Science in China Series A: Mathematics
www.scichina.com math.scichina.com
www.springerlink.com

© 2008 SCIENCE IN CHINA PRESS
Springer

A convergence analysis of the inexact Rayleigh quotient iteration and simplified Jacobi-Davidson method for the large Hermitian matrix eigenproblem

JIA ZhongXiao† & WANG Zhen

Department of Mathematical Sciences, Tsinghua University, Beijing 100084, China
(email: jiazx@tsinghua.edu.cn, wangzhen01@tsinghua.org.cn)

Abstract The inexact Rayleigh quotient iteration (RQI) is used for computing the smallest eigenpair of a large Hermitian matrix. Under certain condition, the method was proved to converge quadratically in literature. However, it is shown in this paper that under the original given condition the inexact RQI may not quadratically converge to the desired eigenpair and even may misconverge to some other undesired eigenpair. A new condition, called the uniform positiveness condition, is given that can fix misconvergence problem and ensure the quadratic convergence of the inexact RQI. An alternative to the inexact RQI is the Jacobi-Davidson (JD) method without subspace acceleration. A new proof of its linear convergence is presented and a sharper bound is established in the paper. All the results are verified and analyzed by numerical experiments.

Keywords: eigenvalue, eigenvector, large Hermite matrix, inexact, RQI, simplified JD, convergence, misconvergence, the uniform positiveness condition

MSC(2000): 65F15

1 Introduction

Consider the problem of computing the smallest eigenvalue λ and the associated eigenvector x of a large Hermitian matrix $A \in \mathbb{C}^{n\times n}$. There are a number of methods to solve this problem, such as the power method, inverse iteration and the Rayleigh quotient iteration (RQI)[1,2]. All the eigenvectors and their approximations are normalized to have a unit length in the 2-norm $\|\cdot\|$ throughout this paper. Assume that u_k is an approximation to x. Then the Rayleigh quotient $\theta_k = u_k^* A u_k$ is a good approximation to λ too. For general matrices, the RQI converges quadratically, and for a Hermitian A it converges cubically. However, its quick convergence requires the exact solution of an increasingly large ill-conditioned linear system at each iteration. Due to the computational cost and the storage requirement, direct solvers are generally very expensive and even impractical. As a result, one has to resort to iterative solvers for getting an inaccurate solution of the linear system at each step. This leads to the inexact RQI, whose convergence becomes difficult and complicated. For some results available, we refer to [3–8] and the references therein. Under certain condition, the inexact RQI may

Received June 27, 2007; accepted December 13, 2007
DOI: 10.1007/s11425-008-0050-y
† Corresponding author
This work was supported by the National Natural Science Foundation of China (Grant Nos. 10471074, 10771116) and the Doctoral Program of the Ministry of Education of China (Grant No. 20060003003)

converge quadratically[8]. Unfortunately, as we will see, under the original given condition, the quadratic convergence of the inexact RQI cannot be guaranteed; even worse, it may not converge to the desired eigenpair and instead may misconverge to some other undesired ones (the second smallest eigenpair of A). This important issue of misconvergence has received little attention and has not been addressed in literature. We consider the convergence of the inexact RQI again and address the important issue of misconvergence. We present a new condition, called the uniform positiveness condition, and prove that under this new condition the inexact RQI converges to the desired eigenpair quadratically.

Besides the simple methods mentioned above, there are other more popular and more practical methods for computing the smallest eigenpair of A, e.g., Lanczos method[9], Arnoldi method[9], Davidson's method[10] and the Jacobi-Davidson (JD) method[11]. Except for the use of subspace acceleration, one of the most attractive features of the JD method is that the convergence is not greatly retarded even if we only solve the correction equations involved inaccurately during iteration. Approximate solutions of the correction equations in the JD method are usually constructed by only a few steps of an iterative solver, e.g., GMRES[12]. Hence the total arithmetic operations may be reduced significantly. The convergence of the inexact JD method has been discussed in a number of papers, e.g., [4, 8, 13, 14].

Eshof[8] focused on the inexact JD method without subspace acceleration and analyzed the relationships between accurate and inaccurate solutions of the correction equations. He used the estimate of the smallest nonzero singular value of $(I-uu^*)(A-\theta I)(I-uu^*)$ to prove the linear convergence of the inexact JD method. In this paper, we make use of orthogonal direct sum decompositions[2] and the angle between x and u_k to give a more detailed analysis about the inexact JD method. We will obtain the bound in [8] and establish a sharper bound than it. Numerically, we find that the convergence is generally much faster than the bound in [8] predicts.

The paper is organized as follows. In Section 2, we first review the inexact RQI and then give a detailed convergence analysis. We present the uniform positiveness condition, prove that under this condition the inexact RQI converges quadratically and discuss how it may misconverge. Finally, we run numerical experiments to confirm the theoretical results. We analyze numerical behavior and give some suggestions on how to use the method. In Section 3, we first review the inexact simplified Jacobi-Davidson method and then prove its linear convergence with numerical experiments supported.

Some notations to be used are introduced. λ_i and x_i, $i=1,2,\ldots,n$, are the eigenvalues and the associated normalized eigenvectors of A with $\lambda_1<\lambda_2\leqslant\cdots\leqslant\lambda_n$. λ_1 and x_1 are denoted by λ and x for brevity. The superscript $*$ denotes the conjugate transpose of a matrix or a vector, I the identity matrix with the order clear in the context, $\|\cdot\|$ the spectral norm of a matrix and the 2-norm of a vector.

2 The inexact RQI

2.1 The method and convergence analysis

Assume that the unit length u_k is a reasonably good approximation to x. Then the Rayleigh quotient $\theta_k=u_k^*Au_k$ is a good approximation to λ too. The RQI computes a new approximate

eigenvector u_{k+1} by solving the linear system

$$(A-\theta_k I)w_{k+1}=u_k,\quad u_{k+1}=\frac{w_{k+1}}{\|w_{k+1}\|},\tag{2.1}$$

and iterates until convergence. It is summarized as follows.

Algorithm 1: The RQI

- Choose a unit length vector u_1 as an approximation to x.
- For $k=1,2,\ldots,$ do
  $\theta_k=u_k^*Au_k$.
  Solve the linear system $(A-\theta_k I)w_{k+1}=u_k$.
  $u_{k+1}=w_{k+1}/\|w_{k+1}\|$.
  Test convergence. If yes, quit.

Though RQI converges cubically, it has an obvious drawback that at each iteration we need the accurate solution of an increasingly ill-conditioned linear system $(A-\theta_k I)w_{k+1}=u_k$. For a large A it is generally impractical to solve it accurately by a direct solver. So we must resort to iterative solvers to get an approximate solution w_{k+1}^i of $(A-\theta_k I)w_{k+1}=u_k$ with relative low accuracy. We call the Rayleigh quotient step outer iteration and solving (2.1) inner iteration. The combination of inner-outer iterations leads to the inexact RQI. To distinguish, the exact RQI means that (2.1) is solved accurately. For the inexact RQI, w_{k+1}^i satisfies

$$(A-\theta_k^i I)w_{k+1}^i=u_k^i+\xi_k d_k,\quad u_{k+1}^i=\frac{w_{k+1}^i}{\|w_{k+1}^i\|},\tag{2.2}$$

where $\xi_k d_k$ is the residual of the approximate solution w_{k+1}^i, d_k is a normalized error vector, and $0<\xi_k<1$ measures the accuracy of w_{k+1}^i and x_i may vary at each outer iteration. The inexact RQI is summarized as Algorithm 2.

Algorithm 2: The inexact RQI

- Choose a unit length vector u_1^i as an approximation to x and $0<\xi<1$.
- For $k=1,2,\ldots,$ do
  $\theta_k^i=(u_k^i)^*Au_k^i$.
  Solve the linear system $(A-\theta_k^i I)w_{k+1}^i=u_k^i$ inaccurately $\|(A-\theta_k^i I)w_{k+1}^i-u_k^i\|=\xi_k\leqslant\xi$.
  $u_{k+1}^i=w_{k+1}^i/\|w_{k+1}^i\|$.
- Test convergence. If yes, quit.

Notay[5] considered the convergence of the inexact RQI for the generalized Hermitian eigenvalue problem. For the exact RQI, one of his results states that if $\theta_1=u_1^*Au_1<\frac{\lambda+\lambda_2}{2}$ then θ_k converges to λ.

We concentrate on the inexact RQI in this section. So without ambiguity we drop the superscript i in u_k^i, w_k^i and θ_k^i.

Theorem 2.1[5]**.** *If w_{k+1} satisfies*

$$\|(A-\theta_k I)w_{k+1}-u_k\|=\xi_k\leqslant\xi<1\tag{2.3}$$

and $\phi_k=\angle(u_k,x)$ denotes the acute angle of u_k and x, then

$$\tan\phi_{k+1}\leqslant\frac{\lambda_n-\lambda}{\lambda_2-\lambda}\frac{\xi}{\sqrt{1-\xi^2}}\sin^2\phi_k+O(\sin^3\phi_k).\tag{2.4}$$

This theorem states that if the approximate solution w_{k+1} satisfies (2.3) then the inexact RQI converges quadratically. As being seen shortly, however, the only requirement (2.3) is not sufficient for the quadratic convergence of the inexact RQI and for its convergence to the desired eigenpair; even worse, it may misconverge to some other eigenpair. To this end, let us make the following orthogonal direct sum decompositions of u_k and d_k:

$$u_k = x\cos\phi_k + e_k\sin\phi_k, \quad e_k \perp x, \tag{2.5}$$

$$d_k = x\cos\psi_k + f_k\sin\psi_k, \quad f_k \perp x, \tag{2.6}$$

where all the vectors have a unit length. Then (2.2) can be written as

$$(A-\theta_k I)w_{k+1} = (\cos\phi_k + \xi_k\cos\psi_k)x + (e_k\sin\phi_k + \xi_k f_k\sin\psi_k). \tag{2.7}$$

The following theorem shows that $\cos\phi_k + \xi_k\cos\psi_k$ plays a crucial role in the convergence of the inexact RQI.

Theorem 2.2. *If w_{k+1} satisfies* (2.3) *and the following uniform positiveness condition holds*

$$\cos\phi_k + \xi_k\cos\psi_k \geqslant c, \tag{2.8}$$

where c is a positive constant independent of k, then

$$\tan\phi_{k+1} \leqslant \xi\frac{\lambda_n-\lambda}{c(\lambda_2-\lambda)}\sin^2\phi_k + O(\sin^3\phi_k). \tag{2.9}$$

Thus, the inexact RQI converges quadratically.

Proof. Premultiplying the two sides of (2.7) by $(A-\theta_k I)^{-1}$ gives

$$w_{k+1} = (\lambda-\theta_k)^{-1}(\cos\phi_k + \xi_k\cos\psi_k)\,x + (A-\theta_k I)^{-1}(e_k\sin\phi_k + \xi_k f_k\sin\psi_k). \tag{2.10}$$

This is an orthogonal direct sum decomposition of w_{k+1} since for a Hermitian A the second term is orthogonal to x as e_k and f_k are so. Note that $\cos\phi_k + \xi_k\cos\psi_k \geqslant c > 0$. We then have

$$\tan\phi_{k+1} = |\lambda-\theta_k|\frac{\|(A-\theta_k I)^{-1}(e_k\sin\phi_k + \xi_k f_k\sin\psi_k)\|}{\cos\phi_k + \xi_k\cos\psi_k}. \tag{2.11}$$

As A is Hermitian, we get

$$\begin{aligned}\lambda-\theta_k =& \lambda - u_k^*Au_k\\ =& \lambda - (x\cos\phi_k + e_k\sin\phi_k)^*A(x\cos\phi_k + e_k\sin\phi_k)\\ =& \lambda - \lambda\cos^2\phi_k - e_k^*Ae_k\sin^2\phi_k\\ =& (\lambda - e_k^*Ae_k)\sin^2\phi_k.\end{aligned}$$

Since $e_k \perp x$, we have

$$\begin{aligned}&\lambda_2-\lambda \leqslant |\lambda - e_k^*Ae_k| \leqslant \lambda_n-\lambda,\\ &(\lambda_2-\lambda)\sin^2\phi_k \leqslant |\lambda-\theta_k| \leqslant (\lambda_n-\lambda)\sin^2\phi_k.\end{aligned} \tag{2.12}$$

Note that

$$\begin{aligned}\|(A-\theta_k I)^{-1}(e_k\sin\phi_k + \xi_k f_k\sin\psi_k)\| &\leqslant \|(A-\theta_k I)^{-1}e_k\|\sin\phi_k + \xi_k\|(A-\theta_k I)^{-1}f_k\|\sin\psi_k\\ &\leqslant (\lambda_2-\lambda)^{-1}(\sin\phi_k + \xi_k\sin\psi_k).\end{aligned}$$

Using $\cos\phi_k + \xi_k\cos\psi_k \geqslant c$ again, we then get

$$\begin{aligned}\tan\phi_{k+1} &\leqslant |\lambda - e_k^* A e_k|\sin^2\phi_k \frac{\sin\phi_k + \xi_k\sin\psi_k}{(\cos\phi_k + \xi_k\cos\psi_k)(\lambda_2 - \lambda)} \\ &\leqslant |\lambda - e_k^* A e_k|\sin^2\phi_k \frac{\xi_k}{c(\lambda_2 - \lambda)} + O(\sin^3\phi_k) \\ &\leqslant \xi_k \frac{\lambda_n - \lambda}{c(\lambda_2 - \lambda)}\sin^2\phi_k + O(\sin^3\phi_k) \\ &\leqslant \xi \frac{\lambda_n - \lambda}{c(\lambda_2 - \lambda)}\sin^2\phi_k + O(\sin^3\phi_k).\end{aligned}$$

We comment that if $\xi_k = 0$ for all k then the inexact RQI is equivalent to the exact RQI and the theorem shows $\tan\phi_{k+1} = O(\sin^3\phi_k)$, the cubic convergence of the exact RQI.

Note that if $\cos\phi_k + \xi_k\cos\psi_k$ is very small, say, $O(\sin^3\phi_k)$, then

$$\tan\phi_{k+1} = O\left(|\lambda - e_k^* A e_k|\sin^2\phi_k \frac{\sin\phi_k + \xi_k\sin\psi_k}{(\lambda_2 - \lambda)\sin^3\phi_k}\right) = O\left(\frac{|\lambda - e_k^* A e_k|}{\lambda_2 - \lambda}\right).$$

So, a serious consequence may happen to the inexact RQI in this case. If it converges, it will converge to some other eigenvalue and eigenvector (usually the second smallest eigenpair).

For $\cos\phi_k \approx 1$ but less than 1, if $\cos\psi_k \approx -1$ and $\xi < 1$ very close to 1, then $\cos\phi_k + \xi_k\cos\psi_k$ is very small. This means that only condition (2.3) cannot ensure the quadratic convergence of the inexact RQI. This possible misconvergence can also be drawn by (2.10): if $\cos\phi_k + \xi_k\cos\psi_k$ is very small then w_{k+1} has a very small component in x but has relatively large components in other eigenvectors. This may cause misconvergence. In particular, if $\cos\phi_k + \xi_k\cos\psi_k = 0$, $w_{k+1} = (A - \theta_k I)^{-1}(e_k\sin\phi_k + \xi_k f_k\sin\psi_k)$, from which we have $w_{k+1} \perp x$. As a result, in exact arithmetic, the inexact RQI usually converges to λ_2 and x_2 rather than our desired λ and x. In summary, even though (2.3) is satisfied, we cannot ensure the quadratic convergence of the inexact RQI and it even may misconverge.

The above analysis shows that the quadratic convergence of the inexact RQI requires that $\cos\phi_k + \xi_k\cos\psi_k$ is not very small and it should be uniformly bounded below by a positive constant far from zero, say 0.01. Since an ξ near 1 may induce the inexact RQI misconverge, we should solve (2.1) with some accuracy, $\xi = 0.1$, for example.

2.2 Numerical experiments

2.2.1 Numerical experiment 1

The matrix A is CAN1054 from [15]. We first use the function *eig* in Matlab to compute the eigenvalues and eigenvectors of A. The starting vector u_1 is then taken to be a normalized perturbation of x such that $\cos\phi_1 = 0.9965$. The Rayleigh quotient $\theta_1 = u_1^* A u_1 = -4.4210$ is closer to the smallest eigenvalue -4.5142 of A, than its second smallest eigenvalue -4.2949. Since $\theta_1 < \frac{\lambda+\lambda_2}{2}$, the exact RQI converges to λ and x. In the inexact RQI the inner tolerance $\xi = 0.99$ and the outer iteration stops when $\|r_k\| = \|(A - \theta_k I)u_k\| \leqslant 10^{-8}$. In this experiment, d_k is chosen to be $-x$ for all k. The results obtained by the exact and inexact RQI are shown in Figure 1, where "o" and "$*$" denote the exact RQI and the inexact RQI respectively.

Figure 1 exhibits the convergence processes of the exact RQI and the inexact RQI. The former converges to λ and x in three outer iterations. For the inexact RQI, since $d_k = -x$

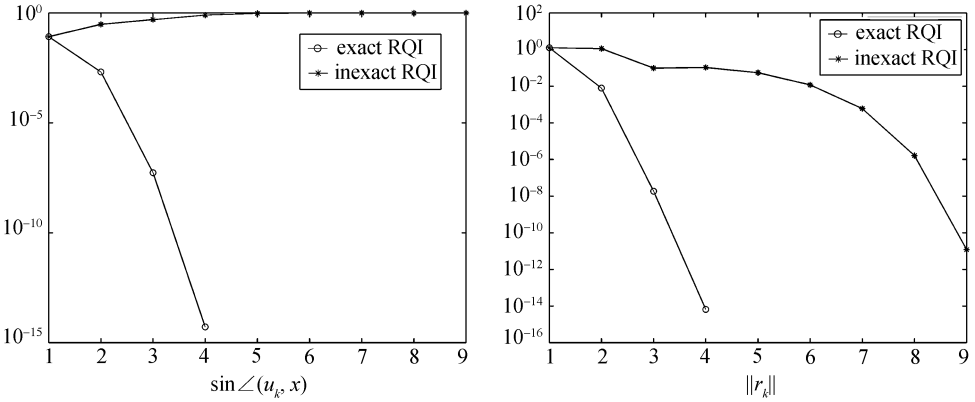

Figure 1 Comparison of the exact RQI with the inexact RQI ($\xi = 0.99$)

and $\xi \approx 1$ mean that the right side of (2.10) has a very small component in x so that the uniform positiveness condition cannot be satisfied, u_k computed by the inexact RQI gradually becomes orthogonal to x numerically, as shown by the left figure. In fact, the inexact RQI finally misconverges to $\lambda_2 = -4.2949$ and x_2, as indicated by the right figure, which shows the residual norms of the two methods.

2.2.2 Numerical experiment 2

In the numerical experiment 1, the uniform positiveness condition fails to hold and the inexact RQI misconverges. We now take $\xi = 0.01$ and other conditions are the same as those in numerical experiment 1. It is seen from (2.11) that $\cos\phi_k + \xi_k\cos\psi_k \geqslant 0.99 - 0.01 = 098$. Therefore, the uniform positiveness condition is satisfied, so the inexact RQI should converge to λ and x quadratically. Table 1 lists $\sin\phi_k$ obtained by the exact and inexact RQI and Figure 2 depicts the convergence processes of the two methods. It is clear

Table 1 $\sin\phi_k$'s of the exact and inexact RQI's

k	Exact RQI	Inexact RQI
1	8.0431e–002	8.0431e–002
2	2.3339e–003	5.4341e–003
3	6.2897e–008	2.7886e–006
4	5.2951e–015	1.0515e–012

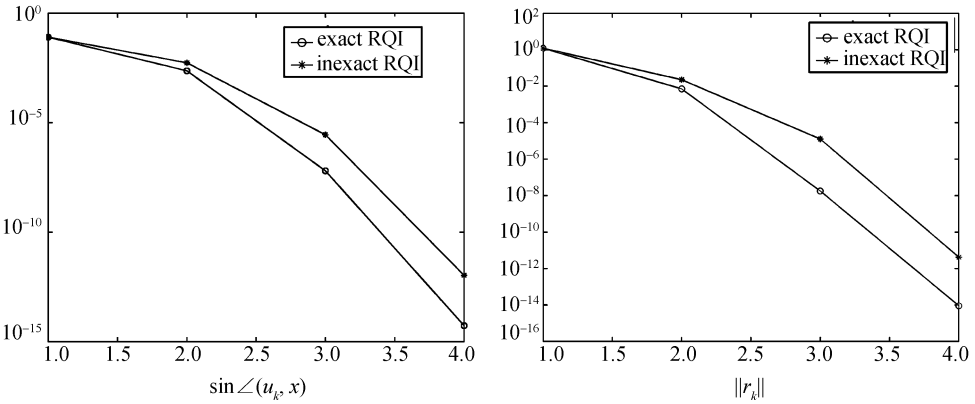

Figure 2 Comparison of the exact RQI with the inexact RQI ($\xi = 0.01$)

that the exact RQI converges cubically while the inexact RQI converges quadratically, but both of them use three outer iterations to achieve the desired accuracy.

In numerical experiments 1 and 2, when d_k are generated randomly in a uniform distribution, the inexact RQI behaves similarly and we can reach the same conclusions. Therefore, numerical experiments suggest that if we know nothing about d_k then we must take ξ to be reasonably small to ensure $\cos\phi_k + \xi_k \cos\psi_k \geqslant c$, so that the inexact RQI converges both quadratically and correctly.

3 The inexact simplified JD method

In this section, we first review the inexact simplified JD method. We then give a new proof of its linear convergence and establish a sharper bound than that in [8]. Finally, we run some numerical experiments to verify the results obtained.

3.1 The method

The simplified JD method is described as Algorithm 3.

Algorithm 3: The simplified JD

- Choose a unit length vector u_1 as approximation to x.
- For $k = 1, 2, \ldots,$ do

  $\theta_k = u_k^* A u_k$.

  $r_k = (A - \theta_k I) u_k$.

  Solve $(I - u_k u_k^*)(A - \theta_k I)(I - u_k u_k^*)\, t_k = -r_k,\ t_k \perp u_k$ for t_k.

  $w_{k+1} = u_k + t_k$.

  $u_{k+1} = w_{k+1}/\|w_{k+1}\|$.

  Test convergence. If yes, quit.

For brevity, we drop the subscript k. Assume that u is an approximation to x. Compute $\theta = u^* A u$ and the residual $r = (A - \theta I)u$. Then the method solves the correction equation

$$(I - uu^*)(A - \theta I)(I - uu^*)\, t = -r, \quad t \perp u, \tag{3.1}$$

for t and computes the new approximate eigenvector $u' = u + t$, $u' = u'/\|u'\|$. Proceed in such a way until convergence.

If (3.1) is solved exactly, then we have

$$(A - \theta I)t - (u^*(A - \theta I)t)u = -r = -(A - \theta I)u.$$

Premultiplying the two sides by $(A - \theta I)^{-1}$ and noting $t \perp u$, we have

$$t = -u + \gamma (A - \theta I)^{-1} u, \quad \gamma = (u^*(A - \theta I)^{-1} u)^{-1}. \tag{3.2}$$

Therefore, the new approximate eigenvector $u' = u + t$ is parallel to $w_{k+1} = (A - \theta I)^{-1} u$ obtained by the inexact RQI (cf. Section 2.1), and the RQI and the simplified JD method are mathematically equivalent.

If (3.1) is solved inexactly, we are led to the inexact simplified JD method[4,8,11]. The correction equation can be solved by iterative solvers, e.g., CG and MINRES, for an approximate solution t, which is used to update the available u. So, t satisfies the equation

$$(I - uu^*)(A - \theta I)(I - uu^*)\, t = -r + \xi \|r\| d, \tag{3.3}$$

where ξ is a small positive number, d is a unit length vector and $\xi\|r\|d$ is the residual of t as the approximate solution of (3.1).

Since $t \perp u$, (3.3) can be written as

$$(A-\theta I)(u+t) = \gamma u + \xi\|r\|d, \quad \gamma = u^*(A-\theta I)t. \tag{3.4}$$

Let us decompose u and d into the following orthogonal direct sums:

$$u = x\cos\phi + e\sin\phi, \quad e \perp x, \tag{3.5}$$

$$d = x\cos\psi + f\sin\psi, \quad f \perp x, \tag{3.6}$$

where all the vectors have the unit length. Substitute them into (3.4). We then get

$$(A-\theta I)(u+t) = (\gamma\cos\phi + \xi\|r\|\cos\psi)\,x + (\gamma\, e\sin\phi + \xi\|r\|f\sin\psi).$$

Therefore, the new approximate eigenvector$u' = u + t$ has the form

$$u' = (\lambda-\theta)^{-1}(\gamma\cos\phi + \xi\|r\|\cos\psi)\,x + (A-\theta I)^{-1}(\gamma\, e\sin\phi + \xi\|r\|f\sin\psi),$$

and

$$\tan\phi' = \tan\angle(u', x) = |\lambda-\theta|\frac{\|(A-\theta I)^{-1}(\gamma\, e\sin\phi + \xi\|r\|f\sin\psi)\|}{|\gamma\cos\phi + \xi\|r\|\cos\psi|}. \tag{3.7}$$

To bound $\tan\phi'$, we need to estimate $\|r\|$, $\cos\psi$ and γ.

Theorem 3.1. *The following results hold*:

$$\|r\| = \|(A-\theta I)e\|\sin\phi + O(\sin^2\phi), \tag{3.8}$$

$$\cos\psi = -(e^*f)\sin\psi\tan\,\phi, \tag{3.9}$$

$$\gamma = ((\lambda - e^*Ae) + \xi(e^*f)\|(A-\theta I)e\|\sin\psi)\tan^2\phi + O(\sin^3\phi). \tag{3.10}$$

Proof. We have

$$\begin{aligned}\|r\|^2 &= \|(A-\theta I)u\|^2\\ &= \|(A-\theta I)(x\cos\phi + e\sin\phi)\|^2\\ &= \|(\lambda-\theta)\,x\cos\phi + (A-\theta I)\,e\sin\phi\|^2\\ &= (\lambda-\theta)^2\cos^2\phi + \|(A-\theta I)e\|^2\sin^2\phi.\end{aligned}$$

It follows from (2.12) and $\lambda_2 - \theta \leqslant \|(A-\theta I)e\| \leqslant \lambda_n - \theta$ that $\|r\| = \|(A-\theta I)e\|\sin\phi + O(\sin^2\phi)$.

Next we consider $\cos\psi$. Premultiplying (3.6) by x^* and making use of (3.5), we get

$$\cos\psi = (x^*d - x^*f\sin\psi) = x^*d = \frac{(u - e\sin\phi)^*d}{\cos\phi}.$$

From (3.3) we have $\xi\|r\|d = (I-uu^*)(A-\theta I)(I-uu^*)t + r$. Premultiplying it by u^* gives

$$\xi\|r\|u^*d = u^*(I-uu^*)(A-\theta I)(I-uu^*)\,t + u^*r = u^*(A-\theta I)u = 0.$$

Therefore, $\cos\psi = -(e^*d)\tan\,\phi = -(e^*f)\sin\psi\tan\,\phi$.

Finally, we estimate γ. From (3.3) we obtain $t = -u + \gamma(A-\theta I)^{-1}u + \xi\|r\|(A-\theta I)^{-1}d$. As $t \perp u$, premultiplying the above by u^* gives

$$\gamma = \frac{1 - \xi\|r\|u^*(A-\theta I)^{-1}d}{u^*(A-\theta I)^{-1}u}. \tag{3.11}$$

It follows from (3.5) and (3.9) that

$$\begin{aligned} u^*(A-\theta I)^{-1}u &= (x\cos\phi + e\sin\phi)^*(A-\theta I)^{-1}(x\cos\phi + e\sin\phi) \\ &= (\lambda-\theta)^{-1}\cos^2\phi + e^*(A-\theta I)^{-1}e\sin^2\phi \\ &= \frac{1}{\lambda - e^*Ae}\tan^{-2}\phi + e^*(A-\theta I)^{-1}e\sin^2\phi. \end{aligned}$$

Since $\|e^*(A-\theta I)^{-1}e\| \leqslant \frac{1}{\lambda_2-\lambda}$, we have

$$u^*(A-\theta I)^{-1}u = \frac{1}{\lambda - e^*Ae}\tan^{-2}\phi + O(\sin^2\phi).$$

Similarly, by (3.5), (3.6) and (2.12), we get

$$\begin{aligned} u^*(A-\theta I)^{-1}d &= (x\cos\phi + e\sin\phi)^*(A-\theta I)^{-1}(x\cos\psi + f\sin\psi) \\ &= (\lambda-\theta)^{-1}\cos\phi\cos\psi + e^*(A-\theta I)^{-1}f\sin\phi\sin\psi \\ &= \frac{1}{\lambda - e^*Ae}\sin^{-2}\phi\cos\phi\cos\psi + e^*(A-\theta I)^{-1}f\sin\phi\sin\psi \\ &= \frac{1}{\lambda - e^*Ae}\sin^{-2}\phi\cos\phi\cos\psi + O(\sin\phi). \end{aligned}$$

Basing on (3.8), (3.9) and the above relation, the numerator of γ is

$$\begin{aligned} &1-\xi\|r\|u^*(A-\theta I)^{-1}d \\ &= 1 - \xi\big(\|(A-\theta I)e\|\sin\phi + O(\sin^2\phi)\big)\left(\frac{1}{\lambda - e^*Ae}\sin^{-2}\phi\cos\phi\cos\psi + O(\sin\phi)\right) \\ &= 1 - \xi\left(\|(A-\theta I)e\|\sin\phi\frac{1}{\lambda - e^*Ae}\sin^{-2}\phi\cos\phi\cos\psi + O(\sin\phi)\right) \\ &= 1 + \xi(e^*f)\|(A-\theta I)e\|\frac{1}{\lambda - e^*Ae}\sin\psi + O(\sin\phi). \end{aligned}$$

Substituting the above relation and the above $u^*(A-\theta I)^{-1}u$ into (3.11),we obtain

$$\begin{aligned} \gamma &= \left(1 + \xi(e^*f)\|(A-\theta I)e\|\frac{1}{\lambda - e^*Ae}\sin\psi + O(\sin\phi)\right) \\ &\quad \times \left((\lambda - e^*Ae)\tan^2\phi + O(\sin^4\phi)\right) \\ &= (\lambda - e^*Ae + \xi(e^*f)\|(A-\theta I)e\|\sin\psi)\tan^2\phi + O(\sin^3\phi). \end{aligned}$$

Based on Theorem 3.1, we can prove the linear convergence of the inexact simplified JD method.

Theorem 3.2. *If t satisfies* (3.3)*, then*

$$\tan\phi' \leqslant \xi\|(A-\theta I)e\|\|(A-\theta I)^{-1}f\||\cos\phi|\sin\psi\sin\phi + O(\sin^2\phi) \tag{3.12}$$

$$\leqslant \xi\frac{\lambda_n - \lambda}{\lambda_2 - \lambda}\sin\phi + O(\sin^2\phi). \tag{3.13}$$

Proof. By Theorem 3.1, the denominator of (3.7) equals

$$\begin{aligned}\gamma\cos\phi+\xi\|r\|\cos\psi &= (\lambda-e^*Ae+\xi(e^*f)\|(A-\theta I)e\|\sin\psi)\sin^2\phi\cos^{-1}\phi\\ &\quad-\xi\|(A-\theta I)e\|(e^*f)\sin^2\phi\sin\psi\cos^{-1}\phi+O(\sin^3\phi)\\ &= (\lambda-e^*Ae)\tan\phi\sin\phi+O(\sin^3\phi).\end{aligned}$$

From (3.10) and (3.8), we get $\gamma=O(\sin^2\phi)$ and $\|r\|=O(\sin\phi)$, respectively. Therefore, $\gamma e\sin\phi$ in the numerator of (3.7) is a lower order term compared to $\xi\|r\|f\sin\psi$. So from

$$\|r\|=\|(A-\theta I)e\|\sin\phi+O(\sin^2\phi)$$

we get that

$$\begin{aligned}&\|(A-\theta I)^{-1}(\gamma e\sin\phi+\xi\|r\|f\sin\psi)\|\\ &\quad\leqslant\gamma\|(A-\theta I)^{-1}e\|\sin\phi+\xi\|r\|\|(A-\theta I)^{-1}f\|\sin\psi\\ &\quad=\xi\|(A-\theta I)e\|\|(A-\theta I)^{-1}f\|\sin\psi\sin\phi+O(\sin^2\phi).\end{aligned}$$

Substituting the above relations into (3.7) gives

$$\begin{aligned}\tan\phi' &\leqslant|\lambda-e^*Ae|\sin^2\phi\frac{\xi\|(A-\theta I)e\|\,\|(A-\theta I)^{-1}f\|\sin\psi\sin\phi+O(\sin^2\phi)}{|\lambda-e^*Ae|\,|\cos^{-1}\phi|\sin^2\phi+O(\sin^3\phi)}\\ &=\frac{\xi\|(A-\theta I)e\|\,\|(A-\theta I)^{-1}f\|\sin\psi\sin\phi+O(\sin^2\phi)}{|\cos^{-1}\phi|+O(\sin\phi)}\\ &=\xi\|(A-\theta I)e\|\,\|(A-\theta I)^{-1}f\||\cos\phi|\sin\psi\sin\phi+O(\sin^2\phi)\\ &\leqslant\xi\frac{\lambda_n-\lambda}{\lambda_2-\lambda}\sin\phi+O(\sin^2\phi),\end{aligned}$$

the last holding as $\|(A-\theta I)e\|\leqslant\lambda_n-\lambda$, $\|(A-\theta I)^{-1}f\|\geqslant 1/(\lambda_2-\lambda)$, $|\cos\phi|\leqslant 1$ and $\sin\psi\leqslant 1$.

Eshof[8] has proved (3.13), but we have given a new proof of it and obtained, though complicated, a sharper bound (3.12).

3.2 Numerical experiments

3.2.1 Numerical experiment 3

We now verify Theorem 3.2 by numerical experiments. We still use the matrix A of Subsection 2.2 and compute its smallest eigenpair, but now ξ should satisfy

$$\xi\frac{\lambda_n-\lambda}{\lambda_2-\lambda}<1 \tag{3.14}$$

so as to ensure the convergence of Algorithm 3. It is seen from this that the inexact simplified JD method may require a much higher accuracy on approximate solutions of (3.1) than the inexact RQI does on that of (2.1). However, we should stress that (3.1) is generally much better conditioned than an increasingly ill-conditioned (2.1), so we expect that Krylov iterative solvers for (3.1) generally converge much faster than they do for (2.1). The computed results of *eig* show $\frac{\lambda_n-\lambda}{\lambda_2-\lambda}\approx 88.28$, so λ is not well separated from λ_2 and this is mildly ill-conditioned problem. We take $\xi=0.01$ and solve (3.3) using MINRES. The starting vector u_1 is a normalized perturbation of x satisfying $\cos\phi_1=0.9958$ and $u_1^*Au_1=-4.3784$. Table 2 reports $\sin\phi_k$'s and $\|r_k\|$'s. The inexact simplified JD method converges to -4.5142 and x in five outer iterations.

Table 2 indicates that $\sin\phi_k$ and $\|r_k\|$ indeed converge to zero linearly but very fast. We investigate why this is the case. Recall in the last proof part of Theorem 3.2 that

Table 2 The inexact simplified JD method

k	$\sin\phi_k$	$\|r_k\|$
1	9.2129e–002	1.5630e+000
2	5.1824e–003	1.8816e–002
3	5.5361e–004	1.5827e–004
4	1.7741e–006	1.2699e–006
5	4.5250e–008	1.1893e–008
6	7.8066e–011	6.7669e–011

$$\tan\phi' \leqslant \xi\|(A-\theta I)e\|\,\|(A-\theta I)^{-1}f\||\cos\phi|\sin\psi\sin\phi + O(\sin^2\phi).$$

$g = \xi\|(A-\theta I)e\|\,\|(A-\theta I)^{-1}f\||\cos\phi|\sin\psi$ is a smaller convergence factor than $\xi\frac{\lambda_n-\lambda}{\lambda_2-\lambda}$. Since g is much less than 1, the method converges very fast. Table 3 lists all the terms in g. It is seen from the table that g is smaller than $\xi\frac{\lambda_n-\lambda}{\lambda_2-\lambda} = 0.8828$ by one order or two orders except $k = 2$. Even for $k = 2$, $g = 0.1270$ is much smaller than 0.8828. The sharper bound indicates that even though ξ is far from satisfying (3.14), the inexact JD method may converge fast.

3.2.2 Numerical experiment 4

We now apply the inexact simplified JD method to some Hermitian and symmetric matrices[15] and compute their smallest eigenpairs. The starting vectors u_1 are normalized perturbations of x, such that $\cos\phi_1 \approx 0.99$, and ξ is taken to be 0.01. Algorithm 3 stops when $\|r_k\| = \|(A-\theta_k I)u_k\| \leqslant 10^{-8}$. We solve the correction equation by using MINRES in Matlab. Table 4 reports the total inner iterations for these problems.

We see from Table 4 that if the eigenvalue problem is not ill-conditioned, then the inexact simplified JD method converges linearly but fast. However, for NOS3, although the method converges in six outer iterations, it needs a lot of inner iterations, much more than those for the other problems. Recall that

$$\tan\phi' \leqslant \xi\frac{\lambda_n-\lambda}{\lambda_2-\lambda}\sin\phi + O(\sin^2\phi).$$

We know that the convergence depends on the distribution of eigenvalues. We found that $\lambda \approx 0.0183$, $\lambda_2 \approx 0.2489$, $\lambda_n \approx 689.9$ and $\frac{\lambda_n-\lambda}{\lambda_2-\lambda} \approx 2992$ is quite large. It is the large quantity that makes the method converge slowly since it requires a relatively smaller ξ and (3.1) is relatively worse-conditioned. We use different ξ to compute the smallest eigenpair of NOS3. Total inner iterations measure the overall performance of the method. Table 5 reports the results.

In the above numerical experiments, all the starting vector u_1 are reasonably good approximations of x satisfying $\cos\phi_1 \approx 0.99$. When $\xi = 0.1$, the correction equations are solved quite poorly. This slows down the convergence of outer iteration and makes the inexact simplified

Table 3 Linear convergence factor of the inexact JD method

k	$\|(A-\theta I)e\|$	$\|(A-\theta I)^{-1}f\|$	$\|\cos\phi\|$	$\sin\psi$	g
1	16.9010	0.4915	0.9957	0.9999	0.0827
2	3.6307	3.4993	1.0000	1.0000	0.1270
3	0.2859	1.3970	1.0000	1.0000	0.0040
4	0.7158	3.8046	1.0000	1.0000	0.0272
5	0.2628	1.1948	1.0000	1.0000	0.0031

Table 4 The inexact JD method

Matrices	Total inner iterations	$\frac{\lambda_n-\lambda}{\lambda_2-\lambda}$	$\frac{\lambda_2-\lambda}{\lvert\lambda\rvert}$
NOS3	364	2992.1	12.608
GR3030	92	129.71	1.4923
ZENIOS	92	30.083	0.11218
BCSPWR10	56	87.062	0.036846
CAN1054	73	88.280	0.048579

Table 5 The inexact simplified JD method

ξ	0.1	0.01	0.001	0.0001
Total inner iterations	406	364	476	629

JD method need 406 inner iterations. When $\xi = 0.01$, the correction equations are solved with higher accuracy, and the method only needs 364 inner iterations. However, when $\xi = 0.001$ and 0.0001, although the correction equations are solved more accurately, the method needs more inner iterations to achieve the higher accuracy ξ. As a result, although outer iterations decrease slowly, total inner iterations increase greatly, making the overall performance of the method poorer. We observe that $\xi = 0.1, 0.01$ and 0.001 are far from satisfying (3.14), but the method not only still converges but also converges faster than for $\xi = 0.0001$. This shows that (3.4) is too stringent.

Acknowledgements The authors cordially thank the referee for his careful reading of the paper and for his comments.

References

1 Golub G H, Loan C V. Matrix Computations. Baltimore-London: The John Hopkins University Press, 1996
2 Parlett B N. The Symmetric Eigenvalue Problem. Philadelphia: SIAM, 1998
3 Berns-Müller J, Spence A. Inexact inverse iteration with variable shift for nonsymmetric generalized eigenvalue problems. *SIAM J Matrix Anal Appl*, **28**: 1069–1082 (2006)
4 Hochstenbach M, Sleijpen G L G. Two-sided and alternating Jacobi-Davidson. *Linear Algebra Appl*, **358**: 145–172 (2003)
5 Notay Y. Convergence analysis of inexact Rayleigh quotient iteration. *SIAM J Matrix Anal Appl*, **24**: 627–644 (2003)
6 Simoncini V, Elden L. Inexact Rayleigh quotient-type methods for eigenvalue computations. *BIT*, **42**: 159–182 (2002)
7 Smit P, Paardekooper M. The effects of inexact solvers in algorithms for symmetric eigenvalue problems. *Linear Algebra Appl*, **287**: 337–357 (1999)
8 Eshof J. The convergence of Jacobi-Davidson iterations for Hermitian eigenproblems. *Numer Linear Algebra Appl*, **9**: 163–179 (2002)
9 Stewart G W. Matrix Algorithms Vol. II. Philadelphia: SIAM, 2001
10 Davidson E. The iterative calculation of a few of the lowest eigenvalues and corresponding eigenvectors of large real-symmetric matrices. *J Comput Phys*, **17**: 87–94 (1975)
11 Sleijpen G L G, Vorst H A. A Jacobi-Davidson iteration method for linear eigenvalue problems. *SIAM J Matrix Anal Appl*, **17**: 401–425 (1996)
12 Saad Y. Iterative Methods for Sparse Linear Systems. Philadelphia: SIAM, 2003
13 Notay Y. Combination of Jacobi-Davidson and conjugate gradients for the partial symmetric eigenproblem. *Numer Linear Algebra Appl*, **9**: 21–44 (2002)
14 Notay Y. Is Jacobi-Davidson faster than Davidson? *SIAM J Matrix Anal Appl*, **36**: 522–543 (2005)
15 Duff I S, Grimes R, Lewis J. Users' Guide for the Harwell-Boeing Sparse Matrix Collection (Release I). Technical Report TR/PA/92/86, CERFACS, 1992