Aplikace matematiky

Gene Howard Golub

Numerical methods for solving linear least squares problems

Aplikace matematiky, Vol. 10 (1965), No. 3, 213–216

Persistent URL: http://dml.cz/dmlcz/102951

NUMERICAL METHODS FOR SOLVING LINEAR LEAST SQUARES PROBLEMS[1])

GENE H. GOLUB

(to topic d)

One of the problems which arises most frequently in a Computer Laboratory is that of finding linear least squares solutions. These problems arise in a variety of contexts, e.g., statistical applications, numerical solution of integral equations of the first kind, etc. Linear least squares problems are particularly difficult to solve because they frequently involve large quantities of data, and they are ill-conditioned by their nature.

Let A be a given $m \times n$ real matrix with $m \geqq n$ and of rank r, and $\boldsymbol{b}$ a given vector. We wish to determine a vector $\hat{\boldsymbol{x}}$ such that

$$\|\boldsymbol{b} - A\hat{\boldsymbol{x}}\| = \min . \tag{1}$$

where $\| \ldots \|$ indicates the euclidean norm. It is well known (cf. [4]) that $\hat{\boldsymbol{x}}$ satisfies the equation

$$A^T A\boldsymbol{x} = A^T\boldsymbol{b} . \tag{2}$$

Unfortunately, the matrix $A^T A$ is frequently ill-conditioned [4] and influenced greatly by roundoff errors. The following example of LÄUCHLI [2] illustrates this well. Suppose

$$A = \begin{bmatrix} 1, & 1, & 1, & 1, & 1 \\ \varepsilon, & 0, & 0, & 0, & 0 \\ 0, & \varepsilon, & 0, & 0, & 0 \\ 0, & 0, & \varepsilon, & 0, & 0 \\ 0, & 0, & 0, & \varepsilon, & 0 \\ 0, & 0, & 0, & 0, & \varepsilon \end{bmatrix},$$

[1]) This report was supported in part by Office of Naval Research Contract Nonr — 225(37) (NR 044-11) at Stanford University.

then

$$A^TA = \begin{bmatrix} 1+\varepsilon^2, & 1, & 1, & 1, & 1, \\ 1, & 1+\varepsilon^2, & 1, & 1, & 1, \\ 1, & 1, & 1+\varepsilon^2, & 1, & 1, \\ 1, & 1, & 1, & 1+\varepsilon^2, & 1, \\ 1, & 1, & 1, & 1, & 1+\varepsilon^2 \end{bmatrix}. \tag{3}$$

Clearly for $\varepsilon \neq 0$, the rank of A^TA is five since the eigenvalues of A^TA are $5 + \varepsilon^2$, $\varepsilon^2, \varepsilon^2, \varepsilon^2, \varepsilon^2$.

Let us assume that the elements of A^TA are computed using double precision arithmetic, and then rounded to single precision accuracy. Now let η be the largest number on the computer such that $\mathrm{fl}(1.0 + \eta) \equiv 1.0$ where $\mathrm{fl}(\ldots)$ indicates the floating point computation. Then if $\varepsilon < \sqrt{\eta}/2$, the rank of the computed representation of (3) will be one. Consequently, no matter how accurate the linear equation solver, it is impossible to solve the normal equations (2).

In [1], HOUSEHOLDER stressed the use of orthogonal transformations for solving linear least squares problems. In this paper, we shall exploit these transformations.

Since the euclidean norm of a vector is unitarily invariant,

$$\|\boldsymbol{b} - A\boldsymbol{x}\| = \|\boldsymbol{c} - QA\boldsymbol{x}\|$$

where $\boldsymbol{c} = Q\boldsymbol{b}$ and Q is an orthogonal matrix. We choose Q so that

$$QA = R = \begin{pmatrix} \tilde{R} \\ \hline O \end{pmatrix}_{\}(m-n)\times n} \tag{4}$$

where $\tilde{R}$ is an upper triangular matrix.

Clearly,

$$\hat{\boldsymbol{x}} = \tilde{R}^{-1}\tilde{\boldsymbol{c}}$$

where $\tilde{\boldsymbol{c}}$ is the first n components of $\boldsymbol{c}$ and consequently,

$$\|\boldsymbol{b} - A\hat{\boldsymbol{x}}\| = \Big(\sum_{j=m+1}^{n} c_j^2\Big)^{\frac{1}{2}}.$$

Since $\tilde{R}$ is an upper triangular matrix and $\tilde{R}^T\tilde{R} = A^TA$, $\tilde{R}^T\tilde{R}$ is simply the Choleski decomposition of A^TA.

There are a number of ways to achieve the decomposition (4); e.g., one could apply a sequence of plane rotations to annihilate the elements below the diagonal of A. A very effective method to realize the decomposition (4) is via Householder transformations [1]. Let $A = A^{(1)}$, and let $A^{(2)}, A^{(3)}, \ldots, A^{(n+1)}$ be defined as follows:

$$A^{(k+1)} = P^{(k)}A^{(k)} \quad (k = 1, 2, \ldots, n).$$

$P^{(k)}$ is a symmetric, orthogonal matrix of the form

$$P^{(k)} = I - 2\mathbf{w}^{(k)}\mathbf{w}^{(k)T}$$

for suitable $\mathbf{w}^{(k)}$ such that $w^{(k)T}\mathbf{w}^{(k)} = 1$. A derivation of $P^{(k)}$ is given in [5]. In order to simplify the calculations, we redefine $P^{(k)}$ as follows:

$$P^{(k)} = I - \beta_k \mathbf{u}^{(k)}\mathbf{u}^{(k)T}$$

where $\sigma_k = (\sum_{i=k}^{m} (a_{i,k}^{(k)})^2)^{\frac{1}{2}}$, $\beta_k = [\sigma_k(\sigma_k + |a_{k,k}^{(k)}|)]^{-1}$, and

$$u_i^{(k)} = 0 \text{ for } i < k\,, \quad u_k^{(k)} = \operatorname{sgn}(a_{k,k}^{(k)})(\sigma_k + |a_{k,k}^{(k)}|)\,, \quad u_i^{(k)} = a_{i,k}^{(k)} \text{ for } i > k\,.$$

Thus

$$A^{(k+1)} = A^{(k)} - \mathbf{u}^{(k)}(\beta_k \mathbf{u}^{(k)T} A^{(k)})\,.$$

After $P^{(k)}$ has been applied to $A^{(k)}$, $A^{(k+1)}$ appears as follows:

$$A^{(k+1)} = \left|\begin{array}{c|c} \tilde{R}^{(k+1)} & \\ \hline & \end{array}\right|$$

where $\tilde{R}^{(k+1)}$ is a $k \times k$ upper triangular matrix which is unchanged by subsequent transformations. Now $a_{k,k}^{(k+1)} = -(\operatorname{sgn} a_{k,k}^{(k)})\,\sigma_k$ so that the rank of A is less than n if $\sigma_k = 0$. Clearly,

$$R = A^{(n+1)}$$

and

$$Q = P^{(n)}P^{(n-1)} \ldots P^{(1)}$$

although one need not compute Q explicitly.

Let $\bar{\mathbf{x}}$ be the intial solution obtained, and let $\hat{\mathbf{x}} = \bar{\mathbf{x}} + \mathbf{e}$. Then

$$\|\mathbf{b} - A\hat{\mathbf{x}}\| = \|\mathbf{r} - A\mathbf{e}\|$$

where

$$\mathbf{r} = \mathbf{b} - A\bar{\mathbf{x}}\,, \text{ the residual vector.}$$

Thus the correction vector $\mathbf{e}$ is itself the solution to a linear least squares problem. Once A has been decomposed then it is a fairly simple matter to compute $\mathbf{r}$ and solve for $\mathbf{e}$. Since $\mathbf{e}$ critically depends upon the residual vector, the components of $\mathbf{r}$ should be computed using double precision inner products and then rounded to single precision accuracy. Naturally, one should continue to iterate as long as improved estimates of $\hat{\mathbf{x}}$ are obtained.

The above iteration technique will converge only if the initial approximation to $\boldsymbol{x}$ is sufficiently accurate. Let

$$\boldsymbol{x}^{(q+1)} = \boldsymbol{x}^{(q)} + \mathbf{e}^{(q)} \quad (q = 0, 1, \ldots) \quad \text{with} \quad \boldsymbol{x}^{(0)} = \boldsymbol{0}.$$

Then if $\|\mathbf{e}^{(1)}\|/\|\boldsymbol{x}^{(1)}\| > c$ and if $c < \frac{1}{2}$, i.e., "at least one bit of the initial solution is correct", one should not iterate since there is little likelihood that the iterative method will converge. Since convergence tends to be linear, one should terminate the procedure as soon as $\|\mathbf{e}^{(k+1)}\| > c\|\mathbf{e}^{(k)}\|$.

References

[1] *A. S. Householder:* Unitary Triangularization of a Nonsymmetric Matrix, J. Assoc. Comput. Mach., Vol. 5 (1958), pp. 339—342.

[2] *P. Läuchli:* Jordan-Elimination und Ausgleichung nach kleinsten Quadraten, Numer. Math., Vol. 3 (1961), pp. 226—240.

[3] *Y. Linnik:* Method of Least Squares and Principles of the Theory of Observations, translated from Russian by R. C. Elandt, Pergamon Press, New York, 1961.

[4] *E. E. Osborne:* On Least Squares Solutions of Linear Equations, J. Assoc. Comput. Mach., Vol. 8 (1961), pp. 628—636.

[5] *J. H. Wilkinson:* Householder's Method for the Solution of the Algebraic Eigenproblem, Comput. J., Vol. 3 (1960), pp. 23—27.

Gene H. Golub, Computation Center, Stanford University, Stanford, California, U.S.A.