ADVANCES IN JACOBI METHODS

Zlatko Drmač
drmac@math.hr
Department of Mathematics
University of Zagreb
Croatia

Vjeran Hari
hari@math.hr
Department of Mathematics
University of Zagreb
Croatia

Ivan Slapničar
ivan.slapnicar@fesb.hr
FESB
University of Split
Croatia

Abstract Two-sided and especially one-sided Jacobi methods for solving eigenproblems of Hermitian positive definite and symmetric indefinite matrices are reviewed. SVD one-sided methods are included in this survey. Special attention is paid to the latest results on accuracy and on asymptotic convergence of scaled iterates by these methods.

Keywords: Jacobi methods, one-sided methods, accuracy, scaled iterates, asymptotic convergence.

Introduction

It is known that diagonalization methods deliver output data with high relative accuracy, a property which is not shared with faster methods such as Divide and Conquer or QR methods which require tridiagonalization or bidiagonalization as a preprocessing step. After the pioneering research of Demmel and Veselić [6], accuracy of other methods for solving different matrix eigenproblems has been inspected (see [25, 41]). Closely related stopping criterions and measures of convergence have also been reinvestigated.

In this overview paper we present the latest research of Jacobi methods for solving Hermitian/symmetric eigenvalue problem and the singular value problem. The new results mostly address accuracy and the asymptotic convergence of scaled iterates.

The paper is divided into three sections. In Section 1 are briefly described two- and one-sided Jacobi methods for Hermitian/symmetric eigenvalue problem. A special attention is paid to J-symmetric Jacobi method of Veselić [51] and its one-sided version which is an excellent tool for accurate eigensolving of indefinite symmetric matrices. In Section 2, their asymptotic convergence is reviewed and several new results are shortly explained. Finally, in Section 3, the latest accuracy results concerning one-sided methods are presented.

1. Two-sided Methods

Here we first give a short introduction to Jacobi-type processes [36, 17]. Then we give a brief description of the the most important methods: Hermitian Jacobi method and J-symmetric Jacobi method. For each of the two methods, we introduce both the two- and one-sided versions.

1.1 Jacobi-type Processes

Jacobi-type methods are iterative processes of the form

$$A^{(k+1)} \;=\; P_k^* \, A^{(k)} \, Q_k, \quad k \geq 0, \tag{1.1}$$

where P_k, Q_k are nonsingular *elementary plane matrices*, P_k^* is the complex transpose of P_k and $A^{(0)} = A$ is the initial matrix of order n. An elementary plane matrix E is a nonsingular matrix which differs from identity matrix I_n in one principal submatrix of order two, denote it by $\hat{E}$, which is called *pivot submatrix* or the (i,j)-*restriction* of E. The pair of indices i, j, $i < j$ which determine position of $\hat{E}$ within E are *pivot indices* and (i,j) is *pivot pair*. In (1.1) pivot indices depend on k, so $i = i(k)$, $j = j(k)$. When emphasis is on pivot indices we shall write $P_{i(k)j(k)}$ instead of P_k (similar for Q_k) and when k is understood just P_{ij}. Transition from $A^{(k)}$ to $A^{(k+1)}$ is called the kth step or iteration of the method.

Jacobi methods are characterized by the requirement

$$a_{i(k)j(k)}^{(k+1)} \;=\; a_{j(k)i(k)}^{(k+1)} \;=\; 0, \quad k \geq 0.$$

which implies that for each k, $\hat{P}_k$ and $\hat{Q}_k$ are computed from $\hat{A}_k$.

Pivot strategy is a rule for selecting pivot pairs. We identify each pivot strategy with a function $\mathrm{I}: \mathbf{N}_0 \to \mathbf{P}_n$, where $\mathbf{N}_0 = \{0, 1, 2, \ldots\}$ and $\mathbf{P}_n = \{(l,m); 1 \leq l < m \leq n\}$, $n \geq 2$. Among different classes of (pivot) strategies we consider here only *periodic strategies* which are periodic functions. This means $\mathrm{I}(k+M) = \mathrm{I}(k)$, $k \geq 0$ for a positive integer M, the *period* of I. A periodic strategy I is *quasi-cyclic* if $\{\mathrm{I}(k); 0 \leq k \leq M-1\} = \mathbf{P}_n$, and a quasi-cyclic strategy is *cyclic* if $M = N \stackrel{\text{def}}{=} n(n-1)/2$. The set of M successive iterations starting with k which is a multiple of M, is usually called a *quasi-cycle* (*cycle* for cyclic strategies). In the sequel the term strategy means periodic strategy. The most common are the row-cyclic and column-cyclic strategies (often referred to as serial) and the parallel ones.

The column-cyclic strategy is defined by $\mathrm{I}_C(k) = (i(k), j(k))$, $k \geq 0$, where as k increases, the pivot pair runs through the column-wise ordering of $\mathbf{P}_n$: $(1,2)$, $(1,3)$, $(2,3)$,…,$(1,n)$, $(2,n)$, …,$(n-1,n)$ of $\mathbf{P}_n$. The row-cyclic strategy I_R is defined in a similar way by the row-wise ordering of $\mathbf{P}_n$: $(1,2)$, $(1,3)$, …,$(1,n)$, $(2,3)$, …,$(n-1,n)$. By parallel strategy we mean a cyclic strategy for which the defining ordering of $\mathbf{P}_n$ can be divided into p subsequences each containing mutually disjoint pairs (pairs (l,m) and (p,q) are disjoint if $\{l,m\} \cap \{p,q\} = \emptyset$). The plane matrices associated with each such subsequence mutually commute, they all can be computed in the same time and all can be applied simultaneously. Best efficiency is obtained when $p \approx n$. Then each subsequence contains around $n/2$ pairs (see [3, 28, 18, 29]). A cycle then consists of p parallel steps and each parallel step comprises $\approx n/2$ sequential steps.

Another interesting periodic strategy is the quasi-cyclic strategy of Mascarenhas [30, 31, 37] which enables cubic asymptotic convergence per quasi-cycle and for which $M \approx 1.25N$.

The notion of convergence depends on the method. For two-sided methods it usually means convergence of the iterated matrix to a diagonal matrix.

1.2 Hermitian Jacobi Method

Let $H = (h_{lm})$ be a (complex) Hermitian matrix of order n. Jacobi method for computing the eigendecomposition of Hermitian matrices generates sequence of Hermitian matrices by the rule (1.1) with $P_k = Q_k$, $k \geq 0$ being unitary matrices. Rewriting (1.1) with these assumptions, yields

$$H^{(k+1)} = V_k^* H^{(k)} V_k, \quad k \geq 0, \quad H^{(0)} = H,$$

where V_k, $k \geq 0$ are unitary plane matrices. If eigenvectors are wanted, then $V^{(k)} = V_0 V_1 \cdots V_{k-1}$ is computed by the rule $V^{(k)} = V^{(k-1)} V_{k-1}$, $k \geq 1$, $V^{(0)} = I_n$. For each $k \geq 0$ pivot submatrix of V_k,

$$\hat{V}_k = \begin{bmatrix} \cos\varphi^{(k)} & -\sin\varphi^{(k)} e^{\imath\omega_k} \\ \sin\varphi^{(k)} e^{-\imath\omega_k} & \cos\varphi^{(k)} \end{bmatrix}$$

is chosen to diagonalize the pivot submatrix $\hat{H}^{(k)}$ of $H^{(k)} = (h_{lm}^{(k)})$. Here $\imath$ denotes the imaginary unit and $\overline{z}$ denotes the complex conjugate of $z \in \mathbf{C}$. Hence, $\varphi^{(k)}$ and ω_k are chosen to produce $h_{ij}^{(k+1)} = 0$. For $h_{ij}^{(k)} = 0$, the k'th step is skipped i.e. $\omega_k = \varphi^{(k)} = 0$ is presumed. Otherwise, the usual choice

$$e^{\imath\omega_k} = \frac{h_{ij}^{(k)}}{|h_{ij}^{(k)}|}, \quad \tan 2\varphi^{(k)} = \frac{2|h_{ij}^{(k)}|}{h_{ii}^{(k)} - h_{jj}^{(k)}}, \quad \varphi^{(k)} \in [-\pi/4\,,\, \pi/4]\,, \tag{1.2}$$

is assumed. If H is real symmetric, ω_k is set zero, so that all $V^{(k)}$ are orthogonal and all $H^{(k)}$ are real symmetric. In this case $|h_{ij}^{(k)}|$ in the relation (1.2) is replaced by $h_{ij}^{(k)}$. Using $c^{(k)} = \cos\varphi^{(k)}$, $s^{(k)} = \sin\varphi^{(k)}$, $t^{(k)} = \tan\varphi^{(k)}$, the transformation formulas read

$$\begin{aligned}
h_{li}^{(k+1)} &= c^{(k)} h_{li}^{(k)} + e^{-\imath\omega_k} s^{(k)} h_{lj}^{(k)} \ = \overline{h_{il}^{(k+1)}}, \quad l \notin \{i,j\} \\
h_{lj}^{(k+1)} &= c^{(k)} h_{lj}^{(k)} - e^{\imath\omega_k} s^{(k)} h_{li}^{(k)} \ = \overline{h_{jl}^{(k+1)}}, \quad l \notin \{i,j\} \\
h_{ii}^{(k+1)} &= h_{ii}^{(k)} + |h_{ij}^{(k)}|\, t^{(k)}, \quad h_{jj}^{(k+1)} = h_{jj}^{(k)} - |h_{ij}^{(k)}|\, t^{(k)} \\
h_{ij}^{(k+1)} &= 0, \qquad h_{lm}^{(k+1)} = h_{lm}^{(k)} \quad \text{whenever} \quad l, m \notin \{i,j\}\,.
\end{aligned} \tag{1.3}$$

As a consequence of the choice of transformation parameters, we have

$$\|\Omega(H^{(k+1)})\|_F^2 = \|\Omega(H^{(k)})\|_F^2 - 2\ |\ h_{ij}^{(k)}\ |^2\,, \quad k \geq 0\,, \tag{1.4}$$

where generally, $\Omega(X) = X - \mathrm{diag}(X)$ stands for the off-diagonal part of X and $\|\cdot\|_F$ is the Frobenius (or Euclidean) matrix norm. By $\|\cdot\|_2$ is denoted the spectral or matrix 2-norm. The quantity $\|\Omega(X)\|_F$ is sometimes referred to as (see [20]) the *off-norm* of X. The quantity $\|\Omega(H^{(k)})\|_F$ can be used to measure the progress of the process.

The serial Jacobi methods are globally convergent, i.e. the sequence of matrices $H^{(k)}$ is convergent to diagonal matrix for any initial Hermitian H (see [16] and [24]). The asymptotically faster quasi-cyclic method defined by Mascarenhas strategy is also globally convergent [37].

If $H \in \mathbf{C}^{n\times n}$ is positive definite, then all $H^{(k)}$, $k \geq 0$ are positive definite and Jacobi method delivers relatively accurate eigenvalues and eigenvectors (almost as accurate as H allows, see [6]). Scaled matrices

$$H_S^{(k)} = [\mathrm{diag}(H^{(k)})]^{-1/2} H^{(k)} [\mathrm{diag}(H^{(k)})]^{-1/2}, \quad k \geq 0. \tag{1.5}$$

play important role concerning accuracy. Their off-diagonal parts, and the corresponding norms

$$A^{(k)} = \Omega(H_S^{(k)}) = H_S^{(k)} - I, \quad \alpha_k = \|A^{(k)}\|_F \quad k \geq 0,$$

are used in terminating of the process (see (2.8)). This will be discussed later. Note that the diagonal elements of $A^{(k)} = (a_{lm}^{(k)})$, $k \geq 0$ are zeros and the off-diagonal elements are given by

$$a_{lm}^{(k)} = h_{lm}^{(k)} / \sqrt{h_{ll}^{(k)} h_{mm}^{(k)}}, \quad l \neq m, \quad k \geq 0. \tag{1.6}$$

The property (1.4) is not shared with the sequence of scaled matrices since

$$H_S^{(k+1)} = (\Delta^{(k+1)})^{-1} (V^{(k)})^* \Delta^{(k)} H_S^{(k)} \Delta^{(k)} V^{(k)} (\Delta^{(k+1)})^{-1}$$

is not a unitary transformation. Actually, the off-norm of scaled matrices can temporarily increase during the process. In [34] it has been proven

$$\begin{aligned} |a_{il}^{(k+1)}|^2 + |a_{jl}^{(k+1)}|^2 &\leq \frac{|a_{il}^{(k)}|^2+|a_{jl}^{(k)}|^2}{1-|a_{ij}^{(k)}|}, \quad l \neq i,j \ , \\ \|A^{(k+1)}\|^2 - \|A^{(k)}\|^2 &\leq |a_{ij}^{(k)}| \frac{\|A^{(k)}\|^2-2|a_{ij}^{(k)}|}{1-|a_{ij}^{(k)}|} \ . \end{aligned} \tag{1.7}$$

Hence, if $\|A^{(k+1)}\| > \|A^{(k)}\|$, then $0 < |a_{ij}^{(k)}| \ \leq \ \frac{1}{2}\|A^{(k)}\|^2$. We see that only quadratically small (in the scaled sense) pivot element can cause the growth of $\|A^{(k)}\|$. The above estimates make it possible to find an upper bound for the finite sequence $\alpha_0, \alpha_1, \ldots, \alpha_N$ provided that α_0 is small enough.

Scaled iterates can be defined in indefinite case, whenever diagonal elements are nonzero. Then the scaling matrix $[\mathrm{diag}(H^{(k)})]^{-1/2}$ should be replaced by $[\mid \mathrm{diag}(H^{(k)}) \mid]^{-1/2} = [\mathrm{diag}(|H^{(k)}|)]^{-1/2}$, the diagonal elements of $H_S^{(k)}$ are ± 1, $A^{(k)} = \Omega(H_S^{(k)})$ and $a_{lm}^{(k)} = h_{lm}^{(k)}/\sqrt{\mid h_{ll}^{(k)} h_{mm}^{(k)} \mid}$, $l \neq m$, $k \geq 0$. Here $|X| = (|x_{lm}|)$ provided that $X = (x_{lm})$.

One-Sided Jacobi Method Let $H = LL^T$ be the Cholesky factorization of positive definite H, and let $L = U\Sigma V^T$ be the SVD of L. Then, since U is orthogonal and Σ diagonal, the decomposition $H = U\Sigma^2 U^T$ is the spectral decomposition of H. Hence, we can diagonalize H in two steps: (1) compute the Cholesky factorization to get L and (2) compute the SVD of L. For the second task we can use the so called *one-sided* (or *right-handed*) Jacobi method. This method appears naturally, when two-sided Jacobi method is applied to $L^T L$. After k steps one obtains the matrix $(LV_0 \cdots V_{k-1})^T LV_0 \cdots V_{k-1}$, hence it is sufficient to iterate the process $G^{(k+1)} = G^{(k)} V_k$, $k \geq 0$, $G^{(0)} = L$. To compute the parameters of $\hat{V}_k$ one needs (see (1.2)) to compute scalar product of pivot columns $(G^{(k)} e_i \mid G^{(k)} e_j)$ and squares of pivot column norms $\|G^{(k)} e_i\|^2$ and $\|G^{(k)} e_j\|^2$ (e_l is the lth column of I_n). Fortunately, using formulas for updating the diagonal elements from (1.3) the squares of norms have to be computed once or twice during the whole process. The effect of kth transformation is to orthogonalize the pivot columns of $G^{(k)}$. Assuming a "convergent" (e.g. the serial) pivot strategy, the sequence $G^{(k)}$ approaches a set of matrices with orthogonal columns. Let $G^{(k)} = U^{(k)} \Sigma^{(k)}$, where $U^{(k)}$ has normalized columns and $\Sigma^{(k)}$ is diagonal positive definite. We can choose any convergent subsequence of $V^{(k)} = V_0 \cdots V_k$ and an appropriate convergent subsequence of $U^{(k)}$ to obtain in the limit $U\Sigma = LV$, the SVD of L. Left singular vectors of L are eigenvectors of H and squares of the singular values of L are the eigenvalues of H. In the context of computing the spectral decomposition of H, right singular vectors are not needed, hence accumulation of rotations can be skipped. This and several other attractive features of the one-sided method are first noted, analyzed and exploited by Veselić and Hari [52]. We shall address some of these features in Subsection 3.1.

1.3 J-Symmetric Jacobi Method

J-symmetric Jacobi method, introduced by Veselić in [51], is a diagonalization method for the generalized eigenvalue problem $Hx = \lambda Jx$ where $H \in \mathbf{R}^{n\times n}$ is symmetric matrix and $J = \mathrm{diag}(I_m, I_{n-m})$ is direct sum of I_m and $-I_{n-m}$. The method generates the sequence of matrices by the rule (1.1) with $P_k = Q_k$, $k \geq 0$. Since transformation matrices need not be orthogonal, we denote them by C_k, $k \geq 0$ so that (1.1) takes form

$$H^{(k+1)} = C_k^T H^{(k)} C_k, \quad k \geq 0, \quad H^{(0)} = H.$$

The method is so designed that $C_k^T J C_k = J$ holds for all $k \geq 0$. We see that transition from $H^{(k)}$ to $H^{(k+1)}$ is made by congruence transformation which preserves the symmetry and inertia of matrices $H^{(k)}$ as well as the eigenvalues of the initial pair (H, J). Since all C_k are $J-orthogonal$, the method is sometimes called J-orthogonal (see [43]). Although the method can be generalized to complex matrices, all known results refer to real matrices.

For $1 \leq i < j \leq m$ and $m+1 \leq i < j \leq n$, C_k is plane rotation and for $1 \leq i \leq m < j \leq n$ it is J-orthogonal plane matrix (*hyperbolic rotation*). Their pivot submatrices have form

$$\hat{C}^{(k)} = \begin{bmatrix} \cos\varphi^{(k)} & -\sin\varphi^{(k)} \\ \sin\varphi^{(k)} & \cos\varphi^{(k)} \end{bmatrix} \quad \text{and} \quad \hat{C}^{(k)} = \begin{bmatrix} \cosh\varphi^{(k)} & \sinh\varphi^{(k)} \\ \sinh\varphi^{(k)} & \cosh\varphi^{(k)} \end{bmatrix},$$

respectively. The angle $\varphi^{(k)}$ is in either case chosen to annihilate the pivot element $h^{(k)}_{i(k)j(k)}$, which leads to very simple angle formulas

$$\tan 2\varphi^{(k)} = \frac{2h_{ij}^{(k)}}{h_{ii}^{(k)} - h_{jj}^{(k)}}, \qquad \tanh 2\varphi^{(k)} = \frac{-2h_{ij}^{(k)}}{h_{jj}^{(k)} + h_{ii}^{(k)}},$$

with $\varphi^{(k)} \in [-\pi/4,\ \pi/4]$ for orthogonal rotation. Using $c^{(k)}$, $s^{(k)}$ and $t^{(k)}$ for both the trigonometric and hyperbolic functions and $\tau = 1$ ($\tau = -1$) for hyperbolic (orthogonal) rotation, we obtain the following transformation formulas (compare to the relation (1.3))

$$
\begin{aligned}
h_{il}^{(k+1)} &= c^{(k)} h_{il}^{(k)} + s^{(k)} h_{jl}^{(k)}, \quad h_{li}^{(k+1)} = h_{il}^{(k+1)}, \quad l \notin \{i,j\} \\
h_{jl}^{(k+1)} &= c^{(k)} h_{jl}^{(k)} + \tau \cdot s^{(k)} h_{il}^{(k)}, \quad h_{lj}^{(k+1)} = h_{jl}^{(k+1)}, \quad l \notin \{i,j\} \\
h_{ii}^{(k+1)} &= h_{ii}^{(k)} + t^{(k)} h_{ij}^{(k)}, \quad h_{jj}^{(k+1)} = h_{jj}^{(k)} + \tau \cdot t^{(k)} h_{ij}^{(k)}, \\
h_{ij}^{(k+1)} &= 0, \qquad h_{pr}^{(k+1)} = h_{pr}^{(k)} \text{ if } \{p,r\} \cap \{i,j\} = \emptyset
\end{aligned}
\tag{1.8}
$$

The transformation defined by $\tau = -1$ ($\tau = 1$) is called kth *trigonometric (hyperbolic) step* or *iteration* of the method. If eigenvectors are wanted, the transformations are collected as product $C_0 C_1 \cdots$. After trigonometric step the relation (1.4) holds. To see what happens after hyperbolic step, let

$$
H^{(k)} = \begin{bmatrix} H_{11}^{(k)} & H_{12}^{(k)} \\ H_{21}^{(k)} & H_{22}^{(k)} \end{bmatrix}, \qquad J = \begin{bmatrix} I_m & \\ & -I_{n-m} \end{bmatrix}, \qquad H_{11}^{(k)} \text{ is } m \text{ by } m,
$$

$$
S^2(H^{(k)}) = \|\Omega(H_{11}^{(k)})\|_F^2 + \|\Omega(H_{22}^{(k)})\|_F^2 - 2\|H_{12}^{(k)}\|_F^2 .
$$

Then for the hyperbolic step holds (see [51])

$$
S^2(H^{(k+1)}) = S^2(H^{(k)}) + 2(h_{ij}^{(k)})^2.
$$

We can say that hyperbolic steps transfer the off-diagonal weight from the blocks $H_{12}^{(k)}$ and $H_{21}^{(k)}$ to diagonal blocks $H_{11}^{(k)}$ and $H_{22}^{(k)}$ and trigonometric steps diagonalize the diagonal blocks. Overall, $H^{(k)}$ converges to a diagonal matrix. In [51] it has been proved that the method converges globally under the serial strategies.

If H is n by n positive definite, then all $H^{(k)}$, $k \geq 0$ are such. Scaled iterate $H_S^{(k)}$ and its off-diagonal part $A^{(k)}$ are defined as above, so that relations (1.5) – (1.6) hold for this method too.

Resemblance with symmetric Jacobi method is also seen from the relations (1.7) which hold not only for trigonometric, but also for hyperbolic steps (see [35]). In particular, (1.7) has form

$$
|a_{il}^{(k+1)}|^2 + |a_{jl}^{(k+1)}|^2 = \frac{|a_{il}^{(k)}|^2 + |a_{jl}^{(k)}|^2 - 2a_{ij}^{(k)} a_{il}^{(k)} a_{jl}^{(k)}}{1 - |a_{ij}^{(k)}|^2} \leq \frac{|a_{il}^{(k)}|^2 + |a_{jl}^{(k)}|^2}{1 - |a_{ij}|}.
$$

One-Sided J-Symmetric Method J-symmetric Jacobi method also has its one-sided version which is an excellent method for accurate eigensolving of symmetric indefinite eigenproblem. Let us briefly derive the algorithm (see [40, 42, 46]).

Let $H = PG_1J_1G_1^TP^T$ be the Bunch-Parlett factorization [4] of a nonsingular, indefinite symmetric matrix H. Here P is permutation, G_1 lower block-triangular with diagonal blocks of order one or two and J_1 diagonal with ± 1 on its diagonal. Using the second permutation P_1 in

$$H = PG_1J_1G_1^TP^T = (PG_1P_1^T)(P_1J_1P_1^T)(P_1G_1^TP^T) = GJG^T$$

one can obtain $J = \mathrm{diag}(I_m, I_{n-m})$. Now $Hx = \lambda x$ can be written as $GJG^T = \lambda x$ and since G is invertible, the last equation is equivalent to $G^TGJG^Tx = \lambda G^Tx$. This can be written as $G^TGz = \lambda Jz$ with $z = JG^Tx$. Since $Gz = GJG^Tx = Hx = \lambda x$, one obtains $x = (Gz)/\lambda$. In conclusion, the initial indefinite symmetric eigenproblem $Hx = \lambda x$ is converted to $J-symmetric$ eigenproblem with positive definite matrix G^TG. To solve the new problem one can apply J-symmetric Jacobi method. As will be seen later, this transition makes sense if accuracy of all eigenvalues and eigenvectors is important. A pleasant fact with the new eigenproblem is that G^TG need not be computed. One simply postmultiplies G by the J-orthogonal transformations,

$$G^{(k+1)} = G^{(k)}C_k\,, \quad k \geq 0, \qquad G^{(0)} = G\,.$$

Using additional vector to store the diagonal of $(G^{(k)})^TG^{(k)}$ and making benefit of simple updates of that vector described in (1.8), only one scalar product of the form $(G^{(k)}e_i \mid G^{(k)}e_j)$ is needed to compute the parameters of C_k. In addition, the matrices C_k need not be accumulated since $G_S^{(k)} = G^{(k)}\Delta_k^{-1}$, $\Delta_k = \mathrm{diag}(\|G^{(k)}e_1\|, \ldots, \|G^{(k)}e_n\|)$ approaches the set of eigenvector matrices of H and diagonal elements of Δ_k^2 approach the eigenvalues of H.

To see that, we have to assume a globally convergent pivot strategy, so that $(G^{(k)})^TG^{(k)} \mapsto \Lambda^2$ as k increases, where Λ is diagonal with positive diagonal elements. Next, note that the sequence $(G_S^{(k)})_k$ is bounded. So, let $(G_S^{(i_k)})_k$ be any convergent subsequence and let $\lim_k G_S^{(i_k)} = Q$. Since $Q^TQ = I_n + N$, where N has zero diagonal, we have

$$I_n + N \;\; = \;\; Q^TQ = \lim_k \Delta_{i_k}^{-1}(G^{(i_k)})^TG^{(i_k)}\Delta_{i_k}^{-1} = \Lambda^2 \lim_k \Delta_{i_k}^{-2}\,,$$

implying $\lim_k \Delta_{i_k} = \Lambda$ and orthogonality of Q. Since $G^{(i_k)} = G_S^{(i_k)}\Delta_{i_k}$, $(G^{(i_k)})_k$ is convergent. Because G is invertible, the product $C^{(i_k)} = C_0 \cdots C_{i_k}$ is also convergent. Let $\lim_{k\to\infty} C^{(i_k)} = C$. Since J-orthogonal matrices make a closed group in the group of regular matrices, C is J-orthogonal. In particular, this means that $CJC^T = J$. We have obtained $Q = GC\Lambda^{-1}$, so we can write

$$\begin{aligned} HQ &= (GJG^T)(GC\Lambda^{-1}) = G(CJC^T)G^TGC\Lambda^{-1} = GCJ\Lambda^2\Lambda^{-1} \\ &= GCJ\Lambda = GC\Lambda^{-1}\Lambda J\Lambda = QJ\Lambda^2\,. \end{aligned}$$

This proves that Q is an eigenvector matrix of H. Since we have the same conclusion for any convergent subsequence of $G_S^{(k)}$ we have proved that $G_S^{(k)}$ approaches the set of eigenvector matrices of H.

Finally, let us mention that the one-sided J-symmetric method is of interest on its own, not just as the part of the compound algorithm. For example, this method can be used to solve the downdating problem, which consists of finding the eigenvalue decomposition of the difference of outer products,

$$H = AA^T - BB^T.$$

The matrix H can be written in the product form

$$H = GJG^T, \quad G = [\ A \quad B\ ], \quad J = \mathrm{diag}(I, -I),$$

and the latter problem can be solved by the the one-sided J-symmetric method. For more details on this problem see [58, 43].

2. Asymptotic Convergence

It is well-known that cyclic Jacobi-type methods converge, under standard conditions, asymptotically quadratically. This means that for large enough k,

$$\|\Omega(H^{(k+N)}\|_F \leq c\,\|\Omega(H^{(k)})\|_F^2\,, \tag{2.1}$$

where $N = n(n-1)/2$. The constant c does not depend on k. By standard conditions we mean either simple eigenvalues or multiple eigenvalues plus ordering of diagonal elements so that those ones converging to the same eigenvalue take successive positions on diagonal plus special (e.g. serial) pivot strategies. By an argument of Wilkinson [56], classical Jacobi method and all kinds of optimal Jacobi methods for symmetric/Hermitian matrices will converge quadratically or better. Since in praxis one uses serial strategies, the relation (2.1) is modified to

$$\|\Omega(H^{((k+1)N)}\|_F \leq c\,\|\Omega(H^{(kN)})\|_F^2\,, \quad k \geq k_0, \tag{2.2}$$

where k_0 is sufficiently large. The relation (2.2) shows that ultimately (in praxis after several cycles) the measure $\|\Omega(H)\|_F$ reduces quadratically per cycle.

The measure $\|\Omega(H^{(k)})\|_F$ appears naturally in asymptotic convergence results, since $\|\Omega(H^{(k)})\|_F = 0$ shows $H^{(k)}$ is diagonal. In addition, $\|\Omega(H^{(k)})\|_F$ is the upper bound for $\max_t |a_{tt}^{(k)} - \lambda_t|$ and therefore measures the absolute distance between diagonal elements and the affiliated eigenvalues.

Note however, that one of the most important properties of Jacobi methods is their relative accuracy. Hence, one would better look for a measure which bounds $\max_t |a_{tt}^{(k)} - \lambda_t|/|\lambda_t|$ or $\max_t |a_{tt}^{(k)} - \lambda_t|/|a_{tt}^{(k)}|$. That measure (see [22]) has form $\|\Omega(DHD)\|_F$, where $D = [\,|\operatorname{diag}(D)|\,]^{-1/2}$ and it also appears in the accuracy estimates from [6]. This has given rise to recent research of the quadratic convergence of scaled iterates by Jacobi methods. The new results will be briefly presented here.

Since one-sided methods can be viewed as intelligent implementations of two-sided methods, their asymptotic convergence amounts to that of two-sided methods.

2.1 Hermitian Jacobi Method

Let us first show using ε notation, the cubic convergence of the row-cyclic Hermitian Jacobi method, when $n = 4$. Suppose the considered cycle is not the first one (hence $h_{34} = 0$). Suppose the eigenvalues are distinct and all off-diagonal elements are $O(\epsilon)$. We assume that ϵ is so small that all angles and their sines have the same order of magnitude as the corresponding pivot elements. Then, we have

$$
\begin{array}{ccccc}
 & \downarrow & \downarrow & & \\
\rightarrow & \bullet & \varepsilon & \varepsilon & \varepsilon \\
\rightarrow & \varepsilon & \bullet & \varepsilon & \varepsilon \\
 & \varepsilon & \varepsilon & \bullet & 0 \\
 & \varepsilon & \varepsilon & 0 & \bullet
\end{array}
\mapsto
\begin{array}{ccccc}
 & \downarrow & & \downarrow & \\
\rightarrow & \bullet & \circ & \varepsilon & \varepsilon \\
 & \circ & \bullet & \varepsilon & \varepsilon \\
\rightarrow & \varepsilon & \varepsilon & \bullet & 0 \\
 & \varepsilon & \varepsilon & 0 & \bullet
\end{array}
\mapsto
\begin{array}{ccccc}
 & \downarrow & & & \downarrow \\
\rightarrow & \bullet & \varepsilon^2 & \circ & \varepsilon \\
 & \varepsilon^2 & \bullet & \varepsilon & \varepsilon \\
 & \circ & \varepsilon & \bullet & \varepsilon^2 \\
\rightarrow & \varepsilon & \varepsilon & \varepsilon^2 & \bullet
\end{array}
\mapsto
$$

$$
\begin{array}{ccccc}
 & & \downarrow & \downarrow & \\
 & \bullet & \varepsilon^2 & \varepsilon^3 & \circ \\
\rightarrow & \varepsilon^2 & \bullet & \varepsilon & \varepsilon \\
\rightarrow & \varepsilon^3 & \varepsilon & \bullet & \varepsilon^2 \\
 & \circ & \varepsilon & \varepsilon^2 & \bullet
\end{array}
\begin{array}{ccccc}
 & & \downarrow & & \downarrow \\
 & \bullet & \varepsilon^2 & \varepsilon^3 & \circ \\
\rightarrow & \varepsilon^2 & \bullet & \circ & \varepsilon \\
 & \varepsilon^3 & \circ & \bullet & \varepsilon^2 \\
\rightarrow & \circ & \varepsilon & \varepsilon^2 & \bullet
\end{array}
\begin{array}{ccccc}
 & & & \downarrow & \downarrow \\
 & \bullet & \varepsilon^2 & \varepsilon^3 & \varepsilon^3 \\
 & \varepsilon^2 & \bullet & \varepsilon^3 & \circ \\
\rightarrow & \varepsilon^3 & \varepsilon^3 & \bullet & \varepsilon^2 \\
\rightarrow & \varepsilon^3 & \circ & \varepsilon^2 & \bullet
\end{array}
\mapsto
$$

$$
\begin{array}{ccccc}
 & \downarrow & \downarrow & & \\
\rightarrow & \bullet & \varepsilon^2 & \varepsilon^3 & \varepsilon^3 \\
\rightarrow & \varepsilon^2 & \bullet & \varepsilon^3 & \varepsilon^5 \\
 & \varepsilon^3 & \varepsilon^3 & \bullet & \circ \\
 & \varepsilon^3 & \varepsilon^5 & \circ & \bullet
\end{array}
\mapsto
\begin{array}{c}
\text{after} \\ \text{the next} \\ \text{cycle}
\end{array}
\begin{array}{ccccc}
 & \downarrow & \downarrow & & \\
\rightarrow & \bullet & \varepsilon^6 & \varepsilon^9 & \varepsilon^{11} \\
\rightarrow & \varepsilon^6 & \bullet & \varepsilon^{11} & \varepsilon^{17} \\
 & \varepsilon^9 & \varepsilon^{11} & \bullet & \circ \\
 & \varepsilon^{11} & \varepsilon^{17} & \circ & \bullet
\end{array}
\mapsto
$$

$$
\begin{array}{c}
\text{after} \\ \text{the next} \\ \text{cycle}
\end{array}
\begin{array}{ccccc}
 & \downarrow & \downarrow & & \\
\rightarrow & \bullet & \varepsilon^{20} & \varepsilon^{31} & \varepsilon^{37} \\
\rightarrow & \varepsilon^{20} & \bullet & \varepsilon^{37} & \varepsilon^{57} \\
 & \varepsilon^{31} & \varepsilon^{37} & \bullet & \circ \\
 & \varepsilon^{37} & \varepsilon^{57} & \circ & \bullet
\end{array}
\qquad
\begin{array}{cl}
\downarrow & \text{Pivot column} \\
\rightarrow & \text{Pivot row} \\
0 & \text{Initial zeros} \\
\circ & \text{Produced zeros} \\
\varepsilon^r & O(\epsilon^r)\ \text{element.}
\end{array}
$$

If the initial matrix has multiple eigenvalues, the above analysis is not correct since almost diagonal matrices with multiple eigenvalues have special structure which influences the asymptotic rate of convergence.

Let

$$\lambda_1 = \cdots = \lambda_{s_1}\ ,\ \lambda_{s_1+1} = \cdots = \lambda_{s_2}\ ,\ldots,\ \lambda_{s_{p-1}+1} = \cdots = \lambda_{s_p} \tag{2.3}$$

be the eigenvalues of Hermitian H and let

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

be the absolute gaps in the spectrum of H. Suppose that H satisfies

$$\|\Omega(H)\|_F \ \leq\ \frac{\delta}{3}\,,$$

and the diagonal elements of H affiliated with the same multiple eigenvalue occupy successive positions on the diagonal. Then H can be partitioned as (H_{ij}) where each block H_{ii} has order $n_i = s_i - s_{i-1}$, $1 \leq i \leq p$, $s_0 = 0$, $s_p = n$. Here n_i is the algebraic multiplicity of λ_{s_i}. We can assume that $p > 1$, since otherwise $H = \lambda_1 I_n$. Let

$$\begin{aligned} \pi_i(H) &= H_{ii}, \qquad \tau_i(H) = [H_{i1},\ldots,H_{i,i-1},H_{i,i+1},\ldots,H_{ip}] \\ \pi(H) &= \mathrm{diag}(H_{11},\ldots,H_{pp}), \quad \tau(H) = H - \pi(H). \end{aligned}$$

The special structure of an almost diagonal Hermitian matrix (see [56], [26, 27] and [19, 20, 21, 23]) is revealed from the following result from [21]. For $1 \leq i \leq p$ holds

$$\|\Omega(H_{ii})\|_F \leq \|H_{ii}\text{-}\lambda_{s_i} I_{n_i}\|_F \leq \frac{1.32}{\delta_i} \sum_{\substack{j=1 \\ j\neq i}}^{p} \|H_{ij}\|_F^2 \leq 0.66 \frac{\|\Omega(H)\|_F^2}{\delta_i}. \tag{2.4}$$

This relation implies that within diagonal blocks the off-diagonal elements are quadratically small compared to those outside the diagonal blocks. When a Jacobi method is applied to such H the relation (2.4) will hold for all iteration matrices i.e. the partition will be preserved (see [21, Lemma 2.2]). This implies that methods defined by the classical and other optimal strategies will never choose at that stage, pivot elements from diagonal blocks and therefore all angles will ultimately tend to zero.

Suppose, $n = 4$, $p = 2$, $n_1 = 3$, $n_2 = 1$. Using the analysis as before and taking into account (2.4), we obtain

$$\begin{array}{ccccc} & \downarrow & \downarrow & & \\ \rightarrow & \bullet & \varepsilon^2 & \varepsilon^2 & \varepsilon \\ \rightarrow & \varepsilon^2 & \bullet & \varepsilon^2 & \varepsilon \\ & \varepsilon^2 & \varepsilon^2 & \bullet & 0 \\ & \varepsilon & \varepsilon & 0 & \bullet \end{array} \quad \begin{array}{ccccc} & \downarrow & & \downarrow & \\ \rightarrow & \bullet & \circ & \varepsilon^2 & \varepsilon \\ & \circ & \bullet & \varepsilon^2 & \varepsilon \\ \rightarrow & \varepsilon^2 & \varepsilon^2 & \bullet & 0 \\ & \varepsilon & \varepsilon & 0 & \bullet \end{array} \quad \begin{array}{ccccc} & \downarrow & & & \downarrow \\ \rightarrow & \bullet & \varepsilon^2 & \circ & \varepsilon \\ & \varepsilon^2 & \bullet & \varepsilon^2 & \varepsilon \\ & \circ & \varepsilon^2 & \bullet & \varepsilon \\ \rightarrow & \varepsilon & \varepsilon & \varepsilon & \bullet \end{array} \ \mapsto$$

$$
\begin{array}{ccccc}
 & & \downarrow & \downarrow & \\
 & \bullet & \varepsilon^2 & \varepsilon^2 & \circ \\
\rightarrow & \varepsilon^2 & \bullet & \varepsilon^2 & \varepsilon \\
\rightarrow & \varepsilon^2 & \varepsilon^2 & \bullet & \varepsilon \\
 & \circ & \varepsilon & \varepsilon & \bullet
\end{array}
\qquad
\begin{array}{ccccc}
 & & \downarrow & & \downarrow \\
 & \bullet & \varepsilon^2 & \varepsilon^2 & \circ \\
\rightarrow & \varepsilon^2 & \bullet & \circ & \varepsilon \\
 & \varepsilon^2 & \circ & \bullet & \varepsilon \\
\rightarrow & \circ & \varepsilon & \varepsilon & \bullet
\end{array}
\qquad
\begin{array}{ccccc}
 & & & \downarrow & \downarrow \\
 & \bullet & \varepsilon^6 & \varepsilon^6 & \varepsilon^3 \\
 & \varepsilon^6 & \bullet & \varepsilon^6 & \circ \\
\rightarrow & \varepsilon^6 & \varepsilon^6 & \bullet & \varepsilon \\
\rightarrow & \varepsilon^3 & \circ & \varepsilon & \bullet
\end{array}
\mapsto
$$

$$
\begin{array}{ccccc}
 & & & \downarrow & \downarrow \\
 & \bullet & \varepsilon^6 & \varepsilon^6 & \varepsilon^3 \\
 & \varepsilon^6 & \bullet & \varepsilon^6 & \varepsilon^7 \\
\rightarrow & \varepsilon^6 & \varepsilon^6 & \bullet & \circ \\
\rightarrow & \varepsilon^3 & \varepsilon^7 & \circ & \bullet
\end{array}
$$

showing again the cubic convergence. Note that cubic reduction inside H_{11} has occurred earlier than indicated above, probably after annihilating $(1,4)$ – element.

In general, when $n > 4$, we have only quadratic convergence for the serial Jacobi methods.

Quadratic convergence considerations of symmetric Jacobi methods in the case of simple and double eigenvalues originated from early works of Wilkinson [55] and Schönhage [39] (see also [54] and [24]).

The harder case of multiple eigenvalues was first considered by Van Kempen [26, 27] and Wilkinson [56]. Later, Hari [21] gave the first complete proof of the quadratic convergence of the serial Jacobi methods. The proof in [21] broadens and completes the considerations of van Kempen from [27], who has not taken into account some quantities which influence the bound. By sophisticated estimates it has been verified in [21] that the result stated in [27] indeed holds. The main result from [21] has the form

$$\|\Omega(H)\|_F \le \frac{\delta}{3} \quad \Rightarrow \quad \|\Omega(H^{(N)})\|_F \le \frac{9}{5}\frac{\|\Omega(H)\|_F^2}{\delta}. \qquad (2.5)$$

It has been shown in [20] that in the case of multiple eigenvalues, large angles can be expected during the process, irrespectively of how tiny the off-diagonal elements are. Hence, the condition on diagonal elements which ensures that the relation (2.4) holds cannot be omitted. If eigenvalues form clusters of small width, one can incorporate perturbation theory (see [21]) to prove the quadratic reduction of $\|\Omega(H^{(kN)}))\|_F$ per cycle. As $\|\Omega(H^{(kN)}))\|_F$ approaches the width of clusters, the asymptotic speed slows down, but further shrinking of $\|\Omega(H^{(kN)}))\|_F$ again increases the speed to reach the quadratic rate.

If δ is tiny however, the estimate (2.5) becomes useless. In praxis this sometimes happens when eigenvalues cluster around zero in such a way that it is difficult or impossible to bound the cluster.

Later, in 1990 Mascarenhas [30, 31] showed that using special quasi-cyclic strategies Jacobi method can perform cubic asymptotic convergence per quasi-cycle. Since his quasi-cycles consist of around 1.25 cycles, this corresponds to the asymptotic rate of order $3^{0.8} \approx 2.41$, thus, between quadratic and cubic speed per cycle. Rhee and Hari [37] proved the global and the cubic (per quasi-cycle) convergence of his method.

Recent Results Although recent research of Jacobi methods has been mostly concentrated on their accuracy properties, closely related are new measures of advancing of the processes. These measures bound the maximum relative distance between diagonal elements and the corresponding eigenvalues. Therefore, they are included in the stopping criterions of the methods. They involve *scaled diagonally dominant* (see [2], [22]) matrices and *relative gaps.*

Suppose $A = D + N$, where D is diagonal and N has zero diagonal. Then $A = (a_{ij})$ is referred to as α–*diagonally dominant* with respect to a norm $\|\cdot\|$ if $\|N\| \leq \alpha \min_{1\leq i\leq n} |a_{ii}|$, with $0 \leq \alpha < 1$. If $A = D + N$ with $|a_{ii}| = 1$, $1 \leq i \leq n$ and Δ_1, Δ_2 are arbitrary nonsingular diagonal matrices, then $B = \Delta_1 A \Delta_2$ is α–*scaled diagonally dominant* (α–s.d.d.) with respect to a given norm, provided that A is α–diagonally dominant with respect to that norm. Note that an α–s.d.d. matrix has nonzero diagonal elements. If A is Hermitian, it is presumed that $\Delta_1 = \Delta_2$ and Δ_1 is real. Such scaling, which is a congruence transformation, will be called symmetric.

Relative gaps are in applications often connected with s.d.d. matrices. Several definitions of relative gaps have been used (cf. [40, 46, 25]). Here we use the one from [22].

Let the eigenvalues of the Hermitian matrix H be ordered as in the relation (2.3). The relative gaps and the minimum relative gap in the spectrum of H are defined as follows

$$\gamma_i = \min_{\substack{1\leq j\leq p \\ j\neq i}} \frac{|\lambda_{s_i} - \lambda_{s_j}|}{|\lambda_{s_i}| + |\lambda_{s_j}|} \quad 1 \leq i \leq p; \qquad \gamma = \min_{1\leq i\leq p} \gamma_i\,. \tag{2.6}$$

The following result can be used in selecting a measure for convergence of Hermitian Jacobi methods.

Proposition 2.1 [22] *Let $H = H^*$ be α-s.d.d. and $H = \Delta_H H_S \Delta_H$, where $\Delta_H = [\ |diag(H)|\ ]^{\frac{1}{2}}$. Let γ_i and γ be as in the relation (2.6). If $\alpha < \gamma/(\gamma+3)$, then*

$$\sum_{j=s_{i-1}+1}^{s_i} |1 - \frac{\lambda_{s_i}}{h_{jj}}|^2 + \|\Omega\,(\pi_i(H_S))\,\|_F^2 \leq \frac{16}{\gamma_i^2}\|\tau_i(H_S)\|_F^4, \ 1 \leq i \leq p. \tag{2.7}$$

If $H = H^*$ is α-s.d.d. and positive definite, then the relation (2.7) holds with constant 4 instead of 16 and under less stringent assumption $\alpha < \gamma/3$. ∎

Summing up the equations (2.7) over i, one obtains, say for positive definite matrix H (see [22, Collorary 3.2(ii)]),

$$\sum_{j=1}^{n} \mid 1-\frac{\lambda_j}{h_{jj}} \mid^2 + \|\Omega\left(\pi(H_S)\right)\|_F^2 \leq \frac{2}{\gamma^2}\|\tau(H_S)\|_F^4. \tag{2.8}$$

Note that $|1-\frac{\lambda_{s_i}}{h_{jj}}| \leq \beta$ implies $|1-\frac{h_{jj}}{\lambda_{s_i}}| \leq \frac{\beta}{1-\beta}$. Hence, if $\|\tau(H_S)\|_F^2/\gamma$ is small, we have accurate estimates for both $|(\lambda_{s_i}-h_{jj})/h_{jj}|$ and $|(h_{jj}-\lambda_{s_i})/\lambda_{s_i}|$. The relation (2.7) for positive definite matrices implies that $\log_{10}(2\|\tau_i(H_S)\|_F^2/\gamma_i)$ indicates how many significant decimal digits are correct in the appropriate diagonal elements as approximations of λ_{s_i}. As a global measure for all diagonals, one can also use $\sqrt{2}\|\tau(H_S)\|_F^2/\gamma$ or larger measures, $\|\tau(H_S)\|_F$ and $\|\Omega(H_S)\|_F$.

Returning to Jacobi methods, note that γ does not change with iteration matrix. Usually, in the first few cycles the partition $(H_{ij}^{(k)})$ is not yet recognized and one prefers to use $\|\Omega(H_S)\|_F$ a simpler although (in rare situations) larger measure than $\|\tau(H_S)\|_F$.

Suppose a serial Jacobi method is applied to a positive definite H producing the sequence $(H^{(k)})_k$. By the result [6] we know that the method computes the eigenvalues and eigenvectors almost with the accuracy that is warranted by the eigenproblem. After several cycles the diagonal elements become approximations of the eigenvalues and lower bounds of absolute or relative gaps can easily be computed from diagonal elements (see Example 3.3 from [22]). If the eigenvalues cluster around zero, some absolute gaps δ_i and therefore δ will be tiny and the result (2.5) becomes useless for practical purposes. The measure $\|\Omega(H^{(k)}\|_F$ alone bounds $\max_t |h_{tt}^{(k)} - \lambda_t|$. Hence, for tiny λ_{s_i}, $\max_{s_{i-1}+1\leq t\leq s_i} |h_{tt}^{(k)} - \lambda_{s_i}|/\lambda_{s_i}$ is bounded by $\|\Omega(H^{(k)}\|_F/\lambda_{s_i}$ which can be large. Thus, the measures involving $\|\Omega(H^{(k)})\|_F$ and absolute gaps will not be appropriate.

On the contrary, for such matrices, all relative gaps γ_i and γ can be large (close to one). As the process advances one can use measures $\|\Omega(H_S)\|_F$ and $\|\Omega(H_S)\|_F^2/\gamma$. This implies that after each cycle one has to compute $\|\Omega(H_S)\|_F$ and perhaps a lower bound of γ. For two-sided Jacobi methods this is appropriate since it requires only $O(n^2)$ flops. However, in the context of one-sided methods, computing $\|\Omega(H_S)\|_F$ requires $O(n^3)$ flops. Hence, during the whole process, one would not like to compute $\|\Omega(H_S)\|_F$ more than once. In such a situation it is necessary to exploit the knowledge that serial Jacobi methods converge asymptotically quadratically. As we have just explained, the quadratic convergence result (2.5) is useless. In addition, in described situation, the relation between $\|\Omega(H_S)\|_F$ and $\|\Omega(H)\|_F$ can be only very roughly estimated, so (2.5) would give even poorer result when translated in terms of $\|\Omega(H_S)\|_F$.

We need a quadratic convergence result for $\|\Omega(H_S)\|_F$ which involves relative instead of absolute gaps. Fortunately, such results can be formulated and derived in the context of two-sided methods and then directly applied to one-sided methods.

In his Ph. D. thesis [32] Matejaš solved this problem for the serial Jacobi methods for positive definite matrices. The easier case of distinct eigenvalues (see [33]) and real matrices assumes form

$$\alpha_N \leq 0.715 \frac{\alpha_0^2}{\gamma} \qquad \text{whenever} \qquad \alpha_0 \leq \frac{1}{4} \min\{\frac{1}{n},\ \gamma\}\ , \tag{2.9}$$

where $\alpha_k = \|A^{(k)}\|_F$ as defined by the relations (1.5) – (1.6). In (2.9) it is assumed that the diagonal elements are decreasingly ordered. The harder case of multiple eigenvalues and complex matrices takes similar form (see [34])

$$\alpha_N \leq \sqrt{\frac{5}{2}} \frac{\alpha_0^2}{\gamma} \qquad \text{whenever} \qquad \alpha_0 \leq \frac{1}{6} \min\{\frac{1}{n},\ \gamma\}\ .$$

A similar result holds for the case of indefinite Hermitian matrix H. In the context of one-sided methods, these results can be used for predicting the number of cycles till convergence. If eigenvectors are wanted one can use appropriate eigenvector perturbation estimates (e.g. [6, Theorem 2.7]) which use condition of scaled matrix (which is in our case close to one) and appropriate relative gaps.

We end this subsection by a brief discussion on stopping criterions. Jacobi methods are predominantly used when accuracy of the output data is important. Therefore, one will probably choose the criterion $\|\Omega(H_S^{(k)}\|_F \leq \text{tol}$, where tolerance tol is chosen by the user. For one-sided methods, where N scalar products are needed to obtain $\|\Omega(H_S^{(k)})\|_F$, tol can be set $f(n)\epsilon$ where f is slowly increasing function of n and ϵ is machine epsilon. For two-sided methods however, one can use a nice stopping criterion of Rutishauser (see [57]) which is almost equivalent to $\|\Omega(H_S^{(k)})\|_F \leq \epsilon$.

2.2 J-Symmetric Jacobi Method

J-Symmetric Jacobi method has similar asymptotic properties. Quadratic convergence is again defined by the relation (2.1) or (2.2). The proof assumes that the diagonal elements of JH approximating the same eigenvalue of (H, J) take successive positions on the diagonal of JH.

Absolute gaps δ_i and δ are defined as above. Drmač and Hari [10] have shown that

$$\|\Omega(H^{(N)})\|_F \leq 3\frac{\|\Omega(H)\|_F^2}{\delta} \quad \text{provided that} \quad \|\Omega(H)\|_F < \frac{\delta}{3m(n-m)},$$

where m is number of positive ones in J. Here the asymptotic assumption is stronger than for the symmetric Jacobi method. In estimating nonorthogonal (hyperbolic) transformations one uses mathematical induction. The number $m(n-m)$ in the denominator is used to compensate a possible gradual increase of $\|\Omega(H^{(k)})\|_F$ during the considered cycle.

J-symmetric method is very well suited to work with the accelerated strategy of Mascarenhas, which yields cubic rate of convergence per quasi-cycle. For the proof, one could combine ideas from [10] and [37].

Since the J-symmetric quadratic convergence result suffers the same shortcomings as its symmetric counterpart (2.5), one is challenged to find a similar remedy. First, one needs a bound for the relative distance between diagonal elements of JH and the eigenvalues of (H, J). This is done for the case of positive definite H. In the following result, nonincreasing ordering of the eigenvalues of the pair (H, J) is assumed. Relative gaps γ_i, $1 \leq i \leq p$ and γ are defined as in (2.6).

Proposition 2.2 [35, Theorem 1(ii)] *Let $H = H^*$ be α-s.d.d. positive definite. Let $H = \Delta H_S \Delta$, where $\Delta = [\ diag(H)\ ]^{\frac{1}{2}}$. Let $\alpha < \gamma/3$ and let the diagonal elements of JH affiliated with the same eigenvalue of (H, J) occupy successive positions on the diagonal. Then for the same partition of $A = H_S - I_n$ and Δ holds*

$$\|A_{ii} - |\lambda_{s_i}|\Delta_{ii}^{-2}\|_F \leq \frac{2}{\gamma_i}\|\tau_i(A)\|_F^2 = \frac{2}{\gamma_i}\sum_{\substack{j=1 \\ j\neq i}}^{p} \|A_{ij}\|_F^2, \qquad 1 \leq i \leq p.$$

■

As earlier, one can deduce

$$\sum_{r=1}^{n} \left|1 - \frac{|\lambda_r|}{h_{rr}}\right|^2 + \|\pi(A)\|_F^2 \leq \frac{2}{\gamma^2}\|\tau(A)\|_F^4 .$$

Thus, scaled diagonally dominant pair (H, J) has the same structure as the pair (H, I_n) which has been discussed earlier. Hence the rotation angles can be large if pivot pair happens to be inside a diagonal block. On the contrary, hyperbolic angles ultimately tend to zero as k increases. This follows from the fact that $m = n_1 + \cdots + n_{r_o}$ for some $2 \leq r_o \leq p-1$ and that is a consequence of the assumption that (H, J) is positive definite. We see that the measure $\|\tau(A)\|_F$ or $\|\Omega(A)\|_F$ should be used for stopping of the process.

The new quadratic convergence result from [35] has form

$$\alpha_N \leq 3.5\frac{\alpha_0^2}{\gamma} \qquad \text{whenever} \qquad \alpha_0 \leq \frac{1}{6}\min\{\frac{1}{n},\ \gamma\}\,.$$

Here $\alpha_k = \|\Omega(A^{(k)})\|_F$, $k \geq 0$ and the result assumes that $\lambda_1 \geq \cdots \geq \lambda_m$, $|\lambda_{m+1}| \geq \cdots \geq |\lambda_n|$, which requires block-permutational similarity of the partition (2.3) and renumbering of the relative gaps.

3. Accuracy

As noticed by Rosanoff et al [38], and theoretically explained by Demmel and Veselić [6], the Jacobi algorithm is more accurate than any algorithm that starts with tridiagonalization of the symmetric matrix (bidiagonalization in the case of SVD computation). In this section we explain this important fact, using the results of Demmel, Veselić, Hari and Drmač. As we will see, if the objective is to compute all eigenvalues with small relative error, the definite and the indefinite case must be treated differently. We first analyze the

3.1 Symmetric Definite Case

Numerical analysis of thie two–stage diagonalization procedure is simple but with far reaching consequences. We start with the analysis of the Cholesky factorization.

In floating point computation, the computed approximation of L is $\tilde{L}$ and we need an estimate for the backward error $\tilde{L}\tilde{L}^T - H$. The following proposition is due to Demmel [7].

Proposition 3.1 *Suppose the Cholesky factorization algorithm has successfully completed all steps in floating point arithmetic with unit roundoff ϵ. If $\tilde{L}$ is the computed lover triangular matrix, then $\tilde{L}\tilde{L}^T = H + \delta H$, where δH is symmetric matrix such that for all $1 \leq i, j \leq n$*

$$|\delta H_{ij}| \leq \eta_C\sqrt{H_{ii}H_{jj}}, \quad \eta_C = \frac{c(n)\epsilon}{1-2c(n)\epsilon}, \ c(n) = \max\{3, n\}. \quad (3.1)$$

■

Note that the backward error δH is bounded entry–wise, rather than norm–wise.

Now, we apply the one-sided Jacobi algorithm to the matrix $\tilde{L}$, that is, we implicitly run the symmetric Jacobi on the matrix $\tilde{L}^T\tilde{L}$. As has been shown in subsection 1.2, this iteration process has form $G^{(k+1)} =$

$G^{(k)}V^{(k)}$, $k \geq 0$, where $V^{(k)}$ is plane Jacobi rotation and $G^{(0)} = \tilde{L}$. The process terminates at index ℓ where the normalized columns of $G^{(\ell)}$ are orthogonal up to a tolerance $O(n\epsilon)$. The following proposition from [9] describes the numerical behaviour of the right–handed Jacobi SVD algorithm.

Proposition 3.2 *Let the cyclic one-sided Jacobi algorithm be applied to $\tilde{L}$ in floating point arithmetic with roundoff ϵ. Let each cycle comprise p parallel steps and let the stopping criterion be satisfied after s cycles. Let $\tilde{G}^{(k)}$, $k = 0, \ldots, \ell = p \cdot s$, $\tilde{G}^{(0)} = \tilde{L}$ be the generated matrices. Then there exists an orthogonal matrix $\hat{V}$ and a backward error $\delta\tilde{L}$ such that $\tilde{G}^{(\ell)} = (\tilde{L} + \delta\tilde{L})\hat{V}$ and*

$$\|e_i^T \delta\tilde{L}\|_2 \leq \eta_J \|e_i^T \tilde{L}\|_2, \;\; 1 \leq i \leq n, \;\; \eta_J \leq (1 + 6\epsilon)^\ell - 1.$$

Further, due to the stopping criterion, the columns of $\tilde{G}^{(\ell)}$ are numerically orthogonal, that is

$$\max_{i,j} \cos \angle(\tilde{G}^{(\ell)} e_i \, , \, \tilde{G}^{(\ell)} e_j) \leq O(n\epsilon). \qquad \blacksquare$$

It is important to note that the error analysis is done row–wise, while the convergence is defined column–wise.

¿From Proposition 3.2 it follows that $\tilde{G}^{(\ell)}$ can be written as $\tilde{G}^{(\ell)} = \tilde{U}\tilde{\Sigma}$, where $\tilde{\Sigma}$ is diagonal with column norms of $\tilde{G}^{(\ell)}$ along its diagonal, and $\tilde{U}$ is numerically orthogonal, $|\tilde{U}^T\tilde{U} - I_n|_{ij} \leq O(n\epsilon)$.

Combining Propositions 3.1 and 3.2, we get

$$\begin{aligned} \tilde{U}\tilde{\Sigma}^2\tilde{U}^T &= (\tilde{L} + \delta\tilde{L})(\tilde{L} + \delta\tilde{L})^T = \tilde{L}\tilde{L}^T + \tilde{L}\delta\tilde{L}^T + \delta\tilde{L}\tilde{L}^T + \delta\tilde{L}\delta\tilde{L}^T \\ &= H + \delta H + E, \;\; E = \tilde{L}\delta\tilde{L}^T + \delta\tilde{L}\tilde{L}^T + \delta\tilde{L}\delta\tilde{L}^T. \end{aligned}$$

The perturbation matrix $\Delta H = \delta H + E$ is symmetric and it holds

$$\max_{i,j} \frac{|\Delta H_{ij}|}{\sqrt{H_{ii}H_{jj}}} \leq \eta \equiv \eta_C + 2\eta_J + O(\eta_J^2).$$

If we set $\tilde{\Lambda} = \tilde{\Sigma}^2$, then we have $H + \Delta H = \tilde{U}\tilde{\Lambda}\tilde{U}^T$. This means that *this variant of the Jacobi diagonalization method computes the eigenvalues and eigenvectors with* <u>*entry–wise small backward error*</u>*. The methods that first tridiagonalize the matrix do not share this important property.*

Let us estimate the forward error in the computed approximations $\tilde{\Lambda}_{ii}$ of the eigenvalues λ_i of H. Let $\tilde{\lambda}_1 \geq \cdots \geq \tilde{\lambda}_n$ be the eigenvalues of $H + \Delta H$ and let the eigenvalues λ_i of H as well as the diagonals of $\tilde{\Lambda}$ be nonincreasingly ordered. First we estimate the maximum relative distance between λ_i and $\tilde{\lambda}_i$.

We can presume that η is small enough so that $\|L^{-1}\Delta H L^{-T}\|_2 < 1$ holds, implying that the positive square root of $I_n + L^{-1}\Delta H L^{-T}$ is well defined. Since the matrix

$$H + \Delta H = L\sqrt{I + L^{-1}\Delta H L^{-T}}\sqrt{I + L^{-1}\Delta H L^{-T}}L^T$$

is similar to

$$\sqrt{I + L^{-1}\Delta H L^{-T}}L^T L\sqrt{I + L^{-1}\Delta H L^{-T}}\,,$$

and $L^T L$ is similar to $H = LL^T$, an application of Ostrowsky's theorem yields $\tilde{\lambda}_i = \lambda_i(1+\theta_i), \quad |\theta_i| \leq \|L^{-1}\Delta H L^{-T}\|_2,\ 1 \leq i \leq n$. The key observation in this estimate is as follows. If we set $D = [\mathrm{diag}(H)]^{1/2}$, then $L^{-1}\Delta H L^{-T} = L^{-1}D(D^{-1}\Delta H D^{-1})DL^{-T}$ and

$$\begin{aligned} \|L^{-1}\Delta H L^{-T}\|_2 &\leq \|L^{-1}D\|_2^2\|D^{-1}\Delta H D^{-1}\|_2 \\ &= \|(D^{-1}HD^{-1})^{-1}\|_2\|D^{-1}\Delta H D^{-1}\|_2 \\ &\leq n\eta\|(D^{-1}HD^{-1})^{-1}\|_2. \end{aligned}$$

The matrix $H_S = D^{-1}HD^{-1}$ has unit diagonal, off-diagonals less than one in modulus, and by the well–known result of van der Sluis [50] it holds

$$\|H_S^{-1}\|_2 \leq \kappa_2(H_S) \leq n \min_{S=\mathrm{diag},\det(S)\neq 0} \kappa_2(SHS) \leq n\kappa_2(H).$$

Here $\kappa_2(X) = \|X\|_2\|X^{-1}\|_2$ is the spectral condition number. Therefore, we can write

$$\tilde{\lambda}_i = \lambda_i(1+\theta_i), \quad |\theta_i| \leq n\eta\|H_S^{-1}\|_2, \quad 1 \leq i \leq n. \tag{3.2}$$

Next, we estimate the relative distance between $\tilde{\lambda}_i$ and $\tilde{\Lambda}_{ii}$. The stopping criterion ensures that $\tilde{U}^T\tilde{U} = I + X$ with $\max_{i,j}|x_{ij}| \leq O(n\epsilon)$, where $X = (x_{ij})$. An easy calculus shows that there exists an orthogonal matrix $\hat{U}$ such that $\tilde{U} = (I+Y)^{1/2}\hat{U}$, where Y is symmetric and $\|Y\|_2 = \|X\|_2 \leq O(n^2\epsilon)$. Hence,

$$H + \Delta H = \tilde{U}\tilde{\Lambda}\tilde{U}^T = (I+Y)^{1/2}\hat{U}\tilde{\Lambda}\hat{U}^T(I+Y)^{1/2},$$

and we can again apply the Ostrowsky's theorem to obtain $\tilde{\lambda}_i = \tilde{\Lambda}_{ii}(1+\theta_i')$, $|\theta_i'| = O(n^2\epsilon))$. This together with (3.2) implies

$$\tilde{\Lambda}_{ii} = \lambda_i\frac{1+\theta_i}{1+\theta_i'} = \lambda_i(1+\eta_i)\,, \quad |\eta_i| \leq O(n^2\epsilon)\|H_S^{-1}\|_2, \quad 1 \leq i \leq n\,.$$

This bound is nearly the best one can hope for in computing with floating point positive definite matrices. For, Demmel [7] has shown that relative entry–wise perturbations of size $1/\|H_S^{-1}\|_2$ can make H exactly singular, and that H can be considered numerically positive definite if $\|H_S^{-1}\|_2 < 1/(n\eta_C)$. Moreover, Veselić and Slapničar [53] have shown that the spectrum of positive definite H is stable under entry–wise floating point perturbations if and only if $\|H_S^{-1}\|_2$ is moderate.

We can conclude that *the forward error in the computed eigenvalues depends on* $\kappa_2(H_S)$*, and not on* $\kappa_2(H)$*, as is in the case of methods based on tridiagonalization. Since* $\kappa_2(H_S) \leq n\kappa_2(H)$*, and since it is possible* $\kappa_2(H_S) \ll \kappa_2(H)$*, the Jacobi method is numerically clearly superior to any method that first tridiagonalizes the matrix. This method is method of choice for computing the eigenvalues of positive definite matrices with high relative accuracy.*

The diagonalization procedure just described (Cholesky factorization followed by the right–handed Jacobi on L) implicitly diagonalizes $L^T L$. This is more than just a nice observation. It actually means that this Jacobi method is preconditioned using one implicit (cost free) step of the Rutishauser LR method. Because of diagonalizing effect of preconditioning, the Jacobi method converges faster, especially if the Cholesky factorization is computed with pivoting. For more discussion see [52] and [14].

3.2 Indefinite Case

Here we provide elements of perturbation analysis and error bounds for the J-symmetric Jacobi method and one-sided J-symmetric compound Jacobi method described in Subsection 1.3.

Let us first consider the J-symmetric eigenvalue problem $Hx = \lambda Jx$ with positive definite H,

$$Q^T H Q = \Lambda, \quad Q^T J Q = J, \quad \Lambda = \mathrm{diag}(\lambda_i).$$

Let $H = DAD$, where $D = \mathrm{diag}(d_{11}, \ldots, d_{nn})$ is diagonal and A is positive definite with unit diagonal. Further, let $\delta H = D\delta A D$. By the result of Veselić and Slapničar [53, 46], if $|\delta H_{ij}| \leq \eta d_{ii} d_{jj}$, and if $\tilde{\eta} \equiv \|\delta A\|_2 \, \|A^{-1}\|_2 \leq n\eta \|A^{-1}\|_2 < 1$, then

$$1 - \tilde{\eta} \leq \frac{\tilde{\lambda}_i}{\lambda_i} \leq 1 + \tilde{\eta}. \tag{3.3}$$

The perturbation result for the invariant subspaces is given by Slapničar and Truhar in [45]. Let us partition the eigenvalue problem as

$$\begin{bmatrix} Q_1^T \\ Q_2^T \end{bmatrix} H \, [Q_1 \; Q_2] = \begin{bmatrix} \Lambda_1 & \\ & \Lambda_2 \end{bmatrix}, \quad \begin{bmatrix} Q_1^T \\ Q_2^T \end{bmatrix} J \, [Q_1 \; Q_2] = \begin{bmatrix} J_1 & \\ & J_2 \end{bmatrix},$$

and let the perturbed problem be partitioned accordingly. Let $\tilde{X}_1$ and X_2 be orthogonal bases for subspaces spanned by the columns of $\tilde{Q}_1$ and JQ_2J_2, respectively. Let $U\Sigma V^*$ be a singular value decomposition of the matrix $X_2^T\tilde{X}_1$. The diagonal entries of the matrix $\sin\Theta(Q_1,\tilde{Q}_1)\equiv\Sigma$ are the sines of canonical angles between subspaces spanned by the columns of Q_1 and $\tilde{Q}_1$ (see [48]). The relative gap is in this case defined by

$$\gamma(\tilde{\Lambda}_1,\Lambda_2)=\min_{i,j}\frac{|\,[\tilde{\Lambda}_1]_{ii}[J_1]_{ii}-[\Lambda_2]_{jj}[J_2]_{jj}\,|}{\sqrt{[\tilde{\Lambda}_1]_{ii}\cdot[\Lambda_2]_{jj}}}. \tag{3.4}$$

Then, if $\|\delta A\|_2\,\|A^{-1}\|_2<1$, we have

$$\|\sin\Theta(Q_1,\tilde{Q}_1)\|_F\leq\|Q\|_2^2\left(\frac{1}{2}\psi+\sqrt{1+\frac{1}{4}\psi^2}\right)\frac{\psi}{\gamma(\tilde{\Lambda}_1,\Lambda_2)}, \tag{3.5}$$

where

$$\psi=\frac{\|\delta A\|_2\,\|A^{-1}\|_F}{\sqrt{1-\|\delta A\|_2\,\|A^{-1}\|_2}}.$$

Clearly, when $J=I$, then Q is orthogonal, and (3.5) is a subspace version of the corresponding eigenvector bounds from [6]. However, when $J\neq I$, then $\|Q\|_2^2\equiv\kappa_2(Q)$ is the spectral condition number. It is a remarkable fact that $\kappa_2(Q)$ is bounded by the condition number of A, the same quantity that governs the accuracy of the computation. In [47], Slapničar and Veselić proved that

$$\kappa(Q)\leq\sqrt{\min_{\Delta J=J\Delta}\kappa_2(\Delta^TA\Delta)}\leq\sqrt{\kappa_2(A)}.$$

We now describe results of numerical analysis. Error analysis of a single hyperbolic rotation is technically more complicated because such transformations are nonorthogonal and possibly of large norm. One step of the method, $H^{(k+1)}=C_k^TH^{(k)}C_k$, in floating point computation is of the form

$$\tilde{H}^{(k+1)}=\tilde{C}_k^T(\tilde{H}^{(k)}+\delta\tilde{H}^{(k)})\tilde{C}_k$$

where $\tilde{C}_k$ is hyperbolic rotation, and the backward perturbation $\delta\tilde{H}^{(k)}$ is bounded as follows: If $\tilde{H}^{(k)}=D_kA_kD_k$, where D_k is diagonal and A_k has unit diagonal, then we can write $\delta\tilde{H}^{(k)}$ as $\delta\tilde{H}^{(k)}=D_k\delta A_kD_k$, where $\|\delta A_k\|_2\leq\alpha_k\varepsilon$. Here $\alpha_k=O(\sqrt{\kappa_2(A_k)}\,\sqrt{n})$. Thus, according to (3.3), the perturbation of the eigenvalues due to single floating point hyperbolic rotation is determined by the value

$$\|\delta A_k\|_2\,\|A_k^{-1}\|_2\leq\alpha_k\varepsilon\|A_k^{-1}\|_2.$$

Accumulating the effect of a total of r rotations, gives the relative error bound for the computed eigenvalues of the order of

$$\varepsilon \sum_{k=0}^{r} \alpha_k \|A_k^{-1}\|_2 + O(n^2)\varepsilon.$$

The error bound for the computed eigenspaces follows by plugging the above bound for $\|\delta A_k\|_2$ into (3.5) and accumulating the effect of a total of r rotations in a similar manner. For details see [43].

During the process, $\|A_k^{-1}\|_2$ and $\kappa_2(A_k)$ tend to one, starting with $\|A^{-1}\|_2$ and $\kappa_2(A)$, respectively. While the theoretical bound for the values of $\|A_k^{-1}\|_2$ and $\kappa_2(A_k)$, $k = 1, 2, \ldots$ is pessimistic, numerical evidence indicates that these values never grow too much above the initial values. For more discussion see [40, 43, 41, 44].

Next, we consider the symmetric indefinite eigenvalue problem $Hx = \lambda x$,

$$Q^T H Q = \Lambda, \quad Q^T Q = I, \quad \Lambda = \mathrm{diag}(\lambda_i) \text{ nonsingular},$$

and the appropriate compound method.

The perturbation bound for the eigenvalues is given by Veselić and Slapničar [53, 46]. Let $|H|_s = \sqrt{H^2}$ be the spectral absolute value of H, and let $|H|_s = DAD$, where $D = \mathrm{diag}(d_{11}, \ldots, d_{nn})$ is diagonal and A is positive definite with unit diagonal. Further, let $\delta H = D\delta AD$. If $|\delta H_{ij}| \leq \eta d_{ii} d_{jj}$, and if $\tilde{\eta} \equiv \|\delta A\|_2 \, \|A^{-1}\|_2 \leq n\eta \|A^{-1}\|_2 < 1$, then the perturbation of the eigenvalues is again bounded by (3.3).

The perturbation theory for the invariant subspaces, given by Truhar and Slapničar [49], assumes the following partition of the eigenvalue decomposition

$$\begin{bmatrix} Q_1^T \\ Q_2^T \end{bmatrix} H \begin{bmatrix} Q_1 & Q_2 \end{bmatrix} = \begin{bmatrix} \Lambda_1 & \\ & \Lambda_2 \end{bmatrix}.$$

Let the perturbed problem be partitioned accordingly. Similarly to (3.5), if $\|\delta A\|_2 \, \|A^{-1}\|_2 < 1$, then

$$\|\sin \Theta(Q_1, \tilde{Q}_1)\|_F \leq \|V\|_2^2 \frac{\|\delta A\|_2 \, \|A^{-1}\|_F}{\sqrt{1 - \|\delta A\|_2 \, \|A^{-1}\|_2}} \frac{1}{\gamma(\tilde{\Lambda}_1, \Lambda_2)}. \tag{3.6}$$

Here $\gamma(\tilde{\Lambda}_1, \Lambda_2)$ is again defined as in (3.4), but without J_1 and J_2, and V is the hyperbolic eigenvector matrix of the pair $(G^T G, J)$.

We now describe numerical analysis of the factorization. For the sake of simplicity, we assume that the matrix is already pivoted so that the Bunch–Parlett symmetric indefinite factorization $PHP^T = GJG^T$ runs with $P = I$. Slapničar [42] has shown that in floating point, the computed matrix G satisfies

$$GJG^T = H + \delta H$$

with symmetric backward perturbation δH bounded entry–wise by

$$|\delta H| \leq O(n)\varepsilon(|H| + |G||G|^T).$$

The effect of this backward error to the eigenvalues of H is given by (3.3) with

$$\tilde{\eta} \leq O(n)\varepsilon\|(\hat{D}^{-1}GVV^TG^T\hat{D}^{-1})^{-1}\|_2, \tag{3.7}$$

where $\hat{D}$ is a diagonal scaling matrix and the rows of the matrix $\hat{D}^{-1}G$ are of unit Euclidean norm. The details of this estimation can be found in [40, 43]. Also, by using (3.6) it can be shown that the error in the computed invariant subspaces, which is due to factorization is bounded by (see [40, 43])

$$\|\sin\Theta(Q_1, \tilde{Q}_1)\|_F \leq O(\|V\|_2^2)\frac{\tilde{\eta}}{\gamma(\tilde{\Lambda}_1, \Lambda_2)}. \tag{3.8}$$

The factorization $PHP^T = GJG^T$ has two more remarkable properties worth mentioning (see [42] for details). First, if the computed factor $\tilde{G}$ is lower triangular (meaning that only 1×1 pivots took place), then the factorization is also forward stable,

$$|\tilde{G} - G| \leq 3n|G|\mathrm{tril}(|G^{-1}|(|H| + |G|\,|G|^T)|G^{-1}|^T)\varepsilon + O(\varepsilon^2).$$

Second, let $G = BD$, where D is diagonal scaling and the columns of B are of unit Euclidean norm. Then the condition number of B is bounded by a function of n, irrespective of the condition number of G,

$$\kappa_2(B) \leq 3.781^n\sqrt{15\,n^2 + n}\,.$$

Let us now consider the iterative part of the method. Let $\tilde{H} = \tilde{G}J\tilde{G}^T$ and $\tilde{G} = (B + \delta B)D$. Then the eigenvalue perturbation can be bounded by (see [53])

$$(1-\eta)^2 \leq \frac{\tilde{\lambda}_i}{\lambda_i} \leq (1+\eta)^2, \quad \eta = \|\delta BB^\dagger\|_2. \tag{3.9}$$

One step of the one–sided method, $G^{(k+1)} = G^{(k)}C_k$, in floating point computation is of the form

$$\tilde{G}^{(k+1)} = (\tilde{G}^{(k)} + \delta\tilde{G}^{(k)})\tilde{C}_k$$

where $\tilde{C}_k$ is hyperbolic rotation, and the backward perturbation $\delta\tilde{G}^{(k)}$ is bounded as follows: If $\tilde{G}^{(k)} = B_k D_k$, where D_k is diagonal and B_k has unit columns, then we can write $\delta\tilde{G}^{(k)}$ as $\delta\tilde{G}^{(k)} = \delta B_k D_k$, where $\|\delta B_k\|_2 \leq \beta_k \varepsilon$. Here β_k is moderate constant. Thus, the perturbation of the eigenvalues due to a single floating point rotation is determined by the value of $\|\delta B_k B_k^\dagger\|_2 \leq \beta_k \varepsilon \|B_k^\dagger\|_2$. Accumulating the effect of a total of r rotations, where $\tilde{G}^{(r)}$ has columns orthogonal up to $O(n\varepsilon)$, gives the relative error bound for the computed eigenvalues of the order of

$$\tilde{\beta} = \varepsilon \sum_{k=0}^{r} \beta_k \|B_k^\dagger\|_2 + O(n^2)\varepsilon. \tag{3.10}$$

By combining (3.3), (3.7), (3.9) and (3.10), it follows that the relative error in the eigenvalues $\tilde{\lambda}_i$ computed by the compound method (indefinite factorization followed by the one-sided J-symmetric Jacobi method) is bounded by

$$\frac{|\tilde{\lambda}_i - \lambda_i|}{|\lambda_i|} \leq \tilde{\eta} + \tilde{\beta} + O(\varepsilon^2). \tag{3.11}$$

Error bound for the computed invariant subspaces is obtained by combining (3.8) with the bound for the errors due to iterative part of the algorithm from [43], essentially giving

$$\|\sin\Theta(Q_1, \tilde{Q}_1)\|_F \leq O(\|V\|_2^2) \frac{\tilde{\eta} + \tilde{\beta}}{\gamma(\tilde{\Lambda}_1, \Lambda_2)}. \tag{3.12}$$

Similarly to the two-sided method, during the process, $\|B_k^\dagger\|_2$ tends to one, starting with $\|B^\dagger\|_2$. While the theoretical bound for the values of $\|B_k^\dagger\|_2$, $k = 1, 2, \ldots$ is again pessimistic, numerical evidence indicates that these values never grow too much above the initial value $\|B^\dagger\|_2$. Moreover, it has been observed in [40, 43], that $\|B^\dagger\|_2$ is in general very low, primarily due to the rank revealing property of the factorization $PHP^T = GJG^T$. Consequently, the final errors in (3.11) and (3.12) are mainly due to the factorization part of the algorithm, that is, to $\tilde{\eta}$. For more discussion see [40, 43, 41, 44].

3.3 Computing SVD

Here we discuss relative accuracy issues in one-sided Jacobi method for computing singular value decomposition of a general $m \times n$ matrix A, $m \geq n$. The simplest variant of the method is simply implicit form of the symmetric Jacobi on $H = A^T A$. More sophisticated versions use preconditioning to enhance numerical accuracy and efficiency (speed).

For instance, the QR factorization with pivoting is an excellent preconditioner in the following sense. If $A = QR$ is the QR factorization with standard Businger–Golub column pivoting ($A \leftarrow AP$, P permutation matrix), then the SVD Jacobi applied to R^T converges much faster than applied to A or R. This is because of implicitly performed step of Rutishauser's LR method (transition from $R^T R$ to RR^T). Moreover, the computed $\tilde{R}$ satisfies $A + \delta A = \hat{Q}\tilde{R}$, where $\hat{Q}$ is orthonormal and $\|\delta A e_i\|_2 \leq \eta_{QR}\|Ae_i\|_2$, $1 \leq i \leq n$. Here η_{QR} is bounded by a modest polynomial times the roundoff unit ϵ. Note that the relative backward error is small in each column of A. For more details see [9].

If we apply the SVD Jacobi on $\tilde{L} = \tilde{R}^T$, then by Proposition 3.2, for some orthogonal $\hat{V}$ it holds $\tilde{R} + \delta\tilde{R} = \hat{V}\tilde{\Sigma}\tilde{U}^T$. Here $\tilde{U}$ is numerically orthogonal, $\tilde{\Sigma}$ is diagonal, and $\|\delta\tilde{R}e_i\|_2 \leq \eta_J\|\tilde{R}e_i\|$, $1 \leq i \leq n$. Combining the results we obtain the relation

$$A + \Delta A = \hat{Q}\hat{V}\tilde{\Sigma}\tilde{U}^T, \quad \Delta A = \delta A + \hat{Q}\delta\tilde{R},$$

where

$$\|\Delta A e_i\|_2 \leq (\eta_{QR} + \eta_J + \eta_{QR}\eta_J)\|Ae_i\|_2, \quad 1 \leq i \leq n.$$

Here we note that the angle of Jacobi rotation underflows if the condition number $\kappa_2(A)$ overflows, and that standard construction of Jacobi rotation can lead to misconvergence of the algorithm. To avoid this, Jacobi rotation must be modified as shown in [11]. Also, instead of $\hat{Q}$, $\hat{V}$ we will have computed numerically orthogonal matrices $\tilde{Q}$, $\tilde{V}$ such that $\|\tilde{Q} - \hat{Q}\|_2$ and $\|\tilde{V} - \hat{V}\|_2$ are bounded by moderate polynomials of the dimensions times the roundoff ϵ.

Let $\sigma_1 \geq \cdots \geq \sigma_n > 0$ and $\tilde{\sigma}_1 \geq \cdots \geq \tilde{\sigma}_n$ be the eigenvalues of A and $A + \Delta A$, respectively. Write $D = \mathrm{diag}(\|Ae_1\|_2, \ldots, \|Ae_1\|_2)$ and

$$A + \Delta A = (I + \Delta A A^\dagger)A = (I + (\Delta A D^{-1})(AD^{-1})^\dagger)A\,.$$

¿From the variational characterization of the singular values, we immediately conclude

$$\max_i \frac{|\tilde{\sigma}_i - \sigma_i|}{\sigma_i} \leq \eta \equiv \|\Delta A A^\dagger\|_2 \leq \sqrt{n}(\eta_{QR} + \eta_J + \eta_{QR}\eta_J)\|A_s^\dagger\|_2,$$

where $A_s = AD^{-1}$. On the other hand, one can show that, for all i, $\tilde{\sigma}_i = \tilde{\Sigma}_{ii}(1 + O(n^2\epsilon))$.

By a theorem due to van der Sluis [50], we know that

$$\kappa_2(A_s) \equiv \|A_s\|_2\|A_s^\dagger\|_2 \leq \sqrt{n} \min_{S=diag,\det(S)\neq 0} \kappa_2(AS) \leq \sqrt{n}\kappa_2(A).$$

Thus, the SVD Jacobi algorithm computes the singular values with small relative backward error in each column of A. This means that small columns are preserved. The relative error in the computed singular values depends on the condition number of the column equilibrated matrix A_s, and not on the condition number of the initial A. These properties are not shared by bidiagonalization based methods. (Recent modification of the bidiagonalization process, due to Barlow [1], improves the accuracy of the bidiagonalization, but not to the level of the Jacobi SVD algorithm.)

The backward error in the Jacobi algorithm can be put into multiplicative form $(I + \Delta A A^{\dagger})A$ with small $\eta = \|\Delta A A^{\dagger}\|$ if $\kappa_2(A_s)$ is moderate. This fact also has important implications to the accuracy of the computed singular vectors. Let σ_i be simple with singular vectors u_i and v_i. If $u_i + \delta u_i$, $v_i + \delta v_i$ are the singular vectors of $A + \Delta A$, corresponding to $\tilde{\sigma}_i$, then

$$\max\{\sin \angle(u_i, u_i + \delta u_i), \sin \angle(v_i, v_i + \delta v_i)\} \leq \frac{O(\eta)}{\mathrm{gap}_i} \tag{3.13}$$

where

$$\mathrm{gap}_i = \min\left\{\min_{j \neq i} \frac{|\sigma_i - \tilde{\sigma}_j|}{\sigma_i}, 2\right\}.$$

Thus, the approximation error for the singular vectors of $A + \Delta A$ depends on the condition number $\kappa_2(A_s)$ and the <u>relative</u> separation of the singular values. Since our computed vectors are close to those of $A + \Delta A$, we can conclude that the SVD Jacobi computes the singular vectors with a bound like (3.13). The same conclusion then holds for the eigenvectors computed by the previously explained diagonalization procedure of symmetric positive definite matrix.

¿From the above analysis we can see that the SVD Jacobi can compute with high relative accuracy the SVD of any matrix A of the form $A = BS$, where S is any diagonal matrix, and B is well conditioned ($\kappa_2(B)$ moderate). A simple device can preserve this property if A is more general, for instance if $A = S_1 B S_2$, where S_1, S_2 are arbitrary diagonal scalings, and B is well conditioned. In that case, the QR factorization in the first step should be computed with column and row pivoting. For the details see [14], [5]. We only note that our theoretical understanding is one step behind numerical experience.

The SVD Jacobi method can be generalized e.g. to the SVD of the product of two matrices. For instance, if $A = BS$, $C = GD$ with well conditioned full column rank B, G with equilibrated columns, and arbitrary diagonal matrices S, D, then the SVD of $AC^T = BSDG^T$ can be computed as follows. First, we compute QR factorization with column pivoting of GDS, $(GDS)P = QR$. Due to pivoting, the matrix R is structured as $R = D_1 R_r$, where D_1 is diagonal and R_r is well conditioned. In fact, $\kappa_2(R_r)$ is bounded by a function of the dimension, for any initial matrix. Then, $AC^T = BPR^T Q^T$, and the product BPR^T is $BPR_r^T D_1$, where BPR_r^T is again well conditioned. Because of that, explicit computation of the product BPR^T will cause no loss of information and the Jacobi SVD will compute accurate SVD of the explicitly computed matrix product. For more details see [13], [12], [15].

In some cases, the matrix A is rather ill conditioned, but with special structure that allows accurate LU decomposition with complete pivoting, $P_1 A P_2 = LDU$, where P_1, P_2 are permutations, D is diagonal, and L, U are well conditioned. This means that computed matrices $\tilde{L}$, $\tilde{D}$, $\tilde{U}$ are such that the SVD of the product $P_1^T \tilde{L}\tilde{D}\tilde{U}P_2^T$ is highly accurate approximation of the SVD of A. On the other hand, $(P_1^T \tilde{L})(\tilde{D}\tilde{U}P_2^T)$ has the structure of the product of two matrices that allow accurate SVD by the Jacobi method. For more details see [8].

References

[1] Barlow J.: More Accurate Bidiagonal Reduction for Computing the Singular Value Decomposition. SIAM J. Matrix. Anal. Appl. To appear.

[2] Barlow J. and Demmel J.: Computing accurate eigensystems of scaled diagonally dominant matrices. SIAM J. Num. Anal., 27 (1990) 762–791.

[3] Brent R. and Luk F.: The Solution of Singular-value and Symmetric Eigenvalue Problems on Multiprocessor Arrays, SIAM Journal of Scientific and Statistical Computing 6 (1985) 69–84.

[4] Bunch J. and Parlett B.: Direct Methods for solving Symmetric indefinite systems of linear equations, SIAM J. Num. Anal. Vol. 8 No. 4 (1971) 639-655.

[5] Cox A. and Higham N. : Stability of Householder QR Factorization for Weighted Least Squares Problems. Proceedings of the 17th Dundee Biennial Conference Numerical Analysis 1997, Pitman Research Notes in Mathematics 380 (1998) 57-73. Ed. Griffiths D. F., Higham D. J. and Watson G. A. Pub. Addison Wesley Longman, Harlow, Essex, UK.

[6] Demmel J. and Veselić K.: Jacobi's method is more accurate than QR. SIAM J. Matrix Anal. Appl. 13 (1992) 1204–1245.

[7] Demmel J.: On floating point errors in Cholesky. Computer Science Department, University of Tennessee 1989. LAPACK Working Note 14.

[8] Demmel J., Gu M., Eisenstat S., Slapničar I., Veselić K. and Drmač Z.: Computing the singular value decomposition with high relative accuracy. Linear Algebra Appl. 299 (1999) 21-80.

[9] Drmač Z.: Computing the Singular and the Generalized Singular Values. Ph.D. thesis, Lehrgebiet Mathematische Physik, Fernuniversität Hagen, 1994.

[10] Drmač Z. and Hari V.: On the Quadratic Convergence of the J–symmetric Jacobi Method. Num. Math. 64 (1993) 147–180.

[11] Drmač Z.: Implementation of Jacobi rotations for accurate singular value computation in floating-point arithmetic. SIAM J. Sci. Comp. 18 (1997) 1200–1222.

[12] Drmač Z.: Accurate computation of the product induced singular value decomposition with applications. SIAM J. Numer. Anal. Vol 35, No. 5 (1998) 1969–1994.

[13] Drmač Z.: A tangent algorithm for computing the generalized singular value decomposition. SIAM J. Numer. Anal. Vol. 35, No. 5 (1998) 1804–1832.

[14] Drmač Z.: A posteriori computation of the singular vectors in a preconditioned Jacobi SVD algorithm. IMA J. Numer. Anal. 19 (1999) 191-213.

[15] Drmač Z.: New accurate algorithms for singular value decomposition of matrix triplets. SIAM J. Matrix Anal. Appl. 21 (2000) 1026–1050.

[16] Forsythe G. and Henrici P.: The Cyclic Jacobi Method for Computing the Principal Values of a Complex Matrix. Trans. Amer. Math. Soc. 94 (1960), 1–23.

[17] Golub G. and van Loan C. : Matrix Computations. The John Hopkins University Press, Baltimore and London 1989.

[18] Hari V. and Veselić K.: On Jacobi methods for singular value decompositions. SIAM J. Sci. Stat. Comput. Vol.8, No.5 (1987).

[19] Hari V.:
On Almost Diagonal Square Matrices With Multiple Singular Values. Radovi Matematički 4 (1988) 209–225.

[20] Hari V.: On Pairs of Almost Diagonal Matrices. Linear Algebra Appl. 148 (1991) 193–223.

[21] Hari V.: On Sharp Quadratic Convergence Bounds for the Serial Jacobi Methods. Numer. Math. 60: 375–406, 1991.

[22] Hari V. and Drmač Z.: On Scaled Almost Diagonal Hermitian Matrix Pairs. SIAM J. Matrix Anal. Appl. Vol. 18. No. 4 (1997) 12pp.

[23] Hari V.: Structure of Almost Diagonal Matrices. Mathematical Communications 4 (1999), 135–158.

[24] Henrici P. and Zimmermann K.: An Estimate for the Norms of Certain Cyclic Jacobi Operators. Linear Algebra Appl. 1 (1968) 289–501.

[25] Ipsen I.: Relative Perturbation Results for Matrix Eigenvalues and Singular Values. Acta Numerica 8 (1998) 151–201.

[26] van Kempen H.: On Quadratic Convergence of the Classical Jacobi Method for Real Symmetric Matrices with Nondinstinct Eigenvalues. Numer. Math. 9 (1966) 11–18.

[27] van Kempen H.: On Quadratic Convergence of the Special Cyclic Jacobi Method. Numer. Math. 9 (1966) 19–22.

[28] Luk F. and Park H.: On the equivalence and convergence of parallel Jacobi SVD algorithms. IEEE Computer (1987).

[29] Luk F. and Park H.: On parallel Jacobi orderings. SIAM J. Sci. Statist. Comput. 10 (1989) 18–26.

[30] Mascarenhas W.: On the Convergence of the Jacobi Method, Poster Presentation, Fourth SIAM Conference on Parallel Processing for Scientific Computing, Chicago, Illinois, December, 1989.

[31] Mascarenhas W.: On the Convergence of the Jacobi Method for Arbitrary Orderings I, SIAM Journal of Scientific and Statistical Computing, submitted, 1990.

[32] Matejaš J.: Quadratic Convergence of Scaled Iterates by Diagonalization Methods. Ph. D. thesis, University of Zagreb, 1999. Croatian language.

[33] Matejaš J.: Quadratic Convergence of Scaled Matrices in Jacobi Method. Numer.Math. 87 (1999) 171–199.

[34] Matejaš J.: Convergence of scaled iterates by Jacobi method. To appear in Linear Algebra Appl. p.p. 1-37.

[35] Matejaš J. and Hari V.: Note on the quadratic convergence of scaled matrices by J-symmetric Jacobi method. Preprint, University of Zagreb, 2002.

[36] Parlett B.: The Symmetric Eigenvalue Problem, Prentice–Hall Inc., Englewood Cliffs, N. J. 1980.

[37] Rhee N. and Hari V.: On the Global and Cubic Convergence of a Quasy-cyclic Jacobi Method. Numer. Math. 66, (1993) 97–122.

[38] Rosanoff R., Gloudeman J. and Levy S.: Numerical Conditions of Stiffness Matrix Formulations for Frame Structures. Proc.of the Second Conference on Matrix Methods in Structural Mechanics, WPAFB Dayton, Ohio, 1968.

[39] Schönhage A.: On Quadratic Convergence of the Jacobi Process. Numer. Math. 6 (1964) 410–412.

[40] Slapničar I.: Accurate Symmetric Eigenreduction by a Jacobi Method. Ph. D. Fernuniversität Hagen, Germany, 1992.

[41] Slapničar I.: Accurate computation of singular values and eigenvalues of symmetric matrices, Mathematical Communications 1 (1996) 153–168.

[42] Slapničar I.: Componentwise Analysis of Direct Factorisation of Real Symmetric and Hermitian Matrices. Linear Algebra Appl. 272 (1997) 227–275.

[43] Slapničar I.:Error Analysis of J-orthogonal Jacobi Method. submitted to Numer. Math.

[44] Slapničar I.: Highly Accurate Symmetric Eigenvalue Decomposition nd Hyperbolic SVD. submitted to Linear Algebra Appl.

[45] Slapničar I. and Truhar N.: Relative perturbation theory for hyperbolic eigenvalue problem. Linear Algebra Appl. 309 (2000) 57–72.

[46] Slapničar I. and Veselić K.: Perturbations of the eigenprojections of a factorised Hermitian matrix. Linear Algebra Appl. 218 (1995) 273–280.

[47] Slapničar I. and Veselić K.: A bound for the condition of a hyperbolic eigenvector matrix. Linear Algebra Appl. 290 (1999) 247–255.

[48] Stewart G. W. and Sun J.-G.: Matrix Perturbation Theory, Academic Press, Boston, 1990.

[49] Truhar N. and Slapničar I.: Relative perturbation bound for invariant subspaces of graded indefinite Hermitian matrices. Linear Algebra Appl. 301 (1999) 171–185.

[50] Van der Sluis A.: Condition numbers and equilibration of matrices. Numer. Math. 14 (1969) 14–23.

[51] Veselić K.: An Eigenreduction Algorithm for Definite Matrix Pairs and its Applications to Overdamped Linear Systems. Num. Math. 64 (1992), 241-269.

[52] Veselić K. and Hari V.: A note on a one-sided Jacobi algorithm. Numer. Math. 56 (1989) 627-633.

[53] Veselić K. and Slapničar I.: Floating-point Perturbations of Hermitian Matrices. Linear Algbra Appl. 195 (1993) 81–116.

[54] Voevodin, V.: Cislennye metody linejnoj algebry. Nauka, Moscow 1966.

[55] Wilkinson J.: Note on the Quadratic Convergence of the Cyclic Jacobi Process. Numer. Math. 4 (1962) 296–300.

[56] Wilkinson J.: Almost Diagonal Matrices with Multiple or Close Eigenvalues. Linear Algebra Appl. 1 (1968) 1–12.

[57] Wilkinson J. and Reinsch C.: Handbook for Automatic Computation, Vol. 2, Linear Algebra. Springer-Verlag, New York 1971.

[58] Zha H.: A Note on The Existence of the Hyperbolic Singular Value Decomposition. Linear Algbra Appl. 240 (1996) 199–205.