In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the Head of my Department or by his representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of Mathematics

The University of British Columbia,
Vancouver 8, Canada.

Date September 8, 1962

ABSTRACT

The purpose of this thesis is to give a survey of the methods currently used to solve the numerical eigenvalue problem for real symmetric matrices. On the basis of the advantages and disadvantages inherent in the various methods, it is concluded that Householder's method is the best.

Since the methods of Givens, Lanczos, and Householder use the Sturm sequence bisection algorithm as the final stage, a complete theoretical discussion of this process is included.

Error bounds from a floating point error analysis (due to Ortega), for the Householder reduction are given. In addition, there is a complete error analysis for the bisection process.

I hereby certify that this abstract is satisfactory.

.....

.....

ACKNOWLEDGEMENTS

The author wishes to express his gratitude to his supervisor Dr. T.E. Hull for his patient and helpful guidance in preparing this thesis.

To the staff of the Computing Centre of the University of British Columbia, the author expresses his thanks for the hours they spent in helping in the writing of a program.

Finally, the author thanks the National Research Council of Canada for their financial support.

TABLE OF CONTENTS

CHAPTER I

Introduction and Summary

According to Lanczos [9], matrix theory has its origin in the solution of simultaneous linear algebraic equations. Once a complete symbolization of algebra was introduced, a general solution of a system of equations by Cramer's rule was discovered. However, the emphasis was still on arithmetic. During the nineteenth century, interest in the operational aspects of mathematics came into focus. Cayley (1859) extended the realm of algebra by showing that a matrix can be regarded as one single algebraic operator. The theory of the characteristic equation was developed by Sylvester and Weierstrass and finally, Frobenius gave a complete algebraic theory. Fredholm (1900) extended the algebraic theory of the characteristic equation to the case of infinitely many variables, thus laying the foundation for the geometric treatment of linear differential and integral operators.

The characteristic equation with the associated eigenvalues and eigenvectors has many fields of application. These include vibrations, atomic and molecular oscillations of particles, boundary value problems, and factor analysis. Evidently then, a knowledge of the methods available for the numerical solution of the eigenvalue problem is important.

The purpose of this thesis is to give an exposition of these methods for real symmetric matrices. The essay has two main sections. We begin Chapter II by discussing, briefly, the determinant and "serial" methods for obtaining eigenvalues. The shortcomings of these methods are pointed out. Then, the more successful methods of Jacobi, Givens, and Lanczos are described in some detail, and, we complete the descriptions by giving a detailed account of Householder's reduction algorithm. Reference to detailed accounts, proofs of convergence, and error analyses are provided where available. The last

section of Chapter II deals with the Sturm sequence algorithm which is used as the final stage in the methods of Givens, Lanczos, and Householder.

Originally, we had planned to obtain a floating point error analysis for the Householder reduction and to present the details in Chapter III. In addition, several numerical experiments were planned. Before this work was completed, James Ortega's paper [10] appeared and we discovered that he had treated, in detail, all that we had planned. As a result, in Chapter III, we first present the basic preliminaries necessary for any floating point error analysis and then limit ourselves to stating the results obtained by Ortega. For completeness of treatment, we also give an error analysis for the Sturm sequence bisection algorithm. In the last section of this essay, we justify our emphasis on the Householder algorithm and indicate an area for further research.

CHAPTER II

METHODS FOR REAL SYMMETRIC MATRICES

Classical Methods

Here, as in the rest of this chapter, we let $A = (a_{ij})$, with $a_{ij} = a_{ji}$ for $i, j = 1,2,\ldots,N$, denote a real symmetric matrix. From a theoretical point of view, it is apparent that the eigenvalues of A could be determined by finding the N real roots of the equation

$$\det (A - \lambda I) = 0$$

which is an N-th degree polynomial equation in λ. Unfortunately, a satisfactory realization (i.e. on an automatic computer) of this theory is not yet feasible for matrices of relatively large order, say $N \sim 100$. The details in support of this statement are to be found in a paper by H.H. Goldstine, F.J. Murray, and J. von Neumann [4]. We sketch, briefly, some of their results.

The direct use of $\det (A - \lambda I)$ involves two problems. These are to determine the coefficients C_i $(i = 1,2,\ldots,N)$ of the equation

$$\det (A - \lambda I) = \lambda^N + C_1 \lambda^{N-1} + C_2 \lambda^{N-2} + \ldots + C_N = 0$$

and then, to determine the N real roots. Goldstine, Murray, and von Neumann divide the known methods for determining the C_i into three classes [4; p.60]. One of these classes is rejected on the grounds that the number of multiplications required is of the order of 2^N [4; p.60] which is a prohibitive figure for $N \sim 100$. The methods in the other two classes are rejected by giving an example where the ratio of the largest to the smallest coefficient is of the order 10^{43} [4; p.61]. They conclude [4; p.61]:

> It is very difficult for us to see how any procedure which gets all the coefficients $C_1,\ldots,C_N$ at one time, ..., can give results with any acceptable precision

unless a very large number of digits are carried throughout.

The authors next discuss the problems inherent in root finding algorithms. Again it is shown [4; p.62] that accuracy would be obtained only if a large number of digits are carried throughout.

The authors go on to discuss the "serial" methods for finding the characteristic values of a matrix. They point out that most of these methods depend upon the spectral decomposition of the matrix A. An example is the power method (see e.g. [16], p.33). The authors also point out that these methods are costly in matrix multiplications and that to get an accuracy of 10^{-5} in the largest determined eigenvalue, 15 decimal digits must be carried to allow for loss of precision due to the inherent instabilities of these methods [4; p.65].

Another difficulty of most of these schemes is that, in case all eigenvalues are desired, we must be aware of the fact that the approximation, λ_i, to the i-th eigenvalue is contaminated by the errors in the previous ones - namely, $\lambda_1, \lambda_2, \ldots, \lambda_{i-1}$ [4; p.65].

Thus, with regard to the classical and serial methods, we believe, as does Givens [3; p.3], that Goldstine, Murray, and von Neumann have shown that these methods are unsuitable for use with automatic computers if all eigenvalues are wanted and the matrix is of relatively high order, say $N \sim 100$.

Recent Methods

We now begin the description of methods currently in use. These methods do not require computation of the coefficients of the characteristic equation. Moreover, they yield the eigenvalues in such a manner that any error in the approximation, λ_i, to the i-th eigenvalue of A is not contaminated by errors in $\lambda_1, \lambda_2, \ldots, \lambda_{i-1}$. Most of the descriptions were taken from a paper by Paul A. White [13].

The Jacobi Method

In 1846, C.G.J. Jacobi [7] introduced a method of reducing A to diagonal form by means of a sequence of simple orthogonal transformations known as plane rotations. If we let $U_{ij}^{(r)}$ be the r-th orthogonal matrix, then we obtain a sequence of transformed matrices $A^{(r)}$ with $A^{(0)} = A$ and $A^{(r)} = U_{ij}^{(r)} A^{(r-1)} U_{ij}^{(r)}$ where

$$U_{ij}^{(r)} = \begin{cases} u_{ii} = \cos\theta_r \\ u_{jj} = \sin\theta_r \\ u_{ij} = \sin\theta_r \\ u_{ji} = \cos\theta_r \\ u_{kk} = 1 \quad , k = 1,\ldots,N \quad k \neq i,j \\ u_{ij} = 0 \qquad \text{otherwise,} \end{cases}$$

the angle, θ_r, being chosen so that the elements in the (i,j) and (j,i) positions become zero. That is, Jacobi's method depends on the choice of U_{ij} such that

$$A^{(r)} = U_{i_r j_r}^{(r)'} U_{i_{r-1} j_{r-1}}^{(r-1)'} \cdots U_{i_1 j_1}^{(1)'} A\, U_{i_1 j_1}^{(1)} \cdots U_{i_{r-1} j_{r-1}}^{(r-1)} U_{i_r j_r}^{(r)} \longrightarrow D$$

where D is a diagonal matrix and $U_{i_r j_r}^{(r)'}$ is the transpose of $U_{i_r j_r}^{(r)}$. Jacobi annihilated (i.e. rotated to zero) the maximum off-diagonal element at each stage, and he was able to prove that after some finite number of steps, M, all off-diagonal elements would be less in magnitude than any preassigned $\epsilon > 0$. J.H. Wilkinson [17] gives the details for this method and also discusses the practical details for actual numerical work. Goldstine, Murray, and von Neumann [4] give a thorough theoretical discussion of the Jacobi method.

There are two evident drawbacks to this method. The first is that scanning the matrix for the largest off-diagonal element at each stage may be time consuming. The second is that the nature of the orthogonal

transformations in no way guarantees that an element once rotated to zero will remain zero throughout successive stages. Hence, the scanning must be done over the entire set of off-diagonal elements. There is, however, a complete error analysis for this method [4]. Because of the drawbacks, White notes that this method has been replaced, in practice, by two variations which we now describe.

The first of these is known as the cyclic Jacobi method. This method removes the first drawback by systematically reducing to zero in turn each element of the first row, regardless of size, provided of course, the element is not already small enough; then, the second row, and then, the third, etc. Because of the second drawback, this procedure is iterated until all off-diagonal elements are sufficiently small. With sufficient restrictions on the rotation angles, G. Forsythe and P. Henrici [2] have been able to prove that this method converges.

The second variation is really a modification of the preceding method and is due to Pope and Tompkins [12]. We start with some threshold value $1 > t > 0$ and reduce to zero first, only those off-diagonal elements whose magnitude exceeds t. Iteration is done until all off-diagonal elements are $< t$. For the second stage, the threshold value is t^2. This procedure is continued until t^2 is sufficiently small. The advantage of this method is that it is faster [13; p.398] than either the classical Jacobi or cyclic Jacobi method. Probably the best of this class of methods would be some combination "cyclic-threshold" procedure.

Givens' Method

A detailed account of the theory and a complete error analysis for Givens' method occurs in an Oak Ridge National Laboratory Report [3]. This method is based on the reduction of A, not to diagonal, but to triple-diagonal form. This form is reached after $\frac{(N-1)(N-2)}{2}$ rotations. The

first rotation is made in the (2,3) plane and the angle of rotation is chosen to make the element a_{13} zero. Then, systematically, the elements in positions (1,4),..., (1,N) are made zero. Then elements (2,4),..., (2,N) are made zero, but the zeros in the first row and column remain unaltered.

Since the rotation may be omitted if an element that is to be made zero is already zero, Givens' method has an advantage over the classical Jacobi method; also, the rotation angle is easier to compute for Givens' method. Moreover, one systematic sweep through the matrix A results in the triple-diagonal form. A drawback, of course, is the necessity for computing the eigenvalues of the resulting triple-diagonal matrix. This is done by a Sturm sequence process which is described in the last section of this chapter.

Lanczos' Method [8]

Lanczos' method, like those of Givens and Householder, reduces a symmetric matrix to triple-diagonal form. The following description is taken from a detailed treatment by J.H. Wilkinson [17]. Starting with an arbitrary vector $\underline{b}_1$, we construct a sequence of orthogonal vectors $\underline{b}_1, \underline{b}_2, \ldots, \underline{b}_N$. $\underline{b}_2$ is taken to be the component of $A\underline{b}_1$ which is orthogonal to $\underline{b}_1$ - that is $\underline{b}_2 = A\underline{b}_1 - \alpha_1 \underline{b}_1$. The requirement that $\underline{b}_1^T \underline{b}_2 = 0$ implies $\alpha_1 = \dfrac{\underline{b}_1^T A \underline{b}_1}{\underline{b}_1^T \underline{b}_1}$.

The vector $\underline{b}_3$ is determined from the equation

$$\underline{b}_3 = A\underline{b}_2 - \alpha_2 \underline{b}_2 - \beta_2 \underline{b}_1$$

under the conditions that $\underline{b}_1^T \underline{b}_3 = 0$, $\underline{b}_2^T \underline{b}_3 = 0$.

This gives $\alpha_2 = \dfrac{\underline{b}_2^T A \underline{b}_2}{\underline{b}_2^T \underline{b}_2}$ and $\beta_2 = \dfrac{\underline{b}_1^T A \underline{b}_2}{\underline{b}_1^T \underline{b}_1} = \dfrac{\underline{b}_1^T \underline{b}_2}{\underline{b}_1^T \underline{b}_1}$.

In general, the vector $\underline{b}_{r+1}$ is determined from

$$\underline{b}_{r+1} = A\underline{b}_r - \alpha_r \underline{b}_r - \beta_r \underline{b}_{r+1} - \gamma_r \underline{b}_{r-2} - \cdots - \omega_r \underline{b}_1$$

such that $\underline{b}_{r+1}$ is orthogonal to each of $\underline{b}_r$, $\underline{b}_{r-1}, \ldots, \underline{b}_1$.

This gives $\alpha_r = \dfrac{\underline{b}_r^T A \underline{b}_r}{\underline{b}_r^T \underline{b}_r}$

$$\beta_r = \frac{\underline{b}_r^T \underline{b}_r}{\underline{b}_{r-1}^T \underline{b}_{r-1}} \quad , \text{ with the other constants being zero.}$$

The last two constants, α_N and β_N, are obtained from $\underline{b}_{N+1} = A\underline{b}_N - \alpha_N \underline{b}_N - \beta_N \underline{b}_{N-1}$ by choosing $\underline{b}_{N+1}$ orthogonal to $\underline{b}_N$ and $\underline{b}_{N-1}$. It can be shown [17; p.139] that $\underline{b}_{N+1}$ is necessarily orthogonal to $\underline{b}_{N-2}, \ldots, \underline{b}_1$ and that $\underline{b}_{N+1}$ is the null vector. The process terminates at this stage.

The above description tacitly assumes that no $\underline{b}_r$ is the null vector. Such an assumption is not valid [17; pp.139-140]. In case $\underline{b}_r$ is the null vector, we replace $\underline{b}_r$ by an arbitrary vector $\underline{c}_r$ which is orthogonal to $\underline{b}_{r-1}, \underline{b}_{r-2}, \ldots, \underline{b}_1$, and continue the process. The only change introduced is that $\beta_r = 0$ [17; p.140]. We now form the matrix B defined as

$$B = (\underline{b}_1 \vdots \underline{b}_2 \vdots \ldots \vdots \underline{b}_N)$$

or

$$B = (\underline{b}_1 \vdots \underline{b}_2 \vdots \ldots \vdots \underline{c}_r \vdots \ldots \vdots \underline{b}_N) \quad - \text{ i.e., the}$$

matrix B has column vectors equal to the above constructed orthogonal set of vectors. It can be shown [17; pp.141-142] that

$$B^{-1}AB = \begin{bmatrix} \alpha_1 & \beta_2 & & & & \\ 1 & \alpha_2 & \beta_3 & & & \\ & 1 & \alpha_3 & \beta_4 & & \\ & & \ddots & \ddots & \ddots & \\ & & & \alpha_{N-1} & \beta_N \\ & & & 1 & \alpha_N \end{bmatrix}$$

or

$$B^{-1}AB = \begin{bmatrix} \alpha_1 & \beta_2 & & & & \\ 1 & \alpha_2 & 0 & & & \\ & 0 & \alpha_3 & \beta_4 & & \\ & & \ddots & \ddots & \ddots & \\ & & & \alpha_{N-1} & \beta_N \\ & & & 1 & \alpha_N \end{bmatrix}$$

if, for example, $\underline{b}_3 = 0$.

Thus $B^{-1}AB$ is similar to A. The eigenvalues of A are found by determining the eigenvalues of $B^{-1}AB$. We can, with a little work, consider a symmetric triple-diagonal matrix C instead of $B^{-1}AB$ which is not symmetric. This matrix C is obtained by normalizing the vectors $\underline{b}_i$ [17; pp.147-151]. Consequently, the Sturm sequence method for symmetric triple-diagonal matrices also applies here.

This completes our descriptions (other than Householder's method) of the methods used for finding the eigenvalues of a real symmetric matrix. We point out that the above are descriptions only and that for numerical work these descriptions hardly suffice. We now consider in detail the method

due to Householder.

Householder's Method [15]

Householder [6] suggested that the orthogonal similarity transformation, used in reducing a symmetric matrix A to triple-diagonal form, be obtained as a product of simple orthogonal matrices, P, given by the form

$$(1) \qquad P = I - 2 \underline{w}\, \underline{w}^T$$

where $\underline{w}$ is a column vector such that

$$(2) \qquad \underline{w}^T \underline{w} = 1.$$

It is easy to show that P is symmetric and orthogonal. The symmetry is obvious, and the orthogonality then follows since

$$(3) \qquad \begin{aligned} P^T P &= (I - 2 \underline{w}\, \underline{w}^T)(I - 2 \underline{w}\, \underline{w}^T) \\ &= I - 4 \underline{w}\, \underline{w}^T + 4 \underline{w}\, \underline{w}^T \\ &= I. \end{aligned}$$

In order to make Householder's method explicit, we begin by defining a column vector $\underline{w}_r$ by

$$(4) \qquad \underline{w}_r^T = (0,\ldots,0,X_r \quad X_{r+1}, \ldots, X_N) \quad ,$$

so that $\underline{w}_r$ is a vector with its first (r-1) components equal to zero. We then take P_r to be a P matrix as given by equation (1) with $\underline{w} = \underline{w}_r$. The transformation of a given (N x N), real symmetric matrix, $A = (a_{ij})$, to triple-diagonal form is effected by (N-2) successive similarity transformations $P_2, P_3, \ldots, P_{N-1}$. If we let $A = A^{(1)}$, then $A^{(r)}$ is defined by the equation

$$(5) \qquad A^{(r)} = P_r A^{(r-1)} P_r$$

where $A^{(r-1)}$ contains (N-r) elements in row (r-1) each of which is to be reduced to zero by the transformation with matrix P_r. This gives us (N-r) equations to be satisfied by the (N-r+1) elements of $\underline{w}_r$. From equation (2),

as applied to $\underline{w}_r$, we obtain

$$X_r^2 + X_{r+1}^2 + \ldots + X_N^2 = 1 .$$

These (N-r+1) equations determine the (N-r+1) elements of $\underline{w}_r$ but, because, as will be shown presently, there is a square involved, we are able to choose a determination which will give the greatest numerical stability or convenience.

Before we consider in detail the algebra that is involved, we prove two simple facts about the transformation with matrix P_r. The first of these is Result 1: The transformation with matrix P_r leaves undisturbed the zeros in rows and columns 1,2,...,r-2.

This result is routine once the form of matrix P_r is shown. Evidently,

$$P_r = \begin{bmatrix} \overbrace{\begin{matrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{matrix}}^{r-1} & & & \\ & & & \\ \begin{matrix} 0 & 0 \end{matrix} & \begin{matrix} 0 & 1 \\ 0 & 0 \end{matrix} & \begin{matrix} 0 & 0 \\ 1-2X_r^2 & -2X_2X_{r+1} \\ -2X_{r+1}X_r & 1-2X_{r+1}^2 \end{matrix} & \begin{matrix} 0 \\ -2X_rX_N \\ -2X_{r+1}X_N \end{matrix} \\ & & & \\ & & \begin{matrix} (1-2X_{N-1}^2) & -2X_{N-1}X_N \\ -2X_NX_r & 1-2X_N^2 \end{matrix} & \end{bmatrix} \Big\} r-1$$

Thus premultiplying any (NXN) matrix B by P_r leaves the first r-2 rows and columns of B unaltered. Now suppose

$$B = \begin{bmatrix} \alpha_1 & \beta_2 & & & & & & \\ \beta_2 & \alpha_2 & \beta_3 & & & 0 & & \\ & \ddots & \ddots & \ddots & & & & \\ & & \beta_{r-2} & \alpha_{r-2} & \beta_{r-1} & 0 & & 0 \\ & & & \beta_{r-1} & X & X & \ldots & X \\ & & 0 & & X & X & \ldots & X \\ & & & & & & & \end{bmatrix} = A^{(r-1)}$$

and we post multiply by P_r. It is clear that row and column (r-1) are the first to be altered. This verifies the first result.

The second simple fact is:

Result 2. If $A^{(r)} = P_r A^{(r-1)} P_r$, where $A^{(r-1)}$ and P_r are as before, then the sum of the squares of the elements of row (r-1) of $A^{(r-1)}$ remains invariant.

Proof. Because of the above discussion, it is sufficient to show the result for

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1N} \\ \vdots & & & \\ a_{N1} & a_{N2} & & a_{NN} \end{bmatrix}, \text{ with } a_{ij} = a_{ji}, \text{ and}$$

$$P_2 = \begin{bmatrix} 1 & 0 & & 0 \\ 0 & 1-2X_2^2 & \cdots & -2X_2X_N \\ \vdots & & & \\ 0 & -2X_NX_2 & \cdots & 1-2X_N^2 \end{bmatrix}.$$

Since $P_2 A P_2 = (I - 2 \underline{w}_2 \underline{w}_2^T) A (I - 2 \underline{w}_2 \underline{w}_2^T)$

$$= A - 2 \underline{w}_2 \underline{w}_2^T A - (\underline{w}_2^T A \underline{w}_2) \underline{w}_2^T$$

$$- 2 A \underline{w}_2 - \underline{w}_2 (\underline{w}_2^T A \underline{w}_2) \underline{w}_2^T,$$

$$(6) \quad P_2 A P_2 = A - 2 \underline{w}_2 \underline{q}^T - 2 \underline{q} \underline{w}_2^T$$

where

$$(7) \quad \underline{q} = A \underline{w}_2 - (\underline{w}_2^T A \underline{w}_2) \underline{w}_2 = \underline{p} - k \underline{w}_2$$

with

$$(8) \quad k = (\underline{w}_2^T A \underline{w}_2) = \underline{w}^T \underline{p} .$$

Consequently, the first row of P_2AP_2 is

$$(a_{11} , a_{12} - 2 q_1 X_2 , a_{13} - 2 q_1 X_3 , \ldots , a_{1N} - 2 q_1 X_N) .$$

Thus the sum of the squares is

$$(a_{11})^2 + (a_{12} - 2 q_1 X_2)^2 + \ldots + (a_{1N} - 2 q_1 X_N)^2$$

$$= a_{11}^2 + a_{12}^2 + \ldots + a_{1N}^2 + 4 \left[q_1^2 X_2^2 + q_1^2 X_3^2 + \ldots + q_1^2 X_N^2 \right]$$

$$- 4 \left[a_{12} q_1 X_2 + a_{13} q_1 X_3 + \ldots + a_{1N} q_1 X_N \right]$$

$$= a_{11}^2 + a_{12}^2 + \ldots + a_{1N}^2$$

since $q_1 = a_{12}X_2 + a_{13} X_3 + \ldots + a_{1N} X_N$ and $X_2^2 + X_3^2 + \ldots + X_N^2 = 1$.

This establishes the second result. (It should be noted that Wilkinson [15; p.24] states that the sum of the squares of the elements in any row must be invariant.)

Because of Result 1 above, the details of the transformation with matrix P_r will be illustrated for any stage r, if we provide the details for $A^{(r)} = A$ and $P_r = P_2$ - i.e., the first stage. Let $A = (a_{ij})$ $i,j = 1,2,\ldots, N$ and let $\underline{w}_2^T = \underline{w}^T = (0, X_2, X_3, \ldots, X_N)$, so that

$$X_2^2 + X_3^2 + \ldots + X_N^2 = 1 .$$

We wish to determine P_2 such that P_2AP_2 has zeros in positions (1,3) , (1,4), ..., (1,N) and in (3,1), (4,1), ..., (N,1) . Since left multiplication of any (N x N) square matrix by P_2 leaves the first row unaltered, $P_2 AP_2$ has the desired zeros if and only if AP_2 has the desired zeros. Thus $\underline{w}$ must be chosen accordingly.

Since $AP_2 = A(1 - 2 \underline{w}_2 \underline{w}_2^T)$

$$= A - 2 \underline{p}_2 \underline{w}_2^T ,$$

the following set of equations must be satisfied

$$(9) \qquad \begin{cases} a_{13} - 2\,p_1\,X_3 = 0 \\ a_{14} - 2\,p_1\,X_4 = 0 \\ \vdots \\ a_{1N} - 2\,p_1\,X_N = 0 \end{cases}$$

Moreover, from Result 2

$$(10) \qquad a_{12} - 2\,p_1\,X_2 = \pm\, S^{\frac{1}{2}} \qquad \text{where}$$

$$S = a_{12}^2 + a_{13}^2 + \ldots + a_{1N}^2 \quad .$$

If we multiply equation (10) by X_2 and the successive equations of (9) by $X_3, X_4, \ldots, X_N$ respectively, and then add the resulting equations we obtain by using equations(3) and (1)

$$(11) \qquad p_1 = \mp\, X_2\, S^{\frac{1}{2}} \quad .$$

Substituting (11) into (10) and solving for X_2^2 , we have

$$(12) \qquad X = \tfrac{1}{2}\left[1 \mp \frac{a_{12}}{S^{\frac{1}{2}}}\right] , \text{ and,}$$

putting (11) into equations (9) we have

$$(13) \qquad X_k = \mp\, \frac{a_{1k}}{2X_2\, S^{\frac{1}{2}}}$$

where k = 3, 4, ... , N. The upper and lower signs go together in equations (10), (12), and (13).

From equation (10), we see that

$$(14) \qquad \begin{cases} \beta_2 = a_{12} - 2\,p_1\,X_2 = \mp\, (S^{\frac{1}{2}}) \text{ and} \\ \alpha_1 = a_{11} \end{cases}$$

where we denote the final resulting triple-diagonal matrix by

$$C = \begin{bmatrix} \alpha_1 & \beta_2 & & 0 & & \\ \beta_2 & \alpha_2 & \beta_3 & & & \\ & \beta_3 & \alpha_3 & \beta_4 & & \\ & & \ddots & \ddots & & \\ & 0 & & \beta_{N-1} & \alpha_{N-1} & \beta_N \\ & & & & \beta_N & \alpha_N \end{bmatrix}$$

The above choice of signs means the X_i's are not uniquely defined (we referred to this before) and consequently, for practical work, we are free to choose that sign which gives greater numerical stability.

Let us digress a moment to ascertain what aspects must be considered so that we obtain accurate results. We refer specifically to a paper by C.T. Fike [1] . In this paper, Fike defines the P- condition, $P_k(A)$, for any real, N-square matrix A and its proper value α_k . He says that " $P_k(A)$ can be regarded as a measure of the practical difficulty attached to the problem of computing the proper value α_k ." Using this P- condition, Fike goes on to show that real symmetric matrices are well conditioned - i.e., there is not too much difficulty in computing a proper value α_k of A and that "similarity transformations made with orthogonal matrices cannot cause a deterioration in the conditioning of the problem." Fike also refers to Householder [5] and Householder and Bauer [6] who suggested that orthogonal similarity transformations are particularly stable in numerical work. Wilkinson [15] also says that it is essential that the matrices, P_r, be as accurately orthogonal as possible.

As Wilkinson goes on to point out, this means that, since we determine X_3 X_4 ,..., X_N , for example, by dividing by X_2, we should choose the sign

in equation (12) such that X_2^2 is as large as possible. If we do this, the resulting equations are

$$(15) \qquad X_2^2 = \tfrac{1}{2}\left[1 + \frac{a_{12}\,SGN(a_{12})}{S^{\frac{1}{2}}}\right]$$

$$(16) \qquad X_k = \frac{a_{1k}\,SGN(a_{12})}{2\,X_2\,S^{\frac{1}{2}}}$$

with $X_2 = \sqrt{X_2^2}$, since the sign is not important, and

$$(17) \qquad \beta_2 = -\,SGN(a_{12})\,S^{\frac{1}{2}}$$

where $SGN(a_{12}) = \begin{cases} +1 & \text{if } a_{12} \geqslant 0 \\ -1 & \text{if } a_{12} < 0 \,. \end{cases}$

If we use equations (12) and (13) then, according to Wilkinson [15; p.24], the equation

$$X_2^2 + X_3^2 + \ldots + X_N^2 = 1$$

is very accurately satisfied. As a final practical detail, we point out that the transformed matrices $A^{(2)}, A^{(3)}, \ldots, A^{(N-1)}$ are obtained by using equations (6), (7), and (8).

Bisection method for the Eigenvalues of a Real Symmetric Triple-Diagonal Matrix

It is evident from the above descriptions of the methods of Lanczos, Givens and Householder, that we need a method for obtaining the eigenvalues of a real symmetric triple-diagonal matrix $C = (c_{ij})$, where

$$c_{ij} = \begin{cases} \alpha_i & j=i \\ 0 & |i-j| > 1 \end{cases}$$

and $c_{i,i+1} = c_{i+1,i} = \beta_{i+1} \qquad i = 1,2, \ldots, N-1$.

We shall consider in detail the bisection method for computing the eigenvalues of the matrix C. This method depends upon the Sturm sequence associated with the matrix $(C - \lambda I)$, λ real. Ortega [11] considers the theory in detail. He even gives a trivial example to show that the theory in Givens' Oak Ridge paper is not quite correct [11; p.26]. We consider the special case where none of the β_i's are zero. In case there are any β_i's exactly zero, they are replaced, in the program written, by the smallest positive number recognized by the machine. According to Wilkinson [17; p.130], exactly zero β_i's are very rare and in case there is a $\beta_i = 0$, he feels that it is not worthwhile to separate C into two matrices. Givens [13; p.401] has apparently shown that such a change can cause a change in the eigenvalues of not more than twice the magnitude of the non zero term replacing the zero. As the error analysis given below will show, such an error is indeed of no consequence to the accuracy of the method.

For the matrix $C - \lambda I$ and $i = 0, 1, 2, \ldots, N$, we consider a sequence f_i of the upper left principal minors defined by:

$$(17) \qquad f_i(\lambda) = \begin{cases} 1 & \text{if } i = 0 \\ (\alpha_1 - \lambda) & \text{if } i = 1 \\ (\alpha_i - \lambda) f_{i-1} - \beta_i^2 f_{i-2} & \text{for } 2 \leq i \leq N. \end{cases}$$

Definition 1. For $i \neq N$, put

$$SGN\left[f_i(\lambda)\right] = \begin{cases} +1 & \text{if } f_i(\lambda) \geq 0 \\ -1 & \text{if } f_i(\lambda) < 0, \end{cases}$$

and for $i = N$

$$SGN\left[f_N(\lambda)\right] = \begin{cases} +1 & \text{if } f_N(\lambda) > 0 \\ -1 & \text{if } f_N(\lambda) < 0 \\ -SGN\left[f_{N-1}(\lambda)\right] & \text{if } f_N(\lambda) = 0. \end{cases}$$

Definition 2. Let $A(\lambda)$ denote the number of agreements in sign of the sequence $\{f_i(\lambda)\}$ $(i = 0, 1, \ldots, N)$ calculated by means of definition 1.

Theorem: Let a real symmetric triple-diagonal matrix be given by C where none of the β_i's are zero. Then, for any real λ, the number of eigenvalues of C that are greater than λ is given by $A(\lambda)$.

Proof: We first establish the following properties of the sequence $\{f_i(\lambda)\}$:

(a) Two consecutive $f_i(\lambda)$ cannot both vanish for the same λ.

(b) If $f_i(\lambda) = 0$, then $f_{i-1}(\lambda) \cdot f_{i+1}(\lambda) < 1$ for $i = 1, \ldots, N-1$.

(c) $f_N(\lambda) = 0$ has no multiple roots. Property (a) is proved by induction. Obviously, both f_0 and f_1 cannot be zero. Hence, we assume the result for $k < N$ and show that f_k and f_{k+1} cannot both be zero. If

$$f_k(\lambda) = (\alpha_k - \lambda) f_{k-1}(\lambda) - \beta_k^2 f_{k-2}(\lambda) = 0, \text{ then}$$

$$f_{k+1}(\lambda) = (\alpha_{k+1} - \lambda) f_k(\lambda) - \beta_{k+1}^2 f_{k-1}(\lambda)$$

$$= -\beta_{k+1}^2 f_{k-1}(\lambda) \neq 0, \text{ since}$$

$\beta_{k+1} \neq 0$ and by the induction hypothesis $f_{k-1}(\lambda) \neq 0$. Property (b) is established by using property (a). If $f_i(\lambda) = 0$ then $f_{i-1}(\lambda) \neq 0$ and consequently, $\left[f_{i-1}(\lambda)\right] \cdot \left[f_{i+1}(\lambda)\right] = -\beta_{i-1}^2 \left[f_{i-1}(\lambda)\right]^2 < 0$.

To prove property (c), we first note that $f_N(\lambda) = \det(C - \lambda I)$. By property (a), if $f_N(\lambda) = 0$ then $f_{N-1}(\lambda) \neq 0$. Therefore $C - \lambda I$ is of rank N-1 since $f_{N-1}(\lambda)$ is the determinant of the matrix obtained from $C - \lambda I$ by deleting the last row and column.

$C - \lambda I$ is symmetric implies that there exists a unitary transformation U such that $U^{-1}(C - \lambda I)U = U^{-1} C U - \lambda I$ is diagonal. That is,

$$U^{-1}(C - \lambda I)U = \begin{bmatrix} \lambda_1 - \lambda & 0 & & 0 \\ 0 & \lambda_2 - \lambda & 0 & 0 \\ & & & \\ 0 & & 0 & \lambda_N - \lambda \end{bmatrix}$$

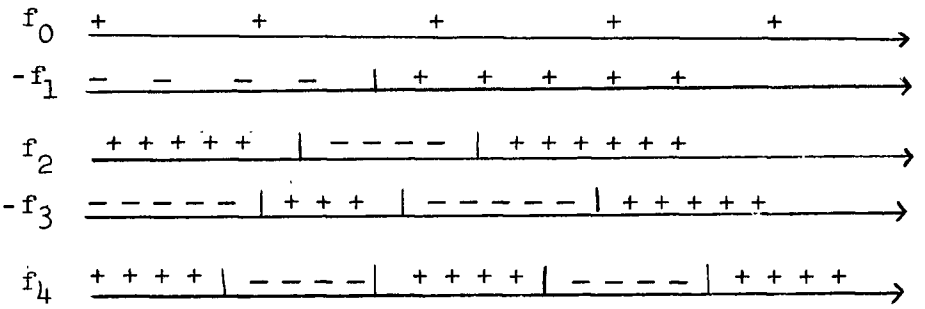
Since $C - \lambda I$ is of rank N-1, at most one of $(\lambda_i - \lambda)$ is zero for $i = 1, 2, \ldots, N$. Thus λ is a simple eigenvalue of C and a simple zero of f_N.

To establish the conclusion of the theorem, we assume

$$(*)\ \left\langle (-1)^N f_N(\lambda),\ (-1)^{N-1} f_{N-1}(\lambda),\ \ldots,\ f_0(\lambda) \right\rangle$$

is a classical Sturm sequence for λ, not a zero of f_N. That this is the case will be proved later. Thus the number of zeros of f_N greater than λ is equal to the number of variations in sign of (*). If we rewrite (*) as $\left\langle f_0, -f_1, +f_2, -f_3, \ldots \right\rangle$ then there is an agreement of signs in $\{f_i\}$ if and only if there is a corresponding difference of signs in (*). Hence, if λ is not a zero of f_N, $A(\lambda)$ gives the number of roots of $f_N(\lambda) = 0$ that are greater than λ. Now suppose λ_0 is a zero of f_N. Then by property (b) and the fact that f_{N-1} is a continuous function, we may conclude that there is some ϵ- neighborhood of λ_0, say $N(\lambda_0, \epsilon)$, such that $f_{N-1}(\lambda) \neq 0$ for $\lambda \in N(\lambda_0, \epsilon)$. Consequently, $\mathrm{SGN}\left[f_{N-1}(\lambda)\right]$ is constant for $\lambda \in N(\lambda_0, \epsilon)$. Moreover, since $f_{N-1}(\lambda_0) \neq 0$, the number of agreements in sign is constant for the sequence $\left\langle f_{N-1}(\lambda), \ldots, f_0 \right\rangle$ provided $\lambda \in N(\lambda_0, \epsilon)$. Since λ_0 is a simple zero of $f_N(\lambda)$, $A(\lambda_0 - \delta) - A(\lambda_0 + \delta) = 1$ for a δ such that $0 < \delta < \epsilon$. Therefore, $\mathrm{SGN}\left[f_N(\lambda_0 - \delta)\right] = \mathrm{SGN}\left[f_{N-1}(\lambda_0 - \delta)\right]$. In the limit, as $\delta \to 0$, we have, using definition 1, that $\mathrm{SGN}\left[f_N(\lambda_0)\right] \neq \mathrm{SGN}\left[f_{N-1}(\lambda_0)\right]$. Thus $A(\lambda_0) = A(\lambda_0 - \delta) - 1$. That is, with our choice of signs, $A(\lambda)$ gives the number of eigenvalues of C greater than λ.

To complete the proof, there remains to show that the sequence (*) is, for not a root of $f_N(\lambda) = 0$, a Sturm sequence. This is verified by induction, but before doing so, we give an illustration for the sequence $\langle f_0, -f_1, f_2, -f_3, f_4 \rangle$. Let us consider the following diagram where the horizontal lines represent the λ-axis and the heavy vertical bars the roots of f_i $(i = 0, 1, 2, 3, 4)$.



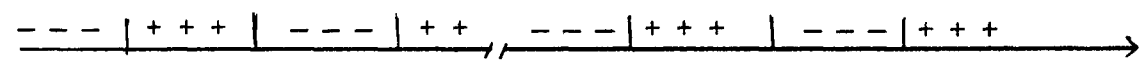
Clearly, if the zeros of f_i are in the relative positions shown, $\langle f_0, -f_1, f_2, -f_3, f_4 \rangle$ is a Sturm sequence. Consequently, for the general sequence (*), we need only show that this relative positioning is necessary. For the initial induction step we consider $\langle f_0, -f_1 \rangle$. From the above diagram, this is clearly a Sturm sequence. Assume that for $K < N$, K even, $\langle f_0, -f_1, \cdots, -f_{K-1}, f_K \rangle$ has the relative positioning mentioned. We wish to show that

$$\langle f_0, -f_1, \cdots, -f_{K-1}, f_K, -f, -f_{K+1} \rangle$$

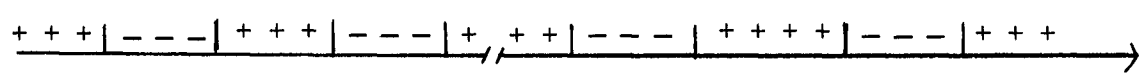
also has the desired relative positioning of the roots of $f_i = 0$ $(i = 1, 2, \ldots, K+1)$.

By the induction hypothesis, the zeros of f_K are positioned relative to those of $-f_{K-1}$ as follows:

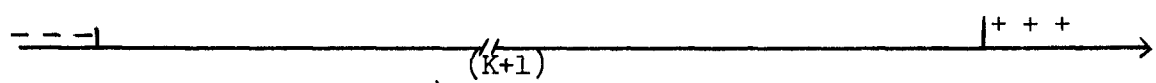
$-f_{K-1}$:



f_K :



$-f_{K+1}$:



Consider the largest root, λ_1 , of $-f_{K+1}(\lambda) = 0$. For λ large enough $-f_{K+1}(\lambda) > 0$. Hence, to the right of $\lambda_1^{(K+1)}$, the sign of $-f_{K+1}$ is positive. For $\lambda_1^{(K)}$ we have that $\left[-f_{K-1}(\lambda_1^{(K)})\right] \cdot \left[-f_{K+1}(\lambda_1^{(K)})\right] < 0$. Consequently $-f_{K+1}(\lambda_1^{(K)}) < 0$. Thus $\lambda_1^{(K)} < \lambda_1^{(K+1)}$. Similarly, $\lambda_{K+1}^{(K+1)} < \lambda_K^{(K)}$.

Because of property (a), we know that no root of $f_K(\lambda) = 0$ is a root of $f_{K+1}(\lambda) = 0$. Consequently, the remaining K-1 roots of $f_{K+1}(\lambda) = 0$ are distributed within the K-1 intervals determined by the K roots of $f_K(\lambda) = 0$. Because of properties (b) and (c) and the induction hypothesis, we may say that $f_{K+1}(\lambda_1^{(K)}) < 0$, $f_{K+1}(\lambda_2^{(K)}) > 0$, $f_{K+1}(\lambda_3^{(K)}) < 0$, ... , $f_{K+1}(\lambda_K^{(K)}) > 0$. That is, there are K-1 distinct zeros of f_{K+1} in the interval $\left[\lambda_K^{(K)} , \lambda_1^{(K)}\right]$. Clearly, the only distribution satisfying the conditions is obtained if two consecutive zeros of f_{K+1} straddle a zero of f_K. The argument for K odd is obtained by an obvious variation in the above argument. This completes the induction and we may conclude that (*) is a Sturm sequence. Moreover, the proof of the theorem is now complete.

We now give a description of the bisection procedure. We assume that the maximum modulus of the α_i's and β_i's is less than 1 - i.e., $\max_i \left\{|\alpha_i| , |\beta_i|\right\} < 1$. We find the eigenvalues λ_i $(i = 1,2,\ldots,N)$

of C starting with the largest and ending with the smallest. Theoretically, this is accomplished by evaluating $A(\lambda)$ at the points P-1, P-2,...,P-k until a k-value is reached such that $A(P-k) > 1$, where P is given by $\max_i \{|\beta_i| + |\alpha_i| + |\beta_{i+1}|\}$ $(i = 1,2,\ldots, N)$. By the above theorem, λ_1 is such that $P-k < \lambda_1 \leq P-k+1$. The interval of unit length $(P-k, P-k+1]$ is divided into two by adding $\frac{1}{2}$ to P-k. Let us put $P-k+1 = \lambda$ so that the interval we consider is $(\lambda-1, \lambda]$. To determine whether $\lambda_1 \in (\lambda-1, \lambda-\frac{1}{2}]$ or $(\lambda-\frac{1}{2}, \lambda]$, we evaluate $A(\lambda-\frac{1}{2})$. Suppose $\lambda_1 \in (\lambda-\frac{1}{2}, \lambda]$ — i.e., $A(\lambda-\frac{1}{2}) > 1$. We now divide $(\lambda-\frac{1}{2}, \lambda]$ by adding $(\frac{1}{2})^2 = \frac{1}{4}$ to $\lambda-\frac{1}{2}$ and evaluate $A(\lambda-\frac{3}{4})$ to determine whether $\lambda_1 \in (\lambda-\frac{1}{2}, \lambda-\frac{3}{4}]$ or $\lambda_1 \in (\lambda-\frac{3}{4}, \lambda]$. Continuing this process, we see that at the j-th stage $\lambda_\ell(1) < \lambda_1 \leq \lambda_u(1)$ where $\lambda_u(1) - \lambda_\ell(1) = (\frac{1}{2})^j$. In the program written, we stop at a stage j=J such that

$$(\tfrac{1}{2})^J \leq 2.10^{-t}$$

where t is the number of digits carried in the mantissa of the floating point number format. To continue the process, we now put $\lambda_u(1) = P$ and repeat the above procedure requiring, of course, that the choice of intervals be made by having $A(\lambda) > 2$. Thus to straddle λ_r by appropriate $\lambda_\ell(r)$ and $\lambda_u(r)$, we begin by putting $\lambda_u(r-1) = P$ and require that our choice of intervals depend upon $A(\lambda)$ being greater than r. The process ends when we have straddled λ_N.

A word on our scaling is in order. It should be clear that if we do not scale then, instead of subtracting 1 from a P-value, we would have to subtract a 1 scaled by 10^s, where s is determined so that in $P-10^s1$ we are subtracting a 1 from the first decimal digit of P. And, instead of adding powers of $\frac{1}{2}$, we would be adding $10^s(\frac{1}{2})^j$ for $j = 1,2,\ldots$. Since the initial scaling method is simpler to code, it's use was adopted. The relative merits of the two methods beyond the coding were not considered.

This completes our descriptions of the more successful methods currently in use. We next give the results of Ortega's error analysis of the Householder reduction and following this, we give a complete error analysis of the bisection procedure.

CHAPTER III

ERROR ANALYSIS

Preliminaries and Notation

The notation used follows that of Wilkinson [14]. Exact mathematical operations will be denoted by their usual symbols: "—" , "+", "x", and " ". The corresponding floating point operations are denoted by fl (x - y), fl (x + y), fl (x . y), and fl (x y) or fl (x/y) where x and y are floating point digital numbers. The floating point arithmetic subroutines used give the result of each operation as the correctly rounded standard floating point number. This implies the following relations for the relative errors that are introduced:

$$\begin{aligned}
fl\ (x \pm y) &= x\ (1 + \epsilon) \pm y\ (1 + \epsilon) \\
fl\ (x.y) &= x.y\ (1 + \epsilon) \\
&= x\ (1 + \epsilon) \cdot y \\
&= x\ (1 + \epsilon)^{\frac{1}{2}} \cdot y\ (1 + \epsilon)^{\frac{1}{2}} \\
fl\ (x/y) &= \left[x/y\right] (1 + \epsilon) \\
&= x\ (1 + \epsilon)\ /y \\
&= x/y\ (1 + \epsilon)
\end{aligned}$$

where $|\epsilon| \leq \frac{1}{2}\ 10^{1-t}$ with t being the number of digits being used in the mantissa of the floating point representation. It should be noted that separate uses of an ϵ do not denote, necessarily, the same thing. For example, the meanings of the two ϵ's in the addition equation are related only by the requirement that both be bounded above in magnitude by $\frac{1}{2}\ 10^{1-t}$. Each of the above relations implies that the result of each floating point arithmetic operation on two floating point digital numbers x and y is the exact result for one or two slightly modified numbers.

For an extended product we have

$$fl\ (x_1 \cdot x_2 \cdot \ldots \cdot x_N) = x_1 \cdot x_2 \cdot \ldots \cdot x_N\ (1+\epsilon_1)\ (1+\epsilon_2) \ldots (1+\epsilon_N)$$

with $|\epsilon_i| \leq \frac{1}{2}\, 10^{1-t}$ for i = 1, 2, ... , N. That is, the extended product of N floating point numbers is an exact result for N slightly modified numbers $x_i\,(1 + \epsilon_i)$. Although for an extended sum there are no useful bounds for the expression

$$\left[\frac{\text{fl}\,(x_1 + x_2 + \ldots + x_N)}{x_1 + x_2 + \ldots + x_N} - 1 \right] ,$$

with the additions done naturally from left to right, it is still true that the calculated sum is the exact sum of modified numbers $x_1, x_2, \ldots, x_N$ differing from the corresponding $x_1, x_2, \ldots, x_N$ by small relative errors. To verify this we start by assuming that

$$\text{fl}\,(x_1 + x_2 + \ldots + x_N) = x_1\,(1 + \epsilon_1^{(N)}) + \ldots + x_N\,(1 + \epsilon_N^{(N)})$$

$$= A \text{ , say.}$$

Then

$$\text{fl}\,(x_1 + x_2 + \ldots + x_N + x_{N+1}) = \text{fl}\,(A + x_{N+1})$$

$$= A\,(1 + \epsilon) + x_{N+1}\,(1 + \epsilon)$$

and a little computation shows that

$$(1 - \tfrac{1}{2}\, 10^{1-t})^N \leq 1 + \epsilon_1^{(N+1)} \leq (1 + \tfrac{1}{2}\, 10^{1-t})^N$$

and for i = 2, 3, ... , N+1

$$(1 - \tfrac{1}{2}\, 10^{1-t})^{N+2-i} \leq 1 + \epsilon_i^{(N+1)} \leq (1 + \tfrac{1}{2}\, 10^{1-t})^{N+2-i} .$$

That is,

$$\text{fl}\,(x_1 + x_2 + \ldots + x_{N+1}) = x_1\,(1 + \epsilon_1^{(N+1)}) + x_2\,(1 + \epsilon_2^{(N+1)}) + \ldots$$

$$\ldots + x_N\,(1 + \epsilon_N^{(N+1)}) + x_{N+1}\,(1 + \epsilon_{N+1}^{(N+1)}) .$$

Consequently, if we assume that the operations take place in the natural order, the last equation implies that the calculated sum is the exact result for numbers x_i such that

$$x_i = x_i\,(1 + \epsilon_i^{(N+1)}) \qquad i = 1, 2, \ldots, N+1 .$$

For an inner product calculated in single precision floating point, we have

$$fl\,(x_1 \cdot y_1 + x_2 \cdot y_2 + \ldots + x_N \cdot y_N) =$$

$$x_1 \cdot y_1\,(1 + \epsilon_1^{(N)}) + x_2 \cdot y_2\,(1 + \epsilon_2^{(N)}) + \ldots + x_N \cdot y_N\,(1 + \epsilon_N^{(N)})$$

with

$$(1 - \tfrac{1}{2}\,10^{1-t})^N \leq 1 + \epsilon_1^{(N)} \leq (1 + \tfrac{1}{2}\,10^{1-t})^N$$

and for $i = 2, 3, \ldots, N$,

$$(1 - \tfrac{1}{2}\,10^{1-t})^{N+2-i} \leq 1 + \epsilon_i^{(N)} \leq (1 + \tfrac{1}{2}\,10^{1-t})^{N+2-r} .$$

For an inner product accumulated in double precision and then rounded, it is clear that

$$fl\,(x_1 \cdot y_1 + \ldots + x_N \cdot y_N) = x_1 \cdot y_1\,(1 + \epsilon) + \ldots$$

$$\ldots + x_N \cdot y_N\,(1 + \epsilon)$$

with $|\epsilon| \leq \tfrac{1}{2}\,10^{1-t}$.

For the square root subroutine used, we assume that

$$fl\,(\sqrt{x}) = \sqrt{x}\,(1 + \epsilon_s) \quad \text{with}$$

$|\epsilon_s| \leq 2\,(\tfrac{1}{2}\,10^{1-t}) = 10^{1-t}$. A detailed analysis of the Newton-Ralphson method for the square root is given by Goldstine, Murray and von Neumann [4]. Using an obvious but crude approximation, we may write

$$fl\,(\sqrt{x}) \sim \sqrt{x\,(1 + 2\epsilon_s)}$$ - that is,

the floating square root subroutine gives a result which is, approximately, the exact square root of a slightly modified number.

It can be noticed that in all of the floating point operations considered so far, we can give a "backwards" error analysis in the sense that we can always say that the result of an operation is, at the worst, a close approximation to an exact result for slightly modified numbers, the modification being given in terms of some multiple of a relative error . As a clarifying example, let us consider the operation $fl\,(x_1 + x_2 + x_3)$,

the additions being done in the natural order. We have

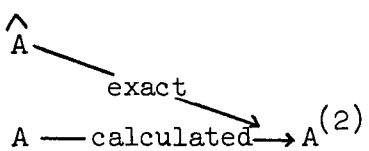
$$fl\,(x_1 + x_2 + x_3) = \left[x_1(1+\epsilon) + x_2(1+\epsilon)\right](1+\epsilon) + x_3(1+\epsilon)$$

$$= x_1(1+\epsilon)^2 + x_2(1+\epsilon)^2 + x_3(1+\epsilon)$$

$$\cong x_1(1+2\epsilon) + x_2(1+2\epsilon) + x_3(1+\epsilon).$$

Consequently, we can claim that the result is very close to an exact result for three modified numbers $x_1(1+2\epsilon)$, $x_2(1+2\epsilon)$, and $x_3(1+\epsilon)$. If we now combine various operations; then, in a rough but natural way, we can give, in many cases, a "backwards" analysis by simply counting the operations. Such an analysis was first attempted for the Householder reduction.

The precise statement of the problem may be illustrated by the first step in the Householder reduction. Let A be the given matrix and $A^{(2)}$ the matrix that results after applying the algorithm to A. Because of rounding errors, $A^{(2)}$ is not an exact result. Using the "backwards" technique, we would like to claim that there exists a symmetric matrix $\hat{A}$ such that if we apply the algorithm to $\hat{A}$ then we get exactly $A^{(2)}$. Schematically, we have



There are two conditions that must be met. First, the elements of $\hat{A}$ are to be related to the corresponding elements of A so that $\hat{a}_{ij}$ differs from a_{ij} by simple multiples of a rounding error. Second, the algorithm must not be altered. There are some difficulties and we now illustrate two of these.

The errors introduced in calculating

$$X_2^2 = \tfrac{1}{2}\left[1 + \frac{a_{12}\ \mathrm{SGN}\ (a_{12})}{S^{\frac{1}{2}}}\right]$$

cannot be attributed to the elements a_{ij} of A in a simple straightforward manner. Let us see why. In accounting for the error made in computing $S^{\frac{1}{2}}$,

we may say that if we used $a_{1i}(1+\epsilon)$ for $i = 2, 3, \ldots, N$ then $\left[\text{calculated } S^{\frac{1}{2}}\right]$ is an exact result for these modified numbers. Let us put $T = \dfrac{a_{12}\,\mathrm{SGN}(a_{12})}{\left[\text{calculated } S^{\frac{1}{2}}\right]}$. Since the $\frac{1}{2}$ and 1 occurring in the formula for X_2^2 is part of the algorithm, we want to attribute all errors made in calculating $X_2^2 = \frac{1}{2}\left[1+T\right]$ to the quantity T - i.e. calculated $X_2^2 = \frac{1}{2}\left[1+\tilde{T}\right]$. The errors occur when we divide $|a_{12}|$ by $\left[\text{calculated } S^{\frac{1}{2}}\right]$, then when this result is added to 1 and finally, when the last result is multiplied by $\frac{1}{2}$. The first and last operations give rise to the following problem. If we modify a_{12} to account for these errors then we have

$$\frac{a_{12}(1+\epsilon)(1+\epsilon)}{a_{12}(1+\epsilon)^2+\ldots.},$$

and we want to claim that the computation is exact for a_{12} so modified. Since the three ϵ's may stand for three different nonzero quantities, some objection to the claim for exactness may be made. We do not know whether this objection is valid or not.

From the expression for X_2^2, it is clear that $T \leq 1$. Because of this fact, it is no longer true that the error made in adding $\left[\text{calculated } T\right]$ to 1 may be attributed to $\left[\text{calculated } T\right]$ in such a fashion that the error is bound by an ϵ. As an example, suppose we are carrying 4 digits and that $\dfrac{|a_{12}|}{\left[\text{calculated } S^{\frac{1}{2}}\right]}$ is .0009. Then the exact result is 1.0009, whereas the calculated result is 1.001. Thus, the error is .0001 and we want to say that $.0001 = .0009(1+\epsilon)$, where $|\epsilon| \leq 5.10^{-4}$. In fact, $\epsilon = -8/9$. Because of these difficulties, the "backwards" error analysis was abandoned.

Error bounds for the Householder Algorithm

The principle of Ortega's analysis is based upon the following observations of Wilkinson [15]. Using the notation introduced in describing Householder's method let $\lambda_i^{(r)}$ be the eigenvalues of $\left[\text{calculated } A^{(r)}\right]$ with $\lambda_i^{(1)}$ the eigenvalues of $A^{(1)} \equiv A$. Let $\left|\lambda_i^{(r)} - \lambda_i^{(r+1)}\right| \leq \delta_i^{(r)}$. Then for the triple-diagonal matrix $\left[\text{calculated } A^{(N-1)}\right]$, we have that

$$\left|\lambda_i^{(1)} - \lambda_i^{(N-1)}\right| \leq \sum_{r=1}^{N-2} \left|\lambda_i^{(r)} - \lambda_i^{(r+1)}\right| \leq \sum_{r=1}^{N-2} \delta_i^{(r)} \quad \text{, for all } i \text{ .}$$

The problem now becomes one of obtaining bounds for $\delta_i^{(r)}$. In order to do this, let Q_{r+1} be the exact orthogonal matrix which would be derived by the Householder algorithm applied to $\left[\text{calculated } A^{(r)}\right]$. For $r = 1, 2, \ldots, N-2$, we define

$$E_r = \left[\text{calculated } \left(P_{r+1} \left[\text{ calculated } A^{(r)}\right] P_{r+1}\right)\right] - \left[\text{exact } \left(Q_{r+1} \left[\text{calculated } A^{(r)}\right] Q_{r+1}\right)\right]$$

$$= \left[\text{calculated } A^{(r+1)}\right] - \left[\text{exact } \left(Q_{r+1} \left[\text{calculated } A^{(r)}\right] Q_{r+1}\right)\right] .$$

Then by definition, the eigenvalues of $\left[\text{calculated } A^{(r+1)}\right]$ are $\lambda_i^{(r+1)}$ and, by definition and similarity, the eigenvalues of exact $\left(Q_{r+1} \left[\text{ cal } A^{(r)}\right] Q_{r+1}\right)$ are $\lambda_i^{(r)}$. Thus, by Lidskii's theorem

$$\left|\lambda_i^{(r+1)} - \lambda_i^{(r)}\right| \leq \text{ max eigenvalue of } E_r \text{ ,}$$

and now the problem is to find bounds for the elements of E_r, $r = 1,2,\ldots, N-2$.

Ortega obtains bounds for $\max_i \left|\lambda_i^{(1)} - \lambda_i^{(n-1)}\right|$ relative to both the spectral norm of A and the Euclidean norm of A, where these norms are defined by:

$$\|A\|_S = \left(\max_i \Lambda_i\right)^{\frac{1}{2}}$$

with Λ_i being the eigenvalues of $A^t A$, and

$$\|A\|_E = \left(\sum_{i,j=1}^{n} a_{ij}^2\right)^{\frac{1}{2}}.$$

The thoroughness of Ortega's analysis may be illustrated by the following points. Since the norms; either spectral or Euclidean, of $\left[\text{calculated } A^{(1)}\right], \ldots, \left[\text{calculated } A^{(n-1)}\right]$ do not necessarily remain that of A, Ortega makes a study of the growth of the norms of $\left[\text{calculated } A^{(r)}\right]$ in terms of the norm of A. He also carries all higher order terms (i.e. terms involving at least ϵ^2) until the final stage and then finds a bound for them. Finally, he considers both normal and double precision floating point inner products for vectors. Before we give Ortega's results, a few words on his notation are in order. He denotes by m_b, m_s, and m_p bounds for the relative errors* in the basic arithmetic operations, square roots, and inner products, respectively.

The results for the spectral norm are:

If $m_s \leq 2m_b$, $Nm_b \leq 10^{-4}$, $N^2 m_b \leq 10^{-3}$, $N^{5/2} m_b \leq 10^{-2}$,

where N is the order of A, and ϵ is the maximum error in any eigenvalue then:

$$\frac{|\epsilon|}{\|A\|_S} \leq \frac{55(N-2)m_s + (3.2\,N^{5/2} + 9.75\,N^2 + 6.0\,N^{3/2} + 157.0\,N - 397)m_b}{1 - 55(N-2)m_s - (3.2\,N^{5/2} + 9.75\,N^2 + 6.0\,N^{3/2} + 157.0N - 397)m_b},$$

for normal floating point inner products.

If $m_s \leq 2m_b$, $Nm_b \leq 10^{-4}$, $N^{3/2} m_b \leq 10^{-2}$ then

$$\frac{|\epsilon|}{\|A\|_S} \leq \frac{55(N-2)m_s + (6.0\,N^{3/2} + 161.1\,N - 348.7)m_b}{1 - 55(N-2)m_s + (6.0\,N^{3/2} + 161.1\,N - 348.7)m_b}$$

for accumulated inner products.

The corresponding results for the Euclidian norm are:

$$\frac{|\epsilon|}{\|A\|_E} \leq \frac{55.5(N-2)m_s + (13.9\,N^2 + 160.9N - 378)m_b}{1 - 55.5(N-2)m_s + (13.9\,N^2 + 160.9\,N - 378)m_b}$$

for normal floating point inner products; and,

$$\frac{|\epsilon|}{\|A\|_E} \leq \frac{55.5(N-2)m_s + 174.8\,(N-2)m_b}{1 - 55.5\,(N-2)m_s + 174.8\,(N-2)m_b},$$ for accumulated inner products.

(* From our earlier discussion we should keep in mind that for addition and inner products we do not generally have true relative errors.)

Denoting the right hand side of these bounds by $F(N, m_b, m_s)$, Ortega has prepared the following tables for comparisons.

Table 1 [10; p.38]

$F(N, m_b, m_s)$ for $m_b = 5 \times 10^{-12}$ and $m_s = 2m_b$

(Spectral Norm)

N	F (N, m_b , m_s)	
	Normal	Accumulated
10	2.10×10^{-8}	1.18×10^{-8}
30	1.63×10^{-7}	4.31×10^{-8}
50	4.82×10^{-7}	7.65×10^{-8}
100	2.25×10^{-6}	1.64×10^{-7}
200	1.13×10^{-5}	3.56×10^{-7}
500	1.03×10^{-4}	1.03×10^{-6}
1000	5.50×10^{-4}	2.31×10^{-6}

Table 2 [10; p.38]

$F(N, m_b, m_s)$ for $m_b = 5 \times 10^{-8}$ and $m_s = 2m_b$

(Spectral Norm)

N	F (N, m_b , m_s)	
	Normal	Accululated
10	2.10×10^{-4}	1.18×10^{-4}
30	1.63×10^{-3}	4.31×10^{-4}
50	4.82×10^{-3}	7.65×10^{-4}
100	2.25×10^{-2}	1.64×10^{-3}
200	——*	3.56×10^{-3}
500	——*	1.03×10^{-2}
1000	——*	2.35×10^{-2}

* No figures of accuracy

Table 3 [10; p.53]

$F(N, m_b, m_s)$ for $m_b = 5 \times 10^{-12}$ and $m_s = 2m_b$

(Euclidean Norm)

N	$F(N, m_b, m_s)$	
	Normal	Accumulated
10	1.76×10^{-8}	1.14×10^{-8}
30	1.01×10^{-7}	4.02×10^{-8}
50	2.43×10^{-7}	6.90×10^{-8}
100	8.32×10^{-7}	1.41×10^{-7}
200	3.07×10^{-6}	2.84×10^{-7}
500	1.82×10^{-5}	7.15×10^{-7}
1000	7.10×10^{-5}	1.44×10^{-6}

Another practical use of the bound $F(N, m_b, m_s)$ is illustrated in the following table.

Table 4 [10; p.40]

Maximum allowable N so that $F(N, m, m) < \delta$ for $m_b = 5 \times 10^{-12}$, $m_s = 2m_b$

δ	N	
	Normal	Accumulated
10^{-7}	23	64
10^{-6}	69	490
10^{-5}	190	3.2×10^3
10^{-4}	490	1.8×10^4
10^{-3}	1270	9.4×10^4

Error bounds for the Bisection Process

To complete the error analysis, we now obtain error bounds for the computed eigenvalues of the symmetric triple-daigonal matrix C. The analysis is essentially that of Wilkinson [14; pp.324-326]. We recall that the elements of C were sealed so that for all i $|\alpha_i| < 1$ and $|\beta_i| < 1$; also, none of the β_i's were zero. Referring to the description of the Sturm sequence bisection method, we see that a sequence $\langle f_0, f_1, \ldots, f_N \rangle$ is calculated. We shall show that this sequence is an exact sequence for a modified matrix $(C' - \lambda I)$. Hence, using Lidskii's theorem, we obtain a bound for $\lambda'_i - \lambda_i$ where λ'_i and λ_i are the eigenvalues of C' and C respectively.

Since $f_r(\lambda) = (\alpha_r - \lambda) f_{r-1}(\lambda) - \beta_r^2 f_{r-2}(\lambda)$

for any trial value λ, we have that

$$fl\left[f_r(\lambda)\right] = \left[\alpha_r(1+\epsilon)(1+\epsilon)(1+\epsilon) - \lambda(1+\epsilon)(1+\epsilon)(1+\epsilon)\right] . f_{r-1}(\lambda) -$$

$$- \beta_r^2 \quad (1+\epsilon)(1+\epsilon)(1+\epsilon) f_{r-2}(\lambda)$$

where $|\epsilon| \leq 5.10^{-t}$. If we assume that $fl(\alpha_r - \lambda)$ is not zero, then the corresponding modified elements α'_i of C' satisfy for all λ, $\alpha'_r - \lambda = (\alpha_r - \lambda)(1+\epsilon)(1+\epsilon)(1+\epsilon)$. In case $fl(\alpha_r - \lambda) = \alpha_r(1+\epsilon_1) - \lambda(1+\epsilon_2) = 0$, we take α'_r to be either α_r or $\alpha_r(1+\epsilon)$. The former in case $\alpha_r = \lambda$ or $\epsilon_1 = 0$; the latter, if $\epsilon_2 = 0$. The β'_i's satisfy $(\beta'_r)^2 = \beta_r^2(1+\epsilon)(1+\epsilon)(1+\epsilon)$.

Because of the scaling, all eigenvalues λ of C satisfy $|\lambda| \leq 3$, since the relation $|\lambda| \leq \max_i \sum_j |a_{ij}|$ holds for any matrix A. Consequently, $|\alpha_r - \lambda| \leq 4$ for all trial values of λ. Thus, $|\alpha'_r - \alpha_r| = \left|(\alpha_r - \lambda)\left\{(1+\epsilon)(1+\epsilon)(1+\epsilon) - 1\right\}\right|$

$$\leq 4\left\{(1+5.10^{-t})^3 - 1\right\} = \delta_1$$

even if $fl(\alpha_r - \lambda) = 0$; and

$$|\beta'_r - \beta_r| = \left|\beta_r\left\{(1+\epsilon)^{\frac{1}{2}}(1+\epsilon)^{\frac{1}{2}}(1+\epsilon)^{\frac{1}{2}} - 1\right\}\right|$$

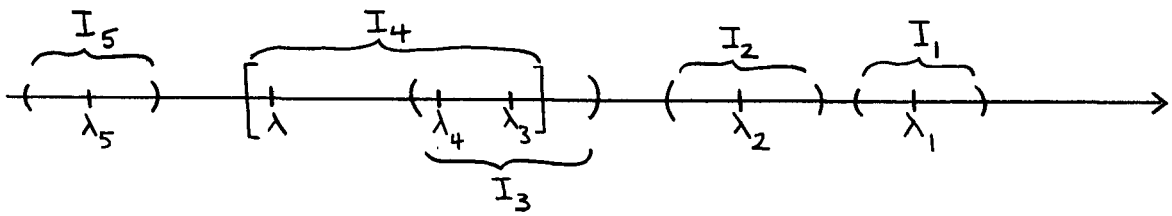
$$\leq 1 \cdot \left[(1+5.10^{-t})^{3/2} - 1\right] = \delta_2.$$

Hence, the eigenvalues of C'-C are bound by $\delta_1 + 2\,\delta_2$ - that is,

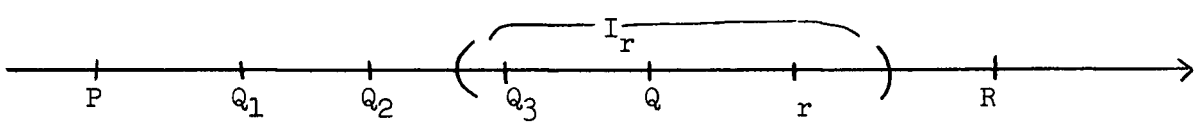
$$|\lambda_i' - \lambda_i| \leq \delta_1 + 2\,\delta_2 .$$

This implies that the eigenvalues λ_i' are in intervals of width $2(\delta_1 + 2\delta_2)$ centered about λ_i .

Let us now consider the Sturm sequence decision process. For any λ , the computed $A(\lambda)$ corresponds to some C' instead of C, but for a λ value outside the above intervals, $A(\lambda)$ is a correct result for C itself. Moreover, we necessarily obtain the correct answer to the question "are there less than r roots greater than ?", if λ is not in the interval about λ_r. This is still true even if some intervals overlap - for example, in the diagram below, $A(\lambda) = 3$ or 4 whenever the λ_i' lie in their permitted regions.



Further consideration of the bisection technique shows that one of the following necessarily happens: (a) we obtain the correct decision at all steps and consequently λ_r really does lie in the interval terminating the bisection process; or, (b) there exists a first step at which the wrong answer is given. By our earlier remarks, this must occur when $A(\lambda)$ is evaluated for λ in some I_r. To make the discussion specific, suppose, as in the following diagram, that the first wrong decision is made at Q, so that the r-th root, λ_r, is placed in PQ instead of QR.



From the description of the bisection technique, it is clear that the process will proceed to Q_1, then to Q_2 - that is, it will move toward λ_r . The next wrong step will occur for a bisection point in I_r - at Q_3 in the diagram.

At Q_3, the r-th root will be placed in Q_2Q_3 or Q_3Q. Thus, at all subsequent stages, the r-th root is placed either in an interval entirely in I_r or in an interval whose right hand end point is in I_r.

Therefore, the center point λ of the final interval in which λ_r is placed satisfies $|\lambda_r - \lambda| \leq 10^{-t} + \delta_1 + 2\delta_2$

$$\cong 10^{-t} + 4\left[3.5.10^{-t}\right] + 2\left[\frac{7}{2} .10^{-t}\right]$$

$$= 68\ 10^{-t} ,$$

since the length of the interval terminating the bisection process does not exceed 2.10^{-t}. The above bound is noteworthy because it is independent of N, the order of the matrix C, and of the root separations.

Conclusion

We have chosen to concentrate on Householder's method because it has certain advantages over the others mentioned above. The big advantage over the Jacobi method (and its variations) is, as has been already mentioned, that Householder's reduction requires a finite number of steps whereas the Jacobi reduction is iterative. This means that we do not have to be concerned with proofs of convergence and rates of convergence. It can also be claimed that Householder's method is more efficient than Givens' or Lanczos' in the sense that fewer multiplications are required. For example, Wilkinson [15; p.25] states that Householder's method requires approximately $2/3N^3$ multiplications; Givens' requires approximately $4/3N^3$; and, Lanczos', with reorthogonalizations, requires approximately $2N^3$. Moreover, there are only 2N square roots in the Householder method compared to $\frac{1}{2}N^2$ in the Givens' method. Heurestically, this advantage in the number of multiplications means that Householder's reduction should yield more accurate results than either Givens' or Lanczos'. However, the major advantage that is hoped to be gained is in the application of Householder's algorithm to the unsymmetric eigenvalue problem. In a recent

paper [18], Wilkinson compared Givens' method with one which used elementary similarity transformations and found that on the matrices (up to order 30) that were tested, the elementary transformation method was just as accurate. But he points out [15; p.26] that the error analysis indicated that the unitary transformations are more stable numerically. Also, C.T. Fike's paper [1], mentioned before, indicates that such should be the case. Consequently, since Householder's reduction retains the mentioned advantages even on the unsymmetric case, it may turn out to be a very important method. Research along these lines is planned for the future.

BIBLIOGRAPHY

[1] Fike, C.T., Note on the practical computation of proper values, J. Assoc. Comp. Mach., vol.6, 1959, pp.360-362.

[2] Forsythe, G. and Henrici, P., The cyclic Jacobi method for computing the principal values of a complex matrix, Trans. Am. Math. Soc., vol.94, no.1, Jan.1960, pp.1-23.

[3] Givens, W., Numerical computation of the characteristic values of a real symmetric matrix, Oak Ridge National Laboratory Report No.1574, 1954.

[4] Goldstine, H.H., Murray, F.J., and von Neumann, J., The Jacobi method for real symmetric matrices, J. Assoc. Comp. Mach., vol.6, Jan. 1959, pp.60-66.

[5] Householder, A.S., Hedrick lectures delivered at the summer meeting of the Mathematical Association of America, August, 1958, cited in Fike, C.T.

[6] Householder, A.S., and Bauer, F.L., On certain methods for expanding the characteristic polynomial - Numerische Matematik, vol. 1, no.1, 1959, pp.29-37.

[7] Jacobi, C.G.J., Uber ein leichtes Verfahren, die in der theorie der säkular störungen vorkommenden gleichungen numerisch augzulösen, J. Reine Angew. Math., 30 (1846), pp.51-95, cited in White, P.A.

[8] Lanczos, C., An iteration method for the solution of the eigenvalue problem of linear differential and integral operators, J. Res. N.B.S, vol.45, 1950, pp.255-282.

[9] Lanczos, C., Applied analysis, Prentice Hall, New Jersey, 1956, pp.49-51.

[10] Ortega, J.M., An error analysis of Householder's method for the symmetric eigenvalue problem, Tech. Report No.18, Appl. Math. Stat. Lab., Stanford University, 1962.

[11] Ortega, J.M., On Sturm sequences for tridiagonal matrices, J. Assoc. Comp. Mach., vol.7, July 1960, pp.260-263.

[12] Pope, D. and Tompkins, C., Maximizing functions of rotation, J. Assoc. Comp. Mach., vol.4, 1957, pp.459-466.

[13] White, P.A., The computation of eigenvalues and eigenvectors of a matrix, J. Soc. Ind. Appl. Math., vol.6, 1958, pp.393-403.

[14] Wilkinson, J.H., Error analysis of floating point computation, Numerische Matematik, vol.2, no.5, 1960, pp.319-340.

[15] Wilkinson, J.H., Householder's method for the solution of the algebraic eigenproblem, The Computer Journal, vol.3, April 1960, pp.23-28.

[16] Wilkinson, J.H., Notes on practical methods of solving linear systems and calculating the eigensystems of matrices, National Physical Laboratory, Teddington, England, 1959.

[17] Wilkinson, J.H., Theory and practice in linear systems and the determination of characteristic values and characteristic vectors, Summer Session (1958) Univ. of Michigan, Ann Arbor, Mich., pp.111-152.

[18] Wilkinson, J.H., Stability of the reduction of a matrix to almost triangular and triangular forms by elementary similarity transformations, J. Assoc. Comp. Mach., vol.6, 1959, pp.336-359.