

## Controlling inner iterations in the Jacobi-Davidson method

***Citation for published version (APA):***

Hochstenbach, M. E., & Notay, Y. (2007). *Controlling inner iterations in the Jacobi-Davidson method*. (CASA-report; Vol. 0701). Technische Universiteit Eindhoven.

***Document status and date:***

Published: 01/01/2007

***Document Version:***

Publisher's PDF, also known as Version of Record (includes final page, issue and volume numbers)

***Please check the document version of this publication:***

- A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
- The final author version and the galley proof are versions of the publication after peer review.
- The final published version features the final layout of the paper including the volume, issue and page numbers.

[Link to publication](#)

***General rights***

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

[www.tue.nl/taverne](http://www.tue.nl/taverne)

***Take down policy***

If you believe that this document breaches copyright please contact us at:

[openaccess@tue.nl](mailto:openaccess@tue.nl)

providing details and we will investigate your claim.

# CONTROLLING INNER ITERATIONS IN THE JACOBI–DAVIDSON METHOD

MICHIEL E. HOCHSTENBACH\* AND YVAN NOTAY†

**Abstract.** The Jacobi–Davidson method is an eigenvalue solver which uses the iterative (and in general inaccurate) solution of inner linear systems to progress, in an outer iteration, towards a particular solution of the eigenproblem. In this paper we prove a relation between the residual norm of the inner linear system and the residual norm of the eigenvalue problem. We show that the latter may be estimated inexpensively during the inner iterations. On this basis, we propose a stopping strategy for the inner iterations to maximally exploit the strengths of the method. These results extend previous results obtained for the special case of Hermitian eigenproblems with the conjugate gradient or the symmetric QMR method as inner solver. The present analysis applies to both standard and generalized eigenproblems, does not require symmetry, and is compatible with most iterative methods for the inner systems. It is also easily extended to other type of inner-outer eigenvalue solvers, as inexact inverse iteration or inexact Rayleigh quotient iteration. The effectiveness of our approach is illustrated by a few numerical experiments, including the comparison of a standard Jacobi–Davidson code with the same code enhanced by our stopping strategy.

**Key words.** eigenvalue, Jacobi–Davidson, inverse iteration, Rayleigh quotient iteration, correction equation, subspace expansion, inner iterations.

**AMS subject classifications.** 65F15, 65F10

**1. Introduction.** The Jacobi–Davidson (JD) method was introduced around a decade ago by Sleijpen and Van der Vorst ([21], see also [23]). It is considered to be one of the best eigenvalue solvers, especially for eigenvalues in the interior of the spectrum. It may be applied to both standard eigenproblems

$$(1.1) \quad A \mathbf{z} = \lambda \mathbf{z}$$

and the generalized eigenproblem (GEP) [4]

$$(1.2) \quad A \mathbf{z} = \lambda B \mathbf{z}$$

or in homogeneous form if  $B$  may be singular [9]

$$(1.3) \quad (\beta A - \alpha B) \mathbf{z} = 0 .$$

In these equations,  $A$ ,  $B$  are  $n \times n$  real or complex matrices, typically large and sparse, so that solving a linear system, for instance of the form  $(A - \zeta I)\mathbf{x} = \mathbf{y}$ , where  $I$  is the identity, by a direct method is unattractive or even infeasible. The JD method computes a few selected eigenpairs  $(\lambda, \mathbf{z})$  or  $(\alpha/\beta, \mathbf{z})$ , for instance those for which the eigenvalue  $\lambda$ , respectively  $\alpha/\beta$ , is closest to a given target  $\tau$ .

JD belongs to the class of subspace methods, which means that approximate eigenvectors are sought in a subspace. Each iteration of these methods has two important phases:

---

\*Service de Métrologie Nucléaire, Université Libre de Bruxelles (C.P. 165-84), 50 Av. F.D. Roosevelt, B-1050 Brussels, Belgium. Research supported by the Belgian FNRS and NSF grant DMS-0405387. Current address: Department of Mathematics and Computing Science, Eindhoven University of Technology, PO Box 513, 5600 MB, The Netherlands, [m.e.hochstenbach@tue.nl](mailto:m.e.hochstenbach@tue.nl), [www.win.tue.nl/~hochsten](http://www.win.tue.nl/~hochsten).

†Service de Métrologie Nucléaire, Université Libre de Bruxelles (C.P. 165-84), 50 Av. F.D. Roosevelt, B-1050 Brussels, Belgium. [ynotay@ulb.ac.be](mailto:ynotay@ulb.ac.be), [homepages.ulb.ac.be/~ynotay](http://homepages.ulb.ac.be/~ynotay). Research supported by the Belgian FRNS, Maître de Recherches.

the subspace extraction, in which a sensible approximate eigenpair is sought, of which the approximate vector is in the search space; and the subspace expansion, in which the search space is enlarged by adding a new basis vector to it, hopefully leading to better approximate eigenpairs in the next extraction phase.

In this paper we analyze the expansion phase. In JD, the expansion vector is computed by solving an inner linear system, often referred to as the *correction equation*, iteratively (and in general inaccurately). We show how the progress in the outer iteration (extraction phase) is connected to the residual norm in the correction equation, and how this progress may be estimated inexpensively during inner iterations.

JD may be seen as an inexact Newton method [3, 22, 30]. In such methods, it is often challenging to find a proper stopping criterion for inner iterations. After some time, the solution of the inner system is accurate enough and any extra work is superfluous. On the other hand, with the exact solution of the linear system, the convergence is quadratic or even cubic in the Hermitian case [21, 23, 15]. To benefit from this, one ought to compute more accurate solutions of inner systems as the outer convergence proceeds.

In [13, 26], stopping criteria were proposed and shown efficient for JD in the specific case of Hermitian eigenproblems with the conjugate gradient (CG) method [8] or the symmetric QMR method [5] as inner solver. They are based on results similar to those presented here, but with a much more restricted scope. The stopping criteria use an inexpensive assessment of the outer convergence during inner iterations to avoid any superfluous work while computing accurate solutions whenever useful. Here we extend this approach and propose a stopping strategy that applies to different forms of the correction equation and different iterative inner solvers. As we will discuss in Section 2.5, our results are also easily extended to inexact inverse iteration [6, 10, 25] and inexact Rayleigh quotient iteration (RQI) [19, 2].

The remainder of this paper is organized as follows. In Section 2 we develop the necessary mathematical results; more precisely, we prove a general theorem in Section 2.1, and consider its application to several contexts: JD for standard eigenproblems (Section 2.2), JD for the GEP with skew projection (as often used in the Hermitian case) (Section 2.3), JD for the GEP with orthogonal projection (as in [4]) (Section 2.4), inexact inverse iteration and inexact RQI (Section 2.5). Stopping criteria are discussed in Section 3. The results of a few numerical experiments and some conclusions are presented in Sections 4 and 5.

## 2. Theoretical results.

**2.1. General theorem.** The following theorem contains our main mathematical result, stated in its general form. How it applies to different forms of inner-outer eigensolvers is made clear in the corollaries given in Sections 2.2 through 2.5.

**THEOREM 2.1.** *Let  $Q, S$  be  $n \times n$  matrices and  $\mathbf{p}$  a vector in  $\mathbb{C}^n$  with two-norm  $\|\mathbf{p}\| = 1$ . For all  $\mathbf{z} \in \mathbb{C}^n$  such that  $\mathbf{p}^* S \mathbf{z} \neq 0$ , letting*

$$(2.1) \quad r = \min_{\xi} \|Q\mathbf{z} - \xi S\mathbf{z}\| ,$$

*there holds*

$$(2.2) \quad \frac{|g - \beta s|}{\sqrt{1 + s^2}} \leq r \leq \begin{cases} \sqrt{g^2 + \beta^2} & \text{if } \beta < g s \\ \frac{g + \beta s}{\sqrt{1 + s^2}} & \text{otherwise} , \end{cases}$$

where

$$(2.3) \quad g = \|(I - \mathbf{p}\mathbf{p}^*)Q\mathbf{z}\| ,$$

$$(2.4) \quad s = \frac{\|(I - \mathbf{p}\mathbf{p}^*)S\mathbf{z}\|}{|\mathbf{p}^*S\mathbf{z}|} ,$$

$$(2.5) \quad \beta = |\mathbf{p}^*Q\mathbf{z}| .$$

Moreover, if  $(I - \mathbf{p}\mathbf{p}^*)Q\mathbf{z} \perp S\mathbf{z}$ ,

$$(2.6) \quad r = \sqrt{g^2 + \frac{\beta^2 s^2}{1 + s^2}} .$$

*Proof.* One has

$$(2.7) \quad r^2 = \|Q\mathbf{z} - \theta S\mathbf{z}\|^2 = \|Q\mathbf{z}\|^2 - |\theta|^2 \|S\mathbf{z}\|^2$$

where

$$\theta = \frac{(S\mathbf{z})^*Q\mathbf{z}}{\|S\mathbf{z}\|^2} .$$

Further, letting

$$\mathbf{g} = (I - \mathbf{p}\mathbf{p}^*)Q\mathbf{z}$$

(thus  $g = \|\mathbf{g}\|$ ), since  $\mathbf{g} \perp \mathbf{p}$ , there holds

$$(2.8) \quad \|Q\mathbf{z}\|^2 = \|\mathbf{g} + \mathbf{p}(\mathbf{p}^*Q\mathbf{z})\|^2 = g^2 + \beta^2 .$$

On the other hand, let

$$\mathbf{s} = (\mathbf{p}^*S\mathbf{z})^{-1} S\mathbf{z} - \mathbf{p}$$

(thus  $s = \|\mathbf{s}\|$ ). Since

$$(2.9) \quad S\mathbf{z} = (\mathbf{p}^*S\mathbf{z})(\mathbf{p} + \mathbf{s}) ,$$

letting

$$\tilde{\beta} = \mathbf{p}^*Q\mathbf{z} ,$$

(thus  $\beta = |\tilde{\beta}|$ ), one has

$$(2.10) \quad \begin{aligned} \|S\mathbf{z}\|^2 \theta &= (\mathbf{p}^*S\mathbf{z})^* ((\mathbf{p} + \mathbf{s})^*Q\mathbf{z}) \\ &= (\mathbf{p}^*S\mathbf{z})^* (\tilde{\beta} + \mathbf{s}^*Q\mathbf{z}) \\ &= (\mathbf{p}^*S\mathbf{z})^* (\tilde{\beta} + \mathbf{s}^*\mathbf{g}) , \end{aligned}$$

the last equality holding because  $\mathbf{s} \perp \mathbf{p}$ , hence  $\mathbf{s}^* = \mathbf{s}^*(I - \mathbf{p}\mathbf{p}^*)$ . Now let

$$\delta = \frac{\mathbf{s}^*\mathbf{g}}{s g}$$

and observe that (2.9) implies, since  $\mathbf{s} \perp \mathbf{p}$ ,

$$(2.11) \quad |\mathbf{p}^* S \mathbf{z}|^2 = \frac{\|S \mathbf{z}\|^2}{\|\mathbf{p} + \mathbf{s}\|^2} = \frac{\|S \mathbf{z}\|^2}{1 + s^2}.$$

One then finds, with (2.7), (2.8), (2.10), and (2.11),

$$\begin{aligned} r^2 &= g^2 + \beta^2 - (1 + s^2)^{-1} \left| \tilde{\beta} + \delta s g \right|^2 \\ &= (1 + s^2)^{-1} \left( g^2 (1 + s^2 (1 - |\delta|^2)) + \beta^2 s^2 - 2 \operatorname{Re} \left( \tilde{\beta}^* \delta \right) s g \right). \end{aligned}$$

When  $\mathbf{g} \perp S \mathbf{z}$ , one has also  $\mathbf{g} \perp \mathbf{s}$  since  $S \mathbf{z}$  is a multiple of  $\mathbf{p} + \mathbf{s}$  and  $\mathbf{g} \perp \mathbf{p}$ . Then,  $\delta = 0$  and the last equality gives (2.6). In the general case  $0 \leq |\delta| \leq 1$  and one obtains

$$\begin{aligned} r^2 &\leq (1 + s^2)^{-1} \max_{0 \leq |\delta| \leq 1} \left( g^2 (1 + s^2) + \beta^2 s^2 - g^2 s^2 |\delta|^2 + 2 \beta s g |\delta| \right), \\ &= g^2 + \beta^2 - (1 + s^2)^{-1} \min_{0 \leq |\delta| \leq 1} (|\delta| g s - \beta)^2 \\ &= \begin{cases} g^2 + \beta^2 & \text{if } \beta < g s \\ g^2 + \beta^2 - (1 + s^2)^{-1} (g s - \beta)^2 & \text{otherwise} \end{cases} \end{aligned}$$

whence the upper bound in (2.2). On the other hand,

$$r^2 \geq (1 + s^2)^{-1} \min_{0 \leq |\delta| \leq 1} \left( g^2 (1 + s^2) + \beta^2 s^2 - g^2 s^2 |\delta|^2 - 2 \beta s g |\delta| \right),$$

which gives the lower bound (obtained for  $|\delta| = 1$ ).  $\blacksquare$

The following lemma helps to understand the relation between the upper bound in (2.2) and the right hand side of (2.6).

LEMMA 2.2. *Let  $g, \beta, s$ , be nonnegative numbers, and define*

$$\begin{aligned} \bar{r} &= \begin{cases} \sqrt{g^2 + \beta^2} & \text{if } \beta < g s \\ \frac{g + \beta s}{\sqrt{1 + s^2}} & \text{otherwise,} \end{cases} \\ \tilde{r} &= \sqrt{g^2 + \frac{\beta^2 s^2}{1 + s^2}}. \end{aligned}$$

Then:

$$(2.12) \quad \tilde{r} \leq \bar{r} \leq \tilde{r} \sqrt{1 + \frac{1}{1 + s^2}}.$$

*Proof.* For  $\beta < g s$ ,

$$\left( \frac{\bar{r}}{\tilde{r}} \right)^2 = \frac{(g^2 + \beta^2)(1 + s^2)}{g^2(1 + s^2) + \beta^2 s^2} = 1 + \frac{\beta^2}{g^2(1 + s^2) + \beta^2 s^2}$$

whence  $\bar{r} \geq \tilde{r}$ , whereas the upper bound holds if and only if  $\frac{\beta^2}{g^2(1 + s^2) + \beta^2 s^2} \leq \frac{1}{1 + s^2}$ , that is, if and only if  $g^2(1 + s^2) + \beta^2 s^2 - \beta^2(1 + s^2) = g^2(1 + s^2) - \beta^2$  is nonnegative, which is always true for  $\beta < g s$ .

On the other hand, for  $\beta \geq g s$ ,

$$\left(\frac{\bar{r}}{\tilde{r}}\right)^2 = \frac{(g + \beta s)^2}{g^2(1 + s^2) + \beta^2 s^2} = 1 + \frac{g s (2\beta - g s)}{g^2(1 + s^2) + \beta^2 s^2}$$

whence  $\bar{r} \geq \tilde{r}$ , whereas the upper bound holds if and only if  $\frac{g s (2\beta - g s)}{g^2(1 + s^2) + \beta^2 s^2} \leq \frac{1}{1 + s^2}$ , that is, if and only if  $g^2(1 + s^2) + \beta^2 s^2 - g s (2\beta - g s)(1 + s^2) = (g(1 + s^2) - \beta s)^2$  is nonnegative. ■

**2.2. JD for standard eigenproblems.** Consider the standard eigenproblem (1.1), let  $\mathbf{u}$  be the current approximation to the sought eigenvector, and assume for convenience  $\|\mathbf{u}\| = 1$ . JD computes an orthogonal correction  $\mathbf{t}$  to  $\mathbf{u}$  by (approximately) solving the correction equation

$$(2.13) \quad (I - \mathbf{u}\mathbf{u}^*)(A - \zeta I)\mathbf{t} = -\mathbf{r} \quad \text{with } \mathbf{t} \perp \mathbf{u},$$

where

$$\mathbf{r} = (A - \theta I)\mathbf{u}$$

is the eigenvalue residual with

$$\theta = \mathbf{u}^* A \mathbf{u}.$$

If we solve this equation exactly, then we find [21]

$$(2.14) \quad \mathbf{t} = -\mathbf{u} + \xi (A - \zeta I)^{-1} \mathbf{u},$$

where  $\xi = (\mathbf{u}^*(A - \zeta I)^{-1}\mathbf{u})^{-1}$  is such that  $\mathbf{t} \perp \mathbf{u}$ . The solution is used to expand the search space. Since  $\mathbf{u}$  is already in this space, the expansion vector is effectively  $(A - \zeta I)^{-1}\mathbf{u}$ , which is the same as for inverse iteration if  $\zeta = \tau$ , and the same as for RQI if  $\zeta = \theta$ . Our analysis applies independently of the choice made for  $\zeta$ . For the sake of completeness, let us mention that, in general, one selects  $\zeta = \tau$  in the initial phase (to enforce convergence towards eigenvalue closest to the target  $\tau$ ), and  $\zeta = \theta$  in the final phase (to benefit from the fast convergence of RQI), see [13, 27] for a discussion.

Since  $\mathbf{t}$  aims at being a correction to  $\mathbf{u}$ , the efficiency of the expansion phase may be measured by assessing the eigenvalue residual norm associated to  $\mathbf{u} + \mathbf{t}$ . This can be done with Theorem 2.1, as made clear in Corollary 2.3 below. Note that this analysis does not take the possible acceleration obtained in the extraction phase by selecting a (hopefully) better eigenvector approximation from the search space into account.

**COROLLARY 2.3** (JD for standard eigenproblems). *Let  $A$  be an  $n \times n$  matrix,  $\zeta$  a complex number, and  $\mathbf{u}$  a vector in  $\mathbb{C}^n$  such that  $\|\mathbf{u}\| = 1$ . Let  $\mathbf{t}$  be a vector in  $\mathbb{C}^n$  such that  $\mathbf{t} \perp \mathbf{u}$ , and let*

$$(2.15) \quad \mathbf{r}_{\text{in}} = -\mathbf{r} - (I - \mathbf{u}\mathbf{u}^*)(A - \zeta I)\mathbf{t}$$

where  $\mathbf{r} = (A - \theta I)\mathbf{u}$  with  $\theta = \mathbf{u}^* A \mathbf{u}$ . Then,

$$(2.16) \quad r_{\text{eig}} = \min_{\xi} \frac{\|(A - \xi I)(\mathbf{u} + \mathbf{t})\|}{\|\mathbf{u} + \mathbf{t}\|}$$

satisfies

$$(2.17) \quad \frac{|g - \beta s|}{1 + s^2} \leq r_{\text{eig}} \leq \begin{cases} \frac{\sqrt{g^2 + \beta^2}}{\sqrt{1 + s^2}} & \text{if } \beta < g s \\ \frac{g + \beta s}{1 + s^2} & \text{otherwise,} \end{cases}$$

where

$$(2.18) \quad g = \|\mathbf{r}_{\text{in}}\| ,$$

$$(2.19) \quad s = \|\mathbf{t}\| ,$$

$$(2.20) \quad \beta = |\theta - \zeta + \mathbf{u}^*(A - \zeta I)\mathbf{t}| .$$

Moreover, if  $\mathbf{r}_{\text{in}} \perp \mathbf{t}$ ,

$$(2.21) \quad r_{\text{eig}} = \sqrt{\frac{g^2}{1 + s^2} + \left(\frac{\beta s}{1 + s^2}\right)^2} .$$

*Proof.* Apply Theorem 2.1 with  $Q = A - \zeta I$ ,  $S = I$ ,  $\mathbf{p} = \mathbf{u}$ ,  $\mathbf{z} = \mathbf{u} + \mathbf{t}$ , and take into account that  $\mathbf{t} \perp \mathbf{u}$  implies  $\|\mathbf{u} + \mathbf{t}\| = \sqrt{1 + s^2}$ , whereas  $(I - \mathbf{p}\mathbf{p}^*)Q\mathbf{z} \perp S\mathbf{z}$  amounts to  $\mathbf{r}_{\text{in}} \perp (\mathbf{u} + \mathbf{t})$ , whence the stated condition since  $\mathbf{r}_{\text{in}} \perp \mathbf{u}$ . ■

The condition  $\mathbf{r}_{\text{in}} \perp \mathbf{t}$  holds in the Hermitian case when solving (2.13) by the (preconditioned) CG method with the zero vector as initial approximation ([8]; see, e.g., [13] for a formal proof in the preconditioned case). This is used in [13] to derive a theoretical result similar to Corollary 2.3 but restricted to that particular context. In [26], the close relation between symmetric QMR and CG is exploited to extend the approach, which nevertheless remains limited to the Hermitian case.

In the general case Corollary 2.3 shows that  $r_{\text{eig}}$  can be accurately estimated. When  $s \ll 1$  (as is the case close to convergence,  $s$  being the norm of a sensible orthogonal correction to  $\mathbf{u}$ ), (2.17) shows that  $r_{\text{eig}} \approx \max(g, \beta s)$ . On the other hand, when  $s$  is relatively large, Lemma 2.2 shows that the upper bound in (2.17) is close to  $\tilde{r}/\sqrt{1 + s^2}$ , which is the right hand side of (2.21). Hence, in all cases,

$$(2.22) \quad r_{\text{eig}} \approx \max\left(\frac{g}{\sqrt{1 + s^2}}, \frac{\beta s}{1 + s^2}\right) .$$

We now illustrate this with two numerical examples. In Example 1, the matrix is *SHERMAN4* from the Harwell-Boeing Collection available on matrix market [11]. The matrix is real, unsymmetric, and of order  $n = 1104$ . Its eigenvalues are real, the smallest ones being (in 3 decimal places) 0.031, 0.085, 0.278, 0.399, 0.432, 0.590. The target eigenvalue is the one closest to  $\tau = 0.5$ , i.e., it is the fifth eigenvalue. Letting  $\mathbf{z}$  be the corresponding eigenvector (with unit norm) and  $\mathbf{v}$  a vector with random entries uniformly distributed in the interval  $(-1, 1)$ , we set  $\mathbf{u} = \|\mathbf{z} + \xi\mathbf{v}\|^{-1}(\mathbf{z} + \xi\mathbf{v})$  with  $\xi = 0.25/\|\mathbf{v}\|$ . We solve the correction equation (2.13) with  $\zeta = \tau = 0.5$ , using full GMRES [18] with the zero vector as initial approximation.

The second matrix (Example 2), which we call *BandRand* for further reference, is a band lower triangular matrix  $A$  of order  $n = 1000$  with prescribed diagonal  $a_{ii} = \sqrt{i}$  and five subdiagonals with random entries uniformly distributed in the interval  $(-1, 1)$ . Because the matrix is lower triangular, the diagonal entries give the eigenvalues, and we target the sixth one  $\sqrt{6}$ . Letting  $\mathbf{z}$  be the corresponding eigenvector (with unit norm) and  $\mathbf{v}$  a vector with

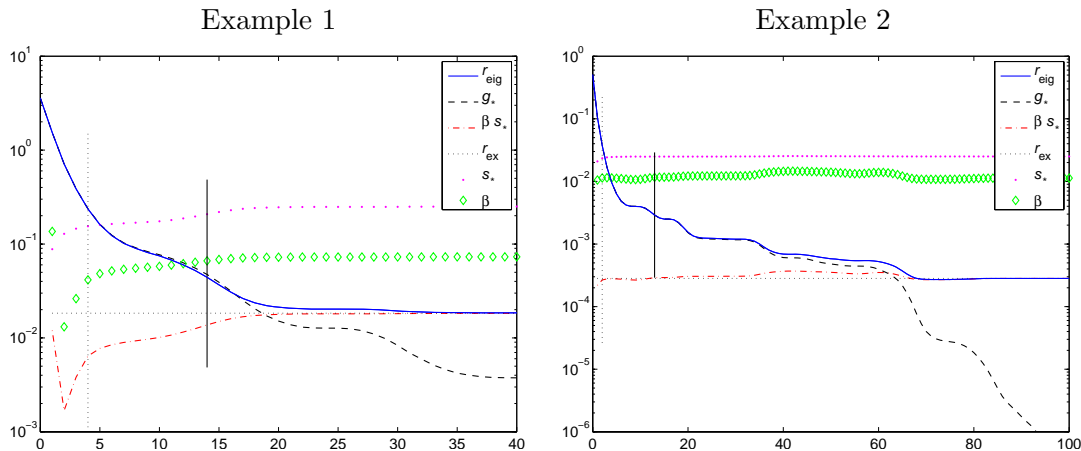


FIG. 2.1. Quantities referred to in Corollary 2.3 when solving the correction equation (2.13) for Examples 1 and 2 with full GMRES;  $g_* = g/\sqrt{1+s^2}$ ,  $s_* = s/(1+s^2)$  and  $r_{\text{ex}}$  is the eigenvalue residual norm (2.16) corresponding to the exact solution of the correction equation (2.13); the dotted vertical line corresponds to the first iteration for which the relative residual error is below 0.1 (i.e.,  $g < 10^{-1}\|\mathbf{r}\|$ ), and the solid vertical line corresponds to the exit point according to criterion (B) proposed in Section 3.

random entries uniformly distributed in the interval  $(-1, 1)$ , we set  $\mathbf{u} = \|\mathbf{z} + \xi\mathbf{v}\|^{-1}(\mathbf{z} + \xi\mathbf{v})$  with  $\xi = 0.025/\|\mathbf{v}\|$ . We solve the correction equation (2.13) with  $\zeta = \theta = \mathbf{u}^*A\mathbf{u}$ , using full GMRES with the zero vector as initial approximation.

In Figure 2.1, we show the different quantities referred to in Corollary 2.3 against the number of inner iterations. After a few iterations, both  $s$  and  $\beta$  converge smoothly towards their asymptotic values (which correspond to (2.19), (2.20) with  $\mathbf{t}$  equal to the exact solution (2.14)). On the other hand,  $g = \|\mathbf{r}_{\text{in}}\|$  decreases as the convergence proceeds, and, as predicted in our analysis (2.22),  $r_{\text{eig}}$  follows closely  $g/\sqrt{1+s^2}$  until  $g/\sqrt{1+s^2} \approx \beta s/(1+s^2)$  which means  $g \approx \beta s/\sqrt{1+s^2}$ . From there,  $r_{\text{eig}}$  stagnates to a value close to  $\beta s/(1+s^2)$ . The latter value is itself close to the eigenvalue residual norm obtained with the exact solution of the correction equation, that is, with an exact inverse iteration step (Example 1), or an exact RQI step (Example 2). We can see this by setting  $g = 0$  in (2.17).

We now discuss the practical computation of  $s$  and  $\beta$  within inner iterations. If  $A$  is Hermitian,  $\mathbf{u}^*(A - \zeta I)\mathbf{t} = \mathbf{r}^*\mathbf{t}$ . With CG or symmetric QMR,  $\beta$  may then be computed from the inner products needed by the method, that is, without any extra work, see [13, 26] for details. In the general case, one may use the following observation. In most iterative methods,  $\mathbf{t}$  is computed as a linear combination of basis vectors  $\mathbf{v}_i$  (sometimes called “search directions”) [17, 29]. Since iterative methods require the multiplication of these basis vectors by the system matrix, which is here  $(I - \mathbf{u}\mathbf{u}^*)(A - \zeta I)$ , the inner products  $\mathbf{u}^*(A - \zeta I)\mathbf{v}_i$  are needed by the method. Storing them then allows us to compute  $\beta$  at no cost from the coefficients of the linear combination (or the recursion) that expresses  $\mathbf{t}$  in the basis vectors.

In most cases, the computation of  $s$  is straightforward, but some iterative methods like GMRES do not form the approximate solution at intermediate steps, and computing it would require significant extra work. However,  $s$  and  $\beta s/(1+s^2)$  do not vary much after the relative residual error in the linear system has been decreased by a modest factor, say 10. This suggests that relevant estimates can be obtained while computing these quantities only a few times. In Section 3 we will illustrate this point when discussing practical stopping criteria in more detail.



**2.3. JD for the GEP with skew projection.** There are two main variants of JD for the GEP. Considering the problem (1.2) with approximate eigenvector  $\mathbf{u}$ , the first variant uses the correction equation with skew projection

$$(2.23) \quad \left(I - \frac{B\mathbf{u}\mathbf{u}^*}{\mathbf{u}^*B\mathbf{u}}\right) (A - \zeta B) \mathbf{t} = -\mathbf{r} \quad \text{with } \mathbf{t} \perp \mathbf{v},$$

where

$$\mathbf{r} = (A - \theta B) \mathbf{u}$$

with

$$\theta = \frac{\mathbf{u}^* A \mathbf{u}}{\mathbf{u}^* B \mathbf{u}}$$

is the eigenvalue residual;  $\mathbf{v}$  is an auxiliary vector that determines the subspace in which the solution is sought, for instance  $\mathbf{v} = B\mathbf{u}$ . The choice made for  $\mathbf{v}$  is unimportant here.

This variant of JD is often preferred for generalized Hermitian eigenvalue problems, that is, when  $A, B$  are Hermitian and  $B$  is positive definite [24]. Indeed, in this case the method is equivalent to standard JD applied to the transformed problem  $B^{-1/2} A B^{-1/2} \mathbf{z} = \lambda \mathbf{z}$ . Our results, however, are for general  $A, B$ .

The exact solution to (2.23) is

$$(2.24) \quad \mathbf{t} = -\mathbf{u} + \xi (A - \zeta B)^{-1} B \mathbf{u},$$

where  $\xi$  is such that  $\mathbf{t} \perp \mathbf{v}$ . Hence, with an approximate solution to (2.23), the expansion vector  $\mathbf{u} + \mathbf{t}$  approximates an inverse iteration step or an RQI step, according to the choice of  $\zeta$ . The following corollary shows how the corresponding eigenvalue residual norm may be estimated.

As for any eigenproblem, the eigenvalue residual norm is meaningful only if the approximate eigenvector is normalized in a certain way. Our results are stated for the case where one searches the solution to (1.2) such that  $\|B\mathbf{z}\| = 1$ . That is, we assume that  $\|B\mathbf{u}\| = 1$  and we estimate the eigenvalue residual norm for  $\frac{\mathbf{u} + \mathbf{t}}{\|B(\mathbf{u} + \mathbf{t})\|}$ . The reason for this is that the statements become somewhat simpler, more transparent, and sometimes computationally slightly more efficient. However, this scaling is no real restriction.

**COROLLARY 2.4** (JD for the GEP, skew projection). *Let  $A, B$  be  $n \times n$  matrices,  $\zeta$  a complex number,  $\mathbf{u}$  a nonzero vector in  $\mathbb{C}^n$  such that  $\|B\mathbf{u}\| = 1$  and  $\mathbf{u}^* B \mathbf{u} \neq 0$ . Set*

$$\mathbf{p} = B\mathbf{u}.$$

*Let  $\mathbf{t}$  be a vector in  $\mathbb{C}^n$ , and let*

$$(2.25) \quad \mathbf{r}_{\text{in}} = -\mathbf{r} - \left(I - \frac{B\mathbf{u}\mathbf{u}^*}{\mathbf{u}^*B\mathbf{u}}\right) (A - \zeta B) \mathbf{t}$$

*where  $\mathbf{r} = (A - \theta B) \mathbf{u}$  with  $\theta = \frac{\mathbf{u}^* A \mathbf{u}}{\mathbf{u}^* B \mathbf{u}}$ . Assume  $\mathbf{p}^* B \mathbf{t} \neq -1$ . Then,*

$$(2.26) \quad r_{\text{eig}} = \min_{\xi} \frac{\|(A - \xi B)(\mathbf{u} + \mathbf{t})\|}{\|B(\mathbf{u} + \mathbf{t})\|}$$

satisfies

$$(2.27) \quad \frac{|g - \beta s|}{\alpha(1 + s^2)} \leq r_{\text{eig}} \leq \begin{cases} \frac{\sqrt{g^2 + \beta^2}}{\alpha\sqrt{1 + s^2}} & \text{if } \beta < g s \\ \frac{g + \beta s}{\alpha(1 + s^2)} & \text{otherwise,} \end{cases}$$

where

$$(2.28) \quad g = \|(I - \mathbf{p}\mathbf{p}^*)\mathbf{r}_{\text{in}}\| = \sqrt{\|\mathbf{r}_{\text{in}}\|^2 - |\mathbf{p}^*\mathbf{r}_{\text{in}}|^2},$$

$$(2.29) \quad \alpha = |1 + \mathbf{p}^*B\mathbf{t}|,$$

$$(2.30) \quad s = \alpha^{-1}\sqrt{\|B\mathbf{t}\|^2 - |\mathbf{p}^*B\mathbf{t}|^2},$$

$$(2.31) \quad \beta = |\theta - \zeta + \mathbf{p}^*\mathbf{r} + \mathbf{p}^*(A - \zeta B)\mathbf{t}|.$$

*Proof.* Apply Theorem 2.1 with  $Q = A - \zeta B$ ,  $S = B$ ,  $\mathbf{z} = \mathbf{u} + \mathbf{t}$ ; take into account that  $(I - \mathbf{p}\mathbf{p}^*)(I - \frac{B\mathbf{u}\mathbf{u}^*}{\mathbf{u}^*B\mathbf{u}}) = I - \mathbf{p}\mathbf{p}^*$ , and that (2.11) implies  $\|B(\mathbf{u} + \mathbf{t})\| = \sqrt{1 + s^2} |\mathbf{p}^*B(\mathbf{u} + \mathbf{t})| = \sqrt{1 + s^2} |1 + \mathbf{p}^*B\mathbf{t}|$ . ■

The practical use of this result raises the same comments as in Section 2.2. Two additional difficulties need to be mentioned. Firstly, computing  $g$  requires an additional inner product. However, in several tests we experienced that the stopping strategy that will be proposed in Section 3 was equally effective using the upper bound  $\|\mathbf{r}_{\text{in}}\|$  instead of the exact value of  $g$ . Secondly,  $\mathbf{p}^*B\mathbf{t}$  is needed to compute  $\alpha$ , and the observation mentioned in Section 2.2 to compute  $\beta$  does not apply because the projector in the correction equation (2.23) is not  $I - \mathbf{p}\mathbf{p}^*$ . However, the comment made about the computation of  $s$  may be extended here, that is, the extra work may be kept modest by computing these quantities only a few times.

**2.4. JD for the GEP with orthogonal projection.** The second main variant of JD for the GEP has been proposed in [4]. Considering the problem (1.3) with approximate eigenvector  $\mathbf{u}$ , it uses the correction equation

$$(2.32) \quad (I - \mathbf{p}\mathbf{p}^*)(\eta A - \zeta B)\mathbf{t} = -\mathbf{r} \quad \text{with } \mathbf{t} \perp \mathbf{v},$$

where

$$\begin{aligned} \eta &= \frac{\mathbf{p}^*B\mathbf{u}}{\sqrt{|\mathbf{p}^*A\mathbf{u}|^2 + |\mathbf{p}^*B\mathbf{u}|^2}}, \\ \zeta &= \frac{\mathbf{p}^*A\mathbf{u}}{\sqrt{|\mathbf{p}^*A\mathbf{u}|^2 + |\mathbf{p}^*B\mathbf{u}|^2}}, \\ \mathbf{r} &= (\eta A - \zeta B)\mathbf{u}, \\ \mathbf{p} &= \frac{(\nu A + \mu B)\mathbf{u}}{\|(\nu A + \mu B)\mathbf{u}\|} \end{aligned}$$

for some given numbers  $\nu, \mu$  (see [4] for a discussion of two possible choices);  $\mathbf{v}$  is an auxiliary vector that determines the subspace in which the solution is sought, but is not of importance here. Note that  $\mathbf{r} \perp \mathbf{p}$ ; the exact solution to (2.32) is

$$(2.33) \quad \mathbf{t} = -\mathbf{u} + \xi(\eta A - \zeta B)^{-1}(\nu A + \mu B)\mathbf{u},$$

where  $\xi$  is such that  $\mathbf{t} \perp \mathbf{v}$ . Therefore, with an approximate solution to (2.23), the expansion vector  $\mathbf{u} + \mathbf{t}$  approximates an RQI step. The following corollary shows how the corresponding eigenvalue residual norm may be estimated. Again, we use a form of normalization: we scale  $\mathbf{u}$  such that  $\|(\nu A + \mu B)\mathbf{u}\| = 1$ , and we estimate the eigenvalue residual norm for  $\frac{\mathbf{u} + \mathbf{t}}{\|(\nu A + \mu B)(\mathbf{u} + \mathbf{t})\|}$ .

**COROLLARY 2.5** (JD for the GEP, orthogonal projection). *Let  $A, B$  be  $n \times n$  matrices,  $\eta, \zeta$  as above,  $\nu, \mu$  be complex numbers,  $\mathbf{u} \in \mathbb{C}^n$  with  $\|(\nu A + \mu B)\mathbf{u}\| = 1$ , and set*

$$\mathbf{p} = (\nu A + \mu B)\mathbf{u} .$$

Let  $\mathbf{t}$  be a vector in  $\mathbb{C}^n$ , and let

$$(2.34) \quad \mathbf{r}_{\text{in}} = -\mathbf{r} - (I - \mathbf{p}\mathbf{p}^*)(\eta A - \zeta B)\mathbf{t}$$

where  $\mathbf{r} = (\eta A - \zeta B)\mathbf{u}$ . Assume that  $\mathbf{p}^*(\nu A + \mu B)\mathbf{t} \neq -1$ . Then,

$$(2.35) \quad r_{\text{eig}} = \min_{\xi} \frac{\|((\eta A - \zeta B) - \xi(\nu A + \mu B))(\mathbf{u} + \mathbf{t})\|}{\|(\nu A + \mu B)(\mathbf{u} + \mathbf{t})\|}$$

satisfies

$$(2.36) \quad \frac{|g - \beta s|}{\alpha(1 + s^2)} \leq r_{\text{eig}} \leq \begin{cases} \frac{\sqrt{g^2 + \beta^2}}{\alpha\sqrt{1 + s^2}} & \text{if } \beta < g s \\ \frac{g + \beta s}{\alpha(1 + s^2)} & \text{otherwise ,} \end{cases}$$

where

$$(2.37) \quad g = \|\mathbf{r}_{\text{in}}\| ,$$

$$(2.38) \quad \alpha = |1 + \mathbf{p}^*(\nu A + \mu B)\mathbf{t}| ,$$

$$(2.39) \quad s = \alpha^{-1} \sqrt{\|(\nu A + \mu B)\mathbf{t}\|^2 - |\mathbf{p}^*(\nu A + \mu B)\mathbf{t}|^2} ,$$

$$(2.40) \quad \beta = |\mathbf{p}^*(\eta A - \zeta B)\mathbf{t}| .$$

*Proof.* Theorem 2.1 applies with  $Q = \eta A - \zeta B$ ,  $S = \nu A + \mu B$ ,  $\mathbf{z} = \mathbf{u} + \mathbf{t}$  and  $g = \|\mathbf{r}_{\text{in}}\|$ . The stated result then follows noting that (2.11) implies  $\|(\nu A + \mu B)(\mathbf{u} + \mathbf{t})\| = \sqrt{1 + s^2} |\mathbf{p}^*(\nu A + \mu B)(\mathbf{u} + \mathbf{t})| = \sqrt{1 + s^2} |1 + \mathbf{p}^*(\nu A + \mu B)\mathbf{t}|$ . ■

Here again, the practical use of this result raises the same comments as in Section 2.2; the observation proposed there to compute  $\mathbf{u}^*(A - \zeta I)\mathbf{t}$  may be used here to obtain  $\mathbf{p}^*(\eta A - \zeta B)\mathbf{t}$  and therefore  $\beta$ ;  $\mathbf{p}^*(\nu A + \mu B)\mathbf{t}$  can also be computed efficiently in a similar manner.

**2.5. Inexact inverse iteration and RQI.** Consider, for the sake of generality, the GEP (1.2) (the standard case is obtained with  $B = I$ ). Let  $\mathbf{u}$  be the current eigenvector approximation. Inverse iteration and RQI compute a new eigenvector approximation by solving

$$(2.41) \quad (A - \zeta B)\mathbf{z} = B\mathbf{u} .$$

RQI uses  $\zeta$  equal to the Rayleigh quotient  $\frac{\mathbf{u}^* A \mathbf{u}}{\mathbf{u}^* B \mathbf{u}}$ , whereas inverse iteration uses some fixed shift. The ‘‘inexact’’ variants are obtained when (2.41) is solved approximately by an iterative method. Theorem 2.1 applies also here, as shown in the following corollary.

**COROLLARY 2.6** (Inexact inverse iteration and RQI). *Let  $A, B$  be  $n \times n$  matrices,  $\zeta$  a complex number,  $\mathbf{u} \in \mathbb{C}^n$  such that  $\|B\mathbf{u}\| = 1$ , and set*

$$\mathbf{p} = B\mathbf{u} .$$

Let  $\mathbf{z}$  be a vector in  $\mathbb{C}^n$ , and let

$$(2.42) \quad \mathbf{r}_{\text{in}} = B\mathbf{u} - (A - \zeta B)\mathbf{z} .$$

Assume  $\mathbf{p}^* B\mathbf{z} \neq 0$ . Then

$$(2.43) \quad r_{\text{eig}} = \min_{\xi} \frac{\|(A - \xi B)\mathbf{z}\|}{\|B\mathbf{z}\|}$$

satisfies

$$\frac{|g - \beta s|}{|\mathbf{p}^* B\mathbf{z}| (1 + s^2)} \leq r_{\text{eig}} \leq \begin{cases} \frac{\sqrt{g^2 + \beta^2}}{|\mathbf{p}^* B\mathbf{z}| \sqrt{1 + s^2}} & \text{if } \beta < g s \\ \frac{g + \beta s}{|\mathbf{p}^* B\mathbf{z}| (1 + s^2)} & \text{otherwise ,} \end{cases}$$

where

$$(2.44) \quad g = \|(I - \mathbf{p}\mathbf{p}^*)\mathbf{r}_{\text{in}}\| = \sqrt{\|\mathbf{r}_{\text{in}}\|^2 - |\mathbf{p}^* \mathbf{r}_{\text{in}}|^2} ,$$

$$(2.45) \quad s = \frac{\sqrt{\|B\mathbf{z}\|^2 - |\mathbf{p}^* B\mathbf{z}|^2}}{|\mathbf{p}^* B\mathbf{z}|} ,$$

$$(2.46) \quad \beta = |1 - \mathbf{p}^* \mathbf{r}_{\text{in}}| .$$

*Proof.* Apply Theorem 2.1 with  $Q = A - \zeta B$ ,  $S = B$ , and take into account that  $(A - \zeta B)\mathbf{z} = B\mathbf{u} - \mathbf{r}_{\text{in}}$  to obtain the given expression for  $\beta$ . ■

To illustrate this result, we consider the same example matrices as in Section 2.2 (and  $B = I$ ), with the same starting approximate eigenvectors and values of  $\zeta$ . Here we solve (2.41) with full GMRES, and, in Figure 2.2, we show the different quantities referred to in Corollary 2.6 against the number of inner iterations. One sees that  $\beta$  converges smoothly towards 1, and that the eigenvalue residual norm  $r_{\text{eig}}$  follows  $g/(|\mathbf{p}^* \mathbf{z}| \sqrt{1 + s^2})$  until  $g \approx s/\sqrt{1 + s^2}$ . Note that  $g/(|\mathbf{p}^* \mathbf{z}| \sqrt{1 + s^2}) = \|(I - \mathbf{p}\mathbf{p}^*)\mathbf{r}_{\text{in}}\|/\|\mathbf{z}\|$ . The importance of  $\|\mathbf{z}\|$  in the convergence process is stressed in [19], where a stopping criterion based on monitoring its value is proposed. Here we see more precisely that the norm  $g$  of the projected inner residual may also play a role, that  $r_{\text{eig}}$  has asymptotic value  $s/(1 + s^2)$ , and stagnates once  $\|(I - \mathbf{p}\mathbf{p}^*)\mathbf{r}_{\text{in}}\|/\|\mathbf{z}\|$  is smaller than the latter value.

**3. Stopping criteria for JD.** We now propose a stopping strategy and give its motivation afterwards. Let  $\varepsilon^{(\text{out})}$  be the (outer) tolerance on the norm of the eigenvalue residual  $\mathbf{r}$ , or some slightly smaller value as discussed below. Let  $\alpha^{(\text{est})}$ ,  $s^{(\text{est})}$ ,  $\beta^{(\text{est})}$  be the values computed according to:

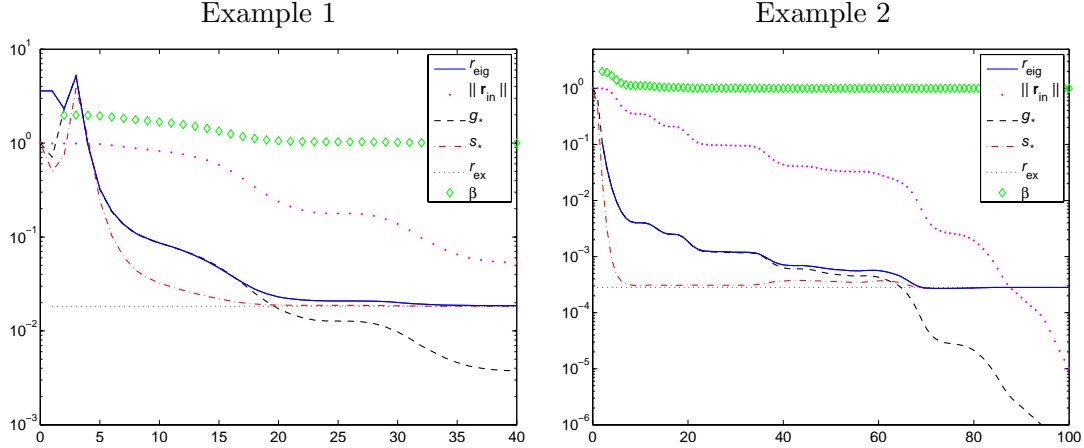


FIG. 2.2. Quantities referred to in Corollary 2.6 when solving (2.6) for Examples 1 and 2 with full GMRES;  $g_* = g/(|\mathbf{P}^* \mathbf{z}| \sqrt{1+s^2})$ ,  $s_* = s/(|\mathbf{P}^* \mathbf{z}|(1+s^2))$  and  $r_{\text{ex}}$  is the eigenvalue residual norm (2.43) corresponding to the exact solution of (2.6).

- (2.19), (2.20) with  $\alpha^{(\text{est})} = 1$  when solving a correction equation of the form (2.13) (with inner residual given by (2.15));
- (2.29), (2.30), (2.31) when solving a correction equation of the form (2.23) (with inner residual given by (2.25));
- (2.38), (2.39), (2.40) when solving a correction equation of the form (2.32) (with inner residual given by (2.34)).

For efficiency reasons, we propose to compute these values only twice in the solution of the inner system: for the first time if  $\|\mathbf{r}_{\text{in}}\| < \tau_1 \|\mathbf{r}\|$ , and we update the values if  $\|\mathbf{r}_{\text{in}}\| < \tau_2 \|\mathbf{r}\|$ , where  $\tau_1 > \tau_2$  are chosen thresholds; we recommend  $\tau_1 = 10^{-1/2}$  and  $\tau_2 = 10^{-1}$ . Let

$$g_k = \|\mathbf{r}_{\text{in}}\|$$

be the current inner residual norm (at inner iteration  $k$ ). Set

$$r_{\text{eig}}^{(k)} = \begin{cases} \sqrt{\frac{g_k^2}{1+s^{(\text{est})2}} + \left(\frac{\beta^{(\text{est})} s^{(\text{est})}}{1+s^{(\text{est})2}}\right)^2} & \text{if (2.13), (2.15) apply and } \mathbf{r}_{\text{in}} \perp \mathbf{t} \\ \frac{\sqrt{g_k^2 + \beta^{(\text{est})2}}}{\alpha^{(\text{est})} \sqrt{1+s^{(\text{est})2}}} & \text{otherwise, if } \beta^{(\text{est})} < g_k s^{(\text{est})} \\ \frac{g_k + \beta^{(\text{est})} s^{(\text{est})}}{\alpha^{(\text{est})} (1+s^{(\text{est})2})} & \text{otherwise.} \end{cases}$$

The  $g_k$  is computed (or estimated) every inner step, but the  $\alpha^{(\text{est})}$ ,  $s^{(\text{est})}$ ,  $\beta^{(\text{est})}$  are only computed twice during the inner iterations. Then we use the three following stopping criteria,

where  $\tau_3$  is another threshold (from experiments  $\tau_3 = 15$  seems a reasonable value):

- (A) **Exit if**  $g_k < \tau_1 \|\mathbf{r}\|$  **and**  $r_{\text{eig}}^{(k)} < \varepsilon^{(out)}$
- (B) **Exit if**  $g_k < \tau_1 \|\mathbf{r}\|$  **and**  $\frac{\beta^{(est)} s^{(est)}}{\alpha^{(est)} (1 + s^{(est)2})} > \frac{\varepsilon^{(out)}}{2}$  **and**  $g_k < \tau_3 \frac{\beta^{(est)} s^{(est)}}{\sqrt{1 + s^{(est)2}}}$
- Exit if**  $g_k < \tau_1 \|\mathbf{r}\|$  **and**  $\frac{\beta^{(est)} s^{(est)}}{\alpha^{(est)} (1 + s^{(est)2})} > \frac{\varepsilon^{(out)}}{2}$  **and**  $k > 1$  **and**
- (C) 
$$\begin{cases} \left(\frac{g_k}{g_{k-1}}\right)^2 > \left(2 - \left(\frac{g_{k-1}}{g_{k-2}}\right)^2\right)^{-1} & \text{if a norm minimizing method} \\ & \text{(GMRES, QMR) is used} \\ g_k > g_{k-1} & \text{otherwise .} \end{cases}$$

We first comment on the condition  $g_k < \tau_1 \|\mathbf{r}\|$ , which is present in all three stopping criteria. To start with, as explained above, this condition is required to have knowledge of  $\alpha^{(est)}$ ,  $s^{(est)}$ ,  $\beta^{(est)}$  (it is the first moment on which they are computed), and hence necessary to be able to check all of (A), (B), and (C). Also, this condition means that we ask for a minimal reduction of the relative residual error. From many test examples we conclude that it seems beneficial to avoid exiting the inner iterations too early. In particular, not exiting early may prevent stagnation in the outer loop, where the extraction phase takes place. The default threshold  $\tau_1 = 10^{-1/2}$  seems to work well in practice, but this may be dynamically decreased if stagnation is detected in the outer loop, as may for instance occur for difficult-to-find interior eigenvalues.

We remark that the residual norm of the approximate eigenpair after the outer loop may sometimes be slightly larger than  $r_{\text{eig}}^{(k)}$ . This may depend on several factors. First, on the extraction process, which works with the entire search space instead of just the span of  $\mathbf{u}$  and  $\mathbf{t}$ ; but of course, one may decide to still choose  $\mathbf{u} + \mathbf{t}$  as approximate eigenvector if (A) is met. Second, on whether or not one has the exact residual norm of the correction equation: with some linear solvers a closely related pseudo-norm of the residual may be easier to compute. Third, it may also depend on possible inaccuracies in  $\alpha$ ,  $s$ , and  $\beta$  (one may decide to recompute them if (A) holds). To prevent a situation where the residual norm of the eigenpair is larger than  $\varepsilon^{(out)}$  we may apply (A) with  $\varepsilon^{(out)}$  slightly smaller than the outer tolerance, for instance taking half of it as also suggested in [26].

The condition  $\beta^{(est)} s^{(est)} / (\alpha^{(est)} (1 + s^{(est)2})) > \varepsilon^{(out)} / 2$  present in criteria (B) and (C) ensures that we do not stop if we can solve the correction equation accurately enough to meet the main criterion (A). Consider for instance the standard eigenvalue problem where  $\alpha = 1$ . If  $\beta^{(est)} s^{(est)} / (\alpha^{(est)} (1 + s^{(est)2})) < \varepsilon^{(out)} / 2$  then from (2.22) it follows that  $r_{\text{eig}}^{(k)}$  will be smaller than  $\varepsilon^{(out)}$  if  $g_k$  is small enough.

In (B), we compare  $g_k$  with  $\beta^{(est)} s^{(est)} / \sqrt{1 + s^{(est)2}}$ . This is motivated by the observations in Section 2.2, in particular by (2.22). This is also in line with the strategy in [26] (there,  $g_k$  is compared to  $r_{\text{eig}}^{(k)}$ , but this amounts to the same comparison since (2.21) holds). Alternatively, one could use a comparison of the ratios  $g_k / g_{k-1}$  and  $r_{\text{eig}}^{(k)} / r_{\text{eig}}^{(k-1)}$  as in [13]. But, following [26], we found the present approach more effective on average. The default threshold  $\tau_3 = 15$  is not far from the one proposed in [26] (when translated to our notation) and is based on many numerical experiments. One explanation of this relatively large value is that

the convergence of Krylov subspace methods is often irregular. The threshold also “absorbs” possible inaccuracies in  $\alpha$ ,  $s$ , and  $\beta$ , that is, using a larger threshold should avoid unnecessary work in the inner iteration in nearly all cases. Criterion (B) is illustrated in Figure 2.1, where the solid vertical lines indicate the exit points (the criterion was checked with  $\alpha^{(\text{est})}$ ,  $s^{(\text{est})}$ ,  $\beta^{(\text{est})}$  equal to the values at the iteration marked by the dotted vertical line).

Criterion (C) is motivated by the possibly irregular convergence of the inner iterative method. There are two main classes of Krylov subspace linear solvers. Galerkin methods may converge irregularly with many peaks, whereas norm minimizing methods present smooth convergence curves, but may stagnate [7]. When a Galerkin type method is used, for instance CG, it may be beneficial to exit when one enters a peak, that is when  $g_k > g_{k-1}$  since it is not known in advance how many iterations will be needed before the residual decreases again. It is slightly more difficult to properly detect a stagnation in a norm minimizing method. Therefore, we use the following idea. Each such method is closely related to a Galerkin method, and the residual norm  $\hat{g}_k$  in the latter is related to the residual norm in the norm minimizing method  $g_k$  by

$$\hat{g}_k = \frac{g_k}{\sqrt{1 - \left(\frac{g_k}{g_{k-1}}\right)^2}}$$

[7, 17, 28]. From this relation, one may check that  $\hat{g}_k > \hat{g}_{k-1}$  if and only if  $\left(\frac{g_k}{g_{k-1}}\right)^2 > \left(2 - \left(\frac{g_{k-1}}{g_{k-2}}\right)^2\right)^{-1}$ . This motivates the first alternative in criterion (C).

**4. Numerical results.** All experiments have been done within the MATLAB environment (version 7.1.0.183 (R14) Service Pack 3), running under Linux on a PC with Intel Pentium 4 CPU at 3.0 GHz.

In the first experiment, we consider again the matrices *SHERMAN4* and *BandRand* described in Section 2.2. We solve the corresponding eigenvalue problem (1.1) for the smallest and the 10 smallest eigenvalues, using the standard JDQR code by Sleijpen [20], and the same code in which our stopping strategy has been implemented. All options are set to their default, except the preconditioning type: the default is left preconditioned GMRES, which does not allow to monitor the true residual norm as needed by our stopping strategy. For our modified version, we therefore select right preconditioned GMRES, and further chose the FGMRES implementation [16] (referred to as “implicit preconditioning” in [20]), which is more convenient for the computation of  $\alpha^{(\text{est})}$ ,  $s^{(\text{est})}$ ,  $\beta^{(\text{est})}$ . For the sake of completeness we include the original JDQR method with all possible preconditioning types in the comparison. Note that because GMRES is a long recurrence method, it needs either to be restarted, or stopped when a maximal number of iteration is reached, regardless of our criteria (A), (B), and (C). Our implementation uses the latter option with the maximal number of iterations set to 15.

The results are given in Table 4.1. One sees that our stopping strategy does not reduce the total number of matrix vector products. The default stopping criterion in JDQR exits quite quickly (in practice the method nearly always uses 5 inner iterations), thus it tends to do fewer inner iterations and more outer ones, trusting on the extraction process in the outer loop. However, such subspace iterations are much more costly than inner GMRES iterations, despite the fact that GMRES is a relatively costly method. Therefore, the code is much faster with our stopping strategy.

Stopping test	Type prec.	1 eigenvalue			10 eigenvalues		
		#MV	#Prec. Sol.	Time	#MV	#Prec. Sol.	Time
Matrix <i>SHERMAN4</i>							
(A),(B),(C)	Right (F)	34	27	0.18	251	235	1.77
standard	Right (F)	34	27	0.36	225	209	3.58
standard	Right	34	36	0.37	225	286	3.58
standard	Left	36	37	0.34	247	304	3.83
(A),(B),(C)	No	157	-	0.45	868	-	3.21
standard	No	180	-	1.10	825	-	6.00
Matrix <i>BandRand</i>							
(A),(B),(C)	No	125	-	0.43	757	-	3.21
standard	No	103	-	0.68	596	-	5.03

TABLE 4.1

Numerical result for the JDQR code with and without our stopping strategy; whenever used, the preconditioning is an incomplete LU preconditioning as obtained with the MATLAB function `luinc` with drop tolerance set to  $10^{-2}$  (such a LU preconditioning makes no sense for the lower triangular matrix *BandRand*); #MV is the number of needed matrix vector multiplications, and #Prec. Sol. is the number of times the preconditioner is applied; Time is elapsed CPU time in seconds; RIGHT (LEFT) means right (left) preconditioned GMRES; (F) means FGMRES implementation.

In the second experiment, we consider generalized Hermitian eigenproblems from the matrix market [11]. Here we compare JDQZ [20], the version of JDQR for the GEP, and JDRPCG\_GEP [12]. This code is a further version of the JDCG code [12] adapted for generalized Hermitian eigenproblems. Note that JDCG was itself built as a simplified version of JDQR for the particular case of Hermitian eigenproblems [13]. JDRPCG\_GEP implements our stopping strategy, uses CG for inner iterations, and, unlike standard implementations, it uses a regular preconditioning for the correction equation (2.23) (not a projected one, see [14] for details). The matrices have moderate size ( $n = 1074$  for *BCSSTK08* and  $n = 2003$  for *BCSSTK13*) but exhibit bad conditioning (the ratio between the largest diagonal element and the smallest eigenvalue is close to  $10^9$ ). For this reason, we do not use JDQZ with the default tolerance ( $10^{-8}$ ), but use  $10^{-6}$  instead. Otherwise, we use all default options in JDQZ, except that we set the “TestSpace” parameter to “SearchSpace”, because this is favorable for generalized Hermitian eigenproblems (JDQZ does not converge with the default choice). This makes also the outer loop similar to the one in JDRPCG\_GEP: both methods use standard Rayleigh–Ritz, making the comparison fair. The JDRPCG\_GEP code is also used with all options to their default, except the tolerance which is adapted to guarantee the same accuracy as obtained with JDQZ (both codes use different criteria to check the outer convergence).

The results are given in Table 4.2. For the second problem, we selected a smaller drop tolerance for LU preconditioning because otherwise JDQZ does not converge with any of the tested options. Note that here and in the table, “no convergence” means stagnation of the outer loop. JDRPCG\_GEP is more robust and always converge if one allows enough matrix vector products. It is also much faster, especially on the second test problem which seems more difficult than the first one.



Code	Type prec.	1 eigenvalue			10 eigenvalues		
		#MV	#Prec. Sol.	Time	#MV	#Prec. Sol.	Time
Matrices <i>BCSSTK08</i> , <i>BCSSTM08</i>							
JDRPCG_GEP	CG like	155	148	0.33	1514	1445	3.31
JDQZ	Right (F)	148	147	0.85	1535	1543	12.3
JDQZ	Right	148	174	0.83	1313	1541	10.9
JDQZ	Left	113	132	0.69	1457	1707	11.5
JDRPCG_GEP	No	12424	-	5.27	(5 eigs found in $10^5$ iter.)		
JDQZ	No	(no convergence)			(no convergence)		
Matrices <i>BCSSTK13</i> , <i>BCSSTM13</i>							
JDRPCG_GEP	CG like	242	229	1.34	1616	1524	9.44
JDQZ	Right (F)	(no convergence)			(no convergence)		
JDQZ	Right	(no convergence)			(no convergence)		
JDQZ	Left	555	649	8.50	3290	3847	57.5

TABLE 4.2

Numerical result for the *JDRPCG\_GEP* and *JDQZ* codes; whenever used, the preconditioning is an incomplete LU preconditioning as obtained with the MATLAB function `luinc` with drop tolerance set to  $10^{-2}$  for the first problem and  $10^{-3}$  for the second problem; #MV is the number of needed matrix vector multiplications, and #Prec. Sol. is the number of times the preconditioner is applied; Time is elapsed CPU time in seconds; RIGHT (LEFT) means right (left) preconditioned GMRES; (F) means FGMRES implementation.

**5. Conclusions.** We considered the JD method for several types of eigenvalue problems. In each case, we showed that the (outer) eigenvalue residual norm may be estimated as

$$r_{\text{eig}} \approx \max \left( \frac{\|\mathbf{r}_{\text{in}}\|}{\alpha \sqrt{1+s^2}}, \frac{\beta s}{\alpha (1+s^2)} \right),$$

where  $\mathbf{r}_{\text{in}}$  is the residual of the correction equation, and where  $\alpha, \beta, s$  are scalar parameters easy to compute or estimate while solving this equation iteratively. Here  $\frac{\beta s}{\alpha (1+s^2)}$  converges quickly to the eigenvalue residual norm obtained with the exact solution of the correction equation, that is, with an exact RQI or inverse iteration step, according to the choice of the shift. Moreover, close to the solution  $\alpha \approx 1$ .

Hence, when solving the correction equation iteratively, one has  $r_{\text{eig}} \approx \frac{\|\mathbf{r}_{\text{in}}\|}{\alpha \sqrt{1+s^2}}$  until a certain point where progress is no longer possible since  $r_{\text{eig}}$  reached its asymptotic value  $\frac{\beta s}{\alpha (1+s^2)}$ , so that a further reduction of  $\|\mathbf{r}_{\text{in}}\|$  is useless. From a theoretical point of view, this shows how and why the JD method converges with an inexact solution of the correction equation. From a practical point of view, this makes it possible to set up stopping criteria for inner iterations that take into account the progress in the outer convergence. We proposed such a strategy and numerical experiments suggest that it can improve both the robustness and the efficiency of the JD method.

## REFERENCES

- [1] Z. BAI, J. DEMMEL, J. DONGARRA, A. RUHE, AND H. A. VAN DER VORST, eds., *Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide*, SIAM, Philadelphia, PA, 2000.
- [2] J. BERNS-MÜLLER, I. G. GRAHAM, AND A. SPENCE, *Inexact inverse iteration for symmetric matrices*, Linear Algebra Appl., 416 (2006), pp. 389–413.
- [3] A. EDELMAN, T. A. ARIAS, AND S. T. SMITH, *The geometry of algorithms with orthogonality constraints*, SIAM J. Matrix Anal. Appl., 20 (1999), pp. 303–353.
- [4] D. FOKKEMA, G. SLEIJPEN, AND H. A. VAN DER VORST, *Jacobi–Davidson style QR and QZ algorithms for the reduction of matrix pencils*, SIAM J. Sci. Comput., 20 (1999), pp. 94–125.
- [5] R. W. FREUND AND N. M. NACHTIGAL, *Software for simplified Lanczos and QMR algorithms*, Appl. Num. Math., 19 (1995), pp. 319–341.
- [6] G. H. GOLUB AND Q. YE, *Inexact inverse iterations for generalized eigenvalue problems*, BIT, 40 (2000), pp. 672–684.
- [7] A. GREENBAUM, *Iterative Methods for Solving Linear Systems*, vol. 17 of Frontiers in Applied Mathematics, SIAM, Philadelphia, PA, 1997.
- [8] M. R. HESTENES AND E. STIEFEL, *Methods of conjugate gradients for solving linear systems*, J. Res. Nat. Bur. Standards, 49 (1952), pp. 409–436.
- [9] M. E. HOCHSTENBACH AND Y. NOTAY, *Homogeneous Jacobi–Davidson*, Tech. Report GANMN 06–04, Université Libre de Bruxelles, Brussels, Belgium, 2006. <http://homepages.ulb.ac.be/~ynotay>.
- [10] Y.-L. LAI, K.-Y. LIN, AND W.-W. LIN, *An inexact inverse iteration for large sparse eigenvalue problems*, Numer. Lin. Alg. Appl., 4 (1997), pp. 425–437.
- [11] *The Matrix Market*. <http://math.nist.gov/MatrixMarket>, a repository for test matrices.
- [12] Y. NOTAY, *JDCG, JDRPCG and JDRPCG\_GEP codes*. Available via <http://homepages.ulb.ac.be/~ynotay/>.
- [13] ———, *Combination of Jacobi–Davidson and conjugate gradients for the partial symmetric eigenproblem*, Numer. Lin. Alg. Appl., 9 (2002), pp. 21–44.
- [14] ———, *Inner iterations in eigenvalue solvers*, Tech. Report GANMN 05–01, Université Libre de Bruxelles, Brussels, Belgium, 2005. <http://homepages.ulb.ac.be/~ynotay>.
- [15] B. N. PARLETT, *The Symmetric Eigenvalue Problem*, SIAM, Philadelphia, PA, 1998. Corrected reprint of the 1980 original.
- [16] Y. SAAD, *A flexible inner-outer preconditioned GMRES algorithm*, SIAM J. Sci. Comput., 14 (1993), pp. 461–469.
- [17] ———, *Iterative Methods for Sparse Linear Systems*, SIAM, Philadelphia, PA, 2003. Second ed.
- [18] Y. SAAD AND M. H. SCHULTZ, *GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 856–869.
- [19] V. SIMONCINI AND L. ELDÉN, *Inexact Rayleigh quotient-type methods for eigenvalue computations*, BIT, 42 (2002), pp. 159–182.
- [20] G. SLEIJPEN, *JDQR and JDQZ codes*. Available via <http://www.math.uu.nl/people/sleijpen>.
- [21] G. SLEIJPEN AND H. A. VAN DER VORST, *A Jacobi–Davidson iteration method for linear eigenvalue problems*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 401–425.
- [22] ———, *The Jacobi–Davidson method for eigenvalue problems and its relation with accelerated inexact Newton schemes*, in Iterative Methods in Linear Algebra II, S. Margenov and P. S. Vassilevski, eds., Series in Computational and Applied Mathematics Vol.3, IMACS, 1996, pp. 377–389.
- [23] ———, *A Jacobi–Davidson iteration method for linear eigenvalue problems*, SIAM Review, 42 (2000), pp. 267–293.
- [24] ———, *Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide*, in Bai et al. [1], 2000, ch. 5.6.
- [25] P. SMIT AND M. H. C. PAARDEKOOPER, *The effects of inexact solvers in algorithms for symmetric eigenvalue problems*, Linear Algebra Appl., 287 (1999), pp. 337–357.
- [26] A. STATHOPOULOS, *Nearly optimal preconditioned methods for Hermitian eigenproblems under limited memory. Part I: Seeking one eigenvalue*. Preprint, Department of Computer Science, College of William & Mary, Williamsburg, USA, 2005.
- [27] D. B. SZYLD, *Criteria for combining inverse and Rayleigh quotient iteration*, SIAM J. Numer. Anal., (1988), pp. 1369–1375.
- [28] H. A. VAN DER VORST, *Computational methods for large eigenvalue problems*, in Handbook of numerical analysis, Vol. VIII, North-Holland, Amsterdam, 2002, pp. 3–179.

- [29] H. A. VAN DER VORST, *Iterative Krylov Methods for Large Linear systems*, Cambridge University Press, Cambridge, 2003.
- [30] K. WU, Y. SAAD, AND A. STATHOPOULOS, *Inexact Newton preconditioning techniques for large symmetric eigenvalue problems*, *Electronic Trans. Numer. Anal.*, 7 (1998), pp. 202–214.