Cluster-robust accuracy bounds for Ritz subspaces

Zhongming Teng a,*,1, Linzhang Lu b,2, Ren-Cang Li c,3

a College of Computer and Information Science, Fujian Agriculture and Forestry University, Fuzhou, Fujian, 350002, PR China
b School of Mathematical Science, Xiamen University & School of Mathematics and Computer Science, Guizhou Normal University, PR China
c Department of Mathematics, University of Texas at Arlington, P.O. Box 19408, Arlington, TX 76019, USA

ARTICLE INFO

Article history:
Received 9 September 2014
Accepted 10 April 2015
Available online 26 April 2015
Submitted by Z. Strakos

MSC:
15A42
65F15

Keywords:
Invariant subspace
Canonical angle
Eigenvalue
Eigenvector
Rayleigh–Ritz procedure
Ritz value
Ritz vector
Majorization

ABSTRACT

Given an approximating subspace for a Hermitian matrix $A$, the Rayleigh–Ritz procedure is commonly used to compute a few approximate eigenvalues (called Ritz values) and corresponding approximate eigenvectors (called Ritz vectors). In this paper, new bounds on the canonical angles between the invariant subspace of $A$ associated with its few extreme (smallest or largest) eigenvalues and its approximating Ritz subspace in terms of the differences between Ritz values and the targeted eigenvalues are obtained. From this result, various bounds are readily available to estimate how accurate the Ritz vectors computed from the approximating subspace may be, based on information on approximation accuracies in the Ritz values. The result is helpful in understanding how Ritz vectors move towards eigenvectors while Ritz values are made to move towards eigenvalues.

© 2015 Elsevier Inc. All rights reserved.
1. Introduction

The Rayleigh–Ritz procedure [1, p. 234] is widely used to find approximate eigenpairs of a Hermitian matrix \( A \in \mathbb{C}^{N \times N} \), given a subspace \( \mathcal{X} \) of \( \mathbb{C}^{N} \) with \( \dim(\mathcal{X}) = \ell \). Let \( \mathcal{X} \) be an orthonormal basis matrix of \( \mathcal{X} \). The basic idea of the procedure goes as follows. Compute the eigen-decomposition of \( X^HAX \):

\[
X^HAX = W\tilde{\Lambda}W^H, \quad \tilde{\Lambda} = \text{diag}(\tilde{\lambda}_1, \ldots, \tilde{\lambda}_\ell),
\]

where \( W = [w_1, \ldots, w_\ell] \in \mathbb{C}^{\ell \times \ell} \) is unitary, and then take each pair \((\tilde{\lambda}_i, Xw_i)\), called a Rayleigh–Ritz pair, as an approximate eigenpair of \( A \). The number \( \tilde{\lambda}_i \) is called a Ritz value and \( Xw_i \) a corresponding Ritz vector.

In the case when \( \mathcal{X} \) is an accurate approximate invariant subspace of \( A \), each Rayleigh–Ritz pair should be a good approximate eigenpair of \( A \). If \( \mathcal{X} \) itself does not well approximate any invariant subspace of \( A \), but contains a subspace that is a good approximation to some invariant subspace of \( A \), then some, but not all, of the Rayleigh–Ritz pairs are expected to approximate well eigenpairs of \( A \). The latter case is more common than the former in eigenvalue computations, where often a subspace is built to contain another subspace that approximates an invariant subspace of \( A \) well, e.g., in the Lanczos method a Krylov subspace is built and usually the Krylov subspace as a whole is not close to any invariant subspace (of the same dimension) but more likely contains a subspace that is a good approximation to an invariant subspace of \( A \).

There are existing results to quantify how good the approximate eigenpairs are. Most results are bounds in terms of the norms of the residual

\[
R(X) := AX - X(X^HAX).
\]

The interested reader is referred to a short summary at the end of [2] for bounds of this kind. The main result of [3] can be interpreted as one of this kind, too. Note \( R(X) = 0 \) if \( \mathcal{X} \) is an exact invariant subspace.

Other results are bounds in terms of the canonical angles between \( \mathcal{X} \) and the invariant subspace which \( \mathcal{X} \) is supposed to approximate. In this regard, Knyazev and Argentati [4] presented the most comprehensive study so far. They obtained several beautiful results in terms of how the vector of eigenvalue differences between the exact eigenvalues and their approximations is weakly majorized by the canonical angles between/from the invariant subspace of interest and/to \( \mathcal{X} \). We will state some of their results to motivate what we will do in Section 3. The results in [4] are basically about estimating the approximation accuracy of (some of) the Ritz values, given information on the approximate accuracy in \( \mathcal{X} \) to an invariant subspace of \( A \). In this paper, we are interested in the converses to these results, i.e., estimating the approximate accuracy in \( \mathcal{X} \), given information on the approximation accuracy of (some of) the Ritz values. Our motivation is from eigenvector computations in Principal Component Analysis in image processing [5,6], where eigenvectors may be computed by optimizing Rayleigh quotients with conjugate gradient type
methods. Recently in the large scale electronic structure calculation, there is an increasing trend to compute a few extreme eigenvalues and corresponding eigenvectors through optimizing Rayleigh quotients. Our main result is helpful in our understanding of the relationships between approximation accuracies in Ritz values and vectors.

Estimates for the accuracy of approximate invariant subspace $\mathcal{X}$ via the approximation accuracy of the Ritz values went back to Weinberger [7] and [8], but these bounds are adversely affected by the presence of clustered, i.e., the bounds are not cluster robust. Some cluster-robust bounds for the angles between the eigenvectors and approximate invariant subspaces in terms of eigenvalue approximation errors can be found in [9]. Cluster robust bounds are also important for accuracy analysis as in [10]. Our results in this paper fall in the similar category – cluster robust bounds. The main difference is that we measure the canonical angles between the invariant subspaces and their approximations.

The rest of this paper is organized as follows. Section 2 collects some preliminaries on the concept of majorization and the canonical angles of two subspaces. In Section 3, we first discuss existing results of Knyazev and Argentati [4] and then let the discussion lead to our main result – Theorem 3.2. Section 4 is devoted to the proof of Theorem 3.2. In Section 5, we present some numerical examples to show the behavior of Theorem 3.2 and to briefly compare our results with one comparable result in [9]. Concluding remarks are given in Section 6.

**Notation.** $\mathbb{C}^{m \times n}$ is the set of all $m \times n$ complex matrices, $\mathbb{C}^n = \mathbb{C}^{n \times 1}$, and $\mathbb{C} = \mathbb{C}^1$. Similarly for $\mathbb{R}^{m \times n}$, $\mathbb{R}^n$, and $\mathbb{R}$, except all involved numbers are real. For a subspace $\mathcal{X}$ of $\mathbb{C}^n$, $\dim(\mathcal{X})$ denotes the dimension of $\mathcal{X}$. For $X \in \mathbb{C}^{m \times n}$, $\text{eig}(X) \in \mathbb{C}^{1 \times n}$ is the vector of the eigenvalues of $X$. For $X \in \mathbb{C}^{m \times n}$, $X$’s submatrices $X_{(k:\ell,i:j)}$, and $X_{(i,j)}$ consist of intersections of row $k$ to row $\ell$, and column $i$ to column $j$, respectively. $\mathbb{R}(X)$ is the column space of $X$, and $X^H$ is its complex conjugate transpose. $I_n$ is the $n \times n$ identity matrix or simply $I$ if its dimension is clear from the context. For $x, y \in \mathbb{R}^{1 \times n}$, notation $x \circ y = [x_1y_1, \ldots, x_ny_n]$ denotes the Hadamard product of $x$ and $y$. In particular, $x^{\circ 2} = x \circ x$.

2. Preliminaries

2.1. Majorization

For $x = [x_1, x_2, \ldots, x_n] \in \mathbb{R}^{1 \times n}$, denote by $x^\downarrow = [x_1^\downarrow, x_2^\downarrow, \ldots, x_n^\downarrow]$, the row vector obtained by rearranging $x_i$ in the descending order, i.e.,

$$x_1^\downarrow \geq x_2^\downarrow \geq \cdots \geq x_n^\downarrow,$$
while \( x^\uparrow = [x_1^\uparrow, x_2^\uparrow, \ldots, x_n^\uparrow] \) is obtained by rearranging \( x_i \) in the ascending order. Given two vectors \( x = [x_1, \ldots, x_n], y = [y_1, \ldots, y_n] \in \mathbb{R}^{1 \times n} \), we say that \( x \) is weakly majorized by \( y \), in symbols \( x \prec_w y \), if \([11, \text{chapter II}]\)

\[
\sum_{i=1}^k x_i^\uparrow \leq \sum_{i=1}^k y_i^\uparrow, \quad \text{for } 1 \leq k \leq n. \tag{2.1}
\]

If, in addition,

\[
\sum_{i=1}^n x_i^\uparrow = \sum_{i=1}^n y_i^\uparrow, \tag{2.2}
\]

we say that the vector \( x \) is majorized by \( y \), in symbols \( x \prec y \).

Majorization provides a succinct way to express numerous useful inequalities involving two vectors \( x \) and \( y \). For example, suppose \( x \) and \( y \) are entrywise nonnegative, i.e., \( x_i \geq 0 \) and \( y_i \geq 0 \), and \( x \prec_w y \). Then, besides those in (2.1) by the definition, we have \([11, \text{p. } 42] \)

\[
\|x\|_p := \left( \sum_{i=1}^n |x_i|^p \right)^{1/p} \leq \|y\|_p := \left( \sum_{i=1}^n |y_i|^p \right)^{1/p} \quad \text{for } 1 \leq p \leq \infty. \tag{2.2}
\]

For \( p = \infty \), this is simply (2.1) for \( k = 1 \): \( \max_i |x_i| \leq \max_i |y_i| \). For the purpose of error estimation in numerical computations, often the inequality (2.2) for \( p = 2 \) and \( \infty \) suffices.

### 2.2. Angles between subspaces

Consider two subspaces \( \mathcal{X} \) and \( \mathcal{Y} \) of \( \mathbb{C}^N \) and suppose

\[
\ell := \dim(\mathcal{X}) \geq \dim(\mathcal{Y}) =: k. \tag{2.3}
\]

Let \( X \in \mathbb{C}^{N \times \ell} \) and \( Y \in \mathbb{C}^{N \times k} \) be orthonormal basis matrices of \( \mathcal{X} \) and \( \mathcal{Y} \), respectively, i.e.,

\[
X^H X = I_\ell, \quad \mathcal{X} = \mathcal{R}(X), \quad \text{and} \quad Y^H Y = I_k, \quad \mathcal{Y} = \mathcal{R}(Y),
\]

and denote by \( \sigma_j \) for \( 1 \leq j \leq k \) in the ascending order, i.e., \( \sigma_1 \leq \cdots \leq \sigma_k \), the singular values of \( Y^H X \). The \( k \) canonical angles \( \theta_j(\mathcal{X}, \mathcal{Y}) \) from \( \mathcal{Y} \) to \( \mathcal{X} \) are defined by

\[
0 \leq \theta_j(\mathcal{X}, \mathcal{Y}) := \arccos \sigma_j \leq \frac{\pi}{2} \quad \text{for } 1 \leq j \leq k.
\]

\footnote{If \( \ell = k \), we may say that these angles are between \( \mathcal{X} \) and \( \mathcal{Y} \).}
They are in the descending order, i.e., \( \theta_1(X, Y) \geq \cdots \geq \theta_k(X, Y) \). Set
\[
\Theta(X, Y) = [\theta_1(X, Y), \ldots, \theta_k(X, Y)] \in \mathbb{R}^{1 \times k}.
\]
It can be seen that angles so defined are independent of the orthonormal basis matrices \( X \) and \( Y \) (which are not unique). A different way to define these angles is through orthogonal projections onto \( X \) and \( Y \) [12].

When \( k = 1 \), i.e., \( Y \) is a vector, there is only one canonical angle from \( Y \) to \( X \) and so we will simply write \( \theta(X, Y) \).

In what follows, for two subspaces \( X \) and \( Y \) of \( \mathbb{C}^N \), we denote by \( \sin(\Theta(X, Y)) \), the vector consisting of the sine of the canonical angles from \( Y \) to \( X \), i.e.,
\[
\sin(\Theta(X, Y)) := [\sin(\theta_1(X, Y)), \ldots, \sin(\theta_k(X, Y))].
\]

**Proposition 2.1.** (See [13, Proposition 2.1]...) Let \( X \) and \( Y \) be two subspaces in \( \mathbb{C}^N \) with \( \dim(X) = \ell \), \( \dim(Y) = k \) and \( \ell \geq k \).

1. For any \( \hat{X} \subseteq X \) with \( \dim(\hat{X}) = \dim(Y) = k \), we have \( \theta_j(X, Y) \leq \theta_j(\hat{X}, Y) \) for \( 1 \leq j \leq k \).
2. There exist \( \hat{X} \subseteq X \) with \( \dim(\hat{X}) = k \leq \ell \) such that \( \theta_j(X, Y) = \theta_j(\hat{X}, Y) \) for \( 1 \leq j \leq k \).

Later, we also need the following proposition which says that the number of zero canonical angles is the same as the dimension of the intersection of two subspaces, a fact that is not hard to prove (see, e.g., [14]).

**Proposition 2.2.** Let \( X \) and \( Y \) be two subspaces in \( \mathbb{C}^N \) satisfying (2.3). Then
\[
\theta_i(X, Y) = 0 \quad \text{for} \quad k - m_0 + 1 \leq i \leq k,
\]
where \( m_0 = \dim(X \cap Y) \).

3. Main result

For the rest of this paper, \( A \) is an \( N \times N \) Hermitian matrix, and has

\[
\begin{align*}
eigenvalues: \quad & \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_N, \quad \text{and} \\
othornormal eigenvectors: \quad & u_1, u_2, \ldots, u_N, \quad \text{and} \\
eigen-decomposition: \quad & A = U \Lambda U^H \quad \text{and} \quad U^H U = I_N.
\end{align*}
\]

(3.1)
As in the introduction, let $\mathcal{X}$ be a subspace of $\mathbb{C}^N$ with $\dim(\mathcal{X}) = \ell$, intended to approximate an invariant subspace of $A$ in the sense that either $\mathcal{X}$ as a whole is an approximate invariant subspace or contains another subspace of dimension less than $\ell$ that is an approximate invariant subspace of $A$. $X \in \mathbb{C}^{N \times \ell}$ is an orthonormal basis matrix of $\mathcal{X}$. Similarly, we introduce notations for $X^HAX$:

\[
\text{eigenvalues (also Ritz values): } \tilde{\lambda}_1 \leq \tilde{\lambda}_2 \leq \cdots \leq \tilde{\lambda}_\ell, \text{ and } \Omega = \text{diag}(\tilde{\lambda}_1, \tilde{\lambda}_2, \ldots, \tilde{\lambda}_\ell),
\]

\[
\text{orthonormal eigenvectors: } w_1, w_2, \ldots, w_\ell, \text{ and } W = [w_1, w_2, \ldots, w_\ell],
\]

\[
\text{eigen-decomposition: } (X^HAX)W = W\Omega
\]

\[
\text{and } W^HW = I_\ell,
\]

\[
\text{Ritz vectors: } \tilde{u}_j = Xw_j \text{ for } 1 \leq j \leq \ell, \text{ and } \tilde{U} = [\tilde{u}_1, \tilde{u}_2, \ldots, \tilde{u}_\ell].
\]

Knyazev and Argentati [4] established several beautiful results that lead to bounds on $\lambda_i - \tilde{\lambda}_i$ in terms of the canonical angles between/from an invariant subspace of $A$ and/to $\mathcal{X}$. Some of them that are relevant to this paper are summarized in Theorem 3.1 below.

**Theorem 3.1.** *(See [4].)* Suppose $\mathcal{Y}$ with $\dim(\mathcal{Y}) = k$ is an invariant subspace of $A$, corresponding to its eigenvalues

\[
\lambda_{\pi_1} \leq \lambda_{\pi_2} \leq \cdots \leq \lambda_{\pi_k},
\]

where $1 \leq \pi_1 < \pi_2 < \cdots < \pi_k \leq N$ and suppose $k \leq \ell$. Let $m = \dim(\mathcal{X} + \mathcal{Y})$, and let the eigenvalues of $P_{\mathcal{X}+\mathcal{Y}}A|_{\mathcal{X}+\mathcal{Y}}$ be $\omega_1 \leq \omega_2 \leq \cdots \leq \omega_m$, where $P_{\mathcal{X}+\mathcal{Y}}$ is the orthogonal projector onto $\mathcal{X} + \mathcal{Y}$ and $P_{\mathcal{X}+\mathcal{Y}}A|_{\mathcal{X}+\mathcal{Y}}$ is the restriction of $P_{\mathcal{X}+\mathcal{Y}}A$ onto $\mathcal{X} + \mathcal{Y}$, and set

\[
\delta := \omega_m - \omega_1 \quad \text{(which is no bigger than } \lambda_N - \lambda_1),
\]

\[
\Delta := [\omega_m - \omega_1, \omega_{m-1} - \omega_2, \ldots, \omega_{m-k+1} - \omega_k],
\]

\[
\hat{\Delta} := [\omega_m - \omega_1, \omega_m - \omega_2, \ldots, \omega_m - \omega_k].
\]

1. *(See [4, Theorem 2.1].)* If $k = \ell$, then

\[
[[\tilde{\lambda}_1 - \lambda_{\pi_1}], \ldots, [\tilde{\lambda}_k - \lambda_{\pi_k}]] \prec_w \delta \sin^2 \Theta(\mathcal{X}, \mathcal{Y}).
\]

2. *(See [4, Theorem 2.2].)* If $k = \ell$ and if $\pi_i = i$ for $1 \leq i \leq \ell$, then

\[
0 \leq [\tilde{\lambda}_1 - \lambda_1, \ldots, \tilde{\lambda}_k - \lambda_k] \prec_w \Delta \circ \sin^2 \Theta(\mathcal{X}, \mathcal{Y})
\]

\[
\prec_w \delta \sin^2 \Theta(\mathcal{X}, \mathcal{Y}).
\]
3. (See [4, Theorem 2.4].) If \( k < \ell \) and if \( \pi_i = i \) for \( 1 \leq i \leq k \), then

\[
0 \leq [\tilde{\lambda}_1 - \lambda_1, \ldots, \tilde{\lambda}_k - \lambda_k] \prec_w \hat{\Delta} \circ \sin^{\circ 2} \Theta(X, y) \prec_w \delta \sin^{\circ 2} \Theta(X, y).
\]

The theory behind computing a few extreme eigenvalues and corresponding eigenvectors through optimizing Rayleigh quotients is the well-known Ky Fan trace minimization principle [15]

\[
\min_{Z^H Z = I_k} \text{trace}(Z^H A Z) = \sum_{i=1}^{k} \lambda_i
\]

and its variations. Conceivably, as \( \text{trace}(X^H A X) \) approaches the targeted value \( \sum_{i=1}^{k} \lambda_i \), \( \mathcal{R}(X) \) should be a good approximate invariant subspace. The question is how good it is. For that purpose, there is an existing well-known inequality in the case \( k = \ell \):

\[
(\lambda_{k+1} - \lambda_{k}) \| \sin \Theta(\mathcal{U}_k, X) \|_2^2 \leq \sum_{i=1}^{k} (\tilde{\lambda}_i - \lambda_i),
\]

first appeared in [8] in 1974 and rediscovered a few times later (see [16–18]), where the invariant subspace \( \mathcal{U}_k \) is defined by

\[
\mathcal{U}_k = \mathcal{R}(U_{(i,1:k)}).
\]

Note \( \| \sin \Theta(\mathcal{U}_k, X) \|_2^2 = \sum_{i=1}^{k} \sin^2 \theta_i(\mathcal{U}_k, X) \) since \( \sin \Theta(\mathcal{U}_k, X) \) is interpreted as a vector.

Given the various results collected in Theorem 3.1 and (3.3), it is natural to expect

\[
(\lambda_{k+1} - \lambda_{k}) \sin^{\circ 2} \Theta(\mathcal{U}_k, X) \prec_w [\tilde{\lambda}_1 - \lambda_1, \ldots, \tilde{\lambda}_k - \lambda_k].
\]

Unfortunately, it fails, as Example 5.1 in Section 5 shows.

Although (3.5) may fail, in what follows we will establish inequalities that resemble (3.5) but only slightly weaker. Compared to (3.3), our results are more general and sharper. Discussions on another existing result in [9] are at the end of this section.

**Theorem 3.2.** Let \( \ell = \dim(X) \geq k = \dim(\mathcal{U}_k) \) and \( m = \dim(X + \mathcal{U}_k) \), and let the eigenvalues of \( P_{X+\mathcal{U}_k}A|_{X+\mathcal{U}_k} \) be

\[
\omega_1 \leq \omega_2 \leq \cdots \leq \omega_m.
\]

Define \( \tilde{\mathcal{U}}_j := \mathcal{R}(X[w_1, w_2, \ldots, w_j]) = \mathcal{R}(\tilde{U}_{(i,1:j)}) \). If \( m > k \), then
\[
\min\{m-k,j\} \sum_{i=1}^{m} (\omega_{k+i} - \omega_{j-i+1}) \sin^2 \theta_i(\mathcal{U}_k, \tilde{\mathcal{U}}_j) \leq \sum_{i=1}^{j} (\tilde{\lambda}_i - \lambda_i) \quad \text{for } 1 \leq j \leq k, \quad (3.6a)
\]
\[
(\omega_{k+1} - \omega_j) \sum_{i=1}^{j} \sin^2 \theta_i(\mathcal{U}_k, \tilde{\mathcal{U}}_j) \leq \sum_{i=1}^{j} (\tilde{\lambda}_i - \lambda_i) \quad \text{for } 1 \leq j \leq k. \quad (3.6b)
\]

The condition \( m > k \) in the theorem is not restrictive at all. In fact, \( m \geq k \) always, and if \( m = k \), then \( \mathcal{X} = \mathcal{U}_k \) and thus \( \Theta(\mathcal{U}_k, \mathcal{X}) = 0 \) and \( \tilde{\lambda}_i = \lambda_i \) for \( 1 \leq i \leq k \), a highly unlikely but otherwise trivial and welcomed case for which the inequalities \((3.6a)\) and \((3.6b)\) trivially hold as equalities.

Suppose \( m > k \). By the Cauchy interlacing inequalities [1, p. 203],
\[
\lambda_i \leq \omega_i \leq \lambda_{N-m+i} \quad \text{for } 1 \leq i \leq m.
\]
Since \( \mathcal{U}_k \subset \mathcal{X} + \mathcal{U}_k \), \( \lambda_i = \omega_i \) for \( 1 \leq i \leq k \). Therefore Theorem 3.2 implies.

**Corollary 3.1.** Assume the conditions of Theorem 3.2, we have
\[
\min\{m-k,j\} \sum_{i=1}^{m} (\lambda_{k+i} - \lambda_{j-i+1}) \sin^2 \theta_i(\mathcal{U}_k, \tilde{\mathcal{U}}_j) \leq \sum_{i=1}^{j} (\tilde{\lambda}_i - \lambda_i) \quad \text{for } 1 \leq j \leq k, \quad (3.7a)
\]
\[
(\lambda_{k+1} - \lambda_j) \sum_{i=1}^{j} \sin^2 \theta_i(\mathcal{U}_k, \tilde{\mathcal{U}}_j) \leq \sum_{i=1}^{j} (\tilde{\lambda}_i - \lambda_i) \quad \text{for } 1 \leq j \leq k. \quad (3.7b)
\]

The inequalities in \((3.7)\) differ from the ones in \((3.6)\) only in replacing \( \omega_i \) with \( \lambda_i \).

**Corollary 3.2.** Assume the conditions of Theorem 3.2, we have
\[
\sum_{i=1}^{k} (\omega_{k+i} - \omega_{k-i+1}) \sin^2 \theta_i(\mathcal{U}_k, \mathcal{X}) \leq \sum_{i=1}^{k} (\tilde{\lambda}_i - \lambda_i), \quad (3.8a)
\]
\[
\sum_{i=1}^{k} (\lambda_{k+i} - \lambda_{k-i+1}) \sin^2 \theta_i(\mathcal{U}_k, \mathcal{X}) \leq \sum_{i=1}^{k} (\tilde{\lambda}_i - \lambda_i). \quad (3.8b)
\]

**Proof.** Let \( j = k \) in \((3.6a)\), we have
\[
\sum_{i=1}^{k} (\tilde{\lambda}_i - \lambda_i) \geq \sum_{i=1}^{k} (\omega_{k+i} - \omega_{k-i+1}) \sin^2 \theta_i(\mathcal{U}_k, \tilde{\mathcal{U}}_k)
\]
\[
\geq \sum_{i=1}^{k} (\omega_{k+i} - \omega_{k-i+1}) \sin^2 \theta_i(\mathcal{U}_k, \mathcal{X}).
\]
The last inequality holds because of Proposition 2.1. This gives (3.8a). The inequality (3.8b) follows from (3.8a) due to the same reason as Corollary 3.1 follows from Theorem 3.2.

The inequality (3.8b) implies (3.3). Next we shall derive an inequality from (3.6b) that only slightly differs from (3.5) that was expected but not true. Let

$$\sum_{i=1}^{j}(\tilde{\lambda}_{t_i} - \lambda_{t_i}) \geq \sum_{i=1}^{j}(\tilde{\lambda}_i - \lambda_i) \geq (\omega_{k+1} - \omega_j) \sum_{i=1}^{j} \sin^2 \theta_i(U_k, \tilde{U}_j) \geq (\lambda_{k+1} - \lambda_k) \sum_{i=1}^{j} \sin^2 \theta_i(U_k, \tilde{U}_j).$$

(3.10)

On the other hand, (3.5) is equivalent to

$$\sum_{i=1}^{j}(\tilde{\lambda}_{t_i} - \lambda_{t_i}) \geq (\lambda_{k+1} - \lambda_k) \sum_{i=1}^{j} \sin^2 \theta_i(U_k, \mathcal{X}).$$

(3.11)

The only difference between (3.10) and (3.11) is the use of $\theta_i(U_k, \tilde{U}_j)$ in (3.10) instead of $\theta_i(U_k, \mathcal{X})$ in (3.11) that contributes to the failure of (3.11) while (3.10) is proved true always.

Among numerous cluster-robust bounds in [9], the one that is most comparable to ours is [9, Lemma 4]

$$\sum_{i=1}^{k} (\lambda_{k+1} - \lambda_i) \sin^2 \theta(u_i, \mathcal{X}) \leq \sum_{i=1}^{k} (\tilde{\lambda}_i - \lambda_i)$$

(3.12)

for the case $\ell = k$ and $\lambda_{k+1} > \lambda_k$. A major difference between this bound and ours is the use of different angles. In (3.12), the involved angles are between eigenvectors $u_i$ and the approximate invariant subspace $\mathcal{X}$, whereas what we have been using so far are the canonical angles between the invariant subspace $U_k$ and its approximation $\mathcal{X}$. In general, although these angles can be related, e.g., by [13, Proposition 2.2]

$$\max_{1 \leq i \leq k} \sin^2 \theta(u_i, \mathcal{X}) \leq \sum_{i=1}^{k} \sin^2 \theta_i(U_k, \mathcal{X}) = \sum_{i=1}^{k} \sin^2 \theta(u_i, \mathcal{X}),$$
a definitive claim as to which one of (3.12) and (3.8b) is always sharper than the other is difficult to establish. Later in Section 5, we will present some numerical examples to compare them. In these examples, (3.8b) is sharper than (3.12).

4. Proof of Theorem 3.2

We need the following lemmas.

Lemma 4.1. (See [18, Lemma 2.3].) Let \( \gamma_1 \geq \gamma_2 \geq \cdots \geq \gamma_m \geq 0 \). If \([\beta_1, \beta_2, \ldots, \beta_m]\) majorizes \([\alpha_1, \alpha_2, \ldots, \alpha_m]\), then

\[
\sum_{i=1}^{m} \gamma_i \beta_i^* \leq \sum_{i=1}^{m} \gamma_i \alpha_i \leq \sum_{i=1}^{m} \gamma_i \beta_i^\dagger.
\]

Lemma 4.2 (Schur’s Theorem). (See [11, p. 35].) For Hermitian matrix \( H \in \mathbb{C}^{n \times n} \),

\[
\text{diag}(H) < \text{eig}(H),
\]

where \( \text{diag}(H) \in \mathbb{R}^{1 \times n} \) is the vector of the diagonal entries of \( H \).

Proof of Theorem 3.2. Without loss of generality, we first assume \( A \) is positive definite. Otherwise we shift \( A \) to \( A - \mu I \) for any \( \mu < \lambda_1 \). Doing so will shift all eigenvalues of \( A \) and Ritz values by the same \( \mu \) and thus keep all the differences \( \lambda_i - \tilde{\lambda}_j \), \( \lambda_i - \lambda_j \) unchanged.

We may also assume \( X + \mathbb{U}_k = \mathbb{C}^N \) since we can simply replace \( A \) by \( P_{X+\mathbb{U}_k} A |_{X+\mathbb{U}_k} \). Doing so changes no Ritz values and Ritz vectors, but changes \( N \) to \( m \) and reduce the set \( \{\lambda_i\}_{i=1}^{N} \) to \( \{\omega_i\}_{i=1}^{m} \). Note that \( \omega_i = \lambda_i \) for \( 1 \leq i \leq k \) regardless, because \( \mathbb{U}_k \subseteq X + \mathbb{U}_k \).

Assume \( A \) is positive definite and \( X + \mathbb{U}_k = \mathbb{C}^N \). The key steps to our proof are 1) the establishment of (4.1) below for \( \sum_{i=1}^{j} (\tilde{\lambda}_i - \lambda_i) \), and 2) bounding the terms in the right-hand side from below.

Recall (3.1) and (3.2) and partition \( W \) and \( U \) as

\[
W = \begin{bmatrix} W_1 & W_2 \end{bmatrix}, \quad U = \begin{bmatrix} U_1 & U_2 & U_3 \end{bmatrix}, \quad \Lambda = \begin{bmatrix} \Lambda_1 & \Lambda_2 & \Lambda_3 \end{bmatrix}.
\]

Let \( \tilde{U}_j = X W_1 \), and write

\[
U^H \tilde{U}_j = \begin{bmatrix} \tilde{U}_1 & \tilde{U}_2 & \tilde{U}_3 \end{bmatrix}, \quad \tilde{U}_{1,2} = \begin{bmatrix} \tilde{U}_1 \\ \tilde{U}_2 \\ \tilde{U}_3 \end{bmatrix}, \quad \Lambda_{1,2} = \begin{bmatrix} \Lambda_1 \\ \Lambda_2 \end{bmatrix}.
\]
We have \( \tilde{U}_j^H A \tilde{U}_j = \tilde{U}_1^H \Lambda_1 \tilde{U}_1 + \tilde{U}_2^H \Lambda_2 \tilde{U}_2 + \tilde{U}_3^H \Lambda_3 \tilde{U}_3 = \tilde{U}_{1,2}^H \Lambda_{1,2} \tilde{U}_{1,2} + \tilde{U}_3^H \Lambda_3 \tilde{U}_3 \). Therefore

\[
\sum_{i=1}^j \tilde{\lambda}_i = \text{trace}(\tilde{U}_j^H A \tilde{U}_j) \\
= \text{trace}(\tilde{U}_{1,2}^H \Lambda_{1,2} \tilde{U}_{1,2}) + \text{trace}(\tilde{U}_3^H \Lambda_3 \tilde{U}_3) \\
= \text{trace}(\tilde{U}_{1,2} \tilde{U}_{1,2}^H \Lambda_{1,2}) + \text{trace}(\tilde{U}_3 \tilde{U}_3^H \Lambda_3)
\]

which yields

\[
\sum_{i=1}^j (\tilde{\lambda}_i - \lambda_i) = -\text{trace}([I_k - \tilde{U}_{1,2} \tilde{U}_{1,2}^H] \Lambda_{1,2}) + \sum_{i=j+1}^k \lambda_i + \text{trace}(\tilde{U}_3 \tilde{U}_3^H \Lambda_3). \tag{4.1}
\]

Next, we bound the terms in the right-hand side of (4.1) from below. It can be verified that \( \tilde{U}_{1,2}^H \tilde{U}_{1,2} + \tilde{U}_3^H \tilde{U}_3 = I_j \) and thus

\[
eig(\tilde{U}_3 \tilde{U}_3^H) = \eig(I_j - \tilde{U}_{1,2} \tilde{U}_{1,2}^H) = [\sin^2 \theta_i(\lambda_k, \tilde{U}_j) \text{ for } 1 \leq i \leq j],
\]

\[
eig(I_k - \tilde{U}_{1,2} \tilde{U}_{1,2}^H) = [\sin^2 \theta_i(\lambda_k, \tilde{U}_j) \text{ for } 1 \leq i \leq j, \underbrace{1, \ldots, 1}_{k-j}], \tag{4.2}
\]

\[
eig(\tilde{U}_3 \tilde{U}_3^H) = [\sin^2 \theta_i(\lambda_k, \tilde{U}_j) \text{ for } 1 \leq i \leq \min\{N-k, j\}, \underbrace{0, \ldots, 0}_{N-k-j}]. \tag{4.3}
\]

Since \( \text{diag}(I_k - \tilde{U}_{1,2} \tilde{U}_{1,2}^H) \prec \eig(I_k - \tilde{U}_{1,2} \tilde{U}_{1,2}^H) \) and \( \text{diag}(\tilde{U}_3 \tilde{U}_3^H) \prec \eig(\tilde{U}_3 \tilde{U}_3^H) \) by Lemma 4.2, we have by Lemma 4.1, (4.2), and (4.3)

\[
\text{trace}([I_k - \tilde{U}_{1,2} \tilde{U}_{1,2}^H] \Lambda_{1,2}) = \text{diag}(I_k - \tilde{U}_{1,2} \tilde{U}_{1,2}^H) \cdot [\lambda_1, \lambda_2, \ldots, \lambda_k]^H \\
\leq \sum_{i=1}^j \lambda_i \sin^2 \theta_{j-i+1}(\lambda_k, \tilde{U}_j) + \sum_{i=j+1}^k \lambda_i \\
= \sum_{i=1}^j \lambda_{j-i+1} \sin^2 \theta_i(\lambda_k, \tilde{U}_j) + \sum_{i=j+1}^k \lambda_i, \tag{4.4}
\]

\[
\text{trace}(\tilde{U}_3 \tilde{U}_3^H \Lambda_3) = \text{diag}(\tilde{U}_3 \tilde{U}_3^H) \cdot [\lambda_{k+1}, \lambda_{k+2}, \ldots, \lambda_N]^H \\
\geq \sum_{i=1}^{\min\{N-k, j\}} \lambda_{k+i} \sin^2 \theta_i(\lambda_k, \tilde{U}_j). \tag{4.5}
\]
Finally, we combine (4.1), (4.4), and (4.5) to get

\[
\sum_{i=1}^{j} (\tilde{\lambda}_i - \lambda_i) \geq -\sum_{i=1}^{j} \lambda_{j-i+1} \sin^2 \theta_i(U_k, \tilde{U}_j) + \sum_{i=1}^{\min\{N-k,j\}} \lambda_{k+i} \sin^2 \theta_i(U_k, \tilde{U}_j)
= \sum_{i=1}^{\min\{N-k,j\}} (\lambda_{k+i} - \lambda_{j-i+1}) \sin^2 \theta_i(U_k, \tilde{U}_j).
\]

The last equality holds because of \(\sin^2 \theta_i(U_k, \tilde{U}_j) = 0\) if \(k \geq j \geq i > N - k\) by Proposition 2.2. This is (3.6a). For (3.6b), we notice that \(\omega_{k+i} - \omega_{j-i+1} \geq \omega_{k+1} - \omega_j\) for \(1 \leq i \leq j \leq k\).

5. Numerical examples

In this section, we present some numerical examples to illustrate our results. In particular, we will demonstrate that our lower bounds are preferred to (3.3), especially in case of tiny \(\lambda_{k+1} - \lambda_k\) or a tight cluster. In these examples, our (3.8b) is sharper than (3.12), a comparable result from [9, Lemma 4].

In the examples below, without loss of generality, we take

\[A = \text{diag}(\lambda_1, \ldots, \lambda_N).\]

Thus the eigenvector matrix \(U = I_N\). Let \(N - k \geq k\) for simplicity.

We shall focus on comparisons among our bounds (3.7a), (3.7b), (3.8b), (3.10), and the existing result (3.3) and (3.12). For this reason, we will measure the following errors, for \(j = 1, \ldots, k\),

\[
\varepsilon_{1,j} = \sum_{i=1}^{j} (\tilde{\lambda}_i - \lambda_i) \leq \varepsilon_{2,j} = \sum_{i=1}^{j} (\tilde{\lambda}_{t_i} - \lambda_{t_i}),
\]

and their lower bounds by (3.7a) and (3.7b)

\[
\varepsilon_{3,j} = \sum_{i=1}^{j} (\lambda_{k+i} - \lambda_{j-i+1}) \sin^2 \theta_i(U_k, \tilde{U}_j),
\]

\[
\varepsilon_{4,j} = (\lambda_{k+1} - \lambda_j) \sum_{i=1}^{j} \sin^2 \theta_i(U_k, \tilde{U}_j),
\]

\[
\varepsilon_{5,j} = (\lambda_{k+1} - \lambda_k) \sum_{i=1}^{j} \sin^2 \theta_i(U_k, \tilde{U}_j),
\]

where \(\{t_i \text{ for } 1 \leq i \leq k\}\) is the permutation of \(\{1, 2, \ldots, k\}\) as determined by (3.9), and the lower bound by (3.8b)
\[ \varepsilon_6 = \sum_{i=1}^{k} (\lambda_{k+i} - \lambda_{k-i+1}) \sin^2 \theta_i(U_k, X). \]  

(5.2)

For comparing our results with the existing (3.3) and (3.12) in [9, Lemma 4], we also introduce

\[ \varepsilon_7 = (\lambda_{k+1} - \lambda_k) \sum_{i=1}^{k} \sin^2 \theta_i(U_k, X), \]  

(5.3)

\[ \varepsilon_8 = \sum_{i=1}^{k} (\lambda_{k+1} - \lambda_i) \sin^2 \theta(u_i, X). \]  

(5.4)

They are the left-hand sides of (3.3) and (3.12), respectively. It is clear that

\[ \varepsilon_{3,1} = \varepsilon_{4,1}, \quad \varepsilon_{3,k} \geq \varepsilon_6, \quad \varepsilon_{4,k} = \varepsilon_{5,k} \geq \varepsilon_7, \]

where the inequalities become equalities when \( \ell = k \).

All \( \varepsilon_{3,j}, \varepsilon_{4,j}, \) and \( \varepsilon_{5,j} \) are lower bounds of \( \varepsilon_{1,j} \) and thus also of \( \varepsilon_{2,j} \) for \( j = 1, \ldots, k \). Consequently, \( \varepsilon_6 \) and the existing \( \varepsilon_7 \) and \( \varepsilon_8 \) are also lower bounds of \( \varepsilon_{1,k} \) and \( \varepsilon_{2,k} \).

**Example 5.1.** We first consider a counter example of (3.5). Let

\[ \lambda_1 = -1.0642, \quad \lambda_2 = -0.2490, \quad \lambda_3 = 1.2347, \quad \lambda_4 = 1.6035, \]

and

\[
X = \begin{bmatrix}
-0.1440 & 0.4026 \\
-0.9444 & -0.3116 \\
-0.2788 & 0.6496 \\
-0.0978 & 0.5646 \\
\end{bmatrix}.
\]

Here \( X \) is only shown with 4 significant digits of \( X \). To make this example repeatable, within MATLAB we correct \( X \) by executing \( \text{qr}(X,0) \) and reset \( X \) accordingly. It is computed (again only 4 significant digits are shown)

\[
\text{eig}(X^HAX) = [-0.2311, 0.9338], \quad \text{i.e.,} \quad \tilde{\lambda}_1 = -0.2311, \quad \tilde{\lambda}_2 = 0.9338.
\]

As in the definition of (3.4), let \( U_2 = \mathcal{R}(U(:,1:2)) \). Then it can be seen that \( \sin \Theta(X, U_2) = [0.9040, 0.1039] \) which consists of the singular values of \( X_{(3:4,:)} \) because

\[
I - X^H U(:,1:2) U^H(:,1:2) X = X^H \left[ I - U(:,1:2) U^H(:,1:2) \right] X \\
= X^H U(:,3:4) U^H(:,3:4) X.
\]
Finally

\[(\lambda_3 - \lambda_2) \sin^2 \theta_1(U_2, X) = 1.2126 > \max_{1 \leq i \leq 2} (\hat{\lambda}_i - \lambda_i) = 1.1828,\]

an opposite of the inequality (3.5). \(\square\)

**Example 5.2.** In this example, we make up a matrix such that \(\lambda_{k+1} - \lambda_k\) is small. Let \(N = 10, \ell = k = 5,\)

\[
\begin{align*}
\lambda_1 &= 0.0826, & \lambda_2 &= 0.4116, & \lambda_3 &= 0.5118, & \lambda_4 &= 0.5518, & \lambda_5 &= 0.5835, \\
\lambda_6 &= 0.5836, & \lambda_7 &= 0.6026, & \lambda_8 &= 0.7196, & \lambda_9 &= 0.7505, & \lambda_{10} &= 0.9962,
\end{align*}
\]

and \(X = \mathcal{R}(X),\) where \(X\) is an \(N \times k\) orthonormal matrix generated by \(\text{qr}(X_0, 0)\) in MATLAB with

\[
X_0 = \begin{bmatrix}
1 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 \\
\frac{1}{N} & \sin 1 & \cos 1 & \tan 1 & \cot 1 \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
\frac{N-k}{N} & \sin(N-k) & \cos(N-k) & \tan(N-k) & \cot(N-k)
\end{bmatrix}
\]

For this example, \(\lambda_{k+1} - \lambda_k = 10^{-4},\) potentially bad news for \(\varepsilon_7 = \varepsilon_{4,k} = \varepsilon_{5,k}\) as lower bounds for \(\varepsilon_{1,k} = \varepsilon_{2,k}.\) Table 5.1 reports the lower bounds \(\varepsilon_{3,j}, \varepsilon_{4,j}\) and \(\varepsilon_{5,j}\) as defined in (5.1b), (5.1c) and (5.1d) on the differences \(\varepsilon_{1,j}\) and \(\varepsilon_{2,j}\) as defined in (5.1a) between Ritz values and the targeted eigenvalues. Our bound \(\varepsilon_{3,j}\) for \(j = 1, \ldots, k\) clearly stands out as the best. In particular, \(\varepsilon_{3,5} = \varepsilon_6\) is three magnitudes sharper than \(\varepsilon_7 = \varepsilon_{4,5} = \varepsilon_{5,5},\) and it is also sharper than \(\varepsilon_8 = 0.315.\) Overall \(\varepsilon_{4,j}\) is very good, too, except for \(j = 5. \quad \square\)

**Example 5.3.** We take \(N = 600\) and

\[
\begin{align*}
\lambda_1 &= -3, & \lambda_2 &= -2.5, & \lambda_3 &= -2, & \lambda_i = \frac{i-4}{N}, & i = 4, \ldots, N.
\end{align*}
\]
There are two eigenvalue clusters: \( \{\lambda_1, \lambda_2, \lambda_3\} \) and \( \{\lambda_4, \ldots, \lambda_N\} \). In this example, we use the Lanczos method [1, chapter 13] to generate the subspace \( \mathcal{X} \). The 10-step Lanczos process generates a Krylov subspace \( \mathcal{X} = \text{span}\{v_0, A^0 v_0, \ldots, A^9 v_0\} \), where \( v_0 \) is an initial vector of all ones. For the example, \( \ell = \dim(\mathcal{X}) = 10 \). As explained in Section 1, usually, \( \mathcal{X} \) as a whole is not close to any invariant subspace, but more likely there is a subspace of \( \mathcal{X} \) which is close to an invariant subspace. Again, we set \( k = 5 \) and seek to bound the angles from \( \mathcal{U}_k \) to \( \mathcal{X} \) by the differences of the first five smallest Ritz values and the targeted eigenvalues.

We computed \( \varepsilon_{1,j}, \varepsilon_{2,j}, \varepsilon_{3,j}, \varepsilon_{4,j}, \) and \( \varepsilon_{5,j} \) for \( j = 1, \ldots, 5 \) in Table 5.2. We also computed \( \varepsilon_6 = 0.602 \times 10^{-2} \) and \( \varepsilon_7 = 0.312 \times 10^{-2} \) which are smaller than \( \varepsilon_{3,k} \) and \( \varepsilon_{4,k} = \varepsilon_{5,k} \) (but not by much), respectively, as expected since \( \ell > k \), and \( \varepsilon_8 = 0.470 \times 10^{-2} \) which is still smaller than our \( \varepsilon_{3,k} \) and \( \varepsilon_6 \). Table 5.2 suggests that our bounds \( \varepsilon_{3,j} \) and \( \varepsilon_{4,j} \) are very sharp for \( j = 1, 2, 3 \) for the example. In fact, the first three eigenvalues are very well-approximated by the corresponding Ritz values. But \( \varepsilon_{5,j} \) is not so good because of the factor \( \lambda_{k+1} - \lambda_k = 1/600 \). All \( \varepsilon_{3,j}, \varepsilon_{4,j} \) and \( \varepsilon_{5,j} \) for \( j = 4, 5 \) are about the same. □

6. Concluding remarks

Knyazev and Argentati [4] established several elegant majorization results on the differences between Ritz values and the eigenvalues of interest of a (large scale) Hermitian matrix \( A \) by the canonical angles between/from an invariant subspace of \( A \) and/to a subspace built usually for the purpose of computing the eigenvalues of interest and their corresponding eigenvectors. With these majorization results, bounds are readily available to tell how accurate (some of) the Ritz values may be, given information on the canonical angles. Our main results can be considered as converses to their results in the sense that our results are about bounding the canonical angles by the differences between Ritz values and the eigenvalues of interest, and bounds are readily available from our results to tell how accurate (some of) the Ritz vectors may be, given information on approximation accuracies in the Ritz values.

Recently, there is an increasing trend of using optimization techniques for solving today’s ever-growing large scale symmetric eigenvalue problems through optimizing Rayleigh quotient matrices because these optimization techniques turn to be more memory and computationally efficient – less memory and more matrix–matrix multiplications.
The idea is to force Ritz values move towards eigenvalues. The result here is helpful in understanding how Ritz vectors move towards eigenvectors while Ritz values are made to move towards eigenvalues.

So far we have made the smallest eigenvalues the focus of our investigation. In the case when the largest eigenvalues are the interested ones, one can simply consider $-A$ instead.

Acknowledgements

The authors are grateful to both reviewers for their constructive comments and suggestions that improve the presentation.

References