Numerical Algorithms

On Complex Falk-Langemeyer Method

--Manuscript Draft--

Manuscript Number:	
Full Title:	On Complex Falk-Langemeyer Method
Article Type:	Original Research
Keywords:	generalized eigenvalue problem; complex Hermitian matrices; definite matrix pair; diagonalization method
Corresponding Author:	Vjeran Hari, Ph. D. University of Zagreb Zagreb, CROATIA
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	University of Zagreb
Corresponding Author's Secondary Institution:	
First Author:	Vjeran Hari, Ph. D.
First Author Secondary Information:	
Order of Authors:	Vjeran Hari, Ph. D.
Order of Authors Secondary Information:	
Funding Information:	Hrvatska Zaklada za Znanost (IP 09 2014 3670) Prof. Vjeran Hari
Abstract:	A new algorithm for the simultaneous diagonalization of two complex Hermitian matrices is derived. It is a proper generalization of the known Falk-Langemeyer algorithm which was originally derived in 1960 for a pair of positive definite matrices. It is proved that the complex Falk-Langemeyer algorithm is well defined for a pair of Hermitian matrices which make a definite pair. Special attention is paid to the stability of the formulas for the transformation parameters in the case when the pivot submatrices are almost proportional. Numerical tests show the high relative accuracy of the method if both matrices are definite and well-behaved, i.e. if they can be well-scaled symmetrically.

Numerical Algorithms manuscript No.
(will be inserted by the editor)

On Complex Falk-Langemeyer Method

Vjeran Hari

Received: date / Accepted: date

Abstract A new algorithm for the simultaneous diagonalization of two complex Hermitian matrices is derived. It is a proper generalization of the known Falk-Langemeyer algorithm which was originally derived in 1960 for a pair of positive definite matrices. It is proved that the complex Falk-Langemeyer algorithm is well defined for a pair of Hermitian matrices which make a definite pair. Special attention is paid to the stability of the formulas for the transformation parameters in the case when the pivot submatrices are almost proportional. Numerical tests show the high relative accuracy of the method if both matrices are definite and well-behaved, i.e. if they can be well-scaled symmetrically.

Keywords generalized eigenvalue problem · complex Hermitian matrices · definite matrix pair · diagonalization method

Mathematics Subject Classification (2000) 65F15

1 Introduction

In 1960 S. Falk and P. Langemeyer [3] proposed a method for the simultaneous diagonalization of two real symmetric positive definite matrices. Their method solves the generalized eigenvalue problem (GEP) $Ax = \lambda Bx$, $x \neq 0$. Later Slapničar and Hari [17] proved the asymptotic quadratic convergence of the method under the serial pivot strategies. In [17] it was also proved that the method was well-defined for a definite pair of symmetric matrices [18]. In 2015 Matejaš [12] considered accuracy properties of the method. Although the paper did not consider the high relative accuracy of the method in the case of positive definite matrices A, B, it provided a very detailed error analysis of the method. Our numerical tests indicate that the method computes

This work has been fully supported by Croatian Science Foundation under the project: IP_09_2014_3670.

Department of Mathematics, Faculty of Science, University of Zagreb, Bijenička cesta 30, 10000 Zagreb, Croatia. E-mail: hari@math.hr

[0] This paper is dedicated to Professor S. Falk

the eigenvalues and eigenvectors of the pair (A,B) to high relative accuracy provided that A and B are well-behaved positive definite matrices. It means that the condition numbers of D_AAD_A and D_BBD_B are small for some diagonal matrices D_A and D_B (see [1,2]). We note that this important property of the FL method is not shared with the QZ, QR and other methods which reduce the problem to the eigenproblem for one symmetric tridiagonal matrix (see [9]). Typically, if the starting matrices are ill-conditioned with respect to matrix inversion or if a positive definitizing shift [10] μ is not known in advance (but still, shifting $A \mapsto A - \mu B$ can cause problems with high relative accuracy of the computed eigenvalues) then the Falk-Langemeyer (FL) method is a good choice. It can be made faster if its inherent parallelism is combined with the BLAS1 `saxpy` computational routine. Also, additional accuracy can be obtained if the floating-point fused multiply and add operation is used, computing $\alpha\beta + \gamma$ with a single rounding, which is now an IEEE-754 standard operator. As a Jacobi-type method, it is very fast and accurate when A and B are nearly diagonal (cf. [11]). This happens in the course of modeling the parameters of a system. Although the global convergence of the FL method has not been considered yet, much is known since it can be linked to the globally convergent HZ method from [9] (see [17], [4, 21]). In conclusion, the FL method is a reliable, accurate and fast Jacobi-type method for the definite GEP. On contemporary CPU and GPU parallel computing machines its main application is to serve as a kernel algorithm for the block Jacobi methods which are used to compute GSVD or solve definite GEP (see[13]). The block Jacobi methods are almost perfectly parallelizable, parallel shared memory versions of the methods are highly scalable, and their speed up almost solely depends on the number of cores used [13]. They compare favorably to the LAPACK `DTGSJA` algorithm.

In this paper, we derive complex FL (CFL) method. Although the obtained formulas are the proper generalizations of the ones in the real case, their derivation is far from trivial. Like in the real case, the formulas for the transformation parameters become useless when the pivot submatrices are proportional. In such a case we provide additional stable formulas. Since the new algorithm is the proper generalization of the real one, the quadratic asymptotic convergence of the CFL can be proved in a straightforward way using the analysis from [17] together with the results from [5, 6]. The global convergence can be proved by linking the method to the complex HZ method from [4], for which the global convergence proof is almost identical to that from [9]. Our main focus in this paper is to derive the complex method and to show that it is well defined for any definite pair of Hermitian matrices. We also provide numerical tests in MATLAB which indicate the high relative accuracy of the method when both matrices A and B are well-behaved positive definite Hermitian matrices.

The paper is organized as follows. In Section 2 we derive the CFL algorithm and show its properties. Is Subsection 2.1 we derive the formulas for the parameters α and β of the transformation matrix. In Subsection 2.2 we define the algorithm and prove its properties. In Section 3 we describe how the numerical tests have been prepared and done. We display the data which strongly indicate the high relative accuracy of the method. The conclusions and proposals for future work are briefly outlined in Section 4.

2 The Derivation of the Complex Falk–Langemeyer Algorithm

Let A and B be two n by n complex Hermitian matrices. The complex Falk–Langemeyer method solves the generalized eigenproblem $Ax = \lambda Bx$ by generating a sequence of "congruent" matrix pairs $(A^{(1)},B^{(1)}),(A^{(2)},B^{(2)}),\ldots$ where $A^{(1)} = A$, $B^{(1)} = B$ and

$$A^{(k+1)} = F_k^* A^{(k)} F_k \ , \quad B^{(k+1)} = F_k^* B^{(k)} F_k \ , \ k \geq 1. \tag{2.1}$$

Here F_k^* denotes the Hermitian transpose of F_k. The transformation matrices are nonsingular *elementary plane matrices* with unit diagonal. Each F_k differs from the identity in only two elements at positions $(i(k),j(k))$ and $(j(k),i(k))$, where $1 \leq i(k) < j(k) \leq n$. The pair $(i(k),j(k))$ is called *pivot pair* and the 2×2 matrix $\hat{F}_k = [e_{i(k)},e_{j(k)}]^* F_k [e_{i(k)},e_{j(k)}]$ is called *pivot submatrix* of F_k. Here $e_1,\ldots,e_n$ are the columns of the identity matrix I_n. For the CFL method we assume

$$\hat{F}_k = \begin{bmatrix} 1 & \alpha_k \\ \beta_k & 1 \end{bmatrix} \ , \quad k \geq 1, \tag{2.2}$$

where the complex scalars α_k and β_k are chosen to satisfy the condition

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

Here, $A^{(k)} = (a_{ij}^{(k)})$, $B^{(k)} = (b_{ij}^{(k)})$, $k \geq 1$. The transition from the pair $(A^{(k)},B^{(k)})$ to the pair $(A^{(k+1)},B^{(k+1)})$ is the *k–th step* of the method. A way how the pivot pairs are selected is called *pivot strategy*. A pivot strategy is *cyclic* if every sequence of $N = n(n-1)/2$ successive pivot pairs contains all pairs from the set $\mathscr{P}_n = \{(p,q); 1 \leq p < q \leq n\}$. For each cyclic strategy, the sequence of N successive steps starting with the matrix pair $(A^{((r-1)N+1)},B^{((r-1)N+1)})$ is referred to as the r'th *cycle*. Two most common cyclic pivot strategies are the *column-cyclic* and the *row-cyclic strategy*. The former is defined by the sequence of pairs $(1,2),(1,3),(2,3),(1,4),(2,4),(3,4),\ldots,(1,n),\ldots$ $,(n-1,n)$ and the latter by $(1,2),(1,3),\ldots,(1,n),(2,3),\ldots,(2,n),\ldots,(n-1,n)$. These two strategies are also called *serial strategies*. Recently, a large set of *generalized serial strategies* has been introduced. It includes the set of weakly-wavefront [16] and many other cyclic strategies (see [8]).

If the eigenvectors are wanted, we have to calculate the sequence of matrices $F^{(1)}$, $F^{(2)},\ldots$, where

$$F^{(1)} = I, \quad F^{(k+1)} = F^{(k)} F_k, \quad k \geq 1\,. \tag{2.3}$$

From the relations (2.1) and (2.3) we obtain for $k \geq 2$

$$F^{(k)} = F_1 \cdots F_{k-1} \qquad \text{and} \qquad A^{(k)} = (F^{(k)})^* A^{(1)} F^{(k)}, \ \ B^{(k)} = (F^{(k)})^* B^{(1)} F^{(k)}\,.$$

2.1 Computation of the transformation parameters

To derive an algorithm for computing α_k, β_k from (2.2), we consider the case of matrices of order two. Since there is just one step to perform, we omit k and use special notation. In particular, we look for α and β which satisfy the following two matrix equations

$$\begin{bmatrix} 1 & \bar{\beta} \\ \bar{\alpha} & 1 \end{bmatrix} \begin{bmatrix} a_1 & a_2 \\ \bar{a}_2 & a_3 \end{bmatrix} \begin{bmatrix} 1 & \alpha \\ \beta & 1 \end{bmatrix} = \begin{bmatrix} a_1' & 0 \\ 0 & a_3' \end{bmatrix}, \qquad \begin{bmatrix} 1 & \bar{\beta} \\ \bar{\alpha} & 1 \end{bmatrix} \begin{bmatrix} b_1 & b_2 \\ \bar{b}_2 & b_3 \end{bmatrix} \begin{bmatrix} 1 & \alpha \\ \beta & 1 \end{bmatrix} = \begin{bmatrix} b_1' & 0 \\ 0 & b_3' \end{bmatrix}.$$

Here a_2, b_2, α and β are complex while the other elements of the matrices are real. The unknowns α and β will be determined from the system of two equations, which are obtained by equating $(1,2)$-elements on the left- and right-hand sides of the above matrix equations. We obtain

$$e_1 = a_1\alpha + a_3\bar{\beta} + \bar{a}_2\alpha\bar{\beta} + a_2 = 0 \tag{2.4}$$

$$e_2 = b_1\alpha + b_3\bar{\beta} + \bar{b}_2\alpha\bar{\beta} + b_2 = 0. \tag{2.5}$$

To solve the above system of equations, we shall use the following quantities

$$\mathfrak{I}_1 = a_1b_2 - a_2b_1 = \begin{vmatrix} a_1 & b_1 \\ a_2 & b_2 \end{vmatrix} \tag{2.6}$$

$$\mathfrak{I}_3 = a_3b_2 - a_2b_3 = \begin{vmatrix} a_3 & b_3 \\ a_2 & b_2 \end{vmatrix} \tag{2.7}$$

$$\mathfrak{I}_2 = \mathfrak{I}_2' + i\mathfrak{I}_2'', \qquad \mathfrak{I}_2',\ \mathfrak{I}_2' \ \text{real} \tag{2.8}$$

$$\mathfrak{I}_2' = a_1b_3 - a_3b_1 = \begin{vmatrix} a_1 & b_1 \\ a_3 & b_3 \end{vmatrix} \tag{2.9}$$

$$i\mathfrak{I}_2'' = a_2\bar{b}_2 - \bar{a}_2b_2 = \begin{vmatrix} a_2 & b_2 \\ \bar{a}_2 & \bar{b}_2 \end{vmatrix} = i\left(-2\begin{vmatrix} \mathrm{Re}(a_2) & \mathrm{Re}(b_2) \\ \mathrm{Im}(a_2) & \mathrm{Im}(b_2) \end{vmatrix}\right). \tag{2.10}$$

Let

$$\begin{bmatrix} \tilde{e}_1 \\ \tilde{e}_2 \end{bmatrix} = \begin{bmatrix} b_2 & -a_2 \\ \bar{b}_2 & -\bar{a}_2 \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \end{bmatrix}. \tag{2.11}$$

Then the relations (2.4), (2.5) and (2.11) imply

$$\tilde{e}_1 = \mathfrak{I}_1\alpha + \mathfrak{I}_3\bar{\beta} - (i\mathfrak{I}_2'')\alpha\bar{\beta} = 0 \tag{2.12}$$

$$\bar{\tilde{e}}_2 = \mathfrak{I}_1\bar{\alpha} + \mathfrak{I}_3\beta - i\mathfrak{I}_2'' = 0. \tag{2.13}$$

The relation (2.11) shows that the system of equations (2.4) - (2.5) implies the system (2.12) - (2.13) in the sense that every solution of the system (2.4) - (2.5) is a solution of the system (2.12) - (2.13). The opposite implication is true only if $\mathfrak{I}_2'' \neq 0$. Geometrically, $\mathfrak{I}_2'' \neq 0$ means that nonzero complex numbers a_2 and b_2 do not lie on a line passing through the origin.

Lemma 2.1 *The following identities hold*

(i) $\begin{vmatrix} \mathfrak{I}_1 & \mathfrak{I}_3 \\ a_1 & a_3 \end{vmatrix} = a_2\mathfrak{I}_2', \qquad \begin{vmatrix} \mathfrak{I}_1 & \mathfrak{I}_3 \\ b_1 & b_3 \end{vmatrix} = b_2\mathfrak{I}_2'$

(ii) $\begin{vmatrix} a_2 & \bar{a}_2 \\ \mathfrak{I}_1 & \bar{\mathfrak{I}}_1 \end{vmatrix} = a_1(i\mathfrak{I}_2''), \qquad \begin{vmatrix} b_2 & \bar{b}_2 \\ \mathfrak{I}_1 & \bar{\mathfrak{I}}_1 \end{vmatrix} = b_1(i\mathfrak{I}_2'')$

(iii) $$\begin{vmatrix} \Im_1 & \bar{\Im}_1 \\ \Im_3 & \bar{\Im}_3 \end{vmatrix} = \Im_2'(i\Im_2'').$$

Proof All identities are implied by the definitions (2.6)– (2.10).

Let

$$\hat{A} = \begin{bmatrix} a_1 & a_2 \\ \bar{a}_2 & a_3 \end{bmatrix}, \hat{B} = \begin{bmatrix} b_1 & b_2 \\ \bar{b}_2 & b_3 \end{bmatrix}, \hat{F} = \begin{bmatrix} 1 & \alpha \\ \beta & 1 \end{bmatrix}, \hat{A}' = \begin{bmatrix} a_1' & \\ & a_3' \end{bmatrix}, \hat{B}' = \begin{bmatrix} b_1' & \\ & b_3' \end{bmatrix}. \quad (2.14)$$

Consider the transformation $(\hat{A},\hat{B}) \to (\hat{A}_\varphi,\hat{B}_\varphi)$ where

$$\begin{bmatrix} \hat{A}_\varphi \\ \hat{B}_\varphi \end{bmatrix} = \begin{bmatrix} \cos\varphi I_2 & -\sin\varphi I_2 \\ \sin\varphi I_2 & \cos\varphi I_2 \end{bmatrix} \begin{bmatrix} \hat{A} \\ \hat{B} \end{bmatrix}, \quad 0 \leq \varphi \leq 2\pi. \quad (2.15)$$

Lemma 2.2 *The solution* (α,β) *of the system (2.4) - (2.5) and the quantities* $\Im_1, \Im_2$ *and* $\Im_3$ *are invariant under the transformation (2.15).*

Proof Let $\hat{A}' = \hat{F}^*\hat{A}\hat{F}$, $\hat{B}' = \hat{F}^*\hat{B}\hat{F}$ where $\hat{A}$, $\hat{B}$, $\hat{F}$, $\hat{A}'$, $\hat{B}'$ are as in the relation (2.14). If $\hat{F}$ simultaneously diagonalizes $\hat{A}$ and $\hat{B}$, then for any φ, $0 \leq \varphi \leq 2\pi$, the matrices

$$\hat{F}^*\hat{A}_\varphi\hat{F} = \cos\varphi\hat{A}' - \sin\varphi\hat{B}' \quad \text{and} \quad \hat{F}^*\hat{B}_\varphi\hat{F} = sin\varphi\hat{A}' + \cos\varphi\hat{B}'$$

are diagonal. From (2.15) it follows that the converse is also true. Namely, if $\hat{F}_\varphi$ simultaneously diagonalizes $\hat{A}_\varphi$ and $\hat{B}_\varphi$ via the congruence transformation, then the relation

$$\begin{bmatrix} \hat{F}_\varphi^*\hat{A}\hat{F}_\varphi \\ \hat{F}_\varphi^*\hat{B}\hat{F}_\varphi \end{bmatrix} = \begin{bmatrix} \cos\varphi I_2 & \sin\varphi I_2 \\ -\sin\varphi I_2 & \cos\varphi I_2 \end{bmatrix} \begin{bmatrix} \hat{F}_\varphi^*\hat{A}_\varphi\hat{F}_\varphi \\ \hat{F}_\varphi^*\hat{B}_\varphi\hat{F}_\varphi \end{bmatrix}$$

shows that it does the same for the matrices $\hat{A}$ and $\hat{B}$. This holds for any $0 \leq \varphi \leq 2\pi$.

If the elements of $\hat{A}_\varphi$ and $\hat{B}_\varphi$ are denoted by $a_r(\varphi)$ and $b_r(\varphi)$, $1 \leq r \leq 3$, then we have

$$[a_r(\varphi)\;\; b_r(\varphi)] = [a_r\;\; b_r]\; R_\varphi, \quad R_\varphi = \begin{bmatrix} \cos\varphi & \sin\varphi \\ -\sin\varphi & \cos\varphi \end{bmatrix}.$$

Hence, for the quantities $\Im_t(\varphi)$, $1 \leq t \leq 3$ associated with the pair $(\hat{A}_\varphi,\hat{B}_\varphi)$, we have $\Im_t(\varphi) = \Im_t \cdot \det(R_\varphi) = \Im_t$, $1 \leq t \leq 3$.

Let

$$\Im = \Im_2^2 + 4\bar{\Im}_1\Im_3.$$

By Lemma 2.1 (iii) we have

$$\begin{aligned} \Im &= (\Im_2')^2 - (\Im_2'')^2 + 2i\Im_2'\Im_2'' + 4\bar{\Im}_1\Im_3 && (2.16) \\ &= (\Im_2')^2 - (\Im_2'')^2 + 2\Im_1\bar{\Im}_3 - 2\bar{\Im}_1\Im_3 + 4\bar{\Im}_1\Im_3 \\ &= (\Im_2')^2 - (\Im_2'')^2 + 2(\Im_1\bar{\Im}_3 + \bar{\Im}_1\Im_3). && (2.17) \end{aligned}$$

The relation (2.17) shows that $\Im$ is real. Recall that the pair $(\hat{A},\hat{B})$ is definite if the matrix $\sigma\hat{A} + \omega\hat{B}$ is positive definite for some real σ and ω.

Lemma 2.3 *Suppose the pair $(\hat{A},\hat{B})$ is definite. Then*

(i) $\Im \geq 0$

(ii) *The following statements are equivalent*

(a) $\Im = 0$

(b) $\Im_1 = \Im_2 = \Im_3 = 0$

(c) $\sigma A + \omega B = 0$ *for some real* σ *and* ω *such that* $|\sigma| + |\omega| > 0$.

Proof Since the pair $(\hat{A},\hat{B})$ is definite, there exists some φ such that $\hat{B}_\varphi$ from the relation (2.15) is positive definite. We can prove the lemma for the pair (A_φ, B_φ) and then invoke the preceding lemma. This shows that in the proof we can assume that $\hat{B}$ is positive definite.

(i) Consider first the case $a_2 = 0$. Since $\Im_2'' = 0$, the relation (2.16) implies

$$\begin{aligned}
\Im &= (\Im_2')^2 + 4a_1\bar{b}_2a_3b_2 = (a_1b_3 - a_3b_1)^2 + 4a_1a_3|b_2|^2 \\
&= (a_1b_3)^2 + (a_3b_1)^2 - 2a_1a_3(b_1b_3 - 2|b_2|^2) \\
&\geq (a_1b_3)^2 + (a_3b_1)^2 - 2|a_1a_3|\max\{b_1b_3 - |b_2|^2, |b_2|^2\} \\
&\geq (a_1b_3)^2 + (a_3b_1)^2 - 2|a_1a_3|b_1b_3 \\
&= (|a_1|b_3 - |a_3|b_1)^2 \geq 0. \qquad (2.18)
\end{aligned}$$

If $b_2 = 0$ then we obtain $\Im = (a_1b_3 - a_3b_1)^2 + 4b_1b_3|a_2|^2 \geq 0$.

Consider now the case $a_2 \neq 0$, $b_2 \neq 0$. Let

$$x = a_1\sqrt{\frac{b_3}{b_1}}, \quad y = a_3\sqrt{\frac{b_1}{b_3}}, \quad z = \frac{b_2}{\sqrt{b_1b_3}},$$

$$a_2 = a_2' + ia_2'', \quad z = z' + iz'', \quad a_2', a_2'', z', z'' \ \text{real}.$$

We have

$$\Im = b_1b_3\{(x-y)^2 - 4(a_2'z'' - a_2''z')^2 + 4\mathrm{Re}[(\bar{a}_2 - x\bar{z})(a_2 - yz)]\}.$$

Hence

$$\begin{aligned}
\frac{1}{4b_1b_3}\Im &= \frac{(x-y)^2}{4} - (a_2'z'' - a_2''z')^2 + (a_2' - z'x)(a_2' - z'y) + (a_2'' - z''x)(a_2'' - z''y) \\
&= \frac{(x-y)^2}{4} + (1 - |z|^2)|a_2|^2 + xy|z|^2 + (a_2'z + a_2''z'')^2 - (a_2'z + a_2''z'')(x+y).
\end{aligned}$$

Let

$$q = (a_2'z' + a_2''z'')/|a_2| = |z|\cos(\angle(a_2,b_2)),$$

where $\angle(a_2,b_2)$ is the (smaller) angle between the radii-vectors determined by the complex numbers a_2 and b_2. Since $|z| < 1$, by the Cauchy-Schwarz inequality, we have $|q| \leq |z| < 1$. We have

$$\begin{aligned}
\frac{1}{4b_1b_3}\Im &= \left[(|a_2|q)^2 - (x+y)(|a_2|q)\right] + xy|z|^2 + \frac{1}{4}(x-y)^2 + (1-|z|^2)|a_2|^2 \\
&= \left(|a_2|q - \frac{x+y}{2}\right)^2 + (1-|z|^2)(|a_2|^2 - a_1a_3). \qquad (2.19)
\end{aligned}$$

If $|a_2|^2 \geq a_1a_3$ we have $\Im \geq 0$. Hence it remains to consider the opposite case. So, let $0 < |a_2|^2 < a_1a_3 = xy$.

If $q(x+y) \leq 0$, we see from the first line of the relation (2.19) that all terms on the right-hand side are nonnegative. So, it remains to consider the case $q(x+y) > 0$, which means that x, y and q are nonzero and have the same sign.

Let $w = a_2/\sqrt{xy}$. Then $|w| < 1$ and we have

$$\left|\frac{x+y}{2} - |a_2|\,q\right| = \frac{|x+y|}{2} - |q|\,|a_2| \geq \sqrt{xy} - |q|\,|w|\sqrt{xy} \geq (1-|w|\,|z|)\sqrt{xy}.$$

Using the obtained inequality in the relation (2.19) we obtain

$$\frac{1}{4b_1b_3}\Im \geq (1-|w|\,|z|)^2xy - (1-|z|^2)(1-|w|^2)xy = (|w|-|z|)^2xy \geq 0. \qquad (2.20)$$

(ii) We shall prove the chain of implications (a) ⇒ (b) ⇒ (c) ⇒ (d).

$(a) \Rightarrow (b)$. We consider first the case $a_2 = 0$. From the condition (a) and the first line of the relation (2.18) we conclude that $a_1a_3 \leq 0$. If $a_1a_3 = 0$ then from the same line we conclude $\Im_2' = 0$. Thus, $a_1/b_1 = a_3/b_3$ implying $a_1 = a_3 = 0$. So, $\hat{A} = 0$ and the condition (b) holds. If $a_1a_3 < 0$ then the first line of the relation (2.18) yields

$$0 = \Im = (|a_1|b_3 + |a_3|b_1)^2 - 4|a_1a_3|\,|b_2|^2 = (|a_1|b_3 - |a_3|b_1)^2 + 4|a_1a_3|\,(b_1b_3 - |b_2|^2).$$

Since $b_1b_3 - |b_2|^2 > 0$ we must have $a_1a_3 = 0$ which contradicts to $a_1a_3 < 0$. We conclude that the case $a_1a_3 < 0$ cannot occur.

If $b_2 = 0$, we have $0 = \Im = (a_1b_3 - a_3b_1)^2 + 4b_1b_3|a_2|^2$, implying $a_2 = 0$ and $\Im_2' = 0$. Hence $\hat{A}$ and $\hat{B}$ are diagonal and proportional. Consequently the condition (b) holds.

Let $a_2 \neq 0$, $b_2 \neq 0$. From the relation (2.19) we see that the case $|a_2|^2 > a_1a_3$ cannot occur.

Let us consider the case $0 < |a_2|^2 = a_1a_3 = xy$. The condition $\Im = 0$, the relation (2.19) and $|a_2| = \sqrt{xy}$ imply

$$0 = \frac{x+y}{2} - |a_2|q = \frac{x+y}{2} - q\sqrt{xy} \quad \Leftrightarrow \quad q\sqrt{xy} = \frac{x+y}{2}$$

which is impossible since $|q| \leq |z| < 1$. Thus that case cannot occur.

It remains to consider the case $0 < |a_2|^2 < a_1a_3 = xy$.

If $q(x+y) \leq 0$, we see from the first line of the relation (2.19) that all terms on the right-hand side are nonnegative and the term $(1-|z|^2)|a_2|^2$ is positive. Hence that case cannot occur.

So, we have $q(x+y) > 0$. It means that the relation (2.20) holds. Now, the condition $\Im = 0$ implies that all inequalities in the relation (2.20) are equalities. That implies

$$\frac{|x|+|y|}{2} = \sqrt{xy}, \qquad |q| = |z|, \qquad |w| = |z|.$$

We first conclude $|x| = |y|$ and then since $xy > 0$ we conclude $x = y$. This means $\Im_2' = 0$ The condition $|q| = |z|$ means $\cos(\angle(a_2, b_2)) = \pm 1$. Hence, if $a_2 \neq b_2$ the line connecting a_2 and b_2 passes through the origin. Therefore, the condition $|w| = |z|$ implies $w = \pm z$.

For a_1, a_3 we have two possibilities: either $a_1 > 0, a_3 > 0$ or $a_1 < 0, a_3 < 0$.

In the first case we have $x = y > 0$, $q > 0$, $\cos(\angle(a_2, b_2)) = 1$ hence $q = |z|$ and $w = z$. Thus $a_2 = \frac{a_1 a_3}{b_1 b_3} b_2$ which implies $\Im''_2 = 0$. We have obtained $\Im_2 = 0$. Now the relation (2.16) implies $4\bar{\Im}_1\Im_3 = 0$. Note that by Lemma 2.1(i) $\Im_1 b_3 = \Im_3 b_1$. Hence, we have $\Im_1 = 0$, $\Im_3 = 0$, $\Im_2 = 0$ and the condition (b) is fulfilled.

In the second case we have $x = y < 0$, $q < 0$, $\cos(\angle(a_2, b_2)) = -1$. Hence $q = -|z|$ and $w = -z$. So we have $a_2 = -\frac{|a_1||a_3|}{b_1 b_3} b_2$ which implies $\Im''_2 = 0$. As in the first case we conclude that the condition (b) is fulfilled.

$(b) \Rightarrow (c)$. If $\hat{B}$ is positive definite, then the condition $\Im_1 = \Im_2 = \Im_3 = 0$ implies

$$a_2 = \mu b_2, \quad a_1 = \mu b_1, \quad a_3 = \mu b_3,$$

with $\mu = a_1/b_1 = a_3/b_3$. Thus, $\hat{A} = \mu\hat{B}$.

If $\hat{A}$ is positive definite, we have $\hat{B} = \nu\hat{A}$ with $\nu = b_1/a_1 = b_3/a_3$.

If neither $\hat{A}$ nor $\hat{B}$ is positive definite, then $\hat{B}_\varphi$ is positive definite for some $0 \leq \varphi < 2\pi$. Then we have $\hat{A}_\varphi = \mu_\varphi \hat{B}_\varphi$ or equivalently $(\mu_\varphi \sin\varphi - \cos\varphi)\hat{A} + (\mu_\varphi \cos\varphi + \sin\varphi)\hat{B} = 0$, where $(\mu_\varphi \sin\varphi - \cos\varphi)^2 + (\mu_\varphi \cos\varphi + \sin\varphi)^2 = 1 + \mu_\varphi^2 > 1$.

$(c) \Rightarrow (a)$. If $s\hat{A} + t\hat{B} = 0$ for some real s and t with $|s| + |t| > 0$, then $\hat{A} = \hat{\mu}\hat{B}$ or $\hat{B} = \hat{\nu}\hat{A}$ for some real $\hat{\mu}$ or $\hat{\nu}$. This implies $\Im_1 = \Im_2 = \Im_3 = 0$ and consequently $\Im = 0$.

Lemma 2.4 *Let* $(\hat{A}, \hat{B})$ *be definite and* $\Im > 0$. *Then*

(i) $\alpha = 0$ *iff* $\Im_3 = 0$
(ii) $\beta = 0$ *iff* $\Im_1 = 0$
(iii) $\alpha = \beta = 0$ *iff* $\Im_1 = \Im_3 = 0$.

Proof By Lemma 2.2 we can assume that $\hat{B}$ is positive definite. We can prove (i) and (ii) simultaneously.

Let $\alpha = 0$ $(\beta = 0)$. Then the equations (2.4) and (2.5) yield

$$\begin{aligned} \bar{\beta} a_3 + a_2 &= 0 \qquad (\alpha a_1 + a_2 = 0), \\ \bar{\beta} b_3 + b_2 &= 0 \qquad (\alpha b_1 + b_2 = 0). \end{aligned}$$

If we multiply the first equation by b_3 (b_1), the second one by $-a_3$ $(-a_1)$ and add them together, we obtain $\Im_3 = 0$ $(\Im_1 = 0)$.

To prove the opposite direction, we multiply the equation (2.4) by b_3 (b_1), the equation (2.5) by $-a_3$ $(-a_1)$ and add them together. We obtain

$$\Im'_2\alpha - \bar{\Im}_3\alpha\bar{\beta} - \Im_3 = 0, \qquad (\Im'_2\bar{\beta} + \bar{\Im}_1\alpha\bar{\beta} + \Im_1 = 0).$$

Hence the assumption $\Im_3 = 0$ $(\Im_1 = 0)$ implies

$$\alpha\Im'_2 = 0 \qquad (\bar{\beta}\Im'_2 = 0).$$

It remains to prove $\Im_2' \neq 0$. Indeed, By Lemma 2.1 $\Im_2' = 0$ and $\Im_3 = 0$ ($\Im_1 = 0$) would imply $\Im_1 = \Im_2 = \Im_3 = 0$. By Lemma 2.3(ii) that would imply $\Im = 0$, which is not true. Hence $\Im_2' \neq 0$ and $\alpha = 0$ ($\beta = 0$).

(iii) If $\alpha = \beta = 0$, then the equations (2.4) and (2.5) are reduced to $a_2 = 0$ and $b_2 = 0$, respectively. Then $\Im_1 = \Im_3 = \Im_2'' = 0$.

Now, suppose that $\Im_1 = \Im_3 = 0$. Note that $\Im_1 = \Im_3 = 0$ can be written as

$$\begin{bmatrix} a_1 & b_1 \\ a_3 & b_3 \end{bmatrix} \begin{bmatrix} b_2 \\ -a_2 \end{bmatrix} = 0. \tag{2.21}$$

Since $(\Im_2)^2 = \Im > 0$, the relation (2.16) implies $\Im_2'\Im_2'' = 0$. The case $\Im_2' = 0$ would together with $\Im_1 = \Im_3 = 0$ imply $0 < \Im = -(\Im_2'')^2$, which is impossible. So, we conclude that $\Im_2'' = 0$ and $\Im_2' \neq 0$. Now, (2.21) implies $a_2 = b_2 = 0$ and the equations (2.4) and (2.5) are reduced to the system

$$\begin{bmatrix} a_1 & a_3 \\ b_1 & b_3 \end{bmatrix} \begin{bmatrix} \alpha \\ \bar{\beta} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \qquad \text{with} \qquad \begin{vmatrix} a_1 & a_3 \\ b_1 & b_3 \end{vmatrix} = \Im_2' \neq 0.$$

We conclude that $\alpha = \beta = 0$.

Lemma 2.5 *Suppose $(\hat{A}, \hat{B})$ is definite and $\Im > 0$. Then the solution (α, β) of the system (2.4) - (2.5) is given by*

$$\alpha = \frac{\Im_3}{\nu}, \qquad \beta = -\frac{\bar{\Im}_1}{\nu}, \tag{2.22}$$

where ν is any nonzero solution of the equation

$$\nu^2 - \Im_2 \nu - \bar{\Im}_1 \Im_3 = 0. \tag{2.23}$$

Proof By Lemma 2.2 we can assume that $\hat{B}$ is positive definite. To solve the system of equations (2.4) - (2.5) we distinguish two cases: $\Im_1\Im_3 = 0$ and $\Im_1\Im_3 \neq 0$.

$\Im_1\Im_3 = 0$. In this case Lemma 2.1(iii) implies $\Im_2'\Im_2'' = 0$. If $\Im_2' = 0$ then by Lemma 2.1(i) one obtains $\Im_1 = \Im_3 = 0$ and consequently by Lemma 2.1(ii) $\Im_2'' = 0$. Thus $\Im = 0$ which contradicts to the assumption $\Im > 0$. So, we must have $\Im_2' \neq 0$ and therefore $\Im_2'' = 0$.

From Lemma 2.4, we know that $\Im_1 = 0$ ($\Im_3 = 0$) implies $\beta = 0$ ($\alpha = 0$). From (2.4) - (2.5) we see that $\beta = 0$ ($\alpha = 0$) implies $\alpha = -b_2/b_1$ ($\beta = -\bar{b}_2/b_3$). By Lemma 2.1(i) we conclude that $-b_2/b_1 = \Im_3/\Im_2'$ ($\bar{b}_2/b_3 = -\bar{\Im}_1/\Im_2'$). Hence we obtain the solution $\alpha = \Im_3/\Im_2' = \Im_3/\Im_2$, $\beta = 0$ ($\alpha = 0$, $\beta = -\bar{\Im}_1/\Im_2' = -\bar{\Im}_1/\Im_2$), where $\Im_2$ ($= \Im_2'$) is the nonzero solution of the equation (2.23). This proves the lemma in the case $\Im_1\Im_3 = 0$.

$\Im_1\Im_3 \neq 0$. In this case Lemma 2.4 implies $\alpha\beta \neq 0$. Furthermore, we cannot have $a_2 = b_2 = 0$ because then the relations (2.6) and (2.7) would imply $\Im_1 = \Im_3 = 0$.

We first consider the case $\Im_2'' = 0$. Using $\Im_2'' = 0$ in the relation (2.12) or (2.13), one obtains $\Im_1\alpha + \Im_3\bar{\beta} = 0$. Therefore, the solution (α, β) can be looked for in the form

$$\alpha = \frac{\Im_3}{\nu}, \qquad \beta = -\frac{\bar{\Im}_1}{\bar{\nu}}, \qquad 0 \neq \nu \in \mathbf{C}.$$

We have obtained the general form of the solution and now we have to insert it into the original system of equations (2.4) - (2.5) and solve for ν. Actually, if $a_2 \neq 0$ ($b_2 \neq 0$) we use the equation e_1 (e_2). Suppose the equation e_1 has been used. After inserting the expressions for α and β in the relation (2.4), after dividing by a_2 and using Lemma 2.1(ii) we obtain $\nu^2 - \mathfrak{I}_2\nu - \bar{\mathfrak{I}}_1\mathfrak{I}_3 = 0$. Note that in the considered case the solutions ν_1 and ν_2 are real. Thus, ν satisfies the quadratic equation (2.23), which proves the lemma.

It remains to consider the case $\mathfrak{I}_2'' \neq 0$. From Lemma 2.4, we see that we can replace the unknowns α, β by ν, μ, where

$$\alpha = \frac{\mathfrak{I}_3}{\nu}, \qquad \beta = -\frac{\bar{\mathfrak{I}}_1}{\mu}, \qquad 0 \neq \nu\mu \in \mathbf{C}. \tag{2.24}$$

Note that the freedom in choosing ν and μ compensates our choice of $\mathfrak{I}_3$ and $\bar{\mathfrak{I}}_1$ in the numerators of the ratios defining α and β, respectively. By inserting α, β from (2.24) into the system (2.12) - (2.13), we obtain

$$\mathfrak{I}_1\mathfrak{I}_3(\nu - \bar{\mu} - \imath\mathfrak{I}_2'') = 0 \tag{2.25}$$

$$\mathfrak{I}_1\bar{\mathfrak{I}}_3\mu - \bar{\mathfrak{I}}_1\mathfrak{I}_3\bar{\nu} = \imath\mu\bar{\nu}\mathfrak{I}_2''. \tag{2.26}$$

Since $\mathfrak{I}_2'' \neq 0$ the solution of the system (2.25) - (2.26) solves the system (2.4) - (2.5). To solve the system (2.25) - (2.26) we divide the first equation by $\mathfrak{I}_1\mathfrak{I}_3$. We obtain

$$\mu = \bar{\nu} + \imath\mathfrak{I}_2''. \tag{2.27}$$

Using Lemma 2.1(iii) and the relation (2.27), one can rewrite the second equation (2.26) as

$$\imath\mu\bar{\nu}\mathfrak{I}_2'' = (\mathfrak{I}_1\bar{\mathfrak{I}}_3 - \bar{\mathfrak{I}}_1\mathfrak{I}_3)\mu + (\mu - \bar{\nu})\bar{\mathfrak{I}}_1\mathfrak{I}_3 = \imath\mathfrak{I}_2'\mathfrak{I}_2''\mu + \imath\mathfrak{I}_2''\bar{\mathfrak{I}}_1\mathfrak{I}_3.$$

After dividing by $\imath\mathfrak{I}_2''$ and using once more (2.27), one obtains

$$0 = \mu\bar{\nu} - \mathfrak{I}_2'\mu - \bar{\mathfrak{I}}_1\mathfrak{I}_3 = \mu(\mu - \imath\mathfrak{I}_2'') - \mathfrak{I}_2'\mu - \bar{\mathfrak{I}}_1\mathfrak{I}_3 = \mu^2 - \mathfrak{I}_2\mu - \bar{\mathfrak{I}}_1\mathfrak{I}_3 \tag{2.28}$$

To obtain the equation for ν, we use the relation (2.27). In the equation (2.28) we replace μ by $\bar{\nu} + \imath\mathfrak{I}_2''$ and then apply the complex conjugation to the obtained equation. We obtain

$$\begin{aligned} 0 &= (\nu - \imath\mathfrak{I}_2'')^2 - (\mathfrak{I}_2' - \imath\mathfrak{I}_2'')(\nu - \imath\mathfrak{I}_2'') - \mathfrak{I}_1\bar{\mathfrak{I}}_3 \\ &= \nu^2 - \mathfrak{I}_2\nu + \imath\mathfrak{I}_2'\mathfrak{I}_2'' - \mathfrak{I}_1\bar{\mathfrak{I}}_3 = \nu^2 - \mathfrak{I}_2\nu - \bar{\mathfrak{I}}_1\mathfrak{I}_3. \end{aligned}$$

Here in the last line we have used Lemma 2.1(iii). If we enumerate the solutions of the obtained equation so that the condition (2.27) is satisfied, we obtain

$$\mu_\pm = \bar{\nu}_\pm + \imath\mathfrak{I}_2'' = \left(\frac{1}{2}\bar{\mathfrak{I}}_2 \pm \frac{1}{2}\sqrt{\mathfrak{I}}\right) + \imath\mathfrak{I}_2'' = \frac{1}{2}\mathfrak{I}_2 \pm \frac{1}{2}\sqrt{\mathfrak{I}} = \nu_\pm.$$

This completes the proof of Lemma 2.5.

Now we can describe the general solution of the system (2.4) - (2.5).

Theorem 2.1 *Let the pair $(\hat{A},\hat{B})$ be definite. Then the solution (α,β) of the system (2.4) - (2.5) has the following form.*

(i) If $\Im > 0$ then $\alpha = \dfrac{\Im_3}{\nu}$, $\beta = -\dfrac{\bar{\Im}_1}{\nu}$, where ν is any nonzero solution of the equation $\nu^2 - \Im_2 \nu - \bar{\Im}_1 \Im_3 = 0$

(ii) If $\Im = 0$ then the equations in the system (2.4)–(2.5) are proportional and there is infinite number of solutions.

(a) Let $\hat{A} \neq 0$. If $|a_1| + |a_2| > 0$ then $\alpha = -\dfrac{\bar{\gamma} a_3 + a_2}{a_1 + \bar{\gamma} \bar{a}_2}$, $\beta = \gamma$, where $\gamma \in \{z \in \mathbf{C}; a_1 + \bar{z} a_2 \neq 0\}$.

If $|a_2| + |a_3| > 0$ then $\alpha = \gamma$, $\beta = -\dfrac{\bar{\gamma} a_1 + \bar{a}_2}{\bar{\gamma} a_2 + a_3}$, where $\gamma \in \{z \in \mathbf{C}; a_3 + \bar{z} a_2 \neq 0\}$.

(b) Let $\hat{B} \neq 0$. The solutions are as in the case (a) provided that a_1, a_2, a_3 are replaced by b_1, b_2, b_3, respectively.

Proof (i) This statement is proved in Lemma 2.5. Note that $\nu = 0$ is a solution of the quadratic equation iff $\bar{\Im}_1 \Im_3 = 0$. In this case the system (2.4) - (2.5) has the solutions given in Lemma 2.4 and they are obtained by the formula (2.22) with $\nu = \Im_2 = \Im_2'$.

(ii) In this case we have by Lemma 2.3 (ii) $\Im_1 = \Im_2 = \Im_3 = 0$ i. e. $\sigma \hat{A} = \omega \hat{B}$ for some real σ and ω with $|\sigma| + |\omega| > 0$. Hence the formulas from (i) (that is from (2.22)) do not exist. Setting $\beta = \gamma$ (or $\alpha = \gamma$) we can use (2.4) or (2.5) to obtain α (β). Note that the case $\hat{A} = \hat{B} = 0$ is not possible since $(\hat{A},\hat{B})$ is definite.

By Vieta's formulas, for the solutions of the equation (2.23), we have

$$\nu_+ \nu_- = -\bar{\Im}_1 \Im_3, \qquad \nu_+ + \nu_- = \Im_2. \tag{2.29}$$

Hence, the conditions $\Im > 0$ and $\bar{\Im}_1 \Im_3 = 0$ imply $\Im_2'' = 0$, $\Im_2' = \Im > 0$ and $\nu_+ = \Im_2'$, $\nu_- = 0$ or $\nu_- = \Im_2'$, $\nu_+ = 0$. Then the solution is unique: $\alpha = \Im_3 / \Im_2'$, $\beta = -\bar{\Im}_1 / \Im_2'$ with $\alpha\beta = 0$.

If $\Im > 0$ and $\bar{\Im}_1 \Im_3 \neq 0$ then we have

$$\alpha_\pm = \frac{\Im_3}{\nu_\pm} = \frac{\Im_3}{-\frac{\bar{\Im}_1 \Im_3}{\nu_\mp}} = \frac{1}{-\frac{\bar{\Im}_1}{\nu_\mp}} = \frac{1}{\beta_\mp}, \qquad (\alpha_+ \beta_+) \cdot (\alpha_- \beta_-) = 1. \tag{2.30}$$

Next we examine the cases $\Im_2' = 0$ and $\Im_2 = 0$. Recall that $\Im_2' = 0$ means that the diagonal parts of $\hat{A}$ and $\hat{B}$ are proportional while $\Im_2'' = 0$ means that a_2 and b_2 lie on a line which passes through the origin. In particular, $\Im_2' = 0$ implies that the solutions of the quadratic equation (2.23) have the same modulus.

Corollary 2.1 *Let the pair $(\hat{A},\hat{B})$ be definite and $\Im > 0$.*

(i) If $\Im_2 = 0$, then the solutions of the system (2.4)–(2.5) have the form

$$\alpha_\pm = \pm \eta \, e^{\imath\tau} \sqrt{\rho}, \qquad \beta_\pm = -\frac{1}{\alpha_\pm} = \mp \eta \, e^{-\imath\tau} / \sqrt{\rho}, \qquad \eta \in \{-1, 1\}, \tag{2.31}$$

where

$$\rho = \begin{cases} a_3/a_1 & \textit{if } a_1a_3 > 0 \\ b_3/b_1 & \textit{otherwise} \end{cases}, \; \tau = \begin{cases} \arg(a_2) & \textit{if } a_2 \neq 0 \\ \arg(b_2) & \textit{otherwise} \end{cases}, \; \eta\, e^{\imath\tau} = e^{\imath \arg(\Im_3)}. \quad (2.32)$$

(ii) If $\Im_2' = 0,\ \Im_2'' \neq 0,$ *then*

$$\alpha_\pm = e^{\imath\Theta_\pm}\sqrt{\rho}, \qquad \beta_\pm = e^{-\imath\Theta_\mp}/\sqrt{\rho}, \tag{2.33}$$

where ρ *is given by (2.32) and for the arguments* $\Theta^\pm$ *it holds*

$$\begin{aligned} \Im_2'' \sin(\arg(\alpha_+) + \arg(\beta_+)) &= \Im_2'' \sin(\Theta_+ - \Theta_-) > 0, \\ \Im_2'' \sin(\arg(\alpha_-) + \arg(\beta_-)) &= \Im_2'' \sin(\Theta_- - \Theta_+) < 0. \end{aligned}$$

Proof Let us first investigate some consequences implied by the condition $\Im_2' = 0$. This condition is equivalent to

$$\sigma \begin{bmatrix} a_1 \\ a_3 \end{bmatrix} + \omega \begin{bmatrix} b_1 \\ b_3 \end{bmatrix} = 0 \quad \text{for some } \sigma, \omega \in \mathbf{R}, \quad |\sigma| + |\omega| > 0.$$

Since $(\hat{A}, \hat{B})$ is definite, the matrix $\sigma_1\hat{A} + \omega_1\hat{B}$ is positive definite for some real σ_1 and ω_1 such that $|\sigma_1| + |\omega_1| > 0$. Obviously, the row vector $[\sigma\ \omega]$ is not proportional to $[\sigma_1\ \omega_1]$. Hence, if $\sigma \neq 0$, then $[a_1, a_3]^T = -\omega/\sigma\ [b_1, b_3]^T$ and we have

$$\left(\omega_1 - \frac{\omega}{\sigma}\sigma_1\right) \begin{bmatrix} b_1 \\ b_3 \end{bmatrix} = \sigma_1 \begin{bmatrix} a_1 \\ a_3 \end{bmatrix} + \omega_1 \begin{bmatrix} b_1 \\ b_3 \end{bmatrix} > 0.$$

We have thus proved that $\sigma \neq 0$ implies $b_1b_3 > 0$. In a similar way one can prove that $\omega \neq 0$ implies $a_1a_3 > 0$.

Since $\Im > 0$, the condition $\Im_2' = 0$ implies $\bar{\Im}_1\Im_3 > (\Im_2''/2)^2 \geq 0$. It means that $\arg(\Im_1) = \arg(\Im_3)$.

Therefore the condition $\Im_2' = 0$ implies three possible cases:

$$\left.\begin{array}{lllll} a_1a_3 > 0, & b_1 = b_3 = 0, & \Im_1 = a_1b_2, & \Im_3 = a_3b_2, & b_2 \neq 0 \\ b_1b_3 > 0, & a_1 = a_3 = 0, & \Im_1 = b_1a_2, & \Im_3 = b_3a_2, & a_2 \neq 0 \\ a_1a_3 > 0, & b_1b_3 > 0, & a_1b_3 = a_3b_1 & \text{implying} & a_3/a_1 = b_3/b_1. \end{array}\right\} \quad (2.34)$$

Hence by Lemma 2.1 (i) we have

$$\frac{\Im_3}{\Im_1} = \frac{|\Im_3|}{|\Im_1|} = \left\{\begin{array}{ll} a_3/a_1 & \text{if } a_1a_3 > 0 \\ b_3/b_1 & \text{otherwise} \end{array}\right\} = \rho.$$

The same conclusion can be drawn from the relation (2.34) and the definitions (2.6), (2.7) of $\Im_1$, $\Im_3$, respectively.

Now consider the condition $\Im_2'' = 0$. It is equivalent to

$$\sigma_2 a_2 + \omega_2 b_2 = 0 \quad \text{for some} \quad \sigma_2, \omega_2 \in \mathbf{R}, \quad |\sigma_2| + |\omega_2| > 0.$$

Hence

$$\Im_3 = a_3b_2 - b_3a_2 = \begin{cases} -a_2\left(a_3\frac{\sigma_2}{\omega_2}+b_3\right), & \omega_2 \neq 0 \\ b_2\left(a_3+\frac{\omega_2}{\sigma_2}b_3\right), & \sigma_2 \neq 0 \end{cases},$$

$$\Im_1 = a_1b_2 - b_1a_2 = \begin{cases} -a_2\left(a_1\frac{\sigma_2}{\omega_2}+b_1\right), & \omega_2 \neq 0 \\ b_2\left(a_1+\frac{\omega_2}{\sigma_2}b_1\right), & \sigma_2 \neq 0 \end{cases}.$$

This shows that a_2, b_2, $\Im_1$ and $\Im_3$ lie on the same line which passes through the origin. Let τ be as in the relation (2.32). If $\bar{\Im}_1\Im_3 > 0$ then $\Im_1$ and $\Im_3$ have the same argument. Hence from the latest relation we have

$$\Im_3 = \eta e^{\imath\tau}|\Im_3|, \quad \bar{\Im}_1 = \eta e^{-\imath\tau}|\Im_1|, \quad \eta e^{\imath\tau} = e^{\imath\arg(\Im_3)}, \quad \eta \in \{-1,1\}. \tag{2.35}$$

(i) If $\Im_2 = 0$, then we have $\bar{\Im}_1\Im_3 = \Im/4 > 0$. By Theorem 2.1 (i) one obtains

$$\alpha_\pm = \frac{\Im_3}{\pm\sqrt{\bar{\Im}_1\Im_3}} \quad \beta_\pm = -\frac{\bar{\Im}_1}{\pm\sqrt{\bar{\Im}_1\Im_3}}. \tag{2.36}$$

It follows that $\alpha_\pm\beta_\pm = -1$. From (2.36), (2.35) and $\bar{\Im}_1\Im_3 = |\Im_1||\Im_3|$ we obtain

$$\alpha_\pm = \frac{e^{\imath\arg(\Im_3)}|\Im_3|}{\pm\sqrt{|\Im_1||\Im_3|}} = \frac{\eta e^{\imath\tau}|\Im_3|}{\pm\sqrt{|\Im_1||\Im_3|}} = \pm\eta e^{\imath\tau}\sqrt{\rho}, \quad \beta_\pm = -\frac{1}{\alpha_\pm},$$

which proves the assertion (2.31).

(ii) If $\Im_2' = 0,\ \Im_2'' \neq 0$, then from the quadratic equation (2.23) for ν, we obtain

$$2\nu_\pm = i\Im_2'' \pm \sqrt{\Im}, \tag{2.37}$$

hence $|\nu_+| = |\nu_-| = \sqrt{\bar{\Im}_1\Im_3} > \sqrt{\Im}/2 > 0$ and $\bar{\nu}_\pm = -\nu_\mp$. Since $\alpha_\pm = \Im_3/\nu_\pm$, $\beta_\pm = -\bar{\Im}_1/\nu_\pm$, we have

$$|\alpha_\pm| = \sqrt{\frac{|\Im_3|}{|\Im_1|}} = \sqrt{\rho} = \frac{1}{|\beta_\pm|}, \qquad \alpha_\pm = \frac{1}{\beta_\mp},$$

where the second equation is part of the relation (2.30). So, we can set

$$\alpha_\pm = e^{i\Theta_\pm}\sqrt{\rho}, \qquad \beta_\pm = e^{-i\Theta_\mp}/\sqrt{\rho}.$$

From the relation (2.37) we obtain

$$e^{i(\Theta_\pm-\Theta_\mp)} = \alpha_\pm\beta_\pm = -\frac{\bar{\Im}_1\Im_3}{(\nu_\pm)^2} = -\frac{|\nu_\pm|^2}{(\nu_\pm)^2} = -\frac{(\nu_\mp)^2}{|\nu_\mp|^2} = -1 + \frac{(\Im_2'')^2}{4\bar{\Im}_1\Im_3} \pm \imath\frac{\Im_2''\sqrt{\Im}}{4\bar{\Im}_1\Im_3}.$$

Hence

$$\Im_2''\sin(\arg(\alpha_\pm)+\arg(\beta_\pm)) = \pm\Im_2''\sin(\Theta_+ - \Theta_-) = \pm\frac{(\Im_2'')^2\sqrt{\Im}}{4|\nu_\pm|^2},$$

which proves the remaining assertion.

Next, we provide conditions which ensure uniqueness of the solution (α,β). We choose the conditions which ensure that $|\alpha|$ and $|\beta|$ are as small as possible. This result can be used to link the complex FL method with the complex HZ method [4, 9], in order to prove the global convergence of the complex FL method.

Corollary 2.2 *Let the pair $(\hat{A},\hat{B})$ be definite and $\Im > 0$. Then the system (2.4) - (2.5) has a unique solution provided that any of the following three conditions is fulfilled.*

(i) $\Im_2' \neq 0$ *and* $|\alpha\beta| < 1$

(ii) $\Im_2' = 0,\ \Im_2'' \neq 0$ *and* $\Im_2'' \sin(\arg(\alpha) + \arg(\beta)) > 0.$

(iii) $\Im_2 = 0$ *and* $\arg(\alpha) = \begin{cases} \arg(b_2) & \text{if } b_2 \neq 0 \\ \arg(a_2) & \text{otherwise} \end{cases}$

Proof (i) Since $\Im > 0$, Theorem 2.1 implies $\alpha_\pm = \Im_3/\nu_\pm$, $\beta_\pm = -\bar{\Im}_1/\nu_\pm$, $\nu_\pm = (\Im_2' + i\Im_2'' \pm \sqrt{\Im})/2$. Here ν_+ and ν_- satisfy the quadratic equation (2.23). The relation (2.29) shows that we have $\nu_+\nu_- = -\bar{\Im}_1\Im_3$ and $\nu_+ + \nu_- = \Im_2$. We consider two cases $\bar{\Im}_1\Im_3 \neq 0$ and $\bar{\Im}_1\Im_3 = 0$.

$\bar{\Im}_1\Im_3 \neq 0$. In this case the relation (2.30) holds. Since $\Im_2' \neq 0$ and $\Im > 0$, one of the solutions ν_+ or ν_- has larger absolute value than the other. If $|\nu_+|$ $(|\nu_-|)$ is larger, we conclude from (2.30) that

$$|\alpha_+\beta_+| < 1,\ |\alpha_-\beta_-| > 1 \qquad (|\alpha_-\beta_-| < 1,\ |\alpha_+\beta_+| > 1)\,.$$

$\bar{\Im}_1\Im_3 = 0$. This case has been already considered (see the paragraph below the relation (2.29)) and we obtained the unique solution $\alpha = \Im_3/\Im_2'$, $\beta = -\bar{\Im}_1/\Im_2'$, which satisfies $\alpha\beta = 0 < 1$.

(ii) The solutions are described in Corollary 2.1(ii). We choose the $+$ solution from the relation (2.33).

(iii) In this case Corollary 2.1(i) implies $\alpha_\pm\beta_\pm = -1$ and the selected solution from (2.31) is (α_+,β_+).

2.1.1 The solutions in the case $\Im = 0$

In practice this case will rarely happen, but if not handled with care, it can cause problems, especially in the presence of rounding errors. What are reasonable choices for α and β in that case?

By Lemma 2.3(ii) the condition $\Im = 0$ is equivalent to the condition $\Im_1 = \Im_2 = \Im_3 = 0$ and also to: $s\hat{A} + t\hat{B} = 0$ for some real s and t such that $|s| + |t| > 0$. Hence $\hat{A} = -(t/s)\hat{B}$ whenever $s \neq 0$ and $\hat{B} = -(s/t)\hat{A}$ whenever $t \neq 0$. Since the pair $(\hat{A},\hat{B})$ is definite, $\sigma\hat{A} + \omega\hat{B}$ is positive definite for some real σ, ω such that $|\sigma| + |\omega| > 0$. Combining these claims we conclude that $\hat{A}$ or $\hat{B}$ has to be definite. If they both are nonzero then they both have to be definite. This implies $a_1a_3 > 0$ or $b_1b_3 > 0$ and at least one of the equations (2.4), (2.5) is nontrivial. We know that in the case $\Im = 0$ these equations are linearly dependant. So, how to solve those equations?

If $a_2 = 0$ and $b_2 = 0$ then we set $\alpha = 0$, $\beta = 0$ and proceed with the next step.

If $|a_2| + |b_2| > 0$, we know from Theorem 2.1(ii) that there is infinite set of solutions (α,β). Here are some natural choices for the solution:

$$\text{(a)} \qquad \alpha = \begin{cases} \pm\frac{a_2}{|a_2|}\frac{a_3}{a_1}, \; |a_2| \geq |b_2| \\ \pm\frac{b_2}{|b_2|}\frac{b_3}{b_1}, \; |a_2| < |b_2| \end{cases}, \qquad \beta = -\frac{1}{\alpha} = \begin{cases} \mp\frac{\bar{a}_2}{|a_2|}\frac{a_1}{a_3}, \; |a_2| \geq |b_2| \\ \mp\frac{\bar{b}_2}{|b_2|}\frac{b_1}{b_3}, \; |a_2| < |b_2| \end{cases},$$

$$\text{(b)} \qquad \alpha = \begin{cases} -\frac{a_2}{a_1}, \; |a_1| \geq |b_1| \\ -\frac{b_2}{b_1}, \; |a_1| < |b_1| \end{cases}, \; \beta = 0 \qquad \text{or} \qquad \alpha = 0, \; \beta = \begin{cases} -\frac{\bar{a}_2}{a_3}, \; |a_3| \geq |b_3| \\ -\frac{\bar{b}_2}{b_3}, \; |a_3| < |b_3| \end{cases}.$$

The first choice, described in (a), is obtained by splitting the equation (2.4) in two equations, $a_1\alpha + a_3\bar{\beta} = 0$, $\bar{a}_2\alpha\bar{\beta} + a_2 = 0$, and then solving the system. The same can be done with the equation (2.5), which gives us the possibility to choose the equation with larger coefficients.

The second choice (b) uses additional condition $\alpha \cdot \beta = 0$. This choice is more attractive to be a part of the complex Falk-Langemeyer method, because in the later stage of the iterative process when the both matrices become almost diagonal, we would like to have small α and β to ensure the quadratic convergence of the algorithm.

Hence, we may set some additional criteria for choosing the solution from the infinite set of solutions. Here they are:

(i) $\quad |\alpha| + |\beta| \to \min,$
(ii) $\quad \alpha \cdot \beta = 0,$
(iii) $\quad (\alpha, \beta)$ is determined from the pivot submatrix of the larger norm.

The first criterion ensures the smallest norm of the transformation matrix $\hat{F}$. The second one ensures the smallest flop count per step of the method. The third one ensures that (α, β) is determined by a more reliable set of input data. Typically the input data are numbers (matrix elements) that are obtained using finite arithmetic. We want them as large as possible to minimize the possibility that they are obtained by sharp cancelations in previous steps.

We see that the choice (b) of the solution complies with all listed requirements.

In particular, if $\hat{A} = 0$ then $\hat{B}$ has to be definite. This means that the first equation (2.4) is trivial (expression e_1 is zero) and we have to solve the second equation (2.5). So, we choose $\alpha = -b_2/b_1$, $\beta = 0$ if $b_1 \geq b_3$ and $\alpha = 0$, $\beta = -\bar{b}_2/b_3$ otherwise. In the case $\hat{B} = 0$, we choose $\alpha = -a_2/a_1$, $\beta = 0$ if $a_1 \geq a_3$ and $\alpha = 0$, $\beta = -\bar{a}_2/a_3$ otherwise.

Influence of rounding errors

Typically, only the computed values of $\Im_1$, $\Im_2$, $\Im_3$ and $\Im$ will be at disposal. If $\Im \approx 0$ then by Lemma 2.3(ii) we shall have $\Im_1 \approx 0$, $\Im_2 \approx 0$, $\Im_3 \approx 0$ and $\|\hat{A} - c\hat{B}\|_2 \approx 0$ for some real c. In such a situation the formulas for computing α and β are prone to large relative errors. The smaller the value of $\Im$ the larger are the relative errors in α and β computed by the standard formulas using ν. How to determine that $\Im$ is small enough to abandon the standard formulas, and how to compute the solution (α, β) ?

In the real computational process on large matrices A, B the case $\Im \approx 0$ will rarely occur. If it does happen then most likely it will appear at the end of the process in the case when the matrix pair has multiple eigenvalues. Therefore, we need a simple and cheap to compute criterion to detect whether $\Im \approx 0$. In the later stage of the process all $|a_2|$ and $|b_2|$ will be small and that will cause $|\Im_1|$, $|\Im_3|$ and $|\Im_2''|$ to be small. The remaining ingredient of $\Im$ is $\Im_2'$ and most of the time (except possibly in the beginning of the process) it determines whether $\Im$ is small. In addition, by Lemma 2.3(ii) we do not expect that small $\Im$ is implied by severe cancelation caused by the numbers $\Im_1$, $\Im_3$ and $\Im_2$, but simply because these numbers are small by absolute value. However, small values of $|\Im_1|$, $|\Im_3|$, $|\Im_2'|$ and $|\Im_2''|$ are caused by severe cancelations or by small $|a_2|$ and $|b_2|$.

Here is one suggestion what to do in general.

(i) We can compute $\|\hat{A}\|_F$ and $\|\hat{B}\|_F$. If one these norms is zero, we use the above simple formulas for α and β. In this case we have $\alpha\beta = 0$.

(ii) If $\|\hat{A}\|_F > 0$ and $\|\hat{B}\|_F > 0$, we normalize $\hat{A}$ and $\hat{B}$ as follows: compute integers $\mu_{\hat{A}}$ and $\mu_{\hat{B}}$ such that $1 \leq 2^{-\mu_{\hat{A}}}\|\hat{A}\|_F \leq 2$ and $1 \leq 2^{-\mu_{\hat{B}}}\|\hat{B}\|_F \leq 2$. Then *renormalize* $\hat{A}$, $\hat{B}$, i.e. make updates: $\hat{A} \leftarrow 2^{-\mu_{\hat{A}}}\hat{A}$ and $\hat{B} \leftarrow 2^{-\mu_{\hat{B}}}\hat{B}$. Note that $\hat{F}$ is invariant under that transformation because it simultaneously diagonalizes $\hat{A}$ and $\hat{B}$ if and only if it simultaneously diagonalizes $2^{-\mu_{\hat{A}}}\hat{A}$ and $2^{-\mu_{\hat{B}}}\hat{B}$.

(iii) Next we compute $\Im$. If $\Im$ is positive and not too tiny, then we compute α and β by the standard formulas using ν. If $\Im$ is negative and $|\Im|$ is not too tiny, then we consider the pair $(\hat{A},\hat{B})$ is not definite and abort the computation. Finally, if $|\Im|$ is tiny, we apply a special procedure to determine how α and β should be computed. Here a tiny $|\Im|$ means a modest multiple of the *unit round-off* (*machine epsilon*) $\mathbf{u}$ multiplied by some reasonable upper bound of $|\Im|$.

The rest of this subsection is devoted to designing that special procedure. The procedure has to determine whether the pair $(\hat{A},\hat{B})$ is definite and how to compute the solution. Such a procedure can be an important part of the CFL algorithm. If the initial pair (A,B) is known to be definite, but it has tiny Crawford constant $c(A,B)$, then the rounding errors can ruin the definiteness of the iterated pair (see [18,19]). By $\mathrm{fl}(x)$ we denote the computed value of x.

Let $\Im_1 = \Im_1' + \imath\Im_1''$, $\Im_3 = \Im_3' + \imath\Im_3''$ and $a_2 = a_2' + \imath a_2''$, $b_2 = b_2' + \imath b_2''$. We have

$$
\begin{aligned}
|\Im| &= |(\Im_2' - \Im_2'')(\Im_2' + \Im_2'') + 4\mathrm{Re}(\bar{\Im}_1\Im_3)| \leq \max\{(\Im_2')^2, (\Im_2'')^2\} + 4|\Im_1'\Im_3' + \Im_1''\Im_3''| \\
&\leq \max\{(|a_1b_3| + |b_1a_3|)^2, 4(|a_2'b_2''| + |a_2''b_2'|)^2\} \\
&\qquad + 4[|a_1a_3||b_2|^2 + |b_1b_3||a_2|^2 + (|a_1b_3| + |b_1a_3|)(|a_2'b_2'| + |a_2''b_2''|)|] \equiv \rho.
\end{aligned}
$$

We consider ρ a reasonable upper bound for $|\mathrm{fl}(\Im)|$. Let ε be a modest multiple of $\mathbf{u}$ (say of $\mathbf{u} \leq \varepsilon \leq 10\mathbf{u}$). Its optimal value can be determined by numerical tests.

If $\mathrm{fl}(\Im) < -\rho\varepsilon$ we consider the initial pair not definite and abort the computation.

If $|\Im|$ is tiny, say of order $\mathbf{u}$ or less, then $\mathrm{fl}(\Im)$ as an approximation of $\Im$ will have large relative error, and in our analysis we shall also use $\mathrm{fl}(\Im)$. Later in the statements of the algorithm $\mathrm{fl}(\Im)$ and $\Im$ will mean the same. Recall that $\Im = 0$ implies $\Im_r = 0$ for all $1 \leq r \leq 3$. Now, if all $\Im_r$ are of order ε, then $|\Im|$ will be of order ε^2. Therefore, if $\rho\varepsilon^2 \leq \mathrm{fl}(\Im)$, we can employ the standard formulas for α, β which use ν.

If $\mathrm{fl}(\Im)$ lies in the interval $(0,\rho\varepsilon^2)$, then severe cancelation(s) take place and the computed ν, α and β will have large relative errors. If $\mathrm{fl}(\Im) \in (-\rho\varepsilon^2, 0)$ we can still speculate that the rounding errors have caused $\mathrm{fl}(\Im)$ to be negative. Therefore, the question arises how else can we compute the solution (α,β) when $|\Im|$ is that tiny?

By adopting the criterions (i)–(ii), we can assume $\alpha\beta = 0$. If $\beta = 0$ the equations (2.4) and (2.5) make a system of linear equations $a_1\alpha = -a_2$, $b_1\alpha = -b_2$ and we can look for the least square (LS) solution.

Let $\tilde{a}_1 = \sqrt{a_1^2 + b_1^2},\ c_1 = a_1/\tilde{a}_1,\ s_1 = b_1/\tilde{a}_1.$ We obtain

$$\| \begin{bmatrix} a_1 \\ b_1 \end{bmatrix} \alpha + \begin{bmatrix} a_2 \\ b_2 \end{bmatrix} \|_2^2 = \| \begin{bmatrix} \tilde{a}_1 \\ 0 \end{bmatrix} \alpha + \begin{bmatrix} c_1 & s_1 \\ -s_1 & c_1 \end{bmatrix} \begin{bmatrix} a_2 \\ b_2 \end{bmatrix} \|_2^2 = \left| \tilde{a}_1 \alpha + \frac{a_1 a_2 + b_1 b_2}{\tilde{a}_1} \right|^2 + \frac{|\Im_1|^2}{a_1^2 + b_1^2},$$

where $\| \cdot \|_2$ stands for the Euclidean vector norm. The solution is

$$\alpha = -\frac{a_1 a_2 + b_1 b_2}{a_1^2 + b_1^2} \qquad \text{with the residual error} \qquad \frac{|\Im_1|}{\sqrt{a_1^2 + b_1^2}}.$$

Since α is a convex sum of $-a_2/a_1$ and $-b_2/b_1$ it lies on the line segment connecting these two points in the complex plane.

The case $\alpha = 0$ is treated in the similar way. We obtain

$$\beta = -\frac{a_3 \bar{a}_2 + b_3 \bar{b}_2}{a_3^2 + b_3^2} \qquad \text{with the residual error} \qquad \frac{|\Im_3|}{\sqrt{a_3^2 + b_3^2}},$$

and β lies on the line segment connecting $-\bar{a}_2/a_3$ and $-\bar{b}_2/b_3$. The above considerations lead us to the following algorithm for computing the solution (α, β):

```
if $\frac{|\Im_1|}{\sqrt{a_1^2 + b_1^2}} \le \frac{\Im_3|}{\sqrt{a_3^2 + b_3^2}}$ then $\alpha = -\frac{a_1 a_2 + b_1 b_2}{a_1^2 + b_1^2},\ \beta = 0$
                                         else $\alpha = 0,\ \beta = -\frac{a_3 \bar{a}_2 + b_3 \bar{b}_2}{a_3^2 + b_3^2}$
endif
```

(2.38)

Note that due to the definiteness of the pair $(\hat{A}, \hat{B})$, we should have $a_1^2 + b_1^2 > 0$ and $a_3^2 + b_3^2 > 0$. If due to rounding errors, $|a_1| + |b_1| = 0$ and $|a_3| + |b_3| > 0$ or vice versa, we can still use the formulas (2.38).

If $\hat{A} = c\hat{B}$ $(\hat{B} = c\hat{A})$ for some real c then the algorithm (2.38) reduces to

```
if $|a_1| + |b_1| \ge |a_3| + |b_3|$ then $\alpha = -a_2/a_1 \ \ (= -b_2/b_1),\ \ \beta = 0$
                            else $\alpha = 0,\ \beta = -\bar{a}_2/a_3 \ \ (= -\bar{b}_2/b_3)$
endif
```

(2.39)

which conforms with the choice (b) of the solution. Here, we have replaced the void condition $(0 \le 0)$ in the algorithm (2.38) by $|a_1| + |b_1| \ge |a_3| + |b_3|$ which reduces to $|a_1| \ge |a_3|$ $(|b_1| \ge |b_3|)$ and ensures that the first requirement $|\alpha| + |\beta| \to \min$ is fulfilled.

The solution of the LS problem is attractive if $\mathrm{fl}(\Im)$ lies in the interval $(-\rho\varepsilon^2, \rho\varepsilon^2)$ because then the residuals are small. Our strategy is to employ it in that case. The narrative is as follows. The rounding errors can cause $\mathrm{fl}(\Im)$ to lie somewhere in the interval $(-\rho\mathbf{u}, 0)$ even if the pair $(\hat{A}, \hat{B})$ is definite. If the pair (A, B) of large matrices is not definite, then we have probably detected a pair of pivot submatrices which is not definite. However, we do not want to neglect the possibility that (A,B) is definite with small Crawford constant. So, if $\mathrm{fl}(\Im) \in (-\rho\varepsilon^2, 0)$, instead of terminating the iteration process, we would rather use the LS solution. If $\mathrm{fl}(\Im) \in (-\rho\varepsilon, -\rho\varepsilon^2)$ it is more likely that $(\hat{A},\hat{B})$ and hence (A,B) is not definite. In such a situation the use of the LS solution will only postpone revealing of that fact.

We cannot say whether the LS solution will accelerate or decelerate the revealing of the fact that the pair (A,B) is not definite. Maybe a finer error analysis could offer an answer. Finally, let us say that in the case $\mathrm{fl}(\Im) \in (\rho\mathbf{u}^2, \rho\mathbf{u})$, we have given preference to the standard solution because numerical tests have shown that it yields smaller residual.

Remark 2.1 If $\Im$ is tiny then the matrices $\hat{A}$, $\hat{B}$ are nearly proportional. A simple calculation shows that the problem

$$\min_{t\in\mathbf{R}} \|\hat{A} - t\hat{B}\|_F \;\rightarrow\; \min \quad \text{has the solution} \quad t^* = \frac{\mathrm{trace}(\hat{A}\hat{B})}{\mathrm{trace}(\hat{B}\hat{B})}.$$

Since for the both matrices we have $1 \leq \|\hat{A}\|_F < 2$ and $1 \leq \|\hat{B}\|_F < 2$, there is no need to additionally consider the associated problem $\min_{t\in\mathbf{R}} \|\hat{B} - t\hat{A}\|_F \;\rightarrow\; \min$.

Hence, instead of checking $|\mathrm{fl}(\Im)| \leq \rho\varepsilon$ or $|\mathrm{fl}(\Im)| \leq \rho\varepsilon^2$, one can alternatively check whether the condition $\|\hat{A} - t^*\hat{B}\|_F \leq \varepsilon\,\|\hat{A}\|$ holds. Hence, if

$\|\hat{A} - t^*\hat{B}\|_F > \varepsilon\,\|\hat{A}\|$ and $\Im < 0$, we can consider that the pair $(\hat{A},\hat{B})$ is not definite

$\|\hat{A} - t^*\hat{B}\|_F > \varepsilon\,\|\hat{A}\|$ and $\Im > 0$, we apply the standard procedure which uses ν

$\|\hat{A} - t^*\hat{B}\|_F \leq \varepsilon\,\|\hat{A}\|$ holds, then we can consider using the alternative formulas (2.38) or (2.39).

This alternative approach for the special procedure seems more attractive because it uses matrix elements in a less complicated manner than using ρ. However, our first numerical tests do not confirm it. What approach is better and how to define ε can probably be resolved in practice through extensive testing.

2.2 The complex Falk-Langemeyer algorithm

We can now write down a pseudo code for the CFL method for a definite pair of Hermitian matrices (A,B) where A and B have dimension n. However, we first make few remarks.

If F is a nonsingular matrix, then the pair (F^*AF, F^*BF) is also definite. So, we have to ensure that each elementary transformation matrix F_k, $k \geq 1$, from the relation (2.1) is nonsingular.

If $\tilde{A},\tilde{B}$ are the principal submatrices of A,B, obtained on the intersection of the same rows and columns, then the pair $(\tilde{A},\tilde{B})$ is definite. This is a consequence of the fact that any principal submatrix of a positive definite matrix is positive definite. So, if the initial pair (A,B) is definite and all the transformation matrices are nonsingular, then each pair of the pivot submatrices $(\hat{A},\hat{B})$ will be definite.

Finally, if $\Im = 0$ is computed from a definite pair $(\hat{A},\hat{B})$, then either the pivot submatrices $\hat{A}$ and $\hat{B}$ are proportional or one of them is zero. In the both cases the nontrivial submatrix is definite. This follows from Lemma 2.3(ii).

Pivot strategy can be chosen in a number of ways. The choice depends on the machine architecture. On conventional computers a good choice is the column– or row–cyclic strategy with possible modifications (cf. [15]).

To ensure faster asymptotic convergence and to have a better insight into the structure which lies within almost diagonal iterated matrices [7,6], we would need the diagonal elements to be specially arranged. Those a_{rr} and b_{rr} for which a_{rr}/b_{rr} approximates the same eigenvalue of (A,B) should occupy successive positions along the diagonal. This can be accomplished by requiring that the quotients $a^{(k+1)}_{i(k)i(k)}/b^{(k+1)}_{i(k)i(k)}$ and $a^{(k+1)}_{j(k)j(k)}/b^{(k+1)}_{j(k)j(k)}$ are always in the prescribed order, say

$$\frac{a^{(k+1)}_{i(k)i(k)}}{b^{(k+1)}_{i(k)i(k)}} \geq \frac{a^{(k+1)}_{j(k)j(k)}}{b^{(k+1)}_{j(k)j(k)}}, \quad k \geq 1.$$

In numerical code it actually means that a check should be made whether the columns of the pivot submatrix $\hat{F}_k$ have to be swapped.

Performing the k^{th} step can include a call to a subroutine similar to the BLAS1 routine ROT.

Convergence criterion requires a thorough investigation. If one of the matrices, A or B is positive definite, and the other is nonsingular then the choice from [4,1] might be a good try. It says to stop the iteration when

$$|a^{(M)}_{pq}| \leq \text{tol} \cdot \sqrt{|a^{(M)}_{pp} a^{(M)}_{qq}|}, \quad |b^{(M)}_{pq}| \leq \text{tol} \cdot \sqrt{|b^{(M)}_{pp} b^{(M)}_{qq}|}, \qquad 1 \leq p < q \leq n. \tag{2.40}$$

This check is typically made at the end of each sweep, i.e. after every batch of $N = n(n-1)/2$ steps. Here tol is a prescribed tolerance (say $tol = c\mathbf{u}$ where c is a modest constant or a slowly growing function of n) and $\mathbf{u}$ is the machine epsilon. The stopping criterion (2.40) will warrant high relative accuracy (HRA) of the computed eigenvalues if the both matrices are positive definite and the matrix pair is well-behaved (see Theorem 3.1). The simplest version of the CFL algorithm is presented below. We assumed $\varepsilon = \mathbf{u}$. We also assume that $\operatorname{sgn}(0) = 1$.

Algorithm 1 *(CFL algorithm) Input data are Hermitian matrices A, B of order n and the logical variable* eivec *whose value determines whether the eigenvectors are to be computed. Output data are almost diagonal matrices A and B obtained by the method (after the convergence criterion has been reached) and, if* eivec *has value true, the matrix F whose columns are approximations of the eigenvectors of* (A,B).

```
1^0 Set $k = 1$, $A^{(k)} = A$, $B^{(k)} = B$. If eivec then set $F^{(k)} = I_n$
2^0 Repeat
    (a) Choose the pivot pair $(i,j)$ $(= (i(k), j(k)))$
    (b) Compute the parameters $(\alpha_k, \beta_k)$ of the transformation matrix $F_k$
    (c) Compute $A^{(k+1)} = F_k^* A^{(k)} F_k$, $B^{(k+1)} = F_k^* B^{(k)} F_k$;
        if eivec then compute $F^{(k+1)} = F^{(k)} F_k$
  until convergence.
```

Algorithm 2 *(2^0 **(b)** part of CFL algorithm, the superscript (k) is omitted and $a'_{ij} = Re(a_{ij})$, $a''_{ij} = Im(a_{ij})$, $b'_{ij} = Re(b_{ij})$, $b''_{ij} = Im(bij)$ is used. The value -1 of the variable job indicates that the computation should be terminated.)*

```
if $|a_{ij}| + |b_{ij}| = 0$ then $\alpha = \beta = 0$ else
(i) renormalize $\hat{A}$, $\hat{B}$ and compute:
    $\Im'_{ij} = a_{ii}b_{jj} - a_{jj}b_{ii};\quad \Im''_{ij} = -2\,(a'_{ij}b''_{ij} - b'_{ij}a''_{ij});\quad \Im_{ij} = \Im'_{ij} + \imath\Im''_{ij};$
    $\Im_i = a_{ii}b_{ij} - a_{ij}b_{ii}; \Im_j = a_{jj}b_{ij} - a_{ij}b_{jj}; \Im = (\Im'_{ij} - \Im''_{ij})\,(\Im'_{ij} + \Im''_{ij}) + 4Re(\bar{\Im}_1\Im_3);$
        $\rho = \max\{(|a_{ii}b_{jj}| + |b_{ii}a_{jj}|)^2, 4(|a'_{ij}b''_{ij}| + |a''_{ij}b'_{ij}|)^2\} +$
          $4\left[|a_{ii}a_{jj}||b_{ij}|^2 + |b_{ii}b_{jj}||a_{ij}|^2 + (|a_{ii}b_{jj}| + |b_{ii}a_{jj}|)(|a'_{ij}b'_{ij}| + |a''_{ij}b''_{ij}|)\right];$
(ii) set job = 0;
    If $\Im > \rho\mathbf{u}^2$ then $\nu = (\Im_{ij} + \mathrm{sgn}(\Im'_{ij})\sqrt{\Im})/2,\ \alpha = \Im_j/\nu,\ \beta = -\bar{\Im}_i/\nu$
    elseif $\Im < -\rho\mathbf{u}$ then job $= -1$
    else if $|\Im_i|\sqrt{a_{jj}^2 + b_{jj}^2} \le |\Im_j|\sqrt{a_{ii}^2 + b_{ii}^2}$
            then  $\alpha = -(a_{ii}a_{ij} + b_{ii}b_{ij})/(a_{ii}^2 + b_{ii}^2),\quad \beta = 0$
            else  $\alpha = 0,\ \beta = -(a_{jj}\bar{a}_{ij} + b_{jj}\bar{b}_{ij})/(a_{jj}^2 + b_{jj}^2)$
         endif
    endif
endif
```

The transformation formulas for the diagonal elements are obtained straightforwardly. We have

$$
\begin{aligned}
a_{ii}^{(k+1)} &= a_{ii}^{(k)} + (|\beta_k|^2 a_{jj}^{(k)} + 2Re(\beta_k a_{ij}^{(k)})),\ \ b_{ii}^{(k+1)} = b_{ii}^{(k)} + (|\beta_k|^2 b_{jj}^{(k)} + 2Re(\beta_k b_{ij}^{(k)})),\\
a_{jj}^{(k+1)} &= a_{jj}^{(k)} + (|\alpha_k|^2 a_{ii}^{(k)} + 2Re(\alpha_k \bar{a}_{ij}^{(k)})),\ \ b_{jj}^{(k+1)} = b_{jj}^{(k)} + (|\alpha_k|^2 b_{ii}^{(k)} + 2Re(\alpha_k \bar{b}_{ij}^{(k)})).
\end{aligned}
$$

The question arises whether it is better to set the pivot elements $a_{ij}^{(k+1)}$ and $b_{ij}^{(k+1)}$ to zero or to compute them. Numerical tests have confirmed that it is better to compute them. The formulas are below

$$
\begin{aligned}
a_{ij}^{(k+1)} &= a_{ij}^{(k)} + (\alpha_k\bar{\beta}_k\bar{a}_{ij}^{(k)} + (\bar{\beta}_k a_{jj}^{(k)} + \alpha_k a_{ii}^{(k)})),\\
b_{ij}^{(k+1)} &= b_{ij}^{(k)} + (\alpha_k\bar{\beta}_k\bar{b}_{ij}^{(k)} + (\bar{\beta}_k b_{jj}^{(k)} + \alpha_k b_{ii}^{(k)})).
\end{aligned}
$$

We have used parentheses to ensure that the updates have the form: new value equals to the old value plus the update. This contributes to the accuracy of the algorithm.

The whole process can be performed in the upper-triangular parts of the complex matrices/arrays $A^{(k)}$ and $B^{(k)}$. We provide the appropriate formulas below.

$$
\left.\begin{array}{ll}
a_{ri}^{(k+1)} = a_{ri}^{(k)} + \beta_k a_{rj}^{(k)}, & b_{ri}^{(k+1)} = b_{ri}^{(k)} + \beta_k b_{rj}^{(k)}\\
a_{rj}^{(k+1)} = a_{rj}^{(k)} + \alpha_k a_{ri}^{(k)}, & b_{rj}^{(k+1)} = b_{rj}^{(k)} + \alpha_k b_{ri}^{(k)}
\end{array}\right\}\quad 1 \le r \le i-1,
$$

$$
\left.\begin{array}{ll}
a_{ir}^{(k+1)} = a_{ir}^{(k)} + \bar{\beta}_k \bar{a}_{rj}^{(k)}, & b_{ir}^{(k+1)} = b_{ir}^{(k)} + \bar{\beta}_k \bar{b}_{rj}^{(k)}\\
a_{rj}^{(k+1)} = a_{rj}^{(k)} + \alpha_k \bar{a}_{ir}^{(k)}, & b_{rj}^{(k+1)} = b_{rj}^{(k)} + \alpha_k \bar{b}_{ir}^{(k)}
\end{array}\right\}\quad i+1 \le r \le j-1,
$$

$$\left.\begin{array}{ll} a_{ir}^{(k+1)} = a_{ir}^{(k)} + \bar{\beta}_k a_{jr}^{(k)}, & b_{ir}^{(k+1)} = b_{ir}^{(k)} + \bar{\beta}_k b_{jr}^{(k)} \\ a_{jr}^{(k+1)} = a_{jr}^{(k)} + \bar{\alpha}_k a_{ir}^{(k)}, & b_{jr}^{(k+1)} = b_{jr}^{(k)} + \bar{\alpha}_k b_{ir}^{(k)} \end{array}\right\} \qquad j+1 \le r \le n,$$

If some of the sets $\{r : 1 \le r \le i-1\}$, $\{r : i+1 \le r \le j-1\}$, $\{r : j+1 \le r \le n\}$ are empty, the corresponding updates of the off-diagonal elements are skipped.

Finally, if the eigenvectors are wanted (the variable eivec has value true), then the update of the matrix $F^{(k)} = (f_{rs}^{(k)})$ has the form

$$f_{ri}^{(k+1)} = f_{ri}^{(k)} + \beta_k f_{rj}^{(k)}, \qquad f_{rj}^{(k+1)} = f_{rj}^{(k)} + \alpha_k f_{ri}^{(k)}, \qquad 1 \le r \le n.$$

Note that each Hermitian matrix H can be represented as $H = \mathrm{Re}(H) + \imath\,\mathrm{Im}(H)$, where the real matrices $\mathrm{Re}(H)$ and $\mathrm{Im}(H)$ are symmetric and skew-symmetric, respectively. Hence H can be represented by the real matrix $\mathbf{H}$ of order n which has, say, $\mathrm{Re}(H)$ in its upper triangle and $\mathrm{Im}(H)$ in its strictly lower triangle. Using the above formulas, one can devise a real algorithm for the complex Falk-Langemeyer method which uses the appropriate real matrices $\mathbf{A}$ and $\mathbf{B}$ instead of A and B, respectively.

In the following proposition we assume the exact (infinite) arithmetic, which means that the LS solution (actually, its derivative, algorithm (2.39)) is used only when $\mathfrak{I} = 0$.

Proposition 2.1 *Let (A,B) be a definite pair of Hermitian matrices and let $(A^{(k)},B^{(k)})$, $k \ge 1$ be the sequence of pairs generated by applying the CFL algorithm to (A,B). Then for each $k \ge 1$ the following assertions hold:*

(i) F_k *is nonsingular*

(ii) $|\alpha_k\beta_k| \le 1$

(iii) $|\alpha_k\beta_k| = 1$ *if and only if* $Re(\mathfrak{I}_{ij}^{(k)}) = 0$ *and* $|a_{ij}^{(k)}| + |b_{ij}^{(k)}| > 0.$

We also have $\alpha_k\beta_k = -1$ *if and only if* $\mathfrak{I}_{ij}^{(k)} = 0.$

Proof Choose any $k \ge 1$ and set $i = i(k), j = j(k)$. In this proof we shall omit the superscript (k) and use $\mathfrak{I}_{ij} = \mathfrak{I}'_{ij} + \imath\mathfrak{I}'_{ij}$.

(i) Note that F_k is singular if and only if $\alpha_k\beta_k = 1$. If $a_{ij} = 0 = b_{ij}$, then $\alpha_k = \beta_k = 0$ and $F_k = I_n$. The algorithm does not break for a definite pair of matrices. Hence $\mathfrak{I} \ge 0$. If $\mathfrak{I} = 0$, then by the algorithm $\alpha_k\beta_k = 0$ and $\det(F_k) = 1$. If $\mathfrak{I} > 0$, then we have the following chain of equivalent statements

$$\begin{aligned} \alpha_k\beta_k = 1 &\Leftrightarrow -\bar{\mathfrak{I}}_i\mathfrak{I}_j = \nu^2 \Leftrightarrow -4\bar{\mathfrak{I}}_i\mathfrak{I}_j = \mathfrak{I}_{ij}^2 + \mathfrak{I} + 2\mathfrak{I}_{ij}\mathrm{sgn}(\mathfrak{I}'_{ij})\sqrt{\mathfrak{I}} \\ &\Leftrightarrow 2\mathfrak{I} + 2|\mathfrak{I}'_{ij}|\sqrt{\mathfrak{I}} + \imath 2\mathrm{sgn}(\mathfrak{I}'_{ij})\mathfrak{I}''_{ij}\sqrt{\mathfrak{I}} = 0 \\ &\Leftrightarrow \mathfrak{I} = 0. \end{aligned}$$

Hence $\alpha_k\beta_k = 1$ and $\mathfrak{I} > 0$ yield a contradiction. This shows that in all cases F_k is nonsingular.

(ii) If $|a_{ij}| + |b_{ij}| = 0$ we have $\alpha_k = \beta_k = 0$. If $\mathfrak{I} = 0$, then by the algorithm $\alpha_k\beta_k = 0$. If $\mathfrak{I} > 0$, then by the relation (2.30) $|\alpha_k^+\beta_k^+| \cdot |\alpha_k^-\beta_k^-| = 1$ and the algorithm chooses $|\nu_k| = \max\{|\nu_k^+|, |\nu_k^-|\}$. Hence $|\alpha_k\beta_k| \le 1$.

(iii) Obviously, we must have $\Im > 0$. By Lemma 2.4 we know that $|\Im_i| \cdot |\Im_j| = 0$ implies $\alpha_k\beta_k = 0$. So, it remains to consider the case $|\Im_i| \cdot |\Im_j| \neq 0$. Then Vieta's formulas (2.29) imply that $\nu_k^+ \neq 0$, $\nu_k^- \neq 0$ so we have $|\alpha_k^+\beta_k^+| \cdot |\alpha_k^-\beta_k^-| = 1$. The algorithm chooses ν_k such that $|\nu_k| = \max\{|\nu_k^+|, |\nu_k^-|\}$ implying $|\alpha_k\beta_k| \leq 1$. Hence we have $|\alpha_k\beta_k| = 1$ if and only if $|\nu_k^+| = |\nu_k^-|$. From $2\nu_k^{\pm} = \Im'_{ij} \pm \sqrt{\Im} + \imath\Im''_{ij}$ and $\Im > 0$, we conclude that $|\nu_k^+| = |\nu_k^-|$ if and only if $\Im'_{ij} = 0$. Thus $|\alpha_k\beta_k| = 1$ implies $\Im'_{ij} = 0$. Conversely, from the algorithm we see that $\Im'_{ij} = 0$ and $\Im > 0$ imply $|\alpha_k\beta_k| = 1$.

To prove the last assertion, note that Corollary 2.1(i) states that $\Im_{ij} = 0$ implies $\alpha_k\beta_k = -1$. Let $\alpha_k\beta_k = -1$. We have already proved that then we must have $\Im'_{ij} = 0$. Now, the relation $\bar{\Im}_i\Im_j = \nu_{\pm}^2$ implies, after simple calculation, $(\Im''_{ij})^2 = \pm\imath\Im''_{ij}\sqrt{\Im}$ which implies $\Im''_{ij} = 0$. Hence, we have $\Im_{ij} = 0$.

Finally, note that the CFL has a nice property with respect to the congruence transformation with a diagonal matrix. Let $D = \mathrm{diag}(d_1, \ldots d_n)$ be nonsingular. Suppose (A, B) is the current matrix pair, (i, j) is the current pivot pair and F is the transformation matrix. Let the pivot submatrices of A, B, F and D be denoted $\hat{A}$, $\hat{B}$, $\hat{F}$ and $\hat{D} = \mathrm{diag}(d_i, d_j)$, respectively. Consider the current step of the CFL method on (A, B) and on $(\tilde{A}, \tilde{B}) = (D^*AD, D^*BD)$. We shall apply tilde to the quantities associated with $(\tilde{A}, \tilde{B})$. An easy calculation reveals

$$\tilde{\Im}_i = |d_i|^2\bar{d}_i d_j\Im_i, \qquad \tilde{\Im}_j = \bar{d}_i d_j|d_j|^2\Im_j, \qquad \tilde{\Im}'_{ij} = |d_i|^2|d_j|^2\Im'_{ij},$$
$$\tilde{\Im}''_{ij} = |d_i|^2|d_j|^2\Im''_{ij}, \qquad \tilde{\Im} = |d_i|^4|d_j|^4\Im.$$

If $\Im > 0$, we have

$$\tilde{\nu} = |d_i|^2|d_j|^2\nu, \qquad \tilde{\alpha} = \frac{d_j}{d_i}\alpha, \qquad \tilde{\beta} = \frac{d_i}{d_j}\beta, \qquad \text{hence} \qquad \tilde{\alpha}\tilde{\beta} = \alpha\beta.$$

This property is in accordance with the relation

$$\hat{D}\hat{\hat{F}} = \begin{bmatrix} d_i & \\ & d_j \end{bmatrix} \begin{bmatrix} 1 & \tilde{\alpha} \\ \tilde{\beta} & 1 \end{bmatrix} = \begin{bmatrix} 1 & \frac{d_i}{d_j}\tilde{\alpha} \\ \frac{d_j}{d_i}\tilde{\beta} & 1 \end{bmatrix} \begin{bmatrix} d_i & \\ & d_j \end{bmatrix} = \begin{bmatrix} 1 & \alpha \\ \beta & 1 \end{bmatrix} \begin{bmatrix} d_i & \\ & d_j \end{bmatrix} = \hat{F}\hat{D},$$

which says that *any diagonal congruence transformation of A and B after (prior to) the current CFL step can be moved prior to (after) the step, provided the transformation is updated in a fair way.*

3 Numerical Tests

Here we present several experiments in MATLAB which deal with HRA of the CFL algorithm. The tests have been made on a PC with Intel(R) Core(TM) i7-2620M CPU and with 8GB installed memory, under the 64-bit operating system Windows 8.1 Enterprise, using MATLAB R2016b.

Our goal is to check numerically whether the derived method can compute the eigenvalues of a pair of positive definite matrices with HRA. First, we have to find some class of "well-behaved" matrix pairs. Roughly speaking, a well behaved pair of matrices is the pair that allows only small relative perturbations of the eigenvalues and eigenvectors if the perturbation matrices are sufficiently small in some norm. Such a pair of matrices obviously has additional properties and its perturbations are also somewhat special. Once we find a well-behaved pair, we can apply to it the method and see how accurately the eigenvalues are computed. A method is HRA on that pair if it generates (in finite arithmetic, at each step and cumulatively) errors that belong to that special kind of perturbations.

Our choice of well-behaved pairs is based on the result of Drmač [2][Theorem 3.2]. It was originally formulated for the case of real positive definite matrices, but it extends straightforwardly to the case of complex Hermitian positive definite matrices. We present it here in the compact form, as in [9, Theorem 5.1].

Theorem 3.1 *([2, Theorem 3.2]) Let A and B be Hermitian positive definite matrices of order n and let* $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ *be the eigenvalues of the pair* (A,B)*. Let* $A_S = D_A^{-1/2} A D_A^{-1/2}$, $B_S = D_B^{-1/2} B D_B^{-1/2}$*, where* $D_A = diag(A)$, $D_B = diag(B)$*. Let* δA *and* δB *be Hermitian perturbations and* $\tilde{\lambda}_1 \geq \tilde{\lambda}_2 \geq \cdots \geq \tilde{\lambda}_n$ *be the eigenvalues of the pair* $(A+\delta A, B+\delta B)$*. Let* $(\delta A)_S = D_A^{-1/2} \delta A D_A^{-1/2}$, $\varepsilon_{A_S} = \|(\delta A)_S\|_2 / \|A_S\|_2$ *and* $(\delta B)_S = D_B^{-1/2} \delta B D_B^{-1/2}$, $\varepsilon_{B_S} = \|(\delta B)_S\|_2 / \|B_S\|_2$*. If*

$$\varepsilon_{A_S} \kappa_2(A_S) = \|(\delta A)_S\|_2 \|A_S^{-1}\|_2 < 1 \quad \textit{and} \quad \varepsilon_{B_S} \kappa_2(B_S) = \|(\delta B)_S\|_2 \|B_S^{-1}\|_2 < 1,$$

then

$$\max_{1 \leq i \leq n} \frac{|\tilde{\lambda}_i - \lambda_i|}{\lambda_i} \leq \frac{\varepsilon_{A_S} \kappa_2(A_S) + \varepsilon_{B_S} \kappa_2(B_S)}{1 - \varepsilon_{B_S} \kappa_2(B_S)}. \tag{3.1}$$

From the theorem it follows that one class of well-behaved pairs is comprised of the pairs of Hermitian positive definite matrices that can be well scaled symmetrically, i.e. for which $\kappa_2(A_S)$ and $\kappa_2(B_S)$ are small numbers. In addition, if the perturbations matrices can be well scaled symmetrically, i.e. if ε_{A_S} and ε_{B_S} are small, then the relative perturbations in all eigenvalues will be small.

Next, we have to find whether the CFL method generates small $\varepsilon_{A_S^{(k)}} \kappa_2(A_S^{(k)})$ and $\varepsilon_{B_S^{(k)}} \kappa_2(B_S^{(k)})$ in each step. Such a proof requires a detailed rounding error analysis which is a demanding task. In each step the method generates errors in α_k, β_k and in the affected matrix elements. From those errors one can form the perturbation matrices. We can denote them $\delta A^{(k)}$ and $\delta B^{(k)}$. The rounding error analysis is also used to show that the perturbation matrices appearing in the process can be moved back, in some way, to $A^{(0)}$ and $B^{(0)}$. That procedure is called *backward error analysis*. Once, all perturbation matrices appearing in the process are moved back to the initial matrices $A^{(0)}$ and $B^{(0)}$ they can be added together to obtain the *accumulated perturbations* or backward errors of $A^{(0)}$ and $B^{(0)}$. We can call them δA and δB as those in the theorem. These are the perturbations that perturb the eigenvalues and eigenvectors of (A,B). Then Theorem 3.1 can be applied to the pair (A,B) and to the accumulated perturbations. If we could estimate the corresponding ε_{A_S} and ε_{B_S} we could conclude whether the method has HRA property.

Applying the Cauchy-Schwarz inequality to the numerator on the right-hand side of the relation (3.1), we obtain

$$\rho_{(A,B)} = \max_{1\leq i\leq n} \frac{|\tilde{\lambda}_i - \lambda_i|}{\lambda_i} / \sqrt{\kappa_2^2(A_S) + \kappa_2^2(B_S)} \leq \frac{\sqrt{\varepsilon_{A_S}^2 + \varepsilon_{B_S}^2}}{1 - \varepsilon_{B_S}\kappa_2(B_S)}. \tag{3.2}$$

If we could prove that the accumulated perturbations δA and δB can be well-scaled symmetrically so that the corresponding ε_{A_S} and ε_{B_S} are tiny, then we would conclude that the method is HRA. Namely, in that case the quantity $\rho_{(A,B)}$ from the relation (3.2) would be tiny. We expect that

$$\rho_{(A,B)} \leq f(n)\,\mathbf{u} \tag{3.3}$$

would hold, where $f(n)$ is a slowly growing function of n. *Hence a strong indication that the method is HRA will be the fact that the relation (3.3) holds for a larger sample of matrix pairs from our class of well-behaved pairs.* We shall call such a sample of matrix pairs Υ.

As the relation (3.2) indicates, the relation (3.3) should hold regardless of the condition numbers $\kappa_2(A^{(0)})$ and $\kappa_2(B^{(0)})$. Therefore, it makes sense to investigate how $\rho_{(A,B)}$ behaves with respect to $\chi_{(A,B)}$, where

$$\chi_{(A,B)} = \sqrt{\kappa_2^2(A^{(0)}) + \kappa_2^2(B^{(0)})}\ .$$

For the given sample of pairs Υ, we shall make the "graph of relative errors" $\mathscr{E}$ for the CFL method. It is defined by

$$\mathscr{E} = \{(\chi_{(A,B)}\ ,\ \rho_{(A,B)}) : (A,B) \in \Upsilon\}.$$

3.1 Implementation details

For a smaller matrix size n, we can compute "nearly exact" eigenvalues λ_i using MATLAB and its variable precision arithmetic (vpa). The eigenvalues $\tilde{\lambda}_i$ are computed by the CFL method using the standard double precision. Hence it will be easy to compute the quantities $\rho_{(A,B)}$ and $\chi_{(A,B)}$. The graph $\mathscr{E}$ will be displayed using MATLAB scatter(x,y,3) function. *The method will be indicated to have HRA property if the y-values of the points on the graph are scattered around the machine epsilon* $\mathbf{u} \approx 2.2 \cdot 10^{-16}$ or below it. For comparisons we shall apply the same accuracy test to the intrinsic MATLAB function eig(A,B).

3.1.1 Matrix pair generation

Let us describe how the pairs of Hermitian positive definite matrices for numerical tests have been generated. The procedure is quite similar to that from [9]. That procedure uses 4 diagonal matrices with positive diagonal elements: Σ, Δ_A, Δ_B, Δ and two unitary matrices U, V of order n. The starting pair $(A^{(0)}, B^{(0)})$ is computed in two steps:

(1) $F = U\Sigma V^*, \quad A = F^*\Delta_A F, \quad B = F^*\Delta_B F,$

(2) $B^{(0)} = D_B^{-1/2} B D_B^{-1/2} \ (= B_S), \quad A^{(0)} = \Delta D_A^{-1/2} A D_A^{-1/2} \Delta \ (= \Delta A_S \Delta),$

where D_A and D_B are the diagonal parts of A and B, exactly as they are defined in Theorem 3.1. The magnitudes of $\kappa_2(A_S^{(0)})$ and $\kappa_2(B_S^{(0)})$ can be controlled by the magnitudes of the diagonal entries of Δ_A, Δ_B, Σ. Indeed, by [20] we have $\kappa_2(A_S^{(0)}) \leq n\kappa_2^2(\Sigma)\kappa_2(\Delta_A)$, $\kappa_2(B_S^{(0)}) \leq n\kappa_2^2(\Sigma)\kappa_2(\Delta_B)$ and almost always $\kappa_2(A_S^{(0)})$ and $\kappa_2(B_S^{(0)})$ are much smaller than these bounds. To simplify construction, we have set $\Delta_B = I_n$.

Note that $\kappa_2(A^{(0)}) \leq \kappa_2(A_S^{(0)})\kappa_2^2(\Delta)$. If the CFL method has HRA property, $\rho_{(A,B)}$ from the relation (3.2) should not depend on $\kappa_2(A^{(0)})$ which is controlled by $\kappa_2(\Delta)$.

If we set $\Delta = I_n$ and $(A^{(0)}, B^{(0)}) = (D_B^{-1/2} A D_B^{-1/2}, B_S)$, then we know the eigenvalues of $(A^{(0)}, B^{(0)})$ in advance. They are the quotients $(\Delta_A)_{ll}/(\Delta_B)_{ll}$, $1 \leq l \leq n$. This can be used when considering the matrix pairs with multiple eigenvalues.

The diagonal matrices are constructed via the MATLAB function `diag(d)`, where `d` is a vector. Vectors are constructed by the MATLAB function `logspace(x1,x2,n)`. We use it to make the diagonal matrices Σ and Δ_A. For the construction of Δ we use our m-function `scalvec(k1,k2,k3,n,k)` which generates vector d of length n, $d = [10^{\mathtt{k1}}, \ldots, 10^{\mathtt{k2}}, \ldots, 10^{\mathtt{k3}}]$. Here `k` determines position of $10^{\mathtt{k2}}$ among the components of d. We have set $k = [n/2]$ where for real t, $[t]$ is the largest integer smaller than or equal to t. To compute Δ, `scalvec` is used within a 3-level loop, controlled by `k1`, `k2` and `k3`. Altogether our main m-file uses a 7-level loop, three for computing Δ, two for Σ and 2 for Δ_A. The unitary matrices U and V are computed using the QR factorization of the random matrices of order n. Say, for computing U the command `[Q,~]=qr(rand(n)+1i*rand(n))` has been used.

Once, we have obtained $A^{(0)}$, $B^{(0)}$, we convert their copies to symbolic type, so that we can use the `vpa` with those copies. We use `vpa` with 80 decimal digits to compute the reference eigenvalues and eigenvectors.

We have made tests for the MATLAB `eig(A,B)` function and for our `zcfl(A,B)` m-function which contains MATLAB code for the CFL method. As a control method, we have used our m-function `zABeig(A,B,dg)` which calls MATLAB functions `eig(A)`, `chol(A)` and `inv(A)`, which all can use `vpa`. Here `dg` stands for the number of decimal digits used by the `vpa`. We have considered only accuracy of the computed eigenvalues.

On input the m-functions accept the pair (A,B) of Hermitian matrices. The m-function `zcfl(A,B)` uses only the upper-triangles of the matrices A and B. On output this m-function yields the eigenvector matrix F, the diagonal matrix of eigenvalues and the number of sweeps needed to terminate the process. We consider output to the control method accurate, and use it to compute the maximum relative error of the computed eigenvalues obtained by `eig(A,B)` and `zcfl(A,B)`.

Altogether, we have generated 15300 pairs of positive definite matrices of order 10. These pairs make the sample Υ for testing the high relative accuracy of the CFL method.

3.1.2 How to code the algorithm?

Although the algorithm for computing the transformation parameters α and β is invariant under the transformation $(A,B) \mapsto (\sigma_A A, \sigma_B B)$ we have first applied just that scaling. It can preclude overflow and avoid working with subnormal numbers. After that we apply the congruence transformation $(A,B) \mapsto (DAD, DBD)$ with a suitably chosen diagonal matrix D. The diagonal entries of D as well as the scalars σ_A and σ_B are computed as powers of 2, so that no rounding error is introduced. The procedure can be described as follows.

If $A=0$ $(B=0)$ then all the eigenvalues of the pair (A,B) are 0 (∞) and any linearly independent set of vectors is a basis of $\mathbf{C}^n$ consisting of the eigenvectors of (A,B). Otherwise, use the following procedure.

(i) Find integers s_A and s_B such that $n2^{s_A} \leq \|A\|_F < n2^{s_A+1}$, $n2^{s_B} \leq \|B\|_F < n2^{s_B+1}$. Then compute $\sigma_A = 2^{s_A}$, $\sigma_B = 2^{s_B}$ and set $\tilde{A} = (\tilde{a}_{rt}) = 2^{-s_A}A$, $\tilde{B} = (\tilde{b}_{rt}) = 2^{-s_B}B$

(ii) Find $s_1,\ldots,s_n$ such that $2^{s_r} \leq \sqrt[4]{\tilde{a}_{rr}^2 + \tilde{b}_{rr}^2} < 2^{s_r+1}$, $1 \leq r \leq n$. Then compute $A^{(0)} = D\tilde{A}D$, $B^{(0)} = D\tilde{B}D$, where $D = \operatorname{diag}(2^{-s_1},\ldots,2^{-s_n})$.

(iii) Initialize the matrix of accumulated transformations $F^{(0)} = D$.

The eigenvalues of the pair (A,B) are $\sigma_B/\sigma_A = 2^{s_b - s_A}$ times the eigenvalues of $(A^{(0)},B^{(0)})$. The eigenvectors of the pair (A,B) are D times the eigenvectors of $(A^{(0)},B^{(0)})$. By this procedure we have achieved that $A^{(0)}$ and $B^{(0)}$ have norms of the same order of magnitude, and $1 \leq \sqrt{(a_{rr}^{(0)})^2 + (b_{rr}^{(0)})^2} < 2$, where $A^{(0)} = (a_{rt}^{(0)})$, $B^{(0)} = (b_{rt}^{(0)})$.

The most important parts of the algorithm are those related to the stopping of the process and to determining whether $\Im$ is sufficiently small to employ the special formulas for α and β.

We have computed the transformation parameters exactly as is described in 2^0 **(b)** part of CFL algorithm in Section 2.2.

As for the stopping criterion, in the case of positive definite matrices A, B, we have used the following procedure: stop the process when

$$|a_{rt}| \leq \sqrt{a_{rr}a_{tt}}\,\mathbf{u}, \qquad |b_{rt}| \leq \sqrt{b_{rr}b_{tt}}\,\mathbf{u}, \qquad 1 \leq r < t \leq n.$$

If the serial (or any cyclic) pivot strategy is used, then it makes sense to set $a_{ij} = 0$ and $b_{ij} = 0$ whenever $|a_{ij}| \leq \sqrt{a_{ii}a_{jj}}\mathbf{u}$ and $|b_{ij}| \leq \sqrt{b_{ii}b_{jj}}\mathbf{u}$, and then proceed with the next step. The process is terminated when all off-diagonal elements of the current iteration matrices A and B are zero.

To justify that stopping procedure one can use the relation (3.1). In the final stage of the process we shall have $\kappa_2(A_S) \approx 1$, $\kappa_2(B_S) \approx 1$. Replacing a_{ij} and a_{ji} (b_{ij} and b_{ji}) by zeros amounts to perturbing the current matrix A (B) by $\delta A = -a_{ij}e_ie_j^T - \bar{a}_{ij}e_je_i^T$ $(\delta B = -b_{ij}e_ie_j^T - \bar{b}_{ij}e_je_i^T)$. Here (i,j) is the pivot pair and $I_n = [e_1,\ldots,e_n]$. Hence the right hand side of the inequality (3.1) will be bounded by a modest multiple of $|a_{ij}|/\sqrt{a_{ii}a_{jj}}$ $(|b_{ij}|/\sqrt{b_{ii}b_{jj}})$. When used in our termination process, the both pivot elements are zeroed and therefore the numerator of the right-hand side of the relation

(3.1) is bounded by $|a_{ij}|/\sqrt{a_{ii}a_{jj}} + |b_{ij}|/\sqrt{b_{ii}b_{jj}}$. Hence the maximum relative error of the eigenvalues is bounded by a modest multiple (or just a fraction) of $\mathbf{u}$.

A quite natural upgrading of that stopping criterion can read: *do the same (replace the pivot elements by zeros) but in addition update also the diagonal elements.* However, the theoretical justification of that upgrading would require a more extensive analysis which shall be omitted here.

In our testings we have encountered positive definite matrix pairs for which the above stopping criterion allows too many sweeps. This occurs when in the final stage of the process there exist pivot submatrices which yield very tiny $|\Im_{ij}|$, $|\Im_i|$ and $|\Im_j|$. These quantities are bounded by a modest multiple of $\mathbf{u}$, and consequently $|\Im|$ is of order $\mathbf{u}^2$. In such a case α_k and β_k are prone to huge relative errors while the LS solution gives the residual which is as large as the pivot elements prior to the step. In another words the both procedures fail to decrease the residual, i.e. we have $|a_{ij}^{(k+1)}| + |b_{ij}^{(k+1)}| \lesssim |a_{ij}^{(k)}| + |b_{ij}^{(k)}|$. A quick solution to this problem can be to locally apply iterative process to each such pair of pivot submatrices $(\hat{A}^{(k)}, \hat{B}^{(k)})$ until the pivot elements can be replaced by zero. The proper solution could be to devise a better criterion when the pivot elements can be replaced by zero in the final stage of the process.

We end the paper by displaying two graphs of the relative errors. The first is made for the intrinsic MATLAB function `eig(A,B)`, the second is made for the m-function `zcfl(A,B)`. Recall that each graph is defined by $\mathscr{E} = \{(\chi_{(A,B)}\ ,\ \rho_{(A,B)}) : (A,B) \in \Upsilon\}$, where the sample of matrix pairs Υ is the same for the both methods. The graphs indicate high relative accuracy of the CFL algorithm.

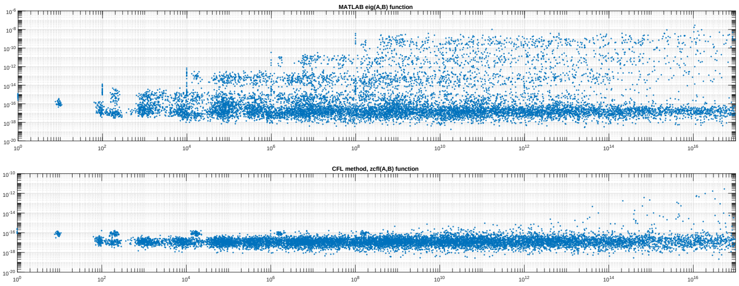

Fig. 1 The graphs of the relative errors for the MATLAB `eig(A,B)` and for `zcfl(A,B)` function.

4 Conclusions and Future Work

The complex version of the Falk-Langemeyer method has been derived. The method treats both matrices in an equal way which is not the case with other methods for solving GEP. It has been shown that the method is well defined for any definite pair of Hermitian matrices. Numerical tests indicate that it computes the eigenvalues of well-behaved matrix pairs to high relative accuracy. It is an excellent choice to be the kernel algorithm for the appropriate block Jacobi methods.

Future work can be concentrated on proving the global and asymptotic quadratic convergence of the method as well as on proving the high relative accuracy of the method. We believe that the quadratic convergence proof will be the same as the one in the case of real matrices. Also, several open problems that have been addressed in this paper should be solved. Finally, since each transformation matrix has the spectral radius that is not smaller than one, the elements of the iterated matrices can become very large. So, a procedure should be included in the algorithm to solve that problem.

Finally, it would be interesting to investigate whether the method is well defined for some larger class of pairs of Hermitian matrices.

References

1. J. Demmel, K. Veselić, *Jacobi's method is more accurate than QR*, SIAM J. Matrix Anal. Appl. **1992**, 13, 1204–1245.
2. Drmač, Z.: A Tangent Algorithm for Computing the Generalized Singular Value Decomposition. SIAM J. Numer. Anal. 35 (5), 1804-1832 (1998)
3. Falk, S., Langemeyer, P.: Das Jacobische Rotations-Verfahren für reel symmetrische Matrizen-Paare I, II. Elektronische Datenverarbeitung 30-43 (1960)
4. Hari, V.: On Cyclic Jacobi Methods for the Positive Definite Generalized Eigenvalue Problem. Ph.D. thesis, University of Hagen (1984)
5. Hari, V.: On Pairs of Almost Diagonal Matrices. Linear Algebra and Its Appl. 148, 193-223 (1991)
6. Hari, V., Drmač, Z: On Scaled Almost Diagonal Hermitian Matrix Pairs. SIAM J. Matrix Anal. Appl. 18 (4), 1000-1012 (1997)
7. Hari, V.: Convergence to Diagonal Form of Block Jacobi-type Methods. Numer. Math. 129 (3), 449-481 (2015)
8. Hari, V., Begović Kovač, E.: Convergence of the Cyclic and Quasi-cyclic Block Jacobi Methods. Electron. T. Numer. Ana. (ETNA) 46, 107-147 (2017)
9. Hari, V.: Globally convergent Jacobi methods for positive definite matrix pairs. Numerical Algorithms, doi.org/10.1007/s11075-017-0435-5 (2018)
10. Higham N. J., Tisseur F., Van Dooren P. M.: Detecting a definite Hermitian pair and a hyperbolic or elliptic quadratic eigenvalue problem, and associated nearness problems, Lin. Alg. and Its Appl. 351-351, 455-474 (2002)
11. Matejaš, J.: Accuracy of the Jacobi Method on Scaled Diagonally Dominant Symmetric Matrices. SIAM. J. Matrix Anal. Appl. 31(1), 133-153 (2009)
12. Matejaš, J.: Accuracy of one step of the Falk-Langemeyer method. Numerical Algorithms 68(4), 645-670 (2015)
13. Novaković, V., Singer, S., Singer, S.: Blocking and Parallelization of the Hari–Zimmermann Variant of the Falk–Langemeyer Algorithm for the Generalized SVD. Parallel Comput. 49, 136-152 (2015)
14. Parlett, B. N.: Symmetric Eigenvalue Problem. Prentice Hall Inc., Englewood Cliffs, N.J. (1980).
15. de Rijk, P. P. M.: A one-sided Jacobi algorithm for computing the singular value decomposition on a vector computer. SIAM J. Sci. Stat. Comp. 10, 359-371 (1989)
16. Shroff, G., Schreiber, R.: On the Convergence of the Cyclyc Jacobi Method for Parallel Block Orderings. SIAM J. Matrix Anal. Appl. 10 (3), 326-346 (1989)
17. Slapničar, I., Hari, V.: On the Quadratic Convergence of the Falk-Langemeyer Method for Definite Matrix Pairs. SIAM J. Matrix Anal. Appl. 12 (1), 84-114 (1991)
18. Stewart, G. W.: Perturbation Bounds for the Definite Generalized Eigenvalue Problem. Lin. Alg. and Its Appl. 23, 69-85 (1960)
19. Stewart, G. W.: Matrix Algorithms, Vol II: Eigensystems. SIAM (2001)
20. van der Sluis, A.: Condition numbers and equilibration of matrices. Numer. Math. 14 (1), 14–23 (1969)
21. Zimmermann, K.: On the Convergence of the Jacobi Process for Ordinary and Generalized Eigenvalue Problems. Ph.D. Thesis, Dissertation No. 4305 ETH, Zürich (1965)