THE LIMITATIONS OF JACOBI METHODS FOR TRIANGULATING SQUARE MATRICES

by

Ying Wang

Department of Mathematics
University of California, Berkeley

October 1975

Abstract

Jacobi methods for symmetric matrices have prompted the search for generalizations to reduce any complex square matrix to upper triangular form using unitary similarity transformations. All attempts have been unsuccessful. It is not the case that all previous investigators have failed to find the right algorithm. It is shown here that there are intrinsic limitations on Jacobi-type procedures.

ACKNOWLEDGMENTS

I am most grateful to my thesis adviser, Professor B.N. Parlett, for his patient guidance. I wish to thank my husband, Yung, for his understanding and encouragement.

This dissertation is dedicated to my parents.

TABLE OF CONTENTS

CHAPTER I

Introduction

In 1846 Jacobi (16), in proving that the eigenvalues of a symmetric matrix are real, gave a constructive procedure for finding them. The method was little appreciated until its rediscovery in 1949 by Goldstine, von Neumann and Murray. Since then an interesting literature has been developed for the modification and generalization of Jacobi's method. For the symmetric and Hermitian case, modifications for various purposes have been developed and are quite successful.

The generalizations to arbitrary matrices are not so satisfactory. The elegance of Jacobi processes lies in their simplicity. The challenge is to produce a Jacobi type algorithm which achieves rapid convergence without sacrificing too much simplicity. The thrust of this thesis is that no such variation can be found.

Greenstadt (11), Lotkin (19), and Huang (15) have produced Jacobi-type procedures for arbitrary matrices. Their limitations can be illustrated by the following example:

$$\begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 10^{-6} & 0 & 0 \end{pmatrix} . \qquad (1.1)$$

The above matrix has distinct and well-separated eigenvalues. Greenstadt's algorithm applied to (1.1) produces a sequence of matrices which cycles. The same matrix is invariant under Lotkin's procedure which reduces to the classical Jacobi method in the Hermitian case. Huang's method claims to produce a monotonically decreasing sequence of the sum of squares of the lower triangular elements. At the first step, the

reduction in this sum is a negligible $10^{-24}/4$, and there is no subsequent improvement. Example (1.1) is merely the simplest of an infinite class of matrices on which all known Jacobi variants fail.

We argue here that there are intrinsic limitations in the Jacobi method; it is not the case that previous investigators have missed the right algorithm.

CHAPTER II

Definition of Jacobi Methods

II.1 Introductory Remark

In mid-twentieth century, von Neumann and Goldstine rediscovered Jacobi's method (1846) for diagonalizing a real symmetric matrix. Active research has been conducted in the modification and generalization of Jacobi method since then. The convergence of Jacobi-type procedures depends strongly on the way in which they are defined. Too narrow a definition, then convergence is not possible as indicated in Section II.2. Too broad a definition, then the basic QR algorithm can be included (Section II.3).

The trouble with including the basic QR algorithm is that it is unacceptably slow compared with the shifted QR algorithm which is the current champion method. Parlett has provided a global convergence theorem (21) for the basic QR algorithm, whereas there is no guarantee of convergence with the usual shift strategies which makes the method so fast. However, no cases are known when the shifted QR fails. The quality in which Jacobi methods surpass other techniques was their simplicity. The essential problem is whether some Jacobi algorithm can be found which achieves convergence without sacrificing too much simplicity.

II.2 Criteria of Jacobi Methods

Jacobi methods produce a sequence of similar matrices, and plane rotations are used as matrices of transformations. At each step, the parameters of the rotation, the plane (p,q) and angle θ, are chosen

according to some rule. In classical Jacobi methods, the plane is determined by the maximum off-diagonal element and the angle θ is chosen to annihilate that element. For cyclic Jacobi methods, the plane (p,q) is chosen according to some predetermined order and the angle is selected to maximize the reduction of the quantity, $N^2(\cdot)$, which in this case is the annihilation of the (p,q) element. Jacobi methods for symmetric matrices have several properties in common. Once the plane (p,q) is selected, the choice for the angle of rotation satisfies the following criteria.

Criterion 1: *The* (p,q) *element of the matrix is annihilated.*

Criterion 2: *The quantity* $N^2(\cdot)$ *is minimized over* θ, *where* $N^2(A) = \sum_{i \neq j} |a_{ij}|^2$ *for* A *symmetric and* $N^2(A) = \sum_{i>j} |a_{ij}|^2$ *for* A *non-symmetric.*

For non-symmetric matrices, these criteria cannot be met simultaneously (see below). Greenstadt's method employed Criterion 1. The quantity, $N^2(\cdot)$, is not only not minimized; it very often increases. There is a class of matrices which the method fails to converge (Section III.4.2). Goldstine-Horwitz and Lotkin produced procedures which employ Criterion 2. Due to algebraic complexity, the minimum has no closed form. Hence approximations are used. These approximations leave certain matrices invariant under the algorithm. From the point of view of convergence proofs, it would be sufficient to reduce $N^2(\cdot)$ by a fixed proportion at each step. Local minimization, even if possible, is a luxury which can be given up. Huang's algorithm used the following relaxed form of Criterion 2.

Criterion 3: *The sequence* $\{N^2(A_k)\}$ *is monotonically decreasing.*

For general matrices, Criteria 1 and 3 cannot both be satisfied simultaneously. Consider a matrix of the following form:

$$\begin{pmatrix} 0 & a_{12} & a_{13} \\ 0 & 0 & a_{23} \\ a_{31} & 0 & 0 \end{pmatrix}.$$

If $|a_{31}|^2 < |a_{12}|^2 + |a_{23}|^2$, then a $(1,3)$ plane rotation designed to annihilate a_{31} will violate Criterion 3. Since Criterion 1 is incompatible with Criterion 3, we are tempted to broaden Criterion 1 in a useful way.

W. Givens in 1954 made a useful observation. The (p,q) element of a matrix can be annihilated by a rotation in the (p,j) or the (k,q) planes for any $k \neq p$ or $j \neq q$. Such rotations are called Givens rotations to distinguish them from those of Jacobi. This suggests the following relaxation of Criterion 1.

Criterion 4: *Given* (p,q), *choose* θ *so that the* (q,r) *element of the matrix is annihilated for some* $r \neq p$.

When trying to triangulate non-symmetric matrices by plane rotations, it seems perverse to exclude Givens rotations just because Jacobi did not need them for the symmetric case. However, as we shall see, this enlargement of the repertoire of rotations has far reaching consequences. Methods not usually associated with Jacobi are seen to be close cousins, if not within the immediate family of Jacobi methods.

II.3 QR Without Shift as a Jacobi Process

II.3.1 Formal Definition

The QR algorithm uses complicated similarity transformations which are based on the Gram-Schmidt factorization.

Theorem 2.1 (Gram-Schmidt): *A rectangular matrix* A, $m \times n$, $m \geq n$, *can always be written in the form* $A = QR$ *where* Q *is unitary, and* R *is upper triangular with non-negative diagonal. The factorization is unique if* $\text{rank}(A) = n$.

The basic (shiftless) algorithm produces a sequence of similar matrices. Each matrix in the sequence is determined by the Gram-Schmidt factorization of the previous matrix. The $k \times k$ term, A_k, is first decomposed to

$$A_k = Q_k R_k \ . \tag{2.1}$$

Then the next matrix in the sequence is formed by multiplying the two factors in the reverse order, i.e.

$$A_{k+1} = R_k Q_k \ . \tag{2.2}$$

A_{k+1} is clearly similar to A_k since $R_k = Q_k^{-1} A_k$. The above, (2.1) and (2.2), constitute the formal definition of the algorithm. Explicit QR factorization requires a large number of operations. Hence in practice A_{k+1} is obtained from A_k by very different means. It was invented by Francis in 1960 (6), and because of his clever use of the Hessenberg form and origin shifts it soon became the champion eigenvalue finder. However it is not a simple algorithm.

II.3.2 Hessenberg Form and the Uniqueness Theorem

Definition 2.2: *A matrix* A *is in upper Hessenberg form if* $a_{ij} = 0$ *whenever* $i > j+1$.

Hessenberg matrices play an important role in the practical QR algorithm.

Lemma 2.3: *The Hessenberg form is invariant under the* QR *algorithm.*

Proof: By the Gram-Schmidt process if A_k is in Hessenberg form then so is Q_k and also R_kQ_k, because R_k is upper triangular.

Q.E.D.

If the starting matrix A is brought into Hessenberg form, then all matrices in the QR sequence are Hessenberg. This reduces the number of operations greatly. To obtain the factors Q_k and R_k at each step of the procedure is a time consuming task. The following fact makes possible the computation of the QR sequence without computing the factors explicitly.

Theorem 2.4 (The Uniqueness Theorem): *Suppose that* $F^{-1}AF = H$ *where* H *is Hessenberg with positive subdiagonal. If* F^* *is unit triangular or* F *is unitary then* F *and* H *are uniquely determined by* A *and the first column of* F.

Proof: Suppose $AF = FH$. The j^{th} column is

$$A\bar{f}_j = \bar{f}_{j+1}h_{j+1,j} + \sum_{i=1}^{j} \bar{f}_i h_{ij}, \quad j = 1,\ldots,n-1$$

where $\bar{f}_i$ is the i^{th} column of F and $H = (h_{ij})$. Show that if columns $\bar{f}_i$, $i = 1,\ldots,j$ are known then $\bar{f}_{j+1}$ and $\bar{h}_j$ are determined.

Let $\bar{g}_k$ be either the k^{th} unit vector, $\bar{e}_k$, if F is unit triangular or $\bar{f}_k$ if F is unitary. Then, in either case $\bar{g}_k^* \bar{f}_j = \delta_{kj}$, $k \leq j$. Define

$$h_{kj} \equiv \bar{g}_k^* (A\bar{f}_j - \sum_{i=1}^{k} \bar{f}_i h_{ij}) , \quad k = 1,\ldots,j$$

$$\bar{d}_{j+1} \equiv A\bar{f}_j - \sum_{i=1}^{j} \bar{f}_i h_{ij} .$$

Then

$$h_{j+1,j} \equiv \bar{e}_{j+1}^* \bar{d}_{j+1} \quad \text{if } F \text{ is unit triangular}$$
$$\equiv \|\bar{d}_{j+1}\|_2 \quad \text{if } F \text{ is unitary}$$

and $$\bar{f}_{j+1} \equiv \bar{d}_{j+1}/h_{j+1,j} .$$ Q.E.D.

II.3.3 The Basic QR algorithm as a Jacobi Method

The formal definition of the QR procedure is best suited for the study of its properties. In this section, the basic QR algorithm on Hessenberg matrices will be presented in a way which can be regarded as a Jacobi method in the sense of using nothing but a sequence of plane rotations.

Lemma 2.5. *A matrix* $A = (a_{ij})$ *can be brought to Hessenberg form by a sequence of Givens rotations.*

Phase I

To effect this reduction the rotation planes may be chosen cyclically in the order

$$(2,3),(2,4),\ldots,(2,n)$$
$$(3,4),\ldots,(3,n)$$
$$\ldots$$
$$(n-1,n)$$

However this cycle only has to be traversed once. The angle θ for the rotation in plane (j,k), $j < k$, is given by

$$\tan \theta = |a_{k,j-1}/a_{j,j-1}| .$$

This choice annihilates element $(k,j-1)$. See (27, Chap. 6).

Definition 2.6. *A Hessenberg matrix* $H = (h_{ij})$ *is unreduced if* $h_{j+1,j} \neq 0$.

If the Hessenberg matrix is reduced, then it is of the following form:

$$\begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix} .$$

The eigenvalues of A are those of A_{11} and A_{22}. Therefore A_{11} and A_{22} can be transformed separately and so, without loss of generality, A can be taken to be unreduced. In fact, $a_{j+1,j}$, $j = 1,\ldots,n-1$, can be assumed to be positive since a unitary normalization can make all non-zero subdiagonal elements positive. This normalization is only needed for simplicity in stating results.

Phase II

Perform a sequence of rotations in the planes $(1,2),(2,3),\ldots,(n-1,n)$. The angle ϕ for the rotation in plane $(j,j+1)$ is given by

$$\begin{aligned}\tan\phi &= |a_{j+1,j-1}/a_{j,j-1}| \quad \text{for } j > 1 \\ &= |a_{j+1,j}/a_{j,j}| \quad \text{for } j = 1 .\end{aligned}$$

For $j > 1$ this annihilates the $(j+1,j-1)$ element introduced on the previous step.

Phase II is repeated indefinitely. We must now show that Phase II is equivalent to the basic QR algorithm on an unreduced Hessenberg matrix.

Theorem 2.7. *Let* H *be an* $n \times n$ *Hessenberg matrix with* $h_{j+1,j} > 0$ *for* $j = 1,2,\ldots,n$. *Then Phase II is equivalent to the basic QR algorithm.*

Proof: Denote the rotations used in Phase II in planes $(1,2),(2,3),\ldots,(n-1,n)$ by F_j, $j = 1,\ldots,n-1$. Define $F \equiv \prod_{k=1}^{n-1} F_k$. By the Uniqueness Theorem, it suffices to show that the first column of F and Q are identical where Q is the unitary factor of the Gram-Schmidt factorization of H. The first column of F is simply the first column of F_1, which is

$$\left(\frac{|a_{11}|}{(|a_{11}|^2+|a_{21}|^2)^{1/2}}, \frac{|a_{21}|}{(|a_{11}|^2+|a_{21}|^2)^{1/2}}, 0, \ldots, 0\right)^T .$$

This is just the first column of Q. Q.E.D.

Hence we have shown that the basic QR procedure can be considered to be of Jacobi type. Even though the procedure does not produce a monotonically decreasing $N^2(\cdot)$, its convergence properties have been established by Parlett (21). The necessary and sufficient conditions for convergence are quite complicated, but when the eigenvalues have distinct absolute value then the algorithm is guaranteed to converge linearly.

CHAPTER III

A Survey of the Jacobi-type Methods

III.1 Introductory Remarks

In 1846, Jacobi (16) gave a constructive proof of the fact that real symmetric matrices have real eigenvalues. It is one of the few efficient methods which existed before the early twentieth century. Since its rediscovery in the late forties, various extensions and modifications of the Jacobi method have been proposed. These Jacobi-like procedures have two central ideas. The matrices of transformation were generally chosen to either annihilate specific elements or to reduce the sum of squares of the absolute values of the off-diagonal. For the classical Jacobi method, these two ideas coincide. To avoid repetition, we shall define some terms which will be used throughout this chapter.

Definition 3.1. *Let* $A = (a_{ij})$ *be an* $n \times n$ *matrix. Define*

$$S^2(A) = \sum_{i \neq j} |a_{ij}|^2$$

$$N_L^2(A) = \sum_{i > j} |a_{ij}|^2$$

$$N_U^2(A) = \sum_{i < j} |a_{ij}|^2 .$$

In several Jacobi-type algorithms, the pivot-pairs are selected in some cyclic ordering.

Definition 3.2. *The order for selecting the pivot pairs is said to be cyclic by rows if the pivot pairs are selected as follows:*

$$(i_0, j_0) = (1,2)$$

$$(i_{k+1},j_{k+1}) = \begin{cases} (i_k,j_k+1) & \text{if } i_k < n-1, \ j_k < n \\ (i_k+1,i_k+2) & \text{if } i_k < n-1, \ j_k = n \\ (1,2) & \text{if } i_k = n-1, \ j_k = n \end{cases}.$$

Similarly for cyclic by columns.

Definition 3.3. *Let* A *be Hermitian. A Jacobi-like procedure is said to be convergent if*

$$\lim_{k\to\infty} A_k = \text{diag}(\lambda_i)$$

where the λ_i*'s are the eigenvalues of* A *in some order. Convergence of a normal matrix to a diagonal matrix is similarly defined.*

Definition 3.4. *A sequence of similar matrices* $\{A_k\}$ *is said to converge to normality if and only if*

$$\lim_{k\to\infty} \|A_k\|_E^2 = \sum_{j=1}^{n} |\lambda_j|^2$$

where λ_j*'s are the eigenvalues of* A.

III.2 Hermitian Matrices

The eigenproblem is simplest for Hermitian matrices (see Appendix). The Jacobi-type procedures for the computation of eigenvalues of this class of matrices use the complex rotations for the transformation. At each step, two elements in symmetrical positions are annihilated simultaneously thus causing a monotonic decrease of the quantity $S^2(A)$. The major difference between the various methods is in the selection of the pivot-pair (i_k,j_k) at each iteration.

III.2.1 The Classical Jacobi Method

In this process, a sequence of matrices A_k is constructed where $A_{k+1} = U_k^* A_k U_k$ and $A_0 = A$. Each $U_k = R(i_k, j_k, \theta_k, \alpha_k)$ is a rotation. If the matrix A is real symmetric, then U_k's are real plane rotations. The pivot-pair (i_k, j_k), is chosen to be the indices of a maximal off-diagonal element of A_k, that is,

$$|a^{(k)}_{i_k,j_k}| = \max_{i \neq j} |a^{(k)}_{ij}| \quad \text{where} \quad A_k = (a^{(k)}_{ij}) .$$

The other parameters θ_k and α_k are chosen to annihilate $a^{(k)}_{i_k,j_k}$ and $a^{(k)}_{j_k,i_k}$. Therefore, at each step, $S^2(A)$ is decreased by the amount $2|a^{(k)}_{i_k,j_k}|^2$.

Theorem 3.5. *The classical Jacobi method converges for real symmetric matrices if each angle* θ *satisfies* $|\theta| \leq \pi/4$.

(See (27), pp. 267.)

Each angle θ is chosen in the procedure to annihilate some element $a^{(k)}_{pq}$. The angle θ therefore satisfies the equation:

$$\tan 2\theta = 2a^{(k)}_{pq}/(a^{(k)}_{pp} - a^{(k)}_{qq}) .$$

So θ can always be taken to lie in the range $(-\pi/4, \pi/4)$.

Theorem 3.6. *For real symmetric matrices, the classical Jacobi method converges quadratically, that is,* $S^2(A_{(r+1)N}) < kS^2(A_{rN})$.

(See (27), pp. 267.)

The Jacobi algorithm is quite suitable for desk calculators. In

a digital computer, the search for the largest element off the diagonal is time consuming. The variations on the classical method have sacrificed maximal reduction in S^2 for faster pivot selection.

III.2.2 The Cyclic Jacobi Method

To simplify the selection of the pivot pair, it is convenient to choose it in some predetermined order. Probably the simplest scheme is to take them sequentially, either by rows or by columns. The convergence properties of this variation of Jacobi's method is quite difficult to prove. Forsythe and Henrici (5) have proved that if the angles of rotation are suitably restricted, then the algorithm converges. More precisely,

Theorem 3.7. *For real symmetric matrices, the cyclic Jacobi method converges if the angles of rotation* θ_k *satisfy* $\theta_k \in J$ *where* J *is a closed interval independent of* k *and interior to the open interval* $(-\pi/2,\pi/2)$.

(See (5).)

Theorem 3.8. *The cyclic Jacobi method ultimately converges quadratically.*

(See (14) and (18).) The disadvantage of this method is that much time may be spent in annihilating elements that are already quite small thus slowing the process considerably.

III.2.3 The Cyclic Jacobi Method with Threshold

To avoid annihilating small elements, a further modification is made. A sequence of monotonically decreasing threshold values is

introduced. The sequence can be finite or infinite. The procedure scans the off-diagonal elements according to some predetermined order, usually cyclic. Only those elements whose absolute values are larger than the current threshold are annihilated. If all off-diagonal elements are less than the current threshold value, then the next number in the threshold sequence is chosen to replace the current threshold. Clearly to assure convergence of the algorithm, the threshold sequence should converge to zero.

Theorem 3.9. *For real symmetric matrices, the cyclic Jacobi method with threshold converges for all monotonic threshold sequences converging to zero.*

(See (22).) All the theorems in this section are quoted for real symmetric matrices. For Hermitian matrices, the proofs in the references can be extended easily.

III.3 Normal Matrices

Normal matrices, like the Hermitian, are unitarily similar to diagonal matrices. However, a direct application of the classical Jacobi method to normal matrices leads to unsatisfactory results. As an example, consider permutation matrices of order n. These matrices are clearly normal and non-Hermitian. When the Jacobi algorithm is applied, the sequence of transformed matrices sometimes cycles, and the quantity $S^2(A)$ does not always decrease. We shall discuss two extensions of the Jacobi method to normal matrices in the following sections. One of the extensions was proposed by Goldstine and Horwitz. Their algorithm attempts to minimize the quantity $S^2(A)$ at each step.

Another simpler algorithm utilizes the fact that the Hermitian and the skew-Hermitian parts of a normal matrix commute (Theorem 0.7).

III.3.1 Goldstine-Horwitz Method

Goldstine and Horwitz (10) used complex rotations $R(i,j,\theta,\alpha)$ to diagonalize normal matrices. The algorithm is locally optimal in the sense that at each iteration the reduction of the quantity $S^2(A)$ is maximal. The parameters i, j, θ, α were chosen specifically to achieve this end. The selection procedure is quite complicated, and we shall not go into the details here. The interested reader can read the original article (10). Goldstine and Horwitz have supplemented their algorithm with a special treatment for matrices like the permutation matrices. With this modification, the convergence of the method can be proved.

Theorem 3.10. *The modified G-H algorithm converges for all normal matrices.*

(See (10).)

Instead of the special method of selection of the pivot-pair, the row cyclic method can be used. Ruhe (24) has proved that this modified G-H procedure converges quadratically. The G-H method reduces to the classical Jacobi method if the matrix is Hermitian.

III.3.2 A Simpler Procedure

There is a more straightforward generalization of the Jacobi process to normal matrices. From Theorems 0.5 and 0.7, matrix A can be written as $A = H+S$ and if A is normal then we have $HS = SH$ where H is

Hermitian and S is skew-Hermitian.

The procedure constructs again a sequence of matrices A_k where $A_{k+1} = U_k^* A_k U_k = U_k^* H U_k + U_k^* S U_k$. The unitary transformation matrices U_k are chosen according to one of the Jacobi-like methods to diagonalize the Hermitian matrix H. Define $U = \prod_{k=1}^{N} U_k$. There is a diagonal matrix Λ such that $\Lambda = U^* H U$. S has been transformed into $S' = U^* S U$ which need not be diagonal if the eigenvalues of H are not distinct. That is, since Λ and S' commute, $S'_{pq} = 0$ implies that $\lambda_{pp} = \lambda_{qq}$. The 2×2 rotations needed to annihilate the non-zero subdiagonal elements of S' leave Λ invariant. Thus the eigenvalues of A can be obtained as the sum of eigenvalues of H and S.

This procedure is in principle a concatenation of two infinite processes. Although this method does not attempt to reduce $S^2(A)$ maximally, it is not always slower than the G-H method in practice.

III.4 Extensions to Arbitrary Matrices

The major attempts to solve the eigenvalue problem for arbitrary matrices by Jacobi-like processes can be divided into two categories. One group was concerned with the triangularization of general matrices. The idea was based on Schur's lemma (Theorem 0.8) which states that any matrix is unitarily similar to a triangular matrix. The other group was interested in reducing arbitrary matrices to normal matrices. The underlying idea was that

$$\inf_{T \text{ invertible}} \|T^{-1}AT\|_E^2 = \sum_{j=1}^{n} |\lambda_j|^2 \qquad \text{(Theorem 0.12)}$$

The eigenvalue problem can then be solved by almost diagonalizing the

almost normal matrices. To avoid concatenation of two infinite processes, Eberlein has produced an algorithm which normalizes and diagonalizes a matrix simultaneously.

III.4.1 Eberlein's Method

From Theorem 0.12 we have seen that a matrix can be brought closer to normality by reducing the Euclidean norm of the matrix. Since this norm is invariant under unitary transformations, complex rotations will not do. Eberlein has thus chosen the unimodular plane shears as transformation matrices. From Theorem 0.12 we have also seen that normality cannot always be obtained. In practice, it is more advantageous to normalize and diagonalize a matrix simultaneously.

In Eberlein's method ((3) and (4)), each iteration of the process can be divided into two steps. The first step, $\tilde{A}_k = S_k^{-1} R_k^* A_k R_k S_k$, reduces the Euclidean norm of the matrix A_k. The second step minimizes the departure of the $\tilde{A}_k$'s from diagonal form, $A_{k+1} = U_k^* \tilde{A}_k U_k$. The R_k and U_k are complex rotations, and S_k are the plane shears, all having the same pivot pair (i_k, j_k). To avoid excessive repetition, we shall define the term commutator as follows: $C(A) = AA^* - A^*A$. The details of the algorithm can be summarized.

(1) At the start of each iteration, the pivot pair is chosen to insure a bounded reduction of $\|C(A_k)\|_E^2$.

(2) The parameters of the complex rotation of $R_k(\theta, \alpha, i_k, j_k)$ are chosen to maximize c_{i_k, j_k} where $C = C(R_k^* A_k R_k) = (c_{i,j})$.

The purpose of this rotation is to prepare the matrix for norm reduction. This is a pre-treatment step.

(3) The parameters of the unimodular plane shears are chosen to minimize $\|\tilde{A}_k\|_E^2$. The correct values for these parameters, where the minimum is attained at that step can be found by solving two simultaneous quartic equations. This is not always an easy task. Eberlein thus gives an explicit approximation to these optimal parameters.

(4) The parameters of the complex rotation U_k are chosen according to one of the methods in Section III.2.

If the starting matrix is normal then Eberlein's method reduces to those mentioned in the previous section.

Eberlein proved that if the pivot pairs are suitably chosen, then the sequence $\{A_k\}$ converges to normality independently of the choices of U_k. However, the global convergence to diagonal form for all matrices cannot be proved.

III.4.2 Greenstadt's Method

This method is the application of cyclic Jacobi method to general matrices. The 2×2 non-trivial principle submatrix of the matrices of transformation can be written in the form

$$\begin{pmatrix} a & -a\bar{\mu}_k \\ a\mu_k & a \end{pmatrix}$$

where $a = (1 + \mu_k\bar{\mu}_k)^{-1/2}$ and μ_k is the root of smaller modulus of the quadratic equation

$$a_{ij}^{(k)}\mu_k^2 - (a_{jj}^{(k)} - a_{ii}^{(k)})\mu_k - a_{ji}^{(k)} = 0 .$$

By the manner in which the parameters a and μ_k are chosen and from

the work of Forsythe and Henrici, the method converges if A is Hermitian (5). For general matrices, however, the result is not so satisfactory. There is one class of matrices which produces either cycling or invariance when this algorithm is applied independent of the pivot strategy. These matrices have equal diagonal elements such that for each pair of off-diagonal elements, a_{ij}, a_{ji}, exactly one of them is zero and such that there are at least two non-zero elements in every row and column (2). For example, take the matrix

$$A = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{pmatrix} .$$

Greenstadt's method produces the following sequence of similar matrices (1):

$$A_1 = \begin{pmatrix} 1 & 0 & -1 \\ -1 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix} \qquad A_2 = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & -1 \\ -1 & 0 & 1 \end{pmatrix}$$

$$A_3 = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix} \qquad A_4 = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & 1 \\ -1 & 0 & 1 \end{pmatrix}$$

$$A_5 = \begin{pmatrix} 1 & 0 & 1 \\ -1 & 1 & 0 \\ 0 & -1 & 1 \end{pmatrix} \qquad A_6 = A$$

Different pivot strategies give similar results.

The difficulty of Greenstadt's method is that not only does it not reduce maximally the quantity $S^2(A)$ as do the classical Jacobi and Goldstine-Horwitz methods, but sometimes it increases it.

III.4.3 Lotkin's Method

As we mentioned in the beginning of this chapter, there are two approaches in Jacobi-type processes. The first is to annihilate some specific elements and the second is to reduce the norm of the lower (or upper) triangular elements a significant amount. Greenstadt concentrated on the first; Lotkin chose the second.

Lotkin specifies that the triangular part with lesser norm be chosen for annihilation. Since the object of the procedure is to have A_k tend to a lower triangular matrix, therefore if $N_u^2(A) \geq N_L^2(A)$ then set $A_0 = A^T$. Let U_k be the matrices of transformation. The 2×2 principal submatrix of U_k is of the following form:

$$\begin{pmatrix} \cos\theta_k e^{i\phi_k} & -\sin\theta_k \\ \sin\theta_k & \cos\theta_k e^{-i\phi_k} \end{pmatrix}.$$

Lotkin intends to choose θ_k and ϕ_k so that $N_u^2(A_{k+1})$ would be a minimum. But the determination of ϕ_k is quite difficult. Lotkin (19) has given explicit approximations for these parameters. When the matrix is Hermitian, it reduces to the classical Jacobi method.

The convergence properties of Lotkin's method have not been established. In fact, Lotkin's method fails to triangularize the following matrix:

$$\begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{pmatrix}.$$

The angle of rotation obtained from his formulae is null. Despite Lotkin's aim to reduce $N_u^2(A_k)$ a maximal amount at each step, his

algorithm has failed to produce a monotonically decreasing sequence $\{N_u^2(A_k)\}$. See Causey (1) for details.

III.5 Summary

We have seen in this chapter the evolution of Jacobi-like procedures. The various processes have given a satisfactory solution to the eigenvalue problem of Hermitian and normal matrices. For general matrices, Eberlein has produced a solution. If we, however, confine ourselves with true, that is unitary, Jacobi processes, then the existing algorithms and their experimental results are not quite satisfactory.

CHAPTER IV

Huang's Algorithm

IV.1 Introductory Remark

Huang's algorithm is also a Jacobi-type procedure for the triangularization of arbitrary matrices. The other methods by Greenstadt and Lotkin, as described in Chapter III, are straightforward generalizations of the original method. Greenstadt's algorithm is essentially the cyclic Jacobi method; it annihilates the lower triangular elements in some predetermined order. Lotkin's procedure generalizes the classical Jacobi method by minimizing the sum of squares of absolute values of upper (or lower) triangular elements. Huang's algorithm differs from these methods in that it is not a simple extension of some existing procedure. The algorithm produces a monotonically decreasing sequence $\{N^2(A_k)\}$ where $N^2(A) = \sum_{i>j} |a_{ij}|^2$. This is a property which neither of the above two procedures enjoy. We shall describe the algorithm in some detail in the next section.

IV.2 Description of the Algorithm

The algorithm uses the complex rotations $R(i,j,\theta,\alpha)$ as transformation matrices. At each step, the element $a_{i+1,i}$ is annihilated with the objective of reducing the quantity $N^2(A)$. When special cases occur, different strategies must be adopted to achieve this aim.

Huang's algorithm can most easily be described inductively. For the sake of clarity, we shall investigate the case $n = 3$ first.

IV.2.1 The n = 3 Case

First we note that any 2×2 matrix can be triangulated by a complex rotation in the (1,2) plane. Clearly if the matrix has complex eigenvalues and the matrix itself is real, then the parameter α will not be null.

Let us now consider a 3×3 matrix $A = (a_{ij})$. If we perform a complex rotation in the (1,2) plane to annihilate a_{21}, then the quantity $N^2(A)$ will decrease by exactly $|a_{21}|^2$. If we perform a (2,3) plane rotation on the resulting matrix $A' = (a'_{ij})$ to annihilate a'_{32}, then $N^2(A')$ will be reduced by $|a'_{32}|^2$. By repeating this process, we obtain a monotonically decreasing sequence $\{N^2(A_k)\}$. The sequence will converge to zero unless at some iteration k, A_k is of the form

$$A_k = \begin{pmatrix} a_{11}^{(k)} & a_{12}^{(k)} & a_{13}^{(k)} \\ 0 & a_{22}^{(k)} & a_{23}^{(k)} \\ a_{31}^{(k)} & 0 & a_{33}^{(k)} \end{pmatrix} . \qquad (4.1)$$

It is in general false that $a_{21}^{(k)}$ and $a_{32}^{(k)}$ are exactly zero for some k. However since $a_{21}^{(i)} \to 0$ and $a_{32}^{(i)} \to 0$ as $i \to \infty$, we have within the limits of computation $a_{21}^{(k)} = 0$ and $a_{32}^{(k)} = 0$ for some k. A matrix of the form (4.1) is the special case referred to above where the rotation parameters are chosen to decrease $N^2(A_k)$ without attempting to annihilate any element. Huang treats this case as follows.

Case 1. $a_{11}^{(k)} \neq a_{33}^{(k)}$. *pivot pair* = (1,3)

After straightforward calculations we have, for any $R(1,3,\alpha,\theta)$

$$\begin{aligned}\Delta_k &= N^2(A_k) - N^2(A_{k+1}) \\ &= 2\cos^2\theta \tan\theta \, \mathrm{Re}(e^{i\alpha} a_{31}^{(k)} (\bar{a}_{11}^{(k)} - \bar{a}_{33}^{(k)})) \\ &\quad - \sin^2\theta (|a_{11}^{(k)} - a_{33}^{(k)}|^2 + |a_{13}^{(k)}|^2 - |a_{31}^{(k)}|^2 + |a_{23}^{(k)}|^2 + |a_{12}^{(k)}|^2 \\ &\qquad - |\cos\theta \xi_{13}^{(k)} - \sin\theta (a_{11}^{(k)} - a_{33}^{(k)})|^2) \\ &\quad \textit{where} \quad \xi_{km} = a_{km} e^{-i\alpha} + a_{mk} e^{i\alpha} \\ &\geq 2\cos^2\theta \tan\theta \, \mathrm{Re}(e^{i\alpha} a_{31}^{(k)} (\bar{a}_{11}^{(k)} - \bar{a}_{33}^{(k)})) - \sin^2\theta \cdot F \\ &\quad \textit{where} \quad F = |a_{11}^{(k)} - a_{33}^{(k)}|^2 + |a_{13}^{(k)}|^2 - |a_{31}^{(k)}|^2 + |a_{23}^{(k)}|^2 + |a_{12}^{(k)}|^2 .\end{aligned}$$

The aim is to choose θ and α so that Δ_k would be maximal or simply positive. Huang did not give an explicit formula for computing the optimal parameters θ and α. But if θ is chosen to satisfy

$$\tan\theta = \begin{cases} \dfrac{\mathrm{Re}(e^{i\alpha} a_{31}^{(k)} (\bar{a}_{11}^{(k)} - \bar{a}_{33}^{(k)}))}{F} & \text{when } F \neq 0 \\ \mathrm{Re}(e^{i\alpha} a_{31}^{(k)} (\bar{a}_{11}^{(k)} - \bar{a}_{33}^{(k)})) & \text{when } F = 0, \end{cases}$$

then Δ_k is greater than zero.

After this transformation, the form (4.1) is destroyed so that the regular procedure can be begun again.

Case 2. *If* $a_{11}^{(k)} = a_{33}^{(k)}$ *and* $a_{11}^{(k)} \neq a_{22}^{(k)}$ *then a pair of rotations is applied. The first rotation is in the* (1,2) *plane. Its purpose is to interchange* $a_{11}^{(k)}$ *with* $a_{22}^{(k)}$ *without changing the value of* $N^2(A_k)$. *The rotation angle is defined by*

$$\tan\theta = \frac{-e^{i\alpha}(a_{11}^{(k)} - a_{22}^{(k)})}{a_{12}^{(k)}} .$$

It is easy to verify that this angle θ does indeed achieve the above aim. After this rotation, the element $a_{32}^{(k+1)}$ of the new matrix is no longer null. The second rotation thus has pivot pair (2,3), and the parameters θ and α are chosen to annihilate $a_{32}^{(k+1)}$.

Case 3. $a_{11}^{(k)} = a_{22}^{(k)} = a_{33}^{(k)}$. *In this case, with no loss of generality, the matrix is of the following form*

$$A_k = \begin{bmatrix} 0 & a_{12}^{(k)} & a_{13}^{(k)} \\ 0 & 0 & a_{23}^{(k)} \\ a_{31}^{(k)} & 0 & 0 \end{bmatrix}. \tag{4.2}$$

The pivot pair in this case is (2,3). *The angle* θ *can be chosen to maximize the reduction of* $N^2(A_k)$.

(1) If $a_{23}^{(k)} = 0$ *then* $\Delta_k = S^2|a_{31}^{(k)}|^2$ *where* $S = \sin\theta$.

Therefore for maximum reduction, take $\theta = \pi/2$.

(2) If $a_{23}^{(k)} \neq 0$ *and* $|a_{31}^{(k)}| \geq |a_{23}^{(k)}|$ *then* $\Delta_k \geq \sin^2\theta \cos^2\theta |a_{31}^{(k)}|^2$, *so take* $\theta = \pi/4$.

(3) If $a_{23}^{(k)} \neq 0$ *and* $|a_{31}^{(k)}| < |a_{23}^{(k)}|$ *then take* $\sin^2\theta = |a_{31}^{(k)}|^2/2|a_{23}^{(k)}|^2$ *for maximum reduction.*

After this rotation, $a_{21}^{(k+1)}$ is no longer zero. A (1,2) plane rotation can be initiated to annihilate $a_{21}^{(k+1)}$. Note that the reduction of $N^2(\cdot)$ is not caused by the (2,3) rotation alone but by the pair of (2,3) and (1,2) plane rotations.

The above procedure for the 3×3 matrices can be summarized as follows:

Step 1. *Repeatedly apply* (1,2) *and* (2,3) *rotations to annihilate* $a_{21}^{(k)}$ *and* $a_{32}^{(k)}$ *respectively. The sequence* $\{N^2(A_k)\}$ *is monotonically decreasing and thus converges. If the limit is zero then the matrix becomes upper triangular, and we are done. If the limit is positive then within the limits of computation for some* k, A_k *must be of the form (4.1). In this case proceed to Step 2.*

Step 2. *(1) If* $a_{11}^{(k)} \neq a_{33}^{(k)}$ *then do a* (1,3) *plane rotation which reduces* $N^2(A_k)$, *and go to Step 1.*

(2) If $a_{11}^{(k)} = a_{33}^{(k)}$ *and* $a_{11}^{(k)} \neq a_{22}^{(k)}$ *then do a* (1,2) *rotation to interchange* $a_{11}^{(k)}$ *and* $a_{22}^{(k)}$, *and do a* (2,3) *rotation to annihilate* $a_{32}^{(k+1)}$, *go to Step 1.*

(3) If $a_{11}^{(k)} = a_{22}^{(k)} = a_{33}^{(k)}$ *then the matrix is of the form (4.2). In this case do a* (2,3) *rotation and go to Step 1.*

We note that adjacent plane rotations are used heavily in this algorithm. From Theorem 0.13 we can see that using an $(i,i+1)$ rotation to annihilate $a_{i+1,i}$ automatically reduces $N^2(A)$ by the amount $|a_{i+1,i}|^2$.

IV.2.2 The n × n Case

In the last section we have seen Huang's algorithm for the $n = 3$ case in detail. For matrices of order n, we shall describe the algorithm briefly and inductively. Let $A = (a_{ij})$ be a matrix of order n.

Initial Step. $n = 2$

A can be trivially triangulated by one rotation.

Inductive Step. Assume that a matrix of order $(n-1)$ can be triangulated

by the algorithm. Consider matrices of order n of the form

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1,n-1} & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2,n-1} & a_{2n} \\ \vdots & & & \vdots & \vdots \\ a_{n-1,1} & & \cdots & a_{n-1,n-1} & \\ a_{n,1} & a_{n,2} & \cdots & a_{n,n-1} & a_{nn} \end{pmatrix}$$

(S_1: the upper principal submatrix of order $(n-1)$; S_2: the lower principal submatrix of order $(n-1)$)

where S_1 and S_2 are the upper and lower principal submatrices of order $(n-1)$ of A respectively. The procedure to triangularize A can be stated as follows.

Step 1. *Apply the procedure repeatedly first to* S_1, *then to the resulting* S_2. *If the resulting matrix is triangular then we are done. Otherwise, at the end of the process, the matrix is within desired degree of accuracy of the form*

$$\begin{pmatrix} a_{11}^{(k)} & a_{12}^{(k)} & \cdots & a_{1n}^{(k)} \\ 0 & a_{22}^{(k)} & \cdots & a_{2n}^{(k)} \\ \vdots & 0 & & \vdots \\ & \vdots & \ddots & \\ a_{n1}^{(k)} & 0 & \cdots 0 & a_{nn}^{(k)} \end{pmatrix} \tag{4.3}$$

with $a_{n1}^{(k)} \neq 0$. *In this case, proceed to Step 2.*

Step 2. *(1) If* $a_{11}^{(k)} \neq a_{nn}^{(k)}$ *then it is possible to do a* $(1,n)$ *plane rotation to reduce* $N^2(A_k)$. *After this rotation, if the matrix is no longer of the form (4.3), proceed to Step 1. If the resulting matrix is still of form (4.3), repeat Step 2.*

(2) *If* $a_{11}^{(k)} = a_{nn}^{(k)}$ *and there exists* k' *such that* $a_{k'k'}^{(k)} \neq a_{11}^{(k)}$ *and* $a_{11}^{(k)} = a_{ii}^{(k)}$, *for all* $i < k$ *then interchange* $a_{11}^{(k)}$ *with* $a_{k'k'}^{(k)}$. *This process does not alter the value* $N^2(A_k)$ *if the interchanges are performed in a particular way with angles of rotation specifically chosen. The resulting matrix has* $a_{n2}^{(k)} \neq 0$, *therefore go back to Step 1 to triangulate the lower principal submatrix* S_2.

(3) *If all the diagonal elements are equal, that is, without loss of generality the matrix can be assumed to have the form*

$$\begin{pmatrix} 0 & a_{12}^{(k)} & \cdots & a_{1n}^{(k)} \\ \vdots & 0 & & \vdots \\ 0 & \vdots & \ddots & \vdots \\ a_{n1}^{(k)} & 0 & \cdots\ 0 & a_{n-1,n}^{(k)} \end{pmatrix}.$$

Then apply a rotation in the $(n-1,n)$ *plane and go to Step 1 since the resulting matrix has* $a_{n-1,1}^{(k+1)}$ *non-zero. The parameters* θ *and* α *of* $R(n-1,n)$ *can be chosen to maximize the reduction of* $N^2(A_k)$ *assuming that* S_1 *can be triangulated in Step 1.*

This concludes our brief description of Huang's procedure. For detailed computation and proofs see (15).

IV.3 Discussion

The ingenuity of the algorithm lies in its success in dealing with the classes of matrices which would either cycle or be stationary under other simpler procedures. The effective use of the inductive method contributes to the simplicity of the algorithm. However due to this inductive nature of the process, Huang's algorithm is a concatenation of infinitely many infinite procedures. Since the algorithm is not concerned with the amount of reduction of the sum of squares of absolute values of the lower triangular elements, except in very few of the steps, much time may be spent on annihilating the already small elements. Therefore, this algorithm would probably be non-competitive with other existing algorithms. It is true that at each step of the procedure, the quantity $N^2(\cdot)$ can be reduced either at the current or at the next step. However, the reduction may be arbitrarily small. In this section, we will examine the procedure in more detail and point out some possible alternatives and improvements.

IV.3.1 Angles of Rotation to Achieve Annihilation

Let us take a closer look at Step 1 of Huang's algorithm. We shall examine the case $n = 3$. In Step 1, when a matrix has reached the form

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ 0 & a_{22} & a_{23} \\ a_{31} & 0 & a_{33} \end{pmatrix}$$

then it was suggested that this matrix will remain stationary under further application of Step 1. This statement is not true. In this

section we will show that the only stationary form of Step 1 is of form (4.2) (all diagonal elements equal) instead.

Let a rotation in the (1,2) plane be applied to the matrix of form (4.1). The element a_{21} will be transformed to a'_{21} in the following way.

$$a'_{21} = -e^{-i\alpha}csa_{11} + e^{-i\alpha}csa_{22} - e^{-2i\alpha}s^2a_{12}$$

where $c = \cos\theta$ and $s = \sin\theta$. In order to prevent an increase in $N^2(A)$, a'_{21} has to remain null. The parameters θ and α thus have to satisfy

$$e^{-i\alpha}cs(a_{22}-a_{11}) - e^{-2i\alpha}s^2a_{12} = 0 .$$

There are two angles of rotation which can achieve this:

(1) Obviously $\theta = 0$ is a solution.

(2) $\tan\theta = e^{i\alpha}(a_{22}-a_{11})/a_{12}$ is another.

The second solution coincides with the first if and only if the two diagonal elements are equal. We have seen then that matrices of form (4.1) are not invariant under Step 1. If $a_{11} = a_{22}$ but $a_{22} \neq a_{33}$ then similarly there is an angle $\theta \neq 0$ such that a (2,3) plane rotation will annihilate a_{32}. Therefore, we have shown that the stationary matrices of Step 1 are of the form (4.2) where all diagonal elements are equal. We note that the non-zero angle of rotation in these cases is the angle used to interchange unequal diagonal elements in Step 2 of Huang's algorithm.

IV.3.2 A Simplification of the Algorithm

We will first describe Step 2-(2) of Huang's algorithm in more detail. The matrices in Step 2 have the form (4.3). If $a_{11} = a_{nn}$ and there exists k such that $a_{kk} \neq a_{11}$ and $a_{11} = a_{ii}$ for all $i < k$, then a_{11} will be interchanged with a_{kk} in the following manner. We will interchange a_{kk} with $a_{k-1,k-1}$, then $a_{k-1,k-1}$ with $a_{k-2,k-2}$ etc. until finally a_{22} is interchanged with a_{11}. This last rotation in the (1,2) plane will cause a_{n2} to become non-zero. In this step, the diagonal elements are being interchanged up the diagonal. It is perfectly reasonable to interchange them down the diagonal. In the latter case the last rotation will be in the (n-1,n) plane, and the element $a_{n-1,1}$ will become non-zero instead. This observation makes Step 2-(1) of Huang's algorithm totally unnecessary. We will describe the simplified method as follows:

Step 1. *Same as Huang's procedure.*

Step 2. *Let* ℓ *be such that* $a_{11}^{(k)} \neq a_{\ell\ell}^{(k)}$ *and* $a_{11}^{(k)} = a_{ii}^{(k)}$ *for all* $i < \ell$ *and* m *be such that* $a_{nn}^{(k)} \neq a_{mm}^{(k)}$ *and* $a_{nn}^{(k)} = a_{jj}^{(k)}$ *for all* $j > m$. *If* ℓ *exists then let* $k = \min((\ell-1),(n-m))$.

Case 1. *If* $k = \ell-1$ *then interchange* $a_{\ell\ell}^{(k)}$ *with* $a_{\ell-1,\ell-1}^{(k)}$, *then* $a_{\ell-1,\ell-1}^{(k)}$ *with* $a_{\ell-2,\ell-2}^{(k)}$, *etc. until finally* $a_{22}^{(k)}$ *is interchanged with* $a_{11}^{(k)}$. *After this step,* $a_{n2}^{(k+1)} \neq 0$. *Go to Step 1.*

Case 2. *If* $k = n-m$ *then interchange* $a_{mm}^{(k)}$ *with* $a_{m+1,m+1}^{(k)}$, *then* $a_{m+1,m+1}^{(k)}$ *with* $a_{m+2,m+2}^{(k)}$ *etc. until finally* $a_{n-1,n-1}^{(k)}$ *is interchanged with*

$a_{nn}^{(k)}$. *After this step,* $a_{n-1,1}^{(k+1)} \neq 0$. *So go back to Step 1.*

If ℓ does not exist then the matrix has equal diagonal elements. In this case go to Step 2-(3) of Huang's scheme.

IV.3.3 The Elimination of the Concatenation of Infinite Processes In the Case n = 3

Huang's procedure is a concatenation of infinitely many infinite processes. This is not a desirable characteristic. In this section we will show that for 3×3 matrices, the algorithm can be changed to yield one infinite process.

From Section IV.3.1 we have seen that the matrices that are invariant under further applications of Step 1 have equal diagonal elements. Huang's scheme for the case $n = 3$ thus reduces to:

Step 1. *Same as described in Section IV.2.1.*

Step 2. *Same as Step 2-(3) in Section IV.2.1.*

Step 1 is itself an infinite process. To eliminate the concatenation of infinite procedures, one has to apply Step 2 prior to the 'end' of Step 1. The question to be answered then is how small must the quantities $|a_{21}|$, $|a_{32}|$, $|D_{12}|$, $|D_{23}|$ be in order that the parameters of the rotation in Step 2 can be chosen to yield a reduction in $N^2(A)$, where $D_{ij} = a_{ii} - a_{jj}$? We have seen in Section IV.2.1 that if a_{21}, a_{32}, D_{12} and D_{23} are null then a rotation can be chosen to reduce $N^2(A)$ provided that a'_{21} is annihilated by the next $(1,2)$ plane rotation.

Let $A = (a_{ij})$ be a 3×3 matrix. Then $N^2(A) = |a_{21}|^2 + |a_{31}|^2 + |a_{32}|^2$. Let

$A' = R_2^*(1,2,\theta_2,\alpha_2)R_1^*(2,3,\theta_1,\alpha_1)AR_1(2,3,\theta_1,\alpha_1)R_2(1,2,\theta_2,\alpha_2)$ where R_2 is chosen to annihilate the (2,1) element of $R_1^*AR_1$. Suppose that the quantities $|a_{21}|$, $|a_{32}|$, $|D_{12}|$ and $|D_{23}|$ are less than some $\epsilon > 0$. The aim is to find this ϵ such that the parameters θ_1 and α_1 can be chosen to yield $N^2(A') < N^2(A)$.

We have, after the rotations R_1 and R_2,

$$\begin{aligned} N^2(A') &= |-e^{-i\alpha_1}csD_{23}+c^2a_{32}-e^{-2i\alpha_1}s^2a_{23}|^2 + |-e^{-i\alpha_1}sa_{21}+ca_{31}|^2 \\ &= c^4|a_{32}|^2 + s^4|a_{23}|^2 + c^2s^2|D_{23}|^2 + 2c^3s\mathrm{Re}(e^{i\alpha_1}a_{32}\bar{D}_{23}) \\ &\quad - 2c^2s^2\mathrm{Re}(e^{2i\alpha_1}a_{32}\bar{a}_{23}) - 2cs^3\mathrm{Re}(e^{i\alpha_1}D_{23}\bar{a}_{23}) \\ &\quad + c^2|a_{31}|^2 + s^2|a_{21}|^2 + 2cs\mathrm{Re}(e^{i\alpha_1}a_{31}\bar{a}_{21}) \end{aligned}$$

where $c = \cos\theta_1$ and $s = \sin\theta_1$.

$$\begin{aligned} \Delta &= N^2(A) - N^2(A') \\ &= (1-c^4)|a_{32}|^2 + c^2|a_{21}|^2 + s^2|a_{31}|^2 - s^4|a_{23}|^2 - c^2s^2|D_{23}|^2 \\ &\quad - 2c^3s\mathrm{Re}(e^{i\alpha_1}a_{32}\bar{D}_{23}) + 2c^2s^2\mathrm{Re}(e^{2i\alpha_1}a_{32}\bar{a}_{23}) \\ &\quad + 2cs^3\mathrm{Re}(e^{i\alpha_1}D_{23}\bar{a}_{23}) - 2cs\mathrm{Re}(e^{i\alpha_1}a_{31}\bar{a}_{21}) \qquad (4.4) \\ &\geq (1-c^4)|a_{32}|^2 + c^2|a_{21}|^2 + s^2|a_{31}|^2 - s^4|a_{23}|^2 - \tfrac{1}{4}|D_{23}|^2 \\ &\quad - |a_{32}||D_{23}| - \tfrac{1}{2}|a_{32}||a_{23}| - |D_{23}||a_{23}| - |a_{31}||a_{21}| \\ &\qquad \text{using } |cs| \leq \tfrac{1}{2} \text{ and } c^2s^2 \leq \tfrac{1}{4} \\ &\geq s^2|a_{31}|^2 - s^4|a_{23}|^2 - \tfrac{5}{4}\epsilon^2 - \epsilon(\tfrac{3}{2}|a_{23}| + |a_{31}|) \quad . \qquad (4.5) \end{aligned}$$

If $\epsilon = 0$ then $\Delta_0 \geq s^2|a_{31}|^2 - s^4|a_{23}|^2$ which is maximum at $\theta_1 = \frac{\pi}{2}$ if $|a_{31}|^2/2|a_{23}|^2 \geq 1$ and at $s^2 = |a_{31}|^2/2|a_{23}|^2$ if $|a_{31}|^2/2|a_{23}|^2 < 1$.

Case 1. *If* $|a_{31}|^2/2|a_{23}|^2 \geq 1$ *then choose* $s = 1$. *From equation (4.4) we have*

$$\Delta = |a_{32}|^2 + |a_{31}|^2 - |a_{23}|^2$$
$$> 0$$

for any magnitude of $|a_{21}|$, $|a_{32}|$, $|D_{12}|$, $|D_{23}|$.

Case 2. *If* $|a_{31}|^2/2|a_{23}|^2 < 1$ *then choose* $s^2 = |a_{31}|^2/2|a_{23}|^2$. *From equation (4.5) we have*

$$\Delta \geq \tfrac{1}{4}|a_{31}|^4/|a_{23}|^2 - \tfrac{5}{4}\epsilon^2 - \epsilon(\tfrac{3}{2}|a_{23}| + |a_{31}|)$$

$\Delta > 0$ *if and only if*

$$f(\epsilon) = \tfrac{5}{4}\epsilon^2 + \epsilon(\tfrac{3}{2}|a_{23}| + |a_{31}|) - \tfrac{1}{4}|a_{31}|^4/|a_{23}|^2$$
$$< 0$$

$f(\epsilon)$ *is a quadratic equation in* ϵ *with roots* ϵ_1, ϵ_2 *real and* $\epsilon_1 > 0$, $\epsilon_2 < 0$. $f(\epsilon) < 0$ *for all* $\epsilon_2 < \epsilon < \epsilon_1$.

Therefore, from the above analysis, the concatenation of infinite procedures can be eliminated in the $n = 3$ case. Note also that if $g(\epsilon) = -\epsilon$ and ϵ_0 is such that $f(\epsilon_0) = g(\epsilon_0)$ then for $|a_{21}| < \epsilon_0$, a pair of rotations R_1, R_2 defined above will reduce the quantity $N^2(A)$ more than if a_{21} is simply annihilated by one rotation. In eliminating the concatenation of infinite processes, we were only interested in a positive reduction of the quantity $N^2(\cdot)$. The magnitude of the reduction is not guaranteed.

IV.3.4 Non-convergence

Huang's algorithm produces a sequence of similar matrices with monotonically decreasing $N^2(\cdot)$. However, the decrement can be arbitrarily small, and because of the limits of computation, it is not always bounded away from zero. This point can be illustrated by the following example.

Let

$$A = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 10^{-9} & 0 & 0 \end{pmatrix}.$$

When Huang's algorithm is applied to the above matrix, the following two rotations are performed:

(1) A (2,3) plane rotation with $\theta = \arcsin(\frac{10^{-9}}{\sqrt{2}})$.

(2) A (1,2) plane rotation is performed to annihilate the (2,1) element of the new matrix.

After this iteration, the quantity, $N^2(\cdot)$, is reduced by

$$\Delta = \frac{|a_{31}|^4}{4|a_{23}|^2} = \frac{10^{-36}}{4}$$

which is positive, but with the limitations of most computing machines, the reduction is regarded as null.

APPENDIX

In this section we shall give an account of the basic theory which is used in previous chapters. The proofs of the theorems will not always be given. Interested readers can either prove them as an exercise or read the cited references. We shall begin with a few elementary definitions of matrix theory.

Let A be an $n \times n$ matrix.

Definition 0.1. A *is symmetric if and only if* $A = A^T$.

Definition 0.2. A *is Hermitian if and only if* $A = A^*$ *(where* $A^* = A^T$ *if* A *is real).* A *is skew-Hermitian if and only if* $A^* = -A$.

Definition 0.3. *A matrix* A *is unitary if and only if* $A^*A = I$, *that is,* $A^{-1} = A^*$.

Definition 0.4. *The function* $\det(A-\lambda I)$ *is a* n^{th} *degree polynomial in* λ. *In the complex number field, it always has* n *roots,* $\lambda_1, \lambda_2, \ldots, \lambda_n$. *The* λ_i*'s are called the eigenvalues of* A. *Corresponding to each* λ_i, *the set of* n *homogeneous linear equations* $Ax_i = \lambda_i x_i$ *has at least one non-trivial solution* x_i. *Such a solution is called an eigenvector corresponding to that eigenvalue.*

Definition 0.5. *Let* T *be any non-singular matrix. If* $B = T^{-1}AT$ *then* A *and* B *are said to be similar. Transforms,* $S^{-1}AS$, *of* A *are called similarity transformations, and the matrix* S *is called the matrix of transformation.*

We shall list below some elementary facts. The proofs can be found in

any introductory linear algebra book.

Theorem 0.1. *The eigenvalues of Hermitian matrices are real.*

Theorem 0.2. *Hermitian matrices have* n *linearly independent eigenvectors.*

Theorem 0.3. *Similarity transformations preserve eigenvalues, that is,* A *and* $T^{-1}AT$ *have the same eigenvalues.*

Definition 0.6. *A matrix* A *is normal if and only if* $AA^* = A^*A$.

One very important characterization of normal matrices is the following.

Theorem 0.4. *A matrix is normal if and only if it is unitarily similar to a diagonal matrix (see (27), pp. 51-52). That is, the matrix* S *in Definition 0.5 is unitary.*

Theorem 0.5. *Any matrix* A *can be written as a sum of a Hermitian matrix and a skew-Hermitian matrix.*

Proof: Let $H = (A+A^*)/2$ and $S = (A-A^*)/2$. Then clearly H is Hermitian, S is skew-Hermitian, and $A = H+S$.

Theorem 0.6. *Let* A *be normal and have eigenvalues* $\mu_j + i\nu_j$. *Then the Hermitian part of* A *has eigenvalues* μ_j, *and the skew-Hermitian part of* A *has eigenvalue* $i\nu_j$.

Proof: By Theorem 1.4, there exists an unitary matrix U such that $U^*AU = \mathrm{diag}(\mu_j+i\nu_j)$. We have then $U^*A^*U = (U^*AU)^* = \mathrm{diag}(\mu_j-i\nu_j)$ and

$$U^*\left(\frac{A+A^*}{2}\right)U = \operatorname{diag}(u_j)$$

$$U^*\left(\frac{A-A^*}{2}\right)U = \operatorname{diag}(iv_j) .$$

Theorem 0.7. *A matrix* A *is normal if and only if its Hermitian part* $(A+A^*)/2$ *and its skew-Hermitian part* $(A-A^*)/2$ *commute.*

Theorem 0.8 (Schur's Lemma). *For any matrix* A, *there exists a unitary matrix* U *such that* U^*AU *is of upper triangular form (see (27), p. 50).*

Definition 0.7. *Let* $A = (a_{ij})$. *The Euclidean norm of* A *is defined by* $\|A\|_E = \left(\sum_{i,j=1}^{n} |a_{ij}|^2\right)^{1/2} = \sqrt{\operatorname{trace}(A^*A)}$. *This is also called the Schur norm and the Frobenius norm.*

Theorem 0.9. $\|\cdot\|_E$ *is invariant under unitary transformations,* $\|UAV\|_E = \|A\|_E$ *for* U, V *unitary.*

Proof:
$$\begin{aligned}\|UAV\|_E^2 &= \operatorname{trace}(V^*A^*U^*UAV)\\ &= \operatorname{trace}(V^*A^*AV)\\ &= \operatorname{trace}(AVV^*A^*)\\ &= \operatorname{trace}(AA^*)\\ &= \operatorname{trace}(A^*A) \quad \text{using} \quad \operatorname{trace}(XY) = \operatorname{trace}(YX)\\ &\qquad\qquad \text{repeatedly}\\ &= \|A\|_E .\end{aligned}$$

Theorem 0.10. *For any matrix* A *with eigenvalues* λ_j, *we have* $\|A\|_E^2 \geq \sum_{j=1}^{n} |\lambda_j|^2$; *equality holds if and only if* A *is normal.*

Proof: This follows directly from Theorems 0.4, 0.8 and 0.9.

Definition 0.8. *An* $n \times n$ *matrix* A *is said to be defective if it does not possess* n *linearly independent eigenvectors.*

Theorem 0.11. *A matrix is defective if and only if it is not diagonalizable (see (27), p. 3).*

Definition 0.9. *Define* $D(A) = (\|A\|_E^2 - \sum_{i=1}^{n} |\lambda_i|^2)^{1/2}$, *and call* $D(A)$ *the departure from normality of* A.

Theorem 0.12. *For any matrix* A, $\inf D(T^{-1}AT)$, *over all invertible* T, *is zero. The infimum is a minimum if and only if* A *is not defective (see (22)).*

This result suggests that one can minimize $D(A)$ to produce a similar almost normal matrix. However it also assures us that normality is not always obtainable.

We have seen that it is theoretically possible to daigonalize a normal matrix and to triangularize a general matrix by a unitary similarity transformation. In general, this matrix of transformation can only be obtained as a product of infinitely many matrices. We shall describe below the particular classes of matrices that characterize Jacobi type methods.

Let R be a linear transformation from the real n-space to itself which rotates the $(i,j)^{th}$ coordinate plane through angle θ. The matrix representation of R differs from I, the identity matrix, only in elements $r_{ii}, r_{ij}, r_{ji}, r_{jj}$ where

$$r_{ii} = r_{jj} = \cos\theta$$
$$r_{ij} = -\sin\theta \quad \text{if} \quad i < j$$
$$r_{ji} = \sin\theta$$

We shall denote this matrix by $R(i,j,\theta)$. In the complex case, we have $R(i,j,\theta,\alpha)$ which differs from I also in the elements r_{ii}, r_{ij}, r_{ji}, r_{jj}. In this case, we have

$$r_{ii} = r_{jj} = \cos\theta$$
$$r_{ij} = -\sin\theta e^{-i\alpha} \quad \text{if} \quad i < j$$
$$r_{ji} = \sin\theta e^{i\alpha}$$

Matrices $R(i,j,\theta)$ are called plane rotations, and $R(i,j,\theta,\alpha)$ are called complex rotations. The pair (i,j) is called the pivot pair of rotation R.

Rotations preserve angles. We have another type of matrix called a plane shear which preserves areas. Plane shears also have determinants equal to one, but they are not unitary. We denote these matrices of transformation by $S(i,j,\theta,\alpha) = (s_{pq})$. We then have

$$s_{pq} = \delta_{pq} \quad \text{if} \quad p, q \neq i, j$$
$$s_{ii} = s_{jj} = \cosh\theta$$
$$s_{ij} = -\bar{s}_{ji} = -ie^{i\alpha}\sinh\theta \quad (\text{assume} \quad i < j) .$$

These matrices can be used to produce almost normal matrices which are similar to arbitrary matrices.

Let U be a real or complex rotation with pivot pair (i,j) and parameters θ, α. Let an arbitrary matrix A be transformed by U. The resulting matrix $A' = U^*AU$ has elements

$$a'_{pq} = a_{pq} \quad \text{for} \quad p, q \neq i,j$$
$$a'_{ip} = \cos\theta\, a_{ip} + e^{i\alpha}\sin\theta\, a_{jp} \quad \text{for} \quad p \neq i, j$$
$$a'_{pi} = \cos\theta\, a_{pi} + e^{-i\alpha}\sin\theta\, a_{pj}$$

$$a'_{jp} = \cos\theta\, a_{jp} - e^{-i\alpha}\sin\theta\, a_{ip} \qquad \text{for } p \neq i, j$$
$$a'_{pj} = \cos\theta\, a_{pj} - e^{i\alpha}\sin\theta\, a_{pi}$$
$$a'_{ii} = \cos^2\theta a_{ii} + e^{i\alpha}\cos\theta\sin\theta a_{ji} + e^{-i\alpha}\cos\theta\sin\theta a_{ij} + \sin^2\theta a_{jj}$$
$$a'_{ij} = -e^{i\alpha}\cos\theta\sin\theta a_{ii} - e^{2i\alpha}\sin^2\theta a_{ji} + \cos^2\theta a_{ij} + e^{i\alpha}\cos\theta\sin\theta a_{jj}$$
$$a'_{ji} = -e^{-i\alpha}\cos\theta\sin\theta a_{ii} - e^{-2i\alpha}\sin^2\theta a_{ij} + \cos^2\theta a_{ji} + e^{-i\alpha}\cos\theta\sin\theta a_{jj}$$
$$a'_{jj} = \sin^2\theta a_{ii} - e^{i\alpha}\cos\theta\sin\theta a_{ji} - e^{-i\alpha}\cos\theta\sin\theta a_{ij} + \cos^2\theta a_{jj}$$

From the above formulas we have the following often used result.

Theorem 0.13. *Let* $U = R(i,j,\theta)$ *or* $U = R(i,j,\theta,\alpha)$. *Let* $A' = U^*AU = (a'_{ij})$. *Then for* $p \neq i, j$, *we have*

$$|a'_{ip}|^2 + |a'_{jp}|^2 = |a_{ip}|^2 + |a_{jp}|^2 \;;$$
$$|a'_{pi}|^2 + |a'_{pj}|^2 = |a_{pi}|^2 + |a_{pj}|^2$$

(see (5)).

Theorem 0.14. *Any* 2×2 *complex unitary matrix may be obtained from the formula*

$$V = \begin{pmatrix} \cos\theta e^{i\alpha} & -\sin\theta e^{i\beta} \\ \sin\theta e^{i\gamma} & \cos\theta e^{i\delta} \end{pmatrix}$$

where $\theta, \alpha, \beta, \gamma, \delta$ *are real and* $\alpha - \beta - \gamma + \delta = 0 \pmod{2\pi}$ *(see (1)).*

Theorem 0.15. *Every orthogonal matrix with determinant* $+1$ *can be expressed as a product of plane rotations (see (20)).*

This suggests that the subset of complex rotations is a large enough subset of 2×2 unitary matrices for them to be effective tools in diagonalizing or triangularizing matrices.

REFERENCES

(1) Robert L. Causey, "Computing eigenvalues of non-Hermitian matrices by methods of Jacobi type," J. Soc. Indust. Appl. Math. 6 (1958), pp. 172-181.

(2) B. Dimsdale, "The non-convergence of a characteristic root method," J. Soc. Indust. Appl. Math. 6 (1958), pp. 23-25.

(3) P.J. Eberlein, "A Jacobi-like method for the automatic computation of eigenvalues and eigenvectors of an arbitrary matrix," J. Soc. Indust. Appl. Math. 10 (1962), pp. 74-88.

(4) P.J. Eberlein and J. Boothroyd, "Solution to the eigenproblem by a norm reducing Jacobi type method," Numerische Mathematik 11, (1968), pp. 1-12.

(5) 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.

(6) J.G.F. Francis, "The QR transformation, Parts I and II," Computer J. 4 (1961, 1962), pp. 265-271, 332-345.

(7) W. Givens, "Computation of plane unitary rotations transforming a general matrix to triangular form," J. Soc. Indust. Appl. Math. 6 (1958), pp. 26-50.

(8) W. Givens, "The characteristic value-vector problem," J. Assoc. Comput. Mach. 4 (1957), pp. 298-307.

(9) 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.

(10) H.H. Goldstine and L.P. Horwitz, "A procedure for the diagonalization of normal matrices," J. Assoc. Comput. Mach. 6 (1959), pp. 176-195.

(11) J. Greenstadt, "A method for finding roots of arbitrary matrices," Math. Tables and Other Aids Comput. 9 (1955), pp. 47-52.

(12) J. Greedstadt, "Some numerical experiments in triangularizing matrices," Numerische Mathematik 4 (1962), pp. 187-195.

(13) R.T. Gregory, "Computing eigenvalues and eigenvectors of a symmetric matrix on the ILLIAC," Math. Tables and Other Aids Comput. 7 (1953), pp. 215-220.

(14) P. Henrici, "On the speed of convergence and cyclic and quasicyclic Jacobi methods for computing eigenvalues of Hermitian matrices," J. Soc. Indust. Appl. Math. 6 (1958), pp. 144-162.

(15) C.P. Huang, "A Jacobi-type method for triangularizing an arbitrary matrix," SIAM J. Num. Anal. 12, 4 (Sept. 1975).

(16) C.G.J. Jacobi, "Ein leichtes verfahren, die in der theorie der Säkularstörungen vorkommenden gleichungen numerisch aufzulösen," J. Reine Angen. Math. 30 (1846), pp. 51-95.

(17) H.P.M. van Kempen, "On the convergence of the classical Jacobi method for real symmetric matrices with non-distinct eigenvalues," Numerische Mathematik 9 (1966), pp. 11-18.

(18) H.P.M. van Kempen, "On the quadratic convergence of the special cyclice Jacobi method," Numerische Mathematik 9 (1966), pp. 19-22.

(19) M. Lotkin, "Characteristic values of arbitrary matrices," Quart. Appl. Math. 14 (1956), pp. 267-275.

(20) L. Mirsky, "On the minimization of matrix norms," Amer. Math. Monthly 65 (1958), pp. 106-107.

(21) B. Parlett, "Global convergence of the basic QR algorithm on Hessenberg matrices," Math. of Comp. 22, 104 (October 1968).

(22) D.A. Pope and C. Tompkins, "Maximizing functions of rotations - experiments concerning speed of diagonalization of symmetric matrices using Jacobi's method," J. Assoc. Comput. Mach. 4 (1957), pp. 459-466.

(23) A. Ruhe, "On the quadratic convergence of a generalization of the Jacobi method to arbitrary matrices," B.I.T. 8 (1968), pp. 210-231.

(24) A. Ruhe, "On the quadratic convergence of the Jacobi method for normal matrices," B.I.T. 7 (1967), pp. 305-313.

(25) H. Rutishauser, "The Jacobi method for real symmetric matrices," Numerische Mathematik 9 (1966), pp. 1-10.

(26) J.H. Wilkinson, "Note on the quadratic convergence of the cyclic Jacobi process," Numerische Mathematik 4 (1962), pp. 296-300.

(27) J.H. Wilkinson, The Algebraic Eigenvalue Problem, Clarendon Press, Oxford, 1965.