LINEAR LEAST SQUARES
AND
QUADRATIC PROGRAMMING

BY

GENE H. GOLUB
MICHAEL A. SAUNDERS

TECHNICAL REPORT NO. CS 134
MAY 1969

COMPUTER SCIENCE DEPARTMENT
School of Humanities and Sciences
STANFORD UNIVERSITY

Linear Least Squares and Quadratic Programming *

Gene H. Golub

Michael A. Saunders

*Invited paper to be presented at the NATO Advanced Study Institute on Integer and Nonlinear Programming, Ile de Bendor, France, June 8-22, 1969. This research was supported, in part, by the National Science Foundation and the Atomic Energy Commission project at Stanford University.

Abstract

Several algorithms are presented for solving linear least squares problems; the basic tool is orthogonalization techniques. A highly accurate algorithm is presented for solving least squares problems with linear inequality constraints. A method is also given for finding the least squares solution when there is a quadratic constraint on the solution.

0. Introduction

One of the most common problems in any computation center is that of finding linear least squares solutions. These problems arise in a variety of areas and in a variety of contexts. For instance, the data may be arriving sequentially from a source and there may be some constraint on the solution. Linear least squares problems are particularly difficult to solve because they frequently involve large quantities of data, and they are ill-conditioned by their very nature.

In this paper, we shall present several numerical algorithms for solving linear least squares problems in a highly accurate manner. In addition, we shall give an algorithm for solving linear least squares problem with linear inequality constraints.

1. Linear least sauares

Let A be a given mxn real matrix of rank r and b a given vector. We wish to determine $\hat{x}$ such that

$$\sum_{i=1}^{m} (b_i - \sum_{i=1}^{n} a_{ij}x_j)^2 = \min.$$

or using matrix notation

$$\|\underset{\sim}{b}-A\underset{\sim}{x}\|_2 = \min. \tag{1.1}$$

If m>n and r<n, then there is no unique solution. Under these conditions, we require amongst those vectors $\underset{\sim}{x}$ which satisfy (1.1) that

$$\|\hat{\underset{\sim}{x}}\|_2 = \min.$$

For r = n , $\hat{x}$ satisfies the normal equations

$$A^TA\hat{\underset{\sim}{x}} = A^Tb \ . \tag{1.2}$$

Unfortunately, the matrix A^TA is frequently ill-conditioned and influenced greatly by roundoff errors. The following example illustrates this well. Suppose

which is clearly of rank 4 . Then

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

and the eigenvalues of A^TA are $4+\varepsilon^2$, ε^2 , ε^2 , ε^2 . 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 $fl(1+\eta) = 1$ where $fl(\ldots)$ indicates the floating point computation. Then if $\varepsilon < \sqrt{\eta}$,

$$fl(A^TA) = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} ,$$

a matrix of rank one, and consequently, no matter how accurate the linear equation solver it will be impossible to solve the normal equations (1.2).

LONGLEY[1967] has given examples in which the solution of the normal equations leads to almost no digits of accuracy of the least squares problem.

2. A matrix decomposition

Now $\|\underset{\sim}{y}\|_2 = (\underset{\sim}{y}^T\underset{\sim}{y})^{1/2}$ so that $\|Q\underset{\sim}{y}\|_2 = \|\underset{\sim}{y}\|_2$ when Q is an orthogonal matrix, viz., $Q^TQ = I$. Thus

$$\|\underset{\sim}{b}-A\underset{\sim}{x}\|_2 = \|\underset{\sim}{c}-QA\underset{\sim}{x}\|_2$$

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

$$QA = R = \begin{bmatrix} \tilde{R} \\ \cdots \\ 0 \end{bmatrix} \begin{matrix} \\ \\ \}(m-n)\times n \end{matrix} \qquad (2.1)$$

where $\tilde{R}$ is an upper triangular matrix (∇). Let

$$\tilde{R} = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ & r_{22} & \cdots & r_{2n} \\ & & \ddots & \\ & & & r_{nn} \\ & & & \\ & & & \end{bmatrix}$$

then

$$\begin{aligned} \|\underset{\sim}{b}-A\underset{\sim}{x}\|_2^2 = {} & (c_1-r_{11}x_1-r_{12}x_2 - \cdots - r_{1n}x_n)^2 \\ & + (c_2-r_{22}x_2 - \cdots - r_{2n}x_n)^2 \\ & + \cdots + (c_n-r_{nn}x_n)^2 \\ & + c_{n+1}^2 + c_{n+2}^2 + \cdots + c_m^2 \ . \end{aligned}$$

Thus $\|\underset{\sim}{b}-A\underset{\sim}{x}\|_2^2$ is minimized when

$$\begin{aligned} r_{11}\hat{x}_1 + r_{12}\hat{x}_2 + \cdots + r_{1n}\hat{x}_n &= c_1 \\ r_{22}\hat{x}_2 + \cdots + r_{2n}\hat{x}_n &= c_2 \\ &\vdots \\ r_{nn}\hat{x}_n &= c_n \end{aligned}$$

i.e., $R\hat{x} = \tilde{c}$, where

$$\tilde{c}^T = (c_1, c_2, \ldots, c_n) \ ,$$

and

$$\|\underset{\sim}{b}-A\underset{\sim}{\hat{x}}\|_2^2 = c_{n+1}^2 + c_{n+2}^2 + \cdots + c_m^2 \ . \qquad (2.2)$$

Then

$$\begin{aligned} R^TR &= [\tilde{R}:0]^T[\tilde{R}:0] = \tilde{R}^T\tilde{R} \\ &= [QA]^T[QA] = A^TA \ , \end{aligned} \qquad (2.3)$$

and thus $\tilde{R}^T\tilde{R}$ is simply the Cholesky decomposition of A^TA .

There are a number of ways to achieve the decomposition of (2.1); 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 (2.1) is via Householder transformations. A matrix P is said to be a Householder transformation if

$$P = I - 2uu^T , \quad u^T u = 1 .$$

Note that 1) $P = P^T$ and 2) $PP^T = I - 2uu^T - 2uu^T + 4uu^T uu^T = I$ so that P is a symmetric, orthogonal transformation.

Let $A^{(1)} = A$ and let $A^{(2)}, A^{(3)}, \ldots, A^{(n+1)}$ be defined as follows:

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

where $P^{(k)} = I - 2w^{(k)} w^{(k)T}$, $w^{(k)T} w^{(k)} = 1$. The matrix $P^{(k)}$ is chosen so that $a_{k+1,k}^{(k+1)} = a_{k+2,k}^{(k+1)} = \ldots = a_{n,k}^{(k+1)} = 0$. Thus after k transformations

$$A^{(k+1)} = \begin{bmatrix} a_{11}^{(2)} & a_{12}^{(2)} & \cdot & \cdot & \cdot & a_{1n}^{(2)} \\ 0 & a_{22}^{(3)} & \cdot & \cdot & \cdot & a_{2n}^{(3)} \\ 0 & & \cdot & \cdot & \cdot & \\ \cdot & & & a_{kk}^{(k+1)} & \cdot & a_{kn}^{(k+1)} \\ \cdot & & & 0 & a_{k+1,k+1}^{(k+1)} & \cdot \\ \cdot & & & 0 & \cdot & \cdot \\ \cdot & & & \cdot & \cdot & \cdot \\ 0 & & & 0 & a_{m,k+1}^{(k+1)} & a_{mn}^{(k+1)} \end{bmatrix}$$

Note that $|a_{kk}^{(k+1)}| = (\sum_{i=k}^{m} (a_{ik}^{(k)})^2)^{1/2}$ since $P^{(k)}$ is an orthogonal transformation. The details of the computation are given in BUSINGER and GOLUB [1965] and GOLUB [1965]. The Householder transformations have been used in a highly effective manner by KALFON et al. [1968] in the implementation of the projection gradient method.

Clearly

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

and

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

although one need not compute Q explicitly. The number of multiplications required to produce R is roughly $mn^2-(n^3/3)$ whereas approximately $mn^2/2$ multiplications are required to form the normal equations (1.2).

3. The practical procedure

It is known that the Cholesky method for solving systems of equations is numerically stable even if no interchanges of rows and columns are performed. Since we are in effect performing a Cholesky decomposition of A^TA no interchanges of the columns of A are needed in most situations. However, numerical experiments have indicated that the accuracy is slightly improved by the interchange strategies outlined below, and consequently, in order to ensure the utmost accuracy one should choose the columns of A by some strategy. In what follows, we shall refer to the matrix $A^{(k)}$ even if some of the columns have been interchanged.

One possibility is to choose at the k^{th} stage the columns of $A^{(k)}$ which will maximize $|a_{kk}^{(k+1)}|$. This is equivalent to searching for the maximum diagonal element in the Cholesky decomposition of A^TA. Let

$$s_j^{(k)} = \sum_{j=k}^{m} (a_{i,j}^{(k)})^2 \qquad \text{for} \qquad j = k,k+1,\ldots,n .$$

Then since $|a_{kk}^{(k+1)}| = (s_k^{(k)})^{1/2}$, one should choose that column for which $s_j^{(k)}$ is maximized. After $A^{(k+1)}$ has been computed, one can compute $s_j^{(k+1)}$ as follows:

$$s_j^{(k+1)} = s_j^{(k)} - (a_{k,j}^{(k+1)})^2 \qquad (j = k+1,\ldots,n)$$

since the orthogonal transformations leave the column lengths invariant. Naturally, the $s_j^{(k)}$'s must be interchanged if the columns of $A^{(k)}$ are interchanged.

The above strategy is useful in determining the rank of a matrix. If the rank of A is r and the arithmetic is performed exactly, then after r transformations

$$A^{(r+1)} = \left[\begin{array}{c|c} \tilde{R}_{rxr} & S_{(n-r)xr} \\ \hline 0 & N \end{array}\right] ,$$

and

$$s_j^{(r+1)} = 0 \qquad \text{for} \qquad j = r+1,\ldots,n$$

which implies $N = 0$. In most situations, however, where rounded arithmetic is used $\|N\| = \varepsilon$. It is not easy to determine bounds on ε when the rank of A is unknown.

The strategy described above is most appropriate when one has a sequence of vectors $\underset{\sim}{b}_1, \underset{\sim}{b}_2, \ldots, \underset{\sim}{b}_p$ for which one desires a least squares estimate. In many problems, there is but one vector $\underset{\sim}{b}$ and one wishes to express it in as few columns of A as possible. Or more precisely, one wishes to determine the k indices such that

$$\sum_{i=1}^{n} (b_i - \sum_{\upsilon=1}^{k} a_{i j_\upsilon} \hat{x}_{j_\upsilon})^2 = \min.$$

We cannot solve this problem, but we shall show how to choose index k when the first k-1 indices are given so that the sum of squares of residuals is maximally reduced. This is the stage-wise regression problem.

We define

$$\tilde{R}^{(k+1)} = \begin{bmatrix} r_{11} & \cdots & r_{1k} \\ & r_{22} \cdots & r_{2k} \\ & \ddots & \vdots \\ & & r_{kk} \end{bmatrix} = \begin{bmatrix} a_{11}^{(2)} & \cdots, & a_{1k}^{(2)} \\ & a_{22}^{(3)} \cdots, & a_{2k}^{(3)} \\ & \ddots & \\ & & a_{kk}^{(k+1)} \end{bmatrix} .$$

Let $\underset{\sim}{c}^{(1)} = b$ and $c^{(k+1)} = P^{(k)} c^{(k)}$. Now $\tilde{R}^{(k)} \underset{\sim}{\hat{x}}^{(k-1)} = \tilde{c}^{(k)}$ where $\hat{x}^{(k-1)}$ is the least squares estimate based on $(k-1)$ columns of A and $\tilde{c}^{(k)T} = (c_1^{(k)}, c_2^{(k)}, \ldots, c_{k-1}^{(k)})$. Thus by (2.2)

$$\|\underset{\sim}{c}^{(k+1)} - \tilde{R}^{(k+1)}\hat{\underset{\sim}{x}}^{(k)}\|_2^2 = \sum_{j=k+1}^{m} (c_j^{(k+1)})^2$$

$$= \sum_{j=k}^{m} (c_j^{(k+1)})^2 - (c_k^{(k+1)})^2$$

$$= \sum_{j=k}^{m} (c_j^{(k)})^2 - (c_k^{(k+1)})^2$$

since length is preserved under an orthogonal transformation. Consequently, we wish to choose that column of $A^{(k)}$ which will maximize $|c_k^{(k+1)}|$. Let

$$t_j^{(k)} = (\sum_{i=k}^{m} a_{ij}^{(k)} c_i^{(k)}) \quad \text{for} \quad j = k+1,\ldots,n .$$

Then since $|c_k^{(k+1)}| = |(\sum_{i=k}^{m} a_{ik}^{(k)} c_i^{(k)})/s_k^{(k)}|$, one should choose that column of $A^{(k)}$ for which $(t_j^{(k)})^2/s_j^{(k)}$ is maximized. After $P^{(k)}$ is applied to $A^{(k)}$, one can adjust $t_j^{(k)}$ as follows:

$$t_j^{(k+1)} = t_j^{(k)} - a_{kj}^{(k+1)} c_k^{(k+1)}$$

In many statistical applications, if $(t_j^{(k)})^2/s_j^{(k)}$ is sufficiently small, then no further transformations are performed.

4. Statistical calculations

In many statistical calculations, it is necessary to compute certain -auxiliary information associated with A^TA. These can readily be obtained from the orthogonal decomposition. Thus

$$\det(A^TA) = (r_{11} \times r_{22} \times \ldots \times r_{nn})^2 .$$

Since

$$A^TA = \tilde{R}^T\tilde{R} \quad , \quad (A^TA)^{-1} = \tilde{R}^{-1}\tilde{R}^{-T} .$$

The inverse of $\tilde{R}$ can be readily obtained since $\tilde{R}$ is an upper triangular matrix. It is possible to calculate $(A^TA)^{-1}$ directly from R. Let

$$(A^TA)^{-1} = X = (\underset{\sim}{x}_1, \underset{\sim}{x}_2, \ldots, \underset{\sim}{x}_n).$$

Then from the relationship

$$\tilde{R}\, X = \tilde{R}^{-T}$$

and by noting that $\{\tilde{R}^{-T}\}_{ii} = 1/r_{ii}$, it is possible to compute $\underset{\sim}{x}_n, \underset{\sim}{x}_{n-1}, \ldots, \underset{\sim}{x}_1$. The number of operations is roughly the same as in the first method but more accurate bounds may be established for this method provided all inner products are accumulated to double precision.

In some applications, the original set of observations are augmented by an additional set of observations. In this case, it is not necessary to begin the calculation from the beginning again if the method of orthogonalization is used. Let $\tilde{R}_1, \underset{\sim}{\tilde{c}}_1$ correspond to the original data after it has been reduced by orthogonal transformations and let $A_2, \underset{\sim}{b}_2$ correspond to the additional observations. Then the up-dated least squares solution can be obtained directly from

$$A = \begin{bmatrix} A_2 \\ \cdots \\ \tilde{R}_1 \end{bmatrix}, \qquad b = \begin{bmatrix} \underset{\sim}{b}_2 \\ \cdots \\ \underset{\sim}{\tilde{c}}_1 \end{bmatrix}.$$

This follows immediately from the fact that the product of two orthogonal transformations is an orthogonal transformation.

The above observation has another implication. One of the arguments frequently advanced for using normal equations is that only $n(n+1)/2$ memory locations are required. By partitioning the matrix A by rows, however, then similarly only $n(n+1)/2$ locations are needed when the method of orthogonalization is used.

In certain statistical applications, it is desirable to remove a row of the matrix A after the least squares solution has been obtained. This can be done in a very simple manner. Consider the matrix

$$A = \begin{bmatrix} \tilde{R} \\ \cdots\cdots \\ i\ \underset{\sim}{\alpha} \end{bmatrix} \quad \text{and} \quad \underset{\sim}{d} = \begin{bmatrix} \underset{\sim}{\tilde{c}} \\ \cdots\cdots \\ i\ \beta \end{bmatrix}$$

where α is the row of A which one wishes to remove, β is the corresponding element of b, and $i = \sqrt{-1}$. Note that

$$S^TS = \tilde{R}^T\tilde{R} - \underset{\sim}{\alpha}^T\underset{\sim}{\alpha} = A^TA - \underset{\sim}{\alpha}^T\underset{\sim}{\alpha} .$$

Let

$$Z_{1,n+1} = \begin{bmatrix} \cos\theta & 0 & \cdots & \sin\theta \\ 0 & 1 & & \\ & & \ddots & \\ & & & 1 \\ \sin\theta & & & 0 \end{bmatrix}$$

$$S^{(1)} = S \quad , \quad \text{and } S^{(2)} = Z_{1,n+1}S^{(1)} \quad .$$

We choose cos θ so that $\{S^{(2)}\}_{n+1,1} = 0$. Thus

$$\{S^{(2)}\}_{1,1} = \sqrt{(r_{11}^2-\alpha_1^2)}$$

$$\{S^{(2)}\}_{1,j} = (r_{11}r_{1j}-\alpha_1\alpha_j)/\sqrt{(r_{11}^2-\alpha_1^2)} \qquad j = 2,3,\ldots,n$$

$$\{S^{(2)}\}_{n+1,j} = i(\alpha_1 r_{1j}-\alpha_j r_{11})/\sqrt{(r_{11}^2-\alpha_1^2)} \qquad j = 2,3,\ldots,n .$$

Note no complex arithmetic is really necessary. The process is continued as follows:

Let

$$Z_{k,n+1} = \begin{bmatrix} 1 & & & & & \\ & 1 & & & & \\ & & \ddots & & & \\ & & & \cos\theta_k & & \sin\theta_k \\ & & & & 1 & \\ & & & \sin\theta_k & & -\cos\theta_k \end{bmatrix} \begin{matrix} \\ \\ \\ k \\ \\ n+1 \end{matrix} \qquad . \qquad (4.1)$$

Then

$$S^{(k+1)} = Z_{k,n+1} S^{(k} \quad , \qquad k = 1,2,\ldots,n \ ,$$

and $\cos\theta_k$ is determined so that $(S^{(k+1)})_{k,n+1} = 0$. Thus roughly $3n^2$ multiplications and divisions and n square roots are required to form the new $\tilde{R}$.

Suppose it is desirable to add an additional variable so that the matrix A is augmented by a vector g (say). The first n columns of $\tilde{R}^{(n)}$ are unchanged. Now one computes

$$h = P^{(n)} \ldots P^{(2)} P^{(1)} g \ .$$

From h one can compute $P^{(n+1)}$ and apply it to $P^{(n)} \ldots P^{(1)} b$. This technique is also useful when an auxiliary serial storage (e.g. magnetic tape) is used.

It is also possible to drop one of the variables in a simple fashion after $\tilde{R}$ has been computed. For example, suppose we wish to drop variable 1, then

$$\tilde{R} = \begin{bmatrix} r_{12} & \cdots & r_{1n} \\ & r_{22} & \cdots & \vdots \\ & & \ddots & \\ & & & r_{nn} \end{bmatrix}_{n\times(n-1)}$$

By using plane rotations, similar to those given by (4.1), it is possible to reduce R to the triangular form again.

5. Gram-Schmidt orthogonalization

In §2, it was shown that it is possible to write

$$QA = R \ . \qquad (5.1)$$

The matrix Q is constructed as a' product of Householder transformations.

From (5.1), we see that

$$A = Q^T R \equiv PS$$

where $P^T P = I_n$, S : $\nwarrow$. Each row of S and each column of P is uniquely determined up to a scalar factor of modulus one. In order to avoid computing square roots, we modify the algorithms so that S is an upper triangular matrix with ones on the diagonal. Thus $P^T P = D$, a diagonal matrix. The calculation of P and S may be calculated in two ways.

a) Classical Gram-Schmidt Algorithm (CGSA)

The elements of S are computed one column at a time. Let

$$A^{(k)} = [\underset{\sim}{p}_1, \underset{\sim}{p}_2, \ldots, \underset{\sim}{p}_{k-1}, \underset{\sim}{a}_k, \ldots, \underset{\sim}{a}_n]$$

and assume

$$\underset{\sim}{p}_i^T \underset{\sim}{p}_j = \delta_{ij} d_i \quad , \quad 1 \leq i,j \leq k-1 \, .$$

At step k , we compute

$$s_{ik} = (\underset{\sim}{p}_i^T \underset{\sim}{a}_k / d_i) \quad , \quad 1 \leq i \leq k-1$$

$$\underset{\sim}{p}_k = \underset{\sim}{a}_k - \sum_{i=1}^{k-1} s_{ik} \underset{\sim}{p}_i \quad , \quad d_k = \|\underset{\sim}{p}_k\|_2^2 \, .$$

b) Modified Gram-Schmidt Algorithm (MGSA)

Here the elements of S are computed one row at a time. We define

$$A^{(k)} = (\underset{\sim}{p}_1, \underset{\sim}{p}_2, \ldots, \underset{\sim}{p}_{k-1}, \underset{\sim}{a}_k^{(k)}, \ldots, \underset{\sim}{a}_n^{(k)})$$

and assume

$$\underset{\sim}{p}_i^T \underset{\sim}{p}_j = \delta_{ij} d_i \, , \; \underset{\sim}{p}_i^T \underset{\sim}{a}_\ell^{(k)} = 0 \quad , \quad 1 \leq i,j \leq k-1 \, , \; k \leq \ell \leq n \, .$$

At step k , we take $\underset{\sim}{p}_k = \underset{\sim}{a}_k^{(k)}$, and compute

$$d_k = \|\underset{\sim}{p}_k\|_2^2 \, , \; s_{k\ell} = (\underset{\sim}{p}_k^T \underset{\sim}{a}_\ell^{(k)})/d_k \, , \; \underset{\sim}{a}_\ell^{(k+1)} = \underset{\sim}{a}_\ell^{(k)} - s_{k\ell} \underset{\sim}{p}_k \, , \quad k+1 \leq \ell < n \, .$$

In both procedures, $s_{kk} = 1$. The two procedures in the absence of roundoff errors, produce the same decomposition. However, they have completely different numerical properties when $n > 2$. If A is at all "ill-conditioned", then using the CGSA, the computed columns of P will soon lose their orthogonality. Consequently, one should never use the CGSA without reorthogonalization, which greatly increases the amount of computation. Reorthogonalization is never needed when using the MGSA. A careful roundoff analysis is given by BJORK [1967]. RICE [1966] has shown experimentally that the MGSA produces excellent results.

The MGSA has the advantages that it is relatively easy to program, and experimentally (cf. JORDAN [1968]), it seems to be slightly more accurate than the Householder procedure. However, it requires roughly $mn^2/2$ operations which is slightly more than that necessary in the Householder procedure. Furthermore, it is not as simple as the Householder procedure to add observations.

6. Sensitivity of the solution

We consider first the inherent sensitivity of the solution of the least squares problem. For this purpose it is convenient to introduce the condition number $\kappa(A)$ of a non-square matrix A . This is defined by

$$\kappa(A) = \sigma_1/\sigma_n \ , \ \sigma_1 = \max_{x\neq 0} \|Ax\|_2 \ / \ \|x\|_2 \ , \ \sigma_n = \min_{x\neq 0} \|Ax\|_2 \ / \ \|x\|_2$$

so that σ_1^2 and σ_n^2 are the greatest and the least eigenvalues of A^TA . From its definition it is clear that $\kappa(A)$ is invariant with respect to unitary transformations. If $\tilde{R}$ is defined as in (2.1) then

$$\sigma_1(\tilde{R}) = \sigma_1(A) \ , \ \sigma_n(\tilde{R}) = \sigma_n(A) \ , \ \kappa(\tilde{R}) = K(A) \ ,$$

while

$$\sigma_1(\tilde{R}) = \|\tilde{R}\|_2 \text{ and } \sigma_n(\tilde{R}) = 1/\|\tilde{R}^{-1}\|_2 \ .$$

The commonest method of solving least squares problems is via the normal equations

$$A^TAx = A^Tb \ . \qquad (6.1)$$

The matrix A^TA is square and we have

$$\kappa(A^TA) = \kappa^2(A) \ .$$

This means that if A has a condition number of the order of $2^{t/2}$ then A^TA has a condition number of order 2^t and it will not be possible using t-digit arithmetic to solve (6.1). The method of orthogonal transformations replaces the least squares problem by the solution of the equations $\tilde{R}x = \tilde{c}$ and $\kappa(\tilde{R}) = \kappa(A)$. It would therefore seem to have substantial advantages since we avoid working with a matrix with condition number $\kappa^2(A)$.

We now show that this last remark is an oversimplification. To this end, we compare the solution of the original system $[A \vdots b]$ with that of a perturbed system. It is convenient to assume that

$$\sigma_1 = \|A\|_2 = \|\underset{\sim}{b}\|_2 = 1 \ ;$$

this is not in any sense a restriction since we can make $\|A\|_2$ and $\|\underset{\sim}{b}\|_2$ of order unity merely by scaling by an appropriate power of two. We now have

$$\kappa(A) = \kappa(\tilde{R}) = \|\tilde{R}^{-1}\|_2 = 1/a_n \ .$$

Consider the perturbed system

$$(A + \varepsilon E \vdots \underset{\sim}{b} + \varepsilon \underset{\sim}{e}) \quad , \|E\|_2 = \|\underset{\sim}{e}\|_2 = 1 \ ,$$

where ε is to be arbitrarily small. The solution $\bar{\underset{\sim}{x}}$ of the perturbed system satisfies the equation

$$(A + \varepsilon E)^T(A + \varepsilon E)\bar{\underset{\sim}{x}} = (A + \varepsilon E)^T(\underset{\sim}{b} + \varepsilon \underset{\sim}{e}) \qquad (6.2)$$

If $\hat{\underset{\sim}{x}}$ is the exact solution of the original system and Q is the exact orthogonal transformation corresponding to A we have

$$QA = \begin{bmatrix} \tilde{R} \\ \cdots \\ 0 \end{bmatrix} \ , \quad Q(A + \varepsilon E) = \begin{bmatrix} \tilde{R} + \varepsilon F \\ \cdots\cdots \\ \varepsilon G \end{bmatrix} \ , \quad Q\underset{\sim}{e} = \begin{bmatrix} \underset{\sim}{f} \\ \cdots \\ \underset{\sim}{g} \end{bmatrix}$$

and

$$\underset{\sim}{r} = \underset{\sim}{b} - A\hat{\underset{\sim}{x}} \ , \quad A^T\underset{\sim}{r} = \underset{\sim}{0}$$

Equation (6.2) therefore becomes

$$(A + \varepsilon E)^T(A + \varepsilon E) = (A^T + \varepsilon E^T)(Ax + r + \varepsilon e)$$

giving

$$\begin{bmatrix} R + \varepsilon F \\ \cdots\cdots \\ \varepsilon G \end{bmatrix}^T \begin{bmatrix} \tilde{R} + \varepsilon F \\ \cdots\cdots \\ \varepsilon G \end{bmatrix} \bar{x} = \begin{bmatrix} R + \varepsilon F \\ \cdots\cdots \\ \varepsilon G \end{bmatrix}^T \left(\begin{bmatrix} \tilde{R} \\ \cdots \\ 0 \end{bmatrix} x + \varepsilon \begin{bmatrix} f \\ \cdots \\ g \end{bmatrix} \right) + \varepsilon E^T r \quad .$$

Neglecting ε^2 where advantageous,

$$(\tilde{R} + \varepsilon F)^T(\tilde{R} + \varepsilon F)\bar{x} = (\tilde{R} + \varepsilon F)^T \tilde{R}\hat{x} + \varepsilon(\tilde{R} + \varepsilon F)^T f + \varepsilon E^T r + O(\varepsilon^2)$$

$$\bar{x} = (\tilde{R} + \varepsilon F)^{-1} \tilde{R}x + \varepsilon(\tilde{R} + \varepsilon F)^{-1} f + \varepsilon(\tilde{R}^T\tilde{R})^{-1} E^T r + O(\varepsilon^2)$$

$$= \hat{x} - \varepsilon\tilde{R}^{-1} F\hat{x} + \varepsilon R^{-1} f + \varepsilon R^{-1} f + \varepsilon(\tilde{R}^T\tilde{R})^{-1} E^T r + O(\varepsilon^2)$$

giving

$$\|\bar{x}-\hat{x}\|_2 \leq \varepsilon\|\tilde{R}^{-1}\|_2\|F\|_2\|\hat{x}\|_2 + \varepsilon\|\tilde{R}^{-1}\|_2\|f\|_2 + \varepsilon\|\tilde{R}^{-1}\|_2^2 \|E\|_2\|r\|_2 + O(\varepsilon^2)$$

$$\leq \varepsilon\kappa(A)\|\hat{x}\|_2 + \varepsilon\kappa(A) + \varepsilon\kappa^2(A)\|r\|_2 + O(\varepsilon^2) \quad .$$

We observe that the bounds include a term $\varepsilon\kappa^2(A)\|r\|_2$. It is easy to verify by means of a 3 x 2 matrix A that this bound is realistic and that an error of this order of magnitude does indeed result from almost any such perturbation E of A . We conclude that although the use of the orthogonal transformation avoids some of the ill effects inherent in the use of the normal equations the value $\kappa^2(A)$ is still relevant to some extent.

When the equations are compatible $\|r\|_2 = 0$ and the term in $\kappa^2(A)$ disappears. In the non-singular linear equation case r is always null and hence it is always $\kappa(A)$ rather than $\kappa^2(A)$ which is relevant.

Since the sensitivity of the solution depends on the condition number, it is frequently desirable to replace the original unknowns x by a new vector of unknowns $D^{-1}x$ where D is a diagonal matrix with non-zero diagonal elements. Thus we wish to find $\hat{y}$ for which

$$\|b-C\hat{y}\|_2 = \min.$$

where $C = AD$ and $\hat{y} = D^{-1}\hat{x}$. Let $\mathcal{D}_n$ be the set of all $n \times n$ diagonal matrices with non-zero diagonal elements. We wish to choose D so that

$$\kappa(A\hat{D}) \leq \kappa(AD) \text{ for all } D \in \mathcal{D}_n \ .$$

Let $\overset{\circ}{D} \in \mathcal{D}_n$ and $\{\overset{\circ}{D}\}_{ii} = 1/\|a_i\|_2$. VAN, DER SLUIS [1968] has shown that

$$\kappa(A\overset{\circ}{D}) \leq \sqrt{n}\, \kappa(A\hat{D}) \ .$$

Therefore in the absence of other information, it would appear that it is best to precondition the matrix A so that all columns of the matrix A have equal length. In practice, one adjusts the exponents of the stored elements of A so that the mantissa of the floating point representation is not changed,

7. Iterative refinement for least squares problems

The iterative refinement method may be used for improving the solution to linear least squares problems. Let

$$\alpha\rho = b - A\hat{x} \quad , \quad \alpha > 0$$

so that

$$\alpha A^T \rho = A^T b - A^T A\hat{x} = 0$$

When $\alpha = 1$, the vector ρ is simply the residual vector r . Thus

$$\left[\begin{array}{c|c} \alpha I & A \\ \hline A^T & 0 \end{array}\right] \left[\begin{array}{c} \rho \\ \hdashline \hat{x} \end{array}\right] = \left[\begin{array}{c} b \\ \hdashline 0 \end{array}\right] \quad , \tag{7.1}$$

or

$$Cy = g \ .$$

One of the standard methods for solving linear equations may now be used to solve (7.1). However, this is quite wasteful of memory space since the dimension of the system to be solved is $(m+n)$. We may simplify this problem somewhat by noting with the aid of (2.3) that

$$\left[\begin{array}{c|c} \alpha I & A \\ \hline A^T & 0 \end{array}\right] = \left[\begin{array}{c|c} \sqrt{\alpha}\, I & 0 \\ \hline \frac{1}{\sqrt{\alpha}} A^T & \frac{1}{\sqrt{\alpha}} \tilde{R}^T \end{array}\right] \left[\begin{array}{c|c} \sqrt{\alpha}\, I & \frac{1}{\sqrt{\alpha}} A \\ \hline 0 & -\frac{1}{\sqrt{\alpha}} \tilde{R} \end{array}\right] \equiv LU \,. \qquad (7.2)$$

Once an approximate solution to $Cy = g$ has been obtained, it is frequently possible to improve the accuracy of the approximate solution. Let $\bar{y}$ be an approximate solution, and let $v = g - C\bar{y}$. Then if $y = \bar{y} + \delta$, δ satisfies the equation

$$C\delta = v \,. \qquad (7.3)$$

Equation (7.3) can be solved approximately from the decomposition (7.2). Of course, it is not possible to solve precisely for δ so that the process may be repeated.

We are now in a position to use the iterative refinement method (cf. MOLER [1967], WILKINSON [1967]) for solving linear equations. Thus one might proceed as follows:

1) Solve for $x^{(0)}$ using one of the orthgonalization procedures outlined in § 2 or 5. $\tilde{R}$ must be saved but it is not necessary to retain Q . Then

$$\rho^{(0)} = \frac{1}{\alpha} (b - Ax^{(0)})$$

2) The vector $y^{(s+1)}$ is determined from the relationship

$$y^{(s+1)} = y^{(s)} + \delta^{(s)}$$

where

$$C\delta^{(s)} = g - Cy^{(s)} \equiv v^{(s)} \,. \qquad (7.4)$$

This calculation is simplified by solving

$$Lz^{(s)} = v^{(s)}$$

$$U\delta^{(s)} = (z)^s \,.$$

The vector $v^{(s)}$ must be calculated using double precision accuracy and then rounding to single precision.

3) Terminate the iteration when $\|\delta^{(s)}\| / \|y^{(s)}\|$ is less than a prescribed number.

Note that the computed residual vector is an approximation to the residual vector when the exact solution $\hat{x}$ is known. This may differ from the residual vectorcomputed from the approximate solution to the least squares problem.

There are three sources of error in the process: (1) computation of the vector $v^{(s)}$, (2) solution of the system of equations for the correction vector $\delta^{(s)}$, and (3) addition of the correction vector to the approximation $y^{(s)}$. It is absolutely necessary to compute the components of the vector $v^{(s)}$ using double precision inner products and then to round to single precision accuracy. The convergence of the iterative refinement process has been discussed in detail by MOLER [1967]. Generally speaking, for a large class of matrices for $k \geq k_o$ all components of $y^{(s)}$ are the correctly rounded single precision approximations to the components of y. There are exceptions to this, however, (cf. KAHAN [1966]). Experimentally, it has been observed, in most instances, that if $\|\delta^{(o)}\|_\infty / \|y^{(o)}\|_\infty \leq 2^{-p}$ where

$$\|v\|_\infty = \max_{1 \leq i \leq n} |v_i| \quad ,$$

then $k_o \geq [t/p]$. We shall return to the subject of iterative refinement when we discuss the solution of linear least squares problem with linear constraints.

A variant of the above procedure has been analyzed by BJÖRCK [1967b], [1968], and he has also given an ALGOL procedure. This has proved to be a very effective method for obtaining highly accurate solutions to linear least squares problems.

8. Least squares problems with constraints

Frequently, one wishes to determine $\hat{x}$ so that $\|b-A\hat{x}\|_2$ is minimized subject to the condition that $G\hat{x} = h$ where G is a $p \times n$ matrix of rank p . One can, of course, eliminate p of the columns of A by Gaussian elimination after a $p \times p$ non-singular submatrix of G has been determined and then solve

the resulting normal equations. This, unfortunately, would not be a numerically stable scheme since no row interchanges between A and G would be permitted.

If one uses Lagrange multipliers, then one must solve the $(n+p)x(n+p)$ system of equations.

$$\left[\begin{array}{c|c} A^TA & G^T \\ \hline G & 0 \end{array}\right]\left[\begin{array}{c} \hat{x} \\ \cdots \\ \lambda \end{array}\right] = \left[\begin{array}{c} A^Tb \\ \cdots \\ h \end{array}\right] \tag{8.1}$$

where λ is the vector of Lagrange multipliers. Since $\hat{x} = (A^TA)^{-1}A^Tb-(A^TA)^{-1}G^T\lambda$,

$$G(A^TA)^{-1}G^T\lambda = Gz-g$$

where

$$z = (A^TA)^{-1}A^Tb .$$

Note z is the least squares solution of the original problem without constraints and one would frequently wish to compare this vector with the final solution $\hat{x}$. The vector z, of course, should be computed by the orthogonalization procedures discussed earlier.

Since $A^TA = \tilde{R}^T\tilde{R}$, $G(A^TA)^{-1}G^T = W^TW$ where $W = \tilde{R}^{-T}G^T$. After W is computed, it should be reduced to a pxp upper triangular matrix K by orthogonalization. The matrix equation

$$K^TK\lambda = Hz-g$$

should be solved by the obvious method. Finally, one computes

$$\hat{x} = z-(A^TA)^{-1}G\lambda$$

where $(A^TA)^{-1}G\lambda$ can be easily computed by using $\tilde{R}^{-1}$.

It is also possible to use the techniques described in §7. Again, let $r = b-A\hat{x}$ so that from (8.1)

$$\left[\begin{array}{c|c|c} I & A & 0 \\ \hline A^T & 0 & G^T \\ \hline 0 & G & 0 \end{array}\right]\left[\begin{array}{c} r \\ \cdots \\ \hat{x} \\ \cdots \\ \lambda \end{array}\right] = \left[\begin{array}{c} b \\ \cdots \\ 0 \\ \cdots \\ h \end{array}\right] \tag{8.2}$$

or

$$Dz=g \quad .$$

Note D is an $(m+n+p)\times(m+n+p)$ matrix. We may simplify the solution of (8.2), however, by noting that

$$\left[\begin{array}{c|c|c} I & A & 0 \\ \hline A^T & 0 & G^T \\ \hline 0 & G & 0 \end{array}\right] = \left[\begin{array}{c|c|c} I & 0 & 0 \\ \hline A^T & -\tilde{R}^T & 0 \\ \hline 0 & B^T & S^T \end{array}\right] \left[\begin{array}{c|c|c} I & A & 0 \\ \hline 0 & \tilde{R} & -B \\ \hline 0 & 0 & S \end{array}\right] \qquad (8.3)$$

where $B = (\tilde{G}\tilde{R}^{-1})^T = PS$ and $P^TP = I$ with $S : \nabla$. The decomposition (8.3) can be used very effectively in conjunction with the method of iterative refinement. BJÖRCK and GOLUB [1967] have given a variant of the above procedure which requires Q and P .

9. Linear least squares solutions with inequality constraints

Again let A,G be given real matrices of orders $m\times n$, $p\times n$, with $m \geq n$, and let b , h be given real vectors of orders m , p . For any vector $\underset{\sim}{x}$ we define

$$r = \underset{\sim}{b}-A\underset{\sim}{x}$$

and we wish to determine an $\underset{\sim}{x}$ such that

$$\underset{U}{r^T}\underset{N}{r} = \min.'$$

subject to

$$G\underset{\sim}{x} \geq \underset{\sim}{h} \quad .$$

Our problem can therefore be stated as follows: find r , $\underset{\sim}{x}$, w such that

$$\underset{\sim}{r} + Ax = \underset{\sim}{b}$$

$$G x - w = h$$

$$\underset{\sim}{w} \geq \underset{\sim}{0}$$

$$\underset{\sim}{r}^T\underset{\sim}{r} = \min.$$

These problems can be solved by quadratic programming but we present an algorithm in this section which leads to a much smaller system of equations and highly accurate results.

If we define

$$f(\underset{\sim}{r},\underset{\sim}{w},\underset{\sim}{x},\underset{\sim}{y},\underset{\sim}{z}) = \frac{1}{2}\,\underset{\sim}{r}^T\underset{\sim}{r} - \underset{\sim}{y}^T(\underset{\sim}{r}+A\underset{\sim}{x}-\underset{\sim}{b}) - \underset{\sim}{z}^T(G\underset{\sim}{x}-\underset{\sim}{w}-\underset{\sim}{h})$$

where we require without loss of generality that $\underset{\sim}{z} \geq \underset{\sim}{0}$, then an equivalent problem is to determine $\underset{\sim}{r},\underset{\sim}{w},\underset{\sim}{x},\underset{\sim}{y},\underset{\sim}{z}$ such that

$$\underset{\sim}{w},\underset{\sim}{z} \geq \underset{\sim}{0}$$

$$f = \min.$$

Equating to zero the partial derivatives of f with respect to $\underset{\sim}{r},\underset{\sim}{x},\underset{\sim}{y},\underset{\sim}{z}$ respectively, we get

$$\underset{\sim}{r} - \underset{\sim}{y} = \underset{\sim}{0}$$

$$-A^T\underset{\sim}{y} - G^T\underset{\sim}{z} = \underset{\sim}{0}$$

$$\underset{\sim}{r} + A\underset{\sim}{x} - \underset{\sim}{b} = \underset{\sim}{0}$$

$$G\underset{\sim}{x} - \underset{\sim}{w} - \underset{\sim}{h} = 8 \quad .$$

Further, let the elements of $\underset{\sim}{w},\underset{\sim}{z}$ be w_i, z_i $(i = 1,2,\ldots,p)$. Then

$$\frac{\partial f}{\partial w_i} = z_i \quad .$$

Now if $w_i > 0$ in the optimal solution, the constraint $w_i \geq 0$ is not -binding and we have

$$\frac{\partial f}{\partial w_i} = 0 \quad ,$$

i.e.,

$$w_i > 0 \Rightarrow z_i = 0 \quad .$$

Since $z_i \geq 0$, this further means that

$$z_i > 0 \Rightarrow w_i = 0 \ .$$

(For otherwise, $z_i > 0 \Rightarrow w_i > 0 \Rightarrow z_i = 0$ which is a contradiction.)

Accordingly, our problem has become one of finding a solution of the system

$$\underset{\sim}{r} + A\underset{\sim}{x} = \underset{\sim}{b} \tag{9.1}$$

$$A^T\underset{\sim}{r} + G^T\underset{\sim}{z} = \underset{\sim}{\theta} \tag{9.2}$$

$$G\underset{\sim}{x} - \underset{\sim}{w} = \underset{\sim}{h} \tag{9.3}$$

such that

$$\underset{\sim}{z} \geq \underset{\sim}{\theta} \ , \quad \underset{\sim}{w} \geq \underset{\sim}{\theta} \ , \quad \underset{\sim}{z}^T\underset{\sim}{w} = 0 \ .$$

We now determine an orthogonal matrix Q and an upper-triangular matrix R such that

$$A = QR \ ,$$

where R is nxn and non-singular if rank(A) = n . Then

$$A^TA = R^TQ^TQR = R^TR \ .$$

Letting $B = (GR^{-1})^T$ and eliminating r from (9.1) and (9.2) it is easily verified that

$$x = \hat{x} + R^{-1}Bz \ , \tag{9.4}$$

where

$$\hat{x} = (R^TR)^{-1}A^Tb$$

is the unconstrained least squares solution (i.e., the solution of (9.1) and (9.2) with $\underset{\sim}{z} = \underset{\sim}{\theta}$). $\hat{\underset{\sim}{x}}$ is found by the methods of §7.

We now determine if $\hat{\underset{\sim}{x}}$ satisfies the original inequalities: if we define $\underset{\sim}{q} = G\hat{\underset{\sim}{x}} - \underset{\sim}{h}$ and find that $\underset{\sim}{q} \geq \underset{\sim}{\theta}$ then the constraints are satisfied and $\hat{x}$ solves the problem.

"Otherwise, we substitute **(9.4)** in **(9.3)** and obtain

$$G(\hat{x} + R^{-1}Bz) - w = h$$

or

$$\left.\begin{aligned} B^T B z + q &= w \\ \text{where we further require'} \quad & \\ z \geq 0 \quad , \quad w \geq 0 \quad , \quad z^T w &= 0 \end{aligned}\right\} \tag{9.5}$$

Thus we find that z,w solve the linear complementarity problem (LCP) defined by (9.5). This is a fundamental mathematical programming problem and several algorithms have been developed for finding solutions (e.g. see . LEMKE [1968], COTTLE [1968], COTTLE and DANTZIG [1968]). The matrix $M = B^T B$ is positive semi-definite, and this is one of the cases when, for example, the principal pivoting method in COTTLE [1968] guarantees termination with a solution, or with an indication that none exists.

Once z has been found it would be a simple matter to substitute into (9.1), (9.2) and find r,x from

$$\left.\begin{aligned} r + A x &= b \\ A^T r &= -G^T z \end{aligned}\right\} \tag{9.6}$$

In practice, however, if we are concerned with the accuracy of our estimate of x we use the solution of the LCP (9.5) only to determine which elements of w are exactly zero. These are the w_i which are non-basic in the solution of (9.5). (There is certainly at least one such w_i, for otherwise we would have $z = 0$, $w \geq 0$, which is the case checked for earlier in determining whether or not $\hat{x}$ solved the problem.)

We now delete from (9.3) those constraints for which w_i is basic, obtaining an $\ell \times n$ system of equations

$$\tilde{G}x = \tilde{h}$$

where $1 \leq \ell \leq p$.

If $\tilde{z}$ is the vector z with the corresponding elements deleted, the remaining step is to solve the system

$$
\begin{aligned}
r + Ax &= b \\
A^T r + \tilde{G}^T \tilde{z} &= \underset{\sim}{0} \\
\tilde{G}x &= h
\end{aligned}
\tag{9.7}
$$

where we are now working with original data and can therefore expect a more accurate solution than could be obtained from (9.6). We can now apply the methods of §8 to this system of equations.

The standard methods for solving the linear complementarity problem employ the elements of $\underset{\sim}{w}$ as the initial set of basic variables, with all elements of $\underset{\sim}{z}$ initially non-basic. In general, it is probable that only a small proportion of the inequalities in the original problem will be constraining the system, which means that only a small proportion of the w_i will be non-zero. Hence it might be expected in general that only a small number of iterations (relative to p) should be required to bring some of the z_i into the basis and reach a feasible solution.

In our particular form of the problem, since the matrix $M = B^T B$ has its largest elements on the diagonal, accuracy can be conserved, to within the limits of the error in forming M , by interchanging rows whenever a column of M is brought into the basis in such a way that the diagonal elements of M become diagonal elements of the basis matrix. This is easily done if the LU decomposition of the basis is calculated each iteration as in the treatment of the simplex method by BARTELS [1968] and BARTELS and GOLUB [1969].

Note that $B = (GR^{-1})^T$ can be determined column by column via repeated back-substitution on the system

$$R^T B = G^T \quad .$$

The algorithm presented here can be used for any quadratic programming problem when a positive definite quadratic form is given. Suppose we wish to determine an x such that

$$
\left.
\begin{aligned}
&\underset{\sim}{x}^T C \underset{\sim}{x} + \underset{\sim}{d}^T \underset{\sim}{x} = \min. \\
\text{subject to} \quad &G\underset{\sim}{x} \geq \underset{\sim}{h}
\end{aligned}
\right\}
\tag{9.8}
$$

Since C is positive definite, we may write

$$C = R^T R$$

where $R(\nabla)$ is the Cholesky factor of C . Such a decomposition can easily be computed. If we now define $\underset{\sim}{b} = -\frac{1}{2} R^{-T} d$ (and calculate b from $R^T b = -\frac{1}{2} d$) we find that

$$\|\underset{\sim}{b} - R\underset{\sim}{x}\|_2^2 = \underset{\sim}{b}^T\underset{\sim}{b} - 2\underset{\sim}{b}^T R\underset{\sim}{x} + \underset{\sim}{x}^T R^T R\underset{\sim}{x}$$

$$= \underset{\sim}{b}^T\underset{\sim}{b} + \underset{\sim}{d}^T\underset{\sim}{x} + \underset{\sim}{x}^T C\underset{\sim}{x}$$

and consequently if we determine an $\underset{\sim}{x}$ such that

$$\|\underset{\sim}{b} - R\underset{\sim}{x}\|_2 = \min.$$

subject to $G\underset{\sim}{x} \geq \underset{\sim}{h}$

then $\underset{\sim}{x}$ will satisfy (9.8) as required.

10. Singular systems

If the rank of A is less than n and if column interchanges are performed to maximize the diagonal elements of R , then

$$A^{(r+1)} = \left[\begin{array}{c|c} \tilde{R}_{rxr} & S_{(n-r)xr} \\ \hline 0 & 0 \end{array}\right]$$

when rank(A) = r . A sequence of Householder transformations may now be applied on the right of $A^{(r+1)}$ so that the elements of $S_{(n-r)xr}$ become annihilated. Thus dropping subscripts and superscripts, we have

$$QAZ = T = \left[\begin{array}{c|c} \tilde{T} & 0 \\ \hline 0 & 0 \end{array}\right]$$

where T is an rxr upper triangular matrix. Now

$$\|b - A\underset{\sim}{x}\|_2 = \|\underset{\sim}{b} - Q^T T Z^T \underset{\sim}{x}\|_2$$

$$= \|c - T\underset{\sim}{y}\|_2$$

where $\underset{\sim}{c} = Qb$ and $\underset{\sim}{y} = Z^T\underset{\sim}{x}$. Since T is of rank r, there is no unique solution so that we impose the condition that $\|\underset{\sim}{\hat{x}}\|_2 = \min$. But $\|\underset{\sim}{y}\|_2 = \|\underset{\sim}{x}\|_2$ since T is orthogonal and $\|\underset{\sim}{y}\|_2 = \min$. when

$$y_{r+1} = y_{r+2} = \ldots = y_m = 0 .$$

Thus

$$\hat{x} = Z \left[\begin{array}{c|c} \tilde{T}^{-1} & 0 \\ \hline 0 & 0 \end{array}\right] Qb$$

This solution has been given by FADEEV, et. al. [1968] and HANSON and LAWSON [1968]. The problem still remains how to numerically determine the rank which will be discussed in §12.

11. Singular value decomposition

Let A be a real, mxn matrix (for notational convenience we assume that $m \geq n$). It is well known (cf. LANCZOS [1951]) that

$$A = U\Sigma V^T \tag{11.1}$$

where

$$UU^T = I_m \quad , \quad VV^T = I_n$$

and

$$\Sigma = \left[\begin{array}{c} \sigma_1 \\ \ddots \\ \hline \sigma_n \\ \hline 0 \end{array}\right] \}(m-n)xn$$

The matrix U consists of the orthonormalized eigenvectors of AA^T , and the matrix V consists of the orthonormalized eigenvectors of A^TA . The

diagonal elements of Σ are the non-negative square roots of the eigenvalues of A^TA ; they are called singular values or principal values of A . We assume

$$\sigma_1 \geq \sigma_2 > \ldots > \sigma_n \geq 0 \quad .$$

Thus if rank(A) = r , $\sigma_{r+1} = \sigma_{r+2} = \ldots = \sigma_n = 0$. The decomposition (11.1) is called the singular value decomposition (SVD).

Let

$$\tilde{A} = \begin{bmatrix} 0 & \\ A^T & 0 \end{bmatrix} \qquad (11.2)$$

It can be shown that the non-zero eigenvalues of $\tilde{A}$ always occur in $\pm$ pairs, viz.

$$\lambda_j(\tilde{A}) = \pm\, \sigma_j(A) \qquad (j = 1,2,\ldots,r) \; . \qquad (11.3)$$

12. Applications of the SVD

The singular value decomposition plays an important role in a number of least squares problems, and we will illustrate this with some examples. Throughout this discussion, we use the Euclidean or Frobenius norm of a matrix, viz.

$$\|A\| = (\Sigma |a_{ij}|^2)^{1/2} \quad .$$

A) Let $\mathcal{U}_n$ be the set of all nxn orthogonal matrices. For an arbitrary nxn real matrix A , determine $Q \in \mathcal{U}_n$ such that

$$\|A-Q\| \leq \|A-X\| \quad \text{for any } X \in \mathcal{U}_n .$$

It has been shown by FAN and HOFFMAN [1955] that if

$$A = U\Sigma V^T \quad , \text{ then } Q = UV^T \quad .$$

B) An important generalization of problem A occurs in factor analysis. For arbitrary $n \times n$ real matrices A and B, determine $Q \in \mathcal{U}_n$ such that

$$\|A-BQ\| \leq \|A-BX\| \quad \text{for any } X \in \mathcal{U}_n \ .$$

It has been shown by GREEN [1952] and by SCHÖNEMANN [1966] that if

$$B^T A = U \Sigma V^T \quad , \quad \text{then } Q = W^T \ .$$

C) Let $\mathcal{M}_{m,n}^{(k)}$ be the set of all $m \times n$ matrices of rank k. Assume $A \in \mathcal{M}_{m,n}^{(r)}$. Determine $B \in \mathcal{M}_{m,n}^{(k)}$ $(k \leq r)$ such that

$$\|A-B\| \leq \|A-X\| \quad \text{for all } X \in \mathcal{M}_{m,n}^{(k)}$$

It has been shown by ECKART and YOUNG [1936] that if

$$A = U \Sigma V^T \ , \quad \text{then } B = U \Omega_k V^T \tag{12.1}$$

where

$$\Omega_k = \begin{bmatrix} \sigma_1 & & 0 \\ & \sigma_2 & \\ & \ddots & \\ 0 & & \sigma_k \\ \hline & 0 & \end{bmatrix} \tag{12.2}$$

Note that

$$\|A-B\| = \|\Sigma - \Omega_k\| = (\sigma_{k+1}^2 + \ldots + \sigma_r^2)^{1/2} \ . \tag{12.3}$$

D) An $n \times m$ matrix X is said to be the pseudo-inverse of an $m \times n$ matrix A if X satisfies the following four properties:

i) $AXA = A$,

ii) $XAX = X$,

iii) $(AX)^T = AX$,

iv) $(XA)^T = xA$.

We denote the pseudo-inverse by A^+ . We wish to determine A^+ numerically. It can be shown (cf. PENROSE [1955]) that A^+ can always be determined and is unique. It is easy to verify that

$$A+ = V\Lambda U^T \tag{12.4}$$

where

$$\Lambda = \begin{bmatrix} \frac{1}{\sigma_1} & & & 0 \\ & \frac{1}{\sigma_2} & & \\ & & \ddots & \\ 0 & & & \frac{1}{\sigma_r} \\ & & & \end{bmatrix}_{n\times n}$$

In recent years there have been a number of algorithms proposed for computing the pseudo-inverse of a matrix. These algorithms usually depend upon a knowledge of the rank of the matrix or upon some suitable chosen parameter. For example in the latter case, if one uses (12.4) to compute the pseudo-inverse, then after one has computed the singular value decomposition numerically it is necessary to determine which of the singular values are zero by testing against some tolerance.

Alternatively, suppose we know that the given matrix A can be represented as

$$A = B+\delta B$$

where δB is a matrix of perturbations and

$$\|\delta B\| \leq \eta \ .$$

Now, we wish to construct a matrix $\hat{B}$ such that

$$\|A-\hat{B}\| \leq \eta$$

and

$$\text{rank}\,(\hat{B}) = \text{minimum}\ .$$

This can be accomplished with the aid of the solution to problem (C). Let

$$B_k = U\Omega_k V^T$$

where Ω_k is defined as in (12.2). Then using (12.3),

$$\hat{B} = B_p$$

if

$$(\sigma_{p+1}^2 + \sigma_{p+2}^2 + \cdots + \sigma_n^2)^{1/2} \leq \eta$$

and

$$(\sigma_p^2 + \sigma_{p+1}^2 + \ldots + \sigma_n^2)^{1/2} > \eta\ .$$

Since rank($\$) = p by construction,

$$\hat{B}^+ = V\Omega_p^+ U^T\ .$$

Thus, we take $\hat{B}^+$ as our approximation to A^+.

E) Let A be a given matrix, and $\underset{\sim}{b}$ be a known vector. Determine $\hat{x}$ so that amongst all $\underset{\sim}{x}$ for which $\|\underset{\sim}{b}-A\underset{\sim}{x}\|_2 = \min$, $\|\hat{x}\|_2 = \min$. It is easy to verify that

$$\hat{\underset{\sim}{x}} = A^+b\ .$$

13. Calculation of the SVD

It was shown by GOLUB and KAHAN[1965] that it is possible to construct a sequence of orthogonal matrices

$$\left\{P^{(k)}\right\}_{k=1}^{n} \quad , \quad \left\{Q^{(k)}\right\}_{k=1}^{n-1}$$

via Householder transformation so that

$$P^{(n)}P^{(n-1)}\cdots P^{(1)}AQ^{(1)}Q^{(2)}\cdots Q^{(n-1)} \equiv P^{T}AQ = J$$

and J is an mxn bi-diagonal matrix of the form

$$\begin{bmatrix} \alpha_1 & \beta_1 & 0 & \cdot & 0 \\ & \alpha_2 & \beta_2 & \cdot & 0 \\ & & \cdot & & \\ & & & \cdot & \beta_{n-1} \\ & & & & \alpha_n \\ \hline & & 0 & & \end{bmatrix} \Big\}(m-n)\times n \quad .$$

The singular values of J are the same as those of A . Thus if the singular value decomposition of

$$J = X\Sigma Y^{T}$$

then

$$A = PX\Sigma Y^{T}Q^{T}$$

so that

$$U = PX \quad , \quad V = QY \quad .$$

GOLUB[1968] has given an algorithm for computing the SVD of J ; the algorithm is based on the highly'effective QR algorithm of FRANCIS [1961,1962] for computing the eigenvalues.

It is not necessary to compute the complete SVD when a vector $\underset{\sim}{b}$ is given. Since $\hat{\underset{\sim}{x}} = V\Sigma^{+}U^{T}\underset{\sim}{b}$, it is only necessary to compute V,Σ and $U^{T}\underset{\sim}{b}$; note, this has a strong "flavor of principal component analysis. An ALGOL" procedure for the SVD has beeen given by GOLUB and REINSCH[1969].

14. Quadratic constraints

We wish to determine $\hat{x}$ so that

$$\|\underset{\sim}{b}-A\underset{\sim}{\hat{x}}\|_2 = \min.$$

when

$$\|\underset{\sim}{\hat{x}}\|_2 = \alpha \quad .$$

Such problems occur in a number of situations, e.g. in the numerical solution of integral equations of the first kind (cf. PHILLIPS [1962]), and in the solution of non-linear least squares problems (cf. MARQUARDT [1963]). Using Lagrange multipliers, we are led to the equation

$$(A^TA-\lambda^*I)\hat{x} = A^Tb$$

where the real constant λ^* is determined as the smallest root of

$$\alpha^2-b^TA(A^TA-\lambda I)^{-2}\,A^Tb = 0 \quad . \tag{14.1}$$

Using the decomposition $A = U\Sigma V^T$ and $\underset{\sim}{c} = U^Tb$, equation (14.1) becomes

$$\alpha^2-\underset{\sim}{c}^T\Sigma(\Sigma^2-\lambda I)^{-2}\Sigma\underset{\sim}{c} = 0 \quad .$$

A combination of bisection and Newton iteration may be used to determine λ^* . It is easily shown that $\lambda^* < \sigma^2_{min}$ (cf. FORSYTHE and GOLUB[1965]).

It is also possible to determine λ^* as a solution to an eigenvalue problem using a technique given by FORSYTHE and GOLUB [1965]. Consider the identity

$$\det\begin{bmatrix} X & Y \\ Z & W \end{bmatrix} = \det(X)\,\det\,(W-ZX^{-1}Y)$$

which is valid for any partitioned matrix with X and W square and $\det(X) \neq 0$. Thus (14.1) is equivalent to the determinantal equation

$$\det\begin{bmatrix} (A^TA-\lambda I)^2 & A^T\underset{\sim}{b} \\ \underset{\sim}{b}^TA & \alpha^2 \end{bmatrix} = 0 \quad .$$

Now there exists a vector p and a number q such that

$$(A^TA-\lambda I)^2 p + A^T b q = 0 \quad , \quad \underset{\sim}{b}^T A \underset{\sim}{p} + \alpha^2 q = 0 \ .$$

A simple elimination shows that λ^* -must satisfy the determinantal equation

$$\det[(A^TA-\lambda I)^2 - \alpha^{-2} A^T \underset{\sim\sim}{bb}^T A] = 0 \qquad (14.2)$$

It is possible to transform (14.2) into a $2n \times 2n$ ordinary eigenvalue problem.

Once λ^* is determined, the solution $\hat{x}$ can be computed from the SVD of A . Thus,

$$\hat{x} = V(\Sigma-\lambda^*\Sigma^{-1})^{-1} \underset{\sim}{c} \ .$$

References

Bartels, R. H. [1968], "A numerical investigation of the simplex method", Technical Report No. CS 104, Computer Science Dept., Stanford University, California.

Bartels, R. H. and G. H. Golub [1969], "The simplex method of linear programming using LU decomposition", Comm. ACM 12, 5, pp. 266-268.

Björck, Å. [1967a], "Solving linear least squares problems by Gram-Schmidt orthogonalization", BIT, 7, pp. 1-21.

Björck, Å. [1967b], "Iterative refinement of linear least squares solution I", BIT, 7, pp. 257-278.

Björck, Å. [1968], "Iterative refinement of linear least squares solutions II", BIT, 8, pp. 8-30.

Björck, Å. and G. H. Golub [1967], "Iterative refinement of linear least squares solutions by Householder transformation", BIT, 7, pp. 322-337.

Businger, P. and G. H. Golub [1965], "Linear least squares solutions by Householder transformations", Num. Math., 7, pp. 269-276.

Cottle, R. W. [1968], "The principal pivoting method of quadratic programming", Mathematics of the Decision Sciences, Part 1, G. B. Dantzig and A. F. Veinott, eds., pp. 144-162.

Cottle, R. W. and G. B. Dantzig [1968], "Complementary pivot theory of mathematical programming", in Mathematics of the Decision Sciences, Part 1, G. B. Dantzig and A. F. Veinott, eds., pp. 115-136.

Eckart, C. and G. Young [1936], "The approximation of one matrix by another of lower rank", Psychometrika, 1, pp. 211-218.

Fadeev, D. K., V. N. Kublanovskaya, and V. N. Fadeeva [1968], "Sur les systemes linearires algebriques de matrices rectangulaires et mal-conditionnees", in Programmation en Mathematriques Numeriques, Editions du Centre National de la Recherche Scientifique, Paris VII.

Fan, K. and A. Hoffman [1955], "Some metric inequalities in the space of matrices'", Proc. Amer. Math. Soc., 6, pp. 111-116.

Forsythe, G. E. and G. H. Golub [1965], "On the stationary values of a second-degree polynomial on the unit sphere", J. SIAM, 13, pp. 1050-1068.

Forsythe, G. E. and C. Moler [1967], Computer Solution of Linear Algebraic Systems, Prentice-Hall, Englewood Cliffs, New Jersey.

Francis, J. [1961, 1962], "The QR transformation. A unitary analogue to the LR transformation'", Comput. J., 4, pp. 265-271.

Golub, G. H. [1965], "Numerical methods for solving linear least squares problems'", Num. Math., 7, pp. 206-216.

Golub, G. H. and W. Kahan [1965], "Calculating the singular values and pseudo-inverse of a matrix", J. SIAM, Numer. Anal. Ser. B, 2, pp. 205-224.

Golub, G. H. and J. Wilkinson [1966], "Iterative refinement of least squares solution", Num. Math., 9, pp. 139-148.

Golub, G. H. [1968], "Least squares, "singular values and matrix approximations", Aplikace Matematiky, 13, pp. 44-51.

Golub, G. H. and C. Reinsch [1969], "Singular value decomposition and least squares solution", Technical Report No. CS 131, Computer Science Dept., Stanford University, California.

Green, B. [1952], "The orthogonal approximation of an oblique structure in factor analysis", Psychometrika, 17, pp. 429-440.

Hanson, R. and C. Lawson [1968], "Extensions and applications of the Householder algorithm for solving linear least squares problems", Jet Propulsion Laboratory.

Householder, A.-S. [1958], "Unitary triangularization of a nonsymmetric matrix", J. Assoc. Comput. Mach., 5, pg. 339-342.

Jordan, T. [1968], "Experiments on error growth associated with some linear least-squares procedures", Math. Comp., 22, pp. 579-588.

Kahan, W. [1966], "Numerical linear algebra", Canad. Math. Bull., 9, pp. 757-801.

Kalfon, P., G. Ribière, and J. Sogno [1968], "Méthode du gradient projeté utilisant la triangularisation unitaire". Publication no. FT/11.3.8/AI Centre National de la Recherche Scientifique Institut Blaise Pascal.

Lanczos, C. [1961], Linear Differential Operators, Van Nostrand, London, Chap. 3.

Lemke, C. E. [1968], "On complementary pivot theory", Mathematics of the Decision Sciences, Part 1, G. B. Dantzig and A. F. Veinott, eds., pp. 95-114.

Longley, J. [1967], "An appraisal of least squares problems for the electronic computer from the point of view of the user", JASA, 62, pp. 819-841.

Marquardt, W. [1963], "An algorithm for least-squares estimation of nonlinear parameters", J. SIAM, 11, pp. 431-441.

Moler, C. B. [1967], "Iterative refinement in floating point", J. Assoc. Comput. Mach., 14, pp. 316-321.

Penrose, R. [1955], "A generalized inverse for matrices", Proc. Cambridge Philos. Soc., 51, pp. 406-413.

Phillips, P. [1962], "A technique for the numerical solution of certain integral equations of the first kind", J. Assoc. Comput. Mach., 9, pp. 84-97.

Rice, J. [1966], "Experiments on Gram-Schmidt Orthogonalization", Math. Comp., 20, pp. 325-328.

Schönemann, P. [1966], "A generalized solution of the orthogonal procrustes problem", Psychometrika, 31, pp. l-10.

van der Sluis, A. [1968], "Condition numbers and equilibration of matrices", (unpublished report from Utrecht University, ERCU 37-4).

Wilkinson, J. H. [1963], Rounding Errors in Algebraic Processes, Prentice-Hall, Englewood Cliffs, New Jersey.

Wilkinson, J. H. [1967], "The solution of ill-conditional linear equations", Mathematical Methods for Digital Computers, Vol. II, A. Ralston, Ph.D. and H. Wilf, Ph.D., eds., John Wiley, New York, pp. 65-93.