f_i are too small. How then must F be chosen to achieve a compromise between stability and fidelity?

3. Ideal and best computable filter

When restricted to filtered estimators, the functions $\langle X \rangle$, $\langle Y \rangle$, $\langle Z \rangle$ and Z reduce to the following expressions:

$$\begin{aligned}
\langle X \rangle &= \|(1 - BA)x\|^2 + \operatorname{tr} B\Sigma^2 B^* \\
&= \sum_{i=1}^{r} \left[(1 - f_i)^2 b_i^2 + f_i^2 \sigma_i^2 \right] / s_i^2 + \sum_{i=r+1}^{n} a_i^2, \\
\langle Y \rangle &= \|(1 - AB)y\|^2 + \operatorname{tr} AB\Sigma^2 B^* A^* \\
&= \sum_{i=1}^{r} \left[(1 - f_i)^2 b_i^2 + f_i^2 \sigma_i^2 \right], \\
\langle Z \rangle &= \frac{\left[\|(1 - AB)y\|^2 + \operatorname{tr}(1 - AB)\Sigma^2(1 - AB)^* \right] \left[m^{-1} \operatorname{tr} \Sigma^2 \right]^2}{\left[\operatorname{tr}(1 - AB)\Sigma^2 \right]^2} \\
&= \frac{\left[\sum_{i=1}^{r} (1 - f_i)^2 (b_i^2 + \sigma_i^2) + \sum_{i=r+1}^{m} \sigma_i^2 \right] \left[m^{-1} \sum_{i=1}^{m} \sigma_i^2 \right]^2}{\left[\sum_{i=1}^{r} (1 - f_i)\sigma_i^2 + \sum_{i=r+1}^{m} \sigma_i^2 \right]^2}, \\
Z &= \frac{\left[\sum_{i=1}^{r} (1 - f_i)^2 c_i^2 + \sum_{i=r+1}^{m} c_i^2 \right] \left[m^{-1} \sum_{i=1}^{m} \sigma_i^2 \right]^2}{\left[\sum_{i=1}^{r} (1 - f_i)\sigma_i^2 + \sum_{i=r+1}^{m} \sigma_i^2 \right]^2}.
\end{aligned}$$

Here we have introduced the notation

$$\begin{aligned}
a_i^2 &= |(V^* x)_i|^2, \\
b_i^2 &= |(U^* y)_i|^2, \\
c_i^2 &= |(U^* z)_i|^2, \\
\sigma_i^2 &= (U^* \Sigma^2 U)_{ii},
\end{aligned}$$

and below we will also use

$$e_i = 1 - f_i.$$

https://doi.org/10.1017/S0334270000002836 Published online by Cambridge University Press

Note that both Z and $\langle Z \rangle$ have the same structure and can be obtained from

$$W = \frac{\left[\sum_{i=1}^{r} e_i^2 m_i^2 + \mu\right]\left[m^{-1} \operatorname{tr} \Sigma^2\right]^2}{\left[\sum_{i=1}^{r} e_i \sigma_i^2 + \nu\right]^2}$$

with a suitable identification of the parameters.

The key to the choice of the best filter is the following observation.

LEMMA 1. *Each of the functions* $\langle X \rangle$ *and* $\langle Y \rangle$ *has only one local extremum at*

$$e_i = \sigma_i^2/(b_i^2 + \sigma_i^2), \qquad i = 1, 2, \ldots, r. \tag{14}$$

If at least one of $\nu\sigma_1^2, \ldots, \nu\sigma_r^2$ *is non-zero,* W *has only one local extremum at*

$$e_i = \frac{\mu}{\nu}\sigma_i^2/m_i^2, \qquad i = 1, 2, \ldots, r.$$

Otherwise, W *has a line of degenerate local extrema along*

$$e_i = k\sigma_i^2/m_i^2,$$

where k *is arbitrary. For each function, the local extremum is a global minimum.*

PROOF. The first order, necessary condition for extrema of $\langle X \rangle$ and $\langle Y \rangle$ is

$$e_i b_i^2 - (1 - e_i)\sigma_i^2 = 0, \tag{15}$$

which has only the solution (14). When (15) is satisfied, then

$$\langle X(e + \delta e) \rangle - \langle X(e) \rangle = \sum_{i=1}^{r} (\delta e_i)^2 (b_i^2 + \sigma_i^2)/s_i^2 > 0,$$

and

$$\langle Y(e + \delta e) \rangle - \langle Y(e) \rangle = \sum_{i=1}^{r} (\delta e_i)^2 (b_i^2 + \sigma_i^2) \geqslant 0,$$

for any δe, so the extrema are global minima.

The first order necessary condition for an entremum of W is

$$e_i m_i^2 \left[\sum_{j=1}^{r} e_j \sigma_j^2 + \nu\right] = \sigma_i^2 \left[\sum_{j=1}^{r} e_j^2 m_j^2 + \mu\right]. \tag{16}$$

Every solution of the equation has the form

$$e_i = k\sigma_i^2/m_i^2, \tag{17}$$

where k is a constant, independent of i. Let us fix k by substituting (17) into (16). We find

$$k\sigma_i^2 \left[\sum_{j=1}^{r} k\sigma_j^4/m_j^2 + \nu\right] = \sigma_i^2 \left[\sum_{j=1}^{r} k^2\sigma_j^4/m_j^2 + \mu\right],$$

https://doi.org/10.1017/S0334270000002836 Published online by Cambridge University Press

and

$$\sigma_i^2\left[k\nu - \mu\right] = 0, \qquad i = 1, 2, \ldots, r,$$

so k is arbitrary if $\nu\sigma_1^2, \ldots, \nu\sigma_r^2$ all vanish, but $k = \mu/\nu$ otherwise. In the first case W has only one local extremum, and in the second it has a line of degenerate local extrema. At the extrema,

$$\frac{W(e)}{\left[m^{-1}\ \mathrm{tr}\ \Sigma^2\right]^2} = k/\left[\sum_{i=1}^{r} e_i\sigma_i^2 + \nu\right].$$

We will now check that

$$\frac{W(e+\delta e) - W(e)}{\left[m^{-1}\ \mathrm{tr}\ \Sigma^2\right]^2} = \frac{\left[\sum_{i=1}^{r}(e_i+\delta e_i)^2 m_i^2 + \mu\right]}{\left[\sum_{i=1}^{r}(e_i+\delta e_i)\sigma_i^2 + \nu\right]^2} - \frac{\left[\sum_{i=1}^{r} e_i^2 m_i^2 + \mu\right]}{\left[\sum_{i=1}^{r} e_i\sigma_i^2 + \nu\right]^2} \geqslant 0$$

for any δe when e is given by (17). After bringing this fraction to a common, positive denominator, and repeatedly using equations (16) and (17), we find its numerator to be

$$\begin{aligned} N &= \left[\sum_{j=1}^{r}\frac{k\sigma_j^4}{m_j^2} + \nu\right]\left[\sum_{i=1}^{r}(\delta e_i)^2 m_i^2\right] - k\left[\sum_{i=1}^{r}\delta e_i\sigma_i^2\right]^2 \\ &= \nu\sum_{i=1}^{r}(\delta e_i)^2 m_i^2 + k\sum_{i>j}\left[(\delta e_j)\sigma_i^2\frac{m_j}{m_i} - (\delta e_i)\sigma_j^2\frac{m_i}{m_j}\right]^2 \\ &\geqslant 0, \end{aligned}$$

so e is the global minimiser of W.

Henceforth we will assume that the conditions given in Lemma 1 for degenerate minima are not satisfied. Thus, we will assume that at least one of

$$\sigma_i\sum_{j=r+1}^{m}\sigma_j^2, \qquad i = 1, 2, \ldots, r,$$

is non-zero, so that both Z and $\langle Z\rangle$ have only one local minimum. With this restriction, we see that the global minima of $\langle X\rangle$, $\langle Y\rangle$ and $\langle Z\rangle$ are attained simultaneously when

$$e_i = \sigma_i^2/\left(b_i^2 + \sigma_i^2\right).$$

This prompts us to assert that *the ideal filter $F^{\square}$ is the minimizer of $\langle X\rangle$, and the best computable filter $F^{\sharp}$ is the minimizer of Z, subject to the constraints*

$$0 \leqslant f_i^{\sharp} \leqslant 1, \qquad i = 1, 2, \ldots, r.$$

https://doi.org/10.1017/S0334270000002836 Published online by Cambridge University Press

We note that Z is computable either if Σ^2 is known or if Σ^2 has the form (10) with σ^2 unknown. In the latter case, Z reduces to Wahba's GCV function.

4. Stability and reduction in variance

The ideal filter can be written

$$f_i^{\square} = \frac{s_i^2}{s_i^2 + \sigma_i^2/a_i^2}, \qquad i = 1, 2, \ldots, r,$$

so

$$f_i^{\square} s_i^{+} = \frac{s_i}{s_i^2 + \sigma_i^2/a_i^2}, \qquad i = 1, 2, \ldots, r,$$

which is certainly a bounded function of s_i. Thus, the ideal estimator is stable.

$F^{\square}$ also achieves a reduction in $\langle X \rangle$ from its value with the Gauss–Markov estimator, obtained when the filter is the projection

$$F^* = \left[\begin{array}{c|c} \begin{matrix} 1 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{matrix} & \\ \hline & \begin{matrix} 0 & & \\ & \ddots & \\ & & 0 \end{matrix} \end{array}\right].$$

A short calculation gives

$$\langle X(F^*) \rangle - \langle X(F^{\square}) \rangle = \sum_{i=1}^{r} \frac{(\sigma^2/s_i^2)^2}{(\sigma^2/s_i^2) + a_i^2} > 0.$$

This difference will be large whenever any $s_i \ll \sigma_i$.

In practice we have the best computable filter, not the ideal, so these properties of stability and reduction in variance cannot be guaranteed, except on average.

5. Algorithm for the constrained minimization of Z

In the final section we focus attention on the special case in which

$$\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_m^2,$$

https://doi.org/10.1017/S0334270000002836 Published online by Cambridge University Press

and we present a simple algorithm for the minimization of

$$Z = \frac{\sum_{i=1}^{r} e_i^2 c_i^2 + \sum_{i=r+1}^{m} c_i^2}{\left[\sum_{i=1}^{r} e_i + m - r\right]^2},$$

subject to the constraints

$$0 \leqslant e_i \leqslant 1, \qquad i = 1, 2, \ldots, r.$$

We can suppose without loss of generality that the numbers c_i^2 have been ordered so that

$$c_1^2 \geqslant c_2^2 \geqslant \cdots \geqslant c_r^2 > 0.$$

Let

$$\alpha_k = \sum_{i=k+1}^{m} c_i^2, \qquad \beta_k = m - k, \qquad k = 0, 1, \ldots, r,$$

and set

$$Z_k = \left[\sum_{i=1}^{k} e_i^2 c_i^2 + \alpha_k\right] \Big/ \left[\sum_{i=1}^{k} e_i + \beta_k\right]^2.$$

Thus, $Z = Z_r$.

The algorithm runs as follows.

Start

Let $k = r$.

Loop

According to Lemma 1, the global minimizer of Z_k is

$$e_i c_i^2 = \alpha_k / \beta_k, \qquad i = 1, 2, \ldots, k. \tag{18}$$

Branch

If $c_i^2 \geqslant \alpha_k/\beta_k$, $i = 1, 2, \ldots, k$, then the constraints are satisfied and the problem is solved, so exit with the current value of k. Otherwise, we must have $c_i^2 < \alpha_k/\beta_k$ for some $i = 1, 2, \ldots, k$.

Continue

Let e^* denote the constrained minimizer of Z_k. Because Z_k does not have any other local minima than (18), at least one constraint must be active at e^*, so $e_i^* = 1$ for some $i = 1, 2, \ldots, k$. Furthermore, the sequences $\{e_i^*\}$

https://doi.org/10.1017/S0334270000002836 Published online by Cambridge University Press

and $\{c_i^2\}$ must be oppositely ordered, so $\{e_i^*\}$ must be increasing, $e_1^* \leqslant e_2^* \leqslant \cdots e_k^*$.

This follows from the rearrangement theorem of Hardy, Littlewood and Pólya [3], because the denominator of Z_k is unaffected by a rearrangement of $\{e_i^*\}$, but the numerator takes its minimum when the sequences $\{e_i^*\}$ and $\{c_i^2\}$ are oppositely ordered. Hence, $e_i^* = e_{i+1}^* = \cdots = e_k^* = 1$.

We now transfer e_k^* from the list of active variables to the list of constrained variables, simply by replacing k by $k - 1$ and returning to the start of the loop.

Exit

The algorithm returns the integer k.

LEMMA 2. *The minimizer of Z_r, subject to the constraints*

$$0 \leqslant e_i \leqslant 1, \qquad i = 1, 2, \ldots, r,$$

is

$$e_i^\# c_i^2 = \alpha_k/\beta_k, \qquad i = 1, 2, \ldots, k, \tag{19}$$

$$e_i^\# = 1, \qquad i = k + 1, \ldots, r,$$

where k is the integer returned by the algorithm.

PROOF. Let e^* denote the constrained minimizer of Z_r. If

$$0 \leqslant e_i^* < 1, \qquad i = 1, 2, \ldots, r,$$

then the constrained minimizer is the global minimizer and is given correctly by (19) with $k = r$. Otherwise, let l be the smallest integer such that

$$e_{l+1}^* = 1.$$

It then follows from the rearrangement theorem that

$$e_1^* \leqslant \cdots \leqslant e_l^* < e_{l+1}^* = \cdots = e_r^* = 1.$$

Hence

$$Z_r(e_1^*, \ldots, e_l^*, e_{l+1}^*, \ldots, e_r^*) = Z_l(e_1^*, \ldots, e_l^*).$$

Now the unconstrained minimizer of Z_l is

$$e_i^* c_i^2 = \alpha_l/\beta_l, \qquad i = 1, 2, \ldots, l,$$

so e^* certainly has the form given in (19). But the algorithm constructs the largest integer k such that the minimizer of Z_k is unconstrained, so we must have

https://doi.org/10.1017/S0334270000002836 Published online by Cambridge University Press

$l \leqslant k$. If $l < k$, then

$$Z_r(e^\#) = Z_k(e_1^\#, \ldots, e_k^\#) < Z_k(e_1^*, \ldots, e_k^*) = Z_l(e_1^*, \ldots, e_l^*) = Z_r(e^*),$$

which contradicts the definition of e^* as the constrained minimizer of Z_r. Hence, $k = l$ and $e^* = e^\#$.

6. Conclusion

We end this note with a warning to anyone faced with the problem of inverting an ill-conditioned linear system of equations when the data is subject to error. Even if both the covariance of the noise and the solution of the linear system were known (!), so that the ideal estimator could be computed, the ratio $|(V^*\delta x)_i/(U^*\delta y)_i|$ could still be as large as $\frac{1}{2}a_i/\sigma_i$, which might not be insignificant. Only the average of the estimates obtained from a large number of data sets will give the correct result. Consequently, it is essential to obtain as much raw data as possible from the experimenter, to invert each data set separately, and finally to average the results. It is folly to accept a single set of averaged data from the experimenter, together with his estimate of the mean and covariance of the noise. The averaging must follow the inversion.

References

[1] P. Craven and G. Wahba, "Smoothing noisy data with spline functions. Estimating the correct degree of smoothing by the method of generalized cross-validation", *Numer. Math.* 31 (1979), 377–403.

[2] G. H. Golub, M. Heath and G. Wahba, "Generalized cross-validation as a method for choosing a good ridge parameter", *Technometrics* 21 (1979), 215–223.

[3] G. H. Hardy, J. E. Littlewood and G. Pólya, *Inequalities* (Cambridge University Press, 2nd edition, 1952), Chapter X.

[4] F. R. de Hoog, "Review of Fredholm equations of the first kind", in *The application and numerical solution of integral equations* (eds. R. S. Anderssen, F. R. de Hoog, and M. A. Lukas) (Sijthoff and Noordhoff, The Netherlands, 1980).

[5] M. A. Lukas, "Regularization", in *The application and numerical solution of integral equations* (eds. R. S. Anderssen, F. R. de Hoog, and M. A. Lukas, Sijthoff and Noordhoff, The Netherlands, 1980).

[6] C. L. Mallows, "Some comments on C_p", *Technometrics* 15 (1973), 661–675.

[7] A. N. Tihonov, "Solution of incorrectly formulated problems and the method of regularization", *Soviet Math. Dokl.* 4 (1963), 1035–1038.

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

Department of Mathematical Physics
University of Adelaide
Adelaide
South Australia 5000

and

Department of Mathematics
University of Queensland
St. Lucia
Queensland 4067