CS 59

ALMOST DIAGONAL MATRICES WITH MULTIPLE OR CLOSE EIGENVALUES

BY

J. H. WILKINSON

TECHNICAL REPORT NO. CS 59
APRIL 3, 1967

This work was supported by the
National Science Foundation and the
Office of Naval Research

COMPUTER SCIENCE DEPARTMENT
School of Humanities and Sciences
STANFORD UNIVERSITY

ALMOST DIAGONAL MATRICES

WITH MULTIPLE OR CLOSE EIGENVALUES

by

J. H. Wilkinson*

ABSTRACT

If $A = D + E$ where D is the matrix of diagonal elements of A, then when A has some multiple or very close eigenvalues E has certain characteristic properties. These properties are considered both for hermitian and non-hermitian A. The properties are important in connexion with several algorithms for diagonalizing matrices by similarity transformations.

*Mathematics Division, National Physical Laboratory, Teddington, Middlesex, England, and Computer Science Department, Stanford University.

This work was supported by N.S.F. and O.N.R.

1. Introduction:

In a number of algorithms for finding eigenvalues of a matrix A_1, the latter is reduced by an iterative sequence of similarity transformations to almost diagonal form. When Al has a multiple eigenvalue this is true of all the transforms (assuming exact computation). We are interested then in the nature of almost diagonal matrices with multiple eigenvalues. It turns out that such matrices have special characteristics which are of considerable interest as regards the convergence of iterative procedures for reducing a matrix to diagonal form.

2. The Hermitian Case:

We first consider hermitian matrices with multiple eigenvalues. Let A be hermitian with eigenvalues $\lambda_1, \lambda_1, \ldots, \lambda_1, \lambda_{r+1}, \lambda_{r+2}, \ldots, \lambda_n$ the root λ_1 being precisely of multiplicity r. (A may have other multiple eigenvalues but this will not affect the argument). Let δ be defined by the relation

$$3\delta = \min_{i = r+1}^{n} |\lambda_i - \lambda_1| \qquad (2.1)$$

and let

$$A = D + E \qquad (2.2)$$

where D is the diagonal of A . Suppose we have

$$\|E\|_F = \epsilon < \delta \text{ (where } F \text{ denote the Frobenius norm } (\Sigma\Sigma|e_{ij}|^2)^{\frac{1}{2}} \qquad (2.3)$$

so that when ϵ is small A may be regarded as almost diagonal. By the Wielandt-Hoffman theorem the λ_i and a_{ii} may be ordered so that

$$\Sigma(\lambda_{p_i}-a_{ii})^2 \leq \epsilon^2 \tag{2.4}$$

Let us permute the rows and columns of A similarly so that the a_{11} associated with the λ_1 eigenvalues are the first r. Without loss of generality we can assume this was true originally and with appropriate numbering of the remaining n-r eigenvalues inequality (2) becomes

$$\sum_1^n(\lambda_i-a_{ii})^2 \leq \epsilon^2 . \tag{2.5}$$

We write

$$A = \begin{bmatrix} F & G \\ G^T & H \end{bmatrix} \tag{2.6}$$

where F is an rxr matrix.

If the eigenvalues of H are $\lambda'_{r+1},\ldots,\lambda'_n$ then since the off-diagonal elements of H are a subset of those of E , we have by the Wielandt-Hoffman theorem [4] with appropriate numbering of the λ'_i

$$\sum_{r+1}^n (\lambda'_i-a_{ii})^2 \leq \epsilon^2 . \tag{2.7}$$

Hence

$$\begin{aligned} |\lambda'_i-\lambda_i| &= |\lambda'_i-a_{ii}+a_{ii}-\lambda_i| \\ &\leq \epsilon+\epsilon = 2\epsilon < 2\delta \end{aligned} \tag{2.8}$$

and

$$\begin{aligned} |\lambda'_i-\lambda_1| &= |\lambda_i-\lambda_1+\lambda'_i-\lambda_i| \\ &\geq |\lambda_i-\lambda_1| - |\lambda'_i-\lambda_i| \\ &\geq 3\delta - 2\delta \\ &= \delta \end{aligned} \tag{2.9}$$

The matrix $H-\lambda_1 I$ is therefore non-singular, i.e. it is of rank $n-r$. Now since A has λ_1 as a r-fold root it, too, is of rank $n-r$. We shall show that this means that F is especially related to G and H . We partition $A-\lambda_1 I$ in the form

$$A-\lambda_1 I = \begin{bmatrix} F-\lambda_1 I & G \\ G^T & H-\lambda_1 I \end{bmatrix} \qquad (2.10)$$

If we premultiply $A-\lambda_1 I$ by

$$\begin{bmatrix} I & -G(H-\lambda_1 I)^{-1} \\ \theta & I \end{bmatrix} \qquad (2.11)$$

its rank is unaltered and hence the derived matrix

$$\begin{bmatrix} F-\lambda_1 I-G(H-\lambda_1 I)^{-1}G^T & \theta \\ G^T & H-\lambda_1 I \end{bmatrix} \qquad (2.12)$$

is also of rank $n-r$. Since $H-\lambda_1 I$ is already of rank $n-r$ this can be true only if

$$F-\lambda_1 I-G(H-\lambda_1 I)^{-1}G^T = \theta \qquad (2.13)$$

i.e. $$F = \lambda_1 I+G(H-\lambda_1 I)^{-1}G^T = \lambda_1 I+M \text{ (say)} \qquad (2.14)$$

Now the elements of G are a subset of those of E and hence

$$\|G\|_E = \|G^T\|_E \leq \epsilon \qquad (2.15)$$

while

$$(H-\lambda_1 I)^{-1} = R \text{ diag } (\lambda'_i-\lambda_1)^{-1} R^H \qquad (2.16)$$

where R is unitary. Hence from the unitary invariance of the Frobenius norm and from (2.9) and (2.15)

$$\|M\|_F \le \|G\|_E \max|(\lambda'_i - \lambda_1)^{-1}| \ \|G^T\|_E$$
$$\le \epsilon^2/\delta \qquad (2.17)$$

We see then that the diagonal elements of F differ from λ_1 by quantities bounded by ϵ^2/δ and its off-diagonal elements are bounded by ϵ^2/δ .

When $\epsilon << \delta$ this means that the largest off-diagonal element of A is never found in F , the matrix with the diagonal elements "associated" with the multiple root λ_1, This has important consequences in connection with the classical Jacobi method [5,9, 14] for diagonalizing hermitian matrices. At each stage in the reduction the largest off-diagonal element in the current matrix is annihilated but theorem shows that after a certain stage such an off-diagonal element is never 'associated' with two elements tending to the same multiple root.

This simple observation removes a difficulty in demonstrating that the classical Jacobi method is always ultimately quadratically convergent [6,9, 10, 14]. A similar remark applies to the serial Jacobi method if a-threshold strategy is used [8]. If at any stage the element which is annihilated is chosen to be one which is not small compared with the current norm of off-diagonal elements then this ensures that from a certain stage the annihilated element will not be associated with two diagonal elements tending to the same multiple root.

3. Pathologically close roots:

In practice when a transformation is made on a matrix having multiple roots, the transformed matrix merely has very close roots because of rounding errors. In discussing the convergence of the Jacobi method the quantity $\min|\lambda_i-\lambda_j|$ is of great importance and the presence of very close roots would appear to be serious. We now show that this is not so.

Suppose the roots of A are

$$\lambda_1,\lambda_2,\ldots,\lambda_r \; ; \; \lambda_{r+1},\ldots,\lambda_n, \qquad (3.1)$$

where

$$\lambda_i = \lambda + \epsilon_i \qquad i = 1,\ldots r \qquad (3.2)$$

and the ϵ_i are very small. The first r roots are therefore pathologically close. Define D and E as in (2.2), but δ by the relation

$$3\delta = \min_{i=r+1}^{n} |\lambda_i-\lambda| \qquad (3.3)$$

and assume that

$$\|E\|_F + (\Sigma\epsilon_i^2)^{\frac{1}{2}} = \epsilon < \delta \ . \qquad (3.4)$$

Now A may be expressed in the form

$$A = R\, D_1\, R^H , \qquad (3.5)$$

where $D_1 = \mathrm{diag}(\lambda_i)$ and D_1 can be separated into D_2 and D_3 where

$$D_2 = \mathrm{diag}(\lambda,\lambda,\ldots,\lambda,\lambda_{r+1},\ldots,\lambda_n) \qquad (3.6)$$

$$D_3 = \mathrm{diag}(\epsilon_1,\epsilon_2,\ldots,\epsilon_r,0,\ldots,0) . \qquad (3.7)$$

Hence

$$A = R(D_2+D_3)R^H = R\ D_2\ R^H + R\ D_3\ R^H = B + C \text{ (say)} \qquad (3.8)$$

The matrix B has λ as an r-fold root and to apply the result of the previous section we require only a bound for the Frobenius norm of its off-diagonal elements. Since B = A-C such a bound is given by

$$\|E\|_F + \|C\|_F = \|E\|_F + (\Sigma\epsilon_i^2)^{\frac{1}{2}} = \epsilon < \delta \ .$$

The Frobenius norm of the off-diagonal elements of B 'associated' with the multiple root is therefore bounded by ϵ^2/δ and hence that of the corresponding elements of A is bounded by $\epsilon^2/\delta + (\Sigma\epsilon_i^2)^{\frac{1}{2}}$.

Suppose for example a matrix A has the roots

$$1\text{-}10^{-10},\ 1,\ 1+10^{-10},\ 2,\ 3, 4$$

and

$$\|E\|_F + 23\ 10^{-10} < 10^{-5} \ .$$

The off-diagonal elements of A associated with the close roots will then have a Frobenius norm bounded by

$$10^{-10} + 2^{\frac{1}{2}}\ 10^{-10}$$

and therefore they will-all be far smaller than the largest off-diagonal element of A . Hence at such a stage in the classical Jacobi method or the threshold serial Jacobi method with a matrix having the root distribution above, the current rotation will not be in a plane associated with the close roots. In fact with the above example one sweep of the threshold serial Jacobi method will reduce the norm of off-diagonal elements from 10^{-5} to 10^{-10}. Provided we do not wish to reduce the norm below this level the presence of the close roots has no adverse influence. (In fact it is beneficial since it ensures that the main weight in the off-diagonal positions is concentrated on fewer elements).

The above results were known to the author as early as 1963 and were used in [14, 15] to establish the cubic convergence of the symmetric Q,R algorithm for a matrix having multiple roots Unfortunately I failed to observe the obvious consequences of the theorem in connection with Jacobi's method.

4. Non-hermitian matrices:

The above proofs may give the impression that the result above is associated specifically with Hermitian matrices. In fact a closely related result is true for any matrix having an r-fold root corresponding to linear divisors, We restrict ourselves to the case when the remaining eigenvalues are distinct though a slightly weaker result can be proved if some of them are multiple eigenvalues.

Again let A have the roots $\lambda_1,\ldots,\lambda_1,\lambda_{r+1},\ldots,\lambda_n$ the first r corresponding to linear divisors. Let

$$A = D + E \text{ (D diagonal)} \tag{4.1}$$

$$\min_{i=r+1}^{n} |\lambda_i-\lambda_j| = 3\delta \tag{4.2}$$

$$\|E\|_\infty=\epsilon < \delta \tag{4.3}$$

Then by Gerschgorin's theorem all eigenvalues of A lie in discs with centres a_{ii} and radii not greater than ϵ. This implies that there must be at least one a_{jj} in each of the discs centred on the λ_i and of radii ϵ. Since these discs are disjoint this means in particular that we can associate an a_{jj} with each of the $\lambda_i (i = r+1,\ldots,n)$. If more than one a_{jj} is associated with any such λ_i we take the first one occuring on the diagonal. Now permute rows and columns of A similarly so that these n-r diagonal elements are in the southeast corner, We can assume that A was in this form originally.

As in the symmetric case $A-\lambda_1 I$ is of rank n-r and partitioning $A-h_1 I$ in the form

$$A-\lambda_1 I = \begin{bmatrix} F-\lambda_1 I & G \\ K & H-\lambda_1 I \end{bmatrix} \qquad (4.4)$$

we have

$$F-\lambda_1 I - G(H-\lambda_1 I)^{-1}K = \theta \qquad (4.5)$$

provided $H-\lambda_1 I$ is non-singular. Now

$$H-\lambda_1 I = \text{diag}(a_{ii}-\lambda_1)+L = D_1+L \text{ (say)} \qquad (4.6)$$

where L is the matrix of off-diagonal elements of H. (These elements are a subset of those of E). Since

$$|a_{ii}-\lambda_1| = |a_{ii}-\lambda_i+\lambda_i-\lambda_1|$$
$$\geq |\lambda_i-\lambda_1|-|a_{ii}-\lambda_i|$$
$$\geq 3\delta - \epsilon \quad (i = r+1,\ldots,n)$$
$$> 2\delta \tag{4.7}$$

D_1 is non-singular and

$$H-\lambda_1 I = D_1 [I + D_1^{-1}L] . \tag{4.8}$$

Now

$$\|D_1^{-1}L\|_\infty \leq \|D_1^{-1}\|_\infty \|L\|_\infty$$
$$\leq \max_{i=r+1}^{n} |a_{ii}-\lambda_1|^{-1} \|L\|_\infty$$
$$\leq \epsilon/2\delta < \tfrac{1}{2} \tag{4.9}$$

and hence

$$\|(H-\lambda_1 I)^{-1}\|_\infty \leq \|D_1^{-1}\|_\infty /(1 - \|D_1^{-1}L\|_\infty)$$
$$\leq \frac{1}{2\delta} / 3$$
$$= \frac{1}{\delta} \tag{4.10}$$

Equation (4.5) therefore gives

$$F = \lambda_1 I + G (H-\lambda_1 I)^{-1}K = \lambda_1 I+M \text{ (say)}$$

where

$$\|M\|_\infty \leq \|G\|_\infty \|(H-\lambda_1 I)^{-1}\|_\infty \|K\|_\infty$$
$$\leq \epsilon^2 / \delta \tag{4.11}$$

This now shows that there was in fact only one a_{jj} associated with each of the λ_i ($i= r+1,\ldots,n$) and the remaining r diagonal elements are all in a disc of radius ϵ^2/δ centred on λ_1. Again off-diagonal elements 'associated' with the multiple eigenvalue are bounded by ϵ^2/δ and are therefore well below the level of the largest off-diagonal elements when $\epsilon << \delta$.

The result is at first sight surprising since the condition of the eigenvalue problem of A seems not to be involved. Indeed a result may be proved which is only marginally weaker even when A is defective (though not as far as λ_1 is concerned). In this respect it is the hypothesis $\|E\|_\infty \leq \epsilon$ which is deceptive. If B has an ill-conditioned eigenvalue problem then in order to derive a similarity transformation $X^{-1}B X = A$ such that A is almost diagonal with $\|E\|_F$ less than a prescribed quantity we shall, in general, have to work to higher precision if B is ill-conditioned than if it is well-conditioned. In the hermitian case the hypothesis does not have this deceptive feature.

5. Pathologically close roots in non-hermitian case:

The deceptive nature of the result becomes apparent as soon as we consider the effect of very close roots, Assume now that A is non-defective and let X be a matrix having as its columns n independent eigenvectors of A . Then we have

$$A = X \, \mathrm{diag}(\lambda_i) X^{-1} \tag{5.1}$$

Using a similar notation to that in paragraph 3 we have in the case of r very close roots

$$A = X D_2 X^{-1} + X D_3 X^{-1} = B + C \tag{5.2}$$

where B now has an r-fold root, In the hermitian case X is unitary and $\|C\|_F = \|D_3\|_F$, but now all we can say is

$$\|C\| \le \|X\| \, \|D_3\| \, \|X^{-1}\| \tag{5.3}$$

and we see that the condition number κ of X with respect to inversion is inevitably involved. It is clear that it is the minimum value of $\|X\| \, \|X^{-1}\|$ for all permissible X that is relevant [1]. It should be emphasized though, that the possession of a multiple root or of a set of very close roots does not imply that $\|X\| \, \|X^{-1}\|$ is necessarily large. Provided the close roots are well-conditioned the fact that the eigenvector problem is ill-conditioned is irrelevant.

6. Iterative refinement of an eigensystem:

The above results have important consequences in connection with procedures for the refinement of a computed eigensystem of a matrix [11, 12, 143. In such procedures one starts with a computed set of eigenvalues and eigenvectors μ_i and x_i . Let X be the matrix having columns x_i and define R and S by the relation

$$AX - X\,\mathrm{diag}(\mu_i) = R \tag{6.1}$$

$$X^{-1}AX - \mathrm{diag}(\mu_i) = X^{-1}R = S \ . \tag{6.2}$$

If the system were exact both R and S would be null. In practice neither R nor S can be computed exactly with the given X because of rounding errors but with well-designed procedures 8 is determined with a low relative error. Hence we have

$$X^{-1}AX = \mathrm{diag}(\mu_i) + \bar{S} + (S-g) \tag{6.3}$$

and if the computed system is accurate $\bar{S}$ is small, and with good procedures for calculating R and $X^{-1}R$ a bound is obtained for $\|S-\bar{S}\|$ which is small compared with $\|\bar{S}\|$ (note $\bar{S}$ is computed explicitly but a bound for the norm only is determined for $S-\bar{S}$). The matrix sum on the right of (6.3) is therefore an almost diagonal matrix which is exactly similar to A . Now when A has a multiple root corresponding to a linear divisor our result shows that provided $\bar{S}$ is small (and hence $S-\bar{S}$ is very small), the off-diagonal elements of $\bar{S}$ associated with the multiple roots will be far smaller than the largest off-diagonal elements of $\bar{S}$. When none of the roots of A is ill-conditioned we shall find typically that if $\|\bar{S}\|_\infty = \epsilon$ then the bound for $\|S-\bar{S}\|$, will be approximately $2^{-t}\epsilon$ (with a t-digit mantissa binary computer), The diagonal elements of $\mathrm{diag}(\mu_i) + \bar{S}$ associated with the multiple roots will differ by quantities of the order of ϵ^2 and the associated off-diagonal elements will be of order ϵ^2. Hence after suitable permutations of rows and columns the right hand side-of (6.3) will have the form

$$\text{Diagonal} + r\{\begin{bmatrix} \epsilon^2 L & \epsilon M \\ \epsilon N & \epsilon P \end{bmatrix} + (\bar{S}-S) \qquad (6.4$$

and the bound for $\|\bar{S}-S\|$ will usually be of order at least as small as ϵ^2. Premultiplication of the first r rows by k ϵ and the first r columns by $1/k\,\epsilon$ then modifies the second matrix to the form

$$\begin{bmatrix} \epsilon^2 L & k\,\epsilon^2 M \\ \frac{1}{k} N & \epsilon\ P \end{bmatrix} \qquad (6.5)$$

and because of this **Gerschgorin's** theorem gives just as fine bounds for multiple roots as for well-separated roots.

Forgetting rounding errors for the moment it is interesting to consider what can be achieved with an approximate matrix X of eigenvectors which can be expressed in the form

$$\bar{X} = X\ (I + \epsilon\ E) \qquad (6.6)$$

where $\|E\|_{al} = 1$ and X is a matrix of exact normalized eigenvectors. We have

$$\begin{aligned} \bar{X}^{-1} A \underline{\bar{X}} &= (I+\epsilon E)^{-1} X^{-1} AX(I+\epsilon E) \\ &= (\ I-\epsilon E+\epsilon^2 E^2 - \cdots)\ \mathrm{diag}(\lambda_i)(\ I+\epsilon E) \\ &= \mathrm{diag}(\lambda_i)+\epsilon F+ \text{ terms in } \epsilon^2 \text{ etc.} \qquad (6.7) \end{aligned}$$

where

$$f_{ij} = -\lambda_j e_{ij} + \lambda_i e_{ij} \qquad (6.8)$$

We see that the elements f_{ij}. is zero whenever $\lambda_i = \lambda_j$. . Hence the off-diagonal elements associated with multiple eigenvalues are of order ϵ^2 .

Notice that when A has eigenvalues which, while not being truly coincident, have separations which are appreciably smaller than ϵ , (6.8) shows that the associated off-diagonal elements are again appreciably smaller than ϵ and a simple application of Gerschgorin's theorem using diagonal similarity transformations gives bounds for the relevant eigenvalues which are of the order of ϵ^2 or of the separations, whichever is the larger, The weakest bounds arise when the separations are themselves of order ϵ. The bounds are then of order ϵ and cannot be improved merely by diagonal similarity transformations.

When the procedure for refining an eigensystem is used iteratively then provided the system is not too ill-conditioned the final eigensystem is "correct to working accuracy." Generally we can assume that the final computed system of vectors satisfies a relation of the form.

$$\bar{X} = X+E \text{ where } \|E\|_\infty \leq n.2^{-t} \|X\|_W \tag{6.8}$$

Hence we have

$$\bar{X}^{-1}A\bar{X} = (X^{-1}-X^{-1}EX^{-1}-1--)AX(I+X^{-1}E) \tag{6.9}$$
$$= \text{diag}(\lambda_i)-X^{-1}E\ \text{diag}(\lambda_i) + \text{diag}(\lambda_i)X^{-1}E+\ldots$$

Equation (6.9) shows the real limitation on the attainable accuracy with computation of a prescribed precision. The off-diagonal elements of $\bar{X}^{-1}A\bar{X}$ are certainly bounded by $2n\ 2^{-t}\|X\|_\infty\|X^{-1}\|_\infty \max|\lambda_i|$ ignoring the quadratic and higher order terms in E. Writing

$$2n\ 2^{-t}\|X\|_\infty\|X^{-1}\|_\infty \max|\lambda_i| = \beta \tag{6.10}$$

the bounds attainable for the eigenvalues using Gerschgorin's theorem and diagonal transformations can be expressed in the following form. Let the eigenvalues be divided into three groups. The first group consists of multiple eigenvalues; the second group consists of eigenvalues with a minimum separation which is less than β and the third group consists of the remainder. For an eigenvalue in the first group having a minimum separation of δ_1

from all other eigenvalues the bound is of the order of β^2/δ_1 . For a member of the second group having operations of up to s with its close neighbours and a minimum separation of order δ_2 from all others the bound is of the order of $s+\beta^2/\delta_2$. For a member of the third group having a minimum separation from <u>all</u> other eigenvalues of δ_3 the bound is of the order of β^2/δ_3 . In general unless $\|X\|_\infty\|X^{-1}\|_\infty$ is quite large the bounds are all appreciably better than $2^{-t}\max|\lambda_i|$ except when δ is of the order of magnitude of β .

This result has been amply confirmed in practise, multiple eigenvalues being found, in general, to the same high precision as well-separated eigenvalues,

REFERENCES

1. Bauer, F. L. , (1963), "Optimally scaled matrices", Numer. Math. 5, 73-87.

2. Gerschgorin, S., (1931), "Über die Abgrenzung der Eigenwerte einer Matrix", Izv. Akad. Nauk. SSSR, Ser. fiz-mat. 6, 749-754.

3. Henrici, P., (1958), 'On the speed of convergence of cyclic and quasi-cyclic Jacobi methods for computing eigenvalues of Hermitian matrices", J. Soc. Industr. Appl. Math., 6, 144-162.

4. Hoffman, A. J. and Wielandt, H. W., (1953), "The variation of the spectrum of a normal matrix", Duke Math. J. 20, 37-39.

5. Jacobi, C. G. J., (1846), "Über ein leichtes Verfahren die in der Theorie der Säculärstörungen vorkommenden Gleichungen numerische aufzulösen". Crelle's J. 30, 51-94 .

6. van Kempken, H. P. M., (1966), 'On the convergence of the classical Jacobi method for real symmetric matrices with non-distinct eigenvalues", Numer. Math. 9, 11-18.

7. van Kempen, H. P. M., (1966), "On the quadradic convergence of the special cyclic Jacobi method", Numer. Math. 9, 19-22.

8. Pope, D. A., and Tompkins, C., (1957), "Maximizing functions of-rotations", J. Ass. Comp. Mach. 4, 459-466.

9. Schönhage, A., (1961), "Zur Konvergenz des Jacobi-Verfahrens", Numer. Math. 3, 374-380.

10. Schönhage, A., (1964), "Zur quadratischen Konvergenz des Jacobi-Verfahrens", Numer. Math. 6, 410-412.

11. Varah, J., (1967), Ph.D. Thesis, Stanford.

12. Wilkinson, J. H., (1961), "Rigorous error bounds for computed eigensystems", Computer J. 4, 230-241.

13. Wilkinson, J. H., (1962), "Note on the quadratic convergence of the cyclic Jacobi process", Numer.. Math. 4,296-300.

14. Wilkinson, J. H., (1965), "The Algebraic Eigenvalue Problem", Oxford University Press.

15. Wilkinson, J. H., (1965), "The Q R algorithm for real symmetric matrices with multiple eigenvalues", Computer J. 8, 85-87.