Newton-Type Methods For Simultaneous Matrix Diagonalization

Rima Khouja, Bernard Mourrain, Jean-Claude Yakoubsohn

Newton-Type Methods For Simultaneous Matrix Diagonalization

Rima Khouja[1*†], Bernard Mourrain[1†] and Jean-Claude Yakoubsohn[2†]

[1*]AROMATH, INRIA Sophia Antipolis Méditérannée, 2004, route des Lucioles, Sophia Antipolis, 06902, France.
[2]Institut de Mathématiques de Toulouse, Université Paul Sabatier, 118, route de Narbonne, Toulouse, 31062, France.

*Corresponding author(s). E-mail(s): rima.khouja@inria.fr;
Contributing authors: bernard.mourrain@inria.fr;
yak@mip.ups-tlse.fr;
†These authors contributed equally to this work.

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 which converges quadratically towards the solution. This construction is not based on the resolution of linear system as it is the case in the classical Newton method. Moreover, we provide a theoretical analysis of this construction to exhibit a condition to get a quadratic convergence. We also propose numerical experiments, which illustrate the theoretical results. This shows that classical QR method would gain in efficiency incorporating the tests given by the theory.

Keywords: Simultaneous diagonalization, Newton-type method, eigenproblem, eigenvalues, high precision computation

MSC Classification: 65F15 , 65H10 , 15A18 , 65-04

1 Introduction

1.1 Our study

Let us consider p *diagonalizable* matrices $M_1, \cdots, M_p$ 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. [25]. The aim of this paper is to numerically compute 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 $\Sigma = (\Sigma_1, \ldots, \Sigma_p)$ and $EMF - \Sigma := (EM_1F - \Sigma_1, \ldots, EM_pF - \Sigma_p)$. 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 $2\,n^2 - n^2 - (n^2 - n) = n$. 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:

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

where $\Sigma' = (\Sigma_0, \Sigma_1, \ldots, \Sigma_p)$, $M_0 \in \mathbb{C}^{n\times n}$ is replacing I_n and Σ_0 is a diagonal matrix replacing I_n in the first equation. When the pencil $M' = (M_0, M_1, \ldots, M_p)$ 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 $M' = (M_0, M_1, \ldots, M_p)$ gives the solution $(FM_0, E\Sigma_0^{-1}, \Sigma\Sigma_0^{-1})$ of (1) for $M = (M_0^{-1}M_1, \ldots, M_0^{-1}M_p)$. 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 analyse the Newton map associated to an iteration. These ideas also are been developed in a technical report for the fast computation of the singular value decomposition [23].

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

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

The technical background to solve this linear system is the Kronecker product, see [24]. In this way the size of the linear system that one needs to invert is n^2.

On the other hand if we consider a Newton map defined by $(E(I_n+X), (I_n+Y)F, \Sigma+S)$ from (E, F, Σ) such that X, Y and S cancel the linear part of the Taylor expansion of $f(E(I_n+X), (I_n+Y)F, \Sigma+S)$, 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 $p = 1, 2$. This is due to the fact that Σ is a diagonal matrix. From these considerations we define and analyse a sequence which converges quadratically towards a solution of the system (1) without inverting a linear system at each step of this Newton-like method. We say that we have a quadratic sequence associated to a system of equation, if the sequence converges quadratically towards a solution.

1.2 Related works

Simultaneous matrix diagonalization is required by many algorithms as it was pointed out in [7]. 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-variable 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 quadratic convergence do not depend on n. Many other papers studies the so-called Jacobi-like methods (see e.g. [28], [29] and references therein).

In [22] an iteration with a proof of 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 $C = (C_1, \ldots, C_s)$, several algorithms based on Riemannian optimization methods (see [2]) have been developped in order to find an *approximate joint diagonalizer* (see e.g. [5, 1, 32, 26]).

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 $(BC_1B^T, \ldots, BC_sB^T)$ over a Riemannian manifold (see [39, 5, 3] 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, 21, 32, 18, 26, 30, 31]). Due to inaccuracies in the computation of the diagonalizer with orthogonality constraints (see. [41]), *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 Sections 3 and 4 the points are updated by using direct and explicit formulas. They have a lower complexity than the Riemannian optimization based algorithms and they are well-adapted to computation with high precision.

Simultaneous diagonalisation of matrix pencils appears in many applications. 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], [19]. In the case of simple roots, this reduces to simultaneous diagonalisation of a matrix pencil.

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 [33]. 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], [34], Sensor array processing [37], Arithmetic Complexity [9], wireless communications [38], multidimensional harmonic retrieval [35], [36], Chemometrics [6], and Principal components analysis [27].

1.3 Outline

The sections 2, 3, 4, 5 are devoted to give conditions to get a quadratic sequence respectively to numerically approximate a solution of the systems

- $FE - I_n = 0$,
- 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. In Section 6 we perform a 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:

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

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

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

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

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

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

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

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

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, $\varphi(\varepsilon, u) \leqslant \varphi(\frac{1}{2}, 1) = 2$. □

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

$$f(E + EX, F + YF) = (F + YF)(E + EX) - I_n \tag{6}$$
$$= Z + (Z + I_n)X + Y(Z + I_n) + Y(Z + I_n)X. \tag{7}$$

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 Equation (8). Moreover we get, by substituting in Equation (7) X and Y by $-\frac{Z}{2}$,

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

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 $i \geqslant 0$*

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

converge quadratially 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 $i \geqslant 0$,

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

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

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

$$\leqslant 2^{-2^{k+1}+1}\varepsilon.$$

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

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

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

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

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

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

We deduce that

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

These properties also holds 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 $\Delta = FME - \Sigma$:

$$\begin{aligned} &(F+YF)(E+EX) - I_n \\ &= Z + (Z+I_n)X + Y(Z+I_n) + Y(Z+I_n)X \\ &(F+YF)M(E+EX) - \Sigma - S \\ &= FME - \Sigma - S + FMEX + YFME + YFMEX \end{aligned} \tag{11}$$

$$= \Delta - S + \Sigma X + Y\Sigma + \Delta X + Y\Delta + Y(\Delta + \Sigma)X \tag{12}$$

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 $\Sigma = \mathrm{diag}(\sigma_1, \cdots, \sigma_n)$, $Z = (z_{i,j})_{1 \leq i,j \leq n}$ and $\Delta = (\delta_{i,j})_{1 \leq i,j \leq n}$ 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

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

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

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

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

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

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

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

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

Since $\mathrm{diag}(\Delta - S - Z\Sigma) = \mathrm{diag}(\Sigma X - X\Sigma) = 0$ 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 and defined the sequences for $i \geqslant 0$,*

$$Z_i = F_i E_i - I_n$$

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

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

$$u := \kappa_0^2 (K_0 + 1)^3 \varepsilon_0 \leqslant 0.136. \tag{21}$$

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

$$\begin{aligned}
\|E_i - E_\infty\| &\leqslant 0.61 \times 2^{1-2^{i+1}} \|E_0\| u \\
\|F_i - F_\infty\| &\leqslant 0.61 \times 2^{1-2^{i+1}} \|F_0\| u.
\end{aligned}$$

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

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

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

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

$$\|\Sigma_i - \Sigma_0\| \leqslant (2 - 2^{2-2^i}) \varepsilon \tag{23}$$

$$\kappa_i \leqslant \frac{\kappa}{1 - 4\kappa\varepsilon} \tag{24}$$

$$K_i \leqslant K + 2\varepsilon \tag{25}$$

These inequalities clearly hold for $i = 0$. Assuming that the induction hypothesis holds for a given i and let us prove it for $i + 1$. First we have

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

Hence

$$\begin{aligned}
\|Z_{i+1}\| &\leqslant (2 + \kappa_i (K_i + 1)(1 + \varepsilon_i)) \kappa_i (K_i + 1) \varepsilon_i^2 \\
&\leqslant 3\kappa_i^2 (K_i + 1)^3 \varepsilon_i^2.
\end{aligned}$$

On the another hand

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

$$\begin{aligned}
\|\Delta_{i+1}\| &\leqslant (2 + \kappa_i (K_i + 1)(K_i + \varepsilon_i)) \kappa_i (K_i + 1) \varepsilon_i^2 \\
&\leqslant 3\kappa_i^2 (K_i + 1)^3 \varepsilon_i^2.
\end{aligned}$$

It follows

$$\begin{aligned}
\varepsilon_{i+1} &\leqslant \frac{3\kappa_0^2(K+1+2\varepsilon)^3}{(1-4\kappa\varepsilon)^2}\varepsilon_i^2 \\
&\leqslant \frac{3\left(1+\frac{u}{8}\right)^3}{\left(1-\frac{u}{2}\right)^2}\kappa^2(K+1)^3\varepsilon_i^2 \quad \text{since} \quad \varepsilon \leqslant \frac{u}{8} \\
&\leqslant \frac{3\left(1+\frac{u}{8}\right)^3}{\left(1-\frac{u}{2}\right)^2}\kappa^2(K+1)^3\varepsilon 2^{2-2^{i+1}}\varepsilon \\
&\leqslant 2^{1-2^{i+1}}\varepsilon \quad \text{since} \quad \frac{3\left(1+\frac{u}{8}\right)^3}{\left(1-\frac{u}{2}\right)^2}\kappa^2(K+1)^3\varepsilon \leqslant 2^{-1} \text{ for } u \leqslant 0.136.
\end{aligned}$$

Next we prove (23) for $i+1$. We have :

$$\begin{aligned}
\|\Sigma_{i+1}-\Sigma_0\| &\leqslant \|S_i\| + \|\Sigma_i - \Sigma_0\| \\
&\leqslant 2^{1-2^i}\varepsilon + (2-2^{2-2^i})\varepsilon = (2-2^{1-2^i})\varepsilon \\
&\leqslant (2-2^{2-2^{i+1}})\varepsilon.
\end{aligned}$$

We then deduce (25) for $i+1$:

$$K_{i+1} := \|\Sigma_{i+1}\| \leqslant \|\Sigma_0\| + (2-2^{2-2^{i+1}})\varepsilon \leqslant K + 2\varepsilon.$$

Let us finally prove (24) for $i+1$. The $\sigma_{i+1,j}$'s are the diagonal values of Σ_{i+1}. The bound [40] implies that

$$|\sigma_{i+1,j} - \sigma_{0,j}| \leqslant \|\Sigma_{i+1} - \Sigma_0\| \leqslant 2\varepsilon \qquad \text{for } 1 \leqslant j \leqslant n.$$

So that for $1 \leqslant j < k \leqslant n$, we obtain using $\kappa\varepsilon \leqslant \frac{u}{8}$:

$$\begin{aligned}
|\sigma_{i+1,k} - \sigma_{i+1,j}| &\geqslant |\sigma_{0,k} - \sigma_{0,j}| - |\sigma_{i+1,k} - \sigma_{0,k}| - |\sigma_{i+1,j} - \sigma_{0,j}| \\
&\geqslant |\sigma_{0,k} - \sigma_{0,j}|(1 - \kappa|\sigma_{i+1,k} - \sigma_{0,k}| - \kappa|\sigma_{i+1,j} - \sigma_{0,j}|) \\
&\geqslant |\sigma_{0,j} - \sigma_{0,k}|(1 - 4\kappa\varepsilon) \\
&\geqslant |\sigma_{0,j} - \sigma_{0,k}|\left(1 - \frac{u}{2}\right) \geqslant 0.
\end{aligned}$$

Finally, we get :

$$\kappa_{i+1} = \leqslant \frac{\kappa}{1-4\kappa\varepsilon}.$$

This completes the proof of the four induction hypotheses (22–25) at order $i+1$.

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

$$\begin{aligned}
\|X_k\| &\leqslant \kappa_k(K_k+1)\varepsilon_k \\
&\leqslant \frac{1+\frac{u}{8}}{1-\frac{u}{2}}\kappa(K+1)\varepsilon 2^{1-2^k} \\
&\leqslant \frac{\left(1+\frac{u}{8}\right)u}{4\left(1-\frac{u}{2}\right)}2^{1-2^k} \\
&\leqslant 0.28 \times 2^{1-2^k}u \quad \text{since } u \leqslant 0.136.
\end{aligned}$$

Consequently,

$$\|W_\infty - I_n\| \leqslant \prod_{i\geqslant 0}(1 + 0.28u2^{1-2^i}) - 1$$

$$\leqslant 0.56u \quad \text{from Lemma 1}$$
$$\leqslant 0.56 \times 0.136 \leqslant 0.0762.$$

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

$$\begin{aligned}\|W_i - W_\infty\| &\leqslant \|W_i\| \left\| 1 - \prod_{k \geqslant i+1} (1 + \|X_k\|) \right\| \\ &\leqslant (1 + \|W_i - I_n\|) \left\| \prod_{k \geqslant 0} (1 + 0.28 \times 2^{1-2^{k+i+1}}) - 1 \right\| \\ &\leqslant (1 + 0.0762) \times 0.56 \times 2^{1-2^{i+1}} u \quad \text{from Lemma 1} \\ &\leqslant 0.61 \times 2^{1-2^{i+1}} u.\end{aligned}$$

We deduce that

$$\|E_i - E_\infty\| \leqslant 0.61 \times 2^{1-2^{i+1}} \|E_0\| u.$$

In the same way we show that F_0 is invertible and

$$\|F_i - F_\infty\| \leqslant 0.61 \times 2^{1-2^{i+1}} \|F_0\| u.$$

The theorem is proved. □

Proposition 4 *The complexity of the Newton iteration in Theorem 3 is in* $\mathcal{O}(n^\omega)$ *where* ω *is the linear algebra exponent of matrix multiplications.*

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 matrix multiplications and diagonal matrix operations, has a complexity in $\mathcal{O}(n^\omega)$. Consequently, the complexity of each iteration is in $\mathcal{O}(n^\omega)$. □

Remark 1 It is possible to generalize this approach in 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 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:

$$\begin{aligned}&(F + \mathrm{YF})M_k(E + \mathrm{EX}) - (\Sigma_k + S_k)\\ &= Z_k - S_k + \Sigma_k X + Y\Sigma_k + Z_k X + \mathrm{YZ}_k + Y(Z_k + \Sigma_k)X \end{aligned} \tag{26}$$

By assuming Z_1, Z_2 are of small norm, the linear system to solve from Equation (26) is the following

$$Z_k - S_k + \Sigma_k X + Y\Sigma_k \;=\; 0, \qquad k = 1, 2 \tag{27}$$

A solution of (27) is given in the following lemma.

Lemma 3 Let $\Sigma_k = \mathrm{diag}(\sigma_1^k, \cdots, \sigma_n^k)$, $Z_k = (z_{i,j}^k)_{1 \le i,j \le n}$ 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 matrices defined by

$$x_{i,i} \;=\; 0 \tag{28}$$

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

$$y_{i,i} \;=\; 0 \tag{30}$$

$$y_{i,j} \;=\; -\frac{\begin{vmatrix} \sigma_i^1 & z_{i,j}^1 \\ \sigma_i^2 & z_{i,j}^2 \end{vmatrix}}{\begin{vmatrix} \sigma_i^1 & \sigma_j^1 \\ \sigma_i^2 & \sigma_j^2 \end{vmatrix}}, \qquad i \neq j \tag{31}$$

$$S_k \;=\; \mathrm{diag}(Z_k), \qquad k = 1, 2. \tag{32}$$

Then we have

$$Z_k - S_k + \Sigma_k X + Y\Sigma_k \;=\; 0, \qquad k = 1, 2 \tag{33}$$

Moreover

$$\|X\|, \|Y\| \;\leqslant\; 2\kappa\varepsilon K \tag{34}$$

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 equation (33) 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 (29), (31). 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 have the equation (33) satisfied. The bounds (34) easily follows from (29), (31). □

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

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

where X_i, Y_i are defined by the formulas (28–31). Let $\varepsilon_0 = \max(\|Z_{0,1}\|, \|Z_{0,2}\|)$, $\kappa_0 = \max\left(1, \max_{i\neq j} \dfrac{1}{\begin{vmatrix} \sigma_{0,i}^1 & \sigma_{0,j}^1 \\ \sigma_{0,i}^2 & \sigma_{0,j}^2 \end{vmatrix}}\right)$ and $K_0 = \max(1, \max_{j,k}|\sigma_{0,j}^k|)$. Assume that

$$u := 4\varepsilon_0 \kappa_0^2 K_0^3 \leqslant 0.094. \tag{35}$$

Then the sequences $(\Sigma_{i,k}, E_i, F_i)_{i\geqslant 0}$ converge quadratially 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 $i \geqslant 0$,

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

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

$$\varepsilon_i \leqslant 2^{1-2^i} \varepsilon \tag{36}$$

$$\|\Sigma_{i,k} - \Sigma_{0,k}\| \leqslant (2 - 2^{2-2^i})\varepsilon \tag{37}$$

$$\kappa_i \leqslant \frac{\kappa}{1 - 8\kappa\varepsilon(K+\varepsilon)} \tag{38}$$

$$K_i \leqslant K + 2\varepsilon \tag{39}$$

These inequalities clearly hold for $i = 0$. Assuming that the induction hypothesis holds for a given i and let us prove it for $i+1$. First, we have

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

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

It follows

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

The proof of (37) is the same as (23) in the proof of Theorem 3, and for (39), $K_i = \max(\|\Sigma_{i,1}\|, \|\Sigma_{i,2}\|) \leqslant K + 2\varepsilon$, as in (25), thus we have:
$|\sigma^k_{i+1,j} - \sigma^k_{0,j}| \leqslant \|\Sigma_{i+1,k} - \Sigma_{0,k}\| \leqslant 2\varepsilon \qquad 1 \leqslant j \leqslant n, k = 1, 2$.

Let us finally prove (38) for $i+1$. First we have:

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

Now,

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

Finally, we get :

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

This completes the proof of the four induction hypotheses (36–39) at order $i+1$.

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,

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

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 for a family of simultaneously diagonalizable matrices.

In this section we present two strategies to solve the system (1) of a family of commuting matrices $(M_i)_{1\leq i\leq p}$ in $\mathcal{W}_n$. The first strategy is trivial and consists of finding the common diagonalizers E and F of the family 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 family can diagonalize the other matrices of the family. We note that, in general, we don't have this property for simultaneously diagonalizable matrices, where, for instance, it is posssible to find a diagonalizer of M_1 which is not a common diagonalizer for the other matrices of the family. 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 are linearly independent i.e. that $\sum_{i=1}^{p} a_i M_i = 0 \Rightarrow a_i = 0, i = 1:p$. Let $E \in GL_n$ and $\Sigma_i \in \mathcal{D}_n'$ be such that

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

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

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

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

$$\begin{pmatrix} EF - I_n \\ FME - \Sigma \end{pmatrix} = 0 \tag{40}$$

where $M = \sum_{i=1}^{p} \alpha_i M_i$. Let $\varepsilon = \|F_0 M E_0 - \Sigma\|$. If

$$n^2\varepsilon \leqslant 0.272$$

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

$$\kappa^2(K_0+1)^3\varepsilon \leqslant \frac{n^2}{2}\varepsilon.$$

The condition $\frac{n^2}{2}\varepsilon \leqslant 0.136$ 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.

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, 5 and 6). 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 3, 4, 7 and 8). 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: *Test1* for Theorem 3 and *Test2* for Theorem 5, considered in this section.

Test1. Let $\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$, $M = E\Sigma E^{-1} + 10^{-\mathrm{e}}A$, where $\mathrm{e} \in \{3, 6\}$. The matrices E, Σ, and $A \in \mathbb{K}^{n\times n}$ are chosen randomly following a standard normal distribution such that E is invertible, Σ is diagonal with n different diagonal entries and A is any random square matrix of size n and Frobenius norm equal to 1. Since M is a small perturbation of $E\Sigma E^{-1}$, more precisely $\|M - E\Sigma E^{-1}\|_{Frob} = 10^{-\mathrm{e}}$, 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:

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

Test2. 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 into a standard normal distribution, 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:
$E_0 = E + 10^{-\mathrm{e}}A$, $F_0 = F + 10^{-\mathrm{e}}B$, $\Sigma_{0,1} = \Sigma_1 + 10^{-\mathrm{e}}C$, $\Sigma_{0,2} = \Sigma_2 + 10^{-\mathrm{e}}D$, where $\mathrm{e} \in \{3, 6\}$, 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 distribution. The residual error reported in this test at iteration k is given by:

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

We notice that the condition established in Theorem 3 (resp. Theorem 5) is reached in *Test1* (resp. *Test2*) for matrices of size 10 with order of perturbation equal to 10^{-6}, and we can see in Tables 1, 2, 5 and 6 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 3, 4, 7 and 8).

Table 1 The computational results throughout 7 iterations of an exemple of implementation of *Test1* with $\mathbb{K} = \mathbb{R}$, $n = 10$ and $\mathrm{e} = 6$.

Iteration	$\kappa^2(K+1)^3\varepsilon \leq 0.136$	err_{res}
1	0.07915	$5.51e-6$
2	$2.52e-6$	$1.76e-10$
3	$9.29e-16$	$6.47e-20$
4	$1.11e-34$	$7.78e-39$
5	$1.83e-72$	$1.28e-76$
6	$4.31e-148$	$3.01e-152$
7	$1.16e-287$	$8.08e-292$

Table 2 The computational results throughout 7 iterations of an exemple of implementation of *Test1* with $\mathbb{K} = \mathbb{C}$, $n = 10$ and $\mathrm{e} = 6$.

Iteration	$\kappa^2(K+1)^3\varepsilon \leq 0.136$	err_{res}
1	0.00735	$1.14e-5$
2	$2.14e-8$	$3.35e-11$
3	$5.11e-19$	$7.99e-22$
4	$6.88e-40$	$1.07e-42$
5	$7.31e-82$	$1.14e-84$
6	$9.70e-166$	$1.51e-168$
7	$4.28e-284$	$6.69e-287$

Table 3 The residual error throughout 7 iterations given by the implementation of *Test1* with $\mathbb{K} = \mathbb{R}, \mathrm{e} = 3$ and $n = 10, 50, 100$.

Iteration	$n = 10$	$n = 50$	$n = 100$
1	0.00857	0.07931	0.03226
2	0.00019	0.05761	0.01380
3	$1.58e-8$	0.00619	0.00061
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 4 The residual error throughout 7 iterations given by the implementation of *Test1* with $\mathbb{K} = \mathbb{C}, \mathrm{e} = 3$ and $n = 10, 50, 100$.

Iteration	$n = 10$	$n = 50$	$n = 100$
1	0.00884	0.00975	0.01600
2	$8.59e-6$	$6.39e-5$	0.00010
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	$2.64e-169$	$8.28e-134$	$1.05e-132$

6.2 Wilkinson polynomial

For $n \in \mathbb{N}^*$, the polynomial given by:

$$P(x) = \prod_{i=1}^{n} (x - i) \tag{41}$$

Table 5 The computational results throughout 7 iterations of an exemple of implementation of *Test2* with $\mathbb{K} = \mathbb{R}$, $n = 10$ and $\mathrm{e} = 6$.

Iteration	$4\kappa^2 K^3 \varepsilon \leq 0.094$	err_{res}
1	0.07650	$6.72e-6$
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 6 The computational results throughout 7 iterations of an exemple of implementation of *Test2* with $\mathbb{K} = \mathbb{C}$, $n = 10$ and $\mathrm{e} = 6$.

Iteration	$4\kappa^2 K^3 \varepsilon \leq 0.094$	err_{res}
1	0.00686	$9.16e-6$
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 7 The residual error throughout 7 iterations given by the implementation of *Test2* with $\mathbb{K} = \mathbb{R}, \mathrm{e} = 3$ and $n = 10, 50, 100$.

Iteration	$n = 10$	$n = 50$	$n = 100$
1	0.02901	0.00457	0.01004
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 *Test2* with $\mathbb{K} = \mathbb{C}, \mathrm{e} = 3$ and $n = 10, 50, 100$.

Iteration	$n = 10$	$n = 50$	$n = 100$
1	0.00733	0.00314	0.00552
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	$1.31e-194$	$1.91e-215$	$4.53e-216$

is the so-called *n-th* Wilkinson polynomial. It is a monic polynomial of degree n with n simple roots from 1 to n. Let $P(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_0$. It is known that the roots of $P(x)$ are the eigenvalues of its companion matrix $C(P)$. It is possible to compute the roots of the Wilkinson polynomial in high precision. The process is to compute by the standard Julia's solver the eigenvalues and the eigenvectors of $C(P)$, then use this as an initial point of the Newton sequences in Section 3 to increase the precision. However, we noticed

that this strategy works only until $n = 19$. For $n \geq 20$ some numerical inaccuracy issues appears in the computation of the initial point. More concretely, if we take for instance $n = 20$, the n eigenvalues given by the standard Julia's solver are as follows:

```
0.9999999999981168 + 0.0im
2.0000000001891918 + 0.0im
2.9999999926196894 + 0.0im
4.000000196012741 + 0.0im
4.999996302203527 + 0.0im
6.000048439601834 + 0.0im
6.999557630040994 + 0.0im
8.002891069857936 + 0.0im
8.986693042189247 + 0.0im
10.049974037139467 + 0.0im
10.886016935269065 + 0.0im
12.358657519230299 + 0.0im
12.561193394139806 + 0.0im
14.51895930872283 - 0.2133045589544431im
14.51895930872283 + 0.2133045589544431im
16.206794587063147 + 0.0im
16.885716688231323 + 0.0im
18.030097274474777 + 0.0im
18.993902180590464 + 0.0im
20.000542093702702 + 0.0im.
```

Since the problem comes from the matrix for which we compute the eigenvalues with double precision, one can think whether we can replace the companion matrix by another matrix which has the same characteristic polynomial (in this case the Wilkinson polynomial). In fact, as discussed by M.Fiedler in [20], we can construct a symmetric matrix whose characteristic polynomial is $P(x)$. We retrieve this construction from [20]: Let $b_1, \ldots, b_{n-1}$ be distinct numbers such that $P(b_i) \neq 0$. Let $Q(x) = \prod_{i=1}^{n-1} (x - b_i)$, and let

$$\begin{aligned} c_i &= -\sqrt{\frac{P(b_i)}{Q'(b_i)}} \\ c^t &= (c_1, \ldots, c_{n-1}) \\ B &= \operatorname{diag}(b_1, \ldots, b_{n-1}) \\ d &= -a_{n-1} - \sum_{i=1}^{n-1} b_i, \end{aligned}$$

then the characteristic polynomial of $A = \begin{pmatrix} B & c \\ c^t & d \end{pmatrix}$ is equal to $P(x)$. Since $P(x)$ is of real coefficients and its roots are simple and real, we can choose the b_i's such that they interlace the roots i.e. $1 < b_1 < 2 < b_2 < \ldots < 19 < b_{n-1} < 20$, so that, as shown in [20], the symmetric matrix is of real coefficients. For instance, we take in our construction $b_i = i + 0.5$, $\forall i \in \{1, \ldots, n-1\}$. Now, by computing the matrix A in high precision (1024 bits) and applying the standard Julia's solver to compute the eigenvalues of A rounded to Float64, we found:

```
1.0000000000000036
```

```
2.000000000000007
2.9999999999999964
4.0
5.0
6.0
7.000000000000011
7.999999999999998
8.999999999999998
9.999999999999998
11.0
12.0
12.999999999999998
13.999999999999996
15.0
16.0
17.000000000000007
18.0
19.000000000000004
19.999999999999996
```

We take these eigenvalues with their eigenvectors as an initial point of the Newton sequences in Section 3. We consider a precision equal to 1024 bits. The residual error is as in the previous subsection. The initial residual error with this initial point is equal to $8.49e-14$. We report the residual error throughout 4 iterations:
iter1: $2.04e-27$
iter2: $3.21e-55$
iter3: $1.16e-110$
iter4: $1.28e-221$
Finally, we find that the 20 eigenvalues computed by the Newton iterations give the 20 roots of the Wilkinson polynomial in high precision. We notice that the process was very fast (taking about 0.3 seconds). This example highlights the importance of the high precision computation in the accuracy of the polynomial's roots.

6.3 QR algorithm with Newton condition

The aim of this experiment is to illustrate the introduction of the condition given by Theorem 3 in an iterative method to compute eigenvalues such as QR method. The practical implementations of eigen solver in linear algebra libraries use many ingredients. For reasons of simplicity we will only consider here the classical basic QR algorithm to compute the eigenvalues (and eigenvectors if the matrix is symmetric) [17]. The QR algorithm consists of generating a sequence $(A_k)_k$ such that $A_0 = A$, at the *k-th* step the QR decomposition of $A_k = Q_k R_k$, where Q_k is an orthogonal matrix and R_k is an upper triangular matrix, is computed; and $A_{k+1} = R_k Q_k$. These sequences converge, under some conditions, to the Schur form of A, such that the diagonal entries of its triangular matrix are the eigenvalues of A. If A is symmetric then the columns of $Q = \prod_k Q_k$ give the eigenvectors of A. The QR decomposition at each step can be computed by using Householder transformations. The classical QR algorithm in its crude form is given in pseudo-code in Algorithm 1.

We can use Algorithm 1 to construct an initial point to the Newton sequence in Section 3. Indeed, since it is sufficient that the initial point verify the condition

Algorithm 1 QR algorithm

Input: $A \in \mathbb{C}^{n\times n}$.
Compute the QR decomposition of A: $A = QR$;
Set $k = 0$, $A_0 = A$, $Q_0 = Q$, $R_0 = R$;
Set $A_1 = R_0 Q_0$;
Set $\text{err}_1 = \|A_1\|_{\text{L,Tri}}$;
while $\text{err}_k = \|A_k\|_{\text{L,Tri}} > threshold$ **do**
    $A_k = Q_k R_k$;
    $A_{k+1} = R_k Q_k$.
end while
Output: $\text{diag}(A_{k^*})$, $\prod_{0\leq k\leq k^*} Q_k$.

established in Theorem 3 to make the Newton sequences converge to the eigenvalue decomposition, we introduce this condition into the QR algorithm (see Algorithms 2 and 3). So that this algorithm stops once the Newton condition is verified, to give the hand to the fast Newton sequence to start iterating. This step reduces noticeably the number of iterations of the QR algorithm (see fig. 1).

Algorithm 2 Test for Newton (`Test_for_Newton`)

Input: $M, \Sigma = \text{diag}(\sigma_1, \ldots, \sigma_n), E, F \in \mathbb{C}^{n\times n}$.
$Z = FE - I_n$, $\Delta = FME - \Sigma$;
$\varepsilon = \max(\|Z\|, \|\Delta\|)$;
$\kappa = \max\left(1, \max_{i\neq j} \dfrac{1}{|\sigma_i - \sigma_j|}\right)$;
$K = \max(1, \max_i |\sigma_i|)$;
Output: $\kappa^2 (K+1)^3 \varepsilon$.

Algorithm 3 QR algorithm with Newton test

Input: $A \in \mathbb{C}^{n\times n}$.
Compute the QR decomposition of A: $A = QR$;
Set $k = 0$, $A_0 = A$, $Q_0 = Q$, $R_0 = R$;
Set $A_1 = R_0 Q_0$, $\Sigma_k = \text{diag}(A_k)$, $E_k = \prod_{0\leq i\leq k-1} Q_i$, $F_k = E_k^*$;
Set $\text{err}_1 = \texttt{Test\_for\_Newton}(A, \Sigma_1, E_1, F_1)$;
while $\text{err}_k = \texttt{Test\_for\_Newton}(A, \Sigma_k, E_k, F_k) > 0.136$ **do**
    $A_k = Q_k R_k$;
    $A_{k+1} = R_k Q_k$.
end while
Output: Σ, E and F.

Going back to the symmetric matrix A of size 20 and characteristic polynomial equal to the Wilkinson polynomial ($n = 20$) in Section 6.2. We apply Algorithm 3 on

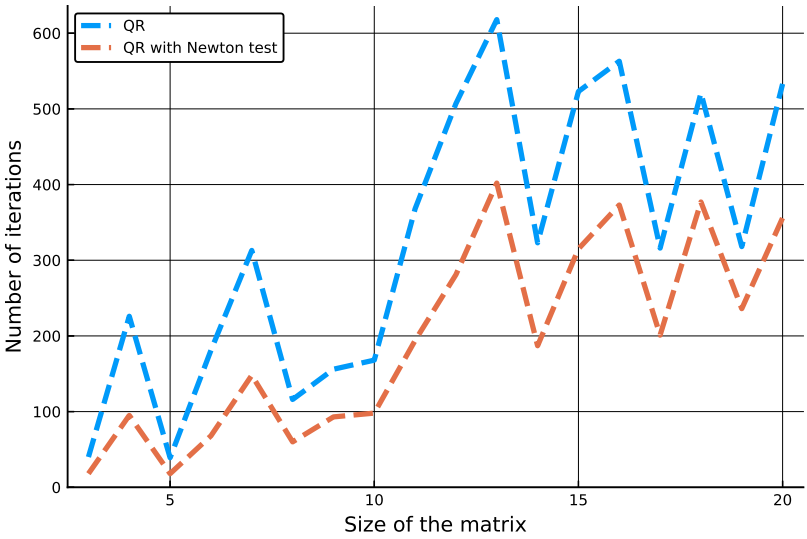

Fig. 1 The number of iterations of respectively Algorithm 1 (with $threshold = 1.e-6$) and Algorithm 3 applied on randomly sampled symmetric positive semi-definite matrices obeying Gaussian distributions of size $n = 3, \ldots, 20$.

A, it needs 230 iterations to provide an initial point satisfying the Newton condition, the initial residual error is $1.45e-5$. Starting from this point the residual error of 6 iterations of the Newton sequences are:
iter1: $3.25e-9$
iter2: $4.07e-19$
iter3: $6.21e-39$
iter4: $1.37e-78$
iter5: $6.68e-158$
iter6: $3.84e-295$
The process took about 0.7 seconds. It took more time than in the previews approach in Section 6.2 and this, not only because there are two more Newton iterations, but also because, as we mentioned before, the QR algorithm implemented in the Julia's solver from which we take the initial point, is more sophisticated. For instance the QR decomposition is applied on a Hessenberg reduction of A. We can also use these techniques to enhance Algorithm 3. However, the main idea that we want to underline here is that the use of the Newton condition in a QR-type algorithm can reduce the number of steps to provide an initial point to the Newton method, and provide an efficient algorithm to compute simple eigenvalues with high precision.

7 Conclusion

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

by incorporating the test given by Theorem 3, the classical QR method gain in efficiency and allows to compute eigenvalues and eigenvectors with high precision.
We focused on the regular case. Some improvements and extension can be considered, such as the treatment of clusters of eigenvalues. Another direction that can be explored, is the construction of higher order methods.

References

[1] P.-A. Absil and K.A. Gallivan. 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] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, Princeton, NJ, 2008.

[3] B. Afsari. Sensitivity analysis for the problem of matrix joint diagonalization. *SIAM Journal on Matrix Analysis and Applications*, 30:1148–1171, 2008.

[4] E. Andruchow, G. Larotonda, L. Recht, and A. Varela. The left invariant metric in the general linear group. *Journal of Geometry and Physics*, 86:241–257, 2014.

[5] Florent Bouchard, Bijan Afsari, Jérôme Malick, and Marco Congedo. Approximate joint diagonalization with Riemannian optimization on the general linear group. *SIAM Journal on Matrix Analysis and Applications*, 41(1):152–170, 2020.

[6] Rasmus Bro. Parafac tutorial and applications. *Chemometrics and Intelligent Laboratory Systems*, 38(2):149–171, 1997.

[7] Angelika Bunse-Gerstner, Ralph Byers, and Volker Mehrmann. A chart of numerical methods for structured eigenvalue problems. *SIAM Journal on Matrix Analysis and Applications*, 13(2):419–453, 1992.

[8] Angelika Bunse-Gerstner, Ralph Byers, and Volker Mehrmann. Numerical methods for simultaneous diagonalization. *SIAM Journal on Matrix Analysis and Applications*, 14(4):927–949, 1993.

[9] Peter Bürgisser, Michael Clausen, and Mohammad A Shokrollahi. *Algebraic complexity theory*, volume 315. Springer Science & Business Media, , 2013.

[10] Jean-François Cardoso and Antoine Souloumiac. Blind beamforming for non gaussian signals. *IEE Proceedings-F*, 140:362–370, 1993.

[11] Jean-François Cardoso and Antoine Souloumiac. Jacobi angles for simultaneous diagonalization. *SIAM Journal on Matrix Analysis and Applications*, 17(1):161–164, 1996.

[12] J Douglas Carroll and Jih-Jie Chang. Analysis of individual differences in multidimensional scaling via an n-way generalization of "Eckart-Young" decomposition. *Psychometrika*, 35(3):283–319, 1970.

[13] Andrzej Cichocki, Danilo Mandic, Lieven De Lathauwer, Guoxu Zhou, Qibin Zhao, Cesar Caiafa, and Huy Anh Phan. Tensor decompositions for signal processing applications: From two-way to multiway component analysis. *IEEE signal processing magazine*, 32(2):145–163, 2015.

[14] Pierre Comon and Christian Jutten. *Handbook of Blind Source Separation: Independent Component Analysis and Applications*. Academic Press, Inc., USA, 1st edition, 2010.

[15] David A. Cox, John B. Little, and Donal O'Shea. *Using Algebraic Geometry*. Number 185 in Graduate Texts in Mathematics. Springer, New York, 2nd edition, 2005.

[16] Lieven De Lathauwer. A link between the canonical decomposition in multilinear algebra and simultaneous matrix diagonalization. *SIAM journal on Matrix Analysis and Applications*, 28(3):642–666, 2006.

[17] James W. Demmel. *Applied Numerical Linear Algebra*. Society for Industrial and Applied Mathematics, USA, 1997.

[18] S.C. Douglas. Self-stabilized gradient algorithms for blind source separation with orthogonality constraints. *IEEE Transactions on Neural Networks*, 11(6):1490–1497, 2000.

[19] Mohamed Elkadi and Bernard Mourrain. *Introduction à la résolution des systèmes polynomiaux*, volume 59 of *Mathèmatiques et Applications*. Springer, , 2007.

[20] Miroslav Fiedler. Expressing a polynomial as the characteristic polynomial of a symmetric matrix. *Linear Algebra and its Applications*, 141:265–270, 1990.

[21] Bernhard N. Flury and Walter Gautschi. An algorithm for simultaneous orthogonal transformation of several positive definite symmetric matrices to nearly diagonal form. *SIAM Journal on Scientific and Statistical Computing*, 7(1):169–184, 1986.

[22] J. V. D. Hoeven and B. Mourrain. Efficient certification of numeric solutions to eigenproblems. In *MACIS*, 2017.

[23] Joris van der Hoeven and Jean-Claude Yakoubsohn. Certified singular value decomposition. Technical Report HAL 01941987, 2018.

[24] Roger A. Horn and Charles R. Johnson. *Topics in Matrix Analysis*. Cambridge University Presss, Cambridge, 1991.

[25] Roger A Horn and Charles R. Johnson. *Matrix analysis*. Cambridge University Press, Cambridge, 2012.

[26] M. Joho and K. Rahbar. Joint diagonalization of correlation matrices by using Newton methods with application to blind signal separation. *Sensor Array and Multichannel Signal Processing Workshop Proceedings, 2002*, pages 403–407, 2002.

[27] Ian Jolliffe. *Principal Component Analysis*, pages 1094–1096. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[28] Xavier Luciani and Laurent Albera. Canonical polyadic decomposition based on joint eigenvalue decomposition. *Chemometrics and Intelligent Laboratory Systems*, 132:152–167, 2014.

[29] Ammar Mesloub, Adel Belouchrani, and Karim Abed-Meraim. Efficient and stable joint eigenvalue decomposition based on generalized givens rotations. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pages 1247–1251. IEEE, 2018.

[30] M. Nikpour, J. Manton, and G. Hori. Algorithms on the Stiefel manifold for joint diagonalisation. *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2:II–1481–II–1484, 2002.

[31] Yasunori Nishimori and Shotaro Akaho. Learning algorithms utilizing quasi-geodesic flows on the Stiefel manifold. *Neurocomput.*, 67:106–135, August 2005.

[32] Kamran Rahbar and James P. Reilly. Geometric optimization methods for blind source separation of signals. In *in Proc. ICA*, pages 375–380, 2000.

[33] Nicholas D Sidiropoulos and Rasmus Bro. On the uniqueness of multilinear decomposition of n-way arrays. *Journal of chemometrics*, 14(3):229–239, 2000.

[34] Nicholas D Sidiropoulos, Lieven De Lathauwer, Xiao Fu, Kejun Huang, Evangelos E Papalexakis, and Christos Faloutsos. Tensor decomposition for signal processing and machine learning. *IEEE Transactions on Signal Processing*, 65(13):3551–3582, 2017.

[35] Mikael Sørensen and Lieven De Lathauwer. Multidimensional harmonic retrieval via coupled canonical polyadic decomposition—part i: Model and identifiability. *IEEE Transactions on Signal Processing*, 65(2):517–527, 2016.

[36] Mikael Sørensen and Lieven De Lathauwer. Multidimensional harmonic retrieval via coupled canonical polyadic decomposition—part ii: Algorithm and multirate sampling. *IEEE Transactions on Signal Processing*, 65(2):528–539, 2016.

[37] Mikael Sørensen, Ignat Domanov, and Lieven De Lathauwer. Coupled canonical polyadic decompositions and multiple shift invariance in array processing. *IEEE Transactions on Signal Processing*, 66(14):3665–3680, 2018.

[38] Mikael Sørensen, Frederik Van Eeghem, and Lieven De Lathauwer. Blind multichannel deconvolution and convolutive extensions of canonical polyadic and block term decompositions. *IEEE Transactions on Signal Processing*, 65(15):4132–4145, 2017.

[39] Wenwu Wang, Saeid Sanei, and Jonathon Chambers. Penalty function-based joint diagonalization approach for convolutive blind separation of nonstationary sources, Jan 2005.

[40] H. Weyl. Das asymptotische verteilungsgesetz der eigenwerte linearer partieller differentialgleichungen (mit einer anwendung auf die theorie der hohlraumstrahlung). *Mathematische Annalen*, 71:441–479, 1912.

[41] A. Yeredor. Non-orthogonal joint diagonalization in the least-squares sense with application in blind source separation. *IEEE Transactions on Signal Processing*, 50(7):1545–1553, 2002.