BIT Numer Math (2013) 53:941–958
DOI 10.1007/s10543-013-0438-0

BIT

On the convergence of Ritz pairs and refined Ritz vectors for quadratic eigenvalue problems

Tsung-Ming Huang · Zhongxiao Jia · Wen-Wei Lin

Received: 29 February 2012 / Accepted: 1 July 2013 / Published online: 11 July 2013

Abstract For a given subspace, the Rayleigh-Ritz method projects the large quadratic eigenvalue problem (QEP) onto it and produces a small sized dense QEP. Similar to the Rayleigh-Ritz method for the linear eigenvalue problem, the Rayleigh-Ritz method defines the Ritz values and the Ritz vectors of the QEP with respect to the projection subspace. We analyze the convergence of the method when the angle between the subspace and the desired eigenvector converges to zero. We prove that there is a Ritz value that converges to the desired eigenvalue unconditionally but the Ritz vector converges conditionally and may fail to converge. To remedy the drawback of possible non-convergence of the Ritz vector, we propose a refined Ritz vector that is mathematically different from the Ritz vector and is proved to converge unconditionally. We construct examples to illustrate our theory.

Communicated by Michiel Hochstenbach.

The first and third authors were supported in part by the National Science Council, the National Center for Theoretical Sciences, the Center of Mathematical Modeling and Scientific Computing, and the Chiao-Da ST Yau Center in Taiwan, and the second author was supported in part by National Basic Research Program of China 2011CB302400 and National Science Foundation of China (No. 11071140).

T.-M. Huang
Department of Mathematics, National Taiwan Normal University, Taipei 116, Taiwan
e-mail: min@ntnu.edu.tw

Z. Jia (✉)
Department of Mathematical Sciences, Tsinghua University, Beijing 100084, China
e-mail: jiazx@tsinghua.edu.cn

W.-W. Lin
Department of Applied Mathematics, National Chiao Tung University, Hsinchu 300, Taiwan
e-mail: wwlin@math.nctu.edu.tw

Keywords Rayleigh-Ritz method · Ritz value · Ritz vector · Refined Ritz vector · Convergence

Mathematics Subject Classification (2010) 15A18 · 65F15 · 65F50

1 Introduction

Consider the numerical solution of the large quadratic eigenvalue problem (QEP)

$$\mathcal{Q}(\lambda)x \equiv \left(\lambda^2 M + \lambda D + K\right)x = 0, \tag{1.1}$$

where $\lambda \in \mathcal{C}$, $x \in \mathcal{C}^n \backslash \{0\}$, M, D and K are $n \times n$ complex matrices with $M = M^H > 0$ Hermitian positive definite. The scalar λ and the nonzero vector x in (1.1) are called an eigenvalue and a corresponding eigenvector of the quadratic pencil $\mathcal{Q}(\lambda)$ or (M, D, K), respectively. The pair (λ, x) is called an eigenpair of (M, D, K). Since $M = M^H > 0$ in (1.1), $\mathcal{Q}(\lambda)$ has $2n$ finite eigenvalues.

QEP (1.1) arises in a wide variety of scientific and engineering applications [2, 28]. The theoretical framework for general matrix polynomials and in particular for quadratic pencils can be found in books by Lancaster [18] and more recently by Gohberg, Lancaster and Rodman [4]. A good survey of mathematical properties, perturbation analysis, and a variety of numerical algorithms for QEPs can be found in the paper by Tisseur and Meerbergen [28].

In practice, a small number of eigenvalues that are nearest to a target τ or located in a prescribed region of the complex plane and the corresponding eigenvectors are often of interest. To this end, we exploit the shift transformation $\lambda_\tau = \lambda - \tau$ with $\det(\mathcal{Q}(\tau)) \neq 0$ to transform (1.1) to a new QEP of the form

$$\mathcal{Q}_\tau(\lambda_\tau)x \equiv \left(\lambda_\tau^2 M_\tau + \lambda_\tau D_\tau + K_\tau\right)x = 0, \tag{1.2}$$

where $M_\tau = M$, $D_\tau = 2\tau M + D$ and $K_\tau = \tau^2 M + \tau D + K$ is nonsingular. So, without loss of generality, throughout the paper, we assume that the eigenvalues to be sought are nonzero.

One kind of classical methods for solving QEP (1.1) is to reformulate it as a certain standard (or generalized) eigenvalue problem via a so-called linearization process and then to apply Krylov subspace based methods or Jacobi-Davidson type methods to solve the corresponding linear eigenvalue problem. Most of these methods fall into the category of the Rayleigh-Ritz method that is widely used for the computation of partial eigenpairs of a standard linear eigenvalue problem from a given projection subspace. As is well known, under the assumption that the angle between a desired eigenvector and the projection subspace tends to zero, there exists a Ritz value that converges to the desired eigenvalue unconditionally but its corresponding Ritz vector may fail to converge; furthermore, when one is concerned with eigenvectors, one can compute certain refined Ritz vectors whose convergence is guaranteed [10, 12, 13, 15, 16]; see also [25].

Over the years, some reliable numerical methods have been proposed that are used to solve large and sparse QEPs directly. Based on certain orthogonal projection conditions, various methods are designed to construct suitable lower dimensional subspaces. Then, the large QEP is projected onto a given subspace to produce a small sized dense QEP which can be solved by the standard QR or QZ algorithm. They fall into the category of the Rayleigh-Ritz method, as will be described in the next paragraph. Methods of this type include the residual inverse iteration method [8, 21, 22], the Jacobi-Davidson method [23, 24], Krylov subspace type methods [6, 19], the nonlinear Arnoldi method [29], second-order Arnoldi (SOAR) type methods [1, 17, 20, 30], the iterated shift-and-invert Arnoldi method [31] and the semiorthogonal generalized Arnoldi (SGA) method [7].

Now we describe the Rayleigh-Ritz method for the QEP. For a given orthonormal matrix $Q \in \mathcal{C}^{n\times m}$ $(m \le n)$, the Rayleigh-Ritz method is to find a scalar $\mu \in \mathcal{C}$ and a unit length vector $\hat{x} \in \mathcal{C}^m$ satisfying the orthogonal projection condition

$$\big(\mu^2 MQ + \mu DQ + KQ\big)\hat{x} \perp \text{span}\{Q\},$$

which amounts to solving the projected QEP

$$\big(\mu^2 \widehat{M} + \mu \widehat{D} + \widehat{K}\big)\hat{x} = 0, \tag{1.3}$$

where

$$\widehat{M} = Q^H M Q, \quad \widehat{D} = Q^H D Q, \quad \widehat{K} = Q^H K Q. \tag{1.4}$$

If $(\mu, \hat{x})$ with $\|\hat{x}\| = 1$ is an eigenpair of $(\widehat{M}, \widehat{D}, \widehat{K})$, i.e., $(\mu^2\widehat{M} + \mu\widehat{D} + \widehat{K})\hat{x} = 0$, then μ and $Q\hat{x}$ are, respectively, called a Ritz value and a corresponding Ritz vector of (M, D, K) with respect to span$\{Q\}$, and $(\mu, Q\hat{x})$ is a Ritz pair of (M, D, K). Since M is Hermitian positive definite, so is $\widehat{M}$ for any given Q. Therefore, we have $2m$ finite Ritz values.

For a *given* Q, the assumption that M is Hermitian positive definite is a *sufficient* condition to ensure the finiteness of both the eigenvalues and the Ritz values. Without this assumption, $\widehat{M}$ would possibly be *singular* for some given orthonormal Q. In this case, there could be some *infinite* Ritz values, the situation would become much more complicated, and the Rayleigh–Ritz method may fail to work. Indeed, as will be seen, some of our important convergence conclusions cannot be drawn, e.g., the bound in Theorem 2.1 may not tend to zero when the subspace span$\{Q\}$ is sufficiently good. In contrast, as will be clear, QEP (1.1) is mathematically equivalent to some standard linear eigenvalue problem provided that M is nonsingular; see (2.1a)–(2.1c). It is well known that the standard Rayleigh–Ritz method for the *linear* eigenvalue problem *always* computes *finite* Ritz values for *any* projection subspace. Therefore, there are some essential differences between the Rayleigh–Ritz method for (1.1) and the method for the linear eigenvalue problem. As is expected, it is nontrivial to establish a convergence theory of the Rayleigh–Ritz method for (1.1). As a key step of our further discussions, we first assume the finiteness of Ritz values for *any* projection subspace span$\{Q\}$. It is simple to justify that for *any* orthonormal Q the Hermitian positive definiteness of M is sufficient to ensure that of $\widehat{M}$. Generally, what we need

in the paper is to assume that $\|\widehat{M}^{-1}\|$ is uniformly bounded independently of Q. This assumption is true if M is Hermitian positive definite, as $\|\widehat{M}^{-1}\| \le \|M^{-1}\|$ for any orthonormal Q. So, purely for simplicity of presentation, we assume that M is Hermitian positive definite throughout the paper. Nevertheless, we must keep it in mind that all the convergence results and claims are true in this paper provided that $\widehat{M}$ is nonsingular and $\|\widehat{M}^{-1}\|$ is bounded.

In this paper we study the convergence of the Ritz value and the corresponding Ritz vector, and extend some of the results in [15, 16, 25] to the Rayleigh-Ritz method for (1.1). Although a number of Rayleigh-Ritz procedures with respect to different subspaces have been used, to our best knowledge, there has been no unified convergence result and general theory. As will be seen later, carrying out this task is indeed nontrivial and complicated. We establish some important results similar to those for the linear eigenvalue problem. It turns out that there exists a Ritz value that converges to the desired eigenvalue unconditionally but the corresponding Ritz vector may fail to converge even if the corresponding projection subspace span$\{Q\}$ contains a sufficiently accurate approximation to the desired eigenvector. It is thus necessary and significant to replace the Ritz vector by a refined Ritz vector that has residual minimization and is mathematically different from the Ritz vector. We prove that the refined Ritz vector converges unconditionally provided that the angles between the desired eigenvector and the subspaces tend to zero. All convergence results are nontrivial generalizations of the known results on the Rayleigh-Ritz method and the refined Rayleigh–Ritz method for the linear eigenvalue problem in [15, 16, 25].

This paper is organized as follows. In Sect. 2, we analyze the convergence for Ritz values and Ritz vectors and prove that the Ritz value is unconditionally convergent but the associated Ritz vector may fail to converge. To remedy this drawback, in Sect. 3, we introduce a refined Ritz vector and prove its unconditional convergence. Finally, we conclude the paper in Sect. 4.

Throughout this paper, the superscripts H and T denote the conjugate transpose and the transpose of a matrix or vector, respectively. I_n is the identity matrix of order n. We denote by $\|\cdot\|$ both Euclidean vector norm and the spectral matrix norm.

2 Convergence of Ritz values and Ritz vectors

Throughout the paper, let (λ_1, x_1) with $\|x_1\| = 1$ be a desired eigenpair of (M, D, K) and assume that λ_1 is simple. Furthermore, we keep in mind the assumption made in the introduction that $\lambda_1 \neq 0$, which is without loss of generality due to the equivalence of (1.1) and (1.2).

We convert QEP (1.1) to a generalized eigenvalue problem (GEP) of the form

$$A\begin{bmatrix}\lambda x\\ x\end{bmatrix} = \lambda B\begin{bmatrix}\lambda x\\ x\end{bmatrix}, \tag{2.1a}$$

or a standard linear eigenvalue problem (LEP) of the form

$$B^{-1}A\begin{bmatrix}\lambda x\\ x\end{bmatrix} = \lambda\begin{bmatrix}\lambda x\\ x\end{bmatrix}, \tag{2.1b}$$

where

$$A = \begin{bmatrix} -D & -K \\ I_n & 0 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} M & 0 \\ 0 & I_n \end{bmatrix}. \tag{2.1c}$$

So λ_1 is an eigenvalue of the matrix pencil (A, B) or the matrix $B^{-1}A$ in (2.1a)–(2.1c) and $v_1 \equiv \begin{bmatrix} \lambda_1 x_1 \\ x_1 \end{bmatrix}/\sqrt{1+|\lambda_1|^2}$ is its corresponding normalized eigenvector. There are numerous linearizations of QEP (1.1). We use (2.1a)–(2.1c) for two reasons. The first is that it is a very commonly used linearization in the literature. The second is that we establish our results in this paper by relating the QEP to such linearization. Other linearizations are certainly possible and useable, but if then we may have to make a very different and more complicated analysis in order to establish the convergence theory of the Rayleigh-Ritz method and refined Ritz vectors for the QEP.

There are unitary matrices $[v_1,\ X]$ and $[y_1,\ Y] \in \mathcal{C}^{2n\times 2n}$ with $v_1, y_1 \in \mathcal{C}^{2n}$ such that

$$\begin{bmatrix} y_1^H \\ Y^H \end{bmatrix} A \begin{bmatrix} v_1 & X \end{bmatrix} = \begin{bmatrix} \alpha & s^H \\ 0 & L \end{bmatrix}, \qquad \begin{bmatrix} y_1^H \\ Y^H \end{bmatrix} B \begin{bmatrix} v_1 & X \end{bmatrix} = \begin{bmatrix} \beta & t^H \\ 0 & N \end{bmatrix}, \tag{2.2}$$

where $L, N \in \mathcal{C}^{(2n-1)\times(2n-1)}$ and $\lambda_1 = \alpha\beta^{-1}$. Since λ_1 is supposed to be simple, it is not an eigenvalue of (L, N).

For a given orthonormal matrix $Q \in \mathcal{C}^{n\times m}$ with $m \le n$, define

$$W = \begin{bmatrix} Q & 0 \\ 0 & Q \end{bmatrix} \tag{2.3}$$

and let $[Q, Q^\perp]$ be unitary with $Q^\perp \in \mathcal{C}^{n\times(n-m)}$. From now on, throughout the paper, let θ_1 be the acute angle between x_1 and the projection subspace $\mathrm{span}\{Q\}$ and

$$q_1 = Q^H x_1, \quad q_1^\perp = \left(Q^\perp\right)^H x_1. \tag{2.4}$$

Then it holds that [25, p. 249, Theorem 2.2]

$$\left\|q_1^\perp\right\| = \sin\theta_1, \quad \|q_1\| = \sqrt{1-\sin^2\theta_1} = \cos\theta_1. \tag{2.5}$$

First of all, we want to show that there is a Ritz value μ_1 that converges to λ_1 unconditionally when $\sin\theta_1 \to 0$. The following perturbation result is needed, which is expressed in terms of the a priori uncomputable $\tan\theta_1$ and is different from Theorem 1 in [27], which is a backward perturbation result in terms of the a posteriori computable residual norm of an approximate eigenpair.

Lemma 2.1 *With λ_1, q_1 and θ_1 defined as above. Let $\widehat{M}, \widehat{D}$ and $\widehat{K}$ be defined in (1.4) and $\widehat{q_1} = q_1/\|q_1\|$. Then there are perturbation matrices $\mathcal{E}_{\widehat{M}}, \mathcal{E}_{\widehat{D}}, \mathcal{E}_{\widehat{K}} \in \mathcal{C}^{m\times m}$ with*

$$\|\mathcal{E}_{\widehat{M}}\| \le \frac{1}{3}\left(m_0 + \frac{1}{|\lambda_1|}d_0 + \frac{1}{|\lambda_1|^2}k_0\right)\tan\theta_1, \tag{2.6a}$$

$$\|\mathcal{E}_{\widehat{D}}\| \le \frac{1}{3}\left(|\lambda_1|m_0 + d_0 + \frac{1}{|\lambda_1|}k_0\right)\tan\theta_1, \tag{2.6b}$$

$$\|\mathcal{E}_{\widehat{K}}\| \le \frac{1}{3}\left(|\lambda_1|^2 m_0 + |\lambda_1|d_0 + k_0\right)\tan\theta_1, \tag{2.6c}$$

such that $(\lambda_1, \widehat{q}_1)$ *is an exact eigenpair of the perturbed* $(\widehat{M}+\mathcal{E}_{\widehat{M}}, \widehat{D}+\mathcal{E}_{\widehat{D}}, \widehat{K}+\mathcal{E}_{\widehat{K}})$, *where*

$$m_0 = \|M\|, \qquad d_0 = \|D\|, \qquad k_0 = \|K\|. \tag{2.7}$$

Proof Recalling (2.4) and (2.5), since

$$0 = \left(\lambda_1^2 M + \lambda_1 D + K\right)x_1 = \left(\lambda_1^2 M + \lambda_1 D + K\right)\begin{bmatrix} Q & Q^\perp\end{bmatrix}\begin{bmatrix} Q^H \\ (Q^\perp)^H\end{bmatrix}x_1,$$

we obtain

$$\lambda_1^2 M Q q_1 + \lambda_1 D Q q_1 + K Q q_1 = -\left(\lambda_1^2 M + \lambda_1 D + K\right)Q^\perp q_1^\perp. \tag{2.8}$$

Pre-multiplying (2.8) by Q^H gives

$$r_1 \equiv \left(\lambda_1^2\widehat{M} + \lambda_1\widehat{D} + \widehat{K}\right)\widehat{q}_1 = -\left(\lambda_1^2 Q^H M + \lambda_1 Q^H D + Q^H K\right)Q^\perp \frac{q_1^\perp}{\|q_1\|}. \tag{2.9}$$

So, noting from (2.5) that $\tan\theta_1 = \frac{\sin\theta_1}{\cos\theta_1} = \frac{\|q_1^\perp\|}{\|q_1\|}$, we have

$$\|r_1\| \le (|\lambda_1|^2 m_0 + |\lambda_1|d_0 + k_0)\tan\theta_1.$$

Define

$$\mathcal{E}_{\widehat{M}} = -\frac{1}{3\lambda_1^2}r_1\widehat{q}_1^H, \qquad \mathcal{E}_{\widehat{D}} = -\frac{1}{3\lambda_1}r_1\widehat{q}_1^H, \qquad \mathcal{E}_{\widehat{K}} = -\frac{1}{3}r_1\widehat{q}_1^H.$$

By (2.9) it is easily seen that $\|\mathcal{E}_{\widehat{M}}\|$, $\|\mathcal{E}_{\widehat{D}}\|$ and $\|\mathcal{E}_{\widehat{K}}\|$ satisfy (2.6a)–(2.6c) and

$$\left[\,\lambda_1^2(\widehat{M}+\mathcal{E}_{\widehat{M}}) + \lambda_1(\widehat{D}+\mathcal{E}_{\widehat{D}}) + (\widehat{K}+\mathcal{E}_{\widehat{K}})\right]\widehat{q}_1 = 0,$$

which completes the proof. □

We may deduce from this lemma that there exists an eigenvalue μ_1 of $(\widehat{M}, \widehat{D}, \widehat{K})$ that converges to λ_1 as $\theta_1 \to 0$. However, things are subtle and by no means trivial here. The difficulty is that, unlike a usual matrix perturbation problem where matrices are *given* and *fixed* and perturbations are allowed to *change*, here the matrix triple $(\widehat{M}, \widehat{D}, \widehat{K})$ and the perturbation triple $(\mathcal{E}_{\widehat{M}}, \mathcal{E}_{\widehat{D}}, \mathcal{E}_{\widehat{K}})$ *change simultaneously* as $\theta_1 \to 0$. This means that there may be a possibility that, as θ_1 changes, the eigenvalue λ_1 of $(\widehat{M}+\mathcal{E}_{\widehat{M}}, \widehat{D}+\mathcal{E}_{\widehat{D}}, \widehat{K}+\mathcal{E}_{\widehat{K}})$ and the eigenvalues of $(\widehat{M}, \widehat{D}, \widehat{K})$ become ill conditioned so swiftly that no eigenvalue of $(\widehat{M}, \widehat{D}, \widehat{K})$ converges to λ_1 though $\theta_1 \to 0$.

Fortunately, by exploiting a theorem of Elsner [3] (also see [26, p. 168]) we can prove that this cannot happen and there is indeed an eigenvalue μ_1 that converges to the desired λ_1 provided that $\theta_1 \to 0$. Elsner's theorem states that, given matrices C and $\tilde{C}$ of order n, for any eigenvalue λ of C there is an eigenvalue $\tilde{\lambda}$ of $\tilde{C}$ such that

$$|\lambda - \tilde{\lambda}| \le \left(\|C\| + \|\tilde{C}\|\right)^{1-\frac{1}{n}} \|C - \tilde{C}\|^{\frac{1}{n}}.$$

For our purpose, define the matrices $\widehat{A}$ and $\widehat{B}$ by

$$\widehat{A} = \begin{bmatrix} -\widehat{D} & -\widehat{K} \\ I_m & 0 \end{bmatrix}, \qquad \widehat{B} = \begin{bmatrix} \widehat{M} & 0 \\ 0 & I_m \end{bmatrix}. \tag{2.10}$$

Then the eigenvalues μ of $(\widehat{M}, \widehat{D}, \widehat{K})$ are equal to those of $(\widehat{A}, \widehat{B})$, whose normalized eigenvectors $\hat{v} \equiv \left[\begin{smallmatrix} \mu\hat{x} \\ \hat{x} \end{smallmatrix}\right]/\sqrt{1+|\mu|^2}$ with $\hat{x}$ the eigenvectors associated with the eigenvalues μ of $(\widehat{M}, \widehat{D}, \widehat{K})$. Since $\widehat{M}$ is Hermitian positive definite, so is $\widehat{B}$. Therefore, all the μ are the eigenvalues of $\widehat{B}^{-1}\widehat{A}$. Furthermore, it holds that $\|\widehat{B}^{-1}\| \le \|B^{-1}\|$ for any given orthonormal Q and Hermitian positive definite M.

From Lemma 2.1, λ_1 is an eigenvalue of $(\widehat{A} + \mathcal{E}_{\widehat{A}}, \widehat{B} + \mathcal{E}_{\widehat{B}})$ with the perturbation matrices

$$\mathcal{E}_{\widehat{A}} = \begin{bmatrix} -\mathcal{E}_{\widehat{D}} & -\mathcal{E}_{\widehat{K}} \\ 0 & 0 \end{bmatrix}, \qquad \mathcal{E}_{\widehat{B}} = \begin{bmatrix} \mathcal{E}_{\widehat{M}} & 0 \\ 0 & 0 \end{bmatrix},$$

i.e., an eigenvalue of $(\widehat{B} + \mathcal{E}_{\widehat{B}})^{-1}(\widehat{A} + \mathcal{E}_{\widehat{A}})$ if $(\widehat{B} + \mathcal{E}_{\widehat{B}})^{-1}$ exists. Since $\widehat{B}$ is Hermitian positive definite and its smallest singular value is bounded by that of B from below, $\widehat{B} + \mathcal{E}_{\widehat{B}}$ must be nonsingular for θ_1 small enough. Moreover, for $\theta_1 \to 0$, it follows from Lemma 2.1 that

$$Lb\|(\widehat{B} + \mathcal{E}_{\widehat{B}})^{-1}\| = \|\widehat{B}^{-1} + O(\mathcal{E}_{\widehat{B}})\| \to \|\widehat{B}^{-1}\| \le \|B^{-1}\| \tag{2.11}$$

is uniformly bounded independent of θ_1. Since $\|\widehat{A}\|$ is always bounded from above as $\|\widehat{D}\| \le \|D\|$ and $\|\widehat{K}\| \le \|K\|$, it follows that $\|\widehat{B}^{-1}\widehat{A}\| \le \|\widehat{B}^{-1}\|\|\widehat{A}\|$ is uniformly bounded independent of θ_1. As a result, for $\theta_1 \to 0$, since $\widehat{A} + \mathcal{E}_{\widehat{A}} \to \widehat{A}$, it follows from (2.11) and Theorem 2.1 that

$$\|(\widehat{B} + \mathcal{E}_{\widehat{B}})^{-1}(\widehat{A} + \mathcal{E}_{\widehat{A}})\| \le \|(\widehat{B} + \mathcal{E}_{\widehat{B}})^{-1}\|\|(\widehat{A} + \mathcal{E}_{\widehat{A}})\|$$

is uniformly bounded independently of θ_1.

Finally, from Theorem 2.1 and $(\widehat{B} + \mathcal{E}_{\widehat{B}})^{-1} = \widehat{B}^{-1} + O(\mathcal{E}_{\widehat{B}})$, it is easily justified that

$$\|\widehat{B}^{-1}\widehat{A} - (\widehat{B} + \mathcal{E}_{\widehat{B}})^{-1}(\widehat{A} + \mathcal{E}_{\widehat{A}})\| = O(\sin\theta_1).$$

Based on Elsner's theorem, we have the following result, which, together with the above discussions, proves the global unconditional convergence of Ritz values when $\theta_1 \to 0$.

Theorem 2.1 *Assume that θ_1 is small enough to make $\widehat{B}+\mathcal{E}_{\widehat{B}}$ nonsingular. There is a Ritz value μ_1 such that*

$$|\mu_1-\lambda_1| \le \left(\left\|\widehat{B}^{-1}\widehat{A}\right\| + \left\|(\widehat{B}+\mathcal{E}_{\widehat{B}})^{-1}(\widehat{A}+\mathcal{E}_{\widehat{A}})\right\|\right)^{1-\frac{1}{2m}} \times \left\|\widehat{B}^{-1}\widehat{A}-(\widehat{B}+\mathcal{E}_{\widehat{B}})^{-1}(\widehat{A}+\mathcal{E}_{\widehat{A}})\right\|^{\frac{1}{2m}}. \tag{2.12}$$

The theorem indicates that as $\theta_1 \to 0$ there is always a Ritz value $\mu_1 \to \lambda_1$ unconditionally. We should comment that bound (2.12) will in general be a too pessimistic overestimate and be for the worst case. If, as usually happens in practice, the condition number of λ_1 as an eigenvalue of $(\widehat{B}+\mathcal{E}_{\widehat{B}})^{-1}(\widehat{A}+\mathcal{E}_{\widehat{A}})$ is bounded, the convergence will be linear in θ_1, much better than that predicted by bound (2.12).

Next, we analyze the convergence of the corresponding Ritz vector $\tilde{x}_1$. Based on decomposition (2.2), we can establish the following result, which is an analogue of Theorem 3.1 in [9] for the standard linear eigenvalue problem. The result will be used when we prove the unconditional convergence of refined Ritz vectors to be introduced in the next section.

Lemma 2.2 *Let $(\mu_1, \tilde{v}_1)$ with $\|\tilde{v}_1\| = 1$ be an approximation to (λ_1, v_1) of the matrix pair (A, B) with $\|v_1\| = 1$. Let*

$$r = A\tilde{v}_1 - \mu_1 B\tilde{v}_1 \tag{2.13}$$

be the residual of $(\mu_1, \tilde{v}_1)$, and define $\mathrm{sep}(\mu_1, (L, N)) := \|(L-\mu_1 N)^{-1}\|^{-1}$. Then

$$\sin\angle(v_1, \tilde{v}_1) \le \frac{\|r\|}{\mathrm{sep}(\mu_1, (L, N))}. \tag{2.14}$$

Proof From (2.2), pre-multiplying (2.13) by Y^H leads to

$$\begin{aligned} Y^H r &= Y^H\left(\alpha y_1 v_1^H + y_1 s^H X^H + YLX^H\right)\tilde{v}_1 \\ &\quad - \mu_1 Y^H\left(\beta y_1 v_1^H + y_1 t^H X^H + YNX^H\right)\tilde{v}_1 \\ &= (L-\mu_1 N)X^H\tilde{v}_1. \end{aligned}$$

Therefore, it follows from $\|X^H\tilde{v}_1\| = \sin\angle(v_1, \tilde{v}_1)$ that (2.14) holds. □

In terms of the a posteriori computable residual r, Theorem 2.2 establishes the relationship between the eigenvector v_1 and its approximation $\tilde{v}_1$ for the generalized eigenvalue problem (2.1a)–(2.1c).

Let $(\mu_1, \widetilde{x}_1)$ be the Ritz pair approximating the desired the desired eigenpair (λ_1, x_1) of (M, D, K), where $\widetilde{x}_1 = Q\hat{x}_1$ and $(\mu_1, \hat{x}_1)$ with $\|\hat{x}_1\| = 1$ is the eigenpair of $(\widehat{M}, \widehat{D}, \widehat{K})$. In terms of θ_1, we attempt to derive one of our main results, an a priori bound for the Ritz vector $\hat{x}_1$ as an approximation to the eigenvector x_1. Note that μ_1 is an eigenvalue of $(\widehat{A}, \widehat{B})$ and $\hat{v}_1 \equiv \left[\begin{smallmatrix}\mu_1\hat{x}_1\\ \hat{x}_1\end{smallmatrix}\right]/\sqrt{1+|\mu_1|^2}$ is its corresponding normalized eigenvector. Similar to (2.2), there are unitary matrices $[\hat{v}_1, \ \widehat{X}]$ and

$[\hat{y}_1,\ \widehat{Y}] \in \mathcal{C}^{2m\times 2m}$ with $\widehat{v}_1, \widehat{y}_1 \in \mathcal{C}^{2m}$ such that

$$\begin{bmatrix} \hat{y}_1^H \\ \widehat{Y}^H \end{bmatrix} \widehat{A} \begin{bmatrix} \hat{v}_1 & \widehat{X} \end{bmatrix} = \begin{bmatrix} \hat{\alpha} & \hat{s}^H \\ 0 & \widehat{L} \end{bmatrix}, \qquad \begin{bmatrix} \hat{y}_1^H \\ \widehat{Y}^H \end{bmatrix} \widehat{B} \begin{bmatrix} \hat{v}_1 & \widehat{X} \end{bmatrix} = \begin{bmatrix} \hat{\beta} & \hat{t}^H \\ 0 & \widehat{N} \end{bmatrix}, \tag{2.15}$$

where $\widehat{L},\ \widehat{N} \in \mathcal{C}^{(2m-1)\times(2m-1)}$ and $\mu_1 = \hat{\alpha}\hat{\beta}^{-1}$. Under the only hypothesis that $\sin\theta_1 \to 0$, it is possible that there is an eigenvalue of $(\widehat{L}, \widehat{N})$ that could be arbitrarily near or even equal to μ_1. For a multiple and derogatory μ_1, that is, μ_1 has more than one trivial or nontrivial Jordan blocks, there are more than one $\tilde{x}_1 = Q\hat{x}_1$ to approximate the unique eigenvector x_1 of (M, D, K). If μ_1 is near an eigenvalue of $(\widehat{L}, \widehat{N})$, we will get a unique $\tilde{x}_1$, but there is no guarantee that it converges to x_1. It leads us to postulate that $\tilde{x}_1$ will converge provided that $\mathrm{sep}(\lambda_1, (\widehat{L}, \widehat{N}))$ is uniformly away from zero independent of θ_1, i.e., $\mathrm{sep}(\lambda_1, (\widehat{L}, \widehat{N})) > c$ with c a positive constant independent of θ_1. We will, quantitatively, show that it is indeed the case. Before proceeding, we need the following lemma.

Lemma 2.3 *Let* $u = \begin{bmatrix} u_2 \\ u_1 \end{bmatrix}$ *and* $\tilde{u} = \begin{bmatrix} \tilde{u}_2 \\ \tilde{u}_1 \end{bmatrix}$ *where* $u_i, \tilde{u}_i \in \mathcal{C}^n$ *for* $i = 1, 2$ *and* $\|u_1\| = \|\tilde{u}_1\| = 1$. *Then*

$$\sin\angle(u_1, \tilde{u}_1) \le \min\{\|u\|, \|\tilde{u}\|\} \sin\angle(u, \tilde{u}).$$

Proof Since $\|u_1\| = 1$, from the definition of $\sin\angle(u, \tilde{u})$, we have

$$\begin{aligned} \sin^2\angle(u, \tilde{u}) &= \min_{\alpha} \left\| \frac{u}{\|u\|} - \alpha\tilde{u} \right\|^2 \\ &= \min_{\alpha} \left(\left\| \frac{u_1}{\|u\|} - \alpha\tilde{u}_1 \right\|^2 + \left\| \frac{u_2}{\|u\|} - \alpha\tilde{u}_2 \right\|^2 \right) \\ &\ge \min_{\alpha} \left\| \frac{u_1}{\|u\|} - \alpha\tilde{u}_1 \right\|^2 \\ &= \frac{1}{\|u\|^2} \min_{\alpha} \|u_1 - \alpha\tilde{u}_1\|^2 \\ &= \frac{1}{\|u\|^2} \sin^2\angle(u_1, \tilde{u}_1). \end{aligned}$$

In the same way, we can also prove that

$$\sin\angle(u_1, \tilde{u}_1) \le \|\tilde{u}\| \sin\angle(u, \tilde{u}).$$

Therefore, the assertion holds. □

Theorem 2.2 *Let* $(\widehat{A}, \widehat{B})$ *be defined in* (2.10) *and it have decomposition* (2.15). *Suppose that the Ritz pair* $(\mu_1, \tilde{x}_1)$ *is used to approximate the desired eigenpair* (λ_1, x_1) *with* $\|\tilde{x}_1\| = \|x_1\| = 1$. *If* $\mathrm{sep}(\lambda_1, (\widehat{L}, \widehat{N})) > 0$, *then*

$$\sin\angle(x_1, \tilde{x}_1) \le \sin\theta_1 + \frac{|\lambda_1|^2 m_0 + |\lambda_1| d_0 + k_0}{\mathrm{sep}(\lambda_1, (\widehat{L}, \widehat{N}))} \tan\theta_1, \tag{2.16}$$

where m_0, d_0 *and* k_0 *are defined in* (2.7).

Proof By the triangle inequality we have

$$\angle(x_1, \tilde{x}_1) \le \angle\big(x_1, QQ^H x_1\big) + \angle\big(QQ^H x_1, \tilde{x}_1\big). \tag{2.17}$$

From (2.4) and (2.5), we have

$$\cos\angle\big(x_1, QQ^H x_1\big) = \frac{|x_1^H QQ^H x_1|}{\|QQ^H x_1\|} = \big\|Q^H x_1\big\| = \cos\theta_1. \tag{2.18}$$

Let $\hat{q}_1 = \frac{Q^H x_1}{\|Q^H x_1\|}$. From (2.17) and (2.18) we get

$$\begin{aligned}\sin\angle(x_1, \tilde{x}_1) &\le \sin\theta_1 + \sin\angle\big(QQ^H x_1, \tilde{x}_1\big) = \sin\theta_1 + \sin\angle(Q\hat{q}_1, Q\hat{x}_1)\\ &= \sin\theta_1 + \sin\angle(\hat{x}_1, \hat{q}_1).\end{aligned} \tag{2.19}$$

From (2.10), it is easily seen that $\big(\mu_1, \hat{v}_1 \equiv \left[\begin{smallmatrix}\mu_1\hat{x}_1\\ \hat{x}_1\end{smallmatrix}\right]\big)$ is an eigenpair of $(\widehat{A}, \widehat{B})$. So we can regard $\big(\lambda_1, \hat{q} \equiv \left[\begin{smallmatrix}\lambda_1\hat{q}_1\\ \hat{q}_1\end{smallmatrix}\right]\big)$ as an approximation of $(\mu_1, \hat{v}_1)$. Then the residual of $(\lambda_1, \hat{q})$ as an approximate eigenpair of $(\widehat{A}, \widehat{B})$ is

$$\begin{aligned}\hat{r} &= \begin{bmatrix}-\widehat{D} & -\widehat{K}\\ I_m & 0\end{bmatrix}\begin{bmatrix}\lambda_1\hat{q}_1\\ \hat{q}_1\end{bmatrix} - \lambda_1\begin{bmatrix}\widehat{M} & 0\\ 0 & I_m\end{bmatrix}\begin{bmatrix}\lambda_1\hat{q}_1\\ \hat{q}_1\end{bmatrix}\\ &= \begin{bmatrix}-(\lambda_1^2\widehat{M} + \lambda_1\widehat{D} + \widehat{K})\hat{q}_1\\ 0\end{bmatrix} \equiv \begin{bmatrix}-\hat{r}_1\\ 0\end{bmatrix}.\end{aligned}$$

By (2.9) in the proof of Theorem 2.1 we have

$$\frac{\|\hat{r}\|}{\|\hat{q}\|} = \frac{\|\hat{r}_1\|}{\|\hat{q}\|} \le \frac{|\lambda_1|^2 m_0 + |\lambda_1| d_0 + k_0}{\|\hat{q}\|}\tan\theta_1. \tag{2.20}$$

From Lemma 2.3, Theorem 2.2 and (2.20), inequality (2.19) satisfies

$$\begin{aligned}\sin\angle(x_1, \tilde{x}_1) &\le \sin\theta_1 + \sin\angle(\hat{x}_1, \hat{q}_1)\\ &\le \sin\theta_1 + \|\hat{q}\|\sin\angle(\hat{v}_1, \hat{q})\\ &\le \sin\theta_1 + \|\hat{q}\|\frac{\|\hat{r}\|/\|\hat{q}\|}{\operatorname{sep}(\lambda_1, (\widehat{L}, \widehat{N}))}\\ &\le \sin\theta_1 + \frac{|\lambda_1|^2 m_0 + |\lambda_1| d_0 + k_0}{\operatorname{sep}(\lambda_1, (\widehat{L}, \widehat{N}))}\tan\theta_1.\end{aligned}$$

□

From Theorem 2.2 we see that $\operatorname{sep}(\lambda_1, (\widehat{L}, \widehat{N})) > 0$ uniformly is a sufficient condition for the convergence of the Ritz vector $\tilde{x}_1$. Furthermore, from Lemma 2.1, since the Ritz value μ_1 approaches the eigenvalue λ_1 as $\theta_1 \to 0$, by the continuity argument we have $\operatorname{sep}(\mu_1, (\widehat{L}, \widehat{N})) \to \operatorname{sep}(\lambda_1, (\widehat{L}, \widehat{N}))$. However, as we have argued above, $\operatorname{sep}(\mu_1, (\widehat{L}, \widehat{N}))$ can be arbitrarily small (and even be exactly zero) when μ_1 is arbitrarily near other eigenvalues (or is associated with a multiple eigenvalue) of $(\widehat{L}, \widehat{N})$. Consequently, while the Ritz value converges unconditionally once $\theta_1 \to 0$,

the corresponding Ritz vector may fail to converge or may converge very slowly or irregularly.

In the following, we give an example to illustrate that the Ritz vector fails to converge to the desired eigenvector.

Example 2.1 Consider QEP (1.1) with

$$M = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{bmatrix}, \qquad D = \begin{bmatrix} -5.5 & -5 & 0 \\ -5 & -11 & -3 \\ 0 & -3 & -4 \end{bmatrix}, \qquad K = \begin{bmatrix} 6 & 6 & 0 \\ 6 & 9 & 2 \\ 0 & 2 & 2 \end{bmatrix}.$$

It is easy to see that M and K are symmetric positive definite and $(1, [0, 0, 1]^T)$ is an eigenpair of the QEP.

Suppose that we have come up with an orthonormal basis

$$Q = \begin{bmatrix} 0 & \frac{8}{\sqrt{73}} \\ 0 & -\frac{3}{\sqrt{73}} \\ 1 & 0 \end{bmatrix}.$$

Then we have $\sin\theta_1 = 0$ exactly, and the projected matrices are

$$\widehat{M} = Q^H M Q = \begin{bmatrix} 2 & -\frac{3}{\sqrt{73}} \\ -\frac{3}{\sqrt{73}} & \frac{34}{73} \end{bmatrix},$$

$$\widehat{D} = Q^H D Q = \begin{bmatrix} -4 & \frac{9}{\sqrt{73}} \\ \frac{9}{\sqrt{73}} & -\frac{211}{73} \end{bmatrix},$$

$$\widehat{K} = Q^H K Q = \begin{bmatrix} 2 & -\frac{6}{\sqrt{73}} \\ -\frac{6}{\sqrt{73}} & \frac{177}{73} \end{bmatrix},$$

from which it follows that

$$\widehat{M} + \widehat{D} + \widehat{K} = 0.$$

Since $\widehat{M} + \widehat{D} + \widehat{K}$ is zero, any nonzero vector $\hat{x}_1$ with $\|\hat{x}_1\| = 1$ is an eigenvector of $(\widehat{M}, \widehat{D}, \widehat{K})$ corresponding to the double eigenvalue one, a Ritz value equal to the desired eigenvalue exactly. However, the Rayleigh-Ritz method itself cannot tell us how to pick up a suitable $\hat{x}_1$. In practice, we might well take $\hat{x}_1 = [1/\sqrt{2}, 1/\sqrt{2}]^T$ and then the approximate eigenvector becomes $[4\sqrt{2}/\sqrt{73},\ -3/\sqrt{146},\ 1/\sqrt{2}]^T$, which has no accuracy as an approximation of the desired eigenvector $[0, 0, 1]^T$ and is completely wrong. Thus the method can fail even though the projection subspace $\mathrm{span}\{Q\}$ contains the desired eigenvector exactly.

In practice, we would not expect $\mathrm{span}\{Q\}$ to contain x_1 exactly. Let us investigate the case that $\mathrm{span}\{Q\}$ contains an enough accurate approximation to x_1, i.e., $\sin\theta_1$ is

very small. We perturb Q by a matrix generated randomly in a normal distribution by $10^{-12} \times \mathsf{randn}(3, 2)$ whose 2-norm is 2.2×10^{-12}, and the resulting

$$\sin\theta_1 = 1.7 \times 10^{-12}.$$

The orthonormalized

$$Q := Q(Q^H Q)^{-1/2} = \begin{bmatrix} -0.000000000001074 & 0.936329177568703 \\ -0.000000000001425 & -0.351123441589302 \\ 1.000000000000000 & 0.000000000000506 \end{bmatrix}$$

and

$$\widehat{M} = \begin{bmatrix} 1.999999999997149 & -0.351123441589253 \\ -0.351123441589253 & 0.465753424656353 \end{bmatrix},$$

$$\widehat{D} = \begin{bmatrix} -3.999999999991449 & 1.053370324770698 \\ 1.053370324770698 & -2.890410958899234 \end{bmatrix},$$

$$\widehat{K} = \begin{bmatrix} 1.999999999997149 & -0.351123441589253 \\ -0.351123441589253 & 0.465753424656353 \end{bmatrix}.$$

We use the Matlab function polyeig.m to solve the projected QEP, and the computed $\mu_1 = 1.000000000009369$ and the associated eigenvector

$$\hat{x}_1 = [0.999982126253304, -0.005978894038382]^T.$$

So the Ritz vector

$$\tilde{x}_1 = Q\hat{x}_1 = [-0.005598212938803,\ 0.002099329850230,\ 0.999982126253300]^T$$

and

$$\sin\angle(x_1, \tilde{x}_1) \approx 0.005979,$$

at least nine orders bigger than $\sin\theta_1$! so $\tilde{x}_1$ is a very poor approximation to x_1 for the given accurate subspace span$\{Q\}$. It is also justified that the residual norm of the Ritz pair $(\mu_1, \tilde{x}_1)$ is

$$\left\| (\mu_1^2 M + \mu_1 D + K)\tilde{x}_1 \right\| \approx 0.011958.$$

The poor accuracy of $\tilde{x}_1$ is due to the fact that there is another Ritz value $\mu = 1.000000000010143$ that is very near to μ_1, so that sep$(\lambda_1, (\widehat{L}, \widehat{N}))$ in (2.16) is tiny.

3 Convergence of refined Ritz vectors

As we have seen in Sect. 2, the Ritz vector may fail to converge or converges very slowly. Since the Ritz value is known to converge to the simple eigenvalue λ_1 when $\sin\theta_1 \to 0$, this suggests us to deal with non-converging Ritz vector by retaining the Ritz value but replacing the Ritz vector with a unit length vector $\tilde{z}_1 \in$ span$\{Q\}$ with a

suitably small residual. Naturally, for a given Ritz value μ_1 we construct $\tilde{z}_1 = Q\hat{z}_1$, where the unit length $\hat{z}_1$ is required to be the optimal solution

$$\hat{z}_1 = \arg \min_{\|z\|=1} \left\| \left(\mu_1^2 M + \mu_1 D + K\right) Qz \right\|. \tag{3.1}$$

The vector $\tilde{z}_1 = Q\hat{z}_1$ is called a refined Ritz vector of (M, D, K) corresponding to μ_1 with respect to span$\{Q\}$. Obviously, $\hat{z}_1$ is the right singular vector of the $n \times m$ rectangular matrix $(\mu_1^2 M + \mu_1 D + K)Q$ associated with its smallest singular value. We can compute $\hat{z}_1$ reliably by a standard SVD algorithm or generally cheaper but still numerically stable cross-product based SVD algorithms; see [11, 17] and also [25]. For a detailed round-off error analysis on the latter ones, we refer to [14].

Before establishing the convergence of the refined Ritz vector $\tilde{z}_1$, we need two lemmas.

Lemma 3.1 *For W defined in (2.3), let (λ_1, x_1) with $\|x_1\| = 1$ be the desired eigenpair of (M, D, K) and $v_1 = \begin{bmatrix} \lambda_1 x_1 \\ x_1 \end{bmatrix} / \sqrt{1 + |\lambda_1|^2}$. Then it holds that*

$$\sin \angle\left(v_1, \text{span}\{W\}\right) = \sin \theta_1. \tag{3.2}$$

Proof By (2.3) and the definition of $\sin\theta_1$, we have

$$\begin{aligned}
&\sin^2 \angle\left(v_1, \text{span}\{W\}\right) \\
&= \frac{1}{1 + |\lambda_1|^2} \min_{u,v \in \text{span}\{Q\}} \left\| \begin{bmatrix} \lambda_1 x_1 \\ x_1 \end{bmatrix} - \begin{bmatrix} u \\ v \end{bmatrix} \right\|^2 \\
&= \frac{1}{1 + |\lambda_1|^2} \min_{u,v \in \text{span}\{Q\}} \left(\|\lambda_1 x_1 - u\|^2 + \|x_1 - v\|^2 \right) \\
&= \frac{|\lambda_1|^2}{1 + |\lambda_1|^2} \min_{u \in \text{span}\{Q\}} \|x_1 - u\|^2 + \frac{1}{1 + |\lambda_1|^2} \min_{v \in \text{span}\{Q\}} \|x_1 - v\|^2 \\
&= \frac{|\lambda_1|^2}{1 + |\lambda_1|^2} \sin^2 \theta_1 + \frac{1}{1 + |\lambda_1|^2} \sin^2 \theta_1 \\
&= \sin^2 \theta_1.
\end{aligned}$$

□

Lemma 3.2 *Let (A, B) be defined in (2.1c). It holds that*

$$\min_{\|z\|=1} \left\| (A - \mu_1 B) \begin{bmatrix} \mu_1 Qz \\ Qz \end{bmatrix} \right\| = \sqrt{1 + |\mu_1|^2} \min_{\|z\|=1} \left\| \left(\mu_1^2 M + \mu_1 D + K\right) Qz \right\| \tag{3.3}$$

and the minimum is attained at $\hat{z}_1$.

Proof Without the minimizations, for any m dimensional vector z, it is direct to verify that the two hand sides are equal. So the assertion holds. □

Theorem 3.1 *Let μ_1 be the Ritz value of (M, D, K) approximating the desired simple eigenvalue λ_1. Suppose* sep$(\mu_1, (L, N)) > 0$, *where L, N are defined in (2.2).*

Then we have

$$\sin\angle(x_1,\tilde{z}_1) < \frac{\sqrt{1+|\lambda_1|^2}(|\lambda_1-\mu_1|(\|B\|+\|A-\mu_1B\|)+\|A-\mu_1B\|\sin\theta_1)}{\cos\theta_1\,\mathrm{sep}(\mu_1,(L,N))}. \tag{3.4}$$

Proof Let $v_1=\left[\begin{smallmatrix}\lambda_1x_1\\ x_1\end{smallmatrix}\right]/\sqrt{1+|\lambda_1|^2}$. From Lemma 2.3, we have

$$\begin{aligned}\sin\angle(x_1,\tilde{z}_1) &\le \sqrt{1+|\mu_1|^2}\sin\angle\left(\begin{bmatrix}\lambda_1x_1\\ x_1\end{bmatrix},\begin{bmatrix}\mu_1Q\hat{z}_1\\ Q\hat{z}_1\end{bmatrix}\right)\\ &= \sqrt{1+|\mu_1|^2}\sin\angle(v_1,\tilde{z}),\end{aligned}$$

where $\tilde{z}=\left[\begin{smallmatrix}\tilde{z}_2\\ \tilde{z}_1\end{smallmatrix}\right]\equiv\left[\begin{smallmatrix}\mu_1Q\hat{z}_1\\ Q\hat{z}_1\end{smallmatrix}\right]/\sqrt{1+|\mu_1|^2}$. Let P_W be the orthogonal projector onto the subspace $\mathrm{span}\{W\}$, where $W=\mathrm{diag}(Q,Q)$. Then

$$P_Wv_1=\begin{bmatrix}\lambda_1QQ^Hx_1\\ QQ^Hx_1\end{bmatrix}.$$

Therefore, we get

$$\|Q^Hx_1\|^{-1}\left(P_Wv_1-\begin{bmatrix}(\lambda_1-\mu_1)QQ^Hx_1\\ 0\end{bmatrix}\right)=\begin{bmatrix}\mu_1Q\frac{Q^Hx_1}{\|Q^Hx_1\|}\\ Q\frac{Q^Hx_1}{\|Q^Hx_1\|}\end{bmatrix}:=\hat{v}_1,$$

which is an approximate eigenvector of the desired form in the left-hand side of (3.3) and $\frac{Q^Hx_1}{\|Q^Hx_1\|}$ is a minimizer candidate for (3.3). Define

$$f=(I_n-P_W)v_1+f_2$$

with

$$f_2=\begin{bmatrix}(\lambda_1-\mu_1)QQ^Hx_1\\ 0\end{bmatrix}.$$

Then from $\cos\theta_1=\|Q^Hx_1\|$ we have

$$\frac{\|f_2\|}{\cos\theta_1}\le|\lambda_1-\mu_1|.$$

From Lemma 3.1 we get $\|(I_n-P_W)v_1\|=\sqrt{1+|\lambda_1|^2}\sin\theta_1$. Therefore, we obtain

$$\begin{aligned}(A-\mu_1B)\hat{v}_1 &= \frac{(A-\mu_1B)(P_Wv_1-f_2)}{\cos\theta_1}\\ &= \frac{(A-\mu_1B)(v_1-f)}{\cos\theta_1}\\ &= \frac{(\lambda_1-\mu_1)Bv_1-(A-\mu_1B)((I_n-P_W)v_1+f_2)}{\cos\theta_1}.\end{aligned}$$

Taking the norms gives

$$\left\|(A-\mu_1 B)\hat{v}_1\right\| \le \frac{\sqrt{1+|\lambda_1|^2}(|\lambda_1-\mu_1|\|B\|+\|A-\mu_1 B\|\sin\theta_1)}{\cos\theta_1} + |\lambda_1-\mu_1|\|A-\mu_1 B\|.$$

From Lemma 3.2, by the optimality property of $\tilde{z}$ we have

$$\begin{aligned}\frac{\|(A-\mu_1 B)\tilde{z}\|}{\sqrt{1+|\mu_1|^2}} &\le \frac{\|(A-\mu_1 B)\hat{v}_1\|}{\sqrt{1+|\mu_1|^2}} \\ &\le \frac{\sqrt{1+|\lambda_1|^2}(|\lambda_1-\mu_1|\|B\|+\|A-\mu_1 B\|\sin\theta_1)}{\sqrt{1+|\mu_1|^2}\cos\theta_1} \\ &\quad + \frac{|\lambda_1-\mu_1|\|A-\mu_1 B\|}{\sqrt{1+|\mu_1^2|}}.\end{aligned}$$

Since $\frac{\|(A-\mu_1 B)\tilde{z}\|}{\sqrt{1+|\mu_1|^2}}$ is a residual norm, it is direct from Theorem 2.2 that

$$\sin\angle(v_1,\tilde{z}) \le \frac{\|(A-\mu_1 B)\tilde{z}\|}{\sqrt{1+|\mu_1|^2}\mathrm{sep}(\mu_1,(L,N))}.$$

Therefore, it holds from Lemma 2.3 that

$$\begin{aligned}\sin\angle(x_1,\tilde{z}_1) &\le \sqrt{1+|\mu_1|^2}\sin\angle(v_1,\tilde{z}) \\ &\le \frac{\sqrt{1+|\lambda_1|^2}(|\lambda_1-\mu_1|\|B\|+\|A-\mu_1 B\|\sin\theta_1)}{\cos\theta_1\mathrm{sep}(\mu_1,(L,N))} \\ &\quad + \frac{|\lambda_1-\mu_1|\|A-\mu_1 B\|}{\mathrm{sep}(\mu_1,(L,N))} \\ &< \frac{\sqrt{1+|\lambda_1|^2}(|\lambda_1-\mu_1|(\|B\|+\|A-\mu_1 B\|)+\|A-\mu_1 B\|\sin\theta_1)}{\cos\theta_1\mathrm{sep}(\mu_1,(L,N))},\end{aligned}$$

which proves (3.4). □

Since μ_1 is shown, as Corollary 2.1 indicates, to converge to λ_1 as $\theta_1 \to 0$, we have $\mathrm{sep}(\mu_1,(L,N)) \to \mathrm{sep}(\lambda_1,(L,N))$, a positive constant independent of θ_1, provided that λ_1 is a simple eigenvalue of (M,D,K). So the refined Ritz vector $\tilde{z}_1$ converges to x_1 once $\sin\theta_1 \to 0$.

We mention that Hochstenbach and Sleijpen [5] proposed a refined Rayleigh–Ritz method for the polynomial eigenvalue problem and derived an a priori bound for the residual norm of the refined Ritz pair as the approximate eigenpair of the problem without invoking any linearization; see Theorem 5.1 there.

We continue Example 2.1 to show considerable merits of refined Ritz vectors. For the case that x_1 lies in $\mathrm{span}\{Q\}$ exactly, recall that $\mu_1 = \lambda_1$ exactly. It is easy to verify

that the smallest singular value of the matrix $(\mu_1^2 M + \mu_1 D + K)Q$ is both exactly zero and simple, the optimal solution $\hat{z}_1 = [1, 0]^T$ in (3.1) and the refined Ritz vector $\tilde{z}_1 = Q\hat{z}_1 = x_1$, exactly the desired eigenvector! So in contrast to the Ritz vector, the refined Ritz vector can pick up the desired eigenvector perfectly.

For the case that span$\{Q\}$ is perturbed in the way described in Example 2.1, the optimal solution in (3.1) is

$$\hat{z}_1 = [1.000000000000000,\ 0.000000000006175]^T$$

and the refined Ritz vector

$$\tilde{z}_1 = [0.000000000004708,\ -0.000000000003593,\ 1.000000000000000]^T.$$

So

$$\sin\angle(x_1, \tilde{z}_1) = 5.9 \times 10^{-12},$$

which is almost as small as $\sin\theta_1 = 1.7 \times 10^{-12}$ and much more accurate than the corresponding Ritz vector $\tilde{x}_1$. Meanwhile, the computed residual norm of the refined approximate eigenpair $(\mu_1, \tilde{z}_1)$ is

$$\left\|(\mu_1^2 M + \mu_1 D + K)\tilde{z}_1\right\| = 1.3 \times 10^{-13},$$

eleven orders smaller than that of the Ritz pair $(\mu_1, \tilde{x}_1)$.

4 Conclusions

Theoretically, we have proved that there exists a Ritz value of (M, D, K) that unconditionally converges to the desired eigenvalue when the angle between the subspace span$\{Q\}$ and the desired eigenvector tends to zero. However, the associated Ritz vector only converges conditionally. To this end, we have proposed the refined Ritz vector that is guaranteed to converge unconditionally. We have presented some examples to demonstrate our theory.

The purpose of this paper is not to present efficient and reliable eigensolvers for QEPs, but rather to establish a general convergence theory of the Rayleigh-Ritz method and to show the unconditional convergence of Ritz values and refined Ritz vectors and the conditional convergence of Ritz vectors. Refined Ritz vectors may become a very valuable component and make great improvement in flexible eigensolvers for QEPs. Numerical experiments in [17] have shown that one can gain very much by replacing Ritz vectors by refined Ritz vectors in second-order Arnoldi type methods and their implicitly restarted algorithms.

Acknowledgements We thank the editor Professor Michiel Hochstenbach and the referee very much for their valuable suggestions and comments that made us improve the presentation of the paper very substantially.

References

1. Bai, Z., Su, Y.: SOAR: a second-order Arnoldi method for the solution of the quadratic eigenvalue problem. SIAM J. Matrix Anal. Appl. **26**, 640–659 (2005)
2. Betcke, T., Higham, N.J., Mehrmann, V., Schroder, C., Tisseur, F.: NLEVP: a collection of nonlinear eigenvalue problems. Available from http://www.mims.manchester.ac.uk/research/numerical-analysis/nlevp.html
3. Elsner, L.: The variation of the spectra of matrices. Linear Algebra Appl. **47**, 127–138 (1982)
4. Gohberg, I.C., Lancaster, P., Rodman, L.: Matrix Polynomials. Academic Press, New York (1982)
5. Hochstenbach, M.E., Sleijpen, G.L.G.: Harmonic and refined Rayleigh–Ritz for the polynomial eigenvalue problem. Numer. Linear Algebra Appl. **15**, 35–54 (2008)
6. Hoffnung, L., Li, R.-C., Ye, Q.: Krylov type subspace methods for matrix polynomials. Linear Algebra Appl. **415**, 52–81 (2006)
7. Huang, W.Q., Li, T., Li, Y.T., Lin, W.-W.: A semiorthogonal generalized Arnoldi method and its variations for quadratic eigenvalue problems. Numer. Linear Algebra Appl. **20**, 259–280 (2013)
8. Huitfeldt, J., Ruhe, A.: A new algorithm for numerical path following applied to an example from hydrodynamical flow. SIAM J. Sci. Stat. Comput. **11**, 1181–1192 (1990)
9. Ipsen, I.C.F.: Absolute and relative perturbation bounds for invariant subspaces of matrices. Linear Algebra Appl. **309**, 45–56 (2000)
10. Jia, Z.: Refined iterative algorithms based on Arnoldi's process for large unsymmetric eigenproblems. Linear Algebra Appl. **259**, 1–23 (1997)
11. Jia, Z.: A refined subspace iteration algorithm for large sparse eigenproblems. Appl. Numer. Math. **32**, 35–52 (2000)
12. Jia, Z.: The refined harmonic Arnoldi method and an implicitly restarted refined algorithm for computing interior eigenpairs of large matrices. Appl. Numer. Math. **42**, 489–512 (2002)
13. Jia, Z.: The convergence of harmonic Ritz values, harmonic Ritz vectors and refined harmonic Ritz vectors. Math. Comput. **74**, 1441–1456 (2005)
14. Jia, Z.: Using cross-product matrices to compute the SVD. Numer. Algorithms **42**, 31–61 (2006)
15. Jia, Z., Stewart, G.W.: On the convergence of Ritz values, Ritz vectors, and refined Ritz vectors. TR-99-08, Institute for Advanced Computer Studies and TR-3986, Department of Computer Science, University of Maryland, College Park (1999)
16. Jia, Z., Stewart, G.W.: An analysis of the Rayleigh-Ritz method for approximating eigenspaces. Math. Comput. **70**, 637–647 (2001)
17. Jia, Z., Sun, Y.: Implicitly restarted generalized second-order Arnoldi type algorithms for the quadratic eigenvalue problem (2013). arXiv:1005.3947v3 [math.NA]
18. Lancaster, P.: Lambda-Matrices and Vibrating Systems. Pergamon Press, Oxford (1966)
19. Li, R.-C., Ye, Q.: A Krylov subspace method for quadratic matrix polynomials with application to constrained least squares problems. SIAM J. Matrix Anal. Appl. **25**, 405–428 (2003)
20. Lin, Y., Bao, L.: Block second-order Krylov subspace methods for large-scale quadratic eigenvalue problems. Appl. Math. Comput. **181**, 413–422 (2006)
21. Meerbergen, K.: Locking and restarting quadratic eigenvalue solvers. SIAM J. Sci. Comput. **22**, 1814–1839 (2001)
22. Neumaier, A.: Residual inverse iteration for the nonlinear eigenvalue problem. SIAM J. Numer. Anal. **22**, 914–923 (1985)
23. Sleijpen, G.L.G., Booten, A.G.L., Fokkema, D.R., van der Vorst, H.A.: Jacobi-Davidson type methods for generalized eigenproblems and polynomial eigenproblems. BIT Numer. Math. **36**, 595–633 (1996)
24. Sleijpen, G.L.G., van der Vorst, H.A., van Gijzen, M.B.: Quadratic eigenproblems are no problem. SIAM News **29**, 8–9 (1996)
25. Stewart, G.W.: Matrix Algorithms II: Eigensystems. SIAM, Philadelphia (2001)
26. Stewart, G.W., Sun, J.-G.: Matrix Perturbation Theory. Academic Press, New York (1990)
27. Tisseur, F.: Backward error and condition of polynomial eigenvalue problems. Linear Algebra Appl. **309**, 339–361 (2000)

28. Tisseur, F., Meerbergen, K.: The quadratic eigenvalue problem. SIAM Rev. **43**, 235–286 (2001)
29. Voss, H.: An Arnoldi method for nonlinear eigenvalue problems. BIT Numer. Math. **44**, 387–401 (2004)
30. Wang, B., Su, Y., Bai, Z.: The second-order biorthogonalization procedure and its application to quadratic eigenvalue problems. Appl. Math. Comput. **172**, 788–796 (2006)
31. Ye, Q.: An iterated shift-and-invert Arnoldi algorithm for quadratic matrix eigenvalue problems. Appl. Math. Comput. **172**, 818–827 (2006)