

CALCULATING THE SINGULAR VALUES AND PSEUDO-INVERSE OF A MATRIX*

G. GOLUB† AND W. KAHAN‡

Abstract. A numerically stable and fairly fast scheme is described to compute the unitary matrices U and V which transform a given matrix A into a diagonal form $\Sigma = U^*AV$, thus exhibiting A 's singular values on Σ 's diagonal. The scheme first transforms A to a bidiagonal matrix J , then diagonalizes J . The scheme described here is complicated but does not suffer from the computational difficulties which occasionally afflict some previously known methods. Some applications are mentioned, in particular the use of the pseudo-inverse $A^I = V\Sigma^IU^*$ to solve least squares problems in a way which dampens spurious oscillation and cancellation.

1. Introduction. This paper is concerned with a numerically stable and fairly fast method for obtaining the following decomposition of a given rectangular matrix A :

$$(1.1) \quad A = U\Sigma V^*,$$

where U and V are unitary matrices and Σ is a rectangular diagonal matrix of the same size as A with nonnegative real diagonal entries. These diagonal elements are called the *singular values* or *principal values* of A ; they are the nonnegative square roots of the eigenvalues of A^*A or AA^* .

Some applications of the decomposition (1.1) will be mentioned in this paper. In particular, the pseudo-inverse A^I of A will be represented in the form

$$(1.2) \quad A^I = V\Sigma^IU^*,$$

where Σ^I is obtained from Σ by replacing each positive diagonal entry by its reciprocal. The properties and applications of A^I are described in papers by Greville [15], Penrose [25], [26], and Ben-Israel and Charnes [3]. The pseudo-inverse's main value, both conceptually and practically, is that it provides a solution for the following least-squares problem.

Of all the vectors \mathbf{x} which minimize the sum of squares $\|\mathbf{b} - A\mathbf{x}\|^2$, which is the shortest (has the smallest $\|\mathbf{x}\|^2 = \mathbf{x}^\mathbf{x}$)?*

The solution is $\mathbf{x} = A^I\mathbf{b}$. If there were only one vector \mathbf{x} which minimized

* Received by the editors July 14, 1964. This paper was first presented at the Symposium on Matrix Computations at Gatlinburg, Tennessee, in April, 1964.

† Computation Center, Stanford University, Stanford, California. The work of the first author was in part supported by Contract Nonr 225(37) at Stanford University and by The Boeing Scientific Research Laboratories, Seattle, Washington.

‡ University of Toronto, Toronto, Ontario. The second author wishes to thank the National Research Council (of Canada) for their support of the Institute of Computer Sciences at the University of Toronto.

$\| \mathbf{b} - A\mathbf{x} \|$ we would save a bit of work by using

$$A^T = (A^*A)^{-1}A^*$$

instead of (1.2), and this is what we often try to do. But if A^*A is (nearly) singular then there will be infinitely many vectors \mathbf{x} which (nearly) minimize $\| \mathbf{b} - A\mathbf{x} \|$ and the last formula will have to be modified in a way which takes A 's rank into account (cf. [4], [6], [7]). The methods considered in this paper simplify the problem of assigning a rank to A .

In the past the conventional way to determine the rank of A was to convert A to a row-echelon form, e.g.,

$$\begin{pmatrix} x & x & x & x & x & \cdot & \cdot & \cdot \\ 0 & x & x & x & x & \cdot & \cdot & \cdot \\ 0 & 0 & x & x & x & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & 0 & 0 & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & 0 & 0 & \cdot & \cdot & \cdot \end{pmatrix}, \quad (\text{rank} = 3),$$

in which x 's represent nonzero elements and 0 's represent zeros. The transformation was accomplished by premultiplying A by a succession either of elementary matrices (cf. [5]) or of unitary matrices (cf. [17]) designed to liquidate the subdiagonal elements of each column in turn. In order to obtain a simple picture like the one above it would have been necessary to perform column-interchanges to ensure that the largest possible numbers were being left on the diagonal (cf. "complete pivoting" as described by Wilkinson [33]). It is certainly possible to arrange that in the row-echelon form of A each row will have its largest element on the diagonal. Consequently the rank of A is just the number r of consecutive nonzero terms on the diagonal of its row-echelon form; all rows after the r th are zero. And Σ , correspondingly, should have just r nonzero singular values on its diagonal.

But in floating-point calculations it may not be so easy to decide whether some number is effectively zero or not. Rather, one will try to determine the rank r by observing whether all rows after the r th are negligible in comparison to the first r , with the expectation that the same will be true of the singular values. Even this criterion is hard to apply, as the following example shows:

$$\begin{pmatrix} 1 & -1 & -1 & -1 & -1 & -1 & \cdot & \cdot & \cdot \\ & 1 & -1 & -1 & -1 & -1 & \cdot & \cdot & \cdot \\ & & 1 & -1 & -1 & -1 & \cdot & \cdot & \cdot \\ & & & 1 & -1 & -1 & \cdot & \cdot & \cdot \\ & & & & 1 & -1 & \cdot & \cdot & \cdot \\ & & & & & 1 & \cdot & \cdot & \cdot \\ & & & & & & 1 & \cdot & \cdot \\ & & & & & & & \cdot & \cdot \\ & & & & & & & & \cdot \end{pmatrix}.$$

If this matrix, already in row-echelon form, has a sufficiently large number

of rows and columns, then, although it may not appear to the naked eye to be deficient in rank, it is violently ill-conditioned (it has a very tiny singular value), as can be seen by applying the matrix to the column vector whose elements are, in turn,

$$1, 2^{-1}, 2^{-2}, 2^{-3}, \dots, 2^{-n}, \dots.$$

On the other hand, when all the -1 's in the matrix are replaced by $+1$'s then the resulting matrix is quite docile. Therefore, it would be very hard to tell, by looking at only the diagonal elements of the row-echelon form, whether or not the original matrix A had a singular value sufficiently small to be deleted during the calculation of A^T . In other words, without looking explicitly at the singular values there seems to be no satisfactory way to assign a rank to A .

The singular values of a matrix A are the nonnegative square roots of the eigenvalues of A^*A or AA^* , whichever has fewer rows and columns (see [1]). But the calculation of A^*A using ordinary floating point arithmetic does serious violence to the smaller singular values as well as to the corresponding eigenvectors which appear in U and V in (1.1). A discussion of these points can be found in a paper by Osborne [24], which also contains a nice proof of the existence of the decomposition (1.1). Since the columns of U are the eigenvectors of AA^* and the columns of V are the eigenvectors of A^*A , there is some possibility that a simple calculation of the decomposition (1.1) could be accomplished by using double-precision arithmetic to deal with A^*A and AA^* directly in some way. Such a scheme would be convenient with a machine like the IBM 7094 which has double-precision hardware. But for most other machines, and especially when a programming language deficient in double-precision facilities is used, the complicated scheme described in this paper seems to be the best we have.

Kogbetliantz [18], Hestenes [16], and Forsythe and Henrici [9] have proposed rotational or Jacobi-type methods for obtaining the decomposition (1.1). Kublanovskaja [19] has suggested a QR -type method. These methods are accurate but are slow in terms of total number of operations.

Our scheme is based upon an idea exploited by Lanczos [20]; the matrix

$$\tilde{A} = \begin{pmatrix} 0 & A \\ A^* & 0 \end{pmatrix}$$

has for its eigenvalues the singular values of A , each appearing with both a positive and a negative sign. The representation \tilde{A} could not be treated directly by a standard eigenvalue-vector program without dealing with the problems which we shall discuss in detail in what follows.

2. A matrix decomposition. In order to facilitate the computation of the singular values and the pseudo-inverse of the complex $m \times n$ matrix A , we

describe a convenient matrix decomposition. We assume throughout our discussion that $m \geq n$ without any loss of generality.

THEOREM 1. *Let A be any $m \times n$ matrix with complex elements. Then A can be decomposed as*

$$A = PJQ^*$$

where P and Q are unitary matrices and J is an $m \times n$ bidiagonal matrix of the form

$$J = \left[\begin{array}{cccccccc} \alpha_1 & \beta_1 & 0 & \cdot & \cdot & \cdot & 0 & \\ & \alpha_2 & \beta_2 & 0 & \cdot & \cdot & \cdot & \\ & \mathbf{0} & \cdot & \cdot & \cdot & \cdot & \cdot & \\ & & & \cdot & \cdot & \cdot & \cdot & \\ & & & & \cdot & \cdot & \cdot & \\ & & & & & \cdot & \beta_{n-1} & \\ & & & & & & \alpha_n & \\ & & & & & & & \mathbf{0} \end{array} \right] \left. \vphantom{\begin{array}{c} \alpha_1 \\ \alpha_2 \\ \mathbf{0} \\ \cdot \\ \cdot \\ \cdot \\ \alpha_n \\ \mathbf{0} \end{array}} \right\} (m - n) \times n$$

Proof. The proof will be a constructive one in which Householder transformations (see [17], [21], [32]) are used. Let $A = A^{(1)}$ and let $A^{(3/2)}, A^{(2)}, \dots, A^{(n)}, A^{(n+1/2)}$ be defined as follows:

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

$$A^{(k+1)} = A^{(k+1/2)} Q^{(k)}, \quad k = 1, 2, \dots, n - 1.$$

$P^{(k)}$ and $Q^{(k)}$ are hermitian, unitary matrices of the form

$$P^{(k)} = I - 2\mathbf{x}^{(k)} \mathbf{x}^{(k)*}, \quad \mathbf{x}^{(k)*} \mathbf{x}^{(k)} = 1,$$

$$Q^{(k)} = I - 2\mathbf{y}^{(k)} \mathbf{y}^{(k)*}, \quad \mathbf{y}^{(k)*} \mathbf{y}^{(k)} = 1.$$

The unitary transformation $P^{(k)}$ is determined so that

$$a_{i,k}^{(k+1/2)} = 0, \quad i = k + 1, \dots, m,$$

and $Q^{(k)}$ is determined so that

$$a_{k,j}^{(k+1)} = 0, \quad j = k + 2, \dots, n,$$

and $A^{(k+1)}$ has the form

$$A^{(k+1)} = \left[\begin{array}{cccccccc} \alpha_1 & \beta_1 & 0 & \cdot & \cdot & & & \\ \mathbf{0} & \alpha_2 & \beta_2 & 0 & \cdot & & & \\ \cdot & \mathbf{0} & \cdot & \cdot & \cdot & & & \\ \cdot & \cdot & \cdot & \cdot & \cdot & & & \\ & & & & & & \mathbf{0} & \\ & & & & & & & \\ & & & & & \alpha_k & \beta_k & \\ & & & & & \mathbf{x} & \mathbf{x} & \cdot \cdot \cdot \\ & & & & & \mathbf{0} & \mathbf{0} & \cdot \cdot \cdot \\ & & & & & \cdot & \cdot & \cdot \cdot \cdot \\ & & & & & \cdot & \cdot & \cdot \cdot \cdot \\ & & & & & \cdot & \cdot & \cdot \cdot \cdot \end{array} \right] .$$

We illustrate the derivation of the formula for $P^{(k)}$. In order not to disturb those elements which have already been annihilated we set

$$x_i^{(k)} = 0, \quad i = 1, 2, \dots, k - 1.$$

Since $P^{(k)}$ is a unitary transformation, length is preserved and consequently

$$(2.1) \quad |\alpha_k|^2 = \sum_{i=k}^m |a_{i,k}^{(k)}|^2.$$

Also, since $P^{(k)}$ is hermitian,

$$P^{(k)} A^{(k+1/2)} = A^{(k)},$$

so that

$$\begin{aligned} (1 - 2|x_k^{(k)}|^2)\alpha_k &= a_{k,k}^{(k)}, \\ -2x_i^{(k)}\bar{x}_k^{(k)}\alpha_k &= a_{i,k}^{(k)}, \quad i = k + 1, \dots, m, \end{aligned}$$

and hence

$$(2.2) \quad |x_k^{(k)}|^2 = \frac{1}{2} \left(1 - \frac{a_{k,k}^{(k)}}{\alpha_k} \right),$$

$$(2.3) \quad x_i^{(k)} = \frac{-a_{i,k}^{(k)}}{2\alpha_k\bar{x}_k^{(k)}}.$$

Equations (2.1), (2.2), and (2.3) define two possible vectors $\mathbf{x}^{(k)}$ to within scalar factors of modulus one. In the interest of numerical stability, let us choose $\text{sgn } \alpha_k$ so that $x_k^{(k)}$ is as large as possible. Thus

$$\alpha_k = -\frac{a_{k,k}^{(k)}}{|a_{k,k}^{(k)}|} \left(\sum_{i=k}^m |a_{i,k}^{(k)}|^2 \right)^{1/2}.$$

Summarizing, we have

$$A^{(k+1/2)} = A^{(k)} - \mathbf{x}^{(k)} \cdot 2(\mathbf{x}^{(k)*} A^{(k)}),$$

with

$$\begin{aligned} s_k &= \left(\sum_{i=k}^m |a_{i,k}^{(k)}|^2 \right)^{1/2}, \\ \alpha_k &= -s_k \left(\frac{a_{k,k}^{(k)}}{|a_{k,k}^{(k)}|} \right), \\ x_i^{(k)} &= 0 \quad \text{for } i < k, \\ x_k^{(k)} &= \left[\frac{1}{2} \left(1 + \frac{|a_{k,k}^{(k)}|}{s_k} \right) \right]^{1/2}, \quad (\text{say}), \\ c_k &= \left(2s_k \frac{a_{k,k}^{(k)}}{|a_{k,k}^{(k)}|} x_k^{(k)} \right)^{-1}, \end{aligned}$$

and

$$x_i^{(k)} = c_k a_{i,k}^{(k)} \quad \text{for } i > k.$$

(If $s_k = 0$, just set $\alpha_k = 0$ and $\mathbf{x}^{(k)} = \mathbf{0}$.) Similarly,

$$A^{(k+1)} = A^{(k+1/2)} - 2(A^{(k+1/2)} \mathbf{y}^{(k)}) \cdot \mathbf{y}^{(k)*},$$

with

$$\begin{aligned} t_k &= \left(\sum_{j=k+1}^n |a_{k,j}^{(k+1/2)}|^2 \right)^{1/2}, \\ \beta_k &= -t_k \cdot \frac{a_{k,k+1}^{(k+1/2)}}{|a_{k,k+1}^{(k+1/2)}|}, \\ y_j^{(k)} &= 0 \quad \text{for } j \leq k, \\ y_{k+1}^{(k)} &= \left[\frac{1}{2} \left(1 + \frac{|a_{k,k+1}^{(k+1/2)}|}{t_k} \right) \right]^{1/2}, \quad (\text{say}), \\ d_k &= \left(2t_k \frac{a_{k,k+1}^{(k+1/2)}}{|a_{k,k+1}^{(k+1/2)}|} y_{k+1}^{(k)} \right)^{-1}, \end{aligned}$$

and

$$y_j^{(k)} = d_k \bar{a}_{k,j} \quad \text{for } j > k + 1.$$

An alternative approach to bidiagonalizing A is to generate the columns of P and Q sequentially as is done by the Lanczos algorithm for tridiagonalizing a symmetric matrix. The equations

$$AQ = PJ \quad \text{and} \quad P^*A = JQ^*$$

can be expanded in terms of the columns \mathbf{p}_i of P and \mathbf{q}_i of Q to yield

$$\left. \begin{aligned} A\mathbf{q}_1 &= \alpha_1 \mathbf{p}_1, \\ A\mathbf{q}_i &= \beta_{i-1} \mathbf{p}_{i-1} + \alpha_i \mathbf{p}_i, \\ \mathbf{p}_{i-1}^* A &= \alpha_{i-1} \mathbf{q}_{i-1}^* + \beta_{i-1} \mathbf{q}_i^*, \\ \mathbf{p}_n^* A &= \alpha_n \mathbf{q}_n^*. \end{aligned} \right\} \quad i = 2, 3, \dots, n,$$

These lead to the following algorithm.

Choose \mathbf{q}_1 arbitrarily with $\|\mathbf{q}_1\| = 1$; then set $\mathbf{w}_1 = A\mathbf{q}_1$;

$$(2.4) \quad \begin{aligned} \alpha_1 &= \|\mathbf{w}_1\|, \mathbf{p}_1 = (\alpha_1)^{-1} \mathbf{w}_1. \text{ Set } \mathbf{z}_i^* = \mathbf{p}_i^* A - \alpha_i \mathbf{q}_i^*, \beta_i = \|\mathbf{z}_i\|, \\ \mathbf{q}_{i+1}^* &= (\beta_i)^{-1} \mathbf{z}_i^* \text{ for } i = 1, 2, \dots, n-1; \text{ set } \mathbf{w}_i = A\mathbf{q}_i - \beta_{i-1} \mathbf{p}_{i-1}, \\ \alpha_i &= \|\mathbf{w}_i\|, \mathbf{p}_i = (\alpha_i)^{-1} \mathbf{w}_i \text{ for } i = 2, \dots, n. \end{aligned}$$

Gaussian elimination to $K - \lambda I$ *without* pivotal interchanges; there will be no trouble here (cf. [33, pp. 285–286]) provided floating point calculation is used and provided λ , if not exactly right, is larger than K 's largest or smaller than K 's smallest eigenvalue by perhaps a unit or so in λ 's last place. The point here is that each nonzero pivot u_i in the elimination process must be of the same sign as $(K - \lambda I)$'s diagonal elements. The result of the elimination process is to express $K - \lambda I = LU$, where

$$L = \begin{pmatrix} 1 & & & & \\ l_1 & 1 & & & \\ & l_2 & 1 & & \\ & & l_3 & \cdot & \\ \mathbf{0} & & & \cdot & \\ & & & & \cdot \\ & & & & l_{n-1} & 1 \end{pmatrix}$$

and

$$U = \begin{pmatrix} u_1 & b_1 & & & \\ & u_2 & b_2 & & \\ & & u_3 & \cdot & \\ \mathbf{0} & & & \cdot & \\ & & & & b_{n-1} \\ & & & & u_n \end{pmatrix}.$$

Here $u_1 = a_1 - \lambda$ and $l_i = b_i/u_i$, $u_{i+1} = a_{i+1} - \lambda - l_i b_i$ for $i = 1, 2, \dots, n - 1$. Next we attempt the solution of $(K - \lambda I)\mathbf{v} = \mathbf{r}$ using for \mathbf{r} a vector whose elements all have the same magnitude but signs chosen to maximize the elements of \mathbf{v} . The choice of sign is accomplished by first solving $L\mathbf{s} = \mathbf{r}$ as follows:

$$s_1 = +1, \\ s_{i+1} = (-l_i s_i) + \text{sgn}(-l_i s_i), \quad i = 1, 2, \dots, n - 1.$$

The solution of $U\mathbf{v} = \mathbf{s}$ for \mathbf{v} and the subsequent normalization of \mathbf{v} complete the calculation. Provided no two pivots u_i have opposite signs one can show that the elements of \mathbf{v} each have the same signs as the corresponding elements of the desired eigenvector despite the rounding errors committed during \mathbf{v} 's calculation. Furthermore, the elements of \mathbf{r} exhibit the same signs as those of $+\mathbf{v}$ or $-\mathbf{v}$, depending upon the sign of the u_i 's. Consequently the cosine of the angle between \mathbf{r} and the correct eigenvector is at least $N^{-1/2}$ in magnitude, and finally we conclude that $K\mathbf{v}$ must differ from $\lambda\mathbf{v}$ by no more than a few units in the last place (cf. the argument in [30]). Now even if \mathbf{v} is contaminated by components of the eigenvectors corresponding to other

To continue the deflation we must so determine P_{j+1} that its application will simultaneously annihilate the spurious element w_j in the j th row and column of the matrix as well as the vector's $(j + 1)$ th element ϕ_{j+1} . But in practice the accumulation of rounding errors will prevent the exact annihilation of both elements; instead we shall have to be satisfied with a P_{j+1} which leaves negligible residuals in place of w_j and ϕ_{j+1} . Wilkinson, having scaled $K - \lambda I$ so that its largest element lies between $\frac{1}{2}$ and 2, would use whichever of the equations

$$w_j c_{j+1} = h_j s_{j+1}, \quad \phi_{j+1} c_{j+1} = -v_{j+2} s_{j+1},$$

contained the largest coefficient $|w_j|$, $|h_j|$, $|\phi_{j+1}|$, or $|v_{j+2}|$ to determine, in conjunction with $c_{j+1}^2 + s_{j+1}^2 = 1$, the values c_{j+1} and s_{j+1} . This method seems to be effective and we believe that it should always work, but since we cannot prove the method's infallibility, our work is incomplete.

Now we can show how to construct a deflation process for the bidiagonal matrix J . The first step is to obtain J 's largest singular value σ ; σ^2 is the largest eigenvalue of the tridiagonal matrix $J^t J$ (see §3). The next step requires the corresponding vectors \mathbf{x} and \mathbf{y} which can be obtained either by solving $J^t J \mathbf{y} = \sigma^2 \mathbf{y}$ for \mathbf{y} and setting $\mathbf{x} = \sigma^{-1} J \mathbf{y}$, or by calculating σ 's eigenvector \mathbf{z} of S in (3.3) and hence obtaining \mathbf{x} and \mathbf{y} from \mathbf{z} 's even and odd components respectively. Both methods for getting \mathbf{x} and \mathbf{y} are numerically stable when performed in floating point. The deflation of J is accomplished by a sequence of 2×2 rotations applied in succession to its first and second columns, its first and second rows, its second and third columns, its second and third rows, its third and fourth columns, \dots , its $(n - 1)$ th and n th rows. The i th rotation applied to rows i and $i + 1$ of J must simultaneously annihilate a spurious subdiagonal element, introduced into row $i + 1$ by the previous column rotation, and the i th element in the current \mathbf{x} -vector. The i th column rotation, except for the first, must annihilate a spurious term introduced by the previous row rotation into the $(i + 1)$ th column just above the first superdiagonal, and simultaneously the transpose of the i th column rotation must liquidate the i th element of the current \mathbf{y} -vector. The first column rotation would when applied to $J^t J - \sigma^2 I$ annihilate the element in its first row and second column. At the end of the deflation process J 's element b_{n-1} should have been replaced by zero. Of course, rounding errors will prevent the rotations from performing their roles exactly upon both the matrix J and the vectors \mathbf{x} and \mathbf{y} , but just as in the deflation of a tridiagonal matrix we are able so to determine the rotations that negligible residuals are left behind in place of the elements we wished to liquidate.

After deflating J we delete its last row and column and repeat the process until J is deflated to a 1×1 matrix or the deflated J becomes negligibly small. At the end we multiply the rotations in reverse order to construct the

matrices X and Y which put J into the form (3.1):

$$J = X\Sigma Y.$$

(If J was complex, a unitary diagonal transformation should be incorporated here.) Finally the matrices P and Q of Theorem 1 are multiplied thus:

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

to exhibit the decomposition (1.1):

$$A = U\Sigma V.$$

The two matrix multiplications PX and QY take most of the work.

5. Applications. The basic decomposition given by (1.1) has many applications in data analysis and applied mathematics. Suppose the matrix A arises from statistical observation, and we wish to replace A by another matrix \hat{A} (say) which has lower rank p and is the best approximation to A in some sense. If we use the Frobenius norm (i.e., $\|A\|^2 = \text{trace } A^*A$) then the problem has been solved [8] as follows.

THEOREM 2. *Let A be an $m \times n$ matrix of rank r which has complex elements. Let S_p be the set of all $m \times n$ matrices of rank $p < r$. Then for all $B \in S_p$,*

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

where

$$\hat{A} = U\hat{\Sigma}V^*$$

and $\hat{\Sigma}$ is obtained from the Σ of (1.1) by setting to zero all but its p largest singular values σ_i .

Proof. Since $A = U\Sigma V^*$ and the Frobenius norm is unitarily invariant,

$$\|A - B\| = \|\Sigma - U^*BV\|.$$

Let $U^*BV = C$. Then

$$\|\Sigma - C\|^2 = \sum_{i=1}^n |\sigma_i - c_{ii}|^2 + \sum_{i \neq j} |c_{ij}|^2 \geq \sum_{i=1}^n |\sigma_i - c_{ii}|^2.$$

Now it is convenient to order the singular values in such a way that $\sigma_i \geq \sigma_{i+1}$. Thus, $\|A - B\|^2$ is minimized if $c_{ii} = \sigma_i$ for $i = 1, 2, \dots, p$, and $c_{ij} = 0$ otherwise, i.e., for $C = \hat{\Sigma}$. Obviously,

$$\|A - \hat{A}\| = (\sigma_{p+1}^2 + \dots + \sigma_r^2)^{1/2}.$$

Finding the vector \mathbf{x} of shortest length which minimizes $\|\mathbf{b} - A\mathbf{x}\|$ is equivalent to finding the vector \mathbf{y} of shortest length which minimizes

$\| \mathbf{c} - J\mathbf{y} \|$, where $\mathbf{c} = P^*\mathbf{b}$ and $\mathbf{y} = Q^*\mathbf{x}$. Here a natural question arises: is there any method which bypasses the complicated scheme in §3 and §4 for exhibiting J 's singular values explicitly, and instead takes advantage of J 's simple bidiagonal form to solve the least squares problem or to calculate J^T ? Such a method, if it exists, must retain provision for intentional perturbations designed to delete, in effect, negligible singular values without inducing too large a discrepancy in J or A . Unfortunately, J 's simple form is deceptive; even J 's rank is hard to estimate without further calculation. For example, if J 's rank r is less than n , then at least $n - r$ of the α_i 's, and possibly more, should vanish; but in practice none of the α_i 's may be negligible even though several may be very small compared with adjacent β_i 's and, in consequence, a few of J 's singular values may be negligible.

Perhaps the recurrence described by Greville [15] can be modified by the introduction of pivoting and then applied to J to calculate J^T . Until this scheme is worked out, the best method we can suggest for solving the least squares problem together with controllable perturbations is the following. Compute explicitly the representation

$$A = U\Sigma V^*,$$

decide which of the singular values are small enough to ignore, replace the remaining singular values by their reciprocals to obtain Σ^T , and finally use

$$A^T = V\Sigma^T U^*$$

to obtain the least squares solution $\mathbf{x} = A^T\mathbf{b}$. Once again, to ignore some singular values $\sigma_{r+1}, \sigma_{r+2}, \dots, \sigma_n$ is equivalent to perturbing A by a matrix whose norm is $(\sum_{i=r+1}^n \sigma_i^2)^{1/2}$.

In some scientific calculations it is preferable that a given square matrix A be perturbed as little as possible (just rounding errors), but instead a perturbation $\delta\mathbf{b}$ in the right-hand side \mathbf{b} of the equation $A\mathbf{x} = \mathbf{b}$ is permissible provided $\| \delta\mathbf{b} \|$ does not exceed a given tolerance ϵ . The substitution

$$\mathbf{y} = V^*\mathbf{x}, \quad \mathbf{c} = U^*\mathbf{b}, \quad \delta\mathbf{c} = U^*\delta\mathbf{b},$$

transforms the perturbed equation $A\mathbf{x} = \mathbf{b} + \delta\mathbf{b}$ into an equivalent diagonal system

$$\Sigma\mathbf{y} = \mathbf{c} + \delta\mathbf{c},$$

in which the permissible perturbation $\delta\mathbf{c}$ still satisfies

$$(5.1) \quad \| \delta\mathbf{c} \| < \epsilon.$$

Subject to this constraint, $\delta\mathbf{c}$ may be chosen to optimize some other criterion. For example, suppose we wish to minimize $\| \mathbf{x} \| = \| \mathbf{y} \|$. Then ideally $\delta\mathbf{c}$

should satisfy $\Sigma^2 \delta \mathbf{c} + \lambda(\mathbf{c} + \delta \mathbf{c}) = \mathbf{0}$ with some suitable positive value of the Lagrange multiplier λ sufficiently small so that (5.1) is satisfied too. But for most practical purposes it is sufficient to use trial and error to determine λ to within a factor of two so that $\delta \mathbf{c} = -(I + \lambda^{-1} \Sigma^2)^{-1} \mathbf{c}$ will satisfy $\delta \mathbf{c}^* \delta \mathbf{c} < \epsilon^2$. The use of such a technique in least squares problems tends to suppress violent oscillation and cancellation which might otherwise detract from the usefulness of the solution \mathbf{x} .

A similar technique is valuable for the solution of the sets of linear equations which approximate integral equations of the form

$$\int A(i, j)x(j) dj = b(i).$$

Here the numerical treatment of the integral equation, in using singular values, is similar to the theoretical treatment found in [29]. Once again, the use of the decomposition $A = U\Sigma V^*$ aids the suppression of spurious oscillations in the function x .

We close with a warning; diagonal transformations can change A 's singular values and A^T in a nontrivial way. Therefore some sort of equilibration may be necessary to allow each row and column of A to communicate its proper significance to the calculation. Two useful forms of equilibration are:

(i) scale each row and column of A in such a way that all the rows have roughly the same norm and so have all the columns;

(ii) scale each row and column of A in such a way that the absolute uncertainty in each element of A does not vary much from element to element. On least squares problems such equilibration is accomplished by weighting each residual in the sum of squares (see [2], [10], [11], [23] on equilibration algorithms, and [14]).

REFERENCES

- [1] A. R. AMIR-MOÉZ AND A. L. FASS, *Elements of Linear Spaces*, Pergamon, New York, 1962, Chap. 12.
- [2] F. L. BAUER, *Optimally scaled matrices*, Numer. Math., 5 (1963), pp. 73-87.
- [3] A. BEN-ISRAEL AND A. CHARNES, *Contributions to the theory of generalized inverses*, J. Soc. Indust. Appl. Math., 11 (1963), pp. 667-699.
- [4] A. BEN-ISRAEL AND S. J. WERSAN, *An elimination method for computing the generalized inverse of an arbitrary complex matrix*, J. Assoc. Comput. Mach., 10 (1963), pp. 532-537.
- [5] G. BIRKHOFF AND S. MACLANE, *A Survey of Modern Algebra*, Macmillan, New York, 1953, Chap. VII, §6.
- [6] J. W. BLATTNER, *Bordered matrices*, J. Soc. Indust. Appl. Math., 10 (1962), pp. 528-536.
- [7] J. C. G. BOOT, *The computation of the generalized inverse of singular or rectangular matrices*, Amer. Math. Monthly, 70 (1963), pp. 302-303.

- [8] C. ECKART AND G. YOUNG, *The approximation of one matrix by another of lower rank*, *Psychometrika*, 1 (1936), pp. 211–218.
- [9] G. E. FORSYTHE AND P. HENRICI, *The cyclic Jacobi method for computing the principal values of a complex matrix*, *Trans. Amer. Math. Soc.*, 94 (1960), pp. 1–23.
- [10] G. E. FORSYTHE AND E. G. STRAUS, *On best conditioned matrices*, *Proc. Amer. Math. Soc.*, 6 (1955), pp. 340–345.
- [11] D. R. FULKERSON AND P. WOLFE, *An algorithm for scaling matrices*, *SIAM Rev.*, 4 (1962), pp. 142–146.
- [12] W. GIVENS, *Numerical computation of the characteristic values of a real symmetric matrix*. Oak Ridge National Laboratory, Report No. 1574, 1954.
- [13] H. H. GOLDSTINE, F. J. MURRAY, AND J. VON NEUMANN, *The Jacobi method for real symmetric matrices*, *J. Assoc. Comput. Mach.*, 6 (1959), pp. 59–96.
- [14] G. H. GOLUB, *Comparison of the variance of minimum variance and weighted least squares regression coefficients*, *Ann. Math. Statist.*, 34 (1963), pp. 984–991.
- [15] T. N. E. GREVILLE, *Some applications of the pseudo-inverse of a matrix*, *SIAM Rev.*, 2 (1960), pp. 15–22.
- [16] M. R. HESTENES, *Inversion of matrices by biorthogonalization and related results*, *J. Soc. Indust. Appl. Math.*, 6 (1958), pp. 51–90.
- [17] A. S. HOUSEHOLDER, *Unitary triangularization of a nonsymmetric matrix*, *J. Assoc. Comput. Mach.*, 5 (1958), pp. 339–342.
- [18] E. G. KOGBETLIANTZ, *Solution of linear equations by diagonalization of coefficients matrix*, *Quart. Appl. Math.*, 13 (1955), pp. 123–132.
- [19] V. N. KUBLANOVSKAJA, *Some algorithms for the solution of the complete problem of eigenvalues*, *V. Vyčisl. Mat.i. Mat. Fiz.*, 1 (1961), pp. 555–570.
- [20] C. LANCZOS, *Linear Differential Operators*, Van Nostrand, London, 1961, Chap. 3.
- [21] D. D. MORRISON, *Remarks on the unitary triangularization of a nonsymmetric matrix*, *J. Assoc. Comput. Mach.*, 7 (1960), pp. 185–186.
- [22] J. M. ORTEGA AND H. F. KAISER, *The LL^t and QR methods for symmetric tri-diagonal matrices*, *Comput. J.*, 6 (1963), pp. 99–101.
- [23] E. E. OSBORNE, *On pre-conditioning of matrices*, *J. Assoc. Comput. Mach.*, 7 (1960), pp. 338–345.
- [24] ———, *On least squares solutions of linear equations*, *Ibid.*, 8 (1961), pp. 628–636.
- [25] R. PENROSE, *A generalized inverse for matrices*, *Proc. Cambridge Philos. Soc.*, 51 (1955), pp. 406–413.
- [26] ———, *On best approximate solutions of linear matrix equations*, *Ibid.*, 52 (1956), pp. 17–19.
- [27] H. RUTISHAUSER, *Deflation bei Bandmatrizen*, *Z. Angew. Math. Phys.*, 10 (1959), pp. 314–319.
- [28] ———, *On Jacobi rotation patterns*, *Proc. Symp. Appl. Math. XV, Experimental Arithmetic, High Speed Computing, and Mathematics*, *Amer. Math. Soc.*, 1963, pp. 219–240.
- [29] F. SMITHIES, *Integral Equations*, Cambridge University Press, Cambridge, 1958, Chap. VIII.
- [30] J. H. WILKINSON, *The calculation of the eigenvectors of codiagonal matrices*, *Comput. J.*, 1 (1958), pp. 148–152.
- [31] ———, *Stability of the reduction of a matrix to almost triangular and triangular forms by elementary similarity transformations*, *J. Assoc. Comput. Mach.*, 6 (1959), pp. 336–359.

- [32] ———, *Householder's method for the solution of the algebraic eigenproblem*, *Comput. J.*, 3 (1960), pp. 23–27.
- [33] ———, *Error analysis of direct methods of matrix inversion*, *J. Assoc. Comput. Mach.*, 8 (1961), pp. 281–330.
- [34] ———, *Error analysis of eigenvalue techniques based on orthogonal transformations*, *J. Soc. Indust. Appl. Math.*, 10 (1962), pp. 162–195.
- [35] ———, *Calculation of the eigenvalues of a symmetric tridiagonal matrix by the method of bisection*, *Numer. Math.*, 4 (1962), pp. 362–367.