Czechoslovak Mathematical Journal, 25 (100) 1975, Praha

NOTE ON A SPLITTING APPROACH TO ILL-CONDITIONED LEAST SQUARES PROBLEMS

ROBERT J. PLEMMONS, Knoxville*)

(Received June 6, 1973)

1. INTRODUCTION

For an $m \times n$ real matrix A of rank n and a real m-vector b, the least squares problem for the linear system $Ax = b$ is to determine the n-vector $\tilde{x}$ such that

$$\|b - A\tilde{x}\|_2 \leqq \|b - Ax\|_2$$

for all n-vectors x. The unique solution is given by

$$\tilde{x} = (A^T A)^{-1} A^T b$$

where "T" denotes the transpose. The traditional approach to the problem is to solve the "normal system"

$$A^T A x = A^T b \tag{1.1}$$

by some standard direct procedure such as Gaussian or Cholesky elimination. However, using the computer this approach is often poor, since the condition number of $A^T A$ is the square of the condition number $\varkappa(A)$, of A. (Here we take the condition number to be the ratio of the largest to the smallest singular value of the matrix.) In fact, using t-digit binary arithmetic one is not able to obtain even an approximate solution to (1.1) unless $\varkappa(A) \leqq 2^{t/2}$.

Several authors have suggested alternatives based on the orthogonal decomposition of A into

$$A = Q^T \begin{bmatrix} R \\ 0 \end{bmatrix}$$

where Q is orthogonal and R is $n \times n$ upper triangular. Writing

$$Q = \begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix} \tag{1.2}$$

*) The research was supported in part by the National Science Foundation Grant 15943.

where Q_1 is $n \times m$, the least squares solution is given by

$$\tilde{x} = R^{-1} Q_1 b . \tag{1.3}$$

One approach of this kind was given by GOLUB [5] and BUSINGER and GOLUB [3] using Householder transformations on A and another employs the Gram-Schmidt algorithm and its modification as suggested by BJÖRCK [2] and by PETERS and WILKINSON [9]. Iterative refinements of these methods can be found in [2] and [6]. Further extensions of the Householder method were given by HANSON and LAWSON [7]. More recently an elimination approach has been suggested by CLINE [4].

In [10] an iterative procedure was suggested for approximating $\tilde{x}$, based on the splitting of the coefficient matrix A into

$$A = M - N , \tag{1.4}$$

where M and A have the same range. This method was developed further and convergence criteria similar to those for the nonsingular case were given in [1]. In this note the use of (1.4) is suggested as a possible direct approach to avoiding the ill-conditioned properties of the normal system (1.1).

2. A SPLITTING APPROACH

First notice that in the splitting (1.4), $M(M^TM)^{-1} M^TN = N$ since the range of N is contained in the range of M and since $M(M^TM)^{-1} M^T$ is the orthogonal projector on the range of M. Also, 1 is not an eigenvalue of $(M^TM)^{-1} M^TN$ since the null space of A is zero. Thus $I - (M^TM)^{-1} M^TN$ is nonsingular. In particular

$$A = M[I - (M^TM)^{-1} M^TN]$$

so that the least squares solution of $Ax = b$ is given by

$$\tilde{x} = [I - (M^TM)^{-1} M^TN]^{-1} (M^TM)^{-1} M^Tb , \tag{2.1}$$

since the right hand side of (2.1) reduces to $(A^TA)^{-1} A^Tb$.

One such choice of M is Q_1^T, where the matrix Q_1 is given in (1.2), for then $Q_1^T = AR^{-1}$ and $R = Q_1A$ since $Q_1Q_1^T = I$, and so A and Q_1^T have the same ranges. In this case $N = Q_1^T - A$ and (2.1) becomes

$$\tilde{x} = [I - (Q_1Q_1^T)^{-1} Q_1(Q_1^T - A)]^{-1} (Q_1Q_1^T)^{-1} Q_1b =$$
$$= [I - I + Q_1A]^{-1} Q_1b = R^{-1}Q_1b .$$

Thus the familiar form given for $\tilde{x}$ by (1.3) is a special case of (2.1).

Quite obviously the value of this splitting approach for a particular ill-conditioned least squares problem depends upon one's being able to choose the matrix M in such a way that its condition number is somewhat less than the condition number of A and with the property that the matrix $I - (M^T M)^{-1} M^T N$ is not difficult to invert. Theoretically the best choice of M is given by $M = Q_1^T$, as mentioned above. However in some cases M can be chosen by observation and by taking into account the structure of A. An example illustrating this situation is given in the next section.

3. EXAMPLE

Let ε be some small positive real number and consider the $m \times n$ matrix

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

originally due to LÄUCHLI [8], with the resulting system $Ax = b$. The matrix $A^T A$ then has the form

$$A^T A = \begin{bmatrix} 1+\varepsilon^2 & 1 & \dots & 1 \\ 1 & 1+\varepsilon^2 & \dots & 1 \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ 1 & 1 & \dots & 1+\varepsilon^2 \end{bmatrix}.$$

As was pointed out in [5] and in [11, p. 135], if a solution of the normal system (1.1) is attempted with $\varepsilon^2 < \beta_0$, the machine precision, then $A^T A$ becomes the singular matrix of all 1's and the resulting least squares problem is rendered unsolvable. In particular, when ε is small and higher powers of ε are ignored the condition number of A is approximated by

$$\varkappa(A) = \frac{\sqrt{n}}{\varepsilon}.$$

Here one would like to choose M with the same range as A such that $\varkappa(M)$ does not involve ε in the denominator. Let C^k denote the k^{th} column of an arbitrary matrix C. Then one can construct M with M^{j+1} independent of ε and having two nonzero entries by setting

$$M^{j+1} = \frac{1}{\varepsilon}(A^{j+1} - A^j)$$

for $j = 1, 2, \ldots, n-1$. Then taking $M^1 = A^1$,

$$M = \begin{bmatrix} 1 & 0 & 0 & \ldots & 0 \\ \varepsilon & -1 & 0 & \ldots & 0 \\ 0 & 1 & -1 & \ldots & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot\cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot\cdot & -1 \\ 0 & 0 & 0 & \ldots & 1 \end{bmatrix}$$

with $\varkappa(M) \approx \sqrt{n}$ and with A and M having the same range. In this case the matrix M is much better conditioned than A for small ε and $M^T M$ is the tri-diagonal matrix

$$M^T M = \begin{bmatrix} 1+\varepsilon^2 & -\varepsilon & 0 & \ldots & 0 & 0 \\ -\varepsilon & 2 & -1 & \ldots & 0 & 0 \\ 0 & -1 & 2 & \ldots & 0 & 0 \\ \cdot & \cdot & \cdot & \cdots & & \\ \cdot & \cdot & \cdot & \cdots & & \\ 0 & 0 & 0 & \ldots & & -1 \\ 0 & 0 & 0 & \ldots & -1 & 2 \end{bmatrix}. \tag{3.1}$$

Moreover, $I - (M^T M)^{-1} M^T N$ is the upper triangular matrix

$$(I - M^T M)^{-1} M^T N = \begin{bmatrix} 1 & 1 & 1 & \ldots & 1 \\ 0 & \varepsilon & \varepsilon & \ldots & \varepsilon \\ 0 & 0 & \varepsilon & \ldots & \varepsilon \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \ldots & \varepsilon \end{bmatrix} \tag{3.2}$$

Inversion of (3.1) and (3.2) and then post-multiplication by $(M^T M)^{-1} M^T b$ yields the least squares solution $\tilde{x}$ to $Ax = b$. In practice, one determines $\tilde{x}$ by elimination and back-substitution so that no matrix inversions are necessary here.

4. REMARKS

(a) It should be mentioned that the example in Section 3 is given only to show that it is sometimes possible to avoid the ill-conditioned problem associated with the normal system $A^T A x = A^T b$ by a judicious choice of M in the splitting (1.4). In general the matrix M may not be so easy to obtain.

(b) Whenever A and M are $m \times n$ with rank n and have the same ranges, it follows that

$$(M^T A)^{-1} M^T = (A^T A)^{-1} A^T .$$

Thus $\tilde{x}$ can be found by computing

$$(M^T A)^{-1} M^T b \, .$$

That this may not always be the best approach is well illustrated by the example in Section 3. Here $\varkappa(M^T A) \approx (A^T A)$ for this particular choice of M.

(c) In the case where the spectral radius of $(M^T M)^{-1} M^T N$ is less than one, the iteration

$$x^{(k+1)} = (M^T M)^{-1} M^T N x^{(k)} + (M^T M)^{-1} M^T b$$

converges to $\tilde{x}$ for each $x^{(0)}$ [10]. In the example above the iteration converges if and only if $0 < \varepsilon < 2$. This iterative method may be useful in solving large, sparse least squares problems.

(d) If the linear system $Ax = b$ is under-determined and A has full row rank then one can compute the solution $\tilde{y}$ of minimum norm in the following manner. The matrix A is split into $A = M - N$ where A and M have the same null spaces. Then $\tilde{y}$ is given by

$$\tilde{y} = M^T(MM^T)^{-1} [I - NM^T(MM^T)^{-1}]^{-1} b \, .$$

References

[1] *A. Berman* and *R. J. Plemmons,* "Cones and iterative methods for best least squares solutions to linear systems", SIAM J. on Numer. Anal. *11* (1974), 145—154.

[2] *A. Björck,* "Solving linear least squares problems by Gram-Schmidt Orthogonalization", BIT, *7* (1967), 1—21.

[3] *P. Businger* and *G. Golub,* "Linear least squares solutions by Householder transformations", Numer. Math., *7* (1965), 269—276.

[4] *A. K. Cline,* "An elimination method for the solution of linear least squares problems", SIAM J. Numer. Anal., *10* (1973), 283—289.

[5] *G. Golub,* "Numerical methods for solving least squares problems", Numer. Math., *7* (1965), 206—216.

[6] *G. Golub* and *J. H. Wilkinson,* "Note on the iterative refinement of least squares solutions", Numer. Math., *9* (1966), 139—148.

[7] *R. J. Hanson* and *C. L. Lawson,* "Extensions and applications of the Householder algorithm for solving linear least squares problems", Math. Comp., *23* (1969), 787—812.

[8] *P. Läuchli,* "Jordon-Elimination und Ausgleichung nach kleinsten Quadraten", Numer. Math., *3* (1961), 226—240.

[9] *G. Peters* and *J. H. Wilkinson,* "The least squares problem and pseudoinverses", Computing J., *13* (1970), 309—316.

[10] *R. J. Plemmons,* "Monotonicity and iterative approximations involving rectangular matrices", Math. Comp., *26* (1972), 853—858.

[11] *J. H. Wilkinson* and *C. Reinsch,* Linear Algebra, Springer-Verlag, New York, 1971.

Author's address: Department of Mathematics and Computer Sciences, University of Tennessee, Knoxville, Tennessee 37916, U.S.A.