# ON CYCLIC JACOBI METHODS FOR THE POSITIVE DEFINITE GENERALIZED EIGENVALUE PROBLEM

# DISSERTATION

zur

Erlangung des Grades eines Dr. rer. nat.
des Fachbereich Mathematik und Informatik
der FERNUNIVERSITÄT – Gesamthochschule – Hagen

vorgelegt von

**Vjeran Hari**

aus Zagreb

Hagen 1984

Berichterstatter: Prof. Dr. K. Veselić
Prof. Dr. S. Falk

Tag der mündlichen Prüfung: 25. Juni 1984

Contents


| | |
|---|---|
| Acknowledgement | iii |
| 0. Introduction | v |
| 1. PRELIMINARIES | 1 |
| 1.1. Generalized Eigenvalue Problem | 1 |
| 1.2. Almost Diagonal Pair | 2 |
| 2. JACOBI METHOD FOR THE PAIR (A,B) | 13 |
| 2.1. A General Jacobi Method | 13 |
| 2.2. The Special Jacobi Method | 22 |
| The Case of Real Matrices | 29 |
| The Method of Falk and Langemeyer | 32 |
| 2.3. Fast Scaled Plane Transformations | 35 |
| The Case of Real Matrices | 38 |
| 2.4. Global Convergence | 41 |
| 3. ASYMPTOTIC CONVERGENCE | 45 |
| 3.1. Assumptions and Notation | 45 |
| 3.2. Preliminaries | 48 |
| 3.3. Simple Eigenvalues | 68 |
| 3.4. Multiple Eigenvalues | 79 |
| 3.5. Qualitative Analysis and Finite Arithmetic | 81 |
| 3.6. Modified Method | 87 |
| Modified Transformation | 88 |
| Quadratic Convergence | 89 |
| 3.7. Numerical Tests | 99 |
| A. The Case of Simple Eigenvalues | 102 |
| B. The Case of Multiple Eigenvalues | 104 |
| Principal Notation | 107 |
| References | 110 |

## Acknowledgement

I would like to express my thanks to all who have helped me to complete this thesis.

First of all I would like to thank my supervisor Prof.Dr. K.Veselić on his help in the choice of subject and on his valuable suggestions and remarks during the writing of the thesis. Prof.Dr.K.Veselić was also the supervisor for my M.S. thesis on the University of Zagreb in 1980. His presence, critical observations and discussions were helpful in my introduction to scientific work. Therefore I heartly thank him for everything.

I also wish to thank Prof.Dr.S.Falk who has accepted to read and review this thesis.

## 0. Introduction

In this thesis we study the asymptotic convergence of a cyclic Jacobi method for the generalized eigenvalue problem

$$Ax = \lambda Bx$$

where A, B are hermitian matrices such that B is positive definite. The method is a generalization of the known Jacobi method for the diagonalization of a hermitian matrix.

Generally, Jacobi methods are not the fastest but there are situations in which they are most appropriate. Such situations are those in which the matrices (to be diagonalized) have already, in a sense, small off-diagonal elements. A typical example is the use of Jacobi methods in the subspace iteration techniques (see [14],[16],[2]).

Historically the first and today the most commonly used Jacobi method for the real generalized eigenvalue problem (real means that A and B are real matrices) is the method proposed by S.Falk and P.Langemeyer in [5], 1960. Six years later, in her Ph.D.thesis [25], K.Zimmermann studied another cyclic Jacobi method for the same problem. In 1979 G.Gose [8] proposed a more general definition of a Jacobi method for that problem.

The convergence of the existing Jacobi methods was studied to same extent in [25] and [8]. In [8] G.Gose proposed the global convergence under some optimal i.e.non-cyclic pivot strategies. In [25] K.Zimmermann proved the convergence of the row- and the column-cyclic Jacobi method provided that

all off-diagonal elements of the matrix $B$ are sufficiently small (see Remark 2.10). Note that we use the term "Jacobi method" for different (Jacobi-like) methods.

The asymptotic convergence has not been investigated theoretically, although it is known from the numerical experiments that it "usually" converges quadratically. In [25] there is an estimate, originally used for the global convergence considerations (see [25], Satz 8, page 32 or here Remark 3.12) which indicates that the quadratic convergence could be expected provided that all the eigenvalues are distinct.

We see that in the real case the problems of the global and the quadratic convergence for the cyclic Jacobi method are still not solved. Another interesting problem concerns the asymptotic rate of convergence in the case of multiple eigenvalues. Since the rate of convergence slows down in the presence of multiple eigenvalues, as the numerical tests indicate, the problem is how to modify the method to be quadratically convergent.

In this thesis we consider a Jacobi method for the complex positive definite generalized eigenvalue problem, which is a generalization of the real method discussed by K.Zimmermann in [25] and to which it is reduces if the matrices are real. The method is globally convergent under the row- and the column-cyclic pivot strategy, but this result is not proved here (see [10]). Here we prove the quadratic convergence for the cyclic pivot strategies provided that the eigenvalues of the problem are simple. We also prove that in the case of multiple eigenvalues the method can be asymptotically modified in such a way that the quadratic convergence persists. The obtained results are applied to both real methods.



The contents is divided into three chapters, each chapter into several sections.

The derivation of the method and the global convergence results are presented in Chapter 2. Here we also prove that the real method considered by K.Zimmermann is simply related to the real method proposed by S.Falk and P.Langemeyer. The quadratic convergence in the case of simple eigenvalues is proved in Chapter 3. In Chapter 1 a pair of "almost diagonal" matrices $A$ and $B$ is investigated. These results are used in the quadratic convergence proof of the modified method which is presented in Section 3.6. Finally, in Section 3.7 we give a brief discussion of a numerical investigation of the asymptotic convergence.

The major contributions of this thesis can be found in Chapter 2, Chapter 3 and Section 1.2. They concern the derivation of a new complex Jacobi method and its asymptotic convergence investigation. The closely related result from Section 1.2 dealing with almost diagonal matrices $A$ and $B$ is a generalization of the known result due to J.H.Wilkinson [23]. In the real case we mention the relation between the two real methods which is proved in Section 2.3. It enables us to formulate the convergence results for the most common Jacobi method due to S.Falk and P.Langemeyer.

The thesis has been written in 1983 during the author's stay with his supervisor Prof.Dr.K.Veselić at the Fernuniversität in Hagen.

# 1. PRELIMINARIES

## 1.1. Generalized Eigenvalue Problem

The generalized eigenvalue problem reads

$$Ax = \lambda Bx, \quad x \neq 0. \tag{1.1.1}$$

Here $A, B$ are given square matrices of order $n$. The vector $x$ is called eigenvector and the number $\lambda$ the corresponding eigenvalue of the matrix pair $(A,B)$.

Two pairs $(A,B)$ and $(A_1,B_1)$ are called equivalent if

$$A_1 = PAQ, \quad B_1 = PBQ \tag{1.1.2}$$

for some nonsingular matrices $P, Q$. The substitution $x = Qy$ transforms (1.1.1) into

$$A_1 y = \lambda B_1 y, \quad y \neq 0.$$

Thus, equivalent matrix pairs have the same eigenvalues and simply related eigenvectors.

In this thesis we deal only with hermitian matrices $A, B$ such that $B$ is positive definite. In this case a pair $Q$, $P = Q^*$ exists such that both matrices $A_1$ and $B_1$ from (1.1.2) are diagonal. The columns of the matrix $Q$ represent a basis of the eigenvectors of the pair $(A,B)$. The diagonality of the matrix $Q^*BQ$ means that the basis of the eigenvectors can be chosen to be B-orthogonal.

It is obvious that all eigenvalues of the pair $(A,B)$ lie in the interval $[-\mu,\mu]$ with

$$\mu = \max_{x \neq 0} \frac{|(Ax|x)|}{(Bx|x)} \tag{1.1.3}$$

where $(\,|\,)$ denotes the usual scalar product in $\mathbb{C}^n$. The number $\mu$ will be called the spectral radius of the pair $(A,B)$ and denoted by $spr(A,B)$.

The notational conventions used in this thesis are quite similar to those in [22]. However, there are some differences and the reader is refered to the list of principal notations on page 107.

In the sequel the letters $A, B$ denote hermitian matrices of order $n$, such that $B$ is positive definite. If not specified otherwise the term pair denotes the pair $(A,B)$.

## 1.2. Almost Diagonal Pair

Here we investigate the structure of "almost diagonal" pair $(A,B)$ in the case of multiple eigenvalues. We generalize the known result of J.H.Wilkinson [23] on almost diagonal hermitian matrices to the case of the matrix pair $(A,B)$.

Let $A = (a_{ij})$, $B = (b_{ij})$ and assume

$$b_{11} = b_{22} = \cdots = b_{nn} = 1, \tag{1.2.1}$$

$$a_{11} \geq a_{22} \geq \cdots \geq a_{nn}, \tag{1.2.2}$$

and

$$\lambda_1 = \lambda_2 = \cdots \lambda_{s_1} > \lambda_{s_1+1} = \cdots = \lambda_{s_2} > \cdots$$
$$\cdots > \lambda_{s_{p-1}+1} = \cdots = \lambda_{s_p}, \tag{1.2.3}$$



where $2 \leq p < n$. Note that $p = 1$ in the relation (1.2.3) corresponds to the trivial case $A = \lambda_1 B$.

For each $i = 1, \ldots, p$

$$n_i = s_i - s_{i-1}, \quad s_0 = 0 \tag{1.2.4}$$

is the multiplicity of $\lambda_{s_i}$, counted as a root of the charcteristic polynomial $\det(\lambda B - A)$. According to the relation (1.2.3) set

$$3\delta_i = \min_{\substack{1 \leq j \leq p \\ j \neq i}} |\lambda_{s_i} - \lambda_{s_j}|, \quad 1 \leq i \leq p, \tag{1.2.5}$$

$$\delta = \min_{1 \leq i \leq p} \delta_i \tag{1.2.6}$$

and

$$C_i = D_i + E_i = A - \lambda_{s_i} B, \quad 1 \leq i \leq p, \tag{1.2.7}$$

where

$$D_i = \operatorname{diag}(C_i) = \operatorname{diag}(a_{11} - \lambda_{s_i} b_{11}, \ldots, a_{nn} - \lambda_{s_i} b_{nn}).$$

In the sequel $\|\cdot\|$ denotes the Euclidean vector or matrix norm, while $\|\cdot\|_2$ denotes the spectral norm. Note that the spectral norm is induced by the Euclidean vector norm. The Euclidean matrix norm is sometimes called the Schur, Frobenius or Hilbert-Schmidt norm.

In the thesis we frequently use the partition

$$A = \begin{bmatrix} A_{11} & \cdots & A_{1p} \\ \vdots & & \vdots \\ A_{p1} & \cdots & A_{pp} \end{bmatrix}, \quad B = \begin{bmatrix} B_{11} & \cdots & B_{1p} \\ \vdots & & \vdots \\ B_{p1} & \cdots & B_{pp} \end{bmatrix}, \tag{1.2.8}$$

where the submatrices $A_{ii}$ and $B_{ii}$ have order $n_i$ $(1 \leq i \leq p)$; $n_i$ being defined by (1.2.4).

1.1.Lemma. Let $A, B$ be such that the relations (1.2.1) and (1.2.2) hold. In addition assume

$$\|E_i\|_2 < \delta_i , \quad 1 \leq i \leq p , \tag{1.2.9}$$

where $E_i$ , $\delta_i$ and $p$ are as in (1.2.7), (1.2.5) and (1.2.3), respectively. Then the inequalities

$$\|A_{ii} - \lambda_{s_i} B_{ii}\| \leq \frac{1}{\delta_i} \sum_{j \neq i} \|A_{ij} - \lambda_{s_i} B_{ij}\|^2 ,$$

$$1 \leq i \leq p \tag{1.2.10}$$

hold, where the submatrices $A_{ij}$ , $B_{ij}$ are as in (1.2.8). In the relation (1.2.10) the Euclidean norm can be replaced by the spectral norm.

Proof. Let $i \in \{1,2,\dots,p\}$ and let $\gamma_1^{(i)}, \dots, \gamma_n^{(i)}$ be the eigenvalues of $C_i$ . Applying the perturbation theorem for the eigenvalues of hermitean matrices (See [22], Sec. 2.44 , or [14], Sec.1.6) to $C_i$ and $D_i$ we obtain

$$|\gamma_j^{(i)} - (a_{jj} - \lambda_{s_i})| \leq \|C_i - D_i\|_2 = \|E_i\|_2 ,$$

$$1 \leq j \leq n . \tag{1.2.11}$$

Since $C_i$ has rank $n - n_i$ there are exactly $n_i$ eigenvalues $\gamma_j^{(i)}$ which are zero. Therefore the inequalities (1.2.11) and (1.2.9) imply that

$$|a_{jj} - \lambda_{s_i}| < \delta_i \tag{1.2.12}$$

holds for at least $n_i$ values of $j$ . Let $S_i$ be the set of all indices $j$ $(1 \leq j \leq n)$ for which the inequality (1.2.12) holds. Than $S_i \subset S = \{1,2,\dots,n\}$ $(1 \leq i \leq p)$ and each $S_i$ has at least $n_i$ elements.



Using the relations (1.2.5) and (1.2.12) we obtain

$$|a_{jj} - \lambda_{s_i}| \geq |\lambda_{s_k} - \lambda_{s_i}| - |a_{jj} - \lambda_{s_k}| >$$

$$> 3 \max\{\delta_i, \delta_k\} - \delta_k \geq$$

$$\geq 2 \max\{\delta_i, \delta_k\} \geq 2\delta_i ,$$

$$j \in S_k , \quad k \neq i , \tag{1.2.13}$$

hence $S_i \cap S_k = \emptyset$ for every $i \neq k$ . Therefore $\bigcup_i S_i$ has exactly $n_1 + n_2 + \dots + n_p = n$ elements from $S$ , hence the sets $S_1, S_2, \dots, S_p$ give a partition of $S$ . Now the relations (1.2.2), (1.2.3) and (1.2.13) imply that

$$S_i = \{s_{i-1} + 1, s_{i-1}+2, \dots, s_i\}, \quad 1 \leq i \leq p .$$

In order to prove the inequalities (1.2.10) we partition the matrix $C_i$ in accordance with (1.2.8). Set $C_i = (C_{jk}^{(i)})$ , where $C_{jk}^{(i)}$ are the corresponding blocks of $C_i$ . Let $e_1, e_2, \dots, e_n$ be the columns of the identity matrix $I_n$ , and let the permutation matrices $P_i$ be defined by their columns as follows :

$$P_i = \begin{cases} I_n , & i = 1 \\ [e_{s_{i-1}+1}, e_{s_{i-1}+2}, \dots, e_{s_i}, e_1, e_2, \dots, e_{s_{i-1}}, e_{s_i+1}, \dots, e_{s_p}], \\ \quad 2 \leq i \leq p-1 \\ [e_{s_{p-1}+1}, \dots, e_{s_p}, e_1, \dots, e_{s_{p-1}}], & i = p . \end{cases}$$

Consequently we find

$$P_i^* C_i P_i = \begin{bmatrix} C_{ii}^{(i)} & G_i \\ G_i^* & K_i \end{bmatrix}, \quad 1 \leq i \leq p , \tag{1.2.14}$$

where

$$G_i = \begin{cases} [c_{12}^{(i)}, \dots, c_{1p}^{(i)}], & i = 1 \\ [c_{i1}^{(i)}, \dots, c_{ii-1}^{(i)}, c_{ii+1}^{(i)}, \dots, c_{ip}], & 2 \le i \le p, \\ [c_{p1}^{(i)}, \dots, c_{p-1p}^{(i)}]. & i = p. \end{cases} \qquad (1.2.15)$$

Let $\varkappa_j^{(i)}$, $j \in S \setminus S_i$ denote the eigenvalues of $K_i$. Applying the perturbation theorem, now to $K_i$ and its diagonal part $\mathrm{diag}(K_i)$, we obtain

$$|(a_{jj} - \lambda_{s_i}) - \varkappa_j^{(i)}| \le \|K_i - \mathrm{diag}(K_i)\|_2, \quad j \in S \setminus S_i. \qquad (1.2.16)$$

Since the spectral norm of a principal submatrix is not larger than the spectral norm of the whole matrix we have

$$\|K_i - \mathrm{diag}(K_i)\|_2 \le \|P_i^*(C_i - D_i)P_i\|_2 = \|E_i\|_2. \qquad (1.2.17)$$

In (1.2.17) the invariance property of the spectral norm has been used. Using the inequalities (1.2.16), (1.2.17), (1.2.13) and (1.2.9) we obtain

$$|\varkappa_j^{(i)}| \ge |a_{jj} - \lambda_{s_i}| - \|E_i\|_2 > 2\delta_i - \delta_i = \delta_i, \quad j \in S \setminus S_i.$$

Thus, $K_i$ is invertible and

$$\|K_i^{-1}\|_2 = \frac{1}{\min\limits_{j \in S \setminus S_i} |\varkappa_j^{(i)}|} \le \frac{1}{\delta_i}, \quad 1 \le i \le p. \qquad (1.2.18)$$

Since

$$\begin{bmatrix} I_{n_i} & -G_i K_i^{-1} \\ 0 & I_{n-n_i} \end{bmatrix} \begin{bmatrix} C_{ii}^{(i)} & G_i \\ G_i^* & K_i \end{bmatrix} \begin{bmatrix} I_{n_i} & 0 \\ -K_i^{-1} G_i^* & I_{n-n_i} \end{bmatrix} =$$



$$= \begin{bmatrix} C_{ii}^{(i)} - G_i K_i^{-1} G_i & 0 \\ 0 & K_i \end{bmatrix}$$

we can use the Sylvester's invertia theorem (see [14], sec.1.5) or the simple rank argument to conclude that

$$C_{ii}^{(i)} = G_i K_i^{-1} G_i^*. \qquad (1.2.19)$$

Let $\|\cdot\|_g$ denote the spectral or Euclidean matrix norm, then the relations (1.2.19) and (1.2.18) imply

$$\|C_{ii}^{(i)}\|_g \le \|G_i K_i^{-1}\|_2 \|G_i^*\|_g \le \|K_i^{-1}\|_2 \|G_i\|_g \|G_i^*\|_g \le \frac{1}{\delta_i} \|G_i\|_g^2,$$

hence the inequalities (1.2.10) follow from the relation (1.2.15). ■

For an arbitrary matrix $C = (c_{ij})$ set

$$S(C) = \|C - \mathrm{diag}(C)\| = \Big[ \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \ne i}}^{n} |c_{ij}|^2 \Big]^{1/2},$$

and for the pair $(A,B)$ set

$$\varepsilon = [S^2(A) + S^2(B)]^{1/2}. \qquad (1.2.20)$$

By the Cauchy–Schwartz inequality and the definition (1.2.20) we have

$$\|E_i\|_2 \le \|E_i\| = S(A - \lambda_{s_i} B) \le \sqrt{1 + \lambda_{s_i}^2}\, \varepsilon \le \sqrt{1 + \mu^2}\, \varepsilon, \quad 1 \le i \le p. \qquad (1.2.21)$$

From (1.2.21) we see that Lemma 1.1 holds provided the assumption (1.2.9) is replaced by any of the following conditions:

$$\|E_i\| < \delta_i, \quad 1 \le i \le p,$$

$$\max_{1\le i\le p} \|E_i\|_2 < \delta ,$$

$$\max_{1\le i\le p} \|E_i\| < \delta , \qquad \varepsilon < \min_{1\le i\le p} \frac{\delta_i}{\sqrt{1+\lambda_{s_i}^2}}$$

or simply by

$$\sqrt{1+\mu^2}\,\varepsilon < \delta$$

where $\mu = \mathrm{spr}(A,B)$.

Under the assumptions of Lemma 1.1 it makes sense to define

$$\tau(A) = \left[\sum_{i=1}^{p}\sum_{\substack{j=1\\ j\neq i}}^{p} \|A_{ij}\|^2\right]^{1/2}$$

and similarly $\tau(B)$.

If the conditions (1.2.1) and (1.2.2) of Lemma 1.1 are not satisfied, define

$$D = \mathrm{diag}\,(1/\sqrt{b_{11}},\dots,1/\sqrt{b_{nn}}) \qquad (1.2.22)$$

and consider the pair (DAD, DBD). It is equivalent to the pair (A,B) and for the matrix DBD the equalities (1.2.1) hold. If (1.2.2.) is still not satisfied for the matrix DAD then there exists a permutation matrix P such that the conditions (1.2.1) and (1.2.2) simultaneously hold for $(P^*DADP, P^*DBDP)$. Since the latter pair is equivalent to the pair (A,B) we can set

$$S(A,B) = [S^2(DAD) + S^2(DBD)]^{1/2} \qquad (1.2.23)$$

and

$$\tau(A,B) = [\tau^2(P^*DADP) + \tau^2(P^*DBDP)]^{1/2}. \qquad (1.2.24)$$

Note that $\tau(A,B)$ is well defined and does not depend on the ordering of the eigenvalues. The definitions of $\varepsilon$ and $S(A,B)$ imply $S(A,B) = \varepsilon$ provided that $b_{ii} = 1$, $1 \le i \le n$.



If $p = n$ then the blocks $A_{ii}$, $B_{ii}$ $(1 \le i \le p)$ have order one, hence $\tau(A,B) = S(A,B)$. If $p = 1$ we can set $\tau(A,B) = 0$, but this case is excluded by the assumption (1.2.3). From the definitions (1.2.23) and (1.2.24) we have

$$\tau(A,B) \le S(A,B).$$

Theorem 1.2. Let A,B be hermitian matrices such that B is positive definite and let $\delta_i$, $E_i$, D be as in (1.2.5), (1.2.7) and (1.2.22) respectively. If

$$\|DE_iD\|_2 < \delta_i , \quad 1 \le i \le p , \qquad (1.2.25)$$

then there is a permutation matrix P such that for the matrices $A' = P^*DADP$, $B' = P^*DBDP$ partitioned as in (1.2.8). the following inequalities hold

$$\|A'_{ii} - \lambda_{s_i} B'_{ii}\| \le \frac{1}{\delta_i}\sum_{\substack{j=1\\ j\neq i}}^{p} \|A'_{ij} - \lambda_{s_i} B'_{ij}\|^2 ,$$

$$1 \le i \le p . \qquad (1.2.26)$$

Moreover, if $\mu, \delta$ and $\tau(A,B)$ are such that the relations (1.1.3), (1.2.6) and (1.2.24) hold then

$$\sum_{i=1}^{p} \|A'_{ii} - \lambda_{s_i} B'_{ii}\|^2 \le [(1+\mu^2)\tau^2(A,B)/\delta]^2. \qquad (1.2.27)$$

Proof: Let P be defined so that the inequalities

$$a'_{11} \ge a'_{22} \ge \dots \ge a'_{nn}$$

hold, where $A' = (a'_{ij})$. For $i = 1,2,\dots,n$ we have $b'_{ii} = 1$ and

$$\|(A' - \lambda_{s_i} B') - \mathrm{diag}(A' - \lambda_{s_i} B')\|_2 =$$

$$= \|P^*DE_iDP\|_2 = \|DE_iD\|_2 < \delta_i .$$

Therefore Lemma 1.1. can be applied to the pair (A',B') to obtain the inequalities (1.2.26). To prove the inequality

(1.2.27) we use the Cauchy–Schwartz inequality and the fact that $\mu = \max_{1\le i\le n}|\lambda_i|$. We have

$$\|A'_{ij} - \lambda_{s_i} B'_{ij}\|^2 \le (1+\lambda_{s_i}^2)(\|A'_{ij}\|^2 + \|B'_{ij}\|^2) \le$$

$$\le (1+\mu^2)(\|A'_{ij}\|^2 + \|B'_{ij}\|^2), \quad 1\le i,j\le p. \qquad (1.2.28)$$

Since $A', B'$ are hermitian matrices the relations (1.2.26), (1.2.28) and (1.2.24) yield

$$\|A'_{ii} - \lambda_{s_i} B'_{ii}\| \le \frac{1+\mu^2}{\delta_i} \sum_{\substack{j=1\\ j\ne i}}^{p} (\|A'_{ij}\|^2 + \|B'_{ij}\|^2) \le$$

$$\le \frac{1+\mu^2}{2\delta_i}\tau^2(A,B), \quad 1\le i\le p. \qquad (1.2.29)$$

Since $\delta_i > \delta$ $(1\le i\le p)$ the inequalities (1.2.26) and (1.2.29) imply

$$2\sum_{i=1}^{p}\|A'_{ii} - \lambda_{s_i} B'_{ii}\|^2 \le \frac{1+\mu^2}{\delta}\tau^2(A,B)\sum_{i=1}^{p}\|A'_{ii} - \lambda_{s_i} B'_{ii}\| \le$$

$$\le \frac{(1+\mu^2)^2}{\delta^2}\tau^2(A,B)\sum_{i=1}^{p}\sum_{\substack{j=1\\ j\ne i}}^{p}(\|A'_{ij}\|^2 + \|B'_{ij}\|^2) =$$

$$= [(1+\mu^2)\tau^2(A,B)/\delta]^2,$$

hence (1.2.27) is proved. ■

The assumption (1.2.25) of Theorem 1.2. can be replaced by any of the following conditions:

$$\|DE_i D\| < \delta_i, \quad 1\le i\le p,$$

$$\max_{1\le i\le p}\|DE_i D\|_2 < \delta,$$

$$\max_{1\le i\le p}\|DE_i D\| < \delta,$$

$$S(A,B) < \min_{1\le i\le p}\frac{\delta_i}{\sqrt{1+\lambda_{s_i}^2}}$$

or simply



$$\sqrt{1+\mu^2}\,S(A,B) < \delta. \qquad (1.2.30)$$

1.3.Corollary. Let the assumption (1.2.30) hold. Then there is an ordering of the eigenvalues such that

$$2\sum_{i=1}^{n}|a_{ii}/b_{ii} - \lambda_i|^2 \le [(1+\mu^2)\tau^2(A,B)/\delta]^2 \le$$

$$\le [(1+\mu^2)S^2(A,B)/\delta]^2. \qquad (1.2.31)$$

Proof: Note that the diagonal elements of $A'$ from Theorem 1.2 are $a_{ii}/b_{ii}$ (modulo permutations) and $b'_{ii} = 1$ $(1\le i\le n)$. Therefore (1.2.31) follows directly from (1.2.27). ■

From Corollary 1.3 we see that $a_{rr}/b_{rr}$ approximates $\lambda_r$ $(1\le r\le n)$ with the error of order $\tau^2(A,B)$. In addition we see from (1.2.26) that the local accuracy of an approximation $a_{rr}/b_{rr}$ depends on $\delta_i$ (local separation of the appropriate eigenvalue $\lambda_{s_i}$) and on $\|A'_{ij} - \lambda_{s_i} B'_{ij}\|^2$ $(1\le j\le p,\ j\ne i)$.

Theorem 1.2 reveals the structure of an almost diagonal pair $(A,B)$. This structure becomes apparent when $S(A,B)\ll\delta$. If $\tau = \tau(A,B)$ then the appropriate diagonal blocks in $A'$ and $B'$ are proportional up to the order $\tau^2$, the constants of proportionality being the eigenvalues.

This structure plays an important role in the asymptotic convergence of the Jacobi methods for the generalized eigenvalue problem (see Chapter 3).

1.4. Remark. If $B = I_n$ then $B' = I_n$ and $A' = P^*AP$. In this case the inequalities (1.2.26) and (1.2.27) reduce to

$$\|A'_{ii} - \lambda_{s_i} I_{n_i}\| \le \frac{1}{\delta_i}\sum_{\substack{j=1\\ j\ne i}}^{p}\|A'_{ij}\|^2, \quad 1\le i\le p \qquad (1.2.32)$$

and

$$\sum_{i=1}^{p} \|A'_{ii} - \lambda_{s_i} I_{n_i}\|^2 \leq \frac{\tau^4(A)}{\delta^2} \leq \frac{\varepsilon^4}{\delta^2}, \qquad (1.2.33)$$

respectively. In [23] Wilkinson has proved a slight modification of (1.2.32). Namely, in [23] only $\delta$ and $\varepsilon = S(A)$ have been used, so that the assumption (1.2.9) reduces to $S(A) < \delta$ and the result reduces to (1.2.32) if $\delta_i$ is replaced by $\delta$.

Later van Kempen [13], using the result of Wilkinson [23] has proved the inequality (1.2.33).



## 2. JACOBI METHOD FOR THE PAIR (A,B)

### 2.1. A General Jacobi Method

Here we define a Jacobi method for the pair (A,B). The method is a generalization of a method defined by Gose in [8] to the complex A and B. Following Gose's lines in [8], we define a general Jacobi method as follows:

1° Take positive constants e, E with $0 < e \leq E$ and a diagonal matrix $D_0$ so that all diagonal elements of $D_0^* B D_0$ lie in the interval $[e,E]$. Then set

$$A^{(1)} = D_0^* A D_0, \quad B^{(1)} = D_0^* B D_0. \qquad (2.1.1)$$

2° For each $k = 1,2,\ldots$

(a) select the pivot pair $(\ell,m)$, $1 \leq \ell < m \leq n$, according to a given pivot strategy;

(b) determine a nonsingular elementary plane matrix $F_k$ so that the $(\ell,m)$-elements of $F_k^* A^{(k)} F_k$ and $F_k^* B^{(k)} F_k$ are zero and the diagonal elements of $F_k^* B^{(k)} F_k$ lie in the interval $[e,E]$;

(c) set

$$A^{(k+1)} = F_k^* A^{(k)} F_k, \quad B^{(k+1)} = F_k^* B^{(k)} F_k. \qquad (2.1.2)$$

An elementary plane matrix $F = (f_{ij})$ differs from the identity matrix only on the positions $(\ell,\ell)$, $(\ell,m)$, $(m,\ell)$ and $(m,m)$. The matrix

$$\hat{F} = \begin{bmatrix} f_{\ell\ell} & f_{\ell m} \\ f_{m\ell} & f_{mm} \end{bmatrix}$$

is called the $(\ell,m)$-restriction of $F$.

The choice of $(\ell,m)$ in $2^{o}(a)$ for any $k = 1,2,\dots$ is called a pivot strategy and the pair $(\ell,m)$ is called the pivot pair. We consider only the cyclic pivot strategies, in which each sequence of

$$N = \frac{1}{2}n(n-1) \tag{2.1.3}$$

successive pairs $(\ell,m)$ contains all pairs $(i,j)$, $1 \le i < j \le n$. Two most common cyclic pivot strategies are the row- and the column-cyclic pivot strategy. In the row-cyclic pivot strategy the pivot pair $(\ell,m)$ runs through the sequence

$$(1,2),\ (1,3),\ \dots,\ (1,n),\ (2,3),\ \dots,\ (2,n),\ \dots,\ (n-1,n)$$

and continues in a cyclic way. The column-cyclic pivot strategy is defined in the same way by the sequence

$$(1,2),\ (1,3),\ (2,3),\ \dots,\ (1,n),\ (2,n),\ \dots,\ (n-1,n).$$

Accordingly, we shall also write

$$\ell = \ell(k), \qquad m = m(k). \tag{2.1.4}$$

In the sequel we shall call the transformation $(A^{(k)},B^{(k)}) \longmapsto (A^{(k+1)},B^{(k+1)})$ the k'th step of a method. Also, the letters $\ell$ and $m$ shall be exclusively used for the pivot indices of the k'th step.

Note that the diagonal matrix $D$ in the relation (2.1.1) can be taken real with positive diagonal elements. The choice of the constants $e, E$ is arbitrary (except for the condition $0 < e \le E$). The requirement $0 < e$ prevents the diagonal elements of $B^{(k)}$ to accumulate at zero and the existence of an upper bound $E$ ensures their moderate size. The choice



$e = E$, especially $e = E = 1$ simplifies the construction of the algorithms (see Sec.2.2).

In the relation (2.1.2) the elementary plane matrix $F_k$ is chosen to satisfy the three conditions:

$$b_{\ell m}^{(k+1)} = 0, \tag{2.1.5}$$

$$a_{\ell m}^{(k+1)} = 0 \tag{2.1.6}$$

and

$$b_{\ell\ell}^{(k+1)},\ b_{mm}^{(k+1)} \in [e, E]. \tag{2.1.7}$$

The existence of $F_k$ is easy to prove. In fact, the $(\ell,m)$-restrictions of $A^{(k)}$ and $B^{(k)}$ are hermitian matrices such that $\hat{B}^{(k)}$ is positive definite. Therefore there is a 2-2 matrix $\hat{F}_k$ such that $\hat{F}_k^{*} A^{(k)} \hat{F}_k$ and $F_k^{*} B^{(k)} F_k$ are diagonal matrices. If the condition (2.1.7) does not hold for the latter matrix then $\hat{F}_k$ can be postmultiplied by a suitable 2-2 diagonal matrix so that (2.1.7) holds.

A special case of elementary plane matrices $F_k$ are complex rotations $R_k$ for which

$$\hat{R}_k = \begin{bmatrix} \cos\varphi_k & e^{i\alpha_k}\sin\varphi_k \\ -e^{-i\alpha_k}\sin\varphi_k & \cos\varphi_k \end{bmatrix},$$

where $\alpha_k$ and $\varphi_k$ are real scalars (angles). Complex rotations are widely used in numerical linear algebra. The reason can be seen from the following result.

2.1. Proposition. Every unitary matrix $\hat{U}$ of order two can be represented by any of the following forms:

(i) $$\hat{U} = \begin{bmatrix} \cos\varphi & e^{i\alpha}\sin\varphi \\ -e^{-i\alpha}\sin\varphi & \cos\varphi \end{bmatrix} \begin{bmatrix} e^{i\alpha'} & \\ & e^{i\alpha''} \end{bmatrix}$$

(ii) $$\hat{U} = \begin{bmatrix} e^{i\alpha'} & \\ & e^{i\alpha''} \end{bmatrix} \begin{bmatrix} \cos\varphi & e^{i\beta}\sin\varphi \\ -e^{-i\beta}\sin\varphi & \cos\varphi \end{bmatrix}$$

(iii) $$\hat{U} = \begin{bmatrix} e^{i\alpha'}\cos\varphi & e^{i\beta'}\sin\varphi \\ -e^{i\beta''}\sin\varphi & e^{i\alpha''}\cos\varphi \end{bmatrix}$$

where

$$\alpha = \beta' - \alpha'', \qquad \beta = \beta' - \alpha'$$

and

$$\alpha' + \alpha'' \equiv \beta' + \beta'' \pmod{2\pi} .$$

Proof: Note that (iii) implies (i) and (ii). Since (iii) is obviously equivalent to $\hat{U}^*\hat{U} = I_2 = \hat{U}\hat{U}^*$, Proposition 2.1 is proved. ∎

The following theorem is a generalization of a result of Gose [8] to complex matrices.

2.2. Theorem. Let

$$\hat{B} = \begin{bmatrix} b_{\ell\ell} & b_{\ell m} \\ b_{m\ell} & b_{mm} \end{bmatrix} \tag{2.1.8}$$

be an arbitrary 2×2 hermitian positive definite matrix. The general nonsingular 2×2 matrix $\hat{F}$ satisfying

$$\hat{F}^*\hat{B}\hat{F} = \begin{bmatrix} b'_{\ell\ell} & 0 \\ 0 & b'_{mm} \end{bmatrix} \tag{2.1.9}$$

has the form



$$\hat{F} = \frac{1}{\cos\gamma} \begin{bmatrix} \frac{1}{\sqrt{b_{\ell\ell}}} & \\ & \frac{1}{\sqrt{b_{mm}}} \end{bmatrix} \begin{bmatrix} \cos\varphi & e^{i\alpha}\sin\varphi \\ -e^{-i\beta}\sin\psi & \cos\psi \end{bmatrix} \begin{bmatrix} e^{i\sigma_\ell}\sqrt{b'_{\ell\ell}} & \\ & e^{i\sigma_m}\sqrt{b'_{mm}} \end{bmatrix} \tag{2.1.10}$$

where $\sigma_\ell$, $\sigma_m$ are real, $\varphi, \psi \in [0, \frac{\pi}{2}]$ and

$$\sin\gamma = \frac{|b_{\ell m}|}{\sqrt{b_{\ell\ell}\cdot b_{mm}}}, \qquad \gamma \in [0, \tfrac{\pi}{2}) . \tag{2.1.11}$$

In addition the equality

$$|\cos\varphi\cos\psi + e^{i(\alpha-\beta)}\sin\varphi\sin\psi| = \cos\gamma \tag{2.1.12}$$

holds.

Proof: The existence of a nonsingular $\hat{F}$ for which (2.1.9) holds is obvious. Let

$$\hat{D} = \frac{1}{\cos\gamma} \begin{bmatrix} e^{i\sigma_\ell}\sqrt{b'_{\ell\ell}} & \\ & e^{i\sigma_m}\sqrt{b'_{mm}} \end{bmatrix}, \tag{2.1.13}$$

where $\gamma$ is as in (2.1.11), $b'_{\ell\ell}$ and $b'_{mm}$ as in (2.1.9) and $\sigma_\ell$, $\sigma_m$ are chosen so that the diagonal elements of $\hat{F}\hat{D}^{-1}$ are real and positive.

For the nonsingular matrix $\hat{F}\hat{D}^{-1}$ there is an upper triangular matrix $\hat{T}$ with positive diagonal elements and an unitary matrix $\hat{U}$ such that

$$\hat{F}\hat{D}^{-1} = \hat{T}\hat{U} . \tag{2.1.14}$$

By the assertion (i) of Proposition 2.1 we can set

$$\hat{T} = \begin{bmatrix} x & y \\ 0 & z \end{bmatrix},$$

$$\hat{U} = \begin{bmatrix} \cos\psi_1 & e^{i\alpha_1}\sin\psi_1 \\ -e^{-i\alpha_1}\sin\psi_1 & \cos\psi_1 \end{bmatrix}\begin{bmatrix} e^{i\beta_1} & 0 \\ 0 & e^{i\beta_2} \end{bmatrix}.$$

$x, z > 0$. (2.1.15)

Adjusting $\beta_1$ and $\beta_2$ if necessary we can assume that $\psi_1 \in (-\frac{\pi}{2}, \frac{\pi}{2}]$. Set

$$\tilde{V} = \hat{T}^* \hat{B} \hat{T}, \quad \tilde{\tilde{V}} = \hat{U}^* \tilde{V} \hat{U}. \tag{2.1.16}$$

The relations (2.1.16), (2.1.14) and (2.1.9) imply

$$\tilde{\tilde{V}} = (\hat{T}\hat{U})^*\hat{B}(\hat{T}\hat{U}) = (\hat{F}\hat{D}^{-1})^*\hat{B}(\hat{F}\hat{D}^{-1}) =$$

$$= \hat{D}^{-1*}(\hat{F}^*\hat{B}\hat{F})\hat{D}^{-1} = \cos^2\gamma\, I_2,$$

hence using again (2.1.16) we obtain

$$\tilde{V} = \hat{U}\tilde{\tilde{V}}\hat{U}^* = \cos^2\gamma\, \hat{U}\hat{U}^* = \cos^2\gamma\, I_2 = \tilde{\tilde{V}}. \tag{2.1.17}$$

From the relations (2.1.16), (2.1.17), (2.1.15) and (2.1.8) we obtain

$$\begin{bmatrix} \cos^2\gamma & \\ & \cos^2\gamma \end{bmatrix} = \begin{bmatrix} x & 0 \\ \bar{y} & z \end{bmatrix}\begin{bmatrix} b_{\ell\ell} & b_{\ell m} \\ \bar{b}_{\ell m} & b_{mm} \end{bmatrix}\begin{bmatrix} x & y \\ 0 & z \end{bmatrix}$$

or in components

$$\begin{aligned} \cos^2\gamma &= x^2 b_{\ell\ell} \\ 0 &= x(y b_{\ell\ell} + z b_{\ell m}) \\ \cos^2\gamma &= (\bar{y} b_{\ell\ell} + z\bar{b}_{\ell m})y + (\bar{y} b_{\ell m} + z b_{mm})z = \\ &= |y|^2 b_{\ell\ell} + z^2 b_{mm} + 2z\,\mathrm{Re}(\bar{y} b_{\ell m}). \end{aligned} \tag{2.1.18}$$

Since $x > 0$ we have

$$x = \frac{1}{\sqrt{b_{\ell\ell}}}\cos\gamma, \qquad y = -z\frac{b_{\ell m}}{b_{\ell\ell}}. \tag{2.1.19}$$



Since $z > 0$ we obtain from (2.1.18), (2.1.19) and (2.1.11)

$$\cos^2\gamma = |y|^2 b_{\ell\ell} + z^2 b_{mm} - 2z^2\frac{|b_{\ell m}|^2}{b_{\ell\ell}} =$$

$$= z^2 b_{mm}\left(1 - \frac{|b_{\ell m}|^2}{b_{\ell\ell} b_{mm}}\right) = z^2 b_{mm}\cos^2\gamma,$$

hence the inequality $z > 0$ implies

$$z = \frac{1}{\sqrt{b_{mm}}} \tag{2.1.20}$$

The equalities (2.1.19), (2.1.20) and (2.1.11) imply

$$y = -\frac{1}{\sqrt{b_{mm}}}\frac{b_{\ell m}}{b_{\ell\ell}} = -\frac{e^{i\alpha_2}}{\sqrt{b_{\ell\ell}}}\sin\gamma, \tag{2.1.21}$$

where

$$\alpha_2 = \arg(b_{\ell m}).$$

Next we calculate the elements of $\hat{T}\hat{U}$. By the equalities (2.1.19), (2.1.21) and (2.1.15) we have

$$(\hat{T}\hat{U})_{\ell\ell} = e^{i\beta_1}(x\cos\psi_1 - y\,e^{-i\alpha_1}\sin\psi_1) =$$

$$= \frac{e^{i\beta_1}}{\sqrt{b_{\ell\ell}}}(\cos\psi_1\cos\gamma + e^{i(\alpha_2-\alpha_1)}\sin\psi_1\sin\gamma).$$

The choice of $\sigma_\ell$ and $\sigma_m$ implies

$$(\hat{T}\hat{U})_{\ell\ell} = (\hat{F}\hat{D}^{-1})_{\ell\ell} > 0$$

hence

$$(\hat{T}\hat{U})_{\ell\ell} = \frac{1}{\sqrt{b_{\ell\ell}}}|\cos\psi_1\cos\gamma + e^{i(\alpha_2-\alpha_1)}\sin\psi_1\sin\gamma|.$$

Using the equalities (2.1.19) and (2.1.21) we obtain

$$(\hat{T}\hat{U})_{\ell m} = e^{i\beta_2}(x\, e^{i\alpha_1}\sin\psi_1 + y\cos\psi_1) =$$

$$= \frac{e^{i(\alpha_1+\beta_2)}}{\sqrt{b_{\ell\ell}}}\,(\sin\psi_1\cos\gamma - e^{i(\alpha_2-\alpha_1)}\cos\psi_1\sin\gamma).$$

Since

$$|\cos\psi_1\cos\gamma + e^{i(\alpha_2-\alpha_1)}\sin\psi_1\sin\gamma|^2 +$$

$$+|\sin\psi_1\cos\gamma - e^{i(\alpha_2-\alpha_1)}\cos\psi_1\sin\gamma|^2 = 1\ ,$$

we can define $\varphi \in [0, \frac{\pi}{2}]$ so that

$$(\hat{T}\hat{U})_{\ell\ell} = \frac{1}{\sqrt{b_{\ell\ell}}}\cos\varphi, \quad (\hat{T}\hat{U})_{\ell m} = \frac{e^{i\alpha}}{\sqrt{b_{\ell\ell}}}\sin\varphi\ ,$$

where

$$\alpha = \arg((\hat{T}\hat{U})_{\ell m})\ .$$

From the relations (2.1.15) and (2.1.20) we obtain

$$(\hat{T}\hat{U})_{m\ell} = -\frac{e^{-i(\alpha_1-\beta_1)}}{\sqrt{b_{mm}}}\sin\psi_1\ ,$$

$$(\hat{T}\hat{U})_{mm} = \frac{e^{i\beta_2}}{\sqrt{b_{mm}}}\cos\psi_1\ .$$

The choice for $\sigma_\ell$, $\sigma_m$ implies $(\hat{T}\hat{U})_{mm} \geq 0$, and since $\cos\psi_1 \geq 0$ we have $\beta_2 = 0$. Therefore we have

$$\hat{T}\hat{U} = \begin{bmatrix} \frac{1}{\sqrt{b_{\ell\ell}}} & \\ & \frac{1}{\sqrt{b_{mm}}} \end{bmatrix}\begin{bmatrix} \cos\varphi & e^{i\alpha}\sin\varphi \\ -e^{-i(\alpha_1-\beta_1)}\sin\psi_1 & \cos\psi_1 \end{bmatrix} \tag{2.1.22}$$

Setting

$$\psi = |\psi_1|, \qquad \beta = \begin{cases} \alpha_1 - \beta_1\ , & \psi_1 \geq 0 \\ \alpha_1 - \beta_1 + \pi, & \psi_1 < 0 \end{cases}$$



the form (2.1.10) is obtained from the relations (2.1.14), (2.1.13) and (2.1.22).

To prove the equality (2.1.12) we apply the determinant to the matrix equalities (2.1.9) and (2.1.10). We have

$$|\det\hat{P}|^2 = \frac{b'_{\ell\ell}\,b'_{mm}}{\det\hat{B}} = \frac{b'_{\ell\ell}\,b'_{mm}}{b_{\ell\ell}\,b_{mm}}\,\frac{1}{\cos^2\gamma}$$

and

$$|\det\hat{P}|^2 = \frac{b'_{\ell\ell}\,b'_{mm}}{b_{\ell\ell}\,b_{mm}}|\cos\varphi\cos\psi + e^{i(\alpha-\beta)}\sin\varphi\sin\psi|^2\cdot\frac{1}{\cos^4\gamma}$$

hence (2.1.12) is obtained by equating the expressions on the right hand sides. ■

Set

$$\eta = \psi - \varphi\ ,$$

where $\varphi$ and $\psi$ are such that the relations (2.1.10) and (2.1.12) hold. Then we have

$$\cos\gamma \leq \cos\varphi\cos\psi + \sin\varphi\sin\psi = \cos\eta\ ,$$

hence

$$|\eta| \leq \gamma\ .$$

2.3. Remark. In the real case we have $|\eta| = \gamma$. Namely, Gose has proved in [8] that in the real case

$$\hat{P} = \frac{1}{\cos\tilde{\gamma}}\begin{bmatrix} \frac{1}{\sqrt{b_{\ell\ell}}} & \\ & \frac{1}{\sqrt{b_{mm}}} \end{bmatrix}\begin{bmatrix} \cos\tilde{\varphi} & \sin\tilde{\varphi} \\ -\sin(\tilde{\varphi}+\tilde{\gamma}) & \cos(\tilde{\varphi}+\tilde{\gamma}) \end{bmatrix}\begin{bmatrix} \sqrt{b'_{\ell\ell}} & \\ & \sqrt{b'_{mm}} \end{bmatrix} \tag{2.1.23}$$

where

$$\sin\tilde{\gamma} = \frac{b_{\ell m}}{\sqrt{b_{\ell\ell}\,b_{mm}}}\ .\ ■$$

From Theorem 2.2 we see that the principal part of each $\hat{F}_k$, namely the matrix

$$\frac{1}{\cos\gamma_k}\begin{bmatrix}\cos\varphi_k & e^{i\alpha_k}\sin\varphi_k\\ -e^{-i\beta_k}\sin\varphi_k & \cos\varphi_k\end{bmatrix}$$

is actually determined from the matrices $\hat{D}_k\hat{A}^{(k)}\hat{D}_k$ and $\hat{D}_k\hat{B}^{(k)}\hat{D}_k$, where $\hat{D}_k = \mathrm{diag}\left(\frac{1}{\sqrt{b^{(k)}_{\ell\ell}}}, \frac{1}{\sqrt{b^{(k)}_{mm}}}\right)$. Therefore, the requirement $e = E = 1$ is not "esentially" restrictive.

## 2.2. The Special Jacobi Method

Here we derive the algorithm of a Jacobi method for the pair $(A,B)$ provided that two additional requirements are met. The first is

$$e = E = 1 \tag{2.2.1}$$

and it implies (see (2.1.1))

$$D_o = \mathrm{diag}\left(\frac{1}{\sqrt{b_{11}}}, \frac{1}{\sqrt{b_{22}}}, \dots, \frac{1}{\sqrt{b_{nn}}}\right).$$

In addition, the conditions (2.2.1) and (2.1.7) imply

$$b^{(k)}_{ii} = 1, \quad 1 \le i \le n, \quad k \ge 1. \tag{2.2.2}$$

The second requirement reads

$$f^{(k)}_{\ell\ell} \ge 0, \quad f^{(k)}_{mm} \ge 0, \quad k \ge 1,$$



so that the $(\ell,m)$-restriction of the looked-for transformation is given by (see (2.1.10) and (2.1.11))

$$\hat{R}_k = \frac{1}{\sqrt{1-|b^{(k)}_{\ell m}|^2}}\begin{bmatrix}\cos\varphi_k & e^{i\alpha_k}\sin\varphi_k\\ -e^{-i\beta_k}\sin\psi_k & \cos\psi_k\end{bmatrix}, \quad k \ge 1, \tag{2.2.3}$$

where $\varphi_k, \psi_k \in [0, \frac{\pi}{2}]$. Now we compute $\hat{F}_k$ explicitly; the subscript $k$ is omitted for simplicity. The matrix $\hat{F}$ can be obtained (cf.[25]) as a product

$$F = \hat{R}_1\hat{D}\hat{R}_2\hat{\Phi} \tag{2.2.4}$$

where $\hat{R}_1$, $\hat{R}_2$ are complex rotations, while $\hat{D}$, $\hat{\Phi}$ are diagonal matrices – $\hat{\Phi}$ being also unitary.

Let $\hat{A}_r = (a^r_{ij})$, $\hat{B}_r = (b^r_{ij})$, $i,j \in \{\ell,m\}$, $1 \le r \le 3$, where

$$\hat{A}_1 = \hat{R}_1^*\hat{A}\hat{R}_1 \qquad \hat{B}_1 = \hat{R}_1^*\hat{B}\hat{R}_1$$
$$\hat{A}_2 = \hat{D}^*\hat{A}_1\hat{D} \qquad \hat{B}_2 = \hat{D}^*\hat{B}_1\hat{D}$$
$$\hat{A}_3 = \hat{R}_2^*\hat{A}_2\hat{R}_2 \qquad \hat{B}_3 = \hat{R}_2^*\hat{B}_2\hat{R}_2$$
$$\hat{A}' = \hat{\Phi}^*\hat{A}_3\hat{\Phi} \qquad \hat{B}' = \hat{\Phi}^*\hat{B}_3\hat{\Phi}$$

and note that

$$\hat{A}' = \hat{F}^*\hat{A}\hat{F}, \quad \hat{B}' = \hat{F}^*\hat{B}\hat{F}.$$

The matrix $\hat{R}_1$ is determined from the requirement $b^1_{\ell m} = 0$. Thus $\hat{R}_1$ is the Jacobi (complex) rotation

$$\hat{R}_1 = \begin{bmatrix}\cos\left(-\frac{\pi}{4}\right) & e^{i\beta_1}\sin\left(-\frac{\pi}{4}\right)\\ -e^{-i\beta_1}\sin\left(-\frac{\pi}{4}\right) & \cos\left(-\frac{\pi}{4}\right)\end{bmatrix} \tag{2.2.5}$$

where

$$\beta_1 = \arg(b_{\ell m}).$$

Setting

$$b_{\ell m} = x\, e^{i\beta_1}, \quad x \geq 0, \tag{2.2.6}$$

we obtain

$$b^1_{\ell\ell} = 1 + x, \quad b^1_{mm} = 1 - x.$$

In order to have $\hat{B}_2 = I_2$, set

$$\hat{D} = \begin{bmatrix} \frac{1}{\sqrt{1+x}} & \\ & \frac{1}{\sqrt{1-x}} \end{bmatrix}. \tag{2.2.7}$$

Using the relations (2.2.5), (2.2.6) and (2.2.7), we obtain after a simple calculation

$$\begin{aligned} a^2_{\ell\ell} &= \frac{1}{1+x}\left(\frac{a_{\ell\ell} + a_{mm}}{2} + u\right), \\ a^2_{mm} &= \frac{1}{1-x}\left(\frac{a_{\ell\ell} + a_{mm}}{2} - u\right) \\ a^2_{\ell m} &= \frac{e^{i\beta_1}}{\sqrt{1-x^2}}\left(\frac{a_{mm} - a_{\ell\ell}}{2} + iv\right), \end{aligned} \tag{2.2.8}$$

where

$$u + iv = e^{-i\beta_1} a_{\ell m}, \quad u,v \text{ real}. \tag{2.2.9}$$

Note that $\hat{B}_3 = I_2$ since $\hat{R}_2$ is unitary. For the matrix $\hat{R}_2$ we assume the form

$$\hat{R}_2 = \begin{bmatrix} \cos\left(\vartheta + \frac{\pi}{4}\right) & e^{i\alpha_1}\sin\left(\vartheta + \frac{\pi}{4}\right) \\ -e^{-i\alpha_1}\sin\left(\vartheta + \frac{\pi}{4}\right) & \cos\left(\vartheta + \frac{\pi}{4}\right) \end{bmatrix} \tag{2.2.10}$$



The natural requirement for $\vartheta$ and $\alpha_1$ is $a^3_{\ell m} = 0$. This choice makes $\hat{R}_2$ the Jacobi notation. From (2.2.8) directly follows

$$\begin{aligned} \operatorname{tg} 2\left(\vartheta + \tfrac{\pi}{4}\right) &= \sigma \frac{2|a^2_{\ell m}|}{a^2_{mm} - a^2_{\ell\ell}}, \\ \alpha_1 &= \beta_1 + \arg\left(\frac{a_{mm} - a_{\ell\ell}}{2} + iv\right) + (1-\sigma)\frac{\pi}{2} \end{aligned} \tag{2.2.11}$$

where $\sigma \in \{1,-1\}$. Now $\sigma$ is determined from the requirement

$$-\frac{\pi}{2} < \alpha_1 - \beta_1 \leq \frac{\pi}{2}.$$

It implies

$$\sigma = \begin{cases} 1, & a_{mm} - a_{\ell\ell} \geq 0 \\ -1, & a_{mm} - a_{\ell\ell} < 0 \end{cases}. \tag{2.2.12}$$

The choice (2.2.12) for $\sigma$ and the choice $-\frac{\pi}{4} < \vartheta \leq \frac{\pi}{4}$ are fundamental for the global convergence proof of the method (see [10]). The choice for $\vartheta$ is also important for the quadratic convergence proof (see Chapter 3).

Using the identity

$$\operatorname{tg} 2\vartheta = -1 / \operatorname{tg}(2\vartheta + \tfrac{\pi}{2})$$

we obtain from (2.2.11), (2.2.12), (2.2.8) and (2.2.9)

$$\begin{aligned} \operatorname{tg} 2\vartheta &= \frac{2u - (a_{\ell\ell} + a_{mm})x}{\sqrt{(a_{mm} - a_{\ell\ell})^2 + 4v^2}\sqrt{1-x^2}}\sigma, \\ -\tfrac{\pi}{4} &< \vartheta \leq \tfrac{\pi}{4}. \end{aligned} \tag{2.2.13}$$

Note that $\hat{Z} = \hat{R}_1 \hat{D} \hat{R}_2$ "diagonalizes" the positive definite matrix $\hat{B}$. By Theorem 2.2 and (2.2.6) we see that

$$\hat{Z} = \frac{1}{\sqrt{1-x^2}} \begin{bmatrix} \cos\varphi & e^{i\alpha}\sin\varphi \\ -e^{-i\beta}\sin\psi & \cos\psi \end{bmatrix} \begin{bmatrix} e^{i\sigma_\ell} & \\ & e^{i\sigma_m} \end{bmatrix} \qquad (2.2.14)$$

holds. By the relations (2.2.5), (2.2.6), (2.2.7) and (2.2.10), after a simple calculation we obtain

$$\begin{aligned}
(\hat{Z})_{\ell\ell} &= \frac{1}{\sqrt{1-x^2}} \frac{1}{2}\left[\sqrt{1-x}\,(c-s) + e^{-i(\alpha_1-\beta_1)}\sqrt{1+x}\,(c+s)\right], \\
(\hat{Z})_{\ell m} &= \frac{1}{\sqrt{1-x^2}} \frac{1}{2}\left[e^{i\alpha_1}\sqrt{1-x}\,(c+s) - e^{i\beta_1}\sqrt{1+x}\,(c-s)\right], \\
(\hat{Z})_{m\ell} &= \frac{1}{\sqrt{1-x^2}} \frac{1}{2}\left[e^{-i\beta_1}\sqrt{1-x}\,(c-s) - e^{-i\alpha_1}\sqrt{1+x}\,(c+s)\right], \\
(\hat{Z})_{mm} &= \frac{1}{\sqrt{1-x^2}} \frac{1}{2}\left[e^{i(\alpha_1-\beta_1)}\sqrt{1-x}\,(c+s) + \sqrt{1+x}\,(c-s)\right],
\end{aligned} \qquad (2.2.15)$$

where $c = \cos\vartheta$, $s = \sin\vartheta$. Comparing the matrix elements from the relation (2.2.14) with those from the relation (2.2.15) we obtain

$$\begin{aligned}
2\cos^2\varphi &= 1 + x\sin 2\vartheta + \sqrt{1-x^2}\cos 2\vartheta\cos(\alpha_1-\beta_1) \\
2\sin^2\varphi &= 1 - x\sin 2\vartheta - \sqrt{1-x^2}\cos 2\vartheta\cos(\alpha_1-\beta_1) \\
2\cos^2\psi &= 1 - x\sin 2\vartheta + \sqrt{1-x^2}\cos 2\vartheta\cos(\alpha_1-\beta_1) \\
2\sin^2\psi &= 1 + x\sin 2\vartheta - \sqrt{1-x^2}\cos 2\vartheta\cos(\alpha_1-\beta_1)
\end{aligned} \qquad (2.2.16)$$

and

$$\begin{aligned}
e^{i\sigma_\ell} &= \left[\sqrt{1-x}\,(c-s) + e^{i(\beta_1-\alpha_1)}\sqrt{1+x}\,(c+s)\right]/(2\cos\varphi) \\
e^{i\sigma_m} &= \left[\sqrt{1+x}\,(c-s) + e^{-i(\beta_1-\alpha_1)}\sqrt{1-x}\,(c+s)\right]/(2\cos\psi) \\
e^{i\alpha}e^{i\sigma_m} &= e^{i\beta_1}\left[e^{-i(\beta_1-\alpha_1)}\sqrt{1-x}\,(c+s) - \sqrt{1+x}\,(c-s)\right]/(2\sin\varphi) \\
e^{-i\beta}e^{i\sigma_\ell} &= e^{-i\beta_1}\left[e^{i(\beta_1-\alpha_1)}\sqrt{1+x}\,(c+s) - \sqrt{1-x}\,(c-s)\right]/(2\sin\psi)
\end{aligned} \qquad (2.2.17)$$

After a simple calculation we obtain from the



equalities (2.2.17)

$$\begin{aligned}
e^{i\alpha} &= \frac{e^{i\beta_1}}{2\cos\psi\sin\varphi}\left[\sin 2\vartheta - x - i\sqrt{1-x^2}\cos 2\vartheta\cos(\alpha_1-\beta_1)\right] \\
e^{-i\beta} &= \frac{e^{-i\beta_1}}{2\cos\varphi\sin\psi}\left[\sin 2\vartheta + x + i\sqrt{1-x^2}\cos 2\vartheta\cos(\alpha_1-\beta_1)\right]
\end{aligned} \qquad (2.2.18)$$

Setting

$$\hat{\Phi} = \begin{bmatrix} e^{-i\sigma_\ell} & \\ & e^{-i\sigma_m} \end{bmatrix}$$

and using the relations (2.2.4) and (2.2.14) we obtain the matrix $\hat{P}_k$ in the form (2.2.3) where $\varphi_k$, $\psi_k$, $\alpha_k$ and $\beta_k$ are as in (2.2.16) and (2.2.18). From the relations (2.2.16) and (2.2.18) we see that only $\cos 2\vartheta$, $\sin 2\vartheta$ and $\cos(\alpha_1-\beta_1)$ are needed in the computation of $\hat{P}$.

If $b_{\ell m} = 0$ (i.e. $x = 0$) then $\beta_1 = \arg(b_{\ell m})$ is not defined. In this case we take $\beta_1 = \arg(a_{\ell m})$, so that the method reduces to the usual Jacobi method for the hermitian matrix $\hat{A}$ (cf. [6], [19]).

We define now the <u>special Jacobi method</u> for the pair $(A,B)$ as follows:

1° Set

$$A^{(1)} = D_o A D_o, \qquad B^{(1)} = D_o B D_o$$

where $D_o = \mathrm{diag}\left(1/\sqrt{b_{11}}, \dots, 1/\sqrt{b_{nn}}\right)$.

2° For $k = 1,2,\dots$

(a) select the pivot pair $(\ell,m)$ according to a given pivot strategy;

(b) determine the nonsingular elementary matrix $F_k$

such that

$$\hat{P}_k = \frac{1}{\sqrt{1-b_k^2}} \begin{bmatrix} \cos\varphi_k & e^{i\alpha_k}\sin\varphi_k \\ -e^{-i\beta_k}\sin\psi_k & \cos\psi_k \end{bmatrix},$$

where

$$2\cos^2\varphi_k = 1 + b_k \sin 2\vartheta_k + t_k \cos 2\vartheta_k \cos\gamma_k ,$$

$$0 \leq \varphi_k < \frac{\pi}{2} ,$$

$$2\cos^2\psi_k = 1 - b_k \sin 2\vartheta_k + t_k \cos 2\vartheta_k \cos\gamma_k ,$$

$$0 \leq \psi_k < \frac{\pi}{2} ,$$

$$e^{i\alpha_k}\sin\varphi_k = \frac{e^{i\arg(b_{\ell m}^{(k)})}}{2\cos\psi_k}(\sin 2\vartheta_k - b_k + it_k\cos 2\vartheta_k \sin\gamma_k) ,$$

$$e^{-i\beta_k}\sin\psi_k = \frac{e^{-i\arg(b_{\ell m}^{(k)})}}{2\cos\varphi_k}(\sin 2\vartheta_k + b_k + it_k\cos 2\vartheta_k \sin\gamma_k)$$

and

$$b_k = |b_{\ell m}^{(k)}|, \qquad t_k = \sqrt{1-b_k^2}, \qquad e_k = a_{mm}^{(k)} - a_{\ell\ell}^{(k)},$$

$$\sigma_k = \begin{cases} 1, & e_k \geq 0 \\ -1, & e_k < 0 \end{cases},$$

$$u_k + iv_k = e^{-i\arg(b_{\ell m}^{(k)})} \cdot a_{\ell m}^{(k)} ,$$

$$\operatorname{tg}\gamma_k = 2\frac{v_k}{|e_k|}, \qquad -\frac{\pi}{2} < \gamma_k < \frac{\pi}{2},$$

$$\operatorname{tg} 2\vartheta_k = \sigma_k \frac{2u_k - (a_{\ell\ell}^{(k)} + a_{mm}^{(k)})b_k}{\sqrt{e_k^2 + 4v_k^2}\cdot t_k}, \qquad -\frac{\pi}{4} < \vartheta_k \leq \frac{\pi}{4} ;$$

(note that $\gamma_k$ used here is different than $\gamma$ from (2.1.8));
if $b_{\ell m}^{(k)} = 0$ and $a_{\ell m}^{(k)} \neq 0$ then in the above formulae



replace $\arg(b_{\ell m}^{(k)})$ by $\arg(a_{\ell m}^{(k)})$ ;

if $b_{\ell m}^{(k)} = 0$ and $a_{\ell m}^{(k)} = 0$ then set $\varphi_k = \psi_k = \alpha_k = \beta_k = 0$ ;

if $e_k = 0$ then set $\vartheta_k = \frac{\pi}{4}$, unless $a_{\ell m}^{(k)} = b_{\ell m}^{(k)} = 0$ ;

(c) perform the transformation

$$A^{(k+1)} = F_k^* A^{(k)} F_k , \qquad B^{(k+1)} = F_k^* B^{(k)} F_k .$$

## The Case of Real Matrices

The presented technique can be applied to the real case giving for each $k \geq 1$ the following formulae :

$$P_k = \begin{bmatrix} \cos(-\frac{\pi}{4}) & \sin(-\frac{\pi}{4}) \\ -\sin(-\frac{\pi}{4}) & \cos(-\frac{\pi}{4}) \end{bmatrix} \begin{bmatrix} (1+b_{\ell m}^{(k)})^{-1/2} & \\ & (1-b_{\ell m}^{(k)})^{-1/2} \end{bmatrix}$$

$$\cdot \begin{bmatrix} \cos(\frac{\pi}{4}+\tilde{\vartheta}_k) & \sin(\frac{\pi}{4}+\tilde{\vartheta}_k) \\ -\sin(\frac{\pi}{4}+\tilde{\vartheta}_k) & \cos(\frac{\pi}{4}+\tilde{\vartheta}_k) \end{bmatrix} =$$

$$= \frac{1}{\sqrt{1-(b_{\ell m}^{(k)})^2}} \begin{bmatrix} \cos\tilde{\varphi}_k & \sin\tilde{\varphi}_k \\ -\sin\tilde{\psi}_k & \cos\tilde{\psi}_k \end{bmatrix}, \qquad (2.2.18)$$

$$\begin{aligned} \cos\tilde{\varphi}_k &= \cos\tilde{\vartheta}_k + \tilde{\xi}_k(\sin\tilde{\vartheta}_k - \tilde{\eta}_k\cos\tilde{\vartheta}_k), \\ \sin\tilde{\varphi}_k &= \sin\tilde{\vartheta}_k - \tilde{\xi}_k(\cos\tilde{\vartheta}_k + \tilde{\eta}_k\sin\tilde{\vartheta}_k), \\ \cos\tilde{\psi}_k &= \cos\tilde{\vartheta}_k - \tilde{\xi}_k(\sin\tilde{\vartheta}_k + \tilde{\eta}_k\cos\tilde{\vartheta}_k), \\ \sin\tilde{\psi}_k &= \sin\tilde{\vartheta}_k + \tilde{\xi}_k(\cos\tilde{\vartheta}_k - \tilde{\eta}_k\cos\tilde{\vartheta}_k), \end{aligned} \qquad (2.2.19)$$

$$\tilde{\xi}_k = \frac{b_{\ell m}^{(k)}}{\sqrt{1+b_{\ell m}^{(k)}}+\sqrt{1-b_{\ell m}^{(k)}}}$$

$$\tilde{\eta}_k = \frac{b_{\ell m}^{(k)}}{(1+\sqrt{1+b_{\ell m}^{(k)}})(1+\sqrt{1-b_{\ell m}^{(k)}})} \qquad (2.2.20)$$

and

$$\operatorname{tg} 2\tilde{\vartheta}_k = \frac{2a_{\ell m}^{(k)} - (a_{\ell\ell}^{(k)} + a_{mm}^{(k)})b_{\ell m}^{(k)}}{(a_{mm}^{(k)} - a_{\ell\ell}^{(k)})\sqrt{1-(b_{\ell m}^{(k)})^2}},$$

$$-\frac{\pi}{4} < \tilde{\vartheta}_k \leq \frac{\pi}{4}. \qquad (2.2.21)$$

If $a_{\ell m}^{(k)} = b_{\ell m}^{(k)} = 0$ we set $\tilde{\vartheta}_k = 0$. If $a_{mm}^{(k)} = a_{\ell\ell}^{(k)}$ and $2a_{\ell m}^{(k)} = (a_{\ell\ell}^{(k)} + a_{mm}^{(k)})b_{\ell m}^{(k)}$ then the matrices $\hat{A}^{(k)}$ and $\hat{B}^{(k)}$ are proportional hence we set $\vartheta_k = \frac{\pi}{4}$, unless $a_{\ell m}^{(k)} = b_{\ell m}^{(k)} = 0$.

In [25] K.Zimmermann has assumed $\hat{P}$ in the form $\hat{R}_1 \hat{D} \hat{R}_2$ but the above formulae are new.

The method defined by the relations (2.2.18)–(2.2.21) is the real analogy of the special Jacobi method and it is due to Zimmermann.

2.4. Proposition. Let $\tilde{\varphi}_k$, $\tilde{\psi}_k$ and $\tilde{\vartheta}_k$ $(k \geq 1)$ be defined by the relations (2.2.19)–(2.2.21). Then for $k \geq 1$

$$\min\{\cos\tilde{\varphi}_k, \cos\tilde{\psi}_k\} > 0, \qquad (2.2.22)$$

$$-1 < \operatorname{tg}\tilde{\varphi}_k \operatorname{tg}\tilde{\psi}_k \leq 1 \qquad (2.2.23)$$

and

$$\operatorname{tg}\tilde{\varphi}_k \operatorname{tg}\tilde{\psi}_k = 1 \quad \text{iff} \quad \tilde{\vartheta}_k = \frac{\pi}{4}. \qquad (2.2.24)$$

Proof: From the relation (2.2.20) it follows that



$b_{\ell m}^{(k)} \longmapsto 1 - \tilde{\xi}_k\tilde{\eta}_k$ is a decreasing and $b_{\ell m}^{(k)} \longmapsto |\tilde{\xi}_k|$ is an increasing function on the interval $[0,1]$. Both functions are even.

Since $|b_{\ell m}^{(k)}| < 1$, $k \geq 1$, and

$$(1-\tilde{\xi}_k\tilde{\eta}_k)(1) = |\xi_k|(1), \quad k \geq 1,$$

we have

$$1 - \tilde{\xi}_k\tilde{\eta}_k > |\xi_k|, \quad k \geq 1. \qquad (2.2.25)$$

From the relations (2.2.19) and (2.2.25) we obtain

$$\min\{\cos\tilde{\varphi}_k, \cos\tilde{\psi}_k\} \geq (1-\tilde{\xi}_k\tilde{\eta}_k)\cos\tilde{\vartheta}_k - |\tilde{\xi}_k \sin\tilde{\vartheta}_k| >$$

$$> |\xi_k|(\cos\tilde{\vartheta}_k - |\sin\tilde{\vartheta}_k|) \geq 0, \quad k \geq 1,$$

which proves (2.2.22). To prove (2.2.23) we use the relation (2.2.19). We have

$$\operatorname{tg}\tilde{\varphi}_k \operatorname{tg}\tilde{\psi}_k = \frac{(1-\tilde{\xi}_k\tilde{\eta}_k)^2 \sin^2\vartheta_k - \tilde{\xi}_k^2\cos^2\vartheta_k}{(1-\tilde{\xi}_k\tilde{\eta}_k)^2\cos^2\vartheta_k - \tilde{\xi}_k^2\sin^2\tilde{\vartheta}_k}, \quad k \geq 1,$$

hence the inequality (2.2.25) implies

$$-1 < -\frac{\tilde{\xi}_k^2}{(1-\tilde{\xi}_k\tilde{\eta}_k)^2} \leq \operatorname{tg}\tilde{\varphi}_k \operatorname{tg}\tilde{\psi}_k \leq 1.$$

To prove the assertion (2.2.24) we note that

$$\operatorname{tg}\tilde{\varphi}_k \operatorname{tg}\tilde{\psi}_k = 1 \quad \text{iff}$$

$$[(1-\tilde{\xi}_k\tilde{\eta}_k)^2 + \xi_k^2]\sin^2\tilde{\vartheta}_k = [(1-\tilde{\xi}_k\tilde{\eta}_k)^2 + \xi_k^2]\cos^2\vartheta_k.$$

Since $(1-\tilde{\xi}_k\tilde{\eta}_k)^2 + \tilde{\xi}_k^2 > 0$ we have $\operatorname{tg}\tilde{\varphi}_k \operatorname{tg}\tilde{\psi}_k = 1$ iff $\tilde{\vartheta}_k = \pm\frac{\pi}{4}$, and since by the definition (2.2.21) $\tilde{\vartheta}_k > -\frac{\pi}{4}$, Proposition 2.4 is proved. ■

## The Method of Falk and Langemeyer

Probably the best known and today most commonly used Jacobi method for the real positive definite generalized eigenvalue problem is due to S.Falk and P.Langemeyer (see [5]). In that method the matrix $F_k = (f_{ij}^{(k)})$ satisfies the conditions (2.1.5), (2.1.6) and (instead of (2.1.7))

$$f_{\ell\ell}^{(k)} = 1 = f_{mm}^{(k)} . \tag{2.2.26}$$

This choice assures a simple algorithm and a decrease in the operation count. The elements $f_{\ell m}^{(k)}$ and $f_{m\ell}^{(k)}$ are determined from the requirements

$$\begin{aligned} a_{\ell m}^{(k+1)} &= f_{\ell m}^{(k)} a_{\ell\ell}^{(k)} + (1 + f_{\ell m}^{(k)} f_{m\ell}^{(k)}) a_{\ell m}^{(k)} + f_{m\ell}^{(k)} a_{mm}^{(k)} = 0 \\ b_{\ell m}^{(k+1)} &= f_{\ell m}^{(k)} b_{\ell\ell}^{(k)} + (1 + f_{\ell m}^{(k)} f_{m\ell}^{(k)}) b_{\ell m}^{(k)} + f_{m\ell}^{(k)} b_{mm}^{(k)} = 0 \end{aligned} \tag{2.2.27}$$

Eliminating the nonlinear term one obtains (see [4],[14],[1])

$$f_{\ell m}^{(k)} = \frac{\zeta_m^{(k)}}{\nu^{(k)}} , \qquad f_{m\ell}^{(k)} = - \frac{\zeta_\ell^{(k)}}{\nu^{(k)}} \tag{2.2.28}$$

where $\nu^{(k)}$ satisfies the quadratic equation

$$\nu^2 - \zeta_{\ell m}^{(k)} \nu - \zeta_\ell^{(k)} \zeta_m^{(k)} = 0 . \tag{2.2.29}$$

In the relations (2.2.28) and (2.2.29) $\zeta_\ell^{(k)}$, $\zeta_m^{(k)}$ and $\zeta_{\ell m}^{(k)}$ are defined as follows :

$$\zeta_\ell^{(k)} = \det \begin{bmatrix} a_{\ell\ell}^{(k)} & a_{\ell m}^{(k)} \\ b_{\ell\ell}^{(k)} & b_{\ell m}^{(k)} \end{bmatrix},$$



$$\zeta_m^{(k)} = \det \begin{bmatrix} a_{mm}^{(k)} & a_{\ell m}^{(k)} \\ b_{mm}^{(k)} & b_{\ell m}^{(k)} \end{bmatrix}$$

$$\zeta_{\ell m}^{(k)} = \det \begin{bmatrix} a_{\ell\ell}^{(k)} & a_{mm}^{(k)} \\ b_{\ell\ell}^{(k)} & b_{mm}^{(k)} \end{bmatrix}$$

The two solutions of the equation (2.2.29) are

$$\nu_{+,-} = \frac{1}{2} \operatorname{sgn}(\zeta_{\ell m}^{(k)}) \left[ |\zeta_{\ell m}^{(k)}| \pm \sqrt{(\zeta_{\ell m}^{(k)})^2 + 4\zeta_\ell^{(k)} \zeta_m^{(k)}} \right]. \tag{2.2.31}$$

By the relation (2.2.30) and by the positive definiteness of $\hat{B}^{(k)}$ we have

$$\begin{aligned} &(\zeta_{\ell m}^{(k)})^2 + 4\zeta_\ell^{(k)} \zeta_m^{(k)} = \\ &= b_{\ell\ell}^{(k)} b_{mm}^{(k)} \left[ \left( a_{\ell\ell}^{(k)} \sqrt{\frac{b_{mm}^{(k)}}{b_{\ell\ell}^{(k)}}} - a_{mm}^{(k)} \sqrt{\frac{b_{\ell\ell}^{(k)}}{b_{mm}^{(k)}}} \right)^2 + \right. \\ &+ 4 \left( a_{\ell\ell}^{(k)} \sqrt{\frac{b_{mm}^{(k)}}{b_{\ell\ell}^{(k)}}} \cdot \frac{b_{\ell m}^{(k)}}{\sqrt{b_{\ell\ell}^{(k)} b_{mm}^{(k)}}} - a_{\ell m}^{(k)} \right) \\ &\left. \cdot \left( a_{mm}^{(k)} \sqrt{\frac{b_{\ell\ell}^{(k)}}{b_{mm}^{(k)}}} \cdot \frac{b_{\ell m}^{(k)}}{\sqrt{b_{\ell\ell}^{(k)} b_{mm}^{(k)}}} - a_{\ell m}^{(k)} \right) \right] \geq \\ &\geq (\zeta_{\ell m}^{(k)})^2 \left( 1 - \frac{(b_{\ell m}^{(k)})^2}{b_{\ell\ell}^{(k)} b_{mm}^{(k)}} \right) \geq 0, \end{aligned} \tag{2.2.32}$$

hence the solutions of the equation (2.2.29) are real.

The choice of the solution is determined from the requirement

$$0 < \det F_k \leq 2 \tag{2.2.33}$$

which yields

$$\nu^{(k)} = \nu_+ = \frac{1}{2}\operatorname{sgn}(\zeta_{\ell m}^{(k)})\left[|\zeta_{\ell m}^{(k)}| + \sqrt{(\zeta_{\ell m}^{(k)})^2 + 4\zeta_\ell^{(k)}\zeta_m^{(k)}}\right]. \tag{2.2.34}$$

If $\zeta_{\ell m}^{(k)} = 0$ then $\zeta_m^{(k)} = \zeta_\ell^{(k)}(b_{mm}^{(k)}/b_{\ell\ell}^{(k)})$; in this case we set $\operatorname{sgn}(\zeta_{\ell m}^{(k)}) = \operatorname{sgn}(\zeta_\ell^{(k)})$ provided that $\zeta_\ell^{(k)} \neq 0$. Therefore we have $\nu^{(k)} = \operatorname{sgn}(\zeta_\ell^{(k)}).\sqrt{\zeta_\ell^{(k)}\zeta_m^{(k)}}$ hence the formulae (2.2.28) yield

$$f_{\ell m}^{(k)} = \sqrt{b_{mm}^{(k)}/b_{\ell\ell}^{(k)}}, \quad f_{m\ell}^{(k)} = -\sqrt{b_{\ell\ell}^{(k)}/b_{mm}^{(k)}}. \tag{2.2.35}$$

If $\zeta_{\ell m}^{(k)} = \zeta_\ell^{(k)} = \zeta_m^{(k)} = 0$ the formulae (2.2.35) are also used (except for the case $a_{\ell m}^{(k)} = 0 = b_{\ell m}^{(k)}$ when we set $f_{\ell m}^{(k)} = 0 = f_{m\ell}^{(k)}$).

2.5. Remark. Suppose that $\det F_k = 0$ i.e. $f_{\ell m}^{(k)} f_{m\ell}^{(k)} = 1$. Using the formulae (2.2.28) and the inequality (2.2.32) we obtain $(\zeta_{\ell m}^{(k)})^2 + 4\zeta_\ell^{(k)}\zeta_m^{(k)} = 0$. Using again (2.2.32) we obtain $\zeta_{\ell m}^{(k)} = 0$ and also $\zeta_\ell^{(k)} = \zeta_m^{(k)} = 0$. The latter condition means the proportionality of the matrices $\hat{A}^{(k)}$ and $\hat{B}^{(k)}$. Since in this case the formulae (2.2.28) do not exist we see that the relations (2.2.28) and (2.2.35) imply (2.2.33).

In the special case when $\hat{A}^{(k)}$ and $\hat{B}^{(k)}$ are proportional the choice (2.2.35) is not uniquely determined by the requirement (2.2.33). Note that the replacement of signs in (2.2.35) leads to another solution. In [2] and [1] it is suggested in this case to take $f_{\ell m}^{(k)} = 0$ and $f_{m\ell}^{(k)} = -b_{\ell m}^{(k)}/b_{mm}^{(k)}$. ■

2.6. Proposition. Let $\hat{A}$, $\hat{B}$ be 2×2 symmetric matrices such that $\hat{B}$ is positive definite and let $\hat{F}$ be a real 2×2 matrix satisfying the following three conditions:



(i) the diagonal elements of $\hat{F}$ are units,

(ii) the matrices $\hat{F}^*\hat{A}\hat{F}$ and $\hat{F}^*\hat{B}\hat{F}$ are diagonal,

(iii) $0 < \det F < 2$.

Then the matrix $\hat{F}$ is uniequely defined by the relations (2.2.28), (2.2.30) and (2.2.34) (if the upper suffix $k$ is neglected).

Proof: Denote by $f_{\ell m}^+$, $f_{m\ell}^+$ and $f_{\ell m}^-$, $f_{m\ell}^-$ the two solutions of the equations (2.2.27) determined by $\nu_+$ and $\nu_-$ from (2.2.31), respectively. Since

$$\nu_+\nu_- = -\zeta_\ell\zeta_m$$

the relation (2.2.28) implies

$$f_{\ell m}^- = 1/f_{m\ell}^+, \quad f_{m\ell}^- = 1/f_{\ell m}^+,$$

hence we have

$$f_{\ell m}^- f_{m\ell}^- = \frac{1}{f_{\ell m}^+ f_{m\ell}^+}$$

So far the conditions (i) and (ii) have been used. Since the condition (iii) is equivalent to $-1 < f_{\ell m}f_{m\ell} < 1$ we see that only the + solution satisfies (iii). ■

In the sequel we shall refer to the method defined by (2.2.28), (2.2.30), (2.2.34) and (2.2.35) as to the method due to S.Falk and P.Langemeyer.

## 2.3. Fast Scaled Plane Transformations

In [15] an idea due to W.M.Gentleman (see [7]) was used in

order to economize the operational count in the similarity transformations with plane rotations. A similar economization is possible here, too. We illustrate it on the special Jacobi method.

We have the sequences

$$A^{(1)}, \quad A^{(2)} = F_1^* A^{(1)} F_1, \quad A^{(3)} = F_2^* A^{(2)} F_2, \ldots$$

$$B^{(1)}, \quad B^{(2)} = F_1^* B^{(1)} F_1, \quad B^{(3)} = F_2^* B^{(2)} F_2, \ldots$$

where $F_1, F_2, \ldots$ are plane transformations from (2.2.3). The (accumulated) transformation is

$$F^{(k)} = F_1 \ldots F_k = F^{(k-1)} F_k, \quad k \geq 1.$$

From the formulae for $\cos\varphi_k$ and $\cos\psi_k$ we see that

$$\min \{ f_{\ell\ell}^{(k)}, f_{mm}^{(k)} \} > 0, \quad k \geq 1,$$

hence we can write

$$F_k = T_k \Delta_k, \quad k \geq 1,$$

where

$$\hat{T}_k = \begin{bmatrix} 1 & f_{\ell m}^{(k)}/f_{mm}^{(k)} \\ f_{m\ell}^{(k)}/f_{\ell\ell}^{(k)} & 1 \end{bmatrix},$$

$$\hat{\Delta}_k = \begin{bmatrix} f_{\ell\ell}^{(k)} & 0 \\ 0 & f_{mm}^{(k)} \end{bmatrix}, \quad k \geq 1.$$

For $k = 1,2,\ldots$ we write

$$D^{(k+1)} = \Delta_1 \ldots \Delta_k = D^{(k)} \Delta_k, \quad D^{(1)} = I_n,$$

$$\tilde{T}_k = D^{(k)} T_k D^{(k)-1}$$

$$\tilde{T}^{(k)} = \tilde{T}_1 \ldots \tilde{T}_k = \tilde{T}^{(k-1)} \tilde{T}_k,$$



and form new sequences of matrices

$$M^{(k)} = D^{(k)-1} A^{(k)} D^{(k)-1}, \quad N^{(k)} = D^{(k)-1} B^{(k)} D^{(k)-1}.$$

Then we have

$$\begin{aligned} \tilde{T}^{(k)} &= T_1 D^{(1)-1} D^{(2)} T_2 D^{(2)-1} D^{(3)} T_3 \ldots \tilde{D}^{(k-1)} D^{(k)} T_k D^{(k)-1} = \\ &= T_1 \Delta_1 T_2 \Delta_2 \ldots \Delta_{k-1} T_k D^{(k)-1} D^{(k+1)} D^{(k+1)-1} = \\ &= F_1 F_2 \ldots F_k D^{(k+1)-1} = \\ &= F^{(k)} D^{(k+1)-1}, \quad k \geq 1, \\ M^{(k+1)} &= D^{(k+1)-1} A^{(k+1)} D^{(k+1)-1} = \\ &= D^{(k+1)-1} \Delta_k T_k^* A^{(k)} T_k \Delta_k D^{(k+1)-1} = \\ &= D^{(k)-1} T_k^* D^{(k)} M^{(k)} D^{(k)} T_k D^{(k)-1} = \\ &= \tilde{T}_k^* M^{(k)} \tilde{T}_k, \quad k \geq 1, \end{aligned}$$

and also

$$N^{(k+1)} = \tilde{T}_k^* N^{(k)} \tilde{T}_k, \quad k \geq 1.$$

Thus, the sequences $M^{(k)}$, $N^{(k)}$, $T^{(k)}$ are obtained by the algorithm

$$\left.\begin{array}{l} D^{(1)} = I_n, \quad M^{(1)} = A^{(1)}, \quad N^{(1)} = B^{(1)}, \quad \tilde{T}^{(1)} = D^{(1)} \\ \tilde{T}_k = D^{(k)} T_k D^{(k)-1}, \quad D^{(k+1)} = D^{(k)} \Delta_k \\ \left\{ \begin{array}{ll} M^{(k+1)} = \tilde{T}_k^* M^{(k)} \tilde{T}_k, & N^{(k+1)} = \tilde{T}_k^* N^{(k)} \tilde{T}_k \\ \tilde{T}^{(k+1)} = \tilde{T}^{(k)} \tilde{T}_k. & \end{array} \right. \end{array} \right\} k \geq 1$$

The $\ell$-part of the algorithm which carries the main part of the operational count is obviously less expensive as the previous method. In fact this new algorithm is a sort of a complex generalization of the algorithm of S.Falk and P.Langemeyer. Note that

$$\tilde{T}_k = \begin{bmatrix} 1 & & * \\ & & \\ * & & 1 \end{bmatrix}, \quad k \geq 1 .$$

In order to skip the first normalization $(A,B) \longmapsto (A^{(1)},B^{(1)})$ we can set $M^{(1)} = A$, $N^{(1)} = B$ and

$$D^{(1)} = D_o = \operatorname{diag}\,(1/\sqrt{b_{11}},\dots,1/\sqrt{b_{nn}}) .$$

The matrix $D^{(k)}$ is actually not needed in the process since it can be easily computed from the diagonal of $N^{(k)}$. The pairs $(A^{(k)},B^{(k)})$ and $(M^{(k)},N^{(k)})$ are for each $k \geq 1$ equivalent hence the eigenvalues are ultimately obtained from the limit matrices $M^{(\infty)}$ and $N^{(\infty)}$. The eigenvectors are obtained from $T^{(\infty)}$ and $N^{(\infty)}$.

## The Case of Real Matrices

Here we prove that the Jacobi method due to K.Zimmermann and the Jacobi method due to S.Falk and P.Langemeyer are simply related.

2.7. Theorem. Let $A$, $B$ be symmetric matrices such that $B$ is positive definite and let the sequences $((A^{(k)},B^{(k)}),k\geq 1)$ and $((A^{(k)'},B^{(k)'}),k\geq 1)$ be generated by the method due to K. Zimmermann and by the method due to S.Falk and P.Langemeyer, respectively. If the corresponding pivot strategies are the same then

$$A^{(k)} = D^{(k)}A^{(k)'}D^{(k)}, \qquad B^{(k)} = D^{(k)}B^{(k)'}D^{(k)}, \quad k\geq 1 \tag{2.3.1}$$

where each $D^{(k)}$ is a diagonal matrix with positive diagonal elements.



Proof: From the second equality in (2.3.1) we see that

$$D^{(k)} = \operatorname{diag}\,(1/\sqrt{b_{11}^{(k)'}},\dots,1/\sqrt{b_{nn}^{(k)'}}), \quad k\geq 1, \tag{2.3.2}$$

where $b_{ii}^{(k)'}$ $(1\leq i\leq n)$ are the diagonal elements of $B^{(k)'}$.

We prove the relation (2.3.1) by induction with respect to $k$. For $k=1$ we have

$$A^{(1)} = D_oAD_o, \qquad B^{(1)} = D_oBD_o,$$
$$A^{(1)'} = A, \qquad B^{(1)'} = B,$$

where

$$D_o = \operatorname{diag}\,(1/\sqrt{b_{11}},\dots,1/\sqrt{b_{nn}}) .$$

Thus $D^{(1)} = D_o$ and (2.3.1) holds for $k=1$.

Suppose that the relation (2.3.1) holds for a fixed $k$ $(k\geq 1)$. We shall prove that then it holds for $k+1$.

The $k$'th step of the two processes are described by the following equalities:

$$\begin{aligned} A^{(k+1)} &= F_k^* A^{(k)} F_k, & B^{(k+1)} &= F_k^* B^{(k)} F_k , \\ A^{(k+1)'} &= F_k'^* A^{(k)'} F_k', & B^{(k+1)'} &= F_k'^* B^{(k)'} F_k' , \end{aligned} \tag{2.3.3}$$

where $F_k$ and $F_k'$ are defined by the relations (2.2.18)–(2.2.21) and by the relations (2.2.26), (2.2.28), (2.2.30), (2.2.34), (2.2.35), respectively. Define the elementary plane matrix $\Delta_k$ by its $(\ell,m)$-restriction

$$\hat{\Delta}_k = \frac{1}{\sqrt{1-(b_{\ell m}^{(k)})^2}} \begin{bmatrix} \cos\tilde{\varphi}_k & 0 \\ 0 & \cos\tilde{\psi}_k \end{bmatrix},$$

where $\tilde{\varphi}_k$ and $\tilde{\psi}_k$ are as in (2.2.19). Then

$$F_k = T_k \Delta_k \ , \qquad (2.3.4)$$

where

$$\hat{T}_k = \begin{bmatrix} 1 & \sin\tilde{\varphi}_k/\cos\tilde{\psi}_k \\ -\sin\tilde{\psi}_k/\cos\tilde{\varphi}_k & 1 \end{bmatrix} . \qquad (2.3.5)$$

The relations (2.3.3) and (2.3.4) imply

$$A^{(k+1)} = (D^{(k)}\Delta_k)^{*}\tilde{T}_k^{*}A^{(k)\prime}\tilde{T}_k(D^{(k)}\Delta_k) , \qquad (2.3.6)$$

where $\tilde{T}_k$ is defined by

$$\tilde{T}_k = \hat{D}^{(k)}\hat{T}_k\hat{D}^{(k)-1} = \begin{bmatrix} 1 & \sqrt{\frac{b_{mm}^{(k)\prime}}{b_{\ell\ell}^{(k)\prime}}}\,\frac{\sin\tilde{\varphi}_k}{\cos\tilde{\psi}_k} \\ \sqrt{\frac{b_{\ell\ell}^{(k)\prime}}{b_{mm}^{(k)\prime}}}\,\frac{\sin\tilde{\psi}_k}{\cos\tilde{\varphi}_k} & 1 \end{bmatrix} \qquad (2.3.7)$$

In (2.3.7) we have used the relations (2.3.5) and (2.3.2). By the assertion (2.2.22) of Proposition 2.4 the matrix $D^{(k)}\Delta_k$ is diagonal with positive diagonal elements. Since the relation (2.3.6) holds with $B^{(k+1)}, B^{(k)\prime}$ instead of $A^{(k+1)}, A^{(k)\prime}$ it suffices to prove

$$\tilde{T}_k = F_k' . \qquad (2.3.8)$$

By the relation (2.2.33) and by the assertion (2.2.23) of Proposition 2.4 we have

$$0 < \det F_k' \leq 2 , \qquad 0 < \det \tilde{T}_k \leq 2 . \qquad (2.3.9)$$

The equality (2.3.8) will follow directly from the relation (2.3.9) and Proposition 2.6 if we prove

(i) $\det \tilde{T}_k = 2$ iff $\det F_k' = 2$

and

(ii) $\tilde{T}_k = F_k'$ if $\det \tilde{T}_k = \det F_k' = 2$ .



To prove (i) we use the following chain of equivalent conditions :

$$\det \tilde{T}_k = 2 \iff \operatorname{tg}\tilde{\varphi}_k \operatorname{tg}\tilde{\psi}_k = 1 \iff \tilde{\vartheta}_k = \frac{\pi}{4}$$

$$\iff \det \begin{bmatrix} a_{\ell\ell}^{(k)} & a_{mm}^{(k)} \\ b_{\ell\ell}^{(k)} & b_{mm}^{(k)} \end{bmatrix} = 0 \iff$$

$$\iff \det \begin{bmatrix} a_{\ell\ell}^{(k)\prime} & a_{mm}^{(k)\prime} \\ b_{\ell\ell}^{(k)\prime} & b_{mm}^{(k)\prime} \end{bmatrix} = 0 \iff$$

$$\iff f_{\ell m}^{(k)\prime} f_{m\ell}^{(k)\prime} = -1 \iff \det F_k' = 2 .$$

Here we have used: the relation (2.3.7), the assertion (2.2.24) of Proposition 2.4, the relation (2.2.21), the induction hypothesis (2.3.1) and the definition (2.2.35) of $F_k'$, respectively.

To prove (ii) we use the definition (2.2.35) and the relations (2.2.19), (2.3.7) in the case $\tilde{\vartheta}_k = \frac{\pi}{4}$ . We obtain $\sin\tilde{\varphi}_k = \cos\tilde{\psi}_k$ and $\sin\tilde{\psi}_k = \cos\tilde{\varphi}_k$ hence it follows $T_k = F_k'$ . ∎

In conclusion, if the fast scaled transformations are used then the Jacobi method due to K.Zimmermann reduces to the method of S.Falk and P.Langemeyer.

## 2.4. Global Convergence

Here we present the global convergence results concerning

the Jacobi methods defined in Section 2.1 and Section 2.2. In order to keep this thesis within reasonable length we give here only the main results and omit the proofs. They can be found in [10].

A Jacobi method is globally convergent if the obtained sequences $(A^{(k)})$, $(B^{(k)})$ are convergent for every initial pair $(A, B)$, and the limit matrices are diagonal.

2.8. Theorem. Let $e, E$ be real numbers such that $0 < e \leq E$. Let the sequence $((A^{(k)},B^{(k)}), k \geq 1)$ be generated by a general Jacobi method defined by $e, E$ and the row- or the column-cyclic pivot strategy. Denote

$$f^{(k)} = \max\{|f_{\ell\ell}^{(k)}|, |f_{mm}^{(k)}|\}, \qquad k \geq 1,$$

where $f_{\ell\ell}^{(k)}$, $f_{mm}^{(k)}$ are the elements of the transformation matrix $F_k$ satisfying the equalities (2.1.2). If

$$\lim_{k\to\infty} \inf f^{(k)} > 0 \tag{2.4.1}$$

then

$$\lim_{k\to\infty} S(A^{(k)},B^{(k)}) = 0$$

where $S(\cdot,\cdot)$ is defined by the relation (1.2.23). In addition, if

$$\lim_{k\to\infty} b_{ii}^{(k)} = d_{ii}, \qquad 1 \leq i \leq n$$

then there is an ordering of the eigenvalues for which

$$\lim_{k\to\infty} a_{ii}^{(k)} = \lambda_i d_{ii}, \qquad 1 \leq i \leq n.$$

Proof: The proof of Theorem 2.8 can be found in [10]. The proof is based on some new results concerning the convergence of cyclic Jacobi-like processes (see [11]). ∎



2.9. Corollary. The row- and the column-cyclic special Jacobi method is globally convergent. For each initial pair $(A,B)$ there is an ordering of the eigenvalues such that

$$\lim_{k\to\infty} A^{(k)} = \operatorname{diag}(\lambda_1,\dots,\lambda_n). \tag{2.4.2}$$

Proof: The proof follows from Theorem 2.7 and the fact that $\lim_{k\to\infty} b_{\ell m}^{(k)} = 0$ (see [10]). ∎

To see that the condition (2.4.1) is not superfluous consider the simple case: $B = I_n$. Then (2.4.1) reads

$$f_{\ell\ell}^{(k)} = f_{mm}^{(k)} = \cos\varphi_k \geq \nu, \qquad k \geq k_0, \tag{2.4.3}$$

where $k_0$ is a positive integer and $\nu$ a real positive number. In [6] G.E.Forsythe and P.Henrici have proved that (2.4.3) is a sufficient and necessary condition for the global convergence of the row- and the column-cyclic Jacobi method for hermitian matrices.

2.10. Remark. In the real case Theorem 2.8.can be applied to a row- or/and column-cyclic general Jacobi method defined by G.Gose [8] or in the special case to the Jacobi method due to K.Zimmermann (see relations (2.2.18)–(2.2.21)). In conclusion, the method due to K.Zimmermann is globally convergent (see [10]).

In [25] K.Zimmermann proved that

$$\lim_{k\to\infty} S(A^{(k)},B^{(k)}) = 0 \quad \text{provided} \quad 4Nt(1+t)^{N-1} < 2^{-(N-n+1)},$$

where $t = S^2(B^{(1)})(1 + S(B^{(1)}))^2(2^N-1)$.

2.11. Remark. Unfortunately Theorem 2.8 cannot be directly applied to the method of S.Falk and P.Langemeyer since the existence of the constants $e,E$ such that $0 < e \leq b_{ii}^{(k)} \leq E$

$(1 \le i \le n,\ k \ge 1)$ holds has not yet been proved. However, using Theorem 2.7 and the global convergence of the Jacobi method due to K.Zimmermann we have

$$\lim_{k\to\infty} \frac{b_{ij}^{(k)}}{\sqrt{b_{ii}^{(k)} b_{jj}^{(k)}}} = 0, \qquad \lim_{k\to\infty} \frac{a_{ij}^{(k)}}{\sqrt{b_{ii}^{(k)} b_{jj}^{(k)}}} = 0, \quad i \neq j,$$

and

$$\lim_{k\to\infty} \frac{a_{ii}^{(k)}}{b_{ii}^{(k)}} = \lambda_i, \qquad 1 \le i \le n,$$

where $\lambda_1, \dots, \lambda_n$ is an ordering of the eigenvalues depending on the initial pair $(A,B)$. Note that the first two equalities are equivalent to

$$\lim_{k\to\infty} S(A^{(k)}, B^{(k)}) = 0$$

(see the definition of $S(\cdot,\cdot)$ in (1.2.23)).



## 3. ASYMPTOTIC CONVERGENCE

### 3.1. Assumptions and Notation

In Chapter 3 we study the asymptotic convergence of the special Jacobi method derived in Section 2.2. The central result is presented in Section 3.3 where the quadratic convergence of the cyclic method is proved in the case of simple eigenvalues. In Section 3.4 we present some new estimates obtained in the case of multiple eigenvalues. In Section 3.5 we use a qualitative analysis of the algorithm in the case of multiple eigenvalues to show that the quadratic convergence cannot be proved in the general case. In Section 3.6 we show that the method can be asymptotically modified to be always quadratically convergent. Finally in Section 3.8 we present the results of the numerical investigation of the asymptotic convergence.

Throughout this chapter we use the following notation:

$$\begin{aligned}
&a_k = |a_{\ell m}^{(k)}|, \qquad b_k = |b_{\ell m}^{(k)}|,\\
&x_k = 1/(1-b_k), \qquad y_k = 1/(1-b_k^2), \qquad t_k = \sqrt{1-b_k^2},\\
&b = \max_{4\le k\le N} b_k, \qquad e_k = a_{mm}^{(k)} - a_{\ell\ell}^{(k)}, \qquad \sigma_k = \begin{cases} 1, & e_k \ge 0 \\ -1, & e_k < 0 \end{cases},\\
&e^{-i \arg (b_{\ell m}^{(k)})} a_{\ell m}^{(k)} = u_k + i v_k,\\
&F_k = R_1^{(k)} D^{(k)} R_2^{(k)} \Phi^{(k)},\\
&\varepsilon_k = S(A^{(k)}, B^{(k)}), \qquad \tau_k = \tau(A^{(k)}, B^{(k)}).
\end{aligned} \tag{3.1.1}$$

Here $N$ is defined by (2.1.3) while $S(\cdot,\cdot)$ and $\tau(\cdot,\cdot)$ are

as in (1.2.23) and (1.2.24). Note that $\tau_k$ is defined only for almost diagonal pairs. The $(\ell,m)$-restriction of $F_k$ has the form

$$\hat{F}_k = \frac{1}{t_k}\begin{bmatrix} \cos\varphi_k & e^{i\alpha_k}\sin\varphi_k \\ -e^{-i\beta_k}\sin\psi_k & \cos\psi_k \end{bmatrix}, \qquad (3.1.2)$$

where $t_k$ is defined by (3.1.1) and

$$2\cos^2\varphi_k = 1 + b_k\sin 2\vartheta_k + t_k\cos 2\vartheta_k\cos\gamma_k, \quad \varphi_k\in[0,\tfrac{\pi}{2})$$

$$2\cos^2\psi_k = 1 - b_k\sin 2\vartheta_k + t_k\cos 2\vartheta_k\cos\gamma_k, \quad \psi_k\in[0,\tfrac{\pi}{2})$$

$$e^{i\alpha_k}\sin\varphi_k = e^{i\arg(b_{\ell m}^{(k)})}[\sin 2\vartheta_k - b_k - i\,t_k\cos 2\vartheta_k\sin\gamma_k]/(2\cos\psi_k)$$

$$e^{-i\beta_k}\sin\psi_k = e^{-i\arg(b_{\ell m}^{(k)})}[\sin 2\vartheta_k + b_k + i\,t_k\cos 2\vartheta_k\sin\gamma_k]/(2\cos\varphi_k) \qquad (3.1.3)$$

holds. Here the angles $\gamma_k$ and $\vartheta_k$ are determined from

$$\operatorname{tg}\gamma_k = 2\frac{v_k}{|e_k|}, \qquad -\frac{\pi}{2} < \gamma_k \le \frac{\pi}{2}, \qquad (3.1.4)$$

and

$$\operatorname{tg} 2\vartheta_k = \sigma_k\frac{2u_k - (a_{\ell\ell}^{(k)} + a_{mm}^{(k)})\,b_k}{\sqrt{e_k^2 + 4v_k^2}\cdot t_k},$$

$$-\frac{\pi}{4} < \vartheta_k \le \frac{\pi}{4}. \qquad (3.1.5)$$

In the relations (3.1.3)–(3.1.5) the notation from (3.1.1) has been used.

If $b_k = 0$ and $a_k = 0$ then the k'th step is skipped, i.e. we set $\vartheta_k = \gamma_k = 0$. If $e_k = 0$ then $\vartheta_k = \frac{\pi}{4}$. If



$b_k = 0$ and $a_k \neq 0$ then $\arg(b_{\ell m}^{(k)})$ in the relation (3.1.3) is replaced by $\arg(a_{\ell m}^{(k)})$, so that $u_k = a_k$, $v_k = 0$. In this case the formulae above are reduced to the formulae for the standard Jacobi algorithm for the matrix $\hat{A}^{(k)}$.

Let $((A^{(k)},B^{(k)}),k\ge 1)$ be the sequence of pairs generated by a cyclic special Jacobi method when applied to a pair $(A,B)$. If

$$\lim_{k\to\infty}\varepsilon_k = 0 \qquad (3.1.6)$$

and if an integer $r_0 \ge 1$ and a constant $c_0 > 0$ exist such that

$$\varepsilon_{(r+1)N+1} \le c_0\varepsilon_{rN+1}^2, \qquad r \ge r_0, \qquad (3.1.7)$$

then the method is <u>quadratically convergent</u> on the pair $(A,B)$. Here $\varepsilon_k$ is defined by (3.1.1). From the relations (3.1.7) and (3.1.6) we see that $\varepsilon_k$ reduces quadratically per cycle provided it is "sufficiently small". In the case of a general cyclic pivot strategy it has not been proved that the relation (3.1.6) holds, but still (3.1.7) can be proved provided $\varepsilon_1$ is small enough. In such a case the relation (3.1.6) is implied by (3.1.7) provided $\varepsilon_1$ is small enough. Therefore it is essential to find the conditions under which (3.1.7) holds. Such conditions are called <u>asymptotic assumptions</u>.

In this chapter we shall frequently use the following two asymptotic assumptions:

$$S(B^{(1)}) < \frac{1}{2N} \qquad (3.1.8)$$

and

$$2\sqrt{1+\mu^2}\,\varepsilon_1 < \delta \qquad (3.1.9)$$

where $\mu = \operatorname{spr}(A,B)$ and $\delta$ are defined by the relations (1.1.3) and (1.2.6). Throughout this thesis it is assumed

$$p \geq 2 \quad \text{and} \quad n \geq 3, \tag{3.1.10}$$

where $p$ is the number of distinct eigenvalues (see (1.2.3)).

## 3.2. Preliminaries

Here we prove some auxiliary results concerning the Jacobi method defined by the relations (3.1.2)–(3.1.5). The pivot strategy is arbitrary if not stated otherwise.

3.1. Lemma. Let $r$ be an integer such that $r \geq 3$ and let $x$ be a nonnegative real number such that

$$2rx < 1.$$

Then

(i) $(1-x)^{-r} \leq 1 + \frac{12}{7} r x$

(ii) $(1-x^2)^{-r} \leq 1 + \frac{72}{67} r x^2$

(iii) $(1+x)^r \leq 1 + \frac{4}{3} r x$

(iv) $(1-x)^{-1/2} \leq 1 + \frac{3}{5} x$.

Proof: (i) Let $(1-x)^{-1} = 1 + \zeta$. Then $\zeta = x(1-x)^{-1} \leq \frac{6}{5} x$. Since $x < \frac{1}{6}$ and $r\zeta < \frac{3}{5}$, we have

$$(1-x)^{-r} = (1+\zeta)^r =$$



$$= 1 + r\zeta\left[1 + \frac{r-1}{2}\zeta + \frac{(r-1)(r-2)}{2\cdot 3}\zeta^2 + \dots + \zeta^{r-2}\right] + \zeta^r \leq$$

$$\leq 1 + r\zeta\left[1 + \left(\frac{r-1}{2}\zeta\right) + \left(\frac{r-1}{2}\zeta\right)^2 + \dots\right] =$$

$$= 1 + \frac{r\zeta}{1 - \frac{r-1}{2}\zeta} \leq 1 + \frac{6}{5}\frac{rx}{1 - \frac{3}{5} r x} \leq 1 + \frac{12}{7} r x.$$

The proofs of the other assertions of Lemma 3.1 are obtained in a similar way. In the proof of (iv) one uses the expansion of the function $(1-x)^{-1/2}$. ∎

3.2. Lemma. Let $a_k$, $b_k$, $\varepsilon_k$ and $\hat{F}_k$ be defined by the relations (3.1.1) and (3.1.2). Then

$$\varepsilon_{k+1} \leq \max\left\{1, \|\hat{F}_k\|_2^2\right\} \cdot \left[\varepsilon_k^2 - 2(a_k^2 + b_k^2)\right], \quad k \geq 1.$$

Proof: For $k \geq 1$ set

$$a^\ell = (a_{\ell 1}^{(k)}, a_{\ell 2}^{(k)}, \dots, a_{\ell \ell-1}^{(k)}, a_{\ell \ell+1}^{(k)}, \dots, a_{\ell m-1}^{(k)}, a_{\ell m+1}^{(k)}, \dots, a_{\ell n}^{(k)}),$$

$$a^m = (a_{m1}^{(k)}, a_{m2}^{(k)}, \dots, a_{m\ell-1}^{(k)}, a_{m\ell+1}^{(k)}, \dots, a_{mm-1}^{(k)}, a_{mm+1}^{(k)}, \dots, a_{mn}^{(k)}),$$

$$a_\ell = (a_{1\ell}^{(k)}, a_{2\ell}^{(k)}, \dots, a_{\ell-1\ell}^{(k)}, a_{\ell+1\ell}^{(k)}, \dots, a_{m-1\ell}^{(k)}, a_{m+1\ell}^{(k)}, \dots, a_{n\ell}^{(k)})^T,$$

$$a_m = (a_{1m}^{(k)}, a_{2m}^{(k)}, \dots, a_{\ell-1m}^{(k)}, a_{\ell+1m}^{(k)}, \dots, a_{m-1m}^{(k)}, a_{m+1m}^{(k)}, \dots, a_{nm}^{(k)})^T,$$

where $a^T$ is the transpose of $a$. Let $a^{\ell'}$, $a^{m'}$, $a_\ell'$, $a_m'$ be the row- and the column-vectors built in the same way, but from the elements of $A^{(k+1)}$.

The transformation

$$A^{(k+1)} = F_k^* A^{(k)} F_k$$

implies

$$\begin{bmatrix} a^{\ell'} \\ a^{m'} \end{bmatrix} = \hat{F}_k^* \begin{bmatrix} a^\ell \\ a^m \end{bmatrix}, \qquad [a_\ell', a_m'] = [a_\ell, a_m] \hat{F}_k$$

hence we obtain

$$\left\| \begin{bmatrix} a\ell' \\ a^{m'} \end{bmatrix} \right\|^2 \leq \|\hat{F}_k^*\|_2^2 \left\| \begin{bmatrix} a^\ell \\ a^m \end{bmatrix} \right\|^2 ,$$

$$\|[a_{\ell'}, a_{m'}]\|^2 \leq \|\hat{F}_k^*\|_2^2 \cdot \|[a_\ell, a_m]\|^2 .$$

All the off-diagonal elements of $A^{(k)}$ changing in the k'th step are those contained in the row-vectors $a^\ell$, $a^m$ and column-vectors $a_\ell$, $a_m$, except $a_{\ell m}^{(k)}$ and $a_{m\ell}^{(k)}$ which are annihilated. Since $\|\hat{F}_k^*\|_2 = \|\hat{F}_k\|_2$ we have

$$S^2(A^{(k+1)}) \leq S^2(A^{(k)}) + (\|\hat{F}_k\|_2^2 - 1)(\|a^\ell\|^2 + \|a^m\|^2 + \|a_\ell\|^2 + \|a_m\|^2) - 2a_k^2 \leq$$

$$\leq \max\left\{1, \|\hat{F}_k\|_2^2\right\} \cdot (S^2(A^{(k)}) - 2a_k^2) . \qquad (3.2.1)$$

The same analysis applies to $B^{(k)}$, hence we have

$$S^2(B^{(k+1)}) \leq \max\left\{1, \|\hat{F}_k\|_2^2\right\} (S^2(B^{(k)}) - 2b_k^2). \qquad (3.2.2)$$

Since $b_{ii}^{(k)} = 1$ $(1 \leq i \leq n;\ k \geq 1)$ we have

$$\varepsilon_k^2 = S^2(A^{(k)}, B^{(k)}) = S^2(A^{(k)}) + S^2(B^{(k)}) ,$$

$$k \geq 1 . \qquad (3.2.3)$$

Lemma 3.2 follows directly from the relations (3.2.3), (3.2.1) and (3.2.2). ■

3.3. Lemma. Let the assumptions (3.1.8) and (3.1.10) hold, and let $a_k$, $b_k$, $b$, $\varepsilon_k$, $x_k$ be as in (3.1.1). Then

$$b < 0.88 \frac{1}{2N} , \qquad (3.2.4)$$

$$\sum_{k=1}^{N} (a_k^2 + b_k^2) \leq 0.877\, \varepsilon_1^2 , \qquad (3.2.5)$$



$$\begin{bmatrix} S^2(A^{(k+1)}) \\ S^2(B^{(k+1)}) \\ \varepsilon_{k+1}^2 \end{bmatrix} \leq x_1 x_2 \cdots x_k \begin{bmatrix} S^2(A^{(1)}) \\ S^2(B^{(1)}) \\ \varepsilon_1^2 \end{bmatrix} \leq$$

$$\leq 1.754 \begin{bmatrix} S^2(A^{(1)}) \\ S^2(B^{(1)}) \\ \varepsilon_1^2 \end{bmatrix}, \quad 1 \leq k \leq N . \qquad (3.2.6)$$

Proof: From the representation $F_k = R_1^{(k)} D^{(k)} R_2^{(k)} \Phi^{(k)}$ (see (3.1.1)) we can conclude that

$$\|F_k\|_2^2 \leq \|D^{(k)}\|_2^2 = 1/(1-b_k) = x_k , \quad k \geq 1. \qquad (3.2.7)$$

Using Lemma 3.2 and the inequality (3.2.7) we have

$$\varepsilon_{k+1}^2 \leq x_k\left[\varepsilon_k^2 - 2(a_k^2 + b_k^2)\right] \leq$$

$$\leq x_k\left\{x_{k-1}\left[\varepsilon_{k-1}^2 - 2(a_{k-1}^2 + b_{k-1}^2)\right] - 2(a_k^2 + b_k^2)\right\} \leq \dots$$

$$\dots \leq x_k x_{k-1} \cdots x_1 \varepsilon_1^2 - 2\sum_{j=1}^{k} x_k x_{k-1} \cdots x_j (a_j^2 + b_j^2) \leq$$

$$\leq (1-b)^{-k} \varepsilon_1^2 - 2\sum_{j=1}^{k} (a_j^2 + b_j^2) , \quad 1 \leq k \leq N . \qquad (3.2.8)$$

Thus

$$\varepsilon_{N+1}^2 \leq (1-b)^{-N} \varepsilon_1^2 - 2\sum_{k=1}^{N} (a_k^2 + b_k^2)$$

and since $\varepsilon_{N+1} \geq 0$ we have

$$\sum_{k=1}^{N} (a_k^2 + b_k^2) \leq \frac{1}{2} (1-b)^{-N} \varepsilon_1^2 . \qquad (3.2.9)$$

To prove the inequality (3.2.4) we set $\varrho^{(k)} = S(B^{(k)})/\sqrt{2}$, $k \geq 1$. Then

$$b_k \leq \varrho^{(k)} , \quad k \geq 1, \qquad (3.2.10)$$

and the inequality (3.2.2) yields

$$\varrho^{(k+1)} \leq \sqrt{x_k}\, \varrho^{(k)} \leq \frac{\varrho^{(k)}}{\sqrt{1 - \varrho^{(k)}}} , \quad k \geq 1 . \qquad (3.2.11)$$

By induction with respect to $k$ we obtain from (3.2.11)

$$\varrho^{(k+1)} \le \frac{\varrho^{(1)}}{\sqrt{1-k\varrho^{(1)}}} \text{ , provided } k\varrho^{(1)} < 1 .$$

The assumption (3.1.8) yileds $N\varrho^{(1)} < \sqrt{2}/4$, hence

$$\varrho^{(k+1)} \le \frac{\frac{\sqrt{2}}{4N}}{\sqrt{1-k\frac{\sqrt{2}}{4N}}} \le \sqrt{\frac{4+\sqrt{2}}{7}} \cdot \frac{1}{2N} , \quad 1 \le k \le N. \qquad (3.2.12)$$

The inequalities (3.2.10) imply $b \le \max \varrho^{(k)}$, and since

$$\sqrt{\frac{4+\sqrt{2}}{7}} < 0.88$$

the assertion (3.2.4) of Lemma 3.3 follows from (3.2.12).

To prove the inequality (3.2.5), we note that the inequality (3.2.4) implies $2Nb < 1$, hence Lemma 3.1(i) can be applied to obtain

$$(1-b)^{-N} \le 1 + \frac{12}{7}Nb < 1 + \frac{6}{7}\sqrt{\frac{4+\sqrt{2}}{7}} < 1.754 . \qquad (3.2.13)$$

Now the relation (3.2.5) follows from the inequalities (3.2.9) and (3.2.13).

Finally, the inequalities (3.2.6) follow from the relations (3.2.1), (3.2.2), (3.2.8) and (3.2.13). ∎

3.4. Lemma. Let the assumptions (3.1.8), (3.1.9) and (3.1.10) hold and let $\delta$ be as in (1.2.6). Then for each pair of indices $(i,j)$, $i < j$, and for each $k$ such that $1 \le k \le N+1$ either

$$|a_{jj}^{(k)} - a_{ii}^{(k)}| > 2.56\delta \qquad (3.2.14)$$

or

$$|a_{jj}^{(k)} - a_{ii}^{(k)}| < 0.44\delta. \qquad (3.2.15)$$



Proof: By the assumption (3.1.9) and the assertion (3.2.6) of Lemma 3.3 we have

$$(1+\mu^2)\varepsilon_k^2 \le 1.754(1+\mu^2)\varepsilon_1^2 < \frac{1}{4}\cdot 1.754\delta^2 < 0.44\delta^2, \quad 1 \le k \le N+1. \qquad (3.2.16)$$

From (3.2.16) we see that the assumption (1.2.30) of Corollary 1.3 is fulfilled for each $k$ satisfying $1 \le k \le N+1$. By Corollary 1.3 and by (3.2.16) we have

$$2\sum_{r=1}^{n} |a_{rr}^{(k)} - \lambda_r^{(k)}|^2 \le [(1+\mu^2)\varepsilon_k^2/\delta]^2 < (0.44\delta)^2, \quad 1 \le k \le N+1, \qquad (3.2.17)$$

where $\lambda_1^{(k)}, \dots, \lambda_n^{(k)}$ is an ordering of the eigenvalues depending on $k$. Let $i<j$ and $1 \le k \le N+1$. If $\lambda_i^{(k)} \ne \lambda_j^{(k)}$ then using the definition of $\delta$ (1.2.6) and the inequality (3.2.17) we obtain

$$|a_{jj}^{(k)} - a_{ii}^{(k)}| \ge |\lambda_i^{(k)} - \lambda_j^{(k)}| - |a_{ii}^{(k)} - \lambda_i^{(k)}| - |\lambda_j^{(k)} - a_{jj}^{(k)}| \ge$$
$$\ge 3\delta - \sqrt{2|a_{ii}^{(k)} - \lambda_i^{(k)}|^2 + 2|a_{jj}^{(k)} - \lambda_j^{(k)}|^2} >$$
$$> 3\delta - 0.44\delta \ge 2.56\delta ,$$

hence (3.2.14) is proved.

If $\lambda_i^{(k)} = \lambda_j^{(k)}$ then using (3.2.17) we obtain

$$|a_{jj}^{(k)} - a_{ii}^{(k)}| \le |a_{jj}^{(k)} - \lambda_j^{(k)} + \lambda_i^{(k)} - a_{ii}^{(k)}| \le$$
$$\le |a_{jj}^{(k)} - \lambda_j^{(k)}| + |\lambda_i^{(k)} - a_{ii}^{(k)}| \le$$
$$\le \sqrt{2|a_{jj}^{(k)} - \lambda_j^{(k)}|^2 + 2|a_{ii}^{(k)} - \lambda_i^{(k)}|^2} < 0.44\delta ,$$

hence also the second alternative (3.2.15) is proved. ∎

By Lemma 3.4 we see that the set

$$S' = \{ k \in \{1,2,\dots,N\};\ |a_{mm}^{(k)} - a_{\ell\ell}^{(k)}| > 2\delta \}$$

is well defined for each pivot strategy, provided the assumptions (3.1.8), (3.1.9) and (3.1.10) hold. In the sequel we use the notation $\sum_k'$ instead of $\sum_{k \in S'}$, for simplicity.

3.5. Lemma. Let the assumptions (3.1.8), (3.1.9) and (3.1.10) hold and let $\varphi_k, \psi_k, e_k, \delta$ and $\mu$ be defined by the relations (3.1.3), (3.1.1), (1.2.6) and (1.1.3). Then

$$\sum_{k=1}^{N}{}' \sin^2 \omega_k \leq 0.62(1+\mu^2)\frac{e_1^2}{\delta^2} \qquad (3.2.18)$$

where $\omega_k \in \{\varphi_k, \psi_k\}$, $1 \leq k \leq N$. If all the eigenvalues $\lambda_i$ are simple then the sum $\sum_k'$ reduces to the usual sum $\sum_k$ and the constant 0.62 in the inequality (3.2.18) can be replaced by 0.4852.

Proof: Using the Cauchy–Schwartz inequality, we obtain from (3.1.5)

$$\operatorname{tg}^2 2\vartheta_k \leq \frac{4 + (a_{\ell\ell}^{(k)} + a_{mm}^{(k)})^2}{(e_k^2 + 4v_k^2)\cdot t_k^2}(u_k^2 + b_k^2) \leq$$

$$\leq \frac{4 + (a_{\ell\ell}^{(k)} + a_{mm}^{(k)})^2}{e_k^2 \cdot t_k^2}(a_k^2 + b_k^2), \quad k \geq 1. \qquad (3.2.19)$$

To estimate $(a_{\ell\ell}^{(k)} + a_{mm}^{(k)})^2$ we use the inequality

$$3\delta \leq 2\mu \qquad (3.2.20)$$

which follows from (3.1.10) and the definitions of $\delta$ and $\mu$ (see (1.2.6) and (1.1.3)). By the inequality (3.2.17) and the triangle inequality we obtain



$$|a_{\ell\ell}^{(k)} + a_{mm}^{(k)}| = |(\lambda_\ell^{(k)} + \lambda_m^{(k)}) + (a_{\ell\ell}^{(k)} - \lambda_\ell^{(k)}) + (a_{mm}^{(k)} - \lambda_m^{(k)})| \leq$$

$$\leq |\lambda_\ell^{(k)} + \lambda_m^{(k)}| + |(a_{\ell\ell}^{(k)} - \lambda_\ell^{(k)}) + (a_{mm}^{(k)} - \lambda_m^{(k)})| \leq$$

$$\leq 2\mu + \sqrt{2(a_{\ell\ell}^{(k)} - \lambda_\ell^{(k)})^2 + 2(a_{mm}^{(k)} - \lambda_m^{(k)})^2} \leq$$

$$\leq 2\mu + 0.44\delta \leq \mu(2 + \tfrac{2}{3}0.44) < \tfrac{7}{3}\mu, \quad k \in S'.$$

The relations (3.1.1), (3.2.14) and (3.2.4) imply

$$e_k^2 \geq 2.56^2\delta^2, \quad k \in S'$$

$$t_k^2 \geq 1 - b^2 \geq 1 - \left(\frac{0.88}{2N}\right)^2 \geq 1 - \left(\frac{0.44}{3}\right)^2 > 0.97848. \qquad (3.2.21)$$

Inserting the obtained inequalities into (3.2.19) we have

$$\sin^2 2\vartheta_k \leq \operatorname{tg}^2 2\vartheta_k \leq \frac{4 + \frac{49}{9}\mu^2}{6.4126\,\delta^2}(a_k^2 + b_k^2) \leq$$

$$\leq 0.85\,\frac{1+\mu^2}{\delta^2}(a_k^2 + b_k^2), \quad k \in S'. \qquad (3.2.22)$$

From the definitions of $\sin\varphi_k$ and $\sin\psi_k$ (see (3.1.3)) we obtain

$$2\max\{\sin^2\varphi_k, \sin^2\psi_k\} \leq 1 - t_k \cos 2\vartheta_k \cos\gamma_k + b_k|\sin 2\vartheta_k| =$$

$$= \frac{\sin^2 2\vartheta_k + \cos^2 2\vartheta_k(\sin^2\gamma_k + b_k^2\cos^2\gamma_k)}{1 + t_k\cos 2\vartheta_k\cos\gamma_k} +$$

$$+ b_k|\sin 2\vartheta_k|, \quad k \geq 1. \qquad (3.2.23)$$

For $k \in S'$ the relations (3.2.22), (3.2.16), (3.1.4) and (3.1.1) imply

$$\sin^2 2\vartheta_k \leq 0.85\,\frac{1+\mu^2}{\delta^2}(a_k^2 + b_k^2) \leq 0.85\cdot\frac{1+\mu^2}{\delta^2}\cdot\frac{e_k^2}{2} <$$

$$< 0.187,$$

$$\cos 2\vartheta_k > \sqrt{1-0.187} = \sqrt{0.813} > 0.9$$

$$\sin \gamma_k \leq 2\frac{|v_k|}{|e_k|} \leq \frac{2}{2.56}\frac{a_k}{\delta} \leq 0.78125\,\frac{a_k}{\delta}$$

$$\cos \gamma_k \geq \sqrt{1-0.78125^2\cdot\frac{1}{2}\cdot 0.44} > 0.93 \quad . \tag{3.2.24}$$

Using the inequalities (3.2.24) and (3.2.21) we obtain for $k \in S'$

$$\frac{1}{1+t_k\cos 2\vartheta_k \cos\gamma_k} \leq \frac{1}{1+0.989\cdot 0.9\cdot 0.93} \leq 0.54711$$

$$\sin^2\gamma_k + b_k^2\cos^2\gamma_k \leq 0.6104\cdot\frac{a_k^2}{\delta^2} + b_k^2$$

$$b_k|\sin 2\vartheta_k| \leq 0.922\ \frac{1+\mu^2}{\delta}\,(a_k^2+b_k^2) \tag{3.2.25}$$

Inserting the inequalities (3.2.25) into (3.2.23) we obtain

$$2\max\{\sin^2\varphi_k, \sin^2\psi_k\} \leq 0.54711\Big[0.85\,\frac{1+\mu^2}{\delta^2}(a_k^2+b_k^2) + 0.6104\,\frac{a_k^2}{\delta^2} + b_k^2\Big] + 0.922\,\frac{\sqrt{1+\mu^2}}{\delta}(a_k^2+b_k^2),$$

$$k \in S'. \tag{3.2.26}$$

Note that the inequality (3.2.20) implies

$$\frac{\sqrt{1+\mu^2}}{\delta} \geq \frac{3}{2}. \tag{3.2.27}$$

By (3.2.26) and (3.2.27) we have

$$2\max\{\sin^2\varphi_k,\sin^2\psi_k\} \leq \frac{1+\mu^2}{\delta^2}(a_k^2+b_k^2)\cdot$$

$$\cdot(0.54711\cdot 0.85 + 0.54711\cdot 0.6104 + 0.922\cdot\frac{2}{3})$$

$$\leq 1.413667\,\frac{1+\mu^2}{\delta^2}(a_k^2+b_k^2), \quad k \in S'. \tag{3.2.28}$$

If $p = n$, i.e. if all the eigenvalues $\lambda_j$ are simple, then the assumption (3.1.10) implies



$$3\delta \leq \mu,$$

hence the inequality

$$\frac{\sqrt{1+\mu^2}}{\delta} \geq 3 \tag{3.2.29}$$

holds. By the inequalities (3.2.29) and (3.2.26) we have

$$2\max\{\sin^2\varphi_k,\sin^2\psi_k\} \leq$$

$$\leq \frac{1+\mu^2}{\delta^2}(a_k^2+b_k^2)(0.54711\cdot 0.85 + 0.54711\cdot 0.6104 +$$

$$+ 0.922\,\tfrac{1}{3}) \leq 1.1064\cdot\frac{1+\mu^2}{\delta^2}(a_k^2+b_k^2), \quad k \geq 1. \tag{3.2.30}$$

From the inequalities (3.2.28) and (3.2.30) we obtain

$$\max\{\sin^2\varphi_k,\sin^2\psi_k\} \leq \begin{cases} 0.70684\,\dfrac{1+\mu^2}{\delta^2}(a_k^2+b_k^2), & p<n \\ 0.5532\ \dfrac{1+\mu^2}{\delta^2}(a_k^2+b_k^2), & p=n \end{cases} \tag{3.2.31}$$

In the case $b_k = 0$, $a_k \neq 0$ we have $\varphi_k = \psi_k = \vartheta_k$ where

$$\operatorname{tg} 2\vartheta_k = 2\frac{|a_k|}{|e_k|} \leq \frac{2}{2.56}\frac{a_k}{\delta}, \quad k \in S'$$

hence

$$\max\{\sin^2\varphi_k, \sin^2\psi_k\} \leq \left(\frac{2}{2.56}\frac{a_k}{\delta}\right)^2, \quad k \in S'.,$$

Thus, in this case the inequality (3.2.31) also holds.

Lemma 3.5 follows from the inequalities (3.2.31) and (3.2.5). ■

Suppose the assumptions (3.1.9) and (3.1.10) hold. Then Theorem 1.2 can be applied to $(A^{(1)}, B^{(1)})$. If in addition

$$a_{11}^{(1)} \geq a_{22}^{(1)} \geq \ldots \geq a_{nn}^{(1)} \tag{3.2.32}$$

then the assertions (1.2.26) and (1.2.27) of Theorem 1.2 hold for the partition (1.2.8) of matrices $A^{(1)}$ and $B^{(1)}$. The problem arises whether this result holds for the whole cycle.

3.6. Lemma. Let the assumptions (3.1.8), (3.1.9), (3.1.10) and (3.2.32) hold. If in addition

$$(1+\mu^2)\varepsilon_1^2 < \frac{\delta}{\mu+1}\delta^2, \tag{3.2.33}$$

where $\varepsilon_1$, $\mu$ and $\delta$ are defined by (3.1.1), (1.1.3) and (1.2.6) respectively, then for $k = 1,2,\dots,N+1$ the estimates

$$\|A_{ii}^{(k)} - \lambda_{s_i} B_{ii}^{(k)}\| \le \frac{1}{\delta_i} \sum_{\substack{j=1 \\ j\neq i}}^{p} \|A_{ij}^{(k)} - \lambda_{s_i} B_{ij}^{(k)}\|^2, \quad 1 \le i \le p \tag{3.2.34}$$

and

$$\sum_{i=1}^{p} \|A_{ii}^{(k)} - \lambda_{s_i} B_{ii}^{(k)}\|^2 \le [(1+\mu^2)\tau_k^2/\delta]^2 \tag{3.2.35}$$

hold. In the estimates (3.2.34) and (3.2.35) $\lambda_{s_i}$, $\delta_i$ and $\tau_k$ are defined by (1.2.3), (1.2.5) and (3.1.1), respectively and the matrices $A^{(k)}$, $B^{(k)}$ are partioned in accordance with (1.2.8).

Proof: Note that the inequality (3.2.16) holds for each pivot strategy. Therefore the condition (1.2.30) is fulfilled for all $1 \le k \le N+1$. By theorem 1.2 for each $k$ $(1 \le k \le N+1)$ there is a permutation matrix $P_k$ such that the inequalities (3.2.34) and (3.2.35) hold for the pair $(P_k^* A^{(k)} P_k, P_k^* B^{(k)} P_k)$. From (3.2.32) we see that $P_1 = I_n$. It remains to prove that $P_k = I_n$ for $k = 2,3,\dots,N+1$. We prove this by induction.

Suppose $P^{(k)} = I_n$ for a fixed $k$, $1 \le k \le N$. We shall prove that then $P^{(k+1)} = I_n$.



From (3.2.17) we obtain

$$|a_{rr}^{(k)} - \lambda_r| < \frac{\sqrt{2}}{2} \cdot 0.44\delta < 0.312\delta, \quad 1 \le r \le n \tag{3.2.36}$$

where $\lambda_1,\dots,\lambda_n$ is an ordering of the eigenvalues. Note that in the k'th step only $a_{\ell\ell}^{(k)}$ and $a_{mm}^{(k)}$ change. For the case $\lambda_\ell \neq \lambda_m$ we have only two possibilities: either

$$|a_{\ell\ell}^{(k+1)} - \lambda_\ell| < 0.312\delta \quad \text{and} \quad |a_{mm}^{(k+1)} - \lambda_m| < 0.312\delta$$

or

$$|a_{\ell\ell}^{(k+1)} - \lambda_m| < 0.312\delta \quad \text{and} \quad |a_{mm}^{(k+1)} - \lambda_\ell| < 0.312\delta. \tag{3.2.37}$$

Any other possibility leads to a contradiction with the assumption (1.2.3) on the multiplicities of the eigenvalues. Therefore it is sufficient to prove that the inequalities in (3.2.37) do not hold in the case $\lambda_\ell \neq \lambda_m$.

Since the inequalities (3.2.37) and (3.2.36) imply

$$|a_{\ell\ell}^{(k+1)} - a_{\ell\ell}^{(k)}| \ge |\lambda_\ell - \lambda_m| - |a_{\ell\ell}^{(k+1)} - \lambda_m| - |\lambda_\ell - a_{\ell\ell}^{(k)}| >$$
$$> 3\delta - 0.312\delta - 0.312\delta = 2.376\delta,$$

it is sufficient to prove that the conditions of Lemma 3.6 imply

$$|a_{\ell\ell}^{(k+1)} - a_{\ell\ell}^{(k)}| \le 2.376\delta. \tag{3.2.38}$$

Using the relation (3.1.2) we obtain

$$a_{\ell\ell}^{(k+1)} = a_{\ell\ell}^{(k)} + [(b_k^2 - \sin^2\varphi_k)\, a_{\ell\ell}^{(k)} + \sin^2\psi_k\, a_{mm}^{(k)} -$$
$$- 2\cos\varphi_k \sin\psi_k\, \mathrm{Re}(e^{-i\beta_k} a_{\ell m}^{(k)})]/(1 - b_k^2),$$

hence

$$|a_{\ell\ell}^{(k+1)} - a_{\ell\ell}^{(k)}| \leq \frac{1}{1-b^2}\left[\max\{|a_{\ell\ell}^{(k)}|, |a_{mm}^{(k)}|\}\cdot\right.$$
$$\cdot(|b_k^2 - \sin^2\varphi_k| + \sin^2\psi_k) +$$
$$\left. + (\sin^2\psi_k + a_k^2)\right]. \qquad (3.2.39)$$

Note that the relation (3.2.21) implies

$$1/(1-b^2) < 1.022 .$$

Since

$$|b_k^2 - \sin^2\varphi_k| + \sin^2\psi_k \leq \begin{cases} \sin^2\varphi_k + \sin^2\psi_k , & \sin^2\varphi_k \geq b_k^2 \\ b_k^2 + \sin^2\psi_k , & \sin^2\varphi_k < b_k^2 \end{cases}$$

we obtain from (3.2.27) and (3.2.31)

$$|b_k^2 - \sin^2\varphi_k| + \sin^2\psi_k \leq 2\cdot 0.707\,\frac{1+\mu^2}{\delta^2}\,(a_k^2 + b_k^2) =$$
$$= 1.414\,\frac{1+\mu^2}{\delta^2}\,(a_k^2 + b_k^2)$$

By the relations (3.2.36) and (3.2.20) we have

$$|a_{rr}^{(k)}| \leq \lambda_r + |a_{rr}^{(k)} - \lambda_r| \leq \mu + 0.312\delta \leq$$
$$\leq \mu(1 + \tfrac{2}{3}\cdot 0.312) < \tfrac{7}{6}\mu,$$

hence

$$\max\{|a_{\ell\ell}^{(k)}|, |a_{mm}^{(k)}|\} < \tfrac{7}{6}\mu.$$

Finally, by (3.2.27) and (3.2.31) we have

$$\sin^2\psi_k + a_k^2 \leq 0.707\,\frac{1+\mu^2}{\delta^2}\,(a_k^2 + b_k^2) +$$
$$+ \frac{4}{9}\,\frac{1+\mu^2}{\delta^2}\,(a_k^2 + b_k^2) \leq 1.152\,\frac{1+\mu^2}{\delta^2}\,(a_k^2 + b_k^2) .$$

Inserting the obtained inequalities into (3.2.39) we obtain



$$|a_{\ell\ell}^{(k+1)} - a_{\ell\ell}^{(k)}| \leq 1.022\,\frac{1+\mu^2}{\delta^2}\,(a_k^2 + b_k^2)\cdot$$
$$\cdot\left[\tfrac{7}{6}\mu\cdot 1.414 + 1.152\right] \leq$$
$$\leq 1.022(1.65\mu + 1.152)\,\frac{1+\mu^2}{\delta^2}\cdot\frac{\varepsilon_k^2}{2} \leq$$
$$\leq 0.85\,\frac{1+\mu^2}{\delta^2}(\mu+1)\cdot\varepsilon_k^2 .$$

The inequalities (3.2.6) and (3.2.33) imply

$$|a_{\ell\ell}^{(k+1)} - a_{\ell\ell}^{(k)}| \leq 0.85\cdot 1.754\,\frac{1+\mu^2}{\delta^2}\,(\mu+1)\cdot\varepsilon_1^2 <$$
$$< 1.5\,\delta ,$$

which proves the inequality (3.2.38) and Lemma 3.6. ■

All the results obtained so far have not depended upon any particular pivot strategy. The following result concerns the two special cyclic pivot strategies and does not depend on any (asymptotic) assumption (cf.[9]).

3.7. Lemma. Let the pairs $(A^{(k)}, B^{(k)})$ and $(\tilde{A}^{(k)}, \tilde{B}^{(k)})$, $k \geq 1$, be generated by the row- and the column-cyclic Jacobi method, respectively. If

$$(A^{(1)}, B^{(1)}) = (\tilde{A}^{(1)}, \tilde{B}^{(1)}) \qquad (3.2.40)$$

then

$$(A^{(rN+1)}, B^{(rN+1)}) = (\tilde{A}^{(rN+1)}, \tilde{B}^{(rN+1)}) ,$$
$$r \geq 1 . \qquad (3.2.41)$$

Proof: Obviously, it suffices to prove the equality (3.2.41) for $r = 1$. Set

$$A^{(k+1)} = F_k^* A^{(k)} F_k , \quad \tilde{A}^{(k+1)} = \tilde{F}_k^* \tilde{A}^{(k)} \tilde{F}_k$$
$$B^{(k+1)} = F_k^* B^{(k)} F_k , \quad \tilde{B}^{(k+1)} = \tilde{F}_k^* \tilde{B}^{(k)} \tilde{F}_k \quad , \; k \geq 1 .$$

From the assumption (3.2.40) we obtain

$$A^{(N+1)} = F^{(N)*}A^{(1)}F^{(N)}, \qquad \tilde{A}^{(N+1)} = \tilde{F}^{(N)*}A^{(1)}\tilde{F}^{(N)},$$

$$B^{(N+1)} = F^{(N)*}B^{(1)}F^{(N)}, \qquad \tilde{B}^{(N+1)} = \tilde{F}^{(N)*}B^{(1)}\tilde{F}^{(N)},$$

where

$$F^{(k)} = F_1F_2\cdots F_k, \qquad \tilde{F}^{(k)} = \tilde{F}_1\tilde{F}_2\cdots\tilde{F}_k, \qquad 1 \le k \le N.$$

Lemma 3.7 will be proved if we show that

$$F^{(N)} = \tilde{F}^{(N)}.$$

Let $F(\ell,m;\hat{F})$ stand for the elementary plane matrix defined by the pivot pair $(\ell,m)$ and its $(\ell,m)$-restriction $\hat{F}$. Since we consider only the first cycle of the two cyclic Jacobi processes, we can set

$$F_k = F_{\ell m} = F(\ell,m;\hat{F}_k),$$

$$\tilde{F}_k = \tilde{F}_{\tilde{\ell}\tilde{m}} = F(\tilde{\ell},\tilde{m};\hat{\tilde{F}}_k), \qquad 1 \le k \le N.$$

Here $(\ell,m)$ and $(\tilde{\ell},\tilde{m})$ are the pivot pairs of the $k$'th step of the row- and the column-cyclic process, respectively.

Since the matrices $F(i,j,;\cdot)$ and $F(q,r;\cdot)$ commute provided $\{i,j\} \cap \{q,r\} = \Phi$, we have

$$F^{(N)} = F_{12}F_{13}\cdots F_{1n}F_{23}\cdots F_{2n}\cdots F_{n-1,n} =$$

$$= F_{12}F_{13}F_{23}F_{14}F_{24}F_{34}\cdots F_{1n}F_{2n}\cdots F_{n-1,n}.$$

On the other hand

$$\tilde{F}^{(N)} = \tilde{F}_{12}\tilde{F}_{13}\tilde{F}_{23}\tilde{F}_{14}\tilde{F}_{24}\tilde{F}_{34}\cdots\tilde{F}_{1n}\tilde{F}_{2n}\cdots\tilde{F}_{n-1,n}$$

and therefore it is sufficient to prove

$$F_{ij} = \tilde{F}_{ij}, \qquad 1 \le i < j \le n, \tag{3.2.42}$$

i.e. that the $(i,j)$-restrictions of $F_{ij}$ and $\tilde{F}_{ij}$ are equal. We shall prove the relation (3.2.42) by induction with respect to $n$.



For $n = 2,3$ the row- and the column-cyclic pivot strategies coincide, hence the relation (3.2.42) holds. Assume that (3.2.42) holds for all starting pairs $(A,B)$ with matrices $A, B$ of order $n-1$, and let the pair $(A^{(1)},B^{(1)})$ be such that the matrices $A^{(1)},B^{(1)}$ are of order $n$. Let the matrices $F_k$, $F^{(k)}$, $\tilde{F}_k$ and $\tilde{F}^{(k)}$ be defined as above with respect to the pair $(A^{(1)},B^{(1)})$. We introduce a new cyclic Jacobi process defined by the following sequence of pivot pairs:

$$(1,2),(1,3),\dots,(1,n-1),(2,3),\dots,(2,n-1),\dots,$$

$$(n-2,n-1),(1,n),(2,n),\dots,(n-1,n).$$

According to this cyclic pivot strategy we define the generated matrix pairs $(A^{(k)'},B^{(k)'})$, and the transformation matrices $F'_k$, $F^{(k)'}$ so that

$$A^{(k+1)'} = F'^{*}_k A^{(k)'}F'_k, \qquad B^{(k+1)'} = F'^{*}_k B^{(k)'}F'_k,$$

$$A^{(k)'} = F^{(k)'*}A^{(1)}F^{(k)'}, \qquad B^{(k)'} = F^{(k)'*}B^{(1)}F^{'(k)},$$

$$F^{(k)'} = F'_1F'_2\cdots F'_k, \qquad F'_k = F'_{\ell',m'} = F(\ell',m',\hat{F}'_k),$$

$$1 \le k \le N,$$

holds.

We prove (3.2.42) in two steps. In the first step we prove

$$F'_{ij} = F_{ij}, \qquad 1 \le i < j \le n, \tag{3.2.43}$$

and in the second step we prove

$$F'_{ij} = \tilde{F}_{ij}, \qquad 1 \le i < j \le n. \tag{3.2.44}$$

First we prove the relation (3.2.43). The proof is divi-

ded in two parts. In the first part we prove by induction with respect to $i$ that

$$F'_{ij} = F_{ij}, \quad 1 \le i < j \le n-1 \qquad (3.2.45)$$

holds. In the second part we prove

$$F'_{in} = F_{in}, \quad 1 \le i \le n-1. \qquad (3.2.46)$$

For $i = 1$ the equalities

$$F'_{ij} = F_{ij}, \quad j = i+1, i+2, \dots, n-1 \qquad (3.2.47)$$

hold. Suppose that (3.2.47) holds for all $i$ satisfying $1 \le i \le i_o \le n-3$. We shall prove that then (3.2.47) holds for $i = i_o + 1$.

Let $A^{(k)} = (a^{(k)}_{rs})$, $A^{(k)'} = (a^{(k)'}_{rs})$ (similarly for $B^{(k)}$ and $B^{(k)'}$), and let

$$q_o = 0, \quad q_r = q_{r-1} + n - r, \quad r = 1, 2, \dots, n-1. \qquad (3.2.48)$$

By the induction hypothesis (with respect to $i$) and by the definition (3.2.48) of $q_i$'s we have

$$a^{(q_{i_o}+1)}_{rs} = a^{(q_{i_o}+1-i_o)'}_{rs}$$

$$b^{(q_{i_o}+1)}_{rs} = b^{(q_{i_o}+1-i_o)'}_{rs}$$

$$i_o + 1 \le r \le s \le n-1. \qquad (3.2.49)$$

Note that the relation (3.2.49) only expresses the fact that the transformation $F_{in}$ $(1 \le i \le i_o)$ has no impact on the values of $a^{(q_{i_o}+1)}_{rs}$, $b^{(q_{i_o}+1)}_{rs}$, $i_o + 1 \le r \le s \le n-1$. From the relation (3.2.49) and from the definition of the pivot strategies we conclude that (3.2.47) holds for $i = i_o + 1$. This proves (3.2.45).



To prove the equalities (3.2.46) we note that the $(1,n)$-restrictions of $A^{(N-n+2)'}$ and of $A^{(q_1)}$ are equal. The same holds for the matrices $B^{(N-n+2)'}$ and $B^{(q_1)}$. Therefore, (3.2.46) holds for $i = 1$. Since $F_{1n}$ commutes with all $F_{rs}$ $(= F'_{rs})$ such that $2 \le r < s \le n-1$ we can conclude that the $(2,n)$-restrictions of $A^{(N-n+3)'}$ and $B^{(N-n+3)'}$ coincide with the $(2,n)$-restrictions of $A^{(q_2)}$ and $B^{(q_2)}$, respectively. Thus $F_{2n} = F'_{2n}$. Continuing with these conclusions we prove other equalities in (3.2.46).

To prove (3.2.44) we use the induction hypothesis (on $n$) which implies

$$F'_{ij} = \tilde{F}_{ij}, \quad 1 \le i < j \le n-1.$$

Therefore $F^{(N-n+1)'} = \tilde{F}^{(N-n+1)}$ and $(A^{(N-n+2)'}, B^{(N-n+2)'}) = (\tilde{A}^{(N-n+2)}, \tilde{B}^{(N-n+2)})$. Since the last $n-1$ transformations of the full cycle are defined (for the both pivot strategies) by the same order of pivot pairs we have

$$F'_{in} = \tilde{F}_{in}, \quad 1 \le i \le n-1.$$

This proves (3.2.44) and completes the proof of Lemma 3.7. ∎

Note that the proof of Lemma 3.7 applies to any Jacobi-like process satisfying the two conditions :

1. The algorithm is defined at each step by the same (unique) rule ;

2. The algorithm uses only the elements of the $2 \times 2$ restrictions of the iterated matrices.

At the end of this section we prove that all obtained results hold for the real Jacobi method due to K.Zimmermann .

3.8. Lemma. All the results stated in Lemma 3.3, Lemma 3.4, Lemma 3.5, Lemma 3.6 and Lemma 3.7 hold for the real Jacobi method defined by the relations (2.2.18)–(2.2.21). In Lemma 3.5 the relation (3.2.18) can be replaced by

$$\sin^2\tilde{\omega}_k \leq \begin{cases} 0.57(1+\mu^2)\dfrac{\varepsilon_1^2}{\delta^2}, & p<n \\ 0.422(1+\mu^2)\dfrac{\varepsilon_1^2}{\delta^2}, & p=n \end{cases} \qquad (3.2.50)$$

where $\tilde{\omega}_k \in \{\tilde{\varphi}_k, \tilde{\psi}_k\}$.

Proof: The proofs of Lemma 3.3, Lemma 3.4 and Lemma 3.7 hold trivially for the real case. To prove Lemma 3.5 we use (2.2.21), (3.2.19) and (3.2.21) to obtain (3.2.22). For $k \in S'$ we have

$$\sin^2\tilde{\vartheta}_k \leq \frac{1}{4}tg^2\tilde{\vartheta}_k \leq \frac{0.85}{4}\cdot\frac{1+\mu^2}{\delta^2}(a_k^2+b_k^2).$$

From the definition (2.2.19) of $\sin\tilde{\varphi}_k$, $\sin\tilde{\psi}_k$ we obtain

$$\begin{aligned}\sin^2\tilde{\omega}_k &\leq \left[|\sin\tilde{\vartheta}_k|(1-\tilde{\xi}_k\tilde{\eta}_k)+\cos\tilde{\vartheta}_k\cdot|\xi_k|\right]^2 \\ &\leq (\sin^2\tilde{\vartheta}_k+\tilde{\xi}_k^2)\left[(1-\tilde{\xi}_k\tilde{\eta}_k)^2+\cos^2\tilde{\vartheta}_k\right] \leq \\ &\leq 2\sin^2\tilde{\vartheta}_k+2\tilde{\xi}_k^2, \quad k\geq 1.\end{aligned}$$

The assertion (3.2.4) of Lemma 3.3 and the relation (3.1.10) imply

$$b_k \leq b < \frac{0.88}{2.3} = \frac{0.88}{6}, \quad 1\leq k\leq N.$$

Using this result in the estimate obtained from the relation (2.2.20) we obtain

$$2\xi_k^2 \leq \frac{2b_k^2}{2+2\sqrt{1-b_k^2}} \leq \frac{b_k^2}{1+\sqrt{1-b^2}} \leq$$



$$\leq \frac{6}{6+\sqrt{35.12}}b_k^2 \leq 0.5031\,b_k^2, \quad 1\leq k\leq N-1.$$

For $k\in S'$ we obtain

$$\sin^2\tilde{\omega}_k \leq \frac{0.85}{2}\frac{1+\mu^2}{\delta^2}(a_k^2+b_k^2)+0.5031\,b_k^2,$$

hence the relations (3.2.27) and (3.2.29) imply

$$\sin^2\omega_k \leq \begin{cases} (0.425+\frac{4}{9}0.5031)\dfrac{1+\mu^2}{\delta^2}(a_k^2+b_k^2) \leq \\ \quad \leq 0.649\dfrac{1+\mu^2}{\delta^2}(a_k^2+b_k^2), & p<n \\ (0.425+\frac{1}{9}0.5031)\dfrac{1+\mu^2}{\delta^2}(a_k^2+b_k^2) \leq \\ \quad \leq 0.481\dfrac{1+\mu^2}{\delta^2}(a_k^2+b_k^2), & p=n. \end{cases} \qquad (3.2.51)$$

The inequality (3.2.51) and the assertion (3.2.5) of Lemma 3.3 imply (3.2.50).

To prove Lemma 3.6 it is sufficient to see that (3.2.38) holds. From the relation (2.2.18) we obtain

$$\begin{aligned}a_{ll}^{(k+1)}-a_{ll}^{(k)} &= y_k\left[(b_k^2-\sin^2\tilde{\varphi}_k)\,a_{ll}^{(k)}+\right. \\ &\left.+\sin^2\psi_k\,a_{mm}^{(k)}-2\cos\tilde{\varphi}_k\sin\tilde{\psi}_k\,a_{lm}^{(k)}\right]\end{aligned}$$

where $y_k$ is defined in (3.1.1). We obtain

$$\begin{aligned}|a_{ll}^{(k+1)}-a_{ll}^{(k)}| &\leq \left[\max\{|a_{ll}^{(k)}|, |a_{mm}^{(k)}|\}\cdot\right. \\ &\left.\cdot(|b_k^2-\sin^2\tilde{\varphi}_k|+\sin^2\tilde{\psi}_k)+\sin^2\tilde{\psi}_k+a_k^2\right]/(1-b^2).\end{aligned}$$

Using the same estimates as in the proof of Lemma 3.6 (replacing $\varphi_k, \psi_k$ by $\tilde{\varphi}_k, \tilde{\psi}_k$, respectively, we prove (3.2.38) for the real method. This proves Lemma 3.7. ∎

Note that Lemma 3.7 holds for the method due to S.Falk and P.Langemeyer.

## 3.3. Simple Eigenvalues

Here we prove the quadratic convergence of the special Jacobi method in the case of simple eigenvalues. Thus, $p = n$, $\tau_k = \varepsilon_k$ $(k \geq 1)$ and (cf.(1.2.6))

$$3\delta = \min_{i \neq j} |\lambda_i - \lambda_j| .$$

3.9. Theorem. Let the assumptions (3.1.8), (3.1.9) and (3.1.10) hold for the pair $(A,B)$, and let the sequence of pairs $((A^{(k)},B^{(k)}), k \geq 1)$ be generated by a cyclic Jacobi method defined by the relations (3.1.2)–(3.1.5). If the eigenvalues of the pair $(A,B)$ are simple then

$$\varepsilon_{N+1} \leq \sqrt{N(1+\mu^2)}\, \frac{\varepsilon_1^2}{\delta} . \tag{3.3.1}$$

Moreover, if the pivot strategy is row- or column-cyclic then

$$\varepsilon_{N+1} \leq \sqrt{1+\mu^2}\, \frac{\varepsilon_1^2}{\delta} . \tag{3.3.2}$$

Proof: The proof uses the technique developed by J.H.Wilkinson in [20].

First we prove that the estimate (3.3.1) holds for an arbitrary pivot strategy. Let $k$ $(1 \leq k \leq N)$ be fixed. Then the pivot pair $(\ell,m)$ is also fixed (cf.(2.1.4)). Consider the elements $a_{\ell m}^{(r)}$, $r = k+1, k+2, \dots, N$. Note that $a_{\ell m}^{(k+1)} = 0$ and that $a_{\ell m}^{(r)}$ changes at most $2(n-2)$ times. Let us denote by $r_1, r_2, \dots, r_s$ $(s < 2n-4)$ those values of $r$ for which $a_{\ell m}^{(r)}$ changes in the $r$'th step. For simplicity we set $z_i = a_{\ell m}^{(r_i+1)}$, $0 \leq i \leq s$, where $r_0 = k$. Then the relations (3.1.2) and (3.1.1) imply



$$\begin{aligned}
z_1 &= \sqrt{y_{r_1}}\,(\,0 \cos\omega'_{r_1} \pm a^{(r_1)} e^{i\nu_1} \sin\omega_{r_1}\,) \\
z_2 &= \sqrt{y_{r_2}}\,(z_1 \cos\omega'_{r_2} \pm a^{(r_2)} e^{i\nu_2} \sin\omega_{r_2}\,) \\
&\vdots \\
z_j &= \sqrt{y_{r_j}}\,(z_{j-1} \cos\omega'_{r_j} \pm a^{(r_j)} e^{i\nu_j} \sin\omega_{r_j}\,) \\
&\qquad 1 \leq j \leq s ,
\end{aligned} \tag{3.3.3}$$

where $\omega'_{r_j}$, $\omega_{r_j} \in \{\varphi_{r_j}, \psi_{r_j}\}$, $\nu_j \in \{\alpha_{r_j}, -\alpha_{r_j}, \beta_{r_j}, -\beta_{r_j}\}$, while $a^{(r_j)}$ is a certain off-diagonal element of $A^{(r_j)}$. From (3.3.3) it follows

$$\begin{aligned}
|z_j| &\leq \sum_{i=1}^{j} \sqrt{y_{r_i} y_{r_{i+1}} \cdots y_{r_j}}\, |a^{(r_i)}| \cdot |\sin\omega_{r_i}| \leq \\
&\leq (1-b^2)^{-j/2} \sum_{i=1}^{j} |a^{(r_i)}| \cdot |\sin\omega_{r_i}| , \\
&\qquad 1 \leq j \leq s .
\end{aligned} \tag{3.3.4}$$

For $k \geq 1$ set

$$A^{(k)} = D_A^{(k)} + E^{(k)} , \quad D_A^{(k)} = \operatorname{diag}(a_{11}^{(k)}, \dots, a_{nn}^{(k)}) .$$

The matrix $E^{(N+1)}$ consists exactly of the elements $z_s$. Note that $s$ is a function of the pivot pair $(\ell,m)$ (hence also of $k$) and the cyclic pivot strategy under consideration. From (3.3.4) we conclude that

$$\begin{aligned}
|E^{(N+1)}| &\leq (1-b^2)^{-(n-2)} (|P^{(2)}| \cdot |\sin\omega_2| + \\
&\quad + |P^{(3)}| \cdot |\sin\omega_3| + |P^{(N)}| \cdot |\sin\omega_N|) ,
\end{aligned} \tag{3.3.5}$$

where each matrix $P^{(k)}$ contains non-zero elements only at those positions of the $\ell$'th and $m$'th row and column which have already been pivot positions. The non-zero elements of $P^{(k)}$ are certain elements of $E^{(k)}$ belonging to the $\ell$'th and

m'th row and column. In (3.3.5) we use the notation $|C| = (|c_{ij}|)$ where $C = (c_{ij})$ is an arbitrary matrix.

By the assertion (3.2.6) of Lemma 3.3 we obtain

$$\| |P^{(k)}| \| = \|P^{(k)}\| \leq S(A^{(k)}) \leq$$

$$\leq \sqrt{1.754}\, S(A^{(1)}), \quad 2 \leq k \leq N. \tag{3.3.6}$$

By the relation (3.1.10) and from the assertion (ii) of Lemma 3.1 we obtain

$$(1-b^2)^{-(n-2)} \leq \left[(1-b^2)^{-N}\right]^{1/n} \leq \left(1+\frac{72}{67}\cdot\frac{1}{2}\cdot\frac{1}{6}\right)^{1/3} \leq$$

$$\leq (12/11)^{1/3} < 1.03. \tag{3.3.7}$$

Using the relations (3.3.5), (3.3.6), (3.3.7) and Lemma 3.5 we obtain

$$S(A^{(N+1)}) = \|E^{(N+1)}\| = \| |E^{(N+1)}| \| \leq$$

$$1.03\sqrt{1.754}\, S(A^{(1)}) \sum_{k=2}^{N} |\sin \omega_k| \leq$$

$$1.36412\, S(A^{(1)}) \cdot \left[(N-1)\sum_{k=2}^{N-1} \sin^2 \omega_k\right]^{1/2}$$

$$0.9502 \sqrt{N(1+\mu^2)}\, \frac{\varepsilon_1}{\delta}\, S(A^{(1)}). \tag{3.3.8}$$

The same analysis applies with $S(B^{(1)})$ and $S(B^{(N+1)})$. Therefore (3.3.1) follows from (3.3.8) and the definition of $\varepsilon_k$ (see (3.1.1), (3.2.3) and (1.2.3)).

Let us prove the second part of Theorem 3.9. By Lemma 3.7 it suffices to prove (3.3.2) only for the row-cyclic pivot strategy. Applying the analysis which was used in proving the equalities (3.3.3) to the row-cyclic pivot strategy, we obtain for the elements of the first row



$$|a_{1j}^{(n)}| \leq \sum_{i=j}^{n-1} \sqrt{y_i y_{i+1} \cdots y_{n-1}}\, |\sin \psi_i| \cdot |a_{i+1,j}^{(i)}|,$$

$$2 \leq j \leq n-1.$$

By the Cauchy–Schwartz inequality we obtain for $2 \leq j \leq n-1$

$$|a_{1j}^{(n)}|^2 \leq \sum_{i=j}^{n-1} |a_{i+1,j}^{(i)}|^2 \cdot \sum_{r=j}^{n-1} y_r y_{r+1} \cdots y_{n-1} \sin^2 \psi_r. \tag{3.3.9}$$

Since $a_{1n}^{(n)} = 0$, the inequality (3.3.9) implies

$$\sum_{j=2}^{n} |a_{1j}^{(n)}|^2 \leq \left[\sum_{j=2}^{n-1}\sum_{i=j}^{n-1} |a_{i+1,j}^{(i)}|^2\right] \cdot \sum_{k=2}^{n-1} y_k \cdots y_{n-1} \sin^2 \psi_k. \tag{3.3.10}$$

First we estimate the sum in brackets. Since

$$a_{ij}^{(j-1)} = a_{ij}^{(i)}, \quad 2 \leq i < j \leq n$$

and

$$|a_{ij}^{(k)}| = |a_{ji}^{(k)}|, \quad 1 \leq i < j \leq n, \quad k \geq 1$$

we have

$$\sum_{j=2}^{n-1}\sum_{i=j}^{n-1} |a_{i+1,j}^{(i)}|^2 = \sum_{j=2}^{n-1}\sum_{i=j+1}^{n} |a_{ij}^{(i-1)}|^2 =$$

$$= \sum_{j=3}^{n}\sum_{i=1}^{j-1} |a_{ij}^{(j-1)}|^2 = \sum_{j=3}^{n}\sum_{i=1}^{j-1} |a_{ij}^{(i)}|^2. \tag{3.3.11}$$

In order to estimate the last sum on the right-hand side of (3.3.11) we use the invariance of the Euclidean norm under unitary transformations. For $2 \leq i \leq j-1$, $3 \leq j \leq n$ we have

$$\left\| \begin{bmatrix} a_{1j}^{(i)} \\ a_{ij}^{(i)} \end{bmatrix} \right\|^2 = \left\| \hat{F}_{i-1}^{*} \begin{bmatrix} a_{1j}^{(i-1)} \\ a_{ij}^{(i-1)} \end{bmatrix} \right\|^2$$

$$\leq \| \Phi^{(i-1)} * R_2^{(i-1)} * D^{(i-1)} * R_1^{(i-1)} * \|_2^2 \left\| \begin{bmatrix} a_{1j}^{(i-1)} \\ a_{ij}^{(i-1)} \end{bmatrix} \right\|^2 =$$

$$= \|D^{(i-1)}\|_2^2 \left\| \begin{matrix} a_{lj}^{(i-1)} \\ a_{ij}^{(i-1)} \end{matrix} \right\|^2 ,$$

where the relation (3.1.1) has been used. By (3.2.7) we obtain

$$|a_{lj}^{(i)}|^2 + |a_{ij}^{(i)}|^2 \le x_{i-1}(|a_{lj}^{(i-1)}|^2 + |a_{ij}^{(i-1)}|^2),$$

$$2 \le i \le j-1 , \quad 3 \le j \le n . \qquad (3.3.12)$$

Iterating the inequalities (3.3.12) for $i = 2,3,\dots,j-1$ we obtain

$$|a_{lj}^{(j-1)}|^2 + \sum_{i=2}^{j-1} |a_{ij}^{(i)}|^2 \le x_1 x_2 \cdots x_{j-2} |a_{lj}^{(1)}|^2 +$$

$$+ \sum_{i=2}^{j-1} x_{i-1} x_i \cdots x_{j-2} |a_{ij}^{(i-1)}|^2 , \quad 3 \le j \le n ,$$

and since

$$a_{ij}^{(i-1)} = a_{ij}^{(1)} , \quad 2 \le i < j \le n ,$$

we have

$$\sum_{j=3}^{n} \sum_{i=2}^{j-1} |a_{ij}^{(i)}|^2 \le x_1 x_2 \cdots x_{n-2} \sum_{j=3}^{n} \sum_{i=1}^{j-1} |a_{ij}^{(1)}|^2 \le$$

$$x_1 x_2 \cdots x_{n-2} \cdot \frac{1}{2} S^2(A^{(1)}) . \qquad (3.3.13)$$

By (3.3.10), (3.3.11) and (3.3.13) we have

$$\sum_{j=2}^{n} |a_{lj}^{(n)}|^2 \le x_1 x_2 \cdots x_{n-2} y_2 y_3 \cdots y_{n-1} \cdot \frac{1}{2} S^2(A^{(1)}) \sum_{k=1}^{n-1} \sin^2 \psi_k \qquad (3.3.14)$$

During the next $N-n+1$ steps the elements of the first row will interact between themselves. Since

$$|a_{l\ell}^{(k+1)}|^2 + |a_{lm}^{(k+1)}|^2 \le x_k(|a_{l\ell}^{(k)}|^2 + |a_{lm}^{(k)}|^2),$$

$$2 \le \ell < m \le n ,$$

which can be obtained similarly as the inequality (3.3.12), we obtain from (3.3.14)



$$\sum_{j=2}^{n} |a_{lj}^{(N+1)}|^2 \le x_n x_{n+1} \cdots x_N \sum_{j=2}^{n} |a_{lj}^{(n)}|^2 \le$$

$$\le x_1 x_2 \cdots x_N y_2 \cdots y_{n-1} \cdot \frac{1}{2} S^2(A^{(1)}) \sum \sin^2 \psi_k . \qquad (3.3.15)$$

Consider now the matrix pair $(A_{n-1}^{(n)}, B_{n-1}^{(n)})$, where $A_{n-1}^{(n)}$, $B_{n-1}^{(n)}$ are the submatrices of $A^{(n)}$, $B^{(n)}$ respectively, obtained on the intersection of the last $n-1$ rows and columns. Applying the result (3.3.15) to the pair $(A_{n-1}^{(n)}, B_{n-1}^{(n)})$ we obtain

$$\sum_{j=3}^{n} |a_{2j}^{(N+1)}|^2 \le x_n x_{n+1} \cdots x_N y_{n-1} \cdots y_{2n-3} \cdot$$

$$\cdot \frac{1}{2} S^2(A^{(n)}) \sum_{k=n+1}^{2n-3} \sin^2 \psi_k .$$

More generally, let $q_i$ $(0 \le i \le n-1)$ be defined by (3.2.48). Apply the inequality (3.3.15) to the principal submatrices of $A^{(q_i+1)}$ and $B^{(q_i+1)}$, obtained on the intersection of the last $n-i$ rows and columns. Then we have

$$\sum_{j=i+2}^{n} |a_{i+1,j}^{(N+1)}|^2 \le x_{q_i+1} \cdots x_N y_{q_i+2} \cdots y_{q_{i+1}} \cdot$$

$$\cdot \frac{1}{2} S^2(A^{(q_i+1)}) \sum_{k=q_i+2}^{q_{i+1}} \sin^2 \psi_k , \quad 0 \le i \le n-3 .$$

Since the assertion (3.2.6) of Lemma 3.3 implies

$$S^2(A^{(q_i+1)}) \le x_1 \cdots x_{q_i} S^2(A^{(1)}) ,$$

we have for $i = 0,1,\dots,n-3$

$$\sum_{j=i+2}^{n} |a_{i+1,j}^{(N+1)}|^2 \le x_1 x_2 \cdots x_N y_{q_i+2} \cdots y_{q_{i+1}}$$

$$\cdot \frac{1}{2} S^2(A^{(1)}) \sum_{k=q_i+2}^{q_{i+1}} \sin^2 \psi_k \le$$

$$\le (1-b)^{-N}(1-b^2)^{-(n-2)} \cdot \frac{1}{2} S^2(A^{(1)}) \sum_{k=q_i+2}^{q_{i+1}} \sin^2 \psi_k \le$$

$$\leq \frac{1}{2}\cdot 1.80662\, S^2(A^{(1)}) \sum_{k=q_{i+1}+2}^{q_{i+2}} \sin^2\varphi_k . \qquad (3.3.16)$$

Here we have used the relations (3.2.13) and (3.3.7).

Since $a_{n-1,n}^{(N+1)} = 0$ we obtain from (3.3.16)

$$S^2(A^{(N+1)}) = 2\sum_{i=0}^{m-3}\sum_{j=i+2}^{n} |a_{i+1,j}^{(N+1)}|^2 \leq$$

$$\leq 1.80662\, S^2(A^{(1)}) \sum_{i=0}^{n-3}\sum_{k=q_{i+1}+2}^{q_{i+2}} \sin^2\varphi_k =$$

$$= 1.80662\, S^2(A^{(1)}) \sum_{\substack{k=1 \\ m\neq \ell+2}}^{N} \sin^2\varphi_k . \qquad (3.3.17)$$

Using Lemma 3.5 in (3.3.17) we obtain

$$S^2(A^{(N+1)}) \leq 1.80662\ 0.4852\,(1+\mu^2)\,\frac{\varepsilon_1^2}{\delta^2}\, S^2(A^{(1)}) \leq$$

$$\leq 0.8766\,(1+\mu^2)\,\frac{\varepsilon_1^2}{\delta^2}\, S^2(A^{(1)}) . \qquad (3.3.18)$$

Since the whole analysis applies to the matrix $B^{(1)}$ we conclude that (3.3.18) holds with $S^2(B^{(1)})$ and $S^2(B^{(N+1)})$ instead of $S^2(A^{(1)})$ and $S^2(A^{(N+1)})$, respectively. Therefore we have

$$\varepsilon_{N+1}^2 \leq 0.8766\,\frac{\varepsilon_1^2}{\delta^2}\,(S^2(A^{(1)}) + S^2(B^{(1)})) \leq$$

$$\leq \left(0.94\sqrt{1+\mu^2}\,\frac{\varepsilon_1^2}{\delta}\right)^2 .$$

The obtained inequality implies the estimate (3.3.2) hence Theorem 3.9 is proved. ■

The estimates (3.3.1) and (3.3.2) are quite analogous to the known estimates for the standard Jacobi method obtained by J.H.Wilkinson in [20].

The factor $\sqrt{1+\mu^2}$ which does not appear in the estimate for the standard Jacobi method originates from the presence of the sum $a_{\ell\ell}^{(k)} + a_{mm}^{(k)}$ in the numerator of the expression for $\operatorname{tg} 2\vartheta_k$ (see the proof of Lemma 3.5). The asymptotic assumption (3.1.9) is approximately $\sqrt{1+\mu^2}$ times stronger than the assumption in [20].



If $\delta$ is very tiny due to a pair of very close eigenvalues, then the estimates (3.3.1) and (3.3.2) imply that $\varepsilon_{N+1}$ is not "essentially smaller" then $\varepsilon_1$. The following result implies that in such situations certain off-diagonal elements of $A^{(N+1)}$ and $B^{(N+1)}$ are still essentially smaller than $\varepsilon_1$.

3.10. Corollary. Let the assumptions (3.1.8), (3.1.9), (3.1.10), (3.2.33) and (3.2.32) hold and let the eigenvalues of the pair $(A,B)$ be simple. Then for the row- and the column-cyclic Jacobi method defined by (3.1.2)–(3.1.5)

$$\sum_{\substack{j=1 \\ j\neq i,n}}^{n} \left(|a_{ij}^{(N+1)}|^2 + |b_{ij}^{(N+1)}|^2\right) \leq 0.4385\,\frac{1+\mu^2}{\delta_i^2}\,\varepsilon_1^4 ,$$

$$1 \leq i \leq n-1 , \qquad (3.3.19)$$

holds where $\delta_i$ is defined by (1.2.5).

Proof: By Lemma 3.6 we see that during the whole cycle the affiliation of the diagonal elements to the eigenvalues does not change. From the proof of Lemma 3.5 we see that (3.2.31) can be replaced by

$$\sin^2\omega_k \leq \begin{cases} 0.70684\,\dfrac{(1+\mu^2)(a_k^2+b_k^2)}{\max\{\delta_\ell^2,\delta_m^2\}} , & p<n \\ 0.5532\,\dfrac{(1+\mu^2)(a_k^2+b_k^2)}{\max\{\delta_\ell^2,\delta_m^2\}} , & p=n \end{cases} ,$$

$$k \in S' , \qquad (3.3.20)$$

where $\omega_k \in \{\varphi_k, \psi_k\}$ and $\delta_i$ is as in (1.2.5). By (3.3.16), (3.3.20) and (3.2.48) we have

$$\sum_{j=i+2}^{n} |a_{i+1,j}^{(N+1)}|^2 \leq 0.90331\, S^2(A^{(1)})\, 0.5532\, \frac{1+\mu^2}{\delta_{i+1}^2} \sum_{k=q+2}^{q+n} (a_k^2 + b_k^2) \leq$$

$$\leq 0.5\, \frac{1+\mu^2}{\delta_{i+1}^2}\, S^2(A^{(1)}) \sum_{k=4}^{N} (a_k^2 + b_k^2)\,, \quad 0 \leq i \leq n-2\,.$$

Using the assertion (3.2.5) of Lemma 3.3 we obtain

$$\sum_{j=i+1}^{n} |a_{ij}^{(N+1)}|^2 \leq 0.4385\, \frac{1+\mu^2}{\delta_i^2}\, \varepsilon_1^2\, S^2(A^{(1)})\,, \quad 1 \leq i \leq n-2.$$

Since the same inequality holds with $b_{ij}^{(N+1)}$ and $S^2(B^{(1)})$ the relation (3.3.19) follows from the definition of $\varepsilon_1$ (see (3.1.1) or (3.2.3)). ■

Note that the assumption (3.2.32) of Lemma 3.10 is not essential (cf. Theorem 1.2).

3.11. Corollary. Let $A, B$ be symmetric matrices such that $B$ is positive definite and let the sequence $((A^{(k)}, B^{(k)}), k \geq 1)$ be generated by a cyclic Jacobi method defined by the relation (2.2.18)–(2.2.21). If the assumptions (3.1.8), (3.1.9) and (3.1.10) hold then the both assertions of Theorem 3.9 are valid. If in addition the assumptions (3.2.33) and (3.2.32) hold then the inequality (3.3.19) is also valid.

Proof: The proofs of Theorem 3.9 and Corollary 3.10 in the case of real matrices are identical to those in the complex case. Having in mind Lemma 3.8 it is possible to decrease the constants $\sqrt{N(1+\mu^2)}/\delta$ and $\sqrt{1+\mu^2}/\delta$ in (3.3.1) and (3.3.2) by the factor $0.422/0.4852$. ■

3.12. Remark. The quadratic convergence of the row- and the column-cyclic real Jacobi method due to K. Zimmermann can be also proved using the estimate of Zimmermann (see [25], Satz



8, p.32), which can be in our notation stated as follows:

$$\varepsilon_{N+1} \leq \left\{ \left[ 1 - \prod_{j=3}^{n} \prod_{i=1}^{j-2} \cos^2 \tilde{\vartheta}_{ij} \right]^{1/2} + (1+t)^{-N} - 1 \right\} \varepsilon_1$$

where $\tilde{\vartheta}_{ij}$ is $\tilde{\vartheta}_k$ from the relation (2.2.21) provided that $i = \ell$ and $j = m$. Here

$$t = \max_{1 \leq k \leq N} \left\{ \frac{1}{\sqrt{1-b_k}} - 1 \right\} = \frac{1}{\sqrt{1-b}} - 1\,,$$

hence the assertion (iv) of Lemma 3.1 implies $t \leq \frac{3}{5} b$. Using the assertion (iii) of Lemma 3.1 and the assertion (3.2.6) of Lemma 3.3 we obtain

$$(1+t)^N - 1 \leq (1 + \tfrac{3}{5}b)^N - 1 \leq 1 + \tfrac{4}{3} N \tfrac{3}{5} b - 1 \leq \tfrac{4}{5} N b \leq$$

$$\leq \tfrac{4}{5} N \tfrac{\sqrt{2}}{2} \sqrt{1.754}\, S(B^{(1)}) \leq \tfrac{4}{5} \sqrt{0.877}\, N\, \varepsilon_1 \leq 0.75\, N\, \varepsilon_1\,.$$

On the other hand by a variant of the Bernoulli's inequality (cf. [12]) and by the relations (3.2.22), (3.2.5) we have

$$1 - \prod_{j=3}^{n} \prod_{i=1}^{j-2} \cos^2 \tilde{\vartheta}_{ij} \leq \sum_{j=3}^{n} \sum_{i=1}^{j-2} \sin^2 \tilde{\vartheta}_{ij} \leq$$

$$\leq \frac{0.85}{4} \cdot \frac{1+\mu^2}{\delta^2} \sum_{k=4}^{N} (a_k^2 + b_k^2) \leq 0.1864\, \frac{1+\mu^2}{\delta^2}\, \varepsilon_1^2 \leq$$

$$\leq (0.432 \sqrt{1+\mu^2}\, \varepsilon_1 / \delta)^2\,.$$

We have proved

$$\varepsilon_{N+1} \leq \left( 0.432\, \frac{\sqrt{1+\mu^2}}{\delta} + 0.75\, N \right) \varepsilon_1^2\,.$$

Since all the eigenvalues are separeted we have $2\mu \geq (n-1)\, 3\delta$ hence $\sqrt{1+\mu^2}/\delta \geq 3(n-1)/2$. Therefore we have

$$\varepsilon_{N+1} \leq \frac{\sqrt{1+\mu^2}}{\delta} (0.432 + 0.25n)\, \varepsilon_1^2\,.$$

Since $0.432 + 0.25n \geq 1$, $n \geq 3$, the obtained result is not better than that from Corollary 3.11. ■

Using Theorem 3.6 and Corollary 3.11 it is easy to formulate the quadratic convergence result for the real method due to S.Falk and P.Langemeyer.

3.13.Corollary. Let $A$, $B$ be symmetric matrices such that $B$ is positive definite and let the sequence $((A^{(k)},B^{(k)}),k\geq 1)$ be generated by a cyclic Jacobi method defined by the relations (2.2.24), (2.2.25), (2.2.27), (2.2.31) and (2.2.32). Let the assumptions

$$S(I_n,B) \leq \frac{1}{2N},$$

$$2\sqrt{1+\mu^2}\, S(A,B) < \delta,$$

and (3.1.10) hold, where $S(\cdot,\cdot)$, $\mu$, $\delta$ are defined by the relations (1.2.23), (1.1.3), (1.2.6), respectively. If the eigenvalues of the pair $(A,B)$ are simple then

$$S(A^{(N+1)},B^{(N+1)}) \leq \frac{\sqrt{N(1+\mu^2)}}{\delta} S^2(A,B).$$

Moreover, if the pivot strategy is row- or column-cyclic then

$$S(A^{(N+1)},B^{(N+1)}) \leq \frac{\sqrt{1+\mu^2}}{\delta} S^2(A,B).$$

Proof: The proof is a direct consequence of Theorem 2.6 and Corollary 3.11. Note that

$$S(I_n,B) = S(D_oBD_o)$$

where $D_o = \mathrm{diag}(1/\sqrt{b_{11}},\dots,1/\sqrt{b_{nn}})$. Also $S(A^{(1)},B^{(1)}) = S(A,B)$ since $A^{(1)} = A$, $B^{(1)} = B$. ■



## 3.4. Multiple Eigenvalues

If the eigenvalues of the pair $(A,B)$ are not simple then generally it is not possible to prove the quadratic convergence of the special Jacobi method. Even for the row-cyclic pivot strategy the numerical experiments show an essential slowdown in the reduction of $\varepsilon_k$ per cycle. In Section 3.5 we show that this slowdown of the assymptotic rate of convergence is due to the special structure of almost diagonal pairs with multiple eigenvalues (see Sec.1.2). These considerations are used in Section 3.6 where the special Jacobi method is modified to be quadratically convergent.

Here we give a brief description of the obtained estimates in the case of multiple eigenvalues. They concern the row- and the column-cyclic Jacobi method from Section 2.2.

Set

$$M = N - \sum_{i=1}^{p} n_i(n_i-1)/2, \quad n_{max} = \max_{1\leq i\leq p} n_i \qquad (3.4.1)$$

where $N$, $n_i$ and $p$ are defined by (2.1.3), (2.2.4) and (1.2.3), respectively.

3.14.Theorem. Let the assumptions (3.1.8), (3.1.9), (3.1.10), (3.2.33) and (3.2.32) hold for the pair $(A,B)$, and let the sequence $((A^{(k)},B^{(k)}),k\geq 1)$ be generated by the row- or the column-cyclic Jacobi method defined by (3.1.2)–(3.1.5). Let $n_{max}$, $M$, $\varepsilon_k$, $\tau_k$, $\delta_i$, $\delta$ and $\mu$ be defined by the relations (3.4.1), (3.1.1), (1.2.5), (1.2.6) and (1.1.3). Then

(i) $$\tau_{N+1} \leq \frac{3}{2}\sqrt{(2.31)^M n_{max}(1+\mu^2)}\,\frac{\varepsilon_1}{\delta}\,\tau_1$$

(ii) $\tau_{N+1} \leq \frac{3}{2}\sqrt{n_{max}(1+\mu^2)}\ \frac{\varepsilon_1^2}{\delta}$ ;

(iii) If $A^{(N+1)}$ and $B^{(N+1)}$ are partitioned according to (1.2.8) then

$$\sum_{j=i+1}^{p} (\|A_{ij}^{(N+1)}\|^2 + \|B_{ij}^{(N+1)}\|^2) \leq 1.12\, n_i \frac{1+\mu^2}{\delta_i^2}\ \varepsilon_1^4 ,$$

$1 \leq i \leq p-1$ ;

(iv) If $n_{max} = 2$ then

$$\varepsilon_{N+1} \leq \frac{18}{17}\sqrt{1+\mu^2}\ \frac{\varepsilon_1^2}{\delta} .$$

Proof: The proof of the assertions (i), (ii) and (iii) is rather long and sophisticated, hence we prove here only (iv).

By Lemma 3.6 we see that $n_{max} = 2$ and the relation (3.2.32) imply $S' \subset \{k \in \{1,2,\dots,N\};\ m \geq \ell+2\}$. It is obvious that one can replace $A^{(N+1)}$, $A^{(1)}$ in the estimate (3.3.17) by $B^{(N+1)}$, $B^{(1)}$, respectively. Now, one can use Lemma 3.5 to estimate the right-hand side of the obtained inequality and find

$$\varepsilon_{N+1}^2 \leq 1.80662\ \varepsilon_1^2 \sum_{\substack{k=1 \\ m \geq \ell+2}}^{N} \sin^2\psi_k \leq 1.80662\ \varepsilon_1^2 \sum_{k=1}^{N}{}' \sin^2\psi_k \leq$$

$$\leq 1.80662\ \varepsilon_1^2\, 0.62(1+\mu^2)\frac{\varepsilon_1^2}{\delta^2} \leq 1.12011(1+\mu^2)\frac{\varepsilon_1^4}{\delta^2} ,$$

which proves (iv). ■

Using Lemma 3.8 it can be proved that Theorem 3.14 holds also for the real Jacobi method due to K.Zimmermann. Taking advantage of the definitions of $S(A,B)$, $\tau(A,B)$ from the relations (1.2.23), (1.2.24) respectively, Theorem 3.14 can be formulated to hold for the Jacobi method due to S.Falk and P.Langemeyer.



## 3.5. Qualitative Analysis and Finite Arithmetic

Here we analyse why the quadratic convergence of the special Jacobi method fails in the case of multiple eigenvalues. We also consider the influence of the multiple eigenvalues on the accuracy of the computed eigenvalues.

We use the Landau's $\mathcal{O}$-symbol. Remember that $x = \mathcal{O}(\eta)$, $\eta > 0$, if there exist constants $c > 0$ and $\eta_0$ such that $|x|/\eta \leq c$, $\eta \in [0,\eta_0]$.

The row- or the column-cyclic Jacobi method is applied on a pair $(A^{(1)},B^{(1)})$ satisfying the following assumptions:

(i) $S(B^{(1)}) \ll \frac{1}{2N}$, $\sqrt{1+\mu^2}\,\varepsilon_1 \ll \min\left\{\frac{1}{2}, \sqrt{\frac{\delta}{\mu+1}}\right\}\delta$ ;

(ii) $a_{11}^{(1)} \geq a_{22}^{(1)} \geq \dots \geq a_{nn}^{(1)}$, $b_{11}^{(1)} = b_{22}^{(1)} = \dots = b_{nn}^{(1)} = 1$ ;

(iii) $n_1 \geq 3$, $\lambda_{s_1} = \mathcal{O}(1)$, $\delta = \mathcal{O}(1)$ ;

(iv) $S(B_{11}^{(1)}) = \mathcal{O}(\varepsilon_1)$ ;

(v) $a_{13}^{(2)} = \mathcal{O}(\varepsilon_1)$, $a_{23}^{(2)} = \mathcal{O}(\varepsilon_1)$ .

Here $A^{(k)} = (a_{ij}^{(k)}) = (A_{rs}^{(k)})$, $B^{(k)} = (b_{ij}^{(k)}) = (B_{rs}^{(k)})$, $k \geq 1$, and the partition is such that the assertions of Lemma 3.6 are valid. The assumptions (i) and (ii) ensure that all the results obtained in this chapter hold for $(A^{(1)},B^{(1)})$. The assumption (iii) ensures that $\delta$ and $\lambda_{s_1}$ are not $\mathcal{O}(\varepsilon_1)$ and that the blocks $A_{11}^{(1)}$ and $B_{11}^{(1)}$ have order $\geq 3$. The next assumption prevents that $S(B_{11}^{(1)}) = \mathcal{O}(\varepsilon_1^2)$.

By (i), (ii) and (iii) we see that Lemma 3.6 can be applied to obtain

$$A_{11}^{(k)} = \lambda_{s_1} B_{11}^{(k)} + E_{11}^{(k)}, \quad \|E_{11}^{(k)}\| = \mathcal{O}(\tau_1^2) = \mathcal{O}(\varepsilon_1^2),$$

$$1 \le k \le N+1 . \tag{3.5.1}$$

By (i) we can assume that (3.5.1) holds for the first few cycles, say for $k = 1,2,\dots,2N+1$. Let us yet consider the assumption (v). From (3.5.1) and from (iv) we have $A_{11}^{(1)} = \mathcal{O}(\varepsilon_1)$. Since (3.3.12) implies

$$|a_{13}^{(2)}|^2 + |a_{23}^{(2)}|^2 \le \frac{1}{1-|b_{12}^{(1)}|} \left(|a_{13}^{(1)}|^2 + |a_{23}^{(1)}|^2\right)$$

we have $|a_{13}^{(2)}|^2 + |a_{23}^{(2)}|^2 = \mathcal{O}(\varepsilon_1^2)$, hence the assumption (v) is realistic. The assumption (v) prevents that $a_{13}^{(2)} = \mathcal{O}(\varepsilon_1^2) = a_{23}^{(2)}$.

Let us consider the second step of the row- or the column-cyclic Jacobi method defined by (3.1.2)–(3.1.5). In the second step $a_{13}^{(2)}$ is annihilated, while $a_{12}^{(2)}$ $(= 0)$ and $a_{23}^{(2)}$ are submitted to change.

3.15. Proposition. Let the assumptions (i)–(v) hold, and let $\gamma_2$ and $\vartheta_2$ be such that (3.1.4) and (3.1.5) hold. Then

$$\operatorname{tg} \gamma_2 = \mathcal{O}(\varepsilon_1^2) / \mathcal{O}(\varepsilon_1^2)$$

$$\operatorname{tg} 2\vartheta_2 = \mathcal{O}(\varepsilon_1^2) / \mathcal{O}(\varepsilon_1^2) .$$

Proof: From the relation (3.5.1) we obtain

$$a_{jj}^{(2)} = \lambda_{s_1} + \mathcal{O}(\varepsilon_1^2), \quad 1 \le j \le n_1, \tag{3.5.2}$$

hence

$$e_2 = a_{33}^{(2)} - a_{11}^{(2)} = \mathcal{O}(\varepsilon_1^2)$$

$$a_{11}^{(2)} + a_{33}^{(2)} = 2\lambda_{s_1} + \mathcal{O}(\varepsilon_1^2) . \tag{3.5.3}$$

From (3.5.1) we also obtain



$$a_{13}^{(2)} = \lambda_{s_1} b_{13}^{(2)} + \mathcal{O}(\varepsilon_1^2) ,$$

hence $b_{13}^{(2)} = \mathcal{O}(\varepsilon_1)$. By the relation (3.1.1) we have

$$u_2 + iv_2 = a_{13}^{(2)} \overline{b_{13}^{(2)}} / |b_{13}^{(2)}| = \lambda_{s_1} |b_{13}^{(2)}| + \mathcal{O}(\varepsilon_1^2) .$$

Thus, we obtain

$$\begin{aligned} u_2 &= \lambda_{s_1} |b_{13}^{(2)}| + \mathcal{O}(\varepsilon_1^2) \\ v_2 &= \mathcal{O}(\varepsilon_1^2) . \end{aligned} \tag{3.5.4}$$

The relation (3.1.4) implies

$$\operatorname{tg} \gamma_2 = 2 v_2 / |e_2| .$$

hence, the first assertion of Proposition 3.15 follows from (3.5.4) and (3.5.3).

The relations (3.1.5), (3.5.2) and (3.5.3) imply

$$\begin{aligned} \operatorname{tg} 2\vartheta_2 &= \sigma_2 \frac{2u_2 - (a_{11}^{(2)} + a_{33}^{(2)}) |b_{13}^{(2)}|}{\sqrt{e_2^2 + 4v_2^2}\sqrt{1-|b_{13}^{(2)}|^2}} = \\ &= \sigma_2 \frac{2\lambda_{s_1}|b_{13}^{(2)}| + \mathcal{O}(\varepsilon_1^2) - 2\lambda_{s_1}|b_{13}^{(2)}| - \mathcal{O}(\varepsilon_1^2)|b_{13}^{(2)}|}{\sqrt{\mathcal{O}(\varepsilon_1^4) + 4\mathcal{O}(\varepsilon_1^4)}\sqrt{1-\mathcal{O}(\varepsilon_1^2)}} = \\ &= \sigma_2 \frac{\mathcal{O}(\varepsilon_1^2) + \mathcal{O}(\varepsilon_1^2)}{\mathcal{O}(\varepsilon_1^2)\,\mathcal{O}(1)} = \frac{\mathcal{O}(\varepsilon_1^2)}{\mathcal{O}(\varepsilon_1^2)} . \quad \blacksquare \end{aligned}$$

Proposition 3.15 implies that $\gamma_2$ and $2\vartheta_2$ can take any value from $\left(-\frac{\pi}{2}, \frac{\pi}{2}\right]$. Then the relations (3.1.2) and (3.1.3) imply that generally

$$\hat{P} = \begin{bmatrix} \mathcal{O}(1) & \mathcal{O}(1) \\ \mathcal{O}(1) & \mathcal{O}(1) \end{bmatrix} .$$

Since $a_{12}^{(2)} = 0$ the transformation formulae obtain

$$a_{12}^{(3)} = (\hat{F}_2)_{11} a_{12}^{(2)} + \overline{(\hat{F}_2)_{21}} a_{23}^{(2)} = O(1)\,O(\varepsilon_1) = O(\varepsilon_1) .$$

During the next $N-2$ steps of the first cycle the contributions to $a_{12}^{(3)}$ can be either $O(\varepsilon_1)$ (provided that the pivot element lies in the block $A_{11}^{(k)}$) or $O(\varepsilon_1^2)$. Generally, we can conclude that $a_{12}^{(N+1)} = O(\varepsilon_1)$, hence $\varepsilon_{N+1} = O(\varepsilon_1)$. Note that the same analysis applies to the matrix $B_{11}^{(1)}$. If $n_1$ is larger than three then there are more elements with chances to remain $O(\varepsilon_1)$ during the whole cycle. If there are more multiple eigenvalues then the probability for $\varepsilon_{N+1}$ to be $O(\varepsilon_1)$ is larger. This probability grows provided the assumption (ii) fails or if the pivot strategy is arbitrary cyclic.

Suppose $\varepsilon_{N+1} = O(\varepsilon_1)$, $\varepsilon_{2N+1} = O(\varepsilon_1)$ and $\tau_1 = O(\varepsilon_1)$. By Theorem 3.14(i) and (ii) we have $\tau_{N+1} = O(\varepsilon_1^2)$, $\tau_{2N+1} = O(\varepsilon_1^3)$. By (3.5.1) we have $a_{ii}^{(2N+1)} = \lambda_{s_1} + O(\tau_{2N+1}^2) = \lambda_{s_1} + O(\varepsilon_1^6)$, $1 \le i \le n_1$. We see that even if $\varepsilon_k$ is stationary per cycle the diagonal elements approach the corresponding eigenvalues with a rate proportional to $\varepsilon_k^2$. This observation together with the global convergence theorem suggests that the function

$$\mathrm{dif}(r) = \sum_{i=1}^{n} |a_{ii}^{(rN+1)} - a_{ii}^{((r-1)N+1)}|, \quad r \ge 1, \qquad (3.5.5)$$

can be used, together with $\varepsilon_{rN+1}$ in the stopping criterion of the process.

Let us now consider the accuracy of the computed eigenvalues if the finite floating point arithmetic is used. By $ep$ we denote the basic relative rounding error for the assumed arithmetic. As has been shown by J.H.Wilkinson (see [21] and



[22]) for the standard arithmetic operations

$$fl(x \circ y) = (x \circ y)(1+\eta), \ |\eta| \le ep, \qquad (3.5.6)$$

holds, where $\circ \in \{+,-,\cdot,/\}$. In (3.5.6) $fl(x)$ denotes the computed (rounded, stored) value of $x$. The relative error of the computed result $fl(x \circ y)$ is $\eta$ and it depends on $x$, $y$ and $\circ$.

3.16. Example. Solve the following problem: let the relation (3.5.6) and the assumptions (i)–(iv) hold; make a qualitative analysis of the relative errors in the computed values of $\mathrm{tg}\,2\vartheta_2$ and $\mathrm{tg}\,\gamma_2$ due to the rounding errors.

Solution: Since $a_{11}^{(2)}$ and $a_{33}^{(2)}$ have been changed we can assume

$$fl(a_{11}^{(2)}) = a_{11}^{(2)}(1+\eta_{11}), \quad |\eta_{11}| \le ep,$$
$$fl(a_{33}^{(2)}) = a_{33}^{(2)}(1+\eta_{33}), \quad |\eta_{33}| \le ep .$$

By (3.5.6) we have

$$\begin{aligned} fl(e_2) &= (fl(a_{33}^{(2)}) - fl(a_{11}^{(2)}))(1+\tilde{\eta}_2), \ |\tilde{\eta}_2| \le ep, \\ &= (a_{33}^{(2)}(1+\eta_{33}) - a_{11}^{(3)}(1+\eta_{11}))(1+\tilde{\eta}_2) = \\ &= (e_2 + a_{33}^{(2)}\eta_{33} - a_{11}^{(3)}\eta_{11})(1+\tilde{\eta}_2) = \\ &= e_2(1+\eta_2), \end{aligned}$$

where

$$\eta_2 = \tilde{\eta}_2 + \frac{a_{33}^{(2)}\eta_{33} - a_{11}^{(2)}\eta_{11}}{e_2}(1+\eta_2) .$$

By (3.5.3) we have $e_2 = O(\varepsilon_1^2)$ hence

$$\eta_2 = O(ep) + \frac{O(ep)}{O(\varepsilon_1^2)}(1+O(ep)) = O(ep) + O(ep/\varepsilon_1^2) = O(ep/\varepsilon_1^2) .$$

Since $e_2$ stands in the denominator of the quotients which determine $tg\,2\vartheta_2$ and $tg\,\gamma_2$ we have

$$fl(tg\,2\vartheta_2) = tg\,2\vartheta_2(1+\eta'), \qquad fl(tg\,\gamma_2) = tg\,\gamma_2(1+\eta'') \tag{3.5.7}$$

where

$$\eta' = \mathcal{O}(ep/\varepsilon_1^2), \quad \eta'' = \mathcal{O}(ep/\varepsilon_1^2), \tag{3.5.8}$$

provided that the expresions in the numerators have not larger relative errors then $fl(e_2)$. ■

Suppose $\varepsilon = 10^{-6}$, $ep = 10^{-16}$. By (3.5.7) and (3.5.8) we can assume that $tg\,2\vartheta_2$ and $tg\,\gamma_2$ have approximately four certain decimal digits. Observing the formulae (3.1.4) we can assume that the elements of $\hat{F}_2$ also have four certain decimal digits. Since $fl(a_{13}^{(3)})$ as a residual of the equation $a_{13}^{(3)} = 0$ has small absolute error we can expect, say $fl(a_{13}^{(3)}) = \mathcal{O}(10^{-8})$.

Therefore, if the algorithm is designed to set the elements $a_{\ell m}^{(k+1)}$ and $b_{\ell m}^{(k+1)}$ zero, it may cause large absolute errors in the matrix elements. This fact can result in large errors in the computed eigenvalues and eigenvectors (see [18],[4]). Also, if B is almost singular the errors in the computed elements of $B^{(k)}$ can cause that $B^{(k)}$ is not anymore positive definite for large k's, i.e.the algorithm fails to work.

To remove this danger we can compute the elements of $\hat{F}_k$ in a higher precision. If this is not possible then we have to compute the elements $a_{\ell m}^{(k+1)}$ and $b_{\ell m}^{(k+1)}$.

Another possibility standing at disposal is to modify



the algorithm in such situation as those assumed in Proposition 3.15 and Example 3.16. The task of such a program should be to improve both, the asymptotic rate of convergence and the accuracy of the computed eigenvalues and eigenvectors. This program is presented in the next section.

## 3.6. Modified Method

We propose here a modification of the special Jacobi method (here called original method) such that the quadratic convergence extends to the case of arbitrary eigenvalues. Our strategy is the following: in the first stage we let the original method work. Multiple (or almost multiple) eigenvalues are indicated by the structure described in Section 1.2 and then the modification is used.

The modified method has the form

$$A^{(k+1)} = H_k^* A^{(k)} H_k, \quad B^{(k+1)} = H_k^* B^{(k)} H_k, \quad k \geq 1, \tag{3.6.1}$$

where the elementary plane matrices $H_k$ are defined below (see the relation (3.6.7)).

Suppose that the pair $(A^{(1)}, B^{(1)})$ is almost diagonal, i.e. $\varepsilon_1 \ll \delta$, where $\varepsilon_1$ and $\delta$ are defined by the relations (3.1.1) and (1.2.6), respectively. Let $\hat{H}_k = I_2 + \hat{H}_k'$, $k \geq 1$. Then the conditions

$$\|\hat{H}_k'\| = \mathcal{O}(\varepsilon_1), \quad 1 \leq k \leq N, \tag{3.6.2}$$

and

$$|a_{\ell m}^{(k+1)}|^2 + |b_{\ell m}^{(k+1)}|^2 = \mathcal{O}(\varepsilon_1^4), \quad 1 \le k \le N, \qquad (3.6.3)$$

are sufficient to prove that $\varepsilon_{N+1} = \mathcal{O}(\varepsilon_1^2)$. Indeed by (3.6.3) and (3.6.2) we have

$$a_{\ell m}^{(k+r)} = \mathcal{O}(\varepsilon_1^2) = b_{\ell m}^{(k+r)}, \quad 1 \le r \le N-k+1,$$

hence all the off-diagonal elements of $A^{(N+1)}$ and $B^{(N+1)}$ are $\mathcal{O}(\varepsilon_1^2)$.

Suppose $H_k = F_k$ $(1 \le k \le N)$, where each $F_k$ is defined by (3.1.2). Then the condition (3.6.3) trivially holds since $a_{\ell m}^{(k+1)} = 0 = b_{\ell m}^{(k+1)}$. Unfortunately the condition (3.6.2) is not always fulfilled. In Section 3.5 it has been shown that the condition (3.6.2) does not necessarily hold for those k's for which $a_{\ell\ell}^{(k)}$ and $a_{mm}^{(k)}$ converge to the identical eigenvalues. In such situations we can set $H_k = G_k$ where $G_k$ is a modified (elementary plane) transformation.

## Modified Transformation

Let $\hat{G}_k$ be 2-2 nonsingular upper-triangular matrix satisfying

$$\hat{G}_k^* \hat{B}^{(k)} \hat{G}_k = I_2$$

or equivalently

$$\hat{B}^{(k)-1} = \hat{G}_k \hat{G}_k^* . \qquad (3.6.4)$$

Then a simple calculation yields

$$\hat{G}_k = \frac{1}{\sqrt{1-b_k^2}} \begin{bmatrix} \sqrt{1-b_k^2} & -b_{\ell m}^{(k)} \\ 0 & 1 \end{bmatrix} . \qquad (3.6.5)$$



Setting

$$\cos\varphi_k' = \sqrt{1-b_k^2}, \quad \sin\varphi_k' = b_k, \quad \alpha_k' = \arg(b_{\ell m}^{(k)}) + \pi,$$
$$\cos\psi_k' = 1, \quad \sin\psi_k' = 0, \quad \beta_k' = 0,$$

we can rewrite the relation (3.6.5) in the form

$$\hat{G}_k = \frac{1}{t_k} \begin{bmatrix} \cos\varphi_k' & e^{i\alpha_k'} \sin\varphi_k' \\ -e^{-i\beta_k'} \sin\psi_k' & \cos\psi_k' \end{bmatrix}, \qquad (3.6.6)$$

where $t_k = \sqrt{1-b_k^2}$. Let $G_k$ be the elementary plane matrix having $\hat{G}_k$ as its $(\ell,m)$-restriction.

We define the transformation of the modified method as follows

$$H_k = \begin{cases} G_k & \text{if} \quad k \ge r_o N + 1 \quad \text{and} \quad |a_{mm}^{(k)} - a_{\ell\ell}^{(k)}| < \delta \\ F_k & \text{otherwise}, \end{cases} \qquad (3.6.7)$$

where $r_o$ satisfies

$$2n(1+\mu^2)\varepsilon_{r_o N+1} < \delta, \quad S(B^{(r_o N+1)}) < \frac{1}{2N}. \qquad (3.6.8)$$

From the definitions (3.6.1), (3.6.7) and (3.6.8) of the modified method we see that it represents a theoretical model not directly applicable in praxis. However, it can serve for modelling the quadratically convergent algorithms.

## Quadratic Convergence

We prove the quadratic convergence of the modified method under an arbitrary cyclic pivot strategy. The asymptotic as-

sumptions are

$$S(B^{(1)}) < \frac{1}{2N}\,, \tag{3.6.9}$$

$$2n(1+\mu^2)\varepsilon_1 < \delta \tag{3.6.10}$$

and

$$1 < p < n\,, \quad n \geq 3\,. \tag{3.6.11}$$

We consider the first cycle of the modified method when applied to the pair $(A^{(1)},B^{(1)})$ ( $(A^{(1)},B^{(1)})$ is "normalized" such that the condition (2.2.2) holds) satisfying the asymptotic assumptions (3.6.9)–(3.6.11). As already indicated by the relation (3.6.5) we use the notation (3.1.1) also for the modified method. Therefore, the quantities such as $\varepsilon_k$, $\tau_k$, $a_k$, $b_k$, $b$, ... are related to the pair $(A^{(k)},B^{(k)})$ obtained by the modified method. Note that the assumptions (3.6.9) and (3.6.10) imply that $r_o = 0$ in (3.6.8). From the relation (3.6.7) it follows that $H_k = G_k$ iff $|a_{mm}^{(k)} - a_{\ell\ell}^{(k)}| < \delta$ (cf. Remark 3.18).

Set

$$S'' = \left\{k \in \{1,2,\dots,N\};\ |a_{mm}^{(k)} - a_{\ell\ell}^{(k)}| < \delta\right\}.$$

3.17. Lemma. Let the assumptions (3.6.9)–(3.6.11) hold and let the pivot strategy be arbitrary. Then the following inequalities hold

$$b < 0.88\,\frac{1}{2N}\,, \tag{3.6.12}$$

$$S^2(B^{(k+1)}) \leq 1.754\,S^2(B^{(1)})\,, \quad 1 \leq k \leq N\,, \tag{3.6.13}$$

$$S^2(A^{(k+1)}) \leq \varepsilon_{k+1}^2 \leq 1.863442\,\varepsilon_1^2\,, \quad 1 \leq k \leq N\,, \tag{3.6.14}$$

$$|a_{\ell m}^{(k+1)}| \leq 0.7721\,\frac{1+\mu^2}{\delta}\,\varepsilon_1^2\,, \quad k \in S''\,. \tag{3.6.15}$$

In addition, if



$$\omega_k = \begin{cases} \max\{\varphi_k,\psi_k\}, & k \notin S'' \\ \varphi_k', & k \in S''\,, \end{cases}$$

then

$$\sum_{k=1}^{N} \sin^2\omega_k \leq 0.676446\,\frac{1+\mu^2}{\delta^2}\,\varepsilon_1^2\,. \tag{3.6.16}$$

Proof: From the equality (3.6.4) we easily obtain

$$\|\hat{G}_k\|_2^2 = \|\hat{B}^{(k)-1}\|_2 = \frac{1}{1-b_k} = x_k\,, \quad k \in S''\,, \tag{3.6.17}$$

hence the proof of the assertions (3.6.12) and (3.6.13) is identical to the related proof for the original method from Lemma 3.3.

Next we prove the assertion (3.6.14).

Let $k \in S''$ be such that the inequality $\sqrt{1+\mu^2}\,\varepsilon_k < \delta$ holds. Note that the assumptions (3.6.9)–(3.6.11) and the assertion (3.2.6) of Lemma 3.3 imply that the smallest element of $S''$ meets the above requirement. For such a $k$ and for each pair $i < j$ (cf. Corollary 1.3) either $|a_{jj}^{(k)} - a_{ii}^{(k)}| < \delta$ or $|a_{jj}^{(k)} - a_{ii}^{(k)}| > \delta$ holds. In the former case the elements $a_{ii}^{(k)}$ and $a_{jj}^{(k)}$ are affiliated to the same eigenvalue. Consequently, $a_{\ell\ell}^{(k)}$ and $a_{mm}^{(k)}$ are affiliated to the same eigenvalue (see the definition of $S''$), say $\lambda$. Then Theorem 1.2 implies

$$\hat{A}^{(k)} = \lambda\hat{B}^{(k)} + \hat{E}''^{(k)}\,, \tag{3.6.18}$$

where

$$\|E''^{(k)}\| \leq \frac{1+\mu^2}{2\delta}\,\tau_k^2 \leq \frac{1+\mu^2}{2\delta}\,\varepsilon_k^2\,. \tag{3.6.19}$$

By the equality (3.6.18) we have

$$\hat{G}_k^*\hat{A}^{(k)}\hat{G}_k = \lambda\hat{G}_k^*\hat{B}^{(k)}\hat{G}_k + \hat{G}_k^*\hat{E}''^{(k)}\hat{G}_k\,.$$

Note that $\hat{G}_k^* \hat{E}^{(k)} \hat{G}_k$ is hermitian and $b_{\ell m}^{(k+1)} = 0$. Using the inequalities (3.6.19) and (3.6.17) we obtain

$$2|a_{\ell m}^{(k+1)}|^2 = 2|(\hat{G}_k^* \hat{E}''^{(k)} \hat{G}_k)_{\ell m}|^2 \leq \|\hat{G}_k^* \hat{E}''^{(k)} \hat{G}_k\|^2 \leq$$
$$\leq \|\hat{G}_k\|_2^4 \cdot \|E''^{(k)}\|^2 \leq \left[\frac{1}{1-b_k} \frac{1+\mu^2}{2\delta} \varepsilon_k^2\right]^2 . \qquad (3.6.20)$$

Since the inequalities

$$S^2(A^{(k+1)}) \leq \frac{1}{1-b_k} S^2(A^{(k)}) - 2a_k^2 + 2|a_{\ell m}^{(k+1)}|^2$$

and

$$S^2(B^{(k+1)}) \leq \frac{1}{1-b_k} S^2(B^{(k)}) - 2b_k^2$$

hold (cf. Lemma 3.2) we set

$$\nu = \frac{1}{1-b} \frac{(1+\mu^2)^2}{4\delta^2}$$

and use the relation (3.6.20) to obtain

$$\varepsilon_{k+1}^2 \leq \frac{1}{1-b} (1 + \nu\varepsilon_k^2)\varepsilon_k^2 - 2(a_k^2 + b_k^2) . \qquad (3.6.21)$$

If $k \notin S''$ then the inequality (3.6.21) also holds (we can set $\nu = 0$ in this case) provided $\sqrt{1+\mu^2}\,\varepsilon_k < \delta$.

We prove by induction with respect to $k$ that

$$\varepsilon_k \leq \left[\frac{1 + 2(k-1)\nu\varepsilon_1^2}{(1-b)^{k-1}}\right]^{1/2} \cdot \varepsilon_1 \leq 1.863442\,\varepsilon_1 ,$$
$$1 \leq k \leq N+1 . \qquad (3.6.22)$$

For $k = 1$ the inequalities in (3.6.22) trivially hold. Note that the relation (3.6.21) also holds for $k = 1$. Suppose that the relation (3.6.22) holds for a $k \in \{1,2,\ldots,N\}$; then we prove that it holds for $k+1$.

The induction hypothesis (3.6.22) and the assumption



(3.6.10) imply that $\sqrt{1+\mu^2}\,\varepsilon_k \leq 2\sqrt{1+\mu^2}\,\varepsilon_1 < \delta$ holds. Hence we can use the relation (3.6.21) to obtain

$$\varepsilon_{k+1}^2 \leq \frac{1}{1-b}\left[1 + \nu\varepsilon_1^2 \cdot \frac{1 + 2(k-1)\nu\varepsilon_1^2}{(1-b)^{k-1}}\right] \frac{1 + 2(k-1)\nu\varepsilon_1^2}{(1-b)^{k-1}} \cdot \varepsilon_1^2 \leq$$
$$\leq \frac{\varepsilon_1^2}{(1-b)^k}\left[1 + 2(k-1)\nu\varepsilon_1^2\right]\left[1 + \nu\varepsilon_1^2 \cdot \frac{1 + 2(k-1)\nu\varepsilon_1^2}{(1-b)^{k-1}}\right] . \qquad (3.6.23)$$

By the assumption (3.6.10) and the definition of $\nu$ we have

$$2(k-1)\nu\varepsilon_1^2 \leq 2N\nu\varepsilon_1^2 \leq n^2 \frac{(1+\mu^2)^2}{4\delta^2} \cdot \varepsilon_1^2 \cdot \frac{1}{1-b} <$$
$$< \frac{1}{16} \frac{1}{1-b} . \qquad (3.6.24)$$

Using the relations (3.6.24), (3.6.12) and (3.2.13) we obtain

$$\frac{1 + 2(k-1)\nu\varepsilon_1^2}{(1-b)^{k-1}} < \frac{1 + \frac{1}{16}}{(1-b)^k} < \frac{17}{16}\left(1 + \frac{6}{7}\sqrt{\frac{4+\sqrt{2}}{7}}\right) <$$
$$< 1.863442 . \qquad (3.6.25)$$

The relations (3.6.25), (3.6.24), (3.6.12) and (3.6.11) imply

$$\frac{1 + 2(k-1)\nu\varepsilon_1^2}{(1-b)^{k-1}} \cdot (1 + 2(k-1)\nu\varepsilon_1^2) <$$
$$< 1.863442\left(1 + \frac{1}{16} \frac{1}{1 - \frac{0.88}{6}}\right) < 2 ,$$

hence the inequality (3.6.23) implies

$$\varepsilon_{k+1} \leq \frac{\varepsilon_1^2}{(1-b)^k} (1 + 2(k-1)\nu\varepsilon_1^2 + 2\nu\varepsilon_1^2) = \frac{1 + 2k\nu\varepsilon_1^2}{(1-b)^k} \cdot \varepsilon_1^2 .$$

This proves the first inequality in (3.6.22). The second inequality follows from the relation (3.6.25) which holds

for all $1 \le k \le N+1$. The relation (3.6.22) implies the assertion (3.6.14).

To prove the assertion (3.6.15) we use the relations (3.6.20), (3.6.12) and (3.6.14). We obtain

$$|a_{\ell m}^{(k+1)}| \le \frac{\sqrt{2}}{2} \frac{1}{1-b} \frac{1+\mu^2}{2\delta} \varepsilon_k^2 \le$$

$$\le \frac{\sqrt{2}}{2} \frac{6}{5.12} \frac{1}{2} \frac{1+\mu^2}{\delta} 1.863442\, \varepsilon_1^2 \le$$

$$\le 0.7721 \frac{1+\mu^2}{\delta} \varepsilon_1^2, \quad k \in S''.$$

Finally, we prove (3.6.16). From the relations (3.6.14), (3.6.10), (3.6.11) and (3.6.12) we obtain

$$1 + \nu \varepsilon_k^2 \le 1 + \frac{(1+\mu^2)^2 \cdot 1.863442\, \varepsilon_1^2}{(1-b)\, 4\delta^2}$$

$$1 + \frac{6}{5.12} \frac{1.863442}{4} \frac{1}{4n^2} < 1 + \frac{0.1365}{N}.$$

Using these inequalities in the relation (3.6.21) we obtain

$$\varepsilon_{N+1}^2 \le \frac{\left(1 + 0.1365 \frac{1}{2N}\right)^N}{(1-b)^N} \cdot \varepsilon_1^2 - 2 \sum_{k=1}^{N} (a_k^2 + b_k^2).$$

By the assertion (iii) of Lemma 3.1 and the assumption (3.6.11) we have

$$\left(1 + 0.1365 \frac{1}{2N}\right)^N \le 1 + \frac{4}{3} N \cdot 0.1365 \frac{1}{2N} < 1 + \frac{2}{3} 0.1365 = 1.091,$$

hence it follows

$$\sum_{k=1}^{N} (a_k^2 + b_k^2) \le \frac{1}{2} \cdot 1.754 \cdot 1.091\, \varepsilon_1^2 \le 0.957 \cdot \varepsilon_1^2. \quad (3.6.27)$$

Suppose $k \in S''$. Then the relations (3.6.5), (3.6.6) and (3.2.27) imply



$$\sin^2 \omega_k = \sin^2 \varphi_k \le b_k^2 \le \frac{4}{9} \frac{1+\mu^2}{\delta^2} \cdot b_k^2 \le$$

$$\le \frac{4}{9} \frac{1+\mu^2}{\delta^2} (a_k^2 + b_k^2).$$

If $k \notin S''$ then the relation (3.2.31) is used. Note that the assumptions (3.6.10), (3.6.11) imply the assumptions (3.1.9), (3.1.10). We conclude that the inequality (3.2.31) holds for all $1 \le k \le N$ hence the assertion (3.6.11) follows from the inequalities (3.6.27) and (3.2.31). ■

3.18. Remark. The relation (3.6.14) and the assumptions (3.6.10), (3.6.11) imply

$$(1+\mu^2) \varepsilon_k^2 \le (1+\mu^2)\, 1.864\, \varepsilon_1^2 \le 1.864 \frac{\delta}{2n} \varepsilon_1 <$$

$$< 1.864 \frac{\delta^2}{4n^2} < \frac{1.864}{36} \delta^2, \quad 1 \le k \le N+1.$$

Following the proof of Lemma 3.4 we obtain that for all $i < j$ and $1 \le k \le N+1$ either $|a_{jj}^{(k)} - a_{ii}^{(k)}| < 0.052\,\delta$ or $|a_{jj}^{(k)} - a_{ii}^{(k)}| > 2.948\,\delta$ holds. Therefore the elements $a_{\ell\ell}^{(k)}$ and $a_{mm}^{(k)}$ are affiliated to the same eigenvalue iff $k \in S''$. ■

3.19. Theorem. Let the assumptions (3.6.9)–(3.6.11) hold and let the sequence $((A^{(k)}, B^{(k)}), k \ge 1)$ be generated by the modified Jacobi method defined by the relations (3.6.1), (3.6.7) and (3.6.8). If the pivot strategy is arbitrary cyclic then

$$\varepsilon_{N+1} \le 2.525 \sqrt{N}\, (1+\mu^2) \frac{\varepsilon_1^2}{\delta}.$$

Proof: The proof resembles the proof of Theorem 3.9 hence

we use the same notation. By an analysis, similar to that in the proof of Theorem 3.9 we obtain

$$z_j = \sqrt{y_{r_j}}\,(z_{j-1}\cos\tilde{\omega}_{r_j} \pm a^{(r_j)} e^{i\nu_j}\sin\omega_{r_j}),$$

$$1 \le j \le s \qquad (3.6.28)$$

where for $j = 1,2,\dots,s$

$$\tilde{\omega}_{r_j}, \omega_{r_j} \in \begin{cases} \{\varphi'_{r_j}, \psi'_{r_j}\}, & r_j \in S'' \\ \{\varphi_{r_j}, \psi_{r_j}\}, & r_j \notin S'' \end{cases},$$

$$\nu_j \in \begin{cases} \{\alpha'_{r_j}, \beta'_{r_j}, -\alpha'_{r_j}, -\beta'_{r_j}\}, & r_j \in S'' \\ \{\alpha_{r_j}, \beta_{r_j}, -\alpha_{r_j}, -\beta_{r_j}\}, & r_j \notin S''. \end{cases}$$

Here $a^{(r_j)}$ is an off-diagonal element of $A^{(r_j)}$ and $y_{r_j} = 1/(1-b^2_{r_j})$. By definition $z_0 = a^{(k+1)}_{\ell m}$ where $\ell, m, k$ are fixed. Note that $z_0 \neq 0$ implies $k \in S''$. By the assertion (3.6.15) of Lemma 3.17 we have

$$|z_0| = |a^{(k+1)}_{\ell m}| \le 0.7721\,\frac{1+\mu^2}{\delta}\,\varepsilon_1^2, \quad k \in S''. \qquad (3.6.29)$$

From the relation (3.6.28) we obtain

$$|z_j| \le \sqrt{y_{r_1} y_{r_2} \cdots y_{r_j}}\,|z_0| + \sum_{i=1}^{j} \sqrt{y_{r_i} y_{r_{i+1}} \cdots y_{r_j}}\,|a^{(r_i)}||\sin\omega_{r_i}| \le$$

$$\le (1-b^2)^{-j/2}\Big(|z_0| + \sum_{i=1}^{j} |a^{(r_i)}||\sin\omega_{r_i}|\Big),$$

$$1 \le j \le s. \qquad (3.6.30)$$

At the end of the full cycle the off-diagonal part



$E^{(N+1)}$ of $A^{(N+1)}$ consist exactly of the elements $z_s$. Note that $r_s$ is the suffix of the last transformation in the first cycle which changed the element at the position $(\ell,m)$. Therefore the relation (3.6.30) implies

$$|E^{(N+1)}| \le (1-b^2)^{-(n-2)}(|Z'| + |P^{(2)}||\sin\omega_2| + \dots$$

$$\dots + |P^{(N)}||\sin\omega_N|) \qquad (3.6.31)$$

where $P^{(k)}$'s are as in the proof of Theorem 3.9, and $Z' = (z'_{ij})$ is hermitian matrix having zeros on the diagonal and $z'_{\ell m} = a^{(k+1)}_{\ell m}$, $\ell < m$, in its upper triangle. Writing (cf.(3.4.1)) $M = N - n_1(n_1-1)/2 - \dots - n_p(n_p-1)/2$ we obtain from the relation (3.6.29)

$$\|\,|Z'|\,\| = \|Z'\| \le \sqrt{2(N-M)}\;0.7721\,\frac{1+\mu^2}{\delta}\,\varepsilon_1^2. \qquad (3.6.33)$$

Since the nonzero elements of $P^{(k)}$ are some off-diagonal elements of $A^{(k)}$, the assertion (3.6.14) of Lemma 3.17 implies

$$\|\,|P^{(k)}|\,\| = \|P^{(k)}\| \le S(A^{(k)}) \le \varepsilon_k \le \sqrt{1.863442}\,\varepsilon_1,$$

$$2 \le k \le N. \qquad (3.6.34)$$

Using Lemma 3.1(ii) and the relations (3.6.11), (3.6.12) we obtain (cf.the relation (3.3.7))

$$(1-b^2)^{-(n-2)} \le [(1-b^2)^{-N}]^{1/n} \le \Big(1+\frac{72}{67}N\cdot b^2\Big)^{1/3} \le$$

$$\le \Big(1+\frac{72}{67}\,\frac{0.88^2}{4\,3}\Big)^{1/3} < 1.023. \qquad (3.6.35)$$

Using the inequalities (3.6.33), (3.6.34), (3.6.35) and (3.6.16) in the inequality (3.6.31) we obtain

$$S(A^{(N+1)}) = \|E^{(N+1)}\| = \||E^{(N+1)}|\| \leq$$

$$\leq 1.023\,(0.7721\sqrt{2N}\,\frac{1+\mu^2}{\delta}\,\varepsilon_1^2 + \sqrt{1.863442}\,\varepsilon_1 \sum_{k=1}^{N} |\sin\omega_k|) \leq$$

$$\leq \Big[1.023\;1.092\sqrt{N}\,\frac{1+\mu^2}{\delta}\,\varepsilon_1^2 + \;1.863442\,\varepsilon_1 \Big(N\sum_{k=1}^{N}\sin^2\omega_k\Big)^{1/2}\Big] \leq$$

$$\leq \;1.023\Big(1.092 + \sqrt{1.863442\;0.676446}\sqrt{N}\,\frac{1+\mu^2}{\delta}\,\varepsilon_1^2\Big) \leq$$

$$\leq \;2.26567\sqrt{N}\,\frac{1+\mu^2}{\delta}\,\varepsilon_1^2 . \qquad (3.6.36)$$

The above analysis applies to the matrix $B^{(1)}$ with $Z' = 0$. The assertion (3.6.13) of Lemma 3.17 implies

$$\||P^{(k)}|\| \leq S(B^{(k)}) \leq \sqrt{1.754}\,S(B^{(1)}), \quad 2 \leq k \leq N,$$

hence we have

$$S(B^{(N+1)}) \leq 1.023\sqrt{1.754}\,S(B^{(1)})\Big(N\sum_{k=1}^{N}\sin^2\omega_k\Big)^{1/2} \leq$$

$$\leq 1.023\,\sqrt{1.754\;0.676446}\;\frac{\sqrt{N(1+\mu^2)}}{\delta}\;\varepsilon_1\,S(B^{(1)}) \leq$$

$$\leq 1.1143125\sqrt{N}\,\frac{1+\mu^2}{\delta}\,\varepsilon_1^2 . \qquad (3.6.37)$$

Using the relations (3.6.36), (3.6.37) and the definition of $\varepsilon_{N+1}$ we finally obtain

$$\varepsilon_{N+1} = (S^2(A^{(N+1)}) + S^2(B^{(N+1)})^{1/2} \leq$$

$$\leq 2.525\sqrt{N}\,\frac{1+\mu^2}{\delta}\,\varepsilon_1^2 ,$$

which proves Theorem 3.19. ■

Global Convergence

3.20. Theorem. The row- and the column-cyclic modified Jacobi method defined by the relations (3.6.1), (3.6.7) and



(3.6.8) is globally convergent.

Proof: By Corollary 2.2 the row- and the column-cyclic original (i.e. special Jacobi) method converges. Therefore there is an integer $r_0 \geq 1$ (depending on the initial pair (A,B)) such that the inequalities (3.6.8) hold. At this point the modification switches on hence we can use Theorem 3.19 and the relations (3.6.8), (3.6.10) to obtain

$$\varepsilon_{(r+1)N+1} \leq 2.525\,\frac{\sqrt{2}}{2}\,\frac{n(1+\mu^2)\varepsilon_{rN+1}}{\delta}\,\varepsilon_{rN+1} \leq$$

$$\leq \frac{2.525\sqrt{2}}{4}\,\varepsilon_{rN+1} \leq 0.9\,\varepsilon_{rN+1}, \quad r \geq r_0 .$$

Thus $\varepsilon_{rN+1} \longrightarrow 0$, $r \longrightarrow \infty$. Using the assertion (3.6.14) of Lemma 3.17 in connection with the r'th cycle as $r \longrightarrow \infty$, we obtain $\varepsilon_k \longrightarrow 0$, $k \longrightarrow \infty$. Since $b_{ii}^{(k)} = 1$ $(1 \leq i \leq n,\ k \geq 1)$, the diagonal elements $a_{ii}^{(k)}$ converge to the appropriate eigenvalues. ■

In the case of real matrices the quadratically convergent modifications can be defined in a similar way.

## 3.7. Numerical Tests

Here we briefly discuss the results of a numerical investigation of the asymptotic convergence of the special Jacobi method from Section 2.2. We also describe some preliminary results concerning the modified method from Section 3.6. We do not describe the algorithms in details ; we rather discuss

the phenomena.

The tests were performed on the IBM-1010 computer of the Fernuniversität-Gesamthochschule in Hagen and on the Hewlett-Packard System 1000 of the Institut Ruđer Bošković in Zagreb. The algorithms are written in the programming language FORTRAN IV (REAL*8) using the floating point arithmetic with the machine precision $ep \approx 10^{-16}$.

The iterated hermitian matrices $A^{(k)}$, $B^{(k)}$ are stored in two n-n arrays. The real (imaginary) part of each matrix is stored in the upper (strictly lower resp.) triangle of the corresponding array. The initial matrices A, B are generated by the following three modes:

1 $\quad A = G_r^* D G_r, \quad B = G_r^* G_r,$

2 $\quad A = F_r + F_r^*, \quad B = G_r^* G_r,$

3 $\quad$ A, B are read from the input file.

Here $G_r$, $F_r$ denote the matrices whose elements are generated by a procedure generating random numbers and D denotes a diagonal matrix whose diagonal elements are read from the input file. We use the first mode when the asymptotic behaviour of the algorithms is investigated. The second mode is used for obtaining additional information concerning the average number of cycles (sweeps) needed for a diagonalization.

A lot of work has been devoted to increase efficiency and reliability of the algorithms. Fast scaled transformations are used except for the diagonal elements of $A^{(k)}$ which are updated after each step as if the normal transformations were used. Several useful devices due to H.Rutishauser (see



[17] or [24] Contribution II/1) and P.J.Eberlein (see [3] or [24] Contribution II/17) are applied. In modelling a quadratically convergent algorithm based on the modified method from Section 3.6 difficulties arrise in making decision when (i.e.in what sweep) and where (at what pivot pair) to implement the modified transformation. Every misuse of the modified transformation is paid by an increase in the total number of sweeps needed to reach the stopping criterion. We use a sequence of tests (enclosed in a loop) which takes into account the current stage of the diagonalization and the estimates from Section 2.2. Although a foolproof criterion for implementing the modification is still lacking the experimental results confirm that in certain situations an essential improvement in the asymptotic convergence can be obtained. Instead of using the (expensive for computation) quantities $S(A^{(k)})$, $S(B^{(k)})$, $\varepsilon_k$, in the algorithm we use $\tilde{S}(A^{(k)})$, $\tilde{S}(B^{(k)})$, $\tilde{\varepsilon}_k$, respectively. The latter quantities denote the sums of the absolute values of the off-diagonal entries of the appropriate arrays. The stopping criterion for the both algorithms is the same: either if no transformation is performed during one full sweep or $\tilde{\varepsilon}_{rN+1} < 10^{-2}\sqrt{ep}\max|a_{jj}^{(rN+1)}|$ holds for an r $(1 \le r \le 20)$ then the process is stopped.

Comperative tests were made on over a hundred matrix pairs. Approximately one half of the matrices were generated by the mode 1, the other half by the mode 2. The order of matrices varied from five to forty. In the asymptotic behaviour of the algorithms we distinguish two cases.

A. The Case of Simple Eigenvalues

In this case the modified method coincides with the original method (i.e.the special Jacobi method). Only if the eigenvalues are distributed in clusters of very small size (pathologically closed eigenvalues) then the two algorithms give different results.

The number of sweeps needed to reach the stopping criterion depends on the magnitude of $\tilde{\varepsilon}_1$ and on the distribution of the eigenvalues. Usually, after $\tilde{\varepsilon}_{rN+1}$ reaches the interval $[0.5, 5]$ five more sweeps are needed. The numbers $\tilde{S}(A^{(rN+1)})$ and $\tilde{S}(B^{(rN+1)})$ have asymptotically the same order of magnitude. The reduction of $\tilde{\varepsilon}_{rN+1}$ with r is asymptotically quadratic. The value of dif(r) from the relation (3.5.5) is asymptotically $O(\tilde{\varepsilon}^2_{(r-1)N+1})$ which agrees with the analysis from Section 3.5.

For the matrix pairs generated by the mode 2 the average number of sweeps needed to reach the stopping criterion is six. For smaller n $(5 \leq n \leq 15)$ this number is five. For larger n $(30 \leq n \leq 40)$ it is seven.

If the eigenvalues form clusters then the total number of sweeps is essentially increased. Usually eight to fourteen sweeps are needed in these cases. The upper bound fourteen is reached provided the eigenvalues form several clusters of different size. If the size of a cluster is smaller than $\sqrt{\varepsilon p}$ then the two algorithms give different results.

We illustrate these conclusions by the following two examples.



3.20. Example. The pair (A,B) is generated by the mode 1 with the diagonal matrix

D = diag(1000,800,500,200,10,-300,-600,-1200) .

The output for the both algorithms is the following:

ORDER OF MATRICES = 8

NUMBER OF SWEEPS = 7

NUMBER OF STEPS = 178

CAU-TIME = 0.4933 SEC (central arithmetic unit time)

| r | $\tilde{S}(A^{(rN+1)})$ | $\tilde{S}(B^{(rN+1)})$ | dif(r) |
|---|---|---|---|
| 1 | 11.962 | 7.325 | 14.115 |
| 2 | 8.249 | 3.537 | 3.340 |
| 3 | 2.057 | 1.666 | 1.911 |
| 4 | 0.227 | 0.218 | 0.569 |
| 5 | 0.358-2 | 0.417-2 | 0.360-1 |
| 6 | 0.200-7 | 0.150-7 | 0.456-4 |
| 7 | 0.956-21 | 0.100-20 | 0.444-15 |

3.21. Example. The pair (A,B) is generated by the mode 1 with the diagonal matrix

D = diag(1,1.00001,0.99991,1.00002,2,2.00001,
1.99991,2.00002,0,0.00001,-0.00001,3,
3.0001,2.99991) .

The output for the both algorithms is the following :

ORDER OF MATRICES = 14

NUMBER OF SWEEPS = 10

NUMBER OF STEPS = 785

CAU-TIME = 2.8667 SEC

| r | $\tilde{S}(A^{(rN+1)})$ | $\tilde{S}(B^{(rN+1)})$ | dif (r) |
|---|---|---|---|
| 1 | 13.637 | 22.001 | 8.900 |
| 2 | 9.273 | 13.120 | 4.005 |
| 3 | 3.933 | 5.664 | 1.925 |
| 4 | 0.900 | 1.812 | 0.685 |
| 5 | 0.873-1 | 0.175 | 0.201 |
| 6 | 0.238-2 | 0.439-2 | 0.621-2 |
| 7 | 0.254-3 | 0.535-3 | 0.892-4 |
| 8 | 0.913-5 | 0.155-4 | 0.227-5 |
| 9 | 0.472-10 | 0.103-9 | 0.999-10 |
| 10 | 0.270-24 | 0.600-24 | 0.0 |

B. The Case of Multiple Eigenvalues

In this case the rate of convergence of the original method is pretty unstable. As in the case A, $\tilde{S}(A^{(rN+1)})$ and $\tilde{S}(B^{(rN+1)})$ have asymptotically the same order of magnitude hence we consider only $\tilde{\varepsilon}_{rN+1}$. For matrices of modest size $(5 \leq n \leq 15)$ the reduction of $\tilde{\varepsilon}_{rN+1}$ with r is quite irregular. For some successive r's, $\tilde{\varepsilon}_{rN+1}$ remains almost constant and for the next few r's it decreases better than quadratically. However, usually $\tilde{\varepsilon}_{rN+1}$ decreases quadratically for $r \geq 4$. For matrices of larger size $(15 \leq n \leq 40)$ with larger multiplicities ($n_i \geq 3$ for some i's) the convergence of $\tilde{\varepsilon}_{rN+1}$ is usually slowed down. The total number of sweeps is generally larger than in the case A due to the slowdown of convergence. For matrices of larger order it increases up to fifteen. In the case of slowdown in the reduction of $\tilde{\varepsilon}_{rN+1}$, the quantity dif (r) has order of magni-



tude $O(\tilde{\varepsilon}^2_{(r-1)N+1})$. This fact is in accordance with the analysis from Section 3.5. Namely, for small $\tilde{\tau}_{rN+1}$ we have

$$|a_{jj}^{(rN+1)} - a_{jj}^{((r-1)N+1)}| =$$

$$= |\lambda_j - a_{jj}^{((r-1)N+1)}| + O(\tau^2_{rN+1}) =$$

$$= O(\tau^2_{(r-1)N+1}) + O(\tau^2_{rN+1}) .$$

The modified algorithm shows better convergence properties. If the multiple eigenvalues form no cluster then the number of sweeps and the CAU-time is usually decreased by 10% to 20%, in some instances up to 30%. The quantity $\tilde{S}(B^{(rN+1)})$ converges (asymptotically) quadratically as in the case A. The quantity $\tilde{S}(A^{(rN+1)})$ converges quadratically to some point (usually up to the order of magnitude $\sqrt{\varepsilon p}$) and then slows down a little. However, this slowdown is not essential; it can prolong the computation for at most one sweep. This slowdown is due to the fact that $a_{\ell m}^{(k+1)} \neq 0$ for some k's in each cycle. If the eigenvalues make clusters then usually the modified algorithm is not faster than the original one. This fact is due to the criterion for switching on the modified step and we believe that the modified algorithm can be yet improved.

We end this section by the following

3.22. Example. The pair (A,B) is generated by the model with the diagonal matrix

$$D = \mathrm{diag}(0,0,0,0,0,0,1,1,1,1,1,1,1,1,-1,-1,-2,-2,-2,-3) .$$

The results are as follows:

ORDER OF MATRICES = 20

| | ORIGINAL METHOD | MODIFIED METHOD |
|---|---|---|
| NUMBER OF SWEEPS | 13 | 9 |
| NUMBER OF STEPS | 2081 | 1654 |
| CAU-TIME | 9.4467 | 7.5133 |

| | Original method | | | Modified method | | |
|---|---|---|---|---|---|---|
| r | $\tilde{S}(A^{(rN+1)})$ | $\tilde{S}(B^{(rN+1)})$ | dif (r) | $\tilde{S}(A^{(rN+1)})$ | $\tilde{S}(B^{(rN+1)})$ | dif (r) |
| 1 | 30.663 | 38.844 | 12.602 | 30.663 | 36.844 | 12.602 |
| 2 | 18.058 | 21.836 | 5.785 | 18.058 | 21.836 | 5.785 |
| 3 | 6.979 | 11.943 | 2.207 | 6.979 | 11.943 | 2.207 |
| 4 | 1.873 | 5.938 | 1.143 | 1.873 | 5.938 | 1.143 |
| 5 | 0.371 | 1.482 | 0.683 | 0.371 | 1.482 | 0.683 |
| 6 | 0.269-1 | 0.107 | 0.129 | 0.269-1 | 0.107 | 0.129 |
| 7 | 0.204-3 | 0.734-3 | 0.218-2 | 0.204-3 | 0.734-3 | 0.218-2 |
| 8 | 0.943-5 | 0.532-4 | 0.512-6 | 0.179-7 | 0.268-7 | 0.512-6 |
| 9 | 0.121-5 | 0.709-5 | 0.276-11 | 0.519-11 | 0.671-17 | 0.444-15 |
| 10 | 0.689-7 | 0.264-6 | 0.127-12 | 0.0 | 0.0 | 0.0 |
| 11 | 0.239-8 | 0.679-8 | 0.254-13 | 0.0 | 0.0 | 0.0 |
| 12 | 0.909-10 | 0.148-9 | 0.155-14 | 0.0 | 0.0 | 0.0 |
| 13 | 0.774-14 | 0.666-14 | 0.0 | 0.0 | 0.0 | 0.0 |



## Principal Notation

Display (relations): (p,q,r) r'th display in Section q of Chapter p

Scalars (real or complex numbers): $\vartheta$, $\varphi$, $\psi$, $\omega$, $\xi$, $\alpha$, $\beta$, $\gamma$, $\sigma$, $\vartheta_k$, $\tilde{\vartheta}_k$, $\vartheta_{ij}$, ... , $x_i$, $y_i$, $a_{ij}$, $b_{ij}$, ...

Complex conjugate: $\bar{\xi}$, $\bar{x}_i$, $\bar{a}_{ij}$, ...

Absolute values (of scalars, vectors, matrices): $|\xi|$, $|x_i|$, $|a_{ij}|$, $|x| = (|x_i|)$, $|A| = (|a_{ij}|)$, ...

Vectors (column): x, y, $x = (x_i)$, $y = (y_i)$, ...

$e_j$ - the j'th column of the identity matrix $I_n$

$a_t$ - the t'th column of the matrix A

Matrices: A, $A = (a_{ij})$, $A = (A_{rt})$ hermitian matrix with elements $a_{ij}$ and blocks $A_{rt}$,

B, $B = (b_{ij})$, $B = (B_{rt})$ positive definite hermitian matrix with elements $b_{ij}$ and blocks $B_{rt}$,

$A^{(k)}$, $A^{(k)} = (a_{ij}^{(k)})$, $A^{(k)} = (A_{rt}^{(k)})$, $B^{(k)}$,... matrices generated by a Jacobi method,

I, $I_n$ the identity matrix (of order n)

C, $C = (c_{ij})$ a general matrix

D, $D^{(k)}$, $D_o$, $\Delta$, $\operatorname{diag}(C) = \operatorname{diag}(c_{11},\dots,c_{nn})$ diagonal matrices

Dimension: n all matrices and vectors are of order n if not stated otherwise, $n \geq 3$

Scalar product: $(x|y) = \sum x_i \bar{y}_i$

Conjugate transpose (matrices, vectors): $C^* = (\bar{c}_{ji})$

Transpose (matrices, vectors): $C^T = (c_{ji})$

Determinants: $\det C$, $\det(C)$

Norms:

$\|x\| = \sqrt{(x|x)}$ Euclidean (vector) norm

$\|C\| = \sqrt{\sum_{i,j} |c_{ij}|^2}$ Euclidean (matrix) norm

$\|C\|_2 = \max_{x \neq 0} \|Cx\|/\|x\|$ spectral norm

Pair: $(A,B)$ A hermitian, B positive definite matrix

Eigenvalues (of the pair $(A,B)$): $\lambda_1, \lambda_2, \dots, \lambda_n$, $\lambda_{s_1} > \lambda_{s_2} > \dots > \lambda_{s_p}$ distinct eigenvalues

Singular values (of a matrix $K$): $\varkappa_1, \varkappa_2, \dots, \varkappa_n$ — the positive square roots of the eigenvalues of $K^*K$

Multiplicities (of the eigenvalues): $n_i = s_i - s_{i-1}$, $1 \leq i \leq n$, $s_o = 0$

Special functions of the eigenvalues: $\delta_i = \min_{\substack{1 \leq l \leq p \\ j \neq l}} |\lambda_{s_i} - \lambda_{s_j}|$, $\delta = \min_{1 \leq i \leq p} \delta_i$

Spectral radius (of the pair $(A,B)$):

$$\mu = \mathrm{spr}(A,B) = \max_{1 \leq i \leq n} |\lambda_i| = \max_{x \neq 0} |(Ax,x)|/(Bx,x)$$

Indices:

$k$ - counts iterates (steps),

$l, m$ - pivot indices, $1 \leq l < m \leq n$, $l = l(k)$, $m = m(k)$

$q_i$ - index function: $q_i = q_{i-1} + (n-i)$, $1 \leq i \leq n-1$, $q_o = 0$

$s_i$ - special indices of the eigenvalues

$i,j,r,t,\dots$ - other indices



Sequences (of matrices, matrix pairs): $(C^{(k)}) = (C^{(k)}, k \geq 1)$, $((A^{(k)}, B^{(k)}), k \geq 1)$

Special functions of matrices:

$S(C) = \|C - \mathrm{diag}(C)\|$

$$\tau(C) = \sqrt{\sum_{\substack{r,t \\ r \neq t}} \|C_{rt}\|^2}$$

$$S(A,B) = \sqrt{S^2(D_o A D_o) + S^2(D_o B D_o)}$$

$D_o = \mathrm{diag}(1/\sqrt{b_{11}}, \dots, 1/\sqrt{b_{nn}})$

$$\tau(A,B) = \sqrt{\tau^2(P^* D_o A D_o P) + \tau^2(P^* D_o B D_o P)}$$

$P$ - appropriate permutation matrix

$\varepsilon_k = S(A^{(k)}, B^{(k)})$, $\tau_k = \tau(A^{(k)}, B^{(k)})$

$$\tilde{S}(C) = \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} (|\mathrm{Re}(c_{ij})| + |\mathrm{Im}(c_{ij})|)$$

$\tilde{\varepsilon}_k = \tilde{S}(A^{(k)}) + \tilde{S}(B^{(k)})$

Special notation used in connection with the special Jacobi method for the pair $(A,B)$: see the relation (3.1.1)

References


[1] Bathe, K.J. and E.L.Wilson: Numerical Methods in Finite Element Analysis. Prentice-Hall, Inc., Englewood Cliffs, N.J., 1976

[2] Bathe, K.J. and E.L.Wilson: Solution Methods for Eigenvalue Problems in Structural Mechanics. Int.J.for Num. Methods in Engineering 6(1973), 213–226.

[3] Eberlein, P.J.: Solution to the Complex Eigenproblem by a Norm-Reducing Jacoby-Type Method. Numer.Math.14(1970), 232–245.

[4] Elsner, L. and Ji-guang Sun: Perturbation Theorems for the Generalized Eigenvalue Problem. Lin.Alg.and Its Appl.48(1982), 341–357.

[5] Falk, S. and P.Langemeyer: Das Jacobische Rotations-Verfahren für reelsymmetrische Matrizen-Paare I,II. Elektronische Datenverarbeitung 1960.

[6] Forsythe, G.E. and P.Henrici: The Cyclic Jacobi Method for Computing the Principal Values of a Complex Matrix. Trans.Amer.Math.Soc.94(1960), 1–23.

[7] Gentleman, W.M.: Least Squares Computations by Givens Transformations without Square Roots. J.Inst.Math.Appl. 12(1973), 392–336.

[8] Gose, G.: Das Jacobi Verfahren für $Ax = \lambda Bx$. ZAMM 59(1979), 93–101.

[9] Hansen, E.R.: On Cyclic Jacobi Methods. SIAM J.Appl. Math. 11 (1963), 448–459.

[10] Hari, V.: On the Global Convergence of Cyclic Jacobi Methods for the Generalized Eigenproblem. Preprint, University of Zagreb.

[11] Hari, V.: On the Convergence to Diagonal Form of Complex Jacobi-Like Processes. Preprint, University of Zagreb.

[12] Henrici, P. and K.Zimmermann: An Estimate for the Norms of Certain Cyclic Jacobi Operators. Lin.Alg.and Its Appl. 1 (1968), 489–501.

[13] Kempen, H.P.M.van : On the Quadratic Convergence of the Serial Cyclic Jacobi Method. Numer.Math. 9 (1966), 19–22.

[14] Parlett, B.N.: Symmetric Eigenvalue Problem. Prentice-Hall, Inc., Englewood Cliffs, N.J., 1980.

[15] Rath, W.: Fast Givens Rotations for Orthogonal Similarity Transformations. Numer.Math. 40 (1982), 47–56.

[16] Rutishauser, H.: Simultaneous Iteration Method for Symmetric Matrices. Numer.Math. 16 (1970), 205–223.

[17] Rutishauser, H.: The Jacobi Method for Real Symmetric Matrices. Numer.Math. 9 (1966), 1–10.

[18] Stewart, G.W.: Perturbation Bounds for the Definite Generalized Eigenvalue Problem. Lin.Alg.and Its Appl. 23 (1979), 69–85.

[19] Voevodin, V.V.: Čislennye metody linejnoj algebry. Izdateľstvo Nauka, Glavnaja redakcija fiziko-matematičeskoj literatury, Moskva 1966.



[20] Wilkinson, J.H.: Note on the Quadratic Convergence of the Cyclic Jacobi Processes. Numer.Math. 4 (1962), 296–300.

[21] Wilkinson, J.H.: Rounding Errors in Algebraic Processes. Prentice-Hall, Inc., Englewood Cliffs, N.J., 1964.

[22] Wilkinson, J.H.: The Algebraic Eigenvalue Problem. Oxford University Press, Oxford 1965.

[23] Wilkinson, J.H.: Almost Diagonal Matrices with Multiple or Close Eigenvalues. Lin.Alg.and Its Appl. 1 (1968), 1–12.

[24] Wilkinson, J.H. and C.Reinsch: Handbook for Automatic Computation. Linear Algebra, Vol.2. Springer-Verlag, New York, Heidelberg, Berlin, 1971.

[25] Zimmermann, K.: On the Convergence of a Jacobi Process for Ordinary and Generalized Eigenvalue Problems. Dissertation No 4305 (1965), Zürich: Eidgenossische Technische Hochschule.

# L e b e n s l a u f

| | |
|---|---|
| Name: | Vjeran Hari |
| Geboren: | 14.November 1949. in Sisak, Jugoslawien |
| Vater: | Ivan Hari, verstorben am 1970 |
| Mutter: | Vjera Hari, geb.Hikec |
| Beruf des Vaters: | Lehrer |
| Staatsangehörigkeit: | jugoslawisch |
| Konfession: | römisch-katholisch |
| Grundsschule: | 1956 – 1964 in Zagreb |
| Gymnasium: | 1964 – 1968 in Zagreb |
| Abitur: | Juni 1968 |
| Studium: | 1968 – 1973 Universität Zagreb |
| Diplom: | Oktober 1973 |
| Magister: | Juli 1980 |
| Wehrdienst: | Juli 1975 – Juli 1976 |
| Tätigkeiten: | wissenschaftlicher Mitarbeiter an der Institut Ruder Bošković in Zagreb, 1.Januar 1974 – 31.December 1974 |

Zagreb, den 12.04.1984