Calcolo (2022) 59:38
https://doi.org/10.1007/s10092-022-00484-3

Newton-type methods for simultaneous matrix diagonalization

Rima Khouja[1] · **Bernard Mourrain**[1] · **Jean-Claude Yakoubsohn**[2]

Received: 10 December 2021 / Revised: 22 August 2022 / Accepted: 20 September 2022 /
Published online: 1 October 2022

Abstract
This paper proposes a Newton-type method to solve numerically the eigenproblem of several diagonalizable matrices, which pairwise commute. A classical result states that these matrices are simultaneously diagonalizable. From a suitable system of equations associated to this problem, we construct a sequence that converges quadratically towards the solution. This construction is not based on the resolution of a linear system as is the case in the classical Newton method. Moreover, we provide a theoretical analysis of this construction and exhibit a condition to get a quadratic convergence. We also propose numerical experiments, which illustrate the theoretical results.

Keywords Simultaneous diagonalization · Newton-type method · eigenproblem · eigenvalues · certification · high precision computation

Mathematics Subject Classification 65F15 · 65H10 · 15A18 · 65-04

Rima Khouja, Bernard Mourrain and Jean-Claude Yakoubsohn have contributed equally to this work.

✉ Rima Khouja
rima.khouja@inria.fr; rima.khouja@univ-lorraine.fr

Bernard Mourrain
bernard.mourrain@inria.fr

Jean-Claude Yakoubsohn
jean-claude.yakoubsohn@math.univ-toulouse.fr

[1] AROMATH, Inria d'Université Côte d'Azur, 2004, route des Lucioles, 06902 Sophia Antipolis, France

[2] Institut de Mathématiques de Toulouse, Université Paul Sabatier, 118, route de Narbonne, 31062 Toulouse, France

1 Introduction

1.1 Our study

Let us consider p *diagonalizable* matrices $M_1, \dots, M_p$ in $\mathbb{C}^{n\times n}$ which pairwise commute. A classical result states that these matrices are simultaneously diagonalizable, i.e., there exists an invertible matrix E and diagonal matrices Σ_i, $1 \leqslant i \leqslant p$, such that $EM_iE^{-1} = \Sigma_i$, $1 \leqslant i \leqslant p$, see e.g. [24]. The aim of this paper is to compute numerically a solution (E, F, Σ) of the system of equations

$$f(E,F,\Sigma) := \begin{pmatrix} FE - I_n \\ FME - \Sigma \end{pmatrix} = 0 \tag{1}$$

where $\Sigma = (\Sigma_1, \dots, \Sigma_p)$ and $FME - \Sigma := (FM_1E - \Sigma_1, \dots, FM_pE - \Sigma_p) = 0$. Notice that this system is multi-linear in the unknowns E, F, Σ. We verify that when $p = 1$ and M_1 is a generic matrix, this system has a solution set of dimension $2\,n^2 + n - 2n^2 = n$ ($n^2 + n^2 + n$ unknowns for E, F, Σ and 2 matrix equations corresponding to $n^2 + n^2$ equations). However, for $p > 1$ and generic matrices M_i, there is no solution. To have a solution, the pencil M must be on the manifold $\mathcal{D}_p$ of p-tuples of simultaneously diagonalizable matrices.

The system (1) can be generalized to the following system:

$$f'\left(E,F,\Sigma'\right) := \begin{pmatrix} FM_0E - \Sigma_0 \\ FME - \Sigma \end{pmatrix} = 0 \tag{2}$$

where $\Sigma' = (\Sigma_0, \Sigma_1, \dots, \Sigma_p)$, $M_0 \in \mathbb{C}^{n\times n}$ is replacing I_n and Σ_0 is a diagonal matrix replacing I_n in the first equation of (1). When the pencil $M' = (M_0, M_1, \dots, M_p)$ contains an invertible matrix, the solutions of the two systems are closely related. If M_0 is invertible, a solution (E, F, Σ') of (2) for $M' = (M_0, M_1, \dots, M_p)$ gives the solution $(FM_0, E\Sigma_0^{-1}, \Sigma\Sigma_0^{-1})$ of (1) for $M = (M_0^{-1}M_1, \dots, M_0^{-1}M_p)$. A similar correspondence between the solution sets can be obtained if a linear combination $M_0' = \sum_{i=1}^{p} \lambda_i M_i$ is invertible.

As (2) can be seen as an homogeneisation of (1) and appears in several contexts and applications, we will also study Newton-type methods for this homogenized system.

To solve the system of equations (1), we propose to apply a Newton-like method and to analyze the Newton map associated to an iteration. These ideas have also been developed for instance in [32] where a Newton method is used for the symmetric eigenvalue problem. A Simultaneous Newton's iteration for ill-conditioned eigenproblem has been introduced in [21]. For more recent references using the Newton-type approach for eigenproblem see for instance [27, 28, 38]. Moreover, similar approach for the fast computation of the singular value decomposition has been presented in a technical report [45].

We say that we have a quadratic sequence associated to a system of equations if the sequence converges quadratically towards a solution.

The classical Newton map defines $(E+X, F+Y, \Sigma+S)$ from (E, F, Σ) in order to cancel the linear part in the Taylor expansion of $f(E+X, F+Y, \Sigma+S)$. An easy computation shows that the perturbations X, Y and S are solutions of such a Sylvester-type linear system

$$\begin{pmatrix} FE - I_n + FX + YE \\ FME - \Sigma - S + XMF + EMY \end{pmatrix} =0. \tag{3}$$

A straight-forward way to solve this linear system is via Kronecker product, see [23]. This leads to a linear system of size $2n^2$, which can be solved in $\mathcal{O}(n^6)$ arithmetic operations.

The construction of the methods studied here is based on perturbations of such type $(E(I_n+X), (I_n+Y)F, \Sigma+S)$ rather than $(E+X, F+Y, \Sigma+S)$. More precisely the perturbations X, Y and S that we consider are perturbations which cancel the linear part of the Taylor expansion of $f(E(I_n+X), (I_n+Y)F, \Sigma+S)$. In this case, we can produce explicit solutions for the linear system in X, Y and S given by:

$$\begin{pmatrix} Z + X + Y \\ \Delta - S + \Sigma X + Y\Sigma \end{pmatrix} =0. \tag{4}$$

where $Z = FE - I_n$ and $\Delta = FME - \Sigma$. We will see that the linear system (4) admits an explicit solution (X, Y, S) with respect to Z and Δ for $p = 1, 2$ in (1). This is because Σ is a diagonal matrix. From these considerations, we define and analyze a sequence that converges quadratically towards a solution of the system (1) without inverting a linear system at each step of this Newton-like method.

1.2 Related works

Simultaneous matrix diagonalization is required by many algorithms as it was pointed out in [7, 19, 25, 30, 46]. A numerical analysis for two normal commuting matrices is proposed in [8] using Jacobi-like methods. Their method adjusts the classical Jacobi method in successively solving $\frac{n(n-1)}{2}$ two-real-variables optimization problems at each sweep of the algorithm. Their main result states a local quadratic convergence and can be summarized in the following way. Let $\mathrm{off}_2(A, B)^2 = \sum_{i\neq j} |A_{i,j}|^2 + |B_{i,j}|^2$. Let $\{\alpha_1, \ldots, \alpha_n\}$ (resp. $\{\beta_1, \ldots, \beta_n\}$) be the set of the eigenvalues of A (resp. B). Let A^k and B^k the matrices obtained at the step k of the Jacobi-like method and $\rho_k = \mathrm{off}_2(A^k, B^k)$. If

$$\rho_0 < \frac{1}{2}\delta := \frac{1}{4}\min_{i\neq j}\left(|\alpha_i - \alpha_j|, |\beta_i - \beta_j|\right),$$

then

$$\rho_{k+1} < 2n(9n-13)\,\frac{\rho_k^2}{\delta}.$$

We will see in Theorems 3 and 5 that the local conditions of the quadratic convergence do not depend on n. Many other papers studied the so-called Jacobi-like methods (see e.g. [31, 33] and references therein).

In [22] a sequence with proof of its convergence towards a numerical solution of the system (1) when $p = 1$, i.e., for M_1, with the assumption of M_1 being a diagonalizable matrix, is presented. It requires matrix inversion. Furthermore, under some extra assumptions, its quadratic convergence is established.

For a pencil of real *symmetric* matrices $C = (C_1, \dots, C_s)$, several algorithms based on Riemannian optimization methods (see [2]) have been developed in order to find an *approximate joint diagonalizer* (see e.g. [1, 5, 26, 36]). The idea is to find a local minimizer $B \in \mathbb{R}^{n\times n}$ of an objective function f which measures the degree of non-diagonality of the pencil $(BC_1B^T, \dots, BC_sB^T)$ over a Riemannian manifold (see [3, 5, 47] for some examples of objective functions). This Riemannian manifold is defined according to the geometric constraints considered on B. For instance, the diagonalizer is supposed to be orthogonal in some of these algorithms after a pre-whitening step (see e.g. [10, 11, 17, 20, 26, 34–36]). Due to inaccuracies in the computation of the diagonalizer with orthogonality constraints (see. [49]), *oblique* constraints, i.e., all the rows of the diagonalizer have unit Euclidean norm, have also been considered instead of the former constraints in more recent works (see e.g. [1, 5]). These algorithms can be used when the pencil of symmetric matrices is simultaneously diagonalizable. In this case, we aim to find a zero of the objective function f. However, these algorithms have a computation complexity higher than the Newton-type algorithm that we propose (see Proposition 4). For instance, most of them combine line search [2, Ch4] or trust region [2, Ch7] methods, and matrix inversions at each iteration (see the exact Riemannian Newton iteration in [1]). Moreover, the points on the Riemannian manifold are updated using a retraction operator (see [2, Ch4] or [5] for an example of a retraction operator on the oblique manifold). In the Newton-type method described in Sects. 3 and 4 the points are updated by using direct and explicit formulas. They have lower complexity than the Riemannian optimization-based algorithms and they are well-adapted to computation with high precision.

Simultaneous matrix diagonalization appears in many applications. For instance, in the solution of multivariate polynomial equations by algebraic methods, the isolated roots of the system are obtained from the computation of common eigenvectors of commuting operators of multiplication in the quotient ring and from their eigenvalues [15, 18]. In the case of simple roots, this reduces to simultaneous diagonalization of a pencil of matrices.

The approach of approximate joint diagonalizer for a pencil of real *symmetric* matrices is used to solve Blind Source Separation (BSS) problem, with potential applications in wide domains of engineering (see e.g. [14]).

Simultaneous matrix diagonalization of pencils of general matrices also appears in the rank (or canonical) decomposition of tensors [16]. Under certain conditions this rank decomposition is unique [39]. In this case simultaneous matrix diagonalization allows to compute this rank decomposition which plays a crucial role in numerous applications such that Psychometric [12], Signal Processing and Machine Learning [13, 40], Sensor array processing [43], Arithmetic Complexity

[9], wireless communications [44], multidimensional harmonic retrieval [41, 42], Chemometrics [6], and Principal components analysis [29].

1.3 Outline

Our *contributions* are a new iteration for the simultaneous diagonalization of matrices, with a local quadratic convergence and its analysis. The iteration is different from a Newton iteration. It does not require to invert a large linear system, but performs simple matrix operations. We analyse the numerical behavior of the method and provide a certification test for the convergence. Sections 2, 3, 4, and 5 are devoted to respectively constructing a sequence to solve numerically:

- $FE - I_n = 0$,
- the system (1) when $p = 1$,
- the system (2) when $p = 1$,
- the system (1) for any p.

Moreover, we provide for these cases, a certification that the sequence converges to a nearby solution, and a test to detect when this convergence is quadratic from an initial point. More precisely, in Sect. 3 we show that a triplet (E_0, F_0, Σ_0) must satisfy a property depending on the quantity $\varepsilon_0 := \max(\kappa_0^2 K_0^2 \|Z_0\|, \kappa_0^2 K_0 \|\Delta_0\|)$ to get a quadratic convergence where

1. $Z_0 = F_0E_0 - I_n$,
2. $\Delta_0 = F_0ME_0 - \Sigma_0$,
3. $\kappa_0 = \max\left(1, \max\limits_{1\leqslant j<k\leqslant n} \dfrac{1}{|\sigma_{0,k} - \sigma_{0,j}|}\right)$,
4. $K_0 = \max_k \left(1, |\sigma_{0,k}|\right)$,

where $\sigma_{0,1}, \dots, \sigma_{0,n}$ denote the diagonal entries of Σ_0. The quantity κ is the condition number of the studied methods. Based on the same methodology as in Sect. 3, Sections 4 and 5 exhibit a certification of the convergence of the sequence constructed to the studied case towards the solution with a sufficient condition on the initial point.

In Sect. 6 we perform numerical experimentation. The final section is for our conclusions and future works.

1.4 Notation and preliminaries

Throughout this work, we will use the infinity vector norm and the corresponding matrix norm. For a given vector $v \in \mathbb{C}^n$ and matrix $M \in \mathbb{C}^{n\times n}$, they are respectively given by:

$$\begin{aligned}\|v\| &= \max\left\{|v_1|, \dots, |v_n|\right\}\\ \|M\| &= \max_{\|v\|=1} \|Mv\|.\end{aligned}$$

Explicitly, $\|M\| = \max\{|m_{i,1}| + \ldots + |m_{i,n}| \,:\, 1 \le i \le n\}$.

For a second matrix $N \in \mathbb{C}^{n\times n}$, we have

$$\begin{aligned}\|M+N\| &\leqslant \|M\| + \|N\| \text{ (sub-additivity)}\\ \|MN\| &\leqslant \|M\|\,\|N\| \text{ (sub-multiplicativity).}\end{aligned}$$

Moreover, for a given matrix $M \in \mathbb{C}^{n\times n}$, we denote by $\|M\|_{\mathrm{L,Tri}}$ and $\|M\|_{Frob}$ the following:

$$\|M\|_{\mathrm{L,Tri}} := \max_{\substack{1 \le i \le n\\ 1 \le j \le i-1}} |m_{i,j}|,$$

i.e the max matrix norm of the lower triangular part of M,

$$\|M\|_{Frob} := \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} |m_{i,j}|^2},$$

i.e., the Frobenius norm of M.

Furthermore, we consider in this paper the regular case of diagonalizable matrices, that is, the matrices are diagonalizable with simple eigenvalues. Thus we will use the following notation

$$\mathcal{W}_n := \left\{M \in \mathbb{C}^{n\times n} \mid M \text{with pairwise distinct eigenvalues}\right\}.$$

It is well-known that $\mathcal{W}_n$ is dense in $\mathbb{C}^{n\times n}$.

The Lie group of $n \times n$ invertible matrices, denoted by GL_n, is the so-called general linear group [4]. We denote by $\mathcal{D}_n$ the vector space of diagonal matrices of size n and $\mathcal{D}'_n$ denotes the subset of $\mathcal{D}_n$ in which the diagonal matrices are of n distinct diagonal entries. Let $E, F \in GL_n$ and $\Sigma \in \mathcal{D}'_n$. The tangent space of GL_n at E (resp. F) is denoted by $T_E GL_n$ (resp. $T_F GL_n$) and the tangent space of $\mathcal{D}'_n$ at Σ is denoted by $T_\Sigma \mathcal{D}'_n$. The perturbation of respectively E, F and Σ that we consider in this paper are of the following form: $E + \dot{E}$, $F + \dot{F}$ and $\Sigma + \dot{\Sigma}$, where $\dot{E}$ and $\dot{F}$ are respectively in $T_E GL_n$ and $T_F GL_n$ and $\dot{\Sigma}$ is in $T_\Sigma \mathcal{D}'_n$.

As GL_n is a Lie group, $\dot{E}$ and $\dot{F}$ can be written as EX and YF such that X, Y are in the Lie algebra of GL_n which is equal to $\mathbb{C}^{n\times n}$ (since this Lie algebra is $T_{I_n} GL_n$ and GL_n is an open subset in $\mathbb{C}^{n\times n}$).

As $\mathcal{D}'_n$ is open in $\mathcal{D}_n$ then $T_\Sigma \mathcal{D}'_n = \mathcal{D}_n$, herein $\dot{\Sigma} = S \in \mathcal{D}_n$.

Finally, the perturbations of E, F and Σ that we consider are as follows:

$E + EX$, $F + YF$ and $\Sigma + S$, such that X and Y are in $\mathbb{C}^{n\times n}$ and S is a diagonal matrix in $\mathbb{C}^{n\times n}$.

For a matrix $M \in \mathbb{C}^{n\times n}$, let diag$(M)$ be the diagonal matrix with the same diagonal as M and let off(M) be the matrix where the diagonal term of M are replaced by 0. We have $M = \text{diag}(M) + \text{off}(M)$. We say that M is an off-matrix if $M = \text{off}(M)$. In addition, let $(\lambda_1, \dots, \lambda_n) \in \mathbb{C}^n$, diag$(\lambda_1, \dots, \lambda_n)$ is the diagonal matrix in $\mathbb{C}^{n\times n}$ of diagonal entries $\lambda_1, \dots, \lambda_n$.

The superscripts $.^t$, $.^*$ and $.^{-1}$ are used respectively for the transpose, Hermitian conjugate, and the inverse matrix.

We state the following lemma which will be used in some of the proofs.

Lemma 1 *Let* $\varphi(\varepsilon, u) = \frac{\prod_{j\geqslant 0}(1+u\varepsilon^{2^j})-1}{\varepsilon u}$. *Given* $\varepsilon \leqslant \frac{1}{2}$, $u \leqslant 1$, *and* $i \geqslant 0$, *we have*

$$\prod_{j\geqslant 0}\left(1 + u\varepsilon^{2^{j+i}}\right) \leqslant 1 + 2u\varepsilon^{2^i} \tag{5}$$

Proof Modulo taking ε^{2^i} instead of ε, it suffices to consider the case when $i = 0$. Now $\varphi(\varepsilon, u)$ is an increasing function in ε and u, since its power series expansion in ε and u admits only positive coefficients. Consequently, $\varphi(\varepsilon, u) \leqslant \varphi(\frac{1}{2}, 1) = 2$. $\square$

2 Newton-type method for the system $FE - I_n = 0$.

Let $f : GL_n \times GL_n \to \mathbb{C}^{n\times n}$, $(E, F) \mapsto FE - I_n$. We consider the following perturbations $E + EX$, $F + YF$ of respectively E and F where $X,\ Y \in \mathbb{C}^{n\times n}$.

To define the Newton sequence we have to solve the linear system obtained by canceling the linear part in the Taylor expansion of $f(E + EX, F + YF)$. The same methodology will be adopted in the next sections for the other considered systems. Hereafter, we detail the computation of the Newton sequence associated to the system $FE - I_n = 0$. Moreover, a sufficient condition on the initial point for the quadratic convergence of this Newton sequence will be established.

Let $Z = FE - I_n$. We observe that

$$f(E + EX, F + YF) = (F + YF)(E + EX) - I_n \tag{6}$$

$$= Z + \left(Z + I_n\right)X + Y\left(Z + I_n\right) + Y\left(Z + I_n\right)X. \tag{7}$$

We assume here that Z is of small norm, i.e., we start from an initial point (E_0, F_0) close from the solution of the system $FE - I_n = 0$.

Consequently, the linear system of first order terms to solve is

$$Z + X + Y = 0. \tag{8}$$

Hence $X = Y = -\frac{Z}{2}$ is a solution of Eq. (8). Moreover we get, by substituting in Eq. (7) X and Y by $-\frac{Z}{2}$,

$$(F+YF)(E+EX)-I_n = Z^2\Big(-\frac{3}{4}I_n+\frac{Z}{4}\Big). \tag{9}$$

Proposition 1 *Let* $Z_0 = F_0E_0 - I_n$. *Define* $X_0 = -\frac{Z_0}{2}$, $E_1 = E_0(I_n + X_0)$, $F_1 = (I_n + X_0)F_0$ *and* $Z_1 = F_1E_1 - I_n$. *Assume that* $\|Z_0\| \leqslant 1$. *Then*

$$\|Z_1\| \leqslant \|Z_0\|^2 \tag{10}$$

Proof It follows easily from (9). □

Theorem 2 *Let* E_0 *and* F_0 *two complex square matrices of size n. Let* $Z_0 = F_0E_0 - I_n$ *and assume that* $\varepsilon = \|Z_0\| < \frac{1}{2}$. *The sequences defined for* $i \geqslant 0$

$$\begin{aligned} Z_i &= F_iE_i - I_n \\ X_i &= -\frac{Z_i}{2} \\ E_{i+1} &= E_i\big(I_n + X_i\big) \\ F_{i+1} &= \big(I_n + X_i\big)F_i \end{aligned}$$

converge quadratically towards the solution of $FE - I_n = 0$. *Each* E_i, *respectively* F_i *are invertible and, if* E_∞ *and* F_∞ *are respectively the limits of sequences* $(E_i)_{i\geqslant 0}$ *and* $(F_i)_{i\geqslant 0}$ *we have for* $i \geqslant 0$,

$$\begin{aligned} \big\|E_i - E_\infty\big\| &\leqslant (1+2\varepsilon)2^{-2^{i+1}+1}\varepsilon\big\|E_0\big\|, \\ \big\|F_i - F_\infty\big\| &\leqslant (1+2\varepsilon)2^{-2^{i+1}+1}\varepsilon\big\|F_0\big\|. \end{aligned}$$

Proof First, by the assumption $\|F_0E_0 - I_n\| = \|Z_0\| < \frac{1}{2}$, we have E_0 and F_0 are invertible. In fact, $E_0F_0 = I_n + E_0F_0 - I_n = I_n + Z_0$ is invertible when $\|Z_0\| < 1$ which is the case since we suppose $\|Z_0\| < \frac{1}{2}$.

Let us prove by induction that $\|Z_k\| \leqslant 2^{-2^k+1}\varepsilon$. Since $\varepsilon < \frac{1}{2}$, we have

$$\begin{aligned} \|Z_{k+1}\| &\leqslant \|Z_k\|^2 \qquad \text{from (10)} \\ &\leqslant \varepsilon 2^{-2^{k+1}+2}\varepsilon \\ &\leqslant 2^{-2^{k+1}+1}\varepsilon. \end{aligned}$$

Consequently $Z_\infty = 0$. Since $X_k = -\frac{Z_k}{2}$ we deduce

$$\|X_k\| \leqslant 2^{-2^k}\varepsilon.$$

It follows $X_\infty = 0$. We have

$$\begin{aligned} E_k &= E_{k-1}\big(I_n + X_{k-1}\big) \\ &= E_0\big(I_n + X_0\big)\cdots\big(I_n + X_{k-1}\big). \end{aligned}$$

Denoting $W_i = \prod_{0\leqslant k\leqslant i}(I_n + X_k)$, $W_\infty = \prod_{k\geqslant 0}(I_n + X_k)$ we compute

$$\begin{aligned}\|W_\infty - I_n\| &\leqslant \prod_{k\geqslant 0}\left(1+2^{-2^k}\varepsilon\right)-1\\ &\leqslant 2\varepsilon \qquad \text{by using Lemma 1.}\end{aligned}$$

Then W_∞ is invertible and $\|W_\infty^{-1}\| \leqslant \dfrac{1}{1-2\varepsilon}$. Let $E_\infty = E_0 W_\infty$. Hence $E_0 = E_\infty W_\infty^{-1}$. In the same way $F_0 = W_\infty^{-1} F_\infty$. Finally, the identity $F_\infty E_\infty - I_n = 0$ permits to conclude that E_0 and F_0 are invertible. In the same way we prove easily that $\|W_i - I_n\| \leqslant 2\varepsilon$. It follows that W_i is invertible. Since $E_i = E_0 W_i$ we deduce that E_i is invertible. Moreover

$$\begin{aligned}\|W_i - W_\infty\| &\leqslant \|W_i\| \left\| 1 - \prod_{k\geqslant i+1}\left(1+\|X_k\|\right)\right\|\\ &\leqslant \left(1+\|W_i - I_n\|\right)\left\|\prod_{k\geqslant 0}\left(1+2^{-2^{k+i+1}}\varepsilon\right)-1\right\|\\ &\leqslant (1+2\varepsilon)2^{-2^{i+1}+1}\varepsilon \qquad \text{by using Lemma 1.}\end{aligned}$$

We deduce that

$$\|E_i - E_\infty\| \leqslant (1+2\varepsilon)2^{-2^{i+1}+1}\varepsilon\|E_0\|.$$

These properties also hold for the F_i's. The theorem is proved. □

3 Newton-like method for diagonalizable matrices.

Let $M \in \mathcal{W}_n$, $\Sigma \in \mathcal{D}'_n$, E, $F \in GL_n$. We aim to construct Newton sequences which converge towards the numerical solution of $f(E,F,\Sigma) = 0$ where $f : GL_n \times GL_n \times \mathcal{D}'_n \to \mathbb{C}^{n\times n} \times \mathbb{C}^{n\times n}$, $(E,F,\Sigma) \mapsto (FE - I_n, FME - \Sigma)$. We consider in the same way as before the perturbations $E + EX$ and $F + YF$ of respectively E and F and in addition the perturbation $\Sigma + S$ of Σ such that $S \in \mathcal{D}_n$. We get with $Z = FE - I_n$ and $\Delta = FME - \Sigma$:

$$\begin{aligned}&(F+YF)(E+EX) - I_n\\ &= Z + \left(Z+I_n\right)X + Y\left(Z+I_n\right) + Y\left(Z+I_n\right)X\end{aligned} \tag{11}$$

$$\begin{aligned}&(F+YF)M(E+EX) - \Sigma - S\\ &= FME - \Sigma - S + FMEX + YFME + YFMEX\\ &= \Delta - S + \Sigma X + Y\Sigma + \Delta X + Y\Delta + Y(\Delta+\Sigma)X\end{aligned} \tag{12}$$

As in the previous section we assume that (E,F,Σ) is sufficiently close to the solution of $f(E,F,\Sigma) = 0$, thus the linear system that we obtain from (11) and (12) is

$$\begin{cases} Z + X + Y & = 0 \\ \Delta - S + \Sigma X + Y\Sigma & = 0 \end{cases}$$

The following lemma gives a solution of this linear system.

Lemma 2 *Let* $\Sigma = \operatorname{diag}(\sigma_1, \dots, \sigma_n)$, $Z = (z_{i,j})_{1 \le i,j \le n}$ *and* $\Delta = (\delta_{i,j})_{1 \le i,j \le n}$ *be given matrices in* $\mathbb{C}^{n \times n}$. *Assume that* $\sigma_i \neq \sigma_j$ *for* $i \neq j$. *Let* S, X *and* Y *be matrices defined by*

$$S = \operatorname{diag}(\Delta - Z\Sigma) \tag{13}$$

$$x_{i,i} = 0 \tag{14}$$

$$x_{i,j} = \frac{-\delta_{i,j} + z_{i,j}\sigma_j}{\sigma_i - \sigma_j}, \qquad i \neq j \tag{15}$$

$$y_{i,i} = - z_{i,i} \tag{16}$$

$$y_{i,j} = \frac{\delta_{i,j} - z_{i,j}\sigma_i}{\sigma_i - \sigma_j}, \qquad i \neq j. \tag{17}$$

Then we have

$$Z + X + Y = 0 \tag{18}$$

$$\Delta - S + \Sigma X + Y\Sigma = 0 \tag{19}$$

Moreover

$$\|X\|, \|Y\| \leqslant \kappa\varepsilon(K + 1) \tag{20}$$

where $\varepsilon \geqslant \max(\|Z\|, \|\Delta\|)$, $\kappa = \max\left(1, \max_{i \neq j} \dfrac{1}{|\sigma_i - \sigma_j|}\right)$ *and* $K = \max(1, \max_i |\sigma_i|)$.

Proof It is easy to verify that $X + Y + Z = 0$. In this way the Eq. (19) is equivalent to

$$\Delta - S - Z\Sigma + \Sigma X - X\Sigma = 0.$$

Since $\operatorname{diag}(\Delta - S - Z\Sigma) = \operatorname{diag}(\Sigma X - X\Sigma) = 0$ the formulas which define X follow easily. The bounds (20) also are obvious to establish. □

In the next theorem we introduce the Newton sequences associated to the system $f(E, F, \Sigma) = 0$ with a sufficient condition on the initial point for its quadratic convergence.

Theorem 3 *Let* $E_0, F_0 \in GL_n$ *and* $\Sigma_0 \in \mathcal{D}'_n$ *be given such that they define the sequences for* $i \geqslant 0$,

$$
\begin{aligned}
Z_i &= F_i E_i - I_n \\
\Delta_i &= F_i M E_i - \Sigma_i \\
S_i &= \operatorname{diag}\left(\Delta_i - Z_i \Sigma_i\right) \\
E_{i+1} &= E_i\left(I_n + X_i\right) \\
F_{i+1} &= \left(I_n + Y_i\right) F_i \\
\Sigma_{i+1} &= \Sigma_i + S_i,
\end{aligned}
$$

where S_i, X_i *and* Y_i *are defined by the formulas* (13–17). *Let us define* $\kappa_0 = \max\left(1, \max_{i \neq j} \dfrac{1}{|\sigma_{0,i} - \sigma_{0,j}|}\right)$, $K_0 = \max(1, \max_i |\sigma_{0,i}|)$ *and* $\varepsilon_0 = \max(\kappa_0^2 K_0^2 \|Z_0\|, \kappa_0^2 K_0 \|\Delta_0\|)$. *Assume that*

$$
\varepsilon_0 \leqslant 0.033. \tag{21}
$$

Then the sequences $(\Sigma_i, E_i, F_i)_{i \geqslant 0}$ *converge quadratically to the solution of* $(FE - I_n, FME - \Sigma) = 0$. *More precisely* E_0 *and* F_0 *are invertible and*

$$
\begin{aligned}
\|E_i - E_\infty\| &\leqslant 8.1 \times 2^{1-2^{i+1}} \|E_0\| \frac{\varepsilon_0}{\kappa K} \\
\|F_i - F_\infty\| &\leqslant 8.1 \times 2^{1-2^{i+1}} \|F_0\| \frac{\varepsilon_0}{\kappa K}. \\
\|\Sigma_i - \Sigma_\infty\| &\leqslant 1.85 \times 2^{1-2^{i}} \frac{\varepsilon_0}{\kappa^2 K}.
\end{aligned}
$$

Proof Let us denote for each $i \geqslant 0$,

$$
\begin{array}{lllll}
\varepsilon & = \varepsilon_0 & \varepsilon_i & = & \max\left(\kappa_i^2 K_i^2 \|Z_i\|, \kappa_i^2 K_i \|\Delta_i\|\right) \\
\kappa & = \kappa_0 & \kappa_i & = & \max\left(1,\ \max_{1 \leqslant j < k \leqslant n} \frac{1}{|\sigma_{i,k} - \sigma_{i,j}|}\right) \\
K & = K_0 & K_i & = & \max_{1 \leq k \leq n}\left(1,\ |\sigma_{i,k}|\right),
\end{array}
$$

where $\sigma_{i,1}, \ldots, \sigma_{i,n}$ denote the diagonal entries of Σ_i. Let us show by induction on i that

$$
\varepsilon_i \leqslant 2^{1-2^i} \varepsilon \tag{22}
$$

$$
\|\Sigma_i - \Sigma_0\| \leqslant \left(2 - 2^{2-2^i}\right) \frac{2a}{\kappa} \varepsilon \tag{23}
$$

with $a = \dfrac{1}{1 - 8\varepsilon}$. These inequalities clearly hold for $i = 0$. Assuming that the induction hypothesis holds for a given i and let us prove it for $i + 1$. We first prove that $\|\Sigma_{i+1} - \Sigma_0\| \leqslant (2 - 2^{2-2^{i+1}}) \dfrac{2a}{\kappa} \varepsilon$ under the assumption $\|\Sigma_i - \Sigma_0\| \leqslant (2 - 2^{2-2^i}) \dfrac{2a}{\kappa} \varepsilon$. To do this, at the first step we show that this implies $K - \dfrac{4a}{\kappa}\varepsilon \leqslant K_i \leq K + \dfrac{4a}{\kappa}\varepsilon$ and $\dfrac{1}{1 + 8a\varepsilon} \kappa \leqslant \kappa_i \leq \dfrac{\kappa}{1 - 8a\varepsilon}$. Let us prove $K - \dfrac{4a}{\kappa}\varepsilon \leqslant K_i \leqslant K + \dfrac{4a}{\kappa}\varepsilon$. We have

$$
\begin{aligned}
K_i := \|\Sigma_i\| \le \|\Sigma_0\| + \|\Sigma_i - \Sigma_0\| \\
&\leqslant K + \left(2 - 2^{2-2^i}\right)\frac{2a}{\kappa}\varepsilon \\
&\leqslant K + \frac{4a}{\kappa}\varepsilon \leqslant K(1+4a\varepsilon).
\end{aligned}
$$

This implies simultaneously $K_i \geqslant K - |K - K_i| \geqslant K - \frac{4a}{\kappa}\varepsilon$ and $K_i \geqslant K(1-4a\varepsilon)$. Let us show that $\kappa_i \le \frac{\kappa}{1-8a\varepsilon}$. In fact, if the $\sigma_{i,j}$'s are the diagonal values of Σ_i, the Weyl's bound [48] implies that

$$|\sigma_{i,j} - \sigma_{0,j}| \leqslant \|\Sigma_i - \Sigma_0\| \leqslant \frac{4a}{\kappa}\varepsilon \text{for } 1 \leqslant j \leqslant n.$$

So that for $1 \leqslant j < k \leqslant n$, we obtain using $1 - 8a\varepsilon \geqslant 0$:

$$
\begin{aligned}
|\sigma_{i,k} - \sigma_{i,j}| &\geqslant |\sigma_{0,k} - \sigma_{0,j}| - |\sigma_{i,k} - \sigma_{0,k}| - |\sigma_{i,j} - \sigma_{0,j}| \\
&\geqslant |\sigma_{0,k} - \sigma_{0,j}|(1 - \kappa|\sigma_{i,k} - \sigma_{0,k}| - \kappa|\sigma_{i,j} - \sigma_{0,j}|) \\
&\geqslant |\sigma_{0,j} - \sigma_{0,k}|(1 - 8a\varepsilon) \geqslant 0.
\end{aligned}
$$

Finally, we get :

$$\kappa_i \leqslant \frac{\kappa}{1-8a\varepsilon}.$$

On the other hand the inequality

$$|\sigma_{i,k} - \sigma_{i,j}| \leqslant |\sigma_{0,k} - \sigma_{0,j}| + |\sigma_{i,k} - \sigma_{0,k}| + |\sigma_{i,j} - \sigma_{0,j}|$$

implies in the same way that above

$$\kappa_i \geqslant \frac{1}{1+8a\varepsilon}\kappa.$$

Next we prove (23) for $i+1$. We know $S_i = diag(\Delta_i - Z_i\Sigma_i)$. Since $\varepsilon_i = \max(\kappa_i^2 K_i^2\|Z_i\|, \kappa_i^2 K_i\|\Delta_i\|)$ and $\kappa_i, K_i \geqslant 1$ then $\|S_i\| \le \frac{2}{\kappa_i}\varepsilon_i \leqslant \frac{2(1+8a\varepsilon)}{\kappa}2^{1-2^i}\varepsilon$.

It follows :

$$
\begin{aligned}
\|\Sigma_{i+1} - \Sigma_0\| &\leqslant \|S_i\| + \|\Sigma_i - \Sigma_0\| \\
&\leqslant \frac{2(1+8a\varepsilon)}{\kappa}2^{1-2^i}\varepsilon + \left(2 - 2^{2-2^i}\right)\frac{2a}{\kappa}\varepsilon \\
&\leqslant \left(2 - 2^{1-2^i}(2-1)\right)\frac{2a}{\kappa}\varepsilon \qquad \text{since } 1 + 8a\varepsilon = a \\
&\leqslant \left(2 - 2^{1-2^i}\right)\frac{2a}{\kappa}\varepsilon
\end{aligned}
$$

But it is easy to see that $2^{1-2^i} \geqslant 2^{2-2^{i+1}}$. Finally we get

$$\|\Sigma_{i+1} - \Sigma_0\| \leqslant \left(2 - 2^{2-2^{i+1}}\right)\frac{2a}{\kappa}\varepsilon.$$

Hence we can also write

$$K_i - \frac{2a}{\kappa_i}\varepsilon \leqslant \|\Sigma_i\| - \|\Sigma_{i+1} - \Sigma_i\| \leqslant K_{i+1} \leqslant \|\Sigma_i\| + \|\Sigma_{i+1} - \Sigma_i\| \leqslant K_i + \frac{2a}{\kappa_i}\varepsilon$$

Using more the Weyl's bound we can easily get that

$$\frac{\kappa_i}{1+4a\varepsilon} \leqslant \kappa_{i+1} \leqslant \frac{\kappa_i}{1-4a\varepsilon}.$$

Now we bound $\kappa_{i+1}^2 K_{i+1}^2 \|Z_{i+1}\|$. We have

$$Z_{i+1} = Z_i X_i + Y_i Z_i + Y_i\left(Z_i + I_n\right)X_i.$$

Since $\|X_i\|, \|Y_i\| \le \kappa_i(\|\Delta_i\| + K_i\|Z_i\|) \leqslant \dfrac{2}{\kappa_i K_i}\varepsilon_i$, we can write

$$\begin{aligned}
\kappa_{i+1}^2 K_{i+1}^2 \|Z_{i+1}\| \leqslant& \frac{\kappa_{i+1}^2 K_{i+1}^2}{\kappa_i^3 K_i^3}4\varepsilon_i^2 + \frac{\kappa_{i+1}^2 K_{i+1}^2}{\kappa_i^4 K_i^4}4\varepsilon_i^3 + \frac{\kappa_{i+1}^2 K_{i+1}^2}{\kappa_i^2 K_i^2}4\varepsilon_i^2 \\
\leqslant& 4\left(2+\varepsilon_i\right)\left(\frac{\kappa_{i+1}K_{i+1}}{\kappa_i K_i}\right)^2\varepsilon_i^2 \\
\leqslant& 4\left(2+\varepsilon_i\right)\left(\frac{1+2a\varepsilon}{1-4a\varepsilon}\right)^2\varepsilon_i^2
\end{aligned}$$

On the other hand

$$\Delta_{i+1} = \Delta_i X_i + Y_i \Delta_i + Y_i\left(\Delta_i + \Sigma_i\right)X_i.$$

Hence

$$\begin{aligned}
\kappa_{i+1}^2 K_{i+1} \|\Delta_{i+1}\| \leqslant& \frac{\kappa_{i+1}^2 K_{i+1}}{\kappa_i^3 K_i^2}4\varepsilon_i^2 + \frac{\kappa_{i+1}^2 K_{i+1}}{\kappa_i^4 K_i^3}4\varepsilon_i^3 + \frac{\kappa_{i+1}^2 K_{i+1}}{\kappa_i^2 K_i}4\varepsilon_i^2 \\
\leqslant& 4\left(2+\varepsilon_i\right)\frac{\kappa_{i+1}^2 K_{i+1}}{\kappa_i^2 K_i}\varepsilon_i^2 \\
\leqslant& 4\left(2+\varepsilon_i\right)\frac{1+2a\varepsilon}{(1-4a\varepsilon)^2}\varepsilon_i^2
\end{aligned}$$

It follows

$$\begin{aligned}\varepsilon_{i+1} \leqslant & 4(2+\varepsilon)\left(\frac{1+2a\varepsilon}{1-4a\varepsilon}\right)^2 \varepsilon_i^2 \\ \leqslant & 8(2+\varepsilon)\left(\frac{1-6\varepsilon}{1-12\varepsilon}\right)^2 \varepsilon\, 2^{1-2^{i+1}} \varepsilon \\ \leqslant & 2^{1-2^{i+1}} \varepsilon \qquad \text{since } 8(2+\varepsilon)\left(\frac{1-6\varepsilon}{1-12\varepsilon}\right)^2 \varepsilon \leqslant 1 \text{ for } \varepsilon \leqslant 0.033.\end{aligned}$$

This completes the proof of the two induction hypothesis (22–23) at order $i+1$. Let $W_i = \prod_{k=0}^{i}(I_n + X_k)$. Since

$$\begin{aligned}\|X_k\| \leqslant & \frac{2}{\kappa_k K_k}\varepsilon_k \\ \leqslant & \frac{2(1+8a\varepsilon)}{\kappa K(1-4a\varepsilon)}\varepsilon 2^{1-2^k} \\ \leqslant & \frac{2}{\kappa K(1-12\varepsilon)}\varepsilon 2^{1-2^k}\end{aligned}$$

Consequently,

$$\begin{aligned}\|W_\infty - I_n\| \leqslant & \prod_{i\geqslant 0}\left(1+\frac{2}{\kappa K(1-12\varepsilon)}\varepsilon 2^{1-2^i}\right) - 1 \\ \leqslant & \frac{4}{\kappa K(1-12\varepsilon)}\varepsilon \quad \text{from Lemma 1} \\ \leqslant & \frac{0.22}{\kappa K} \qquad \text{since } \varepsilon \leqslant 0.033..\end{aligned}$$

Hence W_∞ is invertible and $E_0 = E_\infty W_\infty^{-1}$. This implies that E_0 is invertible. Moreover,

$$\begin{aligned}\|W_i - W_\infty\| \leqslant & \|W_i\| \left\| 1 - \prod_{k\geqslant i+1}(1+\|X_k\|) \right\| \\ \leqslant & \left(1+\|W_i - I_n\|\right)\left\| \prod_{k\geqslant 0}\left(1+\frac{2}{\kappa K(1-12\varepsilon)}\varepsilon \times 2^{1-2^{k+i+1}}\right) - 1 \right\| \\ \leqslant & (1+0.22)\times\frac{4}{\kappa K(1-12\varepsilon)}\times 2^{1-2^{i+1}}\varepsilon \quad \text{from Lemma 1} \\ \leqslant & \frac{8.1}{\kappa K}\times 2^{1-2^{i+1}}\varepsilon.\end{aligned}$$

We deduce that

$$\|E_i - E_\infty\| \leqslant \frac{8.1}{\kappa K}\times 2^{1-2^{i+1}}\|E_0\|\varepsilon.$$

In the same way we show that F_0 is invertible and

$$\|F_i - F_\infty\| \leqslant \frac{8.1}{\kappa K} \times 2^{1-2^{i+1}} \|F_0\| \varepsilon.$$

Finally

$$\begin{aligned}
\|\Sigma_i - \Sigma_\infty\| &\leqslant \sum_{k \geqslant i} \|\Sigma_{k+1} - \Sigma_k\| \\
&\leqslant \sum_{k \geqslant i} \frac{2}{\kappa_k^2 K_k} \varepsilon_k \\
&\leqslant \left(\sum_{k \geqslant 0} 2^{-2^k}\right) 2^{1-2^i} \frac{2}{\kappa^2 K(1-12\varepsilon)(1-8\varepsilon)} \varepsilon \\
&\leqslant 0.82 \times 2.25 \times 2^{1-2^i} \frac{\varepsilon}{\kappa K} \qquad \text{since } \sum_{k \geqslant 0} 2^{-2^k} \leqslant 0.82 \text{ and } \varepsilon \leqslant 0.033. \\
&\leqslant 1.85 \times 2^{1-2^i} \varepsilon_0.
\end{aligned}$$

The theorem is proved. □

Proposition 4 *The complexity of one Newton iteration in Theorem 3 is in $\mathcal{O}(n^3)$.*

Proof The computation of all the entries $x_{i,j}$, $y_{i,j}$ of X_i and Y_i by the formulas (13–17) requires in total $\mathcal{O}(n^2)$ arithmetic operations. The computation of $Z_i, \Delta_i, S_i, E_{i+1}, F_{i+1}$, which requires 6 backward stable matrix multiplications and diagonal matrix operations, has a complexity in $\mathcal{O}(n^3)$. Consequently, the complexity of each iteration is in $\mathcal{O}(n^3)$. □

Remark 1 It is possible to generalize this approach to the case where the diagonal matrices are replaced by Jordan matrices.

4 Newton-like method for two simultaneously diagonalizable matrices

Let M_1, M_2 be two commuting matrices in $\mathcal{W}_n$, thus M_1 and M_2 are simultaneously diagonalizable. We aim to find $E, F \in GL_n$ which diagonalize simultaneously M_1, M_2 so that: $FM_kE = \Sigma_k \mid k \in \{1,2\}$, and $\Sigma_1, \Sigma_2 \in \mathcal{D}'_n$. This equivalent to find the numerical solution of $f(E,F,\Sigma_1,\Sigma_2) = 0$ such that $f : (E,F,\Sigma_1,\Sigma_2) \mapsto (FM_1E - \Sigma_1, FM_1E - \Sigma_1)$

We consider as before the perturbations $E + EX$, $F + YF$ and $\Sigma_k + S_k$ of respectively E, F and Σ_k for $k \in \{1,2\}$. Letting $Z_k = \mathrm{FM}_k E - \Sigma_k$ for $k = 1,2$, we have:

$$\begin{aligned}
&(F + \mathrm{YF})M_k(E + \mathrm{EX}) - (\Sigma_k + S_k) \\
&= Z_k - S_k + \Sigma_k X + Y\Sigma_k + Z_k X + \mathrm{YZ}_k + Y(Z_k + \Sigma_k)X
\end{aligned} \tag{24}$$

By assuming Z_1, Z_2 are of small norm, the linear system to solve from Equation (24) is the following

$$Z_k - S_k + \Sigma_k X + Y\Sigma_k = 0, \qquad k = 1, 2 \tag{25}$$

A solution of (25) is given by the following lemma.

Lemma 3 *Let* $\Sigma_k = \mathrm{diag}(\sigma_1^k, \ldots, \sigma_n^k)$, $Z_k = (z_{i,j}^k)_{1 \le i,j \le n}$ *be given matrices in* $\mathbb{C}^{n \times n}$ *for* $k \in \{1, 2\}$. *Assume that* $\begin{vmatrix} \sigma_j^1 & \sigma_j^2 \\ \sigma_i^1 & \sigma_i^2 \end{vmatrix} \neq 0$ *for* $i \neq j$. *Let* X, Y, *and* S_k *be the matrices defined by*

$$x_{i,i} = 0 \tag{26}$$

$$x_{i,j} = \frac{\begin{vmatrix} \sigma_j^1 & z_{i,j}^1 \\ \sigma_j^2 & z_{i,j}^2 \end{vmatrix}}{\begin{vmatrix} \sigma_i^1 & \sigma_j^1 \\ \sigma_i^2 & \sigma_j^2 \end{vmatrix}}, \qquad i \neq j \tag{27}$$

$$y_{i,i} = 0 \tag{28}$$

$$y_{i,j} = -\frac{\begin{vmatrix} \sigma_i^1 & z_{i,j}^1 \\ \sigma_i^2 & z_{i,j}^2 \end{vmatrix}}{\begin{vmatrix} \sigma_i^1 & \sigma_j^1 \\ \sigma_i^2 & \sigma_j^2 \end{vmatrix}}, \qquad i \neq j \tag{29}$$

$$S_k = \mathrm{diag}(Z_k), \qquad k = 1, 2. \tag{30}$$

Then we have

$$Z_k - S_k + \Sigma_k X + Y\Sigma_k = 0, \qquad k = 1, 2 \tag{31}$$

Moreover

$$\|X\|, \|Y\| \leqslant 2\kappa\varepsilon K \tag{32}$$

where $\varepsilon = \max(\|Z_1\|, \|Z_2\|)$, $\kappa = \max\left(1, \max_{i \neq j} \dfrac{1}{\begin{vmatrix} \sigma_i^1 & \sigma_j^1 \\ \sigma_i^2 & \sigma_j^2 \end{vmatrix}}\right)$, $K = \max(1, \max_{i,k} |\sigma_i^k|)$.

Proof It is easy to verify that the Eq. (31) implies that for $i \neq j$,

$$\sigma_i^k x_{i,j} + \sigma_j^k y_{i,j} + z_{i,j}^k = 0$$

and that the solution of these equations is given by the formula (27), (29). Choosing $x_{i,i} = y_{i,i} = 0$, we take $S_k = \mathrm{diag}(Z_k + \Sigma_k X + Y\Sigma_k) = \mathrm{diag}(Z_k)$ since $\Sigma_k X + Y\Sigma_k$ is an off-matrix, to satisfy the Eq. (31). The bounds (32) follows easily from (27), (29). ☐

Theorem 5 *Let* $E_0, F_0 \in GL_n$ *and* $\Sigma_{0,k} = \operatorname{diag}(\sigma^k_{0,1}, \dots, \sigma^k_{0,n}) \in \mathcal{D}'_n$, $k = 1, 2$, *be given and let define the sequences for* $i \geqslant 0$ *and* $k = 1, 2$ *by*:

$$\begin{aligned} Z_{i,k} &= F_i M_k E_i - \Sigma_{i,k} \\ S_{i,k} &= \operatorname{diag}\left(Z_{i,k}\right) \\ E_{i+1} &= E_i\left(I_n + X_i\right) \\ F_{i+1} &= \left(I_n + Y_i\right)F_i \\ \Sigma_{i+1,k} &= \Sigma_{i,k} + S_{i,k}, \end{aligned}$$

where X_i, Y_i *are defined by the formulas* (26–29). *Let* $\varepsilon_0 = \max(\|Z_{0,1}\|, \|Z_{0,2}\|)$, $\kappa_0 = \max\left(1, \max_{i\neq j} \frac{1}{\begin{vmatrix} \sigma^1_{0,i} & \sigma^1_{0,j} \\ \sigma^2_{0,i} & \sigma^2_{0,j}\end{vmatrix}}\right)$ *and* $K_0 = \max(1, \max_{j,k} |\sigma^k_{0,j}|)$. *Assume that*

$$u := 4\varepsilon_0 \kappa_0^2 K_0^3 \leqslant 0.094. \tag{33}$$

Then the sequences $(\Sigma_{i,k}, E_i, F_i)_{i\geqslant 0}$ *converge quadratically to the solution of* $FM_kE - \Sigma_k$ *for* $k = 1, 2$. *More precisely* E_0 *and* F_0 *are invertible and*

$$\begin{aligned} \|E_i - E_\infty\| &\leqslant 1.46 \times 2^{1-2^{i+1}} \|E_0\| u \\ \|F_i - F_\infty\| &\leqslant 1.46 \times 2^{1-2^{i+1}} \|F_0\| u. \end{aligned}$$

Proof Let us denote for each $i \geqslant 0$,

$$\begin{aligned} \varepsilon &= \varepsilon_0 & \varepsilon_i &= \max(\|Z_{i,1}\|, \|Z_{i,2}\|) \\ \kappa &= \kappa_0 & \kappa_i &= \max\left(1,\ \max_{1\leqslant j<k\leqslant n} \frac{1}{\begin{vmatrix} \sigma^1_{i,j} & \sigma^1_{i,k} \\ \sigma^2_{i,j} & \sigma^2_{i,k}\end{vmatrix}}\right) \\ K &= K_0 & K_i &= \max\left(1, \max_{j,k}\left(|\sigma^k_{i,j}|\right)\right), \end{aligned}$$

where $\sigma^k_{i,1}, \dots, \sigma^k_{i,n}$ are the diagonal entries of $\Sigma_{i,k}$. Let us show by induction on i that

$$\varepsilon_i \leqslant 2^{1-2^i} \varepsilon \tag{34}$$

$$\|\Sigma_{i,k} - \Sigma_{0,k}\| \leqslant \left(2 - 2^{2-2^i}\right)\varepsilon \tag{35}$$

These inequalities clearly hold for $i = 0$. Assuming that the induction hypothesis holds for a given i and let us prove it for $i + 1$. We can notice that $\varepsilon_i \leq 1$. In fact by induction hypothesis, we have $\varepsilon_i \leq 2^{1-2^i}\varepsilon_0$ and from (33) $\varepsilon_0 = \frac{u}{4\kappa_0^2 K_0^3} \leq 1$, since $u \leq 1$

and $\kappa_0, K_0 \geq 1$. As $2^{1-2^i} \leq 1$, $\forall i \geq 0$, we have $\varepsilon_i \leq 1$. We first prove that $\|\Sigma_{i+1,k} - \Sigma_{0,k}\| \leqslant (2 - 2^{2-2^{i+1}})\varepsilon$ under the assumption $\|\Sigma_{i,k} - \Sigma_{0,k}\| \leqslant (2 - 2^{2-2^i})\varepsilon$. To do this, at the first step we show that this implies $K_i \leq K + 2\varepsilon$ and $\kappa_i \leq \dfrac{\kappa}{1 - 8\kappa\varepsilon(K+\varepsilon)}$. Let us prove $K_i \leqslant K + 2\varepsilon$. We have

$$
\begin{aligned}
K_i := \|\Sigma_i\| &\leq \|\Sigma_0\| + \|\Sigma_i - \Sigma_0\| \\
&\leqslant K + \left(2 - 2^{2-2^i}\right)\varepsilon \\
&\leqslant K + 2\varepsilon.
\end{aligned}
$$

Let us show that $\kappa_i \leq \dfrac{\kappa}{1 - 8\kappa\varepsilon(K+\varepsilon)}$. In fact, if the σ_{i,j^k}'s are the diagonal values of Σ_i^k, we have $|\sigma_{i,j}^k - \sigma_{0,j}^k| \leqslant \|\Sigma_{i,k} - \Sigma_{0,k}\| \leqslant 2\varepsilon$ for $1 \leqslant j \leqslant n$ and $k = 1, 2$. It follows :

$$
\begin{aligned}
|\sigma_{i,j}^1\sigma_{i,k}^2 - \sigma_{0,j}^1\sigma_{0,k}^2| =& |\sigma_{i,j}^1\sigma_{i,k}^2 - \sigma_{0,j}^1\sigma_{i,k}^2 + \sigma_{0,j}^1\sigma_{i,k}^2 - \sigma_{0,j}^1\sigma_{0,k}^2| \\
=& |\sigma_{i,k}^2(\sigma_{i,j}^1 - \sigma_{0,j}^1) + \sigma_{0,j}^1(\sigma_{i,k}^2 - \sigma_{0,k}^2)| \\
\leqslant& 2\varepsilon|\sigma_{i,k}^2| + 2\varepsilon|\sigma_{0,j}^1| \\
\leqslant& 2\varepsilon(K + 2\varepsilon) + 2\varepsilon K = 4\varepsilon(K+\varepsilon).
\end{aligned}
$$

Now,

$$
\begin{aligned}
&|\sigma_{i,j}^1\sigma_{i,k}^2 - \sigma_{i,k}^1\sigma_{i+1,j}^2| \geqslant \\
|\sigma_{0,j}^1\sigma_{0,k}^2 - \sigma_{0,k}^1\sigma_{0,j}^2| - |\sigma_{0,j}^1\sigma_{0,k}^2 - \sigma_{i+1,j}^1\sigma_{i,k}^2| - &|\sigma_{i,k}^1\sigma_{i,j}^2 - \sigma_{0,k}^1\sigma_{0,j}^2| \geqslant \\
&|\sigma_{0,j}^1\sigma_{0,k}^2 - \sigma_{0,k}^1\sigma_{0,j}^2|(1 - 8k\varepsilon(K+\varepsilon)).
\end{aligned}
$$

Finally, we get :

$$
\kappa_i \leqslant \frac{\kappa}{1 - 8\kappa\varepsilon(K+\varepsilon)}.
$$

To prove (35) it is sufficient to write

$$
\begin{aligned}
\|\Sigma_{i+1,k} - \Sigma_{0,k}\| &\leqslant \|S_{i,k}\| + \|\Sigma_{i+1,k} - \Sigma_{0,k}\| \\
&\leqslant \varepsilon_i + \left(2 - 2^{2-2^i}\right)\varepsilon \\
&\leqslant \left(2^{1-2^i} + 2 - 2^{2-2^i}\right)\varepsilon \leqslant \left(2 - 2^{2-2^{i+1}}\right)\varepsilon.
\end{aligned}
$$

Let us prove (34). Since we have

$$
Z_{i+1,k} = Z_{i,k}X_i + Y_i Z_{i,k} + Y_i\left(Z_{i,k} + \Sigma_{i,k}\right)X_i.
$$

we deduce

$$\begin{aligned}\|Z_{i+1,k}\| &\leqslant 2\varepsilon_i^2\kappa_i K_i + 2\varepsilon_i^2\kappa_i K_i + 4\varepsilon_i^2\kappa_i^2 K_i^2\left(\varepsilon_i + K_i\right)\\ &\leqslant 4\varepsilon_i^2\kappa_i^2 K_i + 4\varepsilon_i^2\kappa_i^2 K_i^2\left(1 + K_i\right) \quad \text{since } \varepsilon_i \leqslant 1 \text{ and } \kappa_i \geqslant 1\\ &\leqslant 3 \times 4\varepsilon_i^2\kappa_i^2 K_i^3 = 12\varepsilon_i^2\kappa_i^2 K_i^3 \qquad \text{since } K_i \geqslant 1.\end{aligned}$$

It follows

$$\begin{aligned}\varepsilon_{i+1} &\leqslant \frac{12\kappa^2(K+2\varepsilon)^3}{(1-8\kappa\varepsilon(K+\varepsilon))^2}\varepsilon_i^2 \quad \leqslant \frac{12\varepsilon\kappa^2(K+2\varepsilon)^3}{(1-8\kappa\varepsilon(K+\varepsilon))^2}2^{2-2^{i+1}}\varepsilon\\ &\leqslant 3\frac{\left(1+\frac{u}{2}\right)^3}{\left(1-2u\left(1+\frac{u}{4}\right)\right)^2}u2^{2-2^{i+1}}\varepsilon \quad \text{since} \quad \frac{\varepsilon}{K} \leqslant \frac{u}{4}, \kappa\varepsilon \leqslant \frac{u}{4}\\ &\leqslant 2^{1-2^{i+1}}\varepsilon \quad \text{since} \quad 3\frac{\left(1+\frac{u}{2}\right)^3}{\left(1-2u\left(1+\frac{u}{4}\right)\right)^2} \leqslant 2^{-1}\text{for} u \leqslant 0.094.\end{aligned}$$

Let $W_i = \prod_{k=0}^{i}(I_n + X_k)$. Since

$$\begin{aligned}\|X_l\| &\leqslant 2\kappa_l K_l \varepsilon_l\\ &\leqslant 2\frac{\kappa}{1-8\kappa\varepsilon(K+\varepsilon)}(K+2\varepsilon)\varepsilon 2^{1-2^l}\\ &\leqslant \frac{\left(1+\frac{u}{2}\right)u}{2\left(1-2u\left(1+\frac{u}{4}\right)\right)}2^{1-2^l}\\ &\leqslant 0.65 \times 2^{1-2^l}u \quad \text{since} u \leqslant 0.094.\end{aligned}$$

Consequently,

$$\begin{aligned}\|W_\infty - I_n\| &\leqslant \prod_{i\geqslant 0}(1 + 0.65 \times 2^{1-2^i}u) - 1\\ &\leqslant 1.3u \quad \text{from Lemma 1}\\ &\leqslant 1.3 \times 0.094 = 0.1222\end{aligned}$$

Hence W_∞ is invertible and $E_0 = E_\infty W_\infty^{-1}$. This implies that E_0 is invertible. Moreover,

$$
\begin{aligned}
\|W_i - W_\infty\| \leqslant& \|W_i\| \left\| 1 - \prod_{k \geqslant i+1} \left(1 + \|X_k\|\right) \right\| \\
\leqslant& \left(1 + \|W_i - I_n\|\right) \left\| \prod_{k \geqslant 0} \left(1 + 0.059 \times 2^{1-2^{k+i+1}}\right) - 1 \right\| \\
\leqslant& (1 + 0.1222) \times 1.3 \times 2^{1-2^{i+1}} u \\
\leqslant& 1.46 \times 2^{1-2^{i+1}} u.
\end{aligned}
$$

We deduce that

$$
\|E_i - E_\infty\| \leqslant 1.46 \times 2^{1-2^{i+1}} \|E_0\| u.
$$

In the same way we show that F_0 is invertible and

$$
\|F_i - F_\infty\| \leqslant 1.46 \times 2^{1-2^{i+1}} \|F_0\| u.
$$

The theorem is proved. □

5 Convergence of a pencil of simultaneously diagonalizable matrices.

In this section we present two strategies to solve the system (1) of a pencil of commuting matrices $(M_i)_{1 \le i \le p}$ in $\mathcal{W}_n$. The first strategy is trivial and consists of finding the common diagonalizers E and F of the pencil by numerically solving one of the systems $(FE - I_n, FM_1E - \Sigma_1) = 0$ or $(FM_1E - \Sigma_1, FM_2E - \Sigma_1) = 0$ using Theorem 3 or Theorem 5. Next we deduce the remaining diagonal matrices Σ_i using the formulas

$$
\Sigma_{i,k} = \frac{E(:,k)^* M_i E(:,k)}{E(:,k)^* E(:,k)} \qquad 1 \leqslant k \leqslant n, \quad 2 \text{ or } 3 \leqslant i \leqslant p,
$$

where $E(: , k)$ is the *k-th* column in E.

In this strategy we use that a diagonalizer of one or two matrices of the pencil can diagonalize the other matrices of the pencil. We note that, in general, we don't have this property for simultaneously diagonalizable matrices, where, for instance, it is possible to find a diagonalizer of M_1 which is not a common diagonalizer for the other matrices of the pencil. Nevertheless, this property holds here since we suppose that the matrices M_i have simple eigenvalues.

Another strategy is to find a "good" linear combination of the M_i's. This is based on Lemma 4 and Theorem 6.

Lemma 4 *Let us suppose that the M_i commute pairwise and they are linearly independent, i.e., $\sum_{i=1}^{p} a_i M_i = 0 \Rightarrow a_i = 0, i = 1 : p$. Let $E \in GL_n$ and $\Sigma_i \in \mathcal{D}'_n$ be such that*

$$E^{-1}M_iE - \Sigma_i = 0, \quad i = 1 : p.$$

Let $S \in \mathbb{C}^{n\times p}$ and the column i of S is the diagonal of Σ_i. Let $\sigma = (\sigma_1, \dots, \sigma_n)$ and $\Sigma = \operatorname{diag}(\sigma)$. Then the matrix S has a full rank and $\alpha = (S^*S)^{-1}S^*\sigma$ satisfies

$$\sum_{i=1}^{p} \alpha_i E^{-1}M_iE - \Sigma = 0.$$

Proof Since the matrices M_i are simultaneously diagonalizable there exists E be such that $E^{-1}M_iE - \Sigma_i = 0$. The condition

$$\sum_{i=1}^{p} \alpha_i \Sigma_i - \Sigma = 0$$

is written as $S\alpha = \sigma$ where $S \in \mathbb{C}^{n\times p}$. The assumption $\sum_{i=1}^{p} a_i M_i = 0 \Rightarrow a_i = 0, i = 1 : p$ implies that the matrix has a full rank. Consequently,

$$\alpha = (S^*S)^{-1}S^*\sigma.$$

The lemma follows. □

Theorem 6 *Let $M_1, \dots, M_p \in \mathbb{C}^{n\times n}$ be p simultaneously diagonalizable matrices and verify the assumption of linearly independent. Let us consider matrices E_0, F_0 and $\Sigma_{0,i} = \operatorname{diag}(F_0ME_0)$, $i = 1 : p$. Let us define the matrix $S \in \mathbb{C}^{n\times p}$ in which the column i is the diagonal of $\Sigma_{0,i}$. Let $\sigma = \left(1, e^{\frac{2i\pi}{n}}, \dots, e^{\frac{2i(n-1)\pi}{n}}\right)$, $\Sigma = \operatorname{diag}(\sigma)$ and $\alpha = (S^*S)^{-1}S^*\sigma$. We consider the system*

$$\begin{pmatrix} EF - I_n \\ FME - \Sigma \end{pmatrix} = 0 \tag{36}$$

where $M = \sum_{i=1}^{p} \alpha_i M_i$. If

$$n^2 \max\left(\|Z_0\|, \|\Delta_0\|\right) \leqslant 16 \times 0.033$$

then (F_0, E_0, Σ) satisfies the condition (21) of Theorem 3.

Proof In this case the quantity κ defined in the Theorem 3 is equal to

$$\begin{aligned} \kappa =& \frac{1}{2\,|\sin\left(\frac{\pi}{n}\right)|} \\ \leqslant& \frac{n}{4} \quad \text{since } |\sin\left(\frac{\pi}{n}\right)| \geqslant \frac{2}{n} \text{ for } n \geqslant 2. \end{aligned}$$

Since $K_0 = 1$ we get

$$\varepsilon_0 = \max(\kappa_0^2 K_0^2 \|Z_0\|, \kappa_0^2 K_0 \|\Delta_0\|) \leq \frac{n^2}{16} \max(\|Z_0\|, \|\Delta_0\|).$$

The condition

$$\max(\|Z_0\|, \|\Delta_0\|) \leq 0.033 \, \frac{16}{n^2},$$

gives the result. □

6 Numerical illustration

We use a Julia implementation of the Newton sequences in the numerical experiments. The experimentation has been done on a Dell Windows desktop with 8 GB memory and Intel 2.3 GHz CPU. We use the Julia package `ArbNumerics` for the computation in high precision.

6.1 Simulation

In this section we apply the Newton iterations presented in Theorem 3 (resp. Theorem 5) on examples of diagonalizable matrices (resp. of two simultaneously diagonalizable matrices). We validate experimentally the sufficiency of the condition established in Theorem 3 (resp. Theorem 5) to have a quadratic sequence (Tables 1, 2, 3 and 4). On the other hand, as this condition is sufficient but not necessary, we show through some other examples how this Newton sequence starting from an initial point which is not verifying this condition could converge quadratically (Tables 5, 6, 7 and 8). We note that the the computation in the aforementioned tables is done in high precision. Nevertheless, we test also the two Newton-type sequences using machine precision (Tables 9 and 10) and this to show that these sequences have the same numerical behavior of a classical Newton method, i.e., if the solution is in the neighborhood of the initial point the Newton-type iterations will converge

Table 1 The computational results throughout 7 iterations of an example of implementation of *Test-1* with $\mathbb{K} = \mathbb{R}$, $n = 10$ and $e = 6$ in precision 1024

Iteration	$\varepsilon := \max(\kappa_0^2 K_0^2 \|Z_0\|, \kappa_0^2 K_0 \|\Delta_0\|) \leq 0.033$	err_{res}
1	0.00131	9.33*e*–6
2	2.39*e*–8	1.06*e*–10
3	1.68*e*–18	7.49*e*–21
4	2.93*e*–38	1.31*e*–40
5	4.21*e*–78	1.87*e*–80
6	1.17*e*–157	5.24*e*–160
7	4.16*e*–288	6.20*e*–293

Table 2 The computational results throughout 7 iterations of an example of implementation of *Test-1* with $\mathbb{K} = \mathbb{C},\ n = 10$ and $e = 6$ in precision 1024

Iteration	$\varepsilon := \max(\kappa_0^2 K_0^2 \|Z_0\|, \kappa_0^2 K_0 \|\Delta_0\|) \leq 0.033$	err_{res}
1	0.02581	2.76e–4
2	3.49e–6	2.33e–8
3	9.51e–14	6.34e–16
4	5.31e–29	3.54e–31
5	1.96e–59	1.31e–61
6	3.02e–120	2.01e–122
7	4.58e–242	3.05e–244

Table 3 The computational results throughout 7 iterations of an example of implementation of *Test-2* with $\mathbb{K} = \mathbb{R},\ n = 10$ and $e = 6$ in precision 1024

Iteration	$4\kappa^2 K^3 \varepsilon \leq 0.094$	err_{res}
1	7.65e–2	6.72e–6
2	1.73e–7	1.52e–11
3	5.58e–18	4.90e–22
4	5.49e–39	4.82e–43
5	3.10e–81	2.73e–85
6	2.28e–165	2.01e–169
7	2.20e–279	1.94e–283

Table 4 The computational results throughout 7 iterations of an example of implementation of *Test-2* with $\mathbb{K} = \mathbb{C},\ n = 10$ and $e = 6$ in precision 1024

Iteration	$4\kappa^2 K^3 \varepsilon \leq 0.094$	err_{res}
1	6.86e–3	9.16e–6
2	7.14e–9	9.53e–12
3	9.51e–21	1.26e–23
4	6.69e–44	8.92e–47
5	3.77e–90	5.04e–93
6	2.59e–182	3.45e–185
7	1.65e–281	2.20e–284

Table 5 The residual error throughout 7 iterations given by the implementation of *Test-1* with $\mathbb{K} = \mathbb{R}, e = 3$ and $n = 10, 50, 100$ in precision 1024

Iteration	$n = 10$	$n = 50$	$n = 100$
1	8.57e–3	7.93e–2	3.22e–2
2	1.91e–4	5.76e–2	1.38e–2
3	1.58e–8	6.19e–3	6.12e–4
4	4.79e–16	8.74e–5	5.42e–7
5	3.56e–31	1.31e–8	3.83e–13
6	1.39e–61	2.39e–16	1.80e–25
7	1.91e–122	7.03e–32	3.81e–50

Table 6 The residual error throughout 7 iterations given by the implementation of *Test-1* with $\mathbb{K} = \mathbb{C}, e = 3$ and $n = 10, 50, 100$ in precision 1024

Iteration	$n = 10$	$n = 50$	$n = 100$
1	8.84e–3	9.75e–2	1.61e–2
2	8.59e–6	6.39e–5	1.03e–4
3	3.91e–11	3.99e–9	4.68e–9
4	9.87e–22	1.87e–17	3.13e–17
5	7.60e–43	4.42e–34	8.84e–34
6	5.14e–85	2.50e–67	9.45e–67
7	2.64e–169	8.28e–134	1.05e–132

Table 7 The residual error throughout 7 iterations given by the implementation of *Test-2* with $\mathbb{K} = \mathbb{R}, e = 3$ and $n = 10, 50, 100$ in precision 1024

Iteration	$n = 10$	$n = 50$	$n = 100$
1	2.91e–2	4.57e–3	1.01e–2
2	7.97e–5	1.03e–6	1.31e–6
3	4.21e–9	1.69e–11	3.71e–11
4	1.07e–16	2.42e–23	1.23e–22
5	3.92e–33	1.18e–44	1.46e–43
6	2.63e–64	1.02e–89	1.67e–86
7	1.71e–128	3.20e–177	9.01e–172

Table 8 The residual error throughout 7 iterations given by the implementation of *Test-2* with $\mathbb{K} = \mathbb{C}, e = 3$ and $n = 10, 50, 100$ in precision 1024

Iteration	$n = 10$	$n = 50$	$n = 100$
1	7.33e–3	3.14e–3	5.52e–3
2	3.49e–6	7.48e–7	1.35e–6
3	2.91e–12	1.11e–13	1.19e–13
4	2.04e–24	2.54e–27	1.68e–27
5	8.23e–49	3.04e–54	2.19e–54
6	1.88e–97	3.41e–108	1.50e–108
7	1.31e–194	1.91e–215	4.53e–216

Table 9 The residual error throughout 5 iterations given by the implementation of *Test-1* with $\mathbb{K} = \mathbb{R}, e = 3$ and $n = 10, 20, 30$, in double precision

Iteration	$n = 10$	$n = 20$	$n = 30$
1	$4.78e-3$	$1.01e-2$	$1.01e-2$
2	$4.71e-3$	$2.55e-3$	$1.14e-3$
3	$2.29e-5$	$1.97e-5$	$4.08e-7$
4	$1.43e-9$	$2.36e-10$	$2.26e-13$
5	$4.06e-15$	$1.23e-14$	$5.04e-14$
$\Vert M - E_{\texttt{eigen}}\Sigma_{\texttt{eigen}}E_{\texttt{eigen}}^{-1}\Vert_{Frob}$	$9.49e-15$	$2.83e-14$	$7.45e-14$
$\Vert M - E_{\texttt{newton}}\Sigma_{\texttt{newton}}E_{\texttt{newton}}^{-1}\Vert_{Frob}$	$2.96e-15$	$1.01e-14$	$3.42e-14$

Table 10 The residual error throughout 5 iterations given by the implementation of *Test-2* with $\mathbb{K} = \mathbb{R}$, $e = 3$ and $n = 10, 20, 30$, in double precision

Iteration	$n = 10$	$n = 20$	$n = 30$
1	2.71e–3	1.21e—2	4.64e—3
2	1.36e–6	4.91e—6	2.24e—6
3	1.39e–12	2.57e—11	4.74e—11
4	6.16e–15	8.97e–14	1.55e–13
5	7.04e–15	8.09e–14	1.53e–13
$\max(\|M_1 - E\Sigma_1 E^{-1}\|_{Frob}, \|M_2 - E\Sigma_2 E^{-1}\|_{Frob})$	3.74e–15	4.13e–14	8.21e–14

towards this solution with a few number of iterations and the residual error obtained at the end is in double precision.

This allows us to have an heuristic estimation on the numerical dependency of the Newton sequences from this condition to converge. Furthermore, these examples reveal the possibility of achieving computation in such problem with high precision. For example, in the case of a diagonalizable matrix of simple eigenvalues, we can compute its eigenvalues using one of the solvers which works with a double precision. Then we take this point as an initial point for the Newton sequence of Theorem 3 in order to increase the precision. Hereafter, we give some details about the tests: *Test-1* for Theorem 3 and *Test-2* for Theorem 5, considered in this section.

Test-1. Let $\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$, $M = E\Sigma E^{-1} + 10^{-e}A$, where $e \in \{3, 6\}$. The matrices E, Σ, and $A \in \mathbb{K}^{n\times n}$ are chosen randomly following standard normal distributions such that E is invertible, Σ is diagonal with n different diagonal entries and A is a random square matrix obeying normal distribution of size n and Frobenius norm equal to 1. Since M is a small perturbation of $E\Sigma E^{-1}$, more precisely $\|M - E\Sigma E^{-1}\|_{Frob} = 10^{-e}$, M is a diagonalizable matrix of simple eigenvalues. Herein, we apply the Newton iteration of Theorem 3 on M with initial point $E_0 = E$, $F0 = E^{-1}$ and $\Sigma_0 = \Sigma$. The residual error reported in this test at iteration k is given by:

$$\mathrm{err}_{res} = \max\left(\|F_k E_k - I_n\|, \|F_k M E_k - \Sigma_k\|\right).$$

Test-2. Let $\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$, $M_1 = F^{-1}\Sigma_1 E^{-1}$, $M_2 = F^{-1}\Sigma_2 E^{-1}$, where E, F, Σ_1 and $\Sigma_2 \in \mathbb{K}^{n\times n}$ are randomly sampled according to standard normal distributions, such that E and F are invertible, Σ_1 and Σ_2 are diagonal with n different diagonal entries. The Newton iteration in Theorem 5 is applied on M_1 and M_2 with initial point E_0, F_0, $\Sigma_{0,1}$ and $\Sigma_{0,2}$, such that these matrices are obtained by applying a small perturbation on respectively E, F, Σ_1 and Σ_2 as follows:

$E_0 = E + 10^{-e}A$, $F_0 = F + 10^{-e}B$, $\Sigma_{0,1} = \Sigma_1 + 10^{-e}C$, $\Sigma_{0,2} = \Sigma_2 + 10^{-e}D$, where $e \in \{3, 6\}$, A and B (resp. C and D) are random square matrices (resp. random diagonal matrices with different diagonal entries) of size n and Frobenius norm equal to 1, with entries in $\mathbb{K}$ following standard normal distributions. The residual error reported in this test at iteration k is given by:

$$\mathrm{err}_{res} = \max\left(\|F_k M_1 E_k - \Sigma_{k,1}\|, \|F_k M_2 E_k - \Sigma_{k,2}\|\right).$$

We notice that the condition established in Theorem 3 (resp. Theorem 5) is reached in *Test-1* (resp. *Test-2*) for matrices of size 10 with order of perturbation equal to 10^{-6}, and we can see in Tables 1, 2, 3 and 4 that the Newton sequences with initial point verifying the condition in the associated theorem converge quadratically. We can notice also that by increasing the perturbation up to 10^{-3} (the initial point does not verify the condition in the associated theorem), the Newton sequences converge quadratically for different sizes of matrices $n = 10,\ 50,\ 100$ (see Tables 5, 6, 7 and 8). Moreover, we can notice in Table 9 the Newton-type iteration of Theorem 3 applied in double precision converges with a few number of iterations ~ 5 and the final residual error measured with the Frobenius norm is of order machine precision $\sim 10^{-14}$ and it is of the same order obtained by the standard Julia method `eigen` to compute the eigen decomposition. The same remarks are valid for Table 10 where the Newton-type sequence of Theorem 5 needs, in double precision, a few iterations to converges towards the solution given by using the Frobenius norm a residual error of order machine precision.

6.2 Cauchy matrix

In this section we present an example for a Cauchy matrix of size $n = 13$ of entries $a_{i,j} = \frac{1}{i+j}$, $\forall 1 \le i,j \le 13$, that illustrates how the Newton-type iteration can be used to increase the accuracy of the eigenvalues. We take the eigen decomposition given by the standard Julia method `eigen` from the package `LinearAlgebra` as an initial point of Newton sequences in Theorem 3 with 5 iterations. The computation is done with the precision 1024 using `ArbNumerics` package. The initial point given by `eigen` is in double precision. It is converted to the precision 1024 using `ArbNumerics` package, in order to apply Newtons iterations with this precision of 1024 bits. In Table 11 we report the eigenvalues given by `eigen` ($\sigma_{\texttt{eigen}}$) and the eigenvalues rounded to the double precision given by Newton-type sequence ($\sigma_{\texttt{newton}}$) initialized with `eigen`. We also report the relative error $\frac{|\sigma_{\texttt{newton}} - \sigma_{\texttt{eigen}}|}{\sigma_{\texttt{newton}}}$ in order to show the refinement amount realized by the Newton method. As we can see the matrix of this example is ill-conditioned (Cauchy matrices are in general ill-conditioned). There is a cluster of eigenvalues nearby zero. The accuracy enhancement obtained by applying Newton-type iterations can be clearly seen in Table 11, in particular for the first four smallest eigenvalues. For instance, the smallest eigenvalue returned by `eigen` is of order 10^{-17} close to the second smallest eigenvalues of order 10^{-16}. Newton-type method shows that the smallest eigenvalue of the order 10^{-19} yields a large relative error ~ 39.33. This also shows that all the eigenvalues are well-separated.

6.3 Sub-matrix iterations

It is possible to adapt the proposed method, taking into account the condition of the eigenvalue σ_i given by the quantity

Table 11 The relative error between $\sigma_{\texttt{eigen}}$ from the method `eigen` and $\sigma_{\texttt{newton}}$ from the Newton-type method for the Cauchy matrix $\left(\frac{1}{i+j}\right)_{1\le i,j\le 13}$

Eigenvalue	$\sigma_{\texttt{eigen}}$	$\sigma_{\texttt{newton}}$	$\frac{\lvert\sigma_{\texttt{newton}}-\sigma_{\texttt{eigen}}\rvert}{\sigma_{\texttt{newton}}}$
1	2.4030587641505818e-17	5.958203769841865e-19	39.33
2	1.8824087522342697e-16	1.7156976132548192e-16	0.09716
3	2.3152722725223998e-14	2.3178576801522747e-14	0.00111
4	1.9513972147589434e-12	1.951356013568409e-12	$2.11e$-5
5	1.1466969172503778e-10	1.1466967568738049e-10	$1.39e$-7
6	4.991788233415145e-9	4.991788235245136e-9	$3.66e$-10
7	1.6668681228080362e-7	1.666868122813953e-7	$3.54e$-12
8	4.360227301207107e-6	4.360227301206033e-6	$2.46e$-13
9	9.040674871074817e-5	9.040674871075823e-5	$1.11e$-13
10	0.0014925044272821445	0.0014925044272821172	$1.83e$-14
11	0.01955788569925287	0.01955788569925287	$4.81e$-17
12	0.19958813407010345	0.19958813407010337	$4.64e$-16
13	1.3693334145989837	1.3693334145989824	$9.98e$-16

$$\kappa(\sigma_i) = \max_{i\neq j}\left(1, \frac{1}{|\sigma_i - \sigma_j|}\right)$$

Theoretical results imply that the computation of clusters of eigenvalues is ill-conditioned. However, one can apply Theorem 3 on sub-matrices to improve the well-conditioned eigenvalues. We denote

$$\delta = \sqrt{\frac{K\|\Delta_0\|}{0.033}}$$

and p the index such that $\Sigma = \begin{pmatrix} \Sigma_p & \\ & \Sigma_{n-p} \end{pmatrix}$, $\Sigma_p = \mathrm{diag}(\sigma_1, \dots, \sigma_p)$, $\Sigma_{n-p} = \mathrm{diag}(\sigma_{p+1}, \dots, \sigma_n)$ and $|\sigma_i - \sigma_j| > \delta$ for all $1 \leqslant i \le p$ and $i < j \leqslant n$. We adapt Newton iteration to the block associated with the well-conditioned eigenvalues by defining the matrices X, Y and S as follows:

$$\begin{aligned} x_{i,i} &= 0 \\ x_{i,j} &= \begin{cases} \dfrac{-\delta_{i,j} + z_{i,j}\sigma_j}{\sigma_i - \sigma j} & \text{if} \quad |\sigma_i - \sigma_j| > \delta \\ 0 & \text{otherwise} \end{cases} \\ Y &= -Z - X \\ S &= \mathrm{diag}(-\Delta + Z\Sigma). \end{aligned}$$

Table 12 (resp. Table 13) shows the residual error err_{res} as in *Test-1* for the Cauchy matrix of size 200 (resp. the Rosser matrix of size 256 [37]) by applying the

Table 12 The residual error throughout 6 iterations with the Cauchy matrix of size 200

Iteration	$p = 12, \delta = 4.51e - 7$	$p = 5, \delta = 4.51e - 7$
1	2.45e–15	2.35e–15
2	9.63e–26	3.75e–29
3	1.56e–36	1.21e–53
4	1.54e–45	1.81e–83
5	1.15e–54	3.49e–110
6	5.08e–64	8.67e–137

Table 13 The residual error throughout 6 iterations with the Rosser matrix of size 256

Iteration	$p = 11, \delta = 1.11e$–3	$p = 5, \delta = 1.11e$–3
1	7.15e–12	1.65e–12
2	7.18e–20	7.18e–20
3	1.42e–40	1.81e–41
4	1.73e–53	1.56e–85
5	7.17e–66	1.75e–119
6	8.79e–79	8.11e–153

aforementioned sequences, the initial point is given by the Julia method `eigen`. The computation is done in precision 1024.

7 Conclusion

Taking a Newton approach towards systems of equations describing the simultaneous diagonalization problem of diagonalizable matrices, leads us to new algorithmic insights. We exhibit a Newton-type method without solving a linear system at each step as is the case of a classical Newton method. The numerical experiments corroborate the quadratic convergence predicted by the theoretical analysis.

We focused on the regular case. Some improvements and extensions can be considered, such as the treatment of clusters of eigenvalues. Another direction that can be explored, is the construction of higher-order methods.

Declarations

Conflict of Interest The authors declare no conflicting interest that is directly or indirectly related to the work submitted for publication.

References

1. Absil, P.-A., Gallivan, K.A.: Joint diagonalization on the oblique manifold for independent component analysis. In 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings, volume 5, pages V–V, (2006)
2. Absil, P.-A., Mahony, R., Sepulchre, R.: Optimization Algorithms on Matrix Manifolds. Princeton University Press, Princeton (2008)
3. Afsari, B.: Sensitivity analysis for the problem of matrix joint diagonalization. SIAM J. Matrix Anal. Appl. **30**, 1148–1171 (2008)
4. Andruchow, E., Larotonda, G., Recht, L., Varela, A.: The left invariant metric in the general linear group. J. Geom. Phys. **86**, 241–257 (2014)
5. Bouchard, F., Afsari, B., Malick, J., Congedo, M.: Approximate joint diagonalization with Riemannian optimization on the general linear group. SIAM J. Matrix Anal. Appl. **41**(1), 152–170 (2020)
6. Bro, R.: Parafac tutorial and applications. Chemom. Intell. Lab. Syst. **38**(2), 149–171 (1997)
7. Bunse-Gerstner, A., Byers, R., Mehrmann, V.: A chart of numerical methods for structured eigenvalue problems. SIAM J. Matrix Anal. Appl. **13**(2), 419–453 (1992)
8. Bunse-Gerstner, A., Byers, R., Mehrmann, V.: Numerical methods for simultaneous diagonalization. SIAM J. Matrix Anal. Appl. **14**(4), 927–949 (1993)
9. Bürgisser, P., Clausen, M., Shokrollahi, M.A.: Algebraic Complexity Theory, vol. 315. Springer Science & Business Media, Berlin (2013)
10. Cardoso, J.-F., Souloumiac, A.: Blind beamforming for non gaussian signals. IEE Proc.-F **140**, 362–370 (1993)
11. Cardoso, J.-F., Souloumiac, A.: Jacobi angles for simultaneous diagonalization. SIAM J. Matrix Anal. Appl. **17**(1), 161–164 (1996)
12. Carroll, J.D., Chang, J.-J.: Analysis of individual differences in multidimensional scaling via an n-way generalization of Eckart-Young decomposition. Psychometrika **35**(3), 283–319 (1970)
13. Cichocki, A., Mandic, D., De Lathauwer, L., Zhou, G., Zhao, Q., Caiafa, C., Phan, H.A.: Tensor decompositions for signal processing applications: from two-way to multiway component analysis. IEEE Signal Process. Mag. **32**(2), 145–163 (2015)
14. Comon, P., Jutten, C.: Handbook of Blind Source Separation: Independent Component Analysis and Applications, 1st edn. Academic Press Inc., USA (2010)
15. Cox, D.A., Little, J.B., O'Shea, D.: Using Algebraic Geometry. Number 185 in Graduate Texts in Mathematics, 2nd edn. Springer, New York (2005)
16. De Lathauwer, L.: A link between the canonical decomposition in multilinear algebra and simultaneous matrix diagonalization. SIAM J. Matrix Anal. Appl. **28**(3), 642–666 (2006)
17. Douglas, S.C.: Self-stabilized gradient algorithms for blind source separation with orthogonality constraints. IEEE Trans. Neural Netw. **11**(6), 1490–1497 (2000)
18. Elkadi, M., Mourrain, B.: Introduction à la résolution des systèmes polynomiaux. Mathématiques et Applications, vol. 59. Springer, Berlin (2007)
19. Flury, B.D., Neuenschwander, B.E.: Simultaneous diagonalization algorithms with applications in multivariate statistics. In: Approximation and Computation: A Festschrift in Honor of Walter Gautschi, pp. 179–205. Springer, Birkhäuser, Boston, MA (1994)
20. Flury, B.N., Gautschi, W.: An algorithm for simultaneous orthogonal transformation of several positive definite symmetric matrices to nearly diagonal form. SIAM J. Sci. Stat. Comput. **7**(1), 169–184 (1986)
21. Françoise, C.: Simultaneous newton's iteration for the eigenproblem. Computing **5**, 67–74 (1984)
22. Hoeven, J.V.D., Mourrain, B.: Efficient certification of numeric solutions to eigenproblems. In: Blömer, J., Kotsireas, I., Kutsia, T., Simos, D. (eds.) Mathematical Aspects of Computer and Information Sciences. MACIS 2017. Lecture Notes in Computer Science, vol. 10693. Springer, Cham (2017)
23. Horn, R.A., Charles, C.R.: Topics in Matrix Analysis. Cambridge University Press, Cambridge (1991)
24. Horn, R.A., Johnson, C.R.: Matrix Analysis. Cambridge University Press, Cambridge (2012)
25. Jiang, R., Li, D.: Simultaneous diagonalization of matrices and its applications in quadratically constrained quadratic programming. SIAM J. Optim. **26**(3), 1649–1668 (2016)

26. Joho, M., Rahbar, K.: Joint diagonalization of correlation matrices by using Newton methods with application to blind signal separation. In: Sensor Array and Multichannel Signal Processing Workshop Proceedings, 2002, 403–407 (2002)
27. Joho, M.: Newton method for joint approximate diagonalization of positive definite Hermitian matrices. SIAM J. Matrix Anal. Appl. **30**(3), 1205–1218 (2008)
28. Joho, M., Rahbar, K.: Joint diagonalization of correlation matrices by using newton methods with application to blind signal separation. In: Sensor Array and Multichannel Signal Processing Workshop Proceedings, 2002, pp. 403–407. IEEE, (2002)
29. Jolliffe, I.: Principal Component Analysis, pp. 1094–1096. Springer Berlin Heidelberg, Berlin (2011)
30. Laub, A., Heath, M., Paige, C., Ward, R.: Computation of system balancing transformations and other applications of simultaneous diagonalization algorithms. IEEE Trans. Autom. Control **32**(2), 115–122 (1987)
31. Luciani, X., Albera, L.: Canonical polyadic decomposition based on joint eigenvalue decomposition. Chemom. Intell. Lab. Syst. **132**, 152–167 (2014)
32. Mahony, R.E.: The constrained newton method on a lie group and the symmetric eigenvalue problem. Linear Algebra Appl. **248**, 67–89 (1996)
33. Mesloub, A., Belouchrani, A., Abed-Meraim, K.: Efficient and stable joint eigenvalue decomposition based on generalized givens rotations. In: 2018 26th European Signal Processing Conference (EUSIPCO), pp. 1247–1251. IEEE, (2018)
34. Nikpour, M., Manton, J., Hori, G.: Algorithms on the Stiefel manifold for joint diagonalisation. In: 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing, **2**, II–1481–II–1484 (2002)
35. Nishimori, Y., Akaho, S.: Learning algorithms utilizing quasi-geodesic flows on the Stiefel manifold. Neurocomput. **67**, 106–135 (2005)
36. Rahbar, K., Reilly, J.P.: Geometric optimization methods for blind source separation of signals. In: in Proc. ICA, pp. 375–380, (2000)
37. Rosser, B., Lanczos, C., Hestenes, M.R., Karush, W.: Separation of close eigen-values of a real symmetric matrix. J. Res. Natl. Bur. Stand. **47**, 291–297 (1950)
38. Sato, H.: Riemannian newton-type methods for joint diagonalization on the stiefel manifold with application to independent component analysis. Optimization **66**(12), 2211–2231 (2017)
39. Sidiropoulos, N.D., Bro, R.: On the uniqueness of multilinear decomposition of n-way arrays. J. Chemom. **14**(3), 229–239 (2000)
40. Sidiropoulos, N.D., Lathauwer, L.D., Xiao, F., Huang, K., Papalexakis, E.E., Faloutsos, C.: Tensor decomposition for signal processing and machine learning. IEEE Trans. Signal Process. **65**(13), 3551–3582 (2017)
41. Sørensen, M., De Lathauwer, L.: Multidimensional harmonic retrieval via coupled canonical polyadic decomposition-part i: model and identifiability. IEEE Trans. Signal Process. **65**(2), 517–527 (2016)
42. Sørensen, M., De Lathauwer, L.: Multidimensional harmonic retrieval via coupled canonical polyadic decomposition-part ii: algorithm and multirate sampling. IEEE Trans. Signal Process. **65**(2), 528–539 (2016)
43. Sørensen, M., Domanov, I., De Lathauwer, L.: Coupled canonical polyadic decompositions and multiple shift invariance in array processing. IEEE Trans. Signal Process. **66**(14), 3665–3680 (2018)
44. Sørensen, M., Van Eeghem, F., De Lathauwer, L.: Blind multichannel deconvolution and convolutive extensions of canonical polyadic and block term decompositions. IEEE Trans. Signal Process. **65**(15), 4132–4145 (2017)
45. van der Hoeven, J., Yakoubsohn, J.-C..: Certified singular value decomposition. Technical Report HAL 01941987, (2018)
46. Vollgraf, R., Obermayer, K.: Quadratic optimization for simultaneous matrix diagonalization. IEEE Trans. Signal Process. **54**(9), 3270–3278 (2006)
47. Wang, W., Sanei, S., Chambers, J.: Penalty function-based joint diagonalization approach for convolutive blind separation of nonstationary sources. In: IEEE Transactions on Signal Processing, vol.53, pp. 1654–1669 (2005)
48. Weyl, H.: Das asymptotische verteilungsgesetz der eigenwerte linearer partieller differentialgleichungen (mit einer anwendung auf die theorie der hohlraumstrahlung). Math. Ann. **71**, 441–479 (1912)

49. Yeredor, A.: Non-orthogonal joint diagonalization in the least-squares sense with application in blind source separation. IEEE Trans. Signal Process. **50**(7), 1545–1553 (2002)