Journal of Computational Mathematics, Vol.21, No.5, 2003, 657–670.

COMPUTING EIGENVECTORS OF NORMAL MATRICES WITH SIMPLE INVERSE ITERATION *[1)]

Zhen-yue Zhang Tiang-wei Ouyang
(*Department of Mathematics, Zhejiang University, Yuquan Campus, Hangzhou 310027, China*)

Abstract

It is well-known that if we have an approximate eigenvalue $\hat{\lambda}$ of a normal matrix A of order n, a good approximation to the corresponding eigenvector u can be computed by *one* inverse iteration provided the position, say $k_{\max}$, of the largest component of u is known. In this paper we give a detailed theoretical analysis to show relations between the eigenvector u and vector x_k, $k = 1, \cdots, n$, obtained by simple inverse iteration, i.e., the solution to the system $(A - \hat{\lambda}I)x = e_k$ with e_k the kth column of the identity matrix I. We prove that under some weak conditions, the index $k_{\max}$ is of some optimal properties related to the smallest residual and smallest approximation error to u in spectral norm and Frobenius norm. We also prove that the normalized absolute vector $v = |u|/\|u\|_\infty$ of u can be approximated by the normalized vector of $(\|x_1\|_2, \cdots, \|x_n\|_2)^T$. We also give some upper bounds of $|u(k)|$ for those "optimal" indexes such as Fernando's heuristic for $k_{\max}$ without any assumptions. A stable double orthogonal factorization method and a simpler but may less stable approach are proposed for locating the largest component of u.

Key words: Eigenvector, Inverse iteration, Accuracy, Error estimation.

1. Introduction

Let A be a normal matrix of order n. Assume that we have a good approximation $\hat{\lambda}$ to an eigenvalue λ of A, the inverse iteration method

$$(A - \hat{\lambda}I)y_j = z_j, \quad z_{j+1} = y_j/\|y_j\|_\infty$$

is commonly used for computing an eigenvector u of A corresponding λ approximately. In general, the starting vector $z_0 = b$ is chosen at random or to be the vector of all one's and the iteration process converges in several steps [1]. However, there are no practical ways to choose a starting vector b that ensures the rapid convergence, though it is true in theory that one can get an accurate eigenvector to working precision by *a single inverse iteration* if the right vector b is reasonably chosen [7]. In [12], Wilkinson pointed out that for symmetric traditional matrix A, a solution to the homogenous system $(A-\hat{\lambda}I)x = 0$, discarding one of the n equations, say the kth one, will be a good approximation to the eigenvector u provided the kth component $u(k)$ of u is not small. Equivalently, such an approximation, say x_k, can be obtained by one step of inverse iteration $(A - \hat{\lambda}I)x = e_k$ for a properly chosen index k, for example $k = k_{\max}$ corresponding to the largest component $u(k_{\max})$ of u in absolute value. Actually if $u(k)$ is the largest one in absolute value or above average in magnitude, the normalized output $x_k/\|x_k\|$ of a single inverse iteration will yield a residual which archives the *optimal* accuracy in magnitude. (See Corollary 4.1 for details.) It means that the *simple inverse iteration*, a single inverse iteration with right

* Received July 20, 2001; Final revised December 5, 2002.
[1)] The work of this author was supported in part by NSFC (project 19771073), the Special Funds for Major State Basic Research Projects of China (project G19990328), Zhejiang Provincial Natural Science Foundation of China, and Foundation for University Key Teacher by the Ministry of Education, China.

vector $b = e_k$, will give an acceptable approximate eigenvector if the index k is chosen well. Therefore there are two related problems that need to be considered: 1) how to locate the largest component $u(k_{\max})$ of the eigenvector u and, 2) if an index k is approximately estimated to $k_{\max}$, how large the component $u(k)$ is or how close it is to $u(k_{\max})$ in absolute value. In [3], an index corresponding to the largest diagonal entry of the inverse of matrix $A - \hat{\lambda}I$ was suggested as an heuristic for choosing the "optimal" index $k_{\max}$. (The index determined by the heuristic will be denoted as k_d in this paper.) Parlett and Dhillon [9] shown that k_d is asymptotically equal to $k_{\max}$ as $\hat{\lambda}$ tends to the eigenvalue λ. In this paper, we will furthermore discuss such problems for a real symmetric or more generally, normal matrix A by a detailed componentwise analysis of the output $x = x_k$. As shown later, under some weak conditions the index $k_{\max}$ is of some optimal properties such that among all normalized vectors x_k^*, $x_{k_{\max}}^*$ achieves the minimum of residuals both in 2-norm and in ∞-norm. In general, for indexes with some optimal properties, for example $k = k_d$, the corresponding component $u(k)$ is the largest one of u with a factor tightly close to one. For those indexes k corresponding to small components $|u(k)|$, the normalized vector x_k^* may be not a good approximation to u, but the position of its largest component in absolute value also implies the position of large component of u, provided $|u(k)|$ is not small enough. On the other hand, the normalized absolute-valued vector $|u|/\|u\|_\infty$ can also be approximated by the normalized vector of the norm vector $(\|x_1\|, \cdots, \|x_n\|)^T$.

Fernando's approach for determining the index k_d is an application of double factorization (a combination of LDU and UDL factorizations) of the nearly singular tridiagonal matrix $A - \hat{\lambda}I$. (Cf. [9] for careful discussions of the relation between the double triangular factorization and the related eigenvector algorithms.) However double factorization is unstable and the slight danger of overflow and/or underflow still exits. We will propose an orthogonal double factorization based upon QR and QL decompositions to determine k_d stably.

This paper is organized as follows: In Section 2, we first review some error bounds of the residual $\|Ax - \tilde{\lambda}x\|_2$ and the error $\|x - u\|_2$ of the approximate eigenvector x computed by a single inverse iteration with respect to the right vector b. As a deduction, error bounds for x_k obtained by simple inverse iteration are also given. In Section 3 we discuss some optimal properties of $x_{k_{\max}}$ that implies information of locating the largest component of u. A lower bound in terms of $u(k_{\max})$ for the component $u(k_d)$ will be given in Section 4, which shows that $u(k_d)$ is always the largest component of u with a factor tightly close to one. We also shown a simpler way to locating largest component of u. The double orthogonal factorization for determining k_d is proposed in Section 5.

Notations. We define by $\{\lambda_j\}$ the set of eigenvalues of matrix A and by $\{u_j\}$ the corresponding eigenvectors with $\|u_j\|_2 = 1$. The eigenvalue λ_i satisfying $|\lambda_i - \hat{\lambda}| = \min_j |\lambda_j - \hat{\lambda}|$ will be simply denoted as λ. Generally, we always assume that λ is uniquely determined, i.e., if $\lambda_j \neq \lambda$, then $|\lambda_j - \hat{\lambda}| > |\lambda - \hat{\lambda}|$. V_λ denotes the eigenspace spanned by the eigenvectors u_i corresponding to $\lambda_i = \lambda$. (The eigenvalue λ may be multiple.) Specially, if λ is a single eigenvalue, $V_\lambda = \text{spann}\{u\}$, where $u = u_i$. It is also assumed that $\hat{\lambda}$ is not an exact eigenvalue of A. x^H means the conjugate transpose of x and, as we have used, $x(k)$ is the k-th component of vector x.

2. A Review on Inverse Iteration

We focus on a single inverse iteration, i.e., an inverse iteration is viewed as a "direct" method for computing approximately eigenvectors rather than an iteration approach. The nature problem is thus that how good the approximation gotten by one iteration

$$(A - \hat{\lambda}I)x = b \tag{2.1}$$

is for a certain right vector b chosen in practical. To that end, let us first review a well-known

result for the "*optimal*" residuals of approximate vectors when the approximate eigenvalue $\hat{\lambda}$ is fixed.

Theorem 2.1. *Let A be a normal matrix of order n, and $\{\lambda_j\}$ its eigenvalues. Then for a given $\hat{\lambda}$,*

$$\min_{\|x\|_2=1} \|Ax - \hat{\lambda}x\|_2 = \min_j |\lambda_j - \hat{\lambda}|.$$

Proof. Theorem 4.5.1 of [10] gives the lower bound of the residual $\|Ax-\hat{\lambda}x\|_2 \geq \min_j |\lambda_j-\hat{\lambda}|$. The bound is achieved at the eigenvector u corresponding to the eigenvalue λ satisfying $|\lambda-\hat{\lambda}| = \min_j |\lambda_j - \hat{\lambda}|$, completing the proof.

Theorem 2.1 implies what we can expect on approximating an eigenvector $u \in V_\lambda$. In practical, an *optimal* approximation of u is a vector with residual norm $|\lambda - \hat{\lambda}|$ in magnitude, though a smaller approximation error $\|x-u\|_2$ is also possible. The following theorem [7] shows that the optimal approximate magnitude can be achieved by a single inverse iteration if the right vector b is chosen well.

Theorem 2.2. *Let $|\lambda - \hat{\lambda}| = \min_i |\lambda_i - \tilde{\lambda}|$, and let $x^* = x/\|x\|_2$ be the normalized vector of x computed by a single inverse iteration (2.1) with $\|b\|_2 = 1$. Then*

$$|\lambda - \hat{\lambda}| \leq \|Ax^* - \hat{\lambda}x^*\|_2 \leq \frac{|\lambda - \hat{\lambda}|}{|\cos\langle b, V_\lambda\rangle|}, \tag{2.2}$$

where $\langle b, V_\lambda\rangle$ is the angle between vector b and subspace V_λ.

Proof. The left inequality of (2.2) holds obviously. To show the right one, we write vector b in the form

$$b = \cos\langle b, V_\lambda\rangle u + \sum_{\lambda_j \neq \lambda} (b^H u_j) u_j,$$

where $\cos\langle b, V_\lambda\rangle u$ is an orthogonal projection of b in the subspace V_λ with a normal vector $u \in V_\lambda$. It follows that

$$\|x\|_2 = \|(A - \hat{\lambda}I)^{-1} b\|_2 = \left(\frac{\cos^2\langle b, V_\lambda\rangle}{|\lambda - \hat{\lambda}|^2} + \sum_{\lambda_j \neq \lambda} \left| \frac{b^H u_j}{\lambda_j - \hat{\lambda}} \right|^2 \right)^{1/2} \geq \frac{|\cos\langle b, V_\lambda\rangle|}{|\lambda - \hat{\lambda}|}.$$

Therefore we have with $\|Ax - \hat{\lambda}x^*\|_2 = \|b\|_2 = 1$ that

$$\|Ax^* - \hat{\lambda}x^*\|_2 = \frac{1}{\|x\|_2} \leq \frac{|\lambda - \hat{\lambda}|}{|\cos^2\langle b, V_\lambda\rangle|}, \tag{2.3}$$

completing the proof.

Furthermore we have a upper bound of the approximation err of x^* to an eigenvector $u \in V_\lambda$.

Theorem 2.3. *Denote* $\text{gap} = \min_{\lambda_j \neq \lambda} |\lambda_j - \hat{\lambda}|$. *Then under the assumption of Theorem 2.2 we have*

$$\|x^* - su\|_2 \leq \frac{|\lambda - \hat{\lambda}|}{\text{gap}} \tan\langle b, V_\lambda\rangle, \tag{2.4}$$

where s satisfying $|s| = 1$ is a properly chosen constant determined by (2.8).

Proof. As done in the proof of Theorem 2.2, we write

$$x = \frac{\cos\langle b, V_\lambda\rangle}{\lambda - \hat{\lambda}} u + \sum_{\lambda_j \neq \lambda} \frac{b^H u_j}{\lambda_j - \hat{\lambda}} u_j. \tag{2.5}$$

Taking squares of spectrum norms on two sides of (2.5) gives

$$\begin{aligned} \|x\|_2^2 &= \left| \frac{\cos\langle b, V_\lambda\rangle}{\lambda - \hat{\lambda}} \right|^2 + \sum_{\lambda_j \neq \lambda} \left| \frac{b^H u_j}{\lambda_j - \hat{\lambda}} \right|^2 \\ &\leq \left| \frac{\cos\langle b, V_\lambda\rangle}{\lambda - \hat{\lambda}} \right|^2 + \frac{1 - |\cos\langle b, V_\lambda\rangle|^2}{\mathrm{gap}^2}. \end{aligned} \tag{2.6}$$

Then we multiply by $|\lambda - \hat{\lambda}|^2$ the two sides of the equation above and take square roots. It yields that

$$\begin{aligned} |\lambda - \hat{\lambda}| \|x\|_2 &\leq \sqrt{|\cos\langle b, V_\lambda\rangle|^2 + \frac{|\lambda - \hat{\lambda}|^2}{\mathrm{gap}^2}\left(1 - |\cos\langle b, V_\lambda\rangle|^2\right)} \\ &= |\cos\langle b, V_\lambda\rangle| \sqrt{1 + \frac{|\lambda - \hat{\lambda}|^2}{\mathrm{gap}^2} \tan^2\langle b, V_\lambda\rangle}. \end{aligned} \tag{2.7}$$

On the other hand, by (2.5) we have that

$$x^* - su = \left(\frac{\cos\langle b, V_\lambda\rangle}{(\lambda - \hat{\lambda})\|x\|_2} - s \right) u + \frac{1}{\|x\|_2} \sum_{\lambda_j \neq \lambda} \frac{b^H u_j}{\lambda_j - \hat{\lambda}} u_j.$$

Taking spectrum norms again and writing

$$\frac{\cos\langle b, V_\lambda\rangle}{\lambda - \hat{\lambda}} = s \left| \frac{\cos\langle b, V_\lambda\rangle}{\lambda - \hat{\lambda}} \right| \tag{2.8}$$

with $|s| = 1$ gives

$$\begin{aligned} \|x^* - su\|_2^2 &= \left| \frac{\cos\langle b, V_\lambda\rangle}{(\lambda - \hat{\lambda})\|x\|_2} - s \right|^2 + \frac{1}{\|x\|_2^2} \sum_{\lambda_j \neq \lambda} \left| \frac{b^H u_j}{\lambda_j - \hat{\lambda}} \right|^2 \\ &= \left| \frac{|\cos\langle b, V_\lambda\rangle|}{|\lambda - \hat{\lambda}| \|x\|_2} - 1 \right|^2 + \frac{1}{\|x\|_2^2} \left(\|x\|_2^2 - \left| \frac{\cos\langle b, V_\lambda\rangle}{\lambda - \hat{\lambda}} \right|^2 \right) \\ &= 2\left(1 - \frac{|\cos\langle b, V_\lambda\rangle|}{|\lambda - \hat{\lambda}| \, \|x\|_2} \right). \end{aligned} \tag{2.9}$$

Therefore substituting (2.7) into the last equality above we have

$$\begin{aligned} \|x^* - su\|_2^2 &\leq 2\left(1 - \left(1 + \frac{|\lambda - \hat{\lambda}|^2}{\mathrm{gap}^2} \tan^2\langle b, V_\lambda\rangle \right)^{-\frac{1}{2}} \right) \\ &\leq \left(\frac{|\lambda - \hat{\lambda}|}{\mathrm{gap}} \tan\langle b, V_\lambda\rangle \right)^2. \end{aligned} \tag{2.10}$$

Hence the inequality (2.4) holds.

Clearly if the eigenvector $u \in V_\lambda$ is defined above, then $\langle b, V_\lambda\rangle = \langle b, u\rangle$. For simplicity, we always assume that λ is simple, i.e., $V_\lambda = \text{spann}\{u\}$. Similar results are also true for multiple λ.

Remark. Since $|\sin\langle x^*, u\rangle| \le \|x^* - su\|_2$, it follows from Theorem 2.3 that

$$|\sin\langle x^*, u\rangle| \le \frac{|\lambda - \hat{\lambda}|}{\text{gap}} |\tan\langle b, u\rangle|.$$

This estimation also follows from the traditional convergence analysis of power method applying $(A - \hat{\lambda}I)^{-1}$ [6, p.404], or Davis and Kahan' estimation [2] for $|\sin\langle y, u\rangle|$ in terms of the residual $\|Ay - \hat{\lambda}y\|_2$ and Theorem 2.2.

Theorems 2.2 and 2.3 clearly show that a good approximation to the eigenvector u can be obtained by one step of inverse iteration if $\cos\langle b, u\rangle$ is not small enough. However there are no practical ways to chose such vector b, although a randomly chosen vector or $b = e$ with unit components are commonly used [1, 8]. Note that if A is symmetric tridiagonal matrix with positive off diagonals and λ is the (single) extreme eigenvalue (the largest or smallest one) of A, the corresponding eigenvector u is positive or negative [13]. It yields immediately that

$$|\cos\langle e, u\rangle| \ge \frac{1}{\sqrt{n}}, \quad |\tan\langle e, u\rangle| \le \sqrt{n-1},$$

because for positive reals $t_1, \cdots, t_n$ satisfying $\sum_{i=1}^n t_i^2 = 1$, we have $\sum_{i=1}^n t_i = 1$. Therefore a single inverse iteration with $b = e$ can give a good approximation provided $\hat{\lambda}$ is closer to the largest or smallest eigenvalue than to others. For general cases, Wilkinson suggested $b = PLe$, where L is th unit lower triangular matrix of the LU decomposition $A - \hat{\lambda}I = PLU$ with partial column pivoting. Obviously, the vector x solving (2.3) also satisfies $Ux = e$ which can be solved easily by a back-substitution. However it is possible that $\cos\langle b, u\rangle$ is relatively small for such chosen b.

Simple Inverse Iteration. A simpler way is to chose $b = e_k$, the kth column of the identity matrix I. Using $b = e_k$ has the benefit for tridiagonal matrix $A = [\beta_{k-1}, \alpha_k, \beta_k]$ that the tridiagonal system

$$(A - \hat{\lambda}I)x = e_k \tag{2.11}$$

can be easily solved by following down and upper recursion process [5]: First set $\hat{x}(1) = 1$ and determine $\hat{x}(2)$ from the first equation of (2.11), and determine $\hat{x}(3)$ from the second equation and so on, until $\hat{x}(k)$ is computed when the $(k-1)$th equation is used. Similarly form the last $n - k$ equations, we can determine, step by step, $\tilde{x}(n-1)$, $\cdots$, $\tilde{x}(k)$, by setting $\tilde{x}(n) = 1$. Choosing constant c such that $\hat{x}(k) = c\tilde{x}(k)$ and setting

$$\eta_k = \beta_k \hat{x}(k-1) + (\alpha_k - \hat{\lambda})\hat{x}(k) + \beta_{k+1} c\tilde{x}(k+1)$$

gives the solution

$$x_k = (\hat{x}(1), \cdots, \hat{x}(k), c\tilde{x}(k+1), \cdots, c\tilde{x}(n))^T / \eta_k.$$

This approach may give an approximate eigenvector with higher relative accuracy in components if k is certainly chosen. Below is such an example.

Example. Let $n = 200$. The off-diagonals of symmetric tridiagonal matrix A are set to be $1/2$ while the diagonal entries read

$$\alpha(j) = 1 - s(j)s(j+1) - s(j-1)s(j)/4, \quad j = 1, \cdots, n$$

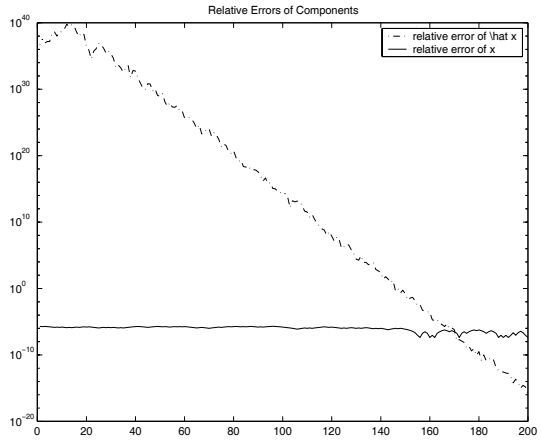

Figure 1: Component-wise relative errors of computed eigenvectors.

with $s(0) = s(n+1) = 0$ and $s(j) = \pm 1$ for $j = 1, \cdots, n$. It can be easily verified that $\lambda = 1$ is an *exact* eigenvalue and the corresponding unnormalized eigenvector u is given by

$$u(j) = s(j)2^j, \quad j = 1, \cdots, n.$$

Clearly u has its largest component $u(n)$ in abstract value, i.e., $k_{\max} = n$. Now we set $\hat{\lambda} = 1 + 10^{-7}$ and $k = n$. All the *relative* errors of components of the computed approximate vector $x = x_k$ are less than 10^{-8} when u and x are normalized. Note that the eigenvector $\hat{x}$ computed by MATLAB function `schur` has quite large relative errors in the most of components. Fig 1.1 plots the component-wise relative errors of the approximations computed by the down and upper recursion (solid line) and MATLAB function `schur` (dashdot line), respectively.

New let's return to the discussion about vectors x_k for normal matrix A. Note that $|\cos\langle e_k, u\rangle| = |u(k)|$. Upper bounds for the x_k's residuals and the approximation errors can be easily obtained by simply applying Theorems 2.2 and 2.3 with $b = e_k$.

Corollary 2.1. *Let A be a normal matrix and x_k the solution of (2.11) with normalized $x_k^* = x_k/\|x_k\|_2$. Then for a suitable constant s_k satisfying $|s_k| = 1$*

$$\|Ax_k^* - \hat{\lambda}x_k^*\|_2 \leq \frac{|\lambda - \hat{\lambda}|}{|u(k)|}, \tag{2.12}$$

$$\|x_k^* - s_k u\|_2 \leq \frac{|\lambda - \hat{\lambda}|}{\text{gap}} \frac{\sqrt{1 - |u(k)|^2}}{|u(k)|}. \tag{2.13}$$

Naturally, an index achieving the minimums of the upper bounds indicates the largest component of u. We denote such an index by $k_{\max}$ (though it may be not unique),

$$|u(k_{\max})| = \max_k |u(k)| = \|u\|_\infty. \tag{2.14}$$

Obviously, $\|u\|_\infty \geq 1/\sqrt{n}$. The following upper bounds are directly obtained by Corollary 2.1.

$$\|Ax_{k_{\max}}^* - \hat{\lambda}x_{k_{\max}}^*\|_2 \leq \frac{|\lambda - \hat{\lambda}|}{\|u\|_\infty} \leq \sqrt{n}|\lambda - \hat{\lambda}|, \tag{2.15}$$

$$\|x_{k_{\max}}^* - s_{k_{\max}} u\|_2 \leq \frac{|\lambda - \hat{\lambda}|}{\text{gap}} \frac{\sqrt{1 - \|u\|_\infty^2}}{\|u\|_\infty} \leq \sqrt{n-1}\frac{|\lambda - \hat{\lambda}|}{\text{gap}}. \tag{2.16}$$

In the following section, we will show that the index $k_{\max}$ is of some optimal properties.

3. Optimal Properties of $k_{\max}$

We denote by $K_{\max}$ the set of all the indexes $k_{\max}$ satisfying (2.14),

$$K_{\max} = \{\, k \,\big|\, |u(k)| = \|u\|_\infty \,\},$$

and by $K_{\max}^C$ the set of all remainder indices. Recalling that $\|Ax_k - \hat{\lambda}x_k\|_2 = 1$, the residual of x_k^* reads $\|Ax_k^* - \hat{\lambda}x_k^*\|_2 = 1/\|x_k\|_2$. Therefore the smallest residual $\min_k \|Ax_k^* - \hat{\lambda}x_k^*\|_2$ is achieved at $k = k_{\text{opt}}$ which is defined by

$$x_{k_{\text{opt}}}\|_2 = \max_k \|x_k\|_2,$$

i.e.,

$$\min_k \frac{\|Ax_k - \hat{\lambda}x_k\|_2}{\|x_k\|_2} = \frac{\|Ax_{k_{\text{opt}}} - \hat{\lambda}x_{k_{\text{opt}}}\|_2}{\|x_{k_{\text{opt}}}\|_2}. \tag{3.1}$$

So does it for the ∞-norm,

$$\min_k \frac{\|Ax_k - \hat{\lambda}x_k\|_\infty}{\|x_k\|_\infty} = \frac{\|Ax_{k_\infty} - \hat{\lambda}x_{k_\infty}\|_\infty}{\|x_{k_\infty}\|_\infty}, \tag{3.2}$$

where k_∞ is defined by

$$\|x_{k_\infty}\|_\infty = \max_k \|x_k\|_\infty.$$

In this section, we will show that under some weak conditions, the vector $x_{k_{\max}}$ for $k_{\max} \in K_{\max}$ also achieve the minimum of residuals both in 2-norm and in ∞-norm. To this end, let us first consider the ∞-norm of x_k, starting at comparing the sizes of vectors x_k. We denote by

$$\eta = \frac{|\lambda - \hat{\lambda}|}{\text{gap}} = \max_{\lambda_j \neq \lambda} \frac{|\lambda - \hat{\lambda}|}{|\lambda_j - \hat{\lambda}|} < 1. \tag{3.3}$$

Theorem 3.1. *For each $k = 1, \cdots, n$, if*

$$|u(k)| < \|u\|_\infty - \frac{1 + \sqrt{1 - \|u\|_\infty^2}}{\|u\|_\infty}\eta, \tag{3.4}$$

then

$$\|x_k\|_\infty < \min_{j \in K_{\max}} \|x_j\|_\infty. \tag{3.5}$$

Proof. We write

$$x_k = \frac{u(k)}{\lambda - \hat{\lambda}}u + \sum_{\lambda_j \neq \lambda} \frac{u_j(k)}{\lambda_j - \hat{\lambda}}u_j \equiv \frac{1}{\lambda - \hat{\lambda}}\big(u(k)u + v_k\big). \tag{3.6}$$

It is easy to show that

$$\|v_k\|_2 = \left\| \sum_{\lambda_j \neq \lambda} \frac{\lambda - \hat{\lambda}}{\lambda_j - \hat{\lambda}} u_j(k) u_j \right\|_2 \leq \eta \sqrt{\sum_{\lambda_j \neq \lambda} |u_j(k)|^2} = \eta\sqrt{1 - |u(k)|^2}. \tag{3.7}$$

Taking ∞-norms on the two sides of (3.6) yields that

$$\|x_k\|_\infty = \frac{1}{|\lambda - \hat{\lambda}|}\big(|u(k)|\|u\|_\infty + \xi_k\big) \tag{3.8}$$

with

$$|\xi_k| \le \eta\sqrt{1 - |u(k)|^2}. \tag{3.9}$$

It follows that for each $k \in K_{\max}^C$ and $k_{\max} \in K_{\max}$,

$$\begin{aligned} \|x_{k_{\max}}\|_\infty - \|x_k\|_\infty &\ge \frac{1}{|\lambda - \hat{\lambda}|}\Big((|u(k_{\max})| - |u(k)|)\|u\|_\infty + \xi_{k_{\max}} - \xi_k\Big) \\ &\ge \frac{1}{|\lambda - \hat{\lambda}|}\Big((\|u\|_\infty - |u(k)|)\|u\|_\infty - \eta(1 + \sqrt{1 - \|u\|_\infty^2})\Big) \\ &> 0. \end{aligned}$$

The last inequality above follows from the condition (3.4). Therefore (3.5) holds, completing the proof.

A similar result for 2-norm is also true.

Theorem 3.2. *If $|u(k)|^2 < \|u\|_\infty^2 - \eta^2$, then*

$$\|x_k\|_2 < \min_{j \in K_{\max}} \|x_j\|_2. \tag{3.10}$$

From Theorems 3.1 and 3.2 we obtain the following important results.

Theorem 3.3. *If $\eta \le \|u\|_\infty$, then*

$$\|u\|_\infty - \frac{1 + \sqrt{1 - \|u\|_\infty^2}}{\|u\|_\infty}\eta \le |u(k_\infty)| \le \|u\|_\infty. \tag{3.11}$$

$$\sqrt{\|u\|_\infty^2 - \eta^2} \le |u(k_{\text{opt}})| \le \|u\|_\infty. \tag{3.12}$$

Corollary 3.1. *Denote by γ the second largest component of u,*

$$\gamma = \max_{k \notin K_{\max}} |u(k)|.$$

$$If \quad \gamma < \|u\|_\infty - \frac{1 + \sqrt{1 - \|u\|_\infty^2}}{\|u\|_\infty}\eta, \quad then \quad |u(k_\infty)| = \|u\|_\infty. \tag{3.13}$$

$$If \quad \gamma < \sqrt{\|u\|_\infty^2 - \eta^2}, \quad then \quad |u(k_{\text{opt}})| = \|u\|_\infty. \tag{3.14}$$

It can be verified that if $\eta \le \|u\|_\infty$, then

$$\|u\|_\infty - \frac{1 + \sqrt{1 - \|u\|_\infty^2}}{\|u\|_\infty}\eta \le \sqrt{\|u\|_\infty^2 - \eta^2}.$$

It is more interesting that if $\gamma < \|u\|_\infty - \frac{1+\sqrt{1-\|u\|_\infty^2}}{\|u\|_\infty}\eta$, then

$$|u(k_\infty)| = |u(k_{\text{opt}})| = |u(k_{\max})|,$$

which implies that

$$k_\infty = k_{\text{opt}} = k_{\max}$$

if $k_{\max}$ is unique. In that case,

$$\|x_{k_{\max}}\|_\infty = \max_k \|x_k\|_\infty \quad \text{and} \quad \|x_{k_{\max}}\|_2 = \max_k \|x_k\|_2.$$

or equivalently, $k_{\max}$ minimizes the residuals in 2-norm and ∞-norm. The following theorem further shows that $k_{\max}$ maximizes the components of $x_{k_{\max}}$, too,

$$|x_{k_{\max}}(k_{\max})| = \max_k |x_{k_{\max}}(k)|.$$

Theorem 3.4. *Let γ be defined in Corollary 3.1. If*

$$\gamma \le \|u\|_\infty - \frac{2\sqrt{1-\|u\|_\infty^2}}{\|u\|_\infty}\eta, \tag{3.15}$$

then for $k \in K_{\max}^C$ and $k_{\max} \in K_{\max}$

$$|x_{k_{\max}}(k)| < |x_{k_{\max}}(k_{\max})|. \tag{3.16}$$

Proof. By (3.6) we get that for each k

$$x_{k_{\max}}(k) = \frac{1}{\lambda - \hat{\lambda}}\big(u(k_{\max})u(k) + v_{k_{\max}}(k)\big).$$

Especially, for $k = k_{\max} \in K_{\max}$

$$x_{k_{\max}}(k_{\max}) = \frac{1}{\lambda - \hat{\lambda}}\big(u(k_{\max})^2 + v_{k_{\max}}(k_{\max})\big).$$

It follows that for $k \in K_{\max}^C$ and $k_{\max} \in K_{\max}$

$$\begin{aligned} |x_{k_{\max}}(k_{\max})| - |x_{k_{\max}}(k)| &\ge \frac{1}{|\lambda - \hat{\lambda}|}\big((\|u\|_\infty - \gamma)\|u\|_\infty - (|v_{k_{\max}}(k_{\max})| + |v_{k_{\max}}(k)|\big) \\ &\ge \frac{1}{|\lambda - \hat{\lambda}|}\big((\|u\|_\infty - \gamma)\|u\|_\infty - \sqrt{2}\|v_{k_{\max}}\|_2\big). \end{aligned}$$

By (3.7) we have

$$\frac{1}{|\lambda - \hat{\lambda}|}\big((\|u\|_\infty - \gamma)\|u\|_\infty - \sqrt{2}\eta\sqrt{1-\|u\|_\infty^2}\big) > 0,$$

completing the proof.

4. Approximately Choice of $k_{\max}$

To locate the largest component of u, or equivalently determine the index $k_{\max}$, Fernando [3] used the index maximizing the diagonal entries of the inverse of matrix $A - \hat{\lambda}I$ as a heuristic for $k_{\max}$. We denote by k_d such index, i.e.,

$$|x_{k_d}(k_d)| = \max_k |(A - \hat{\lambda}I)^{-1}(k,k)| = \max_k |x_k(k)|. \tag{4.1}$$

This is a reasonable choice because be the discussion in the last section, if $k_{\max}$ is unique and the inequality

$$\gamma < \|u\|_\infty - \frac{1+\sqrt{1-\|u\|_\infty^2}}{\|u\|_\infty}\eta$$

holds, then $k_{\max}$ also maximizes $\{|x_k(k)|\}$, i.e., $k_d = k_{\max}$. In this section we will further show some lower bounds for $|u(k_d)|$ in terms of $\|u\|_\infty$ without any assumption. Those bounds will make sure that $x^*_{k_d}$ always achieves the optimal bounds shown in (2.15) and (2.16).

Theorem 4.1. *Let η be defined in (3.3). Then*

$$|u(k_d)|^2 \geq |u(k_{\max})|^2 - \frac{2\eta}{1-\eta}(1 - |u(k_{\max})|^2). \tag{4.2}$$

Proof. Using the expression (3.6) of x_k we have

$$|x_k(k)| = \frac{1}{\lambda - \hat{\lambda}}\big(|u(k)|^2 + \zeta_k\big), \quad |\zeta_k| \leq \eta(1 - |u(k)|^2).$$

Hence, applying the equality with $k = k_{\max}$ and $k = k_d$ yields that

$$\begin{aligned} 0 &\leq \big|x_{k_d}(k_d)\big| - \big|x_{k_{\max}}(k_{\max})\big| \\ &= \frac{1}{|\lambda - \hat{\lambda}|}\big(|u(k_d)|^2 + \zeta_{k_d} - |u(k_{\max})|^2 - \zeta_{k_{\max}}\big) \\ &\leq \frac{1}{|\lambda - \hat{\lambda}|}\big(|u(k_d)|^2 - |u(k_{\max})|^2 + \eta(2 - |u(k_{\max})|^2 - |u(k_d)|^2)\big) \\ &= \frac{1}{|\lambda - \hat{\lambda}|}\big((|u(k_d)|^2 - |u(k_{\max})|^2)(1-\eta) + 2\eta(1 - |u(k_{\max})|^2)\big). \end{aligned}$$

The estimation (4.2) follows immediately.

Remark. It is easy to verify that

$$\left|\frac{u(k_d)}{u(k_{\max})}\right|^2 \geq 1 - \frac{2\eta}{1-\eta}\frac{1 - |u(k_{\max})|^2}{|u(k_{\max})|^2} \geq 1 - \frac{2\eta}{1-\eta}(n-1).$$

Generally we always have that $k_d = k_{\max}$ or $|u(k_d)| = |u(k_{\max})|$.

Now denoting

$$\alpha = \left(1 - \frac{2\eta}{1-\eta}\frac{1 - |u(k_{\max})|^2}{|u(k_{\max})|^2}\right)^{-1/2} \approx 1,$$

$$\beta = \left(1 - \frac{2\eta}{1+\eta}\frac{1}{|u(k_{\max})|^2}\right)^{-1/2} = \sqrt{\frac{1+\eta}{1-\eta}}\alpha \approx 1,$$

we have upper bounds for the residual $\|Ax^*_{k_d} - \hat{\lambda}x^*_{k_d}\|_2$ and the approximate error $\|x^*_{k_d} - s_{k_d}u\|_2$ in terms of the optimal bounds shown in (2.15) and (2.16).

Corollary 4.1.

$$\|Ax^*_{k_d} - \hat{\lambda}x^*_{k_d}\|_2 \leq \frac{|\lambda - \hat{\lambda}|}{|u(k_{\max})|}\alpha, \tag{4.3}$$

$$\|x^*_{k_d} - s_{k_d}u\|_2 \leq \eta\frac{\sqrt{1 - |u(k_{\max})|^2}}{|u(k_{\max})|}\beta, \tag{4.4}$$

where s_{k_d} is that as defined in Corollary 2.1 for $k = k_d$.

Proof. By Theorem 4.1 and the definition of α and β, we have that

$$|u(k_d)|^2 \geq |u(k_{\max})|^2/\alpha^2$$

and

$$1 - |u(k_d)|^2 \le \frac{1+\eta}{1-\eta}(1 - |u(k_{\max})|^2).$$

Applying Corollary 2.1 for $k = k_d$ yields immediately the inequalities (4.3) and (4.4).

Corollary 4.1 shows that x_{k_d} also gives the same residual and approximate error as those given by $x_{k_{\max}}$ up to factors α and β, respectively, which close tightly to one.

A much simpler but less stable way of locating the largest component of u is to use a single vector x_k directly. Though the normalized vector x_k^* may not a good approximation of u if $u(k)$ is not greater than the average in magnitude, it also provides information about the position $k_{\max}$ of the largest component if $u(k)$ is not small enough.

Theorem 4.2. *Let* $|x_k(i_k)| = \max_i |x_k(i)|$. *If* $|u(k)| \ge \sqrt{2}\eta/\|u\|_\infty$, *then*

$$|u(i_k)| \ge |u(k_{\max})| - \frac{\sqrt{2}\eta}{|u(k)|}.$$

Proof. By (3.6),

$$0 \le |x_k(i_k)| - |x_k(k_{\max})| \le \frac{1}{|\lambda - \hat{\lambda}|}\Big(|u(k)|\big(|u(i_k)| - |u(k_{\max})|\big) + \sqrt{2}\eta\Big).$$

It yields the result required.

Below is a small example which substantiates the statement above.

Example. The symmetric matrix A is constructed as follows.

$$\begin{aligned}
&[\mathtt{Q}, \mathtt{R}] = \mathtt{qr}(1 - 2 * \mathtt{rand}(\mathtt{n}));\\
&\mathtt{s} = 10 - 20 * \mathtt{rand}(\mathtt{n}, 1);\\
&\mathtt{A} = \mathtt{Q} * \mathtt{diag}(\mathtt{s}) * \mathtt{Q}^{\mathtt{T}};
\end{aligned}$$

with $n = 100$. We set $\hat{\lambda} = s(1) * (1 + \tau)$ using $\tau = 1.e - 6$. For such an example, the *smallest* component of u in absolute value is $|u(86)| = 1.2574e - 5$ and $\eta = 2.2201e - 5$. If we unfortunately choose $k = 86$, x_k^* gives a large residual $\|Ax_k^* - \hat{\lambda}x_k^*\|_2 = 0.0590$. However, with the index i_k locating the largest component of x_k, $|u(i_k)| = 0.1676$ is the third largest component of u. (The two largest components are 0.1732 and 0.1718.) $x_{i_k}^*$ is almost the optimal approximation to u with residual $\|Ax_{i_k}^* - \hat{\lambda}x_{i_k}^*\|_2 = 4.4470e - 6$ close tightly to the smallest one $\|Ax_{k_{\max}}^* - \hat{\lambda}x_{k_{\max}}^*\|_2 = 4.3019e - 6$. Figure 2.1 plots the components $u(k)$, residuals $\|Ax_k^* - \hat{\lambda}x_k^*\|_2$, and $\|Ax_{i_k}^* - \hat{\lambda}x_{i_k}^*\|_2$ for all k's.

Finally we finish this section by showing the following interesting result about component-wise approximation.

Theorem 4.3. *Let* $\|\cdot\|$ *be the* 2*-norm or* ∞*-norm. If* $\eta < \|u\|_\infty\|u\|$, *then*

$$\left|\frac{\|x_k\|}{\max_k \|x_k\|} - \frac{|u(k)|}{\|u\|_\infty}\right| \le \frac{2\eta}{\|u\|_\infty\|u\| - \eta}.$$

Proof. As that in (3.8), we write

$$\|x_k\| = \frac{1}{|\lambda - \hat{\lambda}|}\big(|u(k)|\,\|u\| + \delta_k\big), \quad |\delta_k| \le \eta\sqrt{1 - u(k)^2}.$$

and have

$$\max_k \|x_k\| = \frac{1}{|\lambda - \hat{\lambda}|}\big(\|u\|_\infty\|u\| + \delta\big), \quad |\delta| \le \eta.$$

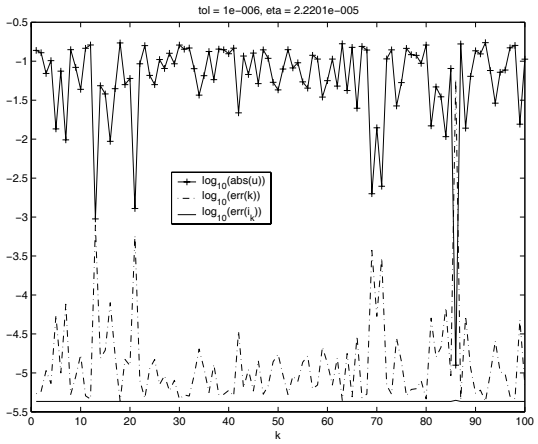

Figure 2: Components of u $(-+-)$, residuals $(-.)$ and refined residuals $(-)$.

It follows that

$$\begin{aligned}\left|\frac{\|x_k\|}{\max_k\|x_k\|}-\frac{|u(k)|}{\|u\|_\infty}\right| &= \left|\frac{|u(k)|\,\|u\|+\delta_k}{\|u\|_\infty\|u\|+\delta}-\frac{|u(k)|}{\|u\|_\infty}\right| \\ &= \left|\frac{\delta_k\|u\|_\infty-\delta|u(k)|}{(\|u\|_\infty\|u\|+\delta)\|u\|_\infty}\right| \\ &= \left|\frac{\eta(\|u\|_\infty+|u(k)|)}{(\|u\|_\infty\|u\|-\eta)\|u\|_\infty}\right| \\ &\le \frac{2\eta}{\|u\|_\infty\|u\|-\eta}.\end{aligned}$$

5. Determination of k_d

For determining k_d, Fernando considered the following nearly homogeneous set of equations [3]

$$(A-\hat{\lambda}I)y_k=\mu_k e_k,\quad y_k(k)=1 \tag{5.1}$$

with real number μ_k. Obviously, $y_k=\mu_k x_k$ with x_k defined in (3.6), and hance $x_k(k)=1/\mu_k$. Therefore the index k minimizing $|\mu_k|$ is just k_d defined in (4.1). Fernando's algorithm of computing μ_k bases upon the LDU and UDL factorizations (double factorization)

$$L_+\mathrm{diag}(d_+)U_+=A-\hat{\lambda}I=U_-\mathrm{diag}(d_-)L_-$$

of the shifted matrix $A-\hat{\lambda}I$ when A is tridiagonal. There are several formulae of evaluating μ_k which are mathematically equivalent each other. One of them reads

$$\mu_k^{-1}=d_+(k)+d_-(k)-(\alpha(k)-\lambda).$$

However algorithms based upon LDU and/or UDL factorizations are unstable even the factorizations exist. The index k located numerically by Fernando's algorithm may differ from the index k_d, or $|u(k)|$ may be relatively small compared with $|u(k_d)|$ or $|u(k_{\max})|$. In the following we will propose a stable algorithm for computing the index k_d using QR and QL factorizations which will be referred as *orthogonal double factorization*.

Let $T = [\beta_i, \alpha_i, \gamma_i]$ be a nonsingular tridiagonal matrix of order n. And let $G_{i,j}$ be the Givens rotation in the (i, j) coordinate plane. The process of the QR factorization $T = QR$ of T can be written as follows

$$G_{n,n-1} \cdots G_{3,2} G_{2,1} T = R$$

with $Q = G_{2,1}^T \cdots G_{n,n-1}^T$. Obviously R is a upper banded matrix of bandwidth three with the kth row

$$[\cdots, 0, r_k, s_k, t_k, 0, \cdots].$$

Writing

$$G_{k,k-1} = \mathrm{diag}(I_{k-2}, \hat{G}_{k,k-1}, I_{n-k-1})$$

for the 2×2 submatrix $\hat{G}_{k,k-1}$ of the Givens rotation $G_{k,k-1}$, we can verify that

$$\hat{G}_{k,k-1} \begin{bmatrix} \hat{r}_{k-1} & \hat{s}_{k-1} & \\ \beta_k & \alpha_k & \gamma_k \end{bmatrix} = \begin{bmatrix} r_{k-1} & s_{k-1} & t_{k-1} \\ & \hat{r}_k & \hat{s}_k \end{bmatrix}.$$

Here $\hat{r}_1 = \alpha_1$ and $\hat{s}_1 = \gamma_1$. Similarly, the QL factorization of T can be written as

$$T = (G_{1,2} \cdots G_{n-1,n})^T L \equiv GL$$

and

$$\hat{G}_{k,k+1} \begin{bmatrix} \beta_k & \alpha_k & \gamma_k \\ & \hat{f}_{k+1} & \hat{g}_{k+1} \end{bmatrix} = \begin{bmatrix} \hat{f}_k & \hat{g}_k & \\ e_{k+1} & f_{k+1} & g_{k+1} \end{bmatrix},$$

where e_{k+1}, f_{k+1} and g_{k+1} are the three nonzero entries of the $(k+1)$th row

$$(\cdots, 0, e_{k+1}, f_{k+1}, g_{k+1}, 0, \cdots)$$

of the lower triangular matrix L with starting $\hat{g}_n = \beta_{n-1}$ and $\hat{f}_n = \alpha_n$.

Theorem 5.1. *Let $T = [\beta_i, \alpha_i, \gamma_i]$ be a nonsingular tridiagonal matrix of order n, and let*

$$QR = T = GL$$

be the QR and QL factorizations, respectively, with upper triangular $R = (r_{ij})$, lower triangular $L = (l_{ij})$, and orthogonal Q and G. Then for each k, $1 \le k \le n$,

$$(T^{-1})_{kk}^{-1} = \alpha_k - \frac{\hat{s}_{k-1}}{\hat{r}_{k-1}} \beta_k - \frac{\hat{f}_{k+1}}{\hat{g}_{k+1}} \gamma_k.$$

The second and the last terms on the right side of the equation above will vanish for $k = 1$ and $k = n$, respectively.

Proof. Considering the first $k-2$ Givens rotations $G_{2,1}, \cdots, G_{k-1,k-2}$ as matrices of order $k-1$ and denote by Q_1 the product

$$Q_1 = G_{k-1,k-2} \cdots G_{21},$$

we have that

$$Q_1 T(1 : k-1, 1 : k-1) = \hat{R}_1,$$

where $\hat{R}_1$ is the upper triangular matrix of $R(1 : k-1, 1 : k-1)$ except the last diagonal entry $\hat{r}_{k-1}$. Similarly for the $(n-k) \times (n-k)$ orthogonal matrix $G_2 = G_{k+1,k+2} \cdots G_{n-1,n}$,

$$G_2 T(k+1 : n, k+1 : n) = \hat{L}_2$$

with $\hat{L}_2 = L(k+1:n, k+1:n)$ except the first diagonal $\hat{g}_{k+1}$. Therefore we have that

$$\mathrm{diag}(Q_1, 1, G_2)T = \begin{bmatrix} \hat{R}_1 & \hat{s}_{k-1} & & \\ & \beta_k & \alpha_k & \gamma_k \\ & & \hat{f}_{k+1} & \hat{L}_2 \end{bmatrix} \equiv B.$$

Note that $\hat{R}_1$ is upper triangular while $\hat{L}_2$ lower triangular. It can be shown that

$$(T^{-1})(k,k) = (B^{-1})(k,k) = (B_k^{-1})(2,2), \tag{5.2}$$

where

$$B_k \equiv B(k-1:k+1, k-1:k+1) = \begin{bmatrix} \hat{r}_{k-1} & \hat{s}_{k-1} & \\ \beta_k & \alpha_k & \gamma_k \\ & \hat{f}_{k+1} & \hat{g}_{k+1} \end{bmatrix}.$$

Applying Cramer's rule gives

$$(B_k^{-1})(2,2) = \frac{\hat{r}_{k-1}\hat{g}_{k+1}}{\det B_k} = \left(\alpha_k - \frac{\hat{s}_{k-1}}{\hat{r}_{k-1}}\beta_k - \frac{\hat{f}_{k+1}}{\hat{g}_{k+1}}\gamma_k\right)^{-1}.$$

By (5.2) the result of the theorem follows.

References

[1] E. Anderson, Z. Bai, C. Bischof, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, S. Ostrouchov, and D. Sorensen, LAPACK Users' Guide, Release 2., SIAM, Philadelphia, 1995

[2] C. Davis and W. Kahan, The rotation of eigenvectors by a perturbation-III, *SIAM J. Numer. Anal.,* **7** (1970), 1-46.

[3] K.V. Fernando, On computing an eigenvector of a tridiagonal matrix, Part 1: Basic Results, *SIAM J. Matrix Anal. Appl.,* **18** (1997), 1013-1034.

[4] G. Meurant, A review on the inverse of symmetric tridiagonal and block tridiagonal matrices, *SIAM J. Matrix Anal. Appl.,* **13** (1992), 707-728.

[5] W. Givens, The characteristic value-vector problem, *J. Assoc. Comput. Math.,* **4** (1957), 298-307.

[6] G.H. Golub and C.F. Van Loan, Matrix Computations, Johns Hopkins University Press, Baltimore and London, 3rd edition, 1996.

[7] I. Ipsen, Computing an eigenvector with inverse iteration, *SIAM Review,* **39** (1997), 254-291.

[8] E. Jessup and I. Ipsen, Improving the accuracy of inverse iteration, *SIAM J. Sci.Stat. Comput.,* **13** (1992), 550-572.

[9] B.N. Parlett and I.S. Dhillon, Fernando's Solution to Wilkinson's Problem: An Application of Double Factorization, *Linear Algebra and Its Applications,* **267** (1997), 247-279.

[10] B.N. Parlett, The Symmetric Eigenvalue Problem, SIAM Press, Philadelphia, 1998.

[11] G.W. Stewart and J. G. Sun, Matrix Perturbation Theory, Academic Press, 1990.

[12] J.H. Wilkinson, The Algebra Eigenvalue Problem, Clarendon, Oxford, 1965.

[13] Zhenyue Zhang, An inverse extreme eigenpair problem and its parallel iterative solution for symmetric tridiagonal matrices, *SIAM J. Numer. Anal.,* **32** (1995), 1620-1634.