Japan J. Indust. Appl. Math. (2018) 35:1007–1035
https://doi.org/10.1007/s13160-018-0310-3

ORIGINAL PAPER

Area 2

Iterative refinement for symmetric eigenvalue decomposition

Takeshi Ogita[1] · **Kensuke Aishima**[2]

Received: 6 January 2018 / Revised: 27 February 2018 / Published online: 11 May 2018

Abstract An efficient refinement algorithm is proposed for symmetric eigenvalue problems. The structure of the algorithm is straightforward, primarily comprising matrix multiplications. We show that the proposed algorithm converges quadratically if a modestly accurate initial guess is given, including the case of multiple eigenvalues. Our convergence analysis can be extended to Hermitian matrices. Numerical results demonstrate excellent performance of the proposed algorithm in terms of convergence rate and overall computational cost, and show that the proposed algorithm is considerably faster than a standard approach using multiple-precision arithmetic.

Keywords Accurate numerical algorithm · Iterative refinement · Symmetric eigenvalue decomposition · Quadratic convergence

Mathematics Subject Classification 65F15 · 15A18 · 15A23

This study was partially supported by CREST, JST, and JSPS KAKENHI Grant numbers 16H03917, 17K14143.

✉ Takeshi Ogita
ogita@lab.twcu.ac.jp

Kensuke Aishima
aishima@hosei.ac.jp

[1] Division of Mathematical Sciences, School of Arts and Sciences, Tokyo Woman's Christian University, 2-6-1 Zempukuji, Suginami-ku, Tokyo 167-8585, Japan

[2] Faculty of Computer and Information Sciences, Hosei University, 3-7-2 Kajino-cho, Koganei-shi, Tokyo 184-8584, Japan

1 Introduction

Let A be a real symmetric $n \times n$ matrix. We are concerned with the standard symmetric eigenvalue problem $Ax = \lambda x$, where $\lambda \in \mathbb{R}$ is an eigenvalue of A and $x \in \mathbb{R}^n$ is an eigenvector of A associated with λ. Solving this problem is important because it plays a significant role in scientific computing. For example, highly accurate computations of a few or all eigenvectors are crucial for large-scale electronic structure calculations in material physics [31,32], in which specific interior eigenvalues with associated eigenvectors need to be computed. Excellent overviews on the symmetric eigenvalue problem can be found in references [25,30].

Throughout this paper, I and O denote the identity and the zero matrices of appropriate size, respectively. For matrices, $\|\cdot\|_2$ and $\|\cdot\|_{\mathrm{F}}$ denote the spectral norm and the Frobenius norm, respectively. Unless otherwise specified, $\|\cdot\|$ means $\|\cdot\|_2$. For legibility, if necessary, we distinguish between the approximate quantities and the computed results, e.g., for some quantity α we write $\widetilde{\alpha}$ and $\widehat{\alpha}$ as an approximation of α and a computed result for α, respectively.

This paper aims to develop a method to compute an accurate result of the eigenvalue decomposition

$$A = XDX^{\mathrm{T}}, \tag{1}$$

where X is an $n \times n$ orthogonal matrix whose i-th columns are eigenvectors $x_{(i)}$ of A (called an eigenvector matrix) and D is an $n \times n$ diagonal matrix whose diagonal elements are the corresponding eigenvalues $\lambda_i \in \mathbb{R}$, i.e., $D_{ii} = \lambda_i$ for $i = 1, \ldots, n$. For this purpose, we discuss an iterative refinement algorithm for (1) together with a convergence analysis. Throughout the paper, we assume that

$$\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n,$$

and the columns of X are ordered correspondingly.

Several efficient numerical algorithms for computing (1) have been developed such as the bisection method with inverse iteration, the QR algorithm, the divide-and-conquer algorithm or the MRRR (multiple relatively robust representations) algorithm via Householder reduction, and the Jacobi algorithm. For details, see [10,11,14,15,25,30] and references cited therein. Since such algorithms have been studied actively in numerical linear algebra for decades, there are highly reliable implementations for them, such as LAPACK routines [4]. We stress that we do not intend to compete with such existing algorithms, i.e., we aim to develop an algorithm to improve the results obtained by any of them. Such a refinement algorithm is useful if the quality of the results is not satisfactory. In other words, our proposed algorithm can be regarded as a supplement to existing algorithms for computing (1).

For symmetric matrices, backward stable algorithms usually provide accurate approximate eigenvalues, which can be seen by the perturbation theory (cf., e.g., [25, p.208]). However, they do not necessarily give accurate approximate eigenvectors, since eigenvectors associated with clustered eigenvalues are very sensitive to perturbations (cf., e.g., [25, Theorem 11.7.1]). To see this, we take a small example

$$A = \begin{bmatrix} 1+\varepsilon & 1 & 1+\varepsilon \\ 1 & 1 & -1 \\ 1+\varepsilon & -1 & 1+\varepsilon \end{bmatrix}.$$

For any ε, the eigenvalues and a normalized eigenvector matrix of A are

$$\begin{cases} \lambda_1 = -1 \\ \lambda_2 = 2 \\ \lambda_3 = 2 + 2\varepsilon \end{cases}, \quad X = [x_{(1)}, x_{(2)}, x_{(3)}] = \begin{bmatrix} 1/\sqrt{3} & 1/\sqrt{6} & 1/\sqrt{2} \\ -1/\sqrt{3} & 2/\sqrt{6} & 0 \\ -1/\sqrt{3} & -1/\sqrt{6} & 1/\sqrt{2} \end{bmatrix},$$

where λ_2 and λ_3 are nearly double eigenvalues for small ε. We set $\varepsilon = 2^{-25} \approx 2.98 \times 10^{-8}$ and adopt the MATLAB built-in function eig in IEEE 754 binary64 floating-point arithmetic, which adopts the LAPACK routine DSYEV, to obtain an approximate eigenvector matrix $\widehat{X} = [\widehat{x}_{(1)}, \widehat{x}_{(2)}, \widehat{x}_{(3)}]$ by some backward stable algorithm. Then, the relative rounding error unit is $2^{-53} \approx 1.11 \times 10^{-16}$, and we have $\|\widehat{X} - X\| \approx 5.46 \times 10^{-9}$, which means 7 or 8 decimal digits are lost. More precisely, we have $\|\widehat{x}_{(1)} - x_{(1)}\| \approx 3.42 \times 10^{-16}$ and $\|\widehat{x}_{(2)} - x_{(2)}\| \approx \|\widehat{x}_{(3)} - x_{(3)}\| \approx 5.46 \times 10^{-9}$. Therefore, we still have room for improvement of approximate eigenvectors $\widehat{x}_{(2)}$ and $\widehat{x}_{(3)}$ in the range of binary64. To improve the accuracy of approximate eigenvectors efficiently, we treat all eigenvectors, since all of them correlate with each other in terms of orthogonality $X^{\mathrm{T}}X = I$ and diagonality $X^{\mathrm{T}}AX = D$.

There exist several refinement algorithms for eigenvalue problems that are based on Newton's method for nonlinear equations (cf. e.g., [3,7,12,29]). Since this sort of algorithm is designed to improve eigenpairs $(\lambda, x) \in \mathbb{R} \times \mathbb{R}^n$ individually, applying such a method to all eigenpairs requires $\mathcal{O}(n^4)$ arithmetic operations. To reduce the computational cost, one may consider preconditioning by Householder reduction of A in ordinary floating-point arithmetic such as $T \approx \widehat{H}^{\mathrm{T}} A \widehat{H}$, where T is a tridiagonal matrix, and $\widehat{H}$ is an approximate orthogonal matrix involving rounding errors. However, this is not a similarity transformation; thus, the original problem is slightly perturbed. Assuming $\widehat{H} = H + \Delta_H$ for some orthogonal matrix H with $\epsilon_H = \|\Delta_H\| \ll 1$, we have $\widetilde{T} = \widehat{H}^{\mathrm{T}} A \widehat{H} = H^{\mathrm{T}}(A + \Delta_A)H$ with $\|\Delta_A\| = \mathcal{O}(\|A\|\epsilon_H)$ and $\epsilon_H \approx \|I - \widehat{H}^{\mathrm{T}}\widehat{H}\|$, where Δ_A can be considered a backward error, and the accuracy of eigenpairs using $\widehat{H}$ is limited by its lack of orthogonality.

A possible approach to achieve an accurate eigenvalue decomposition is to use multiple-precision arithmetic for Householder reduction and the subsequent algorithm. In general, however, we do not know in advance how much arithmetic precision is sufficient to obtain results with desired accuracy. Moreover, the use of such multiple-precision arithmetic for entire computations is often much more time-consuming than ordinary floating-point arithmetic such as IEEE 754 binary64. Therefore, we prefer the iterative refinement approach rather than simply using multiple-precision arithmetic.

Simultaneous iteration or Grassmann–Rayleigh quotient iteration [1] can potentially be used to refine eigenvalue decompositions. However, such methods require higher-precision arithmetic for the orthogonalization of approximate eigenvectors. Thus, we cannot restrict the higher-precision arithmetic to matrix multiplication. Wilkinson [30, Chapter 9, pp.637–647] explained the refinement of eigenvalue decompositions for general square matrices with reference to Jahn's method [6,19]. Such

methods rely on a similarity transformation $C := \widehat{X}^{-1}A\widehat{X}$ with high accuracy for a computed result $\widehat{X}$ for X, which requires an accurate solution of the linear system $\widehat{X}C = A\widehat{X}$ for C, and slightly breaks the symmetry of A due to nonorthogonality of $\widehat{X}$. Davies and Modi [9] proposed a direct method to complete the symmetric eigenvalue decomposition of nearly diagonal matrices. The Davies–Modi algorithm assumes that A is preconditioned to a nearly diagonal matrix such as $\widehat{X}^{\mathrm{T}}A\widehat{X}$, where $\widehat{X}$ is an approximate orthogonal matrix involving rounding errors. Again, as mentioned above, this is not a similarity transformation, and a problem similar to the preconditioning by Householder reduction remains unsolved, i.e., the Davies–Modi algorithm does not refine the nonorthogonality of $\widehat{X}$. See Appendix for details regarding the relationship between the Davies–Modi algorithm and ours.

Given this background, we try to derive a simple and efficient iterative refinement algorithm to simultaneously improve the accuracy of all eigenvectors with quadratic convergence, which requires $\mathcal{O}(n^3)$ operations for each iteration. The proposed algorithm can be regarded as a variant of Newton's method, and therefore, its quadratic convergence is naturally derived.

For a computed eigenvector matrix $\widehat{X}$ for (1), define $E \in \mathbb{R}^{n\times n}$ such that $X = \widehat{X}(I + E)$. Here the existence of E requires that $\widehat{X}$ is nonsingular, which is usually satisfied in practice. We assume that $\widehat{X}$ is modestly accurate, e.g., it is obtained by some backward stable algorithm. Then, we aim to compute a sufficiently precise approximation $\widetilde{E}$ of E using the following two relations.

$$\begin{cases} X^{\mathrm{T}}X = I & \text{(orthogonality)} \\ X^{\mathrm{T}}AX = D & \text{(diagonality)} \end{cases} \tag{2}$$

After obtaining $\widetilde{E}$, we can update $X' := \widehat{X}(I + \widetilde{E})$. If necessary, we iterate the process as $X^{(\nu+1)} := X^{(\nu)}(I + \widetilde{E}^{(\nu)})$. Under some conditions, we prove $\widetilde{E}^{(\nu)} \to O$ and $X^{(\nu)} \to X$, where the convergence rates are quadratic. Our analysis provides a sufficient condition for the convergence of the iterations.

The main benefit of our algorithm is their adaptivity, i.e., they allow the adaptive use of high-precision arithmetic until the desired accuracy of the computed results is achieved. In other words, the use of high-precision arithmetic is mandatory, however, it is primarily restricted to matrix multiplication, which accounts for most of the computational cost. Note that a multiple-precision arithmetic library, such as MPFR [21] with GMP [13], could be used for this purpose. With approaches such as quad-double precision arithmetic [16] and arbitrary precision arithmetic [26], multiple-precision arithmetic can be simulated using ordinary floating-point arithmetic. There are more specific approaches for high-precision matrix multiplication. For example, XBLAS (extra precise BLAS) [20] and other accurate and efficient algorithms for dot products [22,27,28] and matrix products [24] based on error-free transformations are available for practical implementation.

The remainder of the paper is organized as follows. In Sect. 2, we present a refinement algorithm for the symmetric eigenvalue decomposition. In Sect. 3, we provide a convergence analysis of the proposed algorithm. In Sect. 4, we present some numerical results showing the behavior and performance of the proposed algorithm. Finally, we

conclude the paper in Sect. 5. To put our results into context, we review existing work in Appendix.

For simplicity, we basically handle only real matrices. The discussions in this paper can be extended to Hermitian matrices. Moreover, discussion of the standard symmetric eigenvalue problem can readily be extended to the generalized symmetric (or Hermitian) definite eigenvalue problem $Ax = \lambda Bx$, where A and B are real symmetric (or Hermitian) with B being positive definite.

2 Proposed algorithm

Let $A = A^{\mathrm{T}} \in \mathbb{R}^{n\times n}$. The eigenvalues of A are denoted by $\lambda_i \in \mathbb{R}$, $i = 1, \ldots, n$. Then $\|A\| = \max_{1\le i\le n} |\lambda_i|$. Let $X \in \mathbb{R}^{n\times n}$ denote an orthogonal eigenvector matrix comprising normalized eigenvectors of A, and let $\widehat{X}$ denote an approximation of X computed by some numerical algorithm.

In the following, we derive our proposed algorithm. Define $E \in \mathbb{R}^{n\times n}$ such that $X = \widehat{X}(I + E)$. The problem is how to derive a method for computing E. Here, we assume that

$$\epsilon := \|E\| < 1, \tag{3}$$

which implies that $I + E$ is nonsingular and so is $\widehat{X}$.

First, we consider the orthogonality of the eigenvectors, that is, $X^{\mathrm{T}}X = I$. From this, we obtain $I = X^{\mathrm{T}}X = (I + E)^{\mathrm{T}}\widehat{X}^{\mathrm{T}}\widehat{X}(I + E)$ and

$$(I + E)^{-\mathrm{T}}(I + E)^{-1} = \widehat{X}^{\mathrm{T}}\widehat{X}. \tag{4}$$

Using Neumann series expansion, we obtain

$$(I + E)^{-1} = I - E + \Delta_E, \quad \Delta_E := \sum_{k=2}^{\infty}(-E)^k. \tag{5}$$

Here, it follows from (3) that

$$\|\Delta_E\| \le \frac{\epsilon^2}{1-\epsilon}. \tag{6}$$

Substituting (5) into (4) yields

$$E + E^{\mathrm{T}} = I - \widehat{X}^{\mathrm{T}}\widehat{X} + \Delta_1, \tag{7}$$

where $\Delta_1 := \Delta_E + \Delta_E^{\mathrm{T}} + (E - \Delta_E)^{\mathrm{T}}(E - \Delta_E)$. Here, it follows from (3) and (6) that

$$\|\Delta_1\| \le \frac{(3-2\epsilon)\epsilon^2}{(1-\epsilon)^2}. \tag{8}$$

In a similar way to Newton's method (cf. e.g., [5, p. 236]), dropping the second order term Δ_1 from (7) for linearization yields the following matrix equation for $\widetilde{E} = (\widetilde{e}_{ij}) \in \mathbb{R}^{n\times n}$:

$$\widetilde{E} + \widetilde{E}^{\mathrm{T}} = I - \widehat{X}^{\mathrm{T}}\widehat{X}. \tag{9}$$

Next, we consider the diagonalization of A such that $X^{\mathrm{T}}AX = D$. From this, we obtain $D = X^{\mathrm{T}}AX = (I+E)^{\mathrm{T}}\widehat{X}^{\mathrm{T}}A\widehat{X}(I+E)$ and

$$(I+E)^{-\mathrm{T}}D(I+E)^{-1} = \widehat{X}^{\mathrm{T}}A\widehat{X}. \tag{10}$$

Substitution of (5) into (10) yields

$$D - DE - E^{\mathrm{T}}D = \widehat{X}^{\mathrm{T}}A\widehat{X} + \Delta_2, \tag{11}$$

where $\Delta_2 := -D\Delta_E - \Delta_E^{\mathrm{T}}D - (E-\Delta_E)^{\mathrm{T}}D(E-\Delta_E)$. Here, it follows from (3) and (6) that

$$\|\Delta_2\| \le \frac{(3-2\epsilon)\epsilon^2}{(1-\epsilon)^2}\|D\| = \frac{(3-2\epsilon)\epsilon^2}{(1-\epsilon)^2}\|A\|. \tag{12}$$

Similar to the derivation of (9), omission of the second order term Δ_2 from (11) yields the following matrix equation for $\widetilde{E} = (\widetilde{e}_{ij}) \in \mathbb{R}^{n\times n}$ and $\widetilde{D} = (\widetilde{d}_{ij}) \in \mathbb{R}^{n\times n}$:

$$\widetilde{D} - \widetilde{D}\widetilde{E} - \widetilde{E}^{\mathrm{T}}\widetilde{D} = \widehat{X}^{\mathrm{T}}A\widehat{X}. \tag{13}$$

Here, we restrict $\widetilde{D}$ to be diagonal, which implies $\widetilde{d}_{ij} = 0$ for $i \ne j$. We put $\widetilde{\lambda}_i := \widetilde{d}_{ii}$ as approximate eigenvalues. Hereafter, we assume that

$$\widetilde{\lambda}_1 \le \widetilde{\lambda}_2 \le \cdots \le \widetilde{\lambda}_n$$

for the sake of simplicity. If this is not the case, we can find an appropriate permutation τ such that

$$\widetilde{\lambda}_{\tau(1)} \le \widetilde{\lambda}_{\tau(2)} \le \cdots \le \widetilde{\lambda}_{\tau(n)},$$

and we redefine $\widetilde{\lambda}_j := \widetilde{\lambda}_{\tau(j)}$ and $\widehat{x}_{(j)} := \widehat{x}_{(\tau(j))}$ for $j = 1, 2, \ldots, n$. Note that the eigenvalue ordering is not essential in this paper.

Summarizing the above discussion of (9) and (13), we obtain the system of matrix equations for $\widetilde{E} = (\widetilde{e}_{ij})$ and $\widetilde{D} = (\widetilde{d}_{ij}) = \mathrm{diag}(\widetilde{\lambda}_i)$ such that

$$\begin{cases} \widetilde{E} + \widetilde{E}^{\mathrm{T}} = R \\ \widetilde{D} - \widetilde{D}\widetilde{E} - \widetilde{E}^{\mathrm{T}}\widetilde{D} = S \end{cases} \tag{14}$$

$$\Leftrightarrow \begin{cases} \widetilde{e}_{ij} + \widetilde{e}_{ji} = r_{ij} \\ \widetilde{d}_{ij} - \widetilde{\lambda}_i\widetilde{e}_{ij} - \widetilde{\lambda}_j\widetilde{e}_{ji} = s_{ij} \end{cases} \quad \text{for } 1 \le i, j \le n, \tag{15}$$

where $R = (r_{ij})$ and $S = (s_{ij})$ are defined as $R := I - \widehat{X}^{\mathrm{T}}\widehat{X}$ and $S := \widehat{X}^{\mathrm{T}}A\widehat{X}$, respectively.

In fact, (14) is surprisingly easy to solve. First, we focus on the diagonal parts of $\widetilde{E}$. From the first equation in (15), it follows that $\widetilde{e}_{ii} = r_{ii}/2$ for $1 \le i \le n$. Moreover, the second equation in (15) also yields $(1 - 2\widetilde{e}_{ii})\widetilde{\lambda}_i = (1 - r_{ii})\widetilde{\lambda}_i = s_{ii}$. If $r_{ii} \ne 1$, then we have

$$\widetilde{\lambda}_i = \frac{s_{ii}}{1 - r_{ii}} \quad \text{for } 1 \le i \le n. \tag{16}$$

Here, $|r_{ii}| \ll 1$ usually hold due to the assumption that the columns of $\widehat{X}$ are normalized approximately. Note that (16) is equivalent to the Rayleigh quotient $\widetilde{\lambda}_i = (\widehat{x}_{(i)}^{\mathrm{T}} A \widehat{x}_{(i)})/(\widehat{x}_{(i)}^{\mathrm{T}} \widehat{x}_{(i)})$, where $\widehat{x}_{(i)}$ is the i-th column of $\widehat{X}$.

Next, we focus on the off-diagonal parts of $\widetilde{E}$. The combination of (15) and (16) yields

$$\begin{cases} \widetilde{e}_{ij} + \widetilde{e}_{ji} = r_{ij} \\ \widetilde{\lambda}_i \widetilde{e}_{ij} + \widetilde{\lambda}_j \widetilde{e}_{ji} = -s_{ij} \end{cases} \quad \text{for } 1 \le i, j \le n,\ i \ne j,$$

which are simply 2×2 linear systems. Theoretically, for (i, j) satisfying $\widetilde{\lambda}_i \ne \widetilde{\lambda}_j$, the linear systems have unique solutions

$$\widetilde{e}_{ij} = \frac{s_{ij} + \widetilde{\lambda}_j r_{ij}}{\widetilde{\lambda}_j - \widetilde{\lambda}_i}. \tag{17}$$

It is easy to see that, for $i \ne j$, $s_{ij} \to 0$, $r_{ij} \to 0$, $\widetilde{\lambda}_i \to \lambda_i$, and $\widetilde{\lambda}_j \to \lambda_j$ as $\|E\| \to 0$. Thus, for some sufficiently small $\|E\|$, all of $|\widetilde{e}_{ij}|$ for $i \ne j$ as in (17) are also small whenever $\lambda_i \ne \lambda_j$ for $i \ne j$. However, multiple eigenvalues require some care. For multiple eigenvalues $\lambda_i = \lambda_j$ with $i \ne j$, noting the first equation in (15) with adding an artificial condition $\widetilde{e}_{ij} = \widetilde{e}_{ji}$, we choose $\widetilde{e}_{ij}$ as

$$\widetilde{e}_{ij} = \frac{r_{ij}}{2}. \tag{18}$$

Similar to the Newton–Schulz iteration, which is discussed in Appendix, the choice as (18) is quite reasonable for improving the orthogonality of $\widehat{X}$ corresponding to $\widehat{x}_{(i)}$ and $\widehat{x}_{(j)}$, which are the i-th and j-th columns of $\widehat{X}$, respectively. Moreover, the accuracy of $\widehat{x}_{(i)}$ and $\widehat{x}_{(j)}$ is improved as shown in Sect. 3.2.

It remains to detect multiple eigenvalues. According to the perturbation theory in [23, Theorem 2], we have

$$|\lambda_i - \widetilde{\lambda}_i| \le \|S - \widetilde{D}\|_2 + \|A\|_2 \|R\|_2 \quad \text{for } i = 1, \ldots, n.$$

In our algorithm, we set

$$\delta := 2(\|S - \widetilde{D}\|_2 + \|A\|_2 \|R\|_2). \tag{19}$$

For all $i \ne j$, $|\widetilde{\lambda}_i - \widetilde{\lambda}_j| \le \delta$ whenever $\lambda_i = \lambda_j$. In addition, for a sufficiently small $\|E\|$, we see that, for all $i \ne j$, $|\widetilde{\lambda}_i - \widetilde{\lambda}_j| > \delta$ whenever $\lambda_i \ne \lambda_j$ in view of $\delta \to 0$ as $\|E\| \to 0$. In other words, we can identify the multiple eigenvalues using δ in (19) associated with $\widehat{X}$ in some neighborhood of X, which is rigorously proven in Sect. 3.

Finally, we present a refinement algorithm for the eigenvalue decomposition of A in Algorithm 1, which is designed to be applied iteratively.

Let us consider the iterative refinement using Algorithm 1:

$$X^{(0)} \leftarrow \widehat{X} \in \mathbb{R}^{n \times n}, \quad X^{(\nu+1)} \leftarrow \mathsf{RefSyEv}(A, X^{(\nu)}) \quad \text{for } \nu = 0, 1, \ldots . \tag{20}$$

Algorithm 1 RefSyEv: Refinement of approximate eigenvectors of a real symmetric matrix. Higher-precision arithmetic is required for all the computations except the line 6.

Input: $A = A^{\mathrm{T}} \in \mathbb{R}^{n\times n}$, $\widehat{X} \in \mathbb{R}^{n\times n}$
Output: $X' \in \mathbb{R}^{n\times n}$
1: **function** $X' \leftarrow$ RefSyEv$(A, \widehat{X})$
2: $R \leftarrow I - \widehat{X}^{\mathrm{T}}\widehat{X}$
3: $S \leftarrow \widehat{X}^{\mathrm{T}}A\widehat{X}$
4: $\widetilde{\lambda}_i \leftarrow s_{ii}/(1 - r_{ii})$ for $i = 1, \ldots, n$ ▷ Compute approximate eigenvalues.
5: $\widetilde{D} \leftarrow \mathrm{diag}(\widetilde{\lambda}_i)$
6: $\delta \leftarrow 2(\|S - \widetilde{D}\|_2 + \|A\|_2\|R\|_2)$
7: $\widetilde{e}_{ij} \leftarrow \begin{cases} \dfrac{s_{ij} + \widetilde{\lambda}_j r_{ij}}{\widetilde{\lambda}_j - \widetilde{\lambda}_i} & \text{if } |\widetilde{\lambda}_i - \widetilde{\lambda}_j| > \delta \\ r_{ij}/2 & \text{otherwise} \end{cases}$ for $1 \le i, j \le n$ ▷ Compute $\widetilde{E}$.
8: $X' \leftarrow \widehat{X} + \widehat{X}\widetilde{E}$ ▷ Update $\widehat{X}$ by $\widehat{X}(I + \widetilde{E})$.
9: **end function**

Then, $X^{(\nu+1)} = X^{(\nu)}(I + \widetilde{E}^{(\nu)})$ for $\nu = 0, 1, \ldots$, where $\widetilde{E}^{(\nu)} = (\widetilde{e}_{ij}^{(\nu)})$ are the quantities calculated at the line 7 in Algorithm 1. Let $E^{(\nu)}$ be defined such that $X = X^{(\nu)}(I + E^{(\nu)})$. Then, $E^{(\nu)}$ corresponds to the true error of $X^{(\nu)}$ for each ν. Under some condition (the assumption (29) in Lemma 1), we obtain

$$\begin{cases} \|E^{(\nu+1)}\| < \|E^{(\nu)}\| & \text{(monotone convergence)} \\ \|E^{(\nu+1)}\| = \mathcal{O}(\|E^{(\nu)}\|^2) & \text{(quadratic convergence)} \end{cases},$$

which will be shown in Sect. 3.1.

Remark 1 On practical computation of δ at the line 6 in Algorithm 1, one may prefer to use the Frobenius norm rather than the spectral norm, since the former is much easier to compute than the latter. For any real $n \times n$ matrix C, it is known (cf., e.g., [14, p.72]) that $\|C\|_2 \le \|C\|_{\mathrm{F}} \le \sqrt{n}\|C\|_2$. Thus, it may cause some overestimate of δ, and affect the behavior of the algorithm. □

Remark 2 For the generalized symmetric definite eigenvalue problem $Ax = \lambda Bx$ where A and B are real symmetric with B being positive definite, a similar algorithm can readily be derived by replacing line 2 in Algorithm 1 with $R \leftarrow I - \widehat{X}^{\mathrm{T}}B\widehat{X}$. □

3 Convergence analysis

In this section, we prove quadratic convergence of Algorithm 1 under the assumption that the approximate solutions are modestly close to the exact solutions. Our analysis is divided into two parts. First, if we assume that A does not have multiple eigenvalues, then quadratic convergence is proven. Next, we consider a general analysis for any A.

Recall that the error of the approximate solution is expressed as $\|\widehat{X} - X\| = \|\widehat{X}E\|$ in view of $X = \widehat{X}(I + E)$. The refined approximate solution is $X' := \widehat{X}(I + \widetilde{E})$. It then follows that the error of the refined solution is expressed as follows:

$$\|\widehat{X}(I + \widetilde{E}) - X\| = \|\widehat{X}(\widetilde{E} - E)\|.$$

In addition, recall that $\widetilde{E}$ is the solution of the following equations:

$$\widetilde{E} + \widetilde{E}^{\mathrm{T}} = R, \tag{21}$$
$$\widetilde{D} - \widetilde{D}\widetilde{E} - \widetilde{E}^{\mathrm{T}}\widetilde{D} = S, \tag{22}$$

where

$$R := I - \widehat{X}^{\mathrm{T}}\widehat{X}, \tag{23}$$
$$S := \widehat{X}^{\mathrm{T}}A\widehat{X}. \tag{24}$$

However, if $\widetilde{\lambda}_i \approx \widetilde{\lambda}_j$ such that $|\widetilde{\lambda}_i - \widetilde{\lambda}_j| \leq \delta$, where δ is defined as in (19), then (22) is not reflected for the computation of $\widetilde{e}_{ij}$ and $\widetilde{e}_{ji}$. In this case, we choose $\widetilde{e}_{ij} = \widetilde{e}_{ji} = r_{ij}/2$ from (21). Such an exceptional case is considered later in the subsection on multiple eigenvalues.

Briefly, our goal is to prove quadratic convergence

$$\|\widehat{X}(I + \widetilde{E}) - X\| = \mathcal{O}(\|\widehat{X} - X\|^2),$$

which corresponds to

$$\|\widehat{X}(\widetilde{E} - E)\| = \mathcal{O}(\|\widehat{X}E\|^2),$$

as $\widehat{X} \to X$. We would like to prove that

$$\|\widetilde{E} - E\| = \mathcal{O}(\|E\|^2) \tag{25}$$

as $\|E\| \to 0$.

To investigate the relationship between E and $\widetilde{E}$, let ϵ be defined as in (3) and

$$\chi(\epsilon) := \frac{3 - 2\epsilon}{(1 - \epsilon)^2}. \tag{26}$$

Then, we see that

$$E + E^{\mathrm{T}} = R + \Delta_1, \quad \|\Delta_1\| \leq \chi(\epsilon)\epsilon^2 \tag{27}$$

from (7) and (8). In addition, we have

$$D - DE - E^{\mathrm{T}}D = S + \Delta_2, \quad \|\Delta_2\| \leq \chi(\epsilon)\|A\|\epsilon^2 \tag{28}$$

from (11) and (12).

3.1 Simple eigenvalues

We focus on the situation where the eigenvalues of A are all simple and a given $\widehat{X}$ is sufficiently close to an orthogonal eigenvector matrix X. First, we derive a sufficient condition that (17) is chosen for all (i, j), $i \neq j$ in Algorithm 1.

Lemma 1 *Let A be a real symmetric $n \times n$ matrix with simple eigenvalues λ_i, $i = 1, 2, \ldots, n$ and a corresponding orthogonal eigenvector matrix $X \in \mathbb{R}^{n\times n}$. For a given nonsingular $\widehat{X} \in \mathbb{R}^{n\times n}$, suppose that Algorithm 1 is applied to A and $\widehat{X}$ in exact arithmetic, and $\widetilde{D} = \mathrm{diag}(\widetilde{\lambda}_i)$, R, S, and δ are the quantities calculated in Algorithm 1. Define E such that $X = \widehat{X}(I + E)$. If*

$$\epsilon := \|E\| < \min\left(\frac{\min_{i\neq j}|\lambda_i - \lambda_j|}{10n\|A\|}, \frac{1}{100}\right), \tag{29}$$

then we obtain

$$\min_{i\neq j}|\widetilde{\lambda}_i - \widetilde{\lambda}_j| > \delta(= 2(\|S - \widetilde{D}\| + \|A\|\|R\|)). \tag{30}$$

Proof First, it is easy to see that

$$(E - \widetilde{E}) + (E - \widetilde{E})^{\mathrm{T}} = \Delta_1, \quad \|\Delta_1\| \leq \chi(\epsilon)\epsilon^2 \tag{31}$$

from (21), (23), and (27). Hence, we obtain

$$|\widetilde{e}_{ii} - e_{ii}| \leq \frac{\chi(\epsilon)}{2}\epsilon^2 \quad \text{for } i = 1, \ldots, n \tag{32}$$

from the diagonal elements in (31). From (22) and (28), $\widetilde{D} = \mathrm{diag}(\widetilde{\lambda}_i)$ and $D = \mathrm{diag}(\lambda_i)$ are determined as $\widetilde{\lambda}_i = s_{ii}/(1 - 2\widetilde{e}_{ii})$, $\lambda_i = (s_{ii} + \Delta_2(i, i))/(1 - 2e_{ii})$. Thus, we have

$$\begin{aligned}\widetilde{\lambda}_i - \lambda_i &= \frac{s_{ii}(1 - 2e_{ii}) - (s_{ii} + \Delta_2(i, i))(1 - 2\widetilde{e}_{ii})}{(1 - 2e_{ii})(1 - 2\widetilde{e}_{ii})} \\ &= -\frac{(1 - 2\widetilde{e}_{ii})\Delta_2(i, i) + 2(e_{ii} - \widetilde{e}_{ii})s_{ii}}{(1 - 2e_{ii})(1 - 2\widetilde{e}_{ii})} \\ &= -\frac{\Delta_2(i, i)}{1 - 2e_{ii}} - \frac{2(e_{ii} - \widetilde{e}_{ii})s_{ii}}{(1 - 2e_{ii})(1 - 2\widetilde{e}_{ii})}.\end{aligned} \tag{33}$$

For the first term on the right-hand side, we see

$$\left|\frac{\Delta_2(i, i)}{1 - 2e_{ii}}\right| \leq \frac{\chi(\epsilon)}{1 - 2e_{ii}}\|A\|\epsilon^2$$

from (28). Moreover, for the second term,

$$\left|\frac{2(e_{ii} - \widetilde{e}_{ii})s_{ii}}{(1 - 2e_{ii})(1 - 2\widetilde{e}_{ii})}\right| \leq \frac{(1 + 2\epsilon + \chi(\epsilon)\epsilon^2)\chi(\epsilon)}{(1 - 2e_{ii})(1 - 2\widetilde{e}_{ii})}\|A\|\epsilon^2$$

from (24), (28), and (32). In addition, we see

$$\begin{aligned}
&\frac{\chi(\epsilon)}{1-2e_{ii}} + \frac{(1+2\epsilon+\chi(\epsilon)\epsilon^2)\chi(\epsilon)}{(1-2e_{ii})(1-2\widetilde{e}_{ii})} \\
&= \frac{(1-2\widetilde{e}_{ii}) + (1+2\epsilon+\chi(\epsilon)\epsilon^2)}{(1-2e_{ii})(1-2\widetilde{e}_{ii})}\chi(\epsilon) \\
&\leq \frac{2(1+2\epsilon+\chi(\epsilon)\epsilon^2)\chi(\epsilon)}{(1-2\epsilon)(1-2\epsilon-\chi(\epsilon)\epsilon^2)} =: \eta(\epsilon). \qquad (34)
\end{aligned}$$

Combining this with (33), we obtain

$$|\widetilde{\lambda}_i - \lambda_i| \leq \eta(\epsilon)\|A\|\epsilon^2 \quad \text{for } i = 1, \ldots, n. \qquad (35)$$

Hence, noting the definition of δ as in (30), we derive

$$\begin{aligned}
\delta &\leq 2(\|S-D\| + \|A\|\|R\| + \|D-\widetilde{D}\|) \\
&\leq 2(2(2\|A\|\epsilon + \chi(\epsilon)\|A\|\epsilon^2) + \eta(\epsilon)\|A\|\epsilon^2) \\
&\leq 2(4 + 2\chi(\epsilon)\epsilon + \eta(\epsilon)\epsilon)\|A\|\epsilon \\
&< 2(4 + 2\chi(\epsilon)\epsilon + \eta(\epsilon)\epsilon) \cdot \frac{\min_{p\neq q}|\lambda_p - \lambda_q|}{10n}, \qquad (36)
\end{aligned}$$

where the second inequality is due to (27), (28), and (35), and the last inequality is due to (29). In addition, from (26), $\epsilon < 1/100$ in (29), and (34), we see

$$\chi(\epsilon) = \frac{3-2\epsilon}{(1-\epsilon)^2} < 3.05, \quad \eta(\epsilon) = \frac{2(1+2\epsilon+\chi(\epsilon)\epsilon^2)\chi(\epsilon)}{(1-2\epsilon)(1-2\epsilon-\chi(\epsilon)\epsilon^2)} < 7. \qquad (37)$$

Thus, we find that, for all (i, j), $i \neq j$,

$$\begin{aligned}
|\widetilde{\lambda}_i - \widetilde{\lambda}_j| &\geq |\lambda_i - \lambda_j| - 2\eta(\epsilon)\|A\|\epsilon^2 \\
&> \min_{p\neq q}|\lambda_p - \lambda_q| - 2\eta(\epsilon) \cdot \frac{\min_{p\neq q}|\lambda_p - \lambda_q|}{10n} \cdot \frac{1}{100} \\
&> \left(10 - \frac{14}{100}\right) \cdot \frac{\min_{p\neq q}|\lambda_p - \lambda_q|}{10n} \\
&> 2(4 + 2\chi(\epsilon)\epsilon + \eta(\epsilon)\epsilon) \cdot \frac{\min_{p\neq q}|\lambda_p - \lambda_q|}{10n} \\
&> \delta \ (= 2(\|S-\widetilde{D}\| + \|A\|\|R\|)),
\end{aligned}$$

where the first inequality is due to (35), the second inequality is due to (29), the third inequality is due to the second inequality in (37), the fourth inequality is due to (37) and $\epsilon < 1/100$ as in (29), and the last inequality is due to (36), respectively. Thus, we obtain (30). □

The assumption (29) is crucial for the first iteration in the iterative process (20). In the following, monotone convergence of $\|E^{(\nu)}\|$ is proven under the assumption (29) for a given initial guess $\widehat{X} = X^{(0)}$ and $E = E^{(0)}$, so that $\|E^{(\nu+1)}\| < \|E^{(\nu)}\|$ for $\nu = 0, 1, \ldots$. Thus, in the iterative refinement using Algorithm 1, Lemma 1 ensures that $|\widetilde{\lambda}_i^{(\nu)} - \widetilde{\lambda}_j^{(\nu)}| > \delta^{(\nu)}$ for all (i, j), $i \neq j$ as in (30) are consecutively satisfied for $X^{(\nu)}$ in the iterative process. In addition, recall that our aim is to prove the quadratic convergence in the asymptotic regime. To this end, we derive a key lemma that shows (25).

Lemma 2 *Let A be a real symmetric $n \times n$ matrix with simple eigenvalues λ_i, $i = 1, 2, \ldots, n$ and a corresponding orthogonal eigenvector matrix $X \in \mathbb{R}^{n\times n}$. For a given nonsingular $\widehat{X} \in \mathbb{R}^{n\times n}$, suppose that Algorithm 1 is applied to A and $\widehat{X}$ in exact arithmetic, and $\widetilde{E}$ is the quantity calculated in Algorithm 1. Define E such that $X = \widehat{X}(I + E)$. Under the assumption* (29) *in Lemma* 1*, we have*

$$\|\widetilde{E} - E\| < \frac{7}{10}\|E\|, \tag{38}$$

$$\limsup_{\|E\|\to 0} \frac{\|\widetilde{E} - E\|}{\|E\|^2} \leq \frac{6n\|A\|}{\min_{i\neq j} |\lambda_i - \lambda_j|}. \tag{39}$$

Proof Let ϵ, $\chi(\cdot)$, and $\eta(\cdot)$ be defined as in Lemma 1, (26), and (34), respectively. Note that the diagonal elements of $\widetilde{E} - E$ are estimated as in (32). In the following, we estimate the off-diagonal elements of $\widetilde{E} - E$. To this end, define

$$\widetilde{\Delta}_2 := \widetilde{D} - \widetilde{D}E - E^{\mathrm{T}}\widetilde{D} - S. \tag{40}$$

Noting (28), (35), and (40), we see that the off-diagonal elements of $|\Delta_2 - \widetilde{\Delta}_2|$ are less than $2\eta(\epsilon)\|A\|\epsilon^3$. In other words,

$$|\widetilde{\Delta}_2(i, j)| \leq (\chi(\epsilon) + 2\eta(\epsilon)\epsilon)\|A\|\epsilon^2 \tag{41}$$

for $i \neq j$ from (28), where $\widetilde{\Delta}_2(i, j)$ are the (i, j) elements of $\widetilde{\Delta}_2$. In addition, from (22), it follows that

$$\widetilde{D}(E - \widetilde{E}) + (E - \widetilde{E})^{\mathrm{T}}\widetilde{D} = -\widetilde{\Delta}_2. \tag{42}$$

From (31), (41) and (42), we have

$$(e_{ij} - \widetilde{e}_{ij}) + (e_{ji} - \widetilde{e}_{ji}) = \epsilon_1, \quad |\epsilon_1| \leq \chi(\epsilon)\epsilon^2, \tag{43}$$

$$\widetilde{\lambda}_i(e_{ij} - \widetilde{e}_{ij}) + \widetilde{\lambda}_j(e_{ji} - \widetilde{e}_{ji}) = \epsilon_2, \quad |\epsilon_2| \leq (\chi(\epsilon) + 2\eta(\epsilon)\epsilon)\|A\|\epsilon^2. \tag{44}$$

It then follows that

$$e_{ij} - \widetilde{e}_{ij} = \frac{\epsilon_2 - \widetilde{\lambda}_j\epsilon_1}{\widetilde{\lambda}_i - \widetilde{\lambda}_j}, \quad e_{ji} - \widetilde{e}_{ji} = \frac{\epsilon_2 - \widetilde{\lambda}_i\epsilon_1}{\widetilde{\lambda}_j - \widetilde{\lambda}_i}.$$

Therefore, using (35), we obtain

$$|\widetilde{e}_{ij} - e_{ij}| \le \frac{(2\chi(\epsilon) + 2\eta(\epsilon)\epsilon + \chi(\epsilon)\eta(\epsilon)\epsilon^2)\|A\|\epsilon^2}{|\widetilde{\lambda}_i - \widetilde{\lambda}_j|} \le \frac{(2\chi(\epsilon) + 2\eta(\epsilon)\epsilon + \chi(\epsilon)\eta(\epsilon)\epsilon^2)\|A\|\epsilon^2}{|\lambda_i - \lambda_j| - 2\eta(\epsilon)\|A\|\epsilon^2}. \tag{45}$$

Note that $\|\widetilde{E} - E\|^2 \le \|\widetilde{E} - E\|_{\mathrm{F}}^2 = \sum_{i,j} |\widetilde{e}_{ij} - e_{ij}|^2$ and

$$\frac{\chi(\epsilon)}{2}\epsilon^2 \le \frac{(2\chi(\epsilon) + 2\eta(\epsilon)\epsilon + \chi(\epsilon)\eta(\epsilon)\epsilon^2)\|A\|\epsilon^2}{|\lambda_i - \lambda_j| - 2\eta(\epsilon)\|A\|\epsilon^2} \quad (i \ne j)$$

in (32) and (45). Therefore, we obtain

$$\|\widetilde{E} - E\| \le \frac{(2\chi(\epsilon) + 2\eta(\epsilon)\epsilon + \chi(\epsilon)\eta(\epsilon)\epsilon^2)n\|A\|\epsilon^2}{\min_{i\ne j} |\lambda_i - \lambda_j| - 2\eta(\epsilon)\|A\|\epsilon^2}.$$

Combining this with $\chi(0) = 3$ proves (39). Moreover, we have

$$\|\widetilde{E} - E\| < \frac{(2\chi(\epsilon) + 2\eta(\epsilon)\epsilon + \chi(\epsilon)\eta(\epsilon)\epsilon^2)\epsilon}{\frac{\min_{i\ne j}|\lambda_i - \lambda_j|}{n\|A\|\epsilon} - \frac{2\eta(\epsilon)\epsilon}{n}} < \frac{6.4\epsilon}{10 - \frac{14}{100n}} < \frac{7}{10}\epsilon \tag{46}$$

from (29) and (37). □

Using the above lemmas, we obtain a main theorem that states the quadratic convergence of Algorithm 1 if all eigenvalues are simple and a given $\widehat{X}$ is sufficiently close to X.

Theorem 1 *Let A be a real symmetric $n \times n$ matrix with simple eigenvalues λ_i, $i = 1, 2, \ldots, n$ and a corresponding orthogonal eigenvector matrix $X \in \mathbb{R}^{n\times n}$. For a given nonsingular $\widehat{X} \in \mathbb{R}^{n\times n}$, suppose that Algorithm 1 is applied to A and $\widehat{X}$ in exact arithmetic, and X' is the quantity calculated in Algorithm 1. Define E and E' such that $X = \widehat{X}(I + E)$ and $X = X'(I + E')$, respectively. Under the assumption* (29) *in Lemma* 1*, we have*

$$\|E'\| < \frac{5}{7}\|E\|, \tag{47}$$

$$\limsup_{\|E\|\to 0} \frac{\|E'\|}{\|E\|^2} \le \frac{6n\|A\|}{\min_{i\ne j}|\lambda_i - \lambda_j|}. \tag{48}$$

Proof Noting $X'(I + E') = \widehat{X}(I + E)$ $(= X)$ and $X' = \widehat{X}(I + \widetilde{E})$, where $\widetilde{E}$ is the quantity calculated in Algorithm 1, we have

$$X'E' = \widehat{X}(E - \widetilde{E}).$$

Therefore, we obtain

$$E' = (I + \widetilde{E})^{-1}(E - \widetilde{E}). \tag{49}$$

Noting (46) and

$$\|\widetilde{E}\| \le \|\widetilde{E} - E\| + \|E\| \le \frac{17}{10}\|E\| < \frac{1}{50}$$

from (29) and (38), we have

$$\|E'\| \le \frac{\|\widetilde{E} - E\|}{1 - \|\widetilde{E}\|} < \frac{\frac{7}{10}\|E\|}{1 - \frac{1}{50}} = \frac{5}{7}\|E\|. \tag{50}$$

Finally, using (49) and (39), we obtain (48). □

Our analysis indicates that Algorithm 1 may not be convergent for very large n. However, in practice, n is much smaller than $1/\epsilon$ for $\epsilon := \|E\|$ when the initial guess $\widehat{X}$ is computed by some backward stable algorithm, e.g., in IEEE 754 binary64 arithmetic, unless A has nearly multiple eigenvalues. In such a situation, the iterative refinement works well.

Remark 3 For any $\widetilde{\delta} \ge \delta$, we can replace δ by $\widetilde{\delta}$ in Algorithm 1. For example, such cases arise when the Frobenius norm is used for calculating δ instead of the spectral norm as mentioned in Remark 1. In such cases, the quadratic convergence of the algorithm can also be proven in a similar way as in this subsection by replacing the assumption (29) by

$$\epsilon := \|E\| < \min\left(\frac{1}{\rho} \cdot \frac{\min_{i \ne j} |\lambda_i - \lambda_j|}{10n\|A\|}, \frac{1}{100}\right), \tag{51}$$

where $\rho := \widetilde{\delta}/\delta \ge 1$. More specifically, in the convergence analysis, (36) is replaced with

$$\begin{aligned}\widetilde{\delta} &= \rho \cdot 2(\|S - \widetilde{D}\| + \|A\|\|R\|) \le \rho \cdot 2(4 + 2\chi(\epsilon)\epsilon + \eta(\epsilon)\epsilon)\|A\|\epsilon \\ &< 2(4 + 2\chi(\epsilon)\epsilon + \eta(\epsilon)\epsilon) \cdot \frac{\min_{p \ne q} |\lambda_p - \lambda_q|}{10n},\end{aligned}$$

where the last inequality is due to the assumption (51). Therefore, $|\widetilde{\lambda}_i - \widetilde{\lambda}_j| > \widetilde{\delta}$ also hold for $i \ne j$ in the same manner as the proof of Lemma 1. As a result, (38) and (39) are also established even if δ is replaced by $\widetilde{\delta}$. □

3.2 Multiple eigenvalues

Multiple eigenvalues require some care. If $\widetilde{\lambda}_i \approx \widetilde{\lambda}_j$ corresponding to multiple eigenvalues $\lambda_i = \lambda_j$, we might not be able to solve the linear system given by (21) and (22). Therefore, we use equation (21) only, i.e., $\widetilde{e}_{ij} = \widetilde{e}_{ji} = r_{ij}/2$ if $|\widetilde{\lambda}_i - \widetilde{\lambda}_j| \le \delta$.

To investigate the above exceptional process, let us consider a simple case as follows. Suppose λ_i, $i \in \mathcal{M} := \{1, 2, \ldots, p\}$ are multiple, i.e., $\lambda_1 = \cdots = \lambda_p < \lambda_{p+1} < \cdots < \lambda_n$. Then, the eigenvectors corresponding to λ_i, $1 \le i \le p$ are not unique. Suppose $X = [x_{(1)}, \ldots, x_{(n)}] \in \mathbb{R}^{n\times n}$ is an orthogonal eigenvector matrix of A, where $x_{(i)}$ are the normalized eigenvectors corresponding to λ_i for $i = 1, \ldots, n$. Define $X_{\mathcal{M}} := [x_{(1)}, \ldots, x_{(p)}] \in \mathbb{R}^{n\times p}$ and $X_{\mathcal{S}} := [x_{(p+1)}, \ldots, x_{(n)}] \in \mathbb{R}^{n\times(n-p)}$. Then, the columns of $X_{\mathcal{M}}Q$ are also the eigenvectors of A for any orthogonal matrix $Q \in \mathbb{R}^{p\times p}$. Thus, let $\mathcal{V}$ be the set of $n \times n$ orthogonal eigenvector matrices of A and $\mathcal{E} := \{\widehat{X}^{-1}X - I : X \in \mathcal{V}\}$ for a given nonsingular $\widehat{X}$.

The key idea of the proof of quadratic convergence below is to define an orthogonal eigenvector matrix $Y \in \mathcal{V}$ as follows. For any $X_\alpha \in \mathcal{V}$, splitting $\widehat{X}^{-1}X_{\mathcal{M}}$ into the first p rows $V_\alpha \in \mathbb{R}^{p\times p}$ and the remaining $(n-p)$ rows $W_\alpha \in \mathbb{R}^{(n-p)\times p}$, we have

$$\widehat{X}^{-1}X_{\mathcal{M}} = \begin{bmatrix} V_\alpha \\ W_\alpha \end{bmatrix} = \begin{bmatrix} C \\ W_\alpha Q_\alpha^{\mathrm{T}} \end{bmatrix} Q_\alpha \tag{52}$$

in view of the polar decomposition $V_\alpha = CQ_\alpha$, where $C = \sqrt{V_\alpha V_\alpha^{\mathrm{T}}} \in \mathbb{R}^{p\times p}$ is symmetric and positive semidefinite and $Q_\alpha \in \mathbb{R}^{p\times p}$ is orthogonal. Note that, although $X_{\mathcal{M}}Q$ for any orthogonal matrix $Q \in \mathbb{R}^{p\times p}$ is also an eigenvector matrix, the symmetric and positive semidefinite matrix C is independent of Q. In other words, we have

$$\widehat{X}^{-1}(X_{\mathcal{M}}Q) = (\widehat{X}^{-1}X_{\mathcal{M}})Q = \begin{bmatrix} V_\alpha Q \\ W_\alpha Q \end{bmatrix} = \begin{bmatrix} C \\ W_\alpha Q_\alpha^{\mathrm{T}} \end{bmatrix} Q_\alpha Q, \tag{53}$$

where the last equality implies the polar decomposition of $V_\alpha Q$. In addition, if V_α in (52) is nonsingular, the orthogonal matrix Q_α is uniquely determined by the polar decomposition for a fixed $X_\alpha \in \mathcal{V}$. To investigate the features of V_α, we suppose that $\widehat{X}$ is an exact eigenvector matrix. Then, noting $\widehat{X}^{-1} = \widehat{X}^{\mathrm{T}}$, we see that V_α is an orthogonal matrix and $C = I$ and $V_\alpha = Q_\alpha$ in the polar decomposition of V_α in (52). Thus, for any fixed $\widehat{X}$ in some neighborhood of $\mathcal{V}$, the above polar decomposition $V_\alpha = CQ_\alpha$ is unique, where the nonsingular matrix V_α depends on $X_\alpha \in \mathcal{V}$. In the following, we consider any $\widehat{X}$ in such a neighborhood of $\mathcal{V}$.

Recall that, for any orthogonal matrix Q, the last equality in (53) is due to the polar decomposition $V_\alpha Q = C(Q_\alpha Q)$. Hence, we have an eigenvector matrix

$$(X_{\mathcal{M}}Q)(Q_\alpha Q)^{\mathrm{T}} = X_{\mathcal{M}}Q_\alpha^{\mathrm{T}} = \widehat{X}\begin{bmatrix} C \\ W_\alpha Q_\alpha^{\mathrm{T}} \end{bmatrix},$$

which is independent of Q. Thus, we define the unique matrix $Y := [X_{\mathcal{M}}Q_\alpha^{\mathrm{T}}, X_{\mathcal{S}}]$ for all matrices in $\mathcal{V}$, where Y depends only on $\widehat{X}$. Then, the corresponding error term $F = (f_{ij})$ is uniquely determined as

$$\begin{aligned} F &:= \widehat{X}^{-1}Y - I = [\widehat{X}^{-1}X_{\mathcal{M}}Q_\alpha^{\mathrm{T}}, *\,] - I \\ &= \begin{bmatrix} C - I & * \\ * & * \end{bmatrix}, \end{aligned} \tag{54}$$

which implies $f_{ij} = f_{ji}$ corresponding to the multiple eigenvalues $\lambda_i = \lambda_j$. Therefore,

$$f_{ij} = f_{ji} = \frac{r_{ij} + \Delta_1(i,j)}{2} \tag{55}$$

from (27), where $\Delta_1(i,j)$ denote (i,j) elements of Δ_1 for all (i,j). Now, let us consider the situation where $\widehat{X}$ is an exact eigenvector matrix. In (52), noting $\widehat{X}^{-1} = \widehat{X}^{\mathrm{T}}$, we have $W_\alpha = O$ and $C = I$ in the polar decomposition of V_α. Combining the features with (54), we see $F = O$ for the exact eigenvector matrix $\widehat{X}$.

Our aim is to prove $\|F\| \to 0$ in the iterative refinement for $\widehat{X} \approx Y \in \mathcal{V}$, where Y depends on $\widehat{X}$. To this end, for the refined X' as a result of Algorithm 1, we also define an eigenvector matrix $Y' \in \mathcal{V}$ and $F' := (X')^{-1}Y' - I$ such that the submatrices of $(X')^{-1}Y'$ corresponding to the multiple eigenvalues are symmetric and positive definite. Note that the eigenvector matrix Y is changed to Y' corresponding to X' after the refinement.

On the basis of the above observations, we consider general eigenvalue distributions. First of all, define the index sets $\mathcal{M}_k$, $k = 1, 2, \ldots, M$ for multiple eigenvalues $\{\lambda_i\}_{i\in\mathcal{M}_k}$ satisfying the following conditions:

$$\begin{cases} \text{(a) } \mathcal{M}_k \subseteq \{1,2,\ldots,n\} \text{ with } n_k := |\mathcal{M}_k| \geq 2 \\ \text{(b) } \lambda_i = \lambda_j,\ \forall i,j \in \mathcal{M}_k \\ \text{(c) } \lambda_i \neq \lambda_j,\ \forall i \in \mathcal{M}_k,\ \forall j \in \{1,2,\ldots,n\} \setminus \mathcal{M}_k \end{cases}. \tag{56}$$

Using the above definitions, we obtain the following key lemma to prove quadratic convergence.

Lemma 3 *Let A be a real symmetric $n \times n$ matrix with the eigenvalues λ_i, $i = 1, 2, \ldots, n$. Suppose A has multiple eigenvalues with index sets $\mathcal{M}_k$, $k = 1, 2, \ldots, M$, satisfying (56). Let $\mathcal{V}$ be the set of $n \times n$ orthogonal eigenvector matrices of A. For a given nonsingular $\widehat{X} \in \mathbb{R}^{n\times n}$, define $\mathcal{E}$ as*

$$\mathcal{E} := \{\widehat{X}^{-1}X - I : X \in \mathcal{V}\}. \tag{57}$$

In addition, define $Y \in \mathcal{V}$ such that, for all k, the $n_k \times n_k$ submatrices of $\widehat{X}^{-1}Y$ corresponding to $\{\lambda_i\}_{i\in\mathcal{M}_k}$ are symmetric and positive semidefinite. Moreover, define $F \in \mathcal{E}$ such that $Y = \widehat{X}(I + F)$. Then, for any $E_\alpha \in \mathcal{E}$,

$$\|F\| \leq 3\|E_\alpha\|. \tag{58}$$

Furthermore, if there exists some Y such that $\|F\| < 1$, then $Y \in \mathcal{V}$ is uniquely determined.

Proof For any $E_\alpha = (e_{ij}^{(\alpha)}) \in \mathcal{E}$, let E_{diag} denote the block diagonal part of E_α, where the $n_k \times n_k$ blocks of E_{diag} correspond to n_k multiple eigenvalues $\{\lambda_i\}_{i\in\mathcal{M}_k}$, i.e.,

$$E_{\mathrm{diag}}(i,j) := \begin{cases} e_{ij}^{(\alpha)} & \text{if } \lambda_i = \lambda_j \\ 0 & \text{otherwise} \end{cases}$$

for all $1 \le i, j \le n$, where $\lambda_1 \le \cdots \le \lambda_n$. Here, we consider the polar decomposition

$$I + E_{\mathrm{diag}} =: HU_\alpha, \tag{59}$$

where H is a symmetric and positive semidefinite matrix and U_α is an orthogonal matrix. Note that, similarly to C in (52), H is unique and independent of the choice of $X_\alpha \in \mathcal{V}$ that satisfies $X_\alpha = \widehat{X}(I+E_\alpha)$, whereas U_α is not always uniquely determined. Then, we have

$$Y = X_\alpha U_\alpha^{\mathrm{T}} \tag{60}$$

from the definition of Y and

$$\begin{aligned} F &= \widehat{X}^{-1}Y - I \\ &= \widehat{X}^{-1}X_\alpha U_\alpha^{\mathrm{T}} - I \\ &= (E_\alpha + I)U_\alpha^{\mathrm{T}} - I \\ &= (E_\alpha - E_{\mathrm{diag}} + HU_\alpha)U_\alpha^{\mathrm{T}} - I \\ &= (E_\alpha - E_{\mathrm{diag}})U_\alpha^{\mathrm{T}} + H - I, \end{aligned} \tag{61}$$

where the first, second, third, and fourth equalities are consequences of the definition of F, (60), (57), and (59), respectively. Here, we see that

$$\|H - I\| \le \|E_{\mathrm{diag}}\| \tag{62}$$

because all the eigenvalues of H are the singular values of HU_α in (59) that range over the interval $[1 - \|E_{\mathrm{diag}}\|, 1 + \|E_{\mathrm{diag}}\|]$ from the Weyl's inequality for singular values. In addition, note that

$$\|E_{\mathrm{diag}}\| \le \|E_\alpha\|. \tag{63}$$

Therefore, we obtain

$$\|F\| = \|(E_\alpha - E_{\mathrm{diag}})U_\alpha^{\mathrm{T}} + (H - I)\| \le 3\|E_\alpha\|$$

from (61), (62), and (63), giving us (58).

Finally, we prove that Y is unique if $\|F\| < 1$. In the above discussion, if X_α is replaced with some $Y \in \mathcal{V}$, then $E_\alpha = F$, and thus $\|E_{\mathrm{diag}}\| \le \|E_\alpha\| = \|F\| < 1$ in (59). Therefore, $U_\alpha = I$ in (59) due to the uniqueness of the polar decomposition of the nonsingular matrix $I + E_{\mathrm{diag}}$, which implies the uniqueness of Y from (60). □

Moreover, we have the next lemma, corresponding to Lemmas 1 and 2.

Lemma 4 *Let A be a real symmetric $n \times n$ matrix with the eigenvalues λ_i, $i = 1, 2, \ldots, n$. Suppose A has multiple eigenvalues with index sets $\mathcal{M}_k$, $k = 1, 2, \ldots, M$, satisfying* (56)*. For a given nonsingular $\widehat{X} \in \mathbb{R}^{n\times n}$, suppose that Algorithm 1 is*

applied to A and $\widehat{X}$ in exact arithmetic, and $\widetilde{D} = \mathrm{diag}(\widetilde{\lambda}_i)$, $\widetilde{E}$, and δ are the quantities calculated in Algorithm 1. Let F be defined as in Lemma 3. Assume that

$$\epsilon_F := \|F\|0 < \frac{1}{3} \cdot \min\left(\frac{\min_{\lambda_i \neq \lambda_j} |\lambda_i - \lambda_j|}{10n\|A\|}, \frac{1}{100} \right). \tag{64}$$

Then, we obtain

$$\|F - \widetilde{E}\| \leq \frac{(2\chi(\epsilon_F) + 2\eta(\epsilon_F)\epsilon_F + \chi(\epsilon_F)\eta(\epsilon_F){\epsilon_F}^2)n\|A\|\epsilon_F^2}{\min_{\lambda_i \neq \lambda_j} |\lambda_i - \lambda_j| - 2\eta(\epsilon_F)\|A\|\epsilon_F^2}, \tag{65}$$

where $\chi(\cdot)$ and $\eta(\cdot)$ are defined as in (26) and (34), respectively.

Proof First, we see that, for $i \neq j$ corresponding to $\lambda_i \neq \lambda_j$,

$$|\widetilde{e}_{ij} - f_{ij}| \leq \frac{(2\chi(\epsilon_F) + 2\eta(\epsilon_F)\epsilon_F + \chi(\epsilon_F)\eta(\epsilon_F){\epsilon_F}^2)\|A\|\epsilon_F^2}{|\lambda_i - \lambda_j| - 2\eta(\epsilon_F)\|A\|\epsilon_F^2},$$

similar to the proof of (45) in Lemma 2. Concerning the multiple eigenvalues $\lambda_i = \lambda_j$, noting $|\widetilde{\lambda}_i - \widetilde{\lambda}_j| \leq \delta$ and (55), we have

$$|\widetilde{e}_{ij} - f_{ij}| \leq \frac{\chi(\epsilon_F)}{2}\epsilon_F^2 \text{ for } (i, j) \text{ corresponding to } \lambda_i = \lambda_j.$$

Note that, for (i, j) corresponding to $\lambda_i \neq \lambda_j$,

$$\frac{\chi(\epsilon_F)}{2}\epsilon_F^2 \leq \frac{(2\chi(\epsilon_F) + 2\eta(\epsilon_F)\epsilon_F + \chi(\epsilon_F)\eta(\epsilon_F){\epsilon_F}^2)\|A\|\epsilon_F^2}{|\lambda_i - \lambda_j| - 2\eta(\epsilon_F)\|A\|\epsilon_F^2}$$

in the above two inequalities for the elements of $F - \widetilde{E}$. Therefore, we have

$$\begin{aligned} \|F - \widetilde{E}\| \leq \|F - \widetilde{E}\|_{\mathrm{F}} &= \sqrt{\sum_{1 \leq i, j \leq n} |f_{ij} - \widetilde{e}_{ij}|^2} \\ &\leq \frac{(2\chi(\epsilon_F) + 2\eta(\epsilon_F)\epsilon_F + \chi(\epsilon_F)\eta(\epsilon_F){\epsilon_F}^2)n\|A\|\epsilon_F^2}{\min_{\lambda_i \neq \lambda_j} |\lambda_i - \lambda_j| - 2\eta(\epsilon_F)\|A\|\epsilon_F^2} \end{aligned}$$

similar to the proof in Lemma 2. □

On the basis of Lemmas 3 and 4 and Theorem 1, we see the quadratic convergence for a real symmetric matrix A that has multiple eigenvalues.

Theorem 2 *Let A be a real symmetric $n \times n$ matrix with the eigenvalues λ_i, $i = 1, 2, \ldots, n$. Suppose A has multiple eigenvalues with index sets $\mathcal{M}_k$, $k = 1, 2, \ldots, M$, satisfying (56). Let $\mathcal{V}$ be the set of $n \times n$ orthogonal eigenvector matrices of A. For a given nonsingular $\widehat{X} \in \mathbb{R}^{n \times n}$, suppose that Algorithm 1 is applied to A and $\widehat{X}$*

in exact arithmetic, and X' and δ are the quantities calculated in Algorithm 1. Let $Y, Y' \in \mathcal{V}$ be defined such that, for all k, the $n_k \times n_k$ submatrices of $\widehat{X}^{-1}Y$ and $(X')^{-1}Y'$ corresponding to $\{\lambda_i\}_{i\in\mathcal{M}_k}$ are symmetric and positive definite. Define F and F' such that $Y = \widehat{X}(I+F)$ and $Y' = X'(I+F')$, respectively. Furthermore, suppose that (64) in Lemma 4 is satisified for $\epsilon_F := \|F\|$. Then, we obtain

$$\|F'\| < \frac{5}{7}\|F\|, \tag{66}$$

$$\limsup_{\|F\|\to 0} \frac{\|F'\|}{\|F\|^2} \le 3\left(\frac{6n\|A\|}{\min_{\lambda_i\neq\lambda_j}|\lambda_i-\lambda_j|}\right). \tag{67}$$

Proof Let $\widetilde{E}$ and $\widetilde{\lambda}_i, i = 1, 2, \ldots, n$, be the quantities calculated in Algorithm 1. First, note that (65) in Lemma 4 is established. Next, we define $G := (X')^{-1}Y - I$. Then, we have

$$G = (I+\widetilde{E})^{-1}(F-\widetilde{E}) \tag{68}$$

similar to (49). Moreover, similar to (46), we have

$$\begin{aligned}\|\widetilde{E}-F\| &< \frac{(2\chi(\epsilon_F)+2\eta(\epsilon_F)\epsilon_F+\chi(\epsilon_F)\eta(\epsilon_F){\epsilon_F}^2)\epsilon_F}{\frac{\min_{\lambda_i\neq\lambda_j}|\lambda_i-\lambda_j|}{n\|A\|\epsilon_F}-\frac{2\eta(\epsilon_F)\epsilon_F}{n}} \\ &< \frac{6.4\epsilon_F}{3(10-\frac{14}{100n})} < \frac{1}{3}\cdot\frac{7}{10}\epsilon_F\end{aligned}$$

from (65) and (64). Therefore, we see

$$\|G\| < \frac{1}{3}\cdot\frac{5}{7}\epsilon_F$$

in a similar manner as the proof of (50). Using (58), we have

$$\|F'\| \le 3\|G\|. \tag{69}$$

Therefore, we obtain (66). Since we see

$$\limsup_{\epsilon_F\to 0}\frac{\|G\|}{\|F\|^2} \le \frac{6n\|A\|}{\min_{\lambda_i\neq\lambda_j}|\lambda_i-\lambda_j|}$$

from (68) and (65), we obtain (67) from (69). □

In the iterative refinement, Theorem 2 shows that the error term $\|F\|$ is quadratically convergent to zero. Note that $\widehat{X}$ is also convergent to some fixed eigenvector matrix X because Theorem 2 and (65) imply $\|\widetilde{E}\|/\|F\| \to 1$ as $\|F\| \to 0$ in $X' := \widehat{X}(I+\widetilde{E})$, where $\|F\|$ is quadratically convergent to zero.

3.3 Complex case

For a Hermitian matrix $A \in \mathbb{C}^{n\times n}$, we must note that, for any unitary diagonal matrix U, XU is also an eigenvector matrix, i.e., there is a continuum of normalized eigenvector matrices in contrast to the real case. Related to this, note that $R := I - \widehat{X}^{\mathrm{H}}\widehat{X}$ and $S := \widehat{X}^{\mathrm{H}}A\widehat{X}$, and (14) is replaced with $\widetilde{E} + \widetilde{E}^{\mathrm{H}} = R$ in the complex case; thus, the diagonal elements $\widetilde{e}_{ii}$ for $i = 1, \ldots, n$ are not uniquely determined in $\mathbb{C}$. Now, select $\widetilde{e}_{ii} = r_{ii}/2 \in \mathbb{R}$ for $i = 1, \ldots, n$. Then, we can prove quadratic convergence using the polar decomposition in the same way as in the discussion of multiple eigenvalues in the real case. More precisely, we define a normalized eigenvector matrix Y as follows. First, we focus on the situation where all eigenvalues are simple. For a given nonsingular $\widehat{X}$, let Y be defined such that all diagonal elements of $\widehat{X}^{-1}Y$ are positive real numbers. In addition, let $F := \widehat{X}^{-1}Y - I$. Then, we see the quadratic convergence of F in the same way as in Theorem 2.

Corollary 1 *Let $A \in \mathbb{C}^{n\times n}$ be a Hermitian matrix whose eigenvalues λ_i, $i = 1, 2, \ldots, n$, are all simple. Let $\mathcal{V}$ be the set of $n \times n$ unitary eigenvector matrices of A. For a given nonsingular $\widehat{X} \in \mathbb{C}^{n\times n}$, suppose that Algorithm 1 is applied to A and $\widehat{X}$, and a nonsingular X' is obtained. Define $Y, Y' \in \mathcal{V}$ such that all the diagonal elements of $\widehat{X}^{-1}Y$ and $(X')^{-1}Y'$ are positive real numbers. Furthermore, define F and F' such that $Y = \widehat{X}(I + F)$ and $Y' = X'(I + F')$, respectively. If*

$$\|F\| < \frac{1}{3}\min\left(\frac{\min_{\lambda_i\neq\lambda_j}|\lambda_i - \lambda_j|}{10n\|A\|}, \frac{1}{100}\right), \tag{70}$$

then we obtain

$$\|F'\| < \frac{5}{7}\|F\|,$$
$$\limsup_{\|F\|\to 0}\frac{\|F'\|}{\|F\|^2} \le 3\left(\frac{6n\|A\|}{\min_{\lambda_i\neq\lambda_j}|\lambda_i - \lambda_j|}\right).$$

For a general Hermitian matrix having multiple eigenvalues, we define Y in the same manner as in Theorem 2, resulting in the following corollary.

Corollary 2 *Let $A \in \mathbb{C}^{n\times n}$ be a Hermitian matrix with the eigenvalues λ_i, $i = 1, 2, \ldots, n$. Suppose A has multiple eigenvalues with index sets $\mathcal{M}_k$, $k = 1, 2, \ldots, M$, satisfying (56). For a given nonsingular $\widehat{X} \in \mathbb{C}^{n\times n}$, let Y, Y', F, F', and δ be defined as in Corollary 1. Suppose that, for all k, the $n_k \times n_k$ submatrices of $\widehat{X}^{-1}Y$ and $(X')^{-1}Y'$ corresponding to $\{\lambda_i\}_{i\in\mathcal{M}_k}$ are Hermitian and positive definite. Furthermore, suppose that (70) is satisfied. Then, we obtain*

$$\|F'\| < \frac{5}{7}\|F\|,$$
$$\limsup_{\|F\|\to 0}\frac{\|F'\|}{\|F\|^2} \le 3\left(\frac{6n\|A\|}{\min_{\lambda_i\neq\lambda_j}|\lambda_i - \lambda_j|}\right).$$

Note that $\widehat{X}$ in Corollaries 1 and 2 is convergent to some fixed eigenvector matrix X of A in the same manner as in the real case.

4 Numerical results

Here, we present numerical results to demonstrate the effectiveness of the proposed algorithm (Algorithm 1). The numerical experiments discussed in this section were conducted using MATLAB R2016b on a PC with two CPUs (3.0 GHz Intel Xeon E5-2687W v4) and 1 TB of main memory. Let $\mathbf{u}$ denote the relative rounding error unit ($\mathbf{u} = 2^{-24}$ for IEEE binary32, and $\mathbf{u} = 2^{-53}$ for binary64). To realize multiple-precision arithmetic, we adopt Advanpix Multiprecision Computing Toolbox version 4.2.3 [2], which utilizes well-known, fast, and reliable multiple-precision arithmetic libraries including GMP and MPFR. In all cases, we use the MATLAB function norm for computing the spectral norms $\|R\|$ and $\|S - \widetilde{D}\|$ in Algorithm 1 in binary64 arithmetic, and approximate $\|A\|$ by $\max_{1 \le i \le n} |\widetilde{\lambda}_i|$. We discuss numerical experiments for some dozens of seeds for the random number generator, and all results are similar to those provided in this section. Therefore, we adopt the default seed as a typical example using the MATLAB command `rng('default')` for reproducibility.

4.1 Convergence property

Here, we confirm the convergence property of the proposed algorithm for various eigenvalue distributions.

4.1.1 Various eigenvalue distributions

We first generate real symmetric and positive definite matrices using the MATLAB function randsvd from Higham's test matrices [17] by the following MATLAB command.

```
>> A = gallery('randsvd',n,-cnd,mode);
```

The eigenvalue distribution and condition number of A can be controlled by the input arguments `mode` $\in \{1, 2, 3, 4, 5\}$ and `cnd` $=: \alpha \ge 1$, as follows:

1. one large: $\lambda_1 \approx 1, \lambda_i \approx \alpha^{-1}, i = 2, \ldots, n$
2. one small: $\lambda_n \approx \alpha^{-1}, \lambda_i \approx 1, i = 1, \ldots, n-1$
3. geometrically distributed: $\lambda_i \approx \alpha^{-(i-1)/(n-1)}, i = 1, \ldots, n$
4. arithmetically distributed: $\lambda_i \approx 1 - (1 - \alpha^{-1})(i-1)/(n-1), i = 1, \ldots, n$
5. random with uniformly distributed logarithm: $\lambda_i \approx \alpha^{-r(i)}, i = 1, \ldots, n$, where $r(i)$ are pseudo-random values drawn from the standard uniform distribution on $(0, 1)$.

Here, $\kappa(A) \approx$ `cnd` for `cnd` $< \mathbf{u}^{-1} \approx 10^{16}$. Note that for `mode` $\in \{1, 2\}$, there is a cluster of eigenvalues that are not exactly but nearly multiple due to rounding errors when A is generated using randsvd, i.e., all clustered eigenvalues are slightly separated.

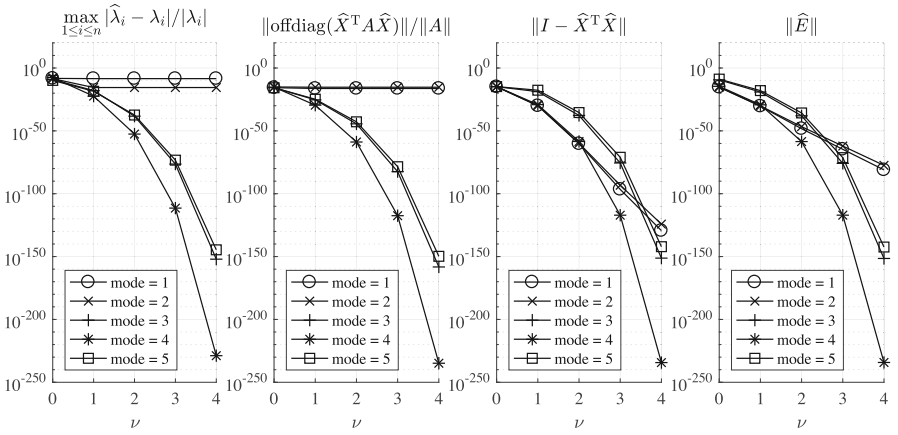

Fig. 1 Results of iterative refinement by Algorithm 1 (RefSyEv) in sufficiently long precision arithmetic for symmetric and positive definite matrices generated by randsvd with $n = 10$ and $\kappa(A) \approx 10^8$

We start with small examples such as $n = 10$, since they are sufficiently illustrative to observe the convergence behavior of the algorithm. Moreover, we set $\mathtt{cnd} = 10^8$ to generate moderately ill-conditioned problems in binary64. We compute $X^{(0)}$ as an initial approximate eigenvector matrix using the MATLAB function eig for the eigenvalue decomposition in binary64 arithmetic, which adopts the LAPACK routine DSYEV. Therefore, $X^{(0)}$ suffers from rounding errors. To see the behavior of Algorithm 1 precisely, we use multiple-precision arithmetic with sufficiently long precision to simulate the exact arithmetic in Algorithm 1. Then, we expect that Algorithm 1 (RefSyEv) works effectively for $\mathtt{mode} \in \{3, 4, 5\}$, but does not for $\mathtt{mode} \in \{1, 2\}$. We also use the built-in function eig in the multiple-precision toolbox to compute the eigenvalues λ_i, $i = 1, 2, \ldots, n$ for evaluation. The results are shown in Fig. 1, which provides $\max_{1\le i\le n} |\widehat{\lambda}_i - \lambda_i|/|\lambda_i|$ as the maximum relative error of the computed eigenvalues $\widehat{\lambda}_i$, $\|\mathrm{offdiag}(\widehat{X}^{\mathrm{T}} A \widehat{X})\|/\|A\|$ as the diagonality of $\widehat{X}^{\mathrm{T}} A \widehat{X}$, $\|I - \widehat{X}^{\mathrm{T}}\widehat{X}\|$ as the orthogonality of a computed eigenvector matrix $\widehat{X}$, and $\|\widehat{E}\|$, where $\widehat{E}$ is a computed result of $\widetilde{E}$ in Algorithm 1. Here, offdiag$(\cdot)$ denotes the off-diagonal part. The horizontal axis shows the number of iterations ν of Algorithm 1.

In the case of $\mathtt{mode} \in \{3, 4, 5\}$, all the quantities decrease quadratically in every iteration, i.e., we observe the quadratic convergence of Algorithm 1, as expected. On the other hand, in the case of $\mathtt{mode} \in \{1, 2\}$, the algorithm fails to improve the accuracy of the initial approximate eigenvectors because the test matrices for $\mathtt{mode} \in \{1, 2\}$ have nearly multiple eigenvalues and the assumption (29) for the convergence of Algorithm 1 is not satisfied for these eigenvalues.

4.1.2 Multiple eigenvalues

We deal with the case where A has exactly multiple eigenvalues. It is not trivial to generate such matrices using of floating-point arithmetic because rounding errors slightly perturb the eigenvalues and multiplicity is broken. To avoid this, we utilize a

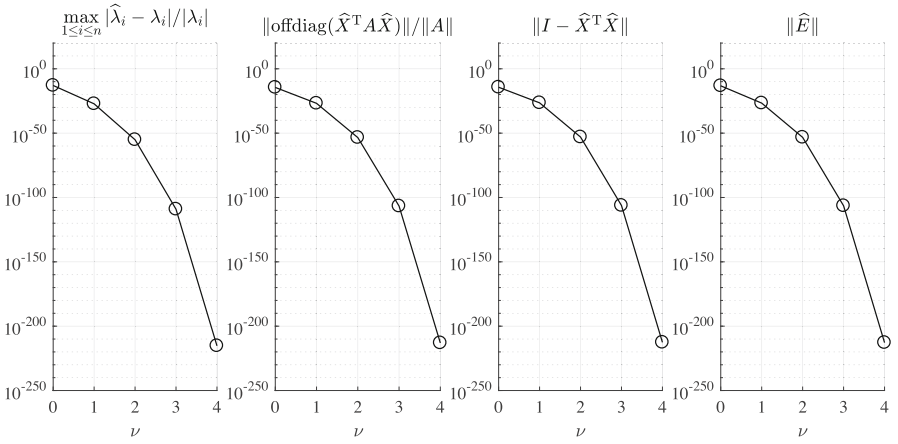

Fig. 2 Results of iterative refinement by Algorithm 1 (RefSyEv) in sufficiently long precision arithmetic for a symmetric 256×256 matrix A having exactly 10-fold eigenvalues

Hadamard matrix H of order n whose elements are 1 or -1 with $H^{\mathrm{T}}H = nI$. For a given $k < n$, let $A = \frac{1}{n} H D H^{\mathrm{T}}$ where

$$D = \mathrm{diag}(\overbrace{-1, \ldots, -1}^{k}, 1, 2, \ldots, n-k).$$

Then A has k-fold eigenvalues $\lambda_i = -1, i = 1, \ldots, k$, and $\lambda_{i+k} = i, i = 1, \ldots, n-k$. We set $n = 256$ and $k = 10$, where A is exactly representable in binary64 format, and compute $X^{(0)}$ using eig in binary64 arithmetic. We apply Algorithm 1 to A and $X^{(\nu)}$ for $\nu = 0, 1, 2, \ldots$. The results are shown in Fig. 2. As can be seen, Algorithm 1 converges quadratically for matrices with multiple eigenvalues, which is consistent with our convergence analysis (Theorem 2).

4.2 Computational speed

To evaluate the computational speed of the proposed algorithm (Algorithm 1), we compare the computing time of Algorithm 1 to that of an approach using multiple-precision arithmetic, which is called "MP-approach". We also compare the results of Algorithm 1 to those of a method combining the Davies–Modi algorithm [9] with the Newton–Schulz iteration [18, Section 8.3], which is called "NS + DM method" and explained in Appendix. Note that the timing should be observed for reference because the computing times for Algorithm 1 and the NS + DM method strongly depend on the implementation of accurate matrix multiplication. Thus, we adopt an efficient method proposed by Ozaki et al. [24] that utilizes fast matrix multiplication routines such as xGEMM in BLAS. In the multiple-precision toolbox, the MRRR algorithm [11] via Householder reduction is implemented sophisticatedly with parallelism to solve symmetric eigenvalue problems.

Table 1 Results for a pseudo-random real symmetric matrix, $n = 100$

Algorithm 1	`eig` (binary64)	$\nu = 1$	$\nu = 2$	$\nu = 3$
$\|\widehat{E}\|$	5.6×10^{-14}	1.8×10^{-27}	3.9×10^{-54}	2.0×10^{-107}
Elapsed time (s)	0.01	0.35	0.31	0.51
(accumulated)	0.01	0.36	0.67	1.18
NS + DM	`eig` (binary64)	$\nu = 1$	$\nu = 2$	$\nu = 3$
$\|I - \widehat{U}\|$	5.5×10^{-14}	2.2×10^{-28}	6.7×10^{-53}	4.4×10^{-106}
Elapsed time (s)	0.01	0.48	0.52	1.02
(accumulated)	0.01	0.49	1.01	2.04
MP-approach	`mp.Digits(d)`	`d` = 34	`d` = 55	`d` = 108
Elapsed time (s)		0.11	0.53	0.62

Table 2 Results for a pseudo-random real symmetric matrix, $n = 500$

Algorithm 1	`eig` (binary64)	$\nu = 1$	$\nu = 2$	$\nu = 3$
$\|\widehat{E}\|$	5.1×10^{-13}	1.3×10^{-25}	2.5×10^{-50}	9.5×10^{-100}
Elapsed time (s)	0.04	1.32	3.54	9.35
(accumulated)	0.04	1.37	4.91	14.25
NS + DM	`eig` (binary64)	$\nu = 1$	$\nu = 2$	$\nu = 3$
$\|I - \widehat{U}\|$	5.1×10^{-13}	1.7×10^{-27}	5.1×10^{-50}	1.9×10^{-99}
Elapsed time (s)	0.04	3.39	6.84	22.31
(accumulated)	0.04	3.43	10.26	32.58
MP-approach	`mp.Digits(d)`	`d` = 34	`d` = 52	`d` = 101
Elapsed time (s)		3.64	39.35	48.12

As test matrices, we generate pseudo-random real symmetric $n \times n$ matrices with $n \in \{100, 500, 1000\}$ using the MATLAB function randn such as `B = randn(n)` and `A = B + B'`. We use eig in binary64, and iteratively refine the computed eigenvectors using Algorithm 1 three times. We do it similarly for the NS+DM method. In the multiple-precision toolbox, we can control the arithmetic precision `d` in decimal digits using the command `mp.Digits(d)`. Corresponding to the results of Algorithm 1, we adjust `d` for $\nu = 1, 2, 3$. For timing fairness, we adopt `d` = max(`d`, 34) because the case for `d` = 34 is faster than that of `d` < 34. In Tables 1, 2 and 3, we show $\|\widehat{E}\|$ and the measured computing time. For the NS+DM method, we show $\|I - \widehat{U}\|$, where $\widehat{U}$ is the correction term in the Davies–Modi algorithm, so that $\|I - \widehat{U}\|$ is comparable to $\|\widehat{E}\|$ in Algorithm 1. As shown by $\|\widehat{E}\|$, Algorithm 1 quadratically improves the accuracy of the computed eigenvectors. Compared to $\|I - \widehat{U}\|$, the behavior of the NS+DM method is similar to those of Algorithm 1. As expected, Algorithm 1 is much faster than the NS+DM method. The accuracy of the results obtained using the

Table 3 Results for a pseudo-random real symmetric matrix, $n = 1000$

Algorithm 1	`eig` (binary64)	$v = 1$	$v = 2$	$v = 3$
$\|\widehat{E}\|$	7.5×10^{-13}	2.8×10^{-25}	1.2×10^{-49}	2.2×10^{-98}
Elapsed time (s)	0.10	5.21	14.45	40.20
(accumulated)	0.10	5.31	19.76	59.96
NS + DM	`eig` (binary64)	$v = 1$	$v = 2$	$v = 3$
$\|I - \widehat{U}\|$	7.5×10^{-13}	6.2×10^{-27}	2.1×10^{-48}	4.0×10^{-97}
Elapsed time (s)	0.10	14.34	33.11	121.54
(accumulated)	0.10	14.44	47.55	169.09
MP-approach	`mp.Digits(d)`	`d` = 34	`d` = 51	`d` = 100
Elapsed time (s)		17.23	291.24	356.39

MP-approach corresponds to the arithmetic precision `d`. As can be seen, Algorithm 1 is considerably faster than the MP-approach for greater n.

5 Conclusion

We have proposed a refinement algorithm (Algorithm 1) for eigenvalue decomposition of real symmetric matrices that can be applied iteratively. Quadratic convergence of the proposed algorithm was proven for a sufficiently accurate initial guess, similar to Newton's method.

The proposed algorithm benefits from the availability of efficient matrix multiplication in higher-precision arithmetic. The numerical results demonstrate the excellent performance of the proposed algorithm in terms of convergence rate and measured computing time.

In practice, it is likely that ordinary floating-point arithmetic, such as IEEE 754 binary32 or binary64, is used for calculating an approximation $\widehat{X}$ to an eigenvector matrix X of a given symmetric matrix A. As done in the numerical experiments, we can use $\widehat{X}$ as an initial guess $X^{(0)}$ in Algorithm 1. However, if A has nearly multiple eigenvalues, it is difficult to obtain a sufficiently accurate $X^{(0)}$ in ordinary floating-point arithmetic such that Algorithm 1 works well. Our future work is to overcome this problem.

Acknowledgements The first author would like to express his sincere thanks to Professor Chen Greif at the University of British Columbia for his valuable comments and helpful suggestions.

A Appendix: Relation to previous work

First, we briefly explain the Davies–Modi algorithm [9], which is relevant to this study. The Davies–Modi algorithm assumes that a real symmetric matrix A is transformed into a nearly diagonal symmetric matrix $S = \widehat{X}^{\mathrm{T}} A \widehat{X}$ by some nearly orthogonal matrix $\widehat{X}$. Under that assumption, the method aims to compute an accurate eigenvector matrix U of S. The idea is seen in Jahn's method [6,19] as shown by Davies and Smith [8], which derives an SVD algorithm based on the Davies–Modi algorithm. To demonstrate the relationship between the Davies–Modi and the algorithm proposed in this work, we explain the Davies–Modi algorithm in the same manner as in reference [8]. Note that U^{T} is written as

$$U^{\mathrm{T}} = e^{Y} = I + Y + \frac{1}{2}Y^2 + \frac{1}{6}Y^3 + \cdots \tag{71}$$

for some skew-symmetric matrix Y. To compute Y with high accuracy, we define a diagonal matrix D_S whose diagonal elements are the eigenvalues of S. Then,

$$D_S = U^{\mathrm{T}} S U = S + YS - SY + \frac{1}{2}(Y^2 S + SY^2) - YSY + \mathcal{O}(\|Y\|^3). \tag{72}$$

Here we define $S_0 = \mathrm{diag}(s_{11}, \ldots, s_{nn})$, and $S_1 = S - S_0$ corresponds to the off-diagonal entries of S. Under a mild assumption $S_1 = \mathcal{O}(\|Y\|)$, we see

$$\begin{aligned} D_S = {} & S_0 + (S_1 + YS_0 - S_0 Y) \\ & + \left(YS_1 - S_1 Y + \frac{1}{2}(Y^2 S_0 + S_0 Y^2) - YS_0 Y \right) + \mathcal{O}(\|Y\|^3). \end{aligned} \tag{73}$$

For the first-order approximation, we would like to solve

$$S_1 + \widetilde{Y}_1 S_0 - S_0 \widetilde{Y}_1 = O. \tag{74}$$

To construct an orthogonal matrix $e^{-\widetilde{Y}_1}$, suppose that $\widetilde{Y}_1$ is a skew-symmetric matrix. Then, $(\widetilde{Y}_1)_{ij} = -(\widetilde{Y}_1)_{ji} = s_{ij}/(s_{ii} - s_{jj})$ for $i \neq j$. To compute U with high accuracy, we construct the second-order approximation $\widetilde{U}^{\mathrm{T}} = I + \widetilde{Y}_1 + \widetilde{Y}_2 + \widetilde{Y}_1^2/2$ for skew-symmetric matrices $\widetilde{Y}_1$ and $\widetilde{Y}_2$. To compute $\widetilde{Y}_2$, letting

$$T = \frac{1}{2}(\widetilde{Y}_1 S_1 - S_1 \widetilde{Y}_1), \quad T_0 = \mathrm{diag}(t_{11}, \ldots, t_{nn}), \quad T_1 = T - T_0, \tag{75}$$

we solve $T_1 + \widetilde{Y}_2 S_0 - S_0 \widetilde{Y}_2 = O$ for a skew-symmetric matrix $\widetilde{Y}_2$. Note that $\widetilde{Y}_2$ is computed in the same manner as $\widetilde{Y}_1$ in (74). Thus, we obtain $\widetilde{U}^{\mathrm{T}}$, where its second-order approximation is proven rigorously. Since the Davies–Modi algorithm is based on matrix multiplication, it requires $\mathcal{O}(n^3)$ operations.

In the following, we discuss the refinement of $\widehat{X}$. Suppose $\widehat{X}$ is computed using ordinary floating-point arithmetic. The main problem is that, since the Davies–Modi algorithm is applied to $S = \widehat{X}^{\mathrm{T}} A \widehat{X}$ to compute the eigenvalue decomposition of S,

$\widehat{X}$ is assumed to be an orthogonal matrix. Generally, however, this is not the case because $\widehat{X}$ suffers from rounding errors. Therefore, even if the computation of $\widehat{X}^{\mathrm{T}} A \widehat{X}$ is performed in exact arithmetic, A and S are not similar unless $\widehat{X}$ is orthogonal. Then, the eigenvalues of A are slightly perturbed, which may cause a significant change of the eigenvectors associated with clustered eigenvalues. To overcome this problem, we must refine the orthogonality of $\widehat{X}$ as preconditioning in higher-precision arithmetic. Recall that we intend to restrict the use of higher-precision arithmetic to matrix multiplications for better computational efficiency. Hence, we may use the Newton–Schulz iteration [18, Section 8.3] such as

$$Z = \frac{1}{2}\widehat{X}(3I - \widehat{X}^{\mathrm{T}}\widehat{X}), \tag{76}$$

where all the singular values of $\widehat{X}$ are quadratically convergent to 1. After this reorthogonalization, for an eigenvector matrix U of $T = Z^{\mathrm{T}} A Z$, the columns of $X' := ZU$ are expected to be sufficiently accurate approximation of the eigenvectors of A. Of course, in practice, we must derive a certain method to compute an approximate eigenvector matrix $\widetilde{U}$ of T. If T is nearly diagonal, it is effective to apply a diagonal shift to T. Similarly, the Jacobi and Davies–Modi algorithms would be able to compute $\widetilde{U}$ accurately using higher-precision arithmetic. As a result, we can obtain a sufficiently accurate eigenvector matrix $X' := Z\widetilde{U}$.

It is worth noting that our approach can reproduce the first-order approximation step in the Davies–Modi algorithm as follows. Since we see

$$\widehat{X}^{\mathrm{T}} A \widehat{X} = S = U D_S U^{\mathrm{T}} = e^{-Y} D_S e^{Y} = D_S + D_S Y - Y D_S + \mathcal{O}(\|Y\|^2) \tag{77}$$

from (71), we obtain $S = \widetilde{D}_S + \widetilde{D}_S\widetilde{Y}_1 - \widetilde{Y}_1\widetilde{D}_S$ as the linearization process in the same manner as (13). It is easy to see that $\widetilde{D}_S = \mathrm{diag}(s_{11}, \ldots, s_{nn})$ and $\widetilde{Y}_1$ is equal to the solution in (74). Hence, if $\widehat{X}$ in Algorithm 1 is an orthogonal matrix and the diagonal elements of S are sufficiently separated, then $\widetilde{E}$ is equal to the skew-symmetric matrix $\widetilde{Y}_1$ in view of $R = O$. In other words, from the viewpoint of the Davies–Modi algorithm, the second-order approximation step is removed and the skew-symmetry of $\widetilde{E}$ is not assumed in Algorithm 1. Instead, condition (9) is integrated to improve orthogonality. If we assume $\widetilde{E} = \widetilde{E}^{\mathrm{T}}$ in (9), we have

$$X' = \widehat{X}(I + \widetilde{E}) = \widehat{X}\left(I + \frac{1}{2}(I - \widehat{X}^{\mathrm{T}}\widehat{X})\right), \tag{78}$$

which is equivalent to the Newton–Schulz iteration (76). In other words, if all eigenvalues are considered clustered, $\widehat{X}$ is refined by the Newton–Schulz iteration. For orthogonality, we require the only relation (9), while an iteration function for polar decomposition must be an odd function, such as the Newton–Schulz iteration.

In summary, to combine the Newton–Schulz iteration and the Davies–Modi algorithm sophisticatedly, we remove unnecessary conditions from both, resulting in Algorithm 1, which can be considered an ideal method to refine orthogonality and diagonality simultaneously.

References

1. Absil, P.A., Mahony, R., Sepulchre, R., Van Dooren, P.: A Grassmann-Rayleigh quotient iteration for computing invariant subspaces. SIAM Rev. **44**, 57–73 (2006)
2. Advanpix: Multiprecision Computing Toolbox for MATLAB, Code and documentation. http://www.advanpix.com/ (2016)
3. Ahuesa, M., Largillier, A., D'Almeida, F.D., Vasconcelos, P.B.: Spectral refinement on quasi-diagonal matrices. Linear Algebra Appl. **401**, 109–117 (2005)
4. Anderson, E., Bai, Z., Bischof, C., Blackford, S., Demmel, J., Dongarra, J., Du Croz, J., Greenbaum, A., Hammarling, S., McKenney, A., Sorensen, D.: LAPACK Users' Guide, 3rd edn. SIAM, Philadelphia (1999)
5. Atkinson, K., Han, W.: Theoretical Numerical Analysis, 3rd edn. Springer, New York (2009)
6. Collar, A.R.: Some notes on Jahn's method for the improvement of approximate latent roots and vectors of a square matrix. Quart. J. Mech. Appl. Math. **1**, 145–148 (1948)
7. Davies, P.I., Higham, N.J., Tisseur, F.: Analysis of the Cholesky method with iterative refinement for solving the symmetric definite generalized eigenproblem. SIAM J. Matrix Anal. Appl. **23**, 472–493 (2001)
8. Davies, P.I., Smith, M.I.: Updating the singular value decomposition. J. Comput. Appl. Math. **170**, 145–167 (2004)
9. Davies, R.O., Modi, J.J.: A direct method for completing eigenproblem solutions on a parallel computer. Linear Algebra Appl. **77**, 61–74 (1986)
10. Demmel, J.W.: Applied Numerical Linear Algebra. SIAM, Philadelphia (1997)
11. Dhillon, I.S., Parlett, B.N.: Multiple representations to compute orthogonal eigenvectors of symmetric tridiagonal matrices. Linear Algebra Appl. **387**, 1–28 (2004)
12. Dongarra, J.J., Moler, C.B., Wilkinson, J.H.: Improving the accuracy of computed eigenvalues and eigenvectors. SIAM J. Numer. Anal. **20**, 23–45 (1983)
13. GMP: GNU Multiple Precision Arithmetic Library, Code and documentation. http://gmplib.org/ (2015)
14. Golub, G.H., Van Loan, C.F.: Matrix Computations, 4th edn. The Johns Hopkins University Press, Baltimore (2013)
15. Gu, M., Eisenstat, S.C.: A divide-and-conquer algorithm for the symmetric tridiagonal eigenproblem. SIAM J. Matrix Anal. Appl. **16**, 172–191 (1995)
16. Hida, Y., Li, X.S., Bailey, D.H.: Algorithms for quad-double precision floating point arithmetic. In: Proceedings of the 15th IEEE Symposium on Computer Arithmetic, pp. 155–162. IEEE Computer Society Press (2001)
17. Higham, N.J.: Accuracy and Stability of Numerical Algorithms, 2nd edn. SIAM, Philadelphia (2002)
18. Higham, N.J.: Functions of Matrices: Theory and Computation. SIAM, Philadelphia (2008)
19. Jahn, H.A.: Improvement of an approximate set of latent roots and modal columns of a matrix by methods akin to those of classical perturbation theory. Quart. J. Mech. Appl. Math. **1**, 131–144 (1948)
20. Li, X.S., Demmel, J.W., Bailey, D.H., Henry, G., Hida, Y., Iskandar, J., Kahan, W., Kang, S.Y., Kapur, A., Martin, M.C., Thompson, B.J., Tung, T., Yoo, D.: Design, implementation and testing of extended and mixed precision BLAS. ACM Trans. Math. Softw. **28**, 152–205 (2002)
21. MPFR: The GNU MPFR Library, Code and documentation. http://www.mpfr.org/ (2013)
22. Ogita, T., Rump, S.M., Oishi, S.: Accurate sum and dot product. SIAM J. Sci. Comput. **26**, 1955–1988 (2005)
23. Oishi, S.: Fast enclosure of matrix eigenvalues and singular values via rounding mode controlled computation. Linear Algebra Appl. **324**, 133–146 (2001)
24. Ozaki, K., Ogita, T., Oishi, S., Rump, S.M.: Error-free transformations of matrix multiplication by using fast routines of matrix multiplication and its applications. Numer. Algorithms **59**, 95–118 (2012)
25. Parlett, B.N.: The symmetric eigenvalue problem, vol. 20, 2nd edn. Classics in Applied Mathematics. SIAM, Philadelphia (1998)
26. Priest, D.M.: Algorithms for arbitrary precision floating point arithmetic. In: Proceedings of the 10th Symposium on Computer Arithmetic, pp. 132–145. IEEE Computer Society Press (1991)
27. Rump, S.M., Ogita, T., Oishi, S.: Accurate floating-point summation part I: faithful rounding. SIAM J. Sci. Comput. **31**, 189–224 (2008)
28. Rump, S.M., Ogita, T., Oishi, S.: Accurate floating-point summation part II: sign, K-fold faithful and rounding to nearest. SIAM J. Sci. Comput. **31**, 1269–1302 (2008)

29. Tisseur, F.: Newton's method in floating point arithmetic and iterative refinement of generalized eigenvalue problems. SIAM J. Matrix Anal. Appl. **22**, 1038–1057 (2001)
30. Wilkinson, J.H.: The Algebraic Eigenvalue Problem. Clarendon Press, Oxford (1965)
31. Yamamoto, S., Fujiwara, T., Hatsugai, Y.: Electronic structure of charge and spin stripe order in $La_{2-x}\,Sr_x\,Ni\,O_4$ ($x = \frac{1}{3}, \frac{1}{2}$). Phys. Rev. B **76**, 165114 (2007)
32. Yamamoto, S., Sogabe, T., Hoshi, T., Zhang, S.-L., Fujiwara, T.: Shifted conjugate-orthogonal-conjugate-gradient method and its application to double orbital extended Hubbard model. J. Phys. Soc. Jpn. **77**, 114713 (2008)