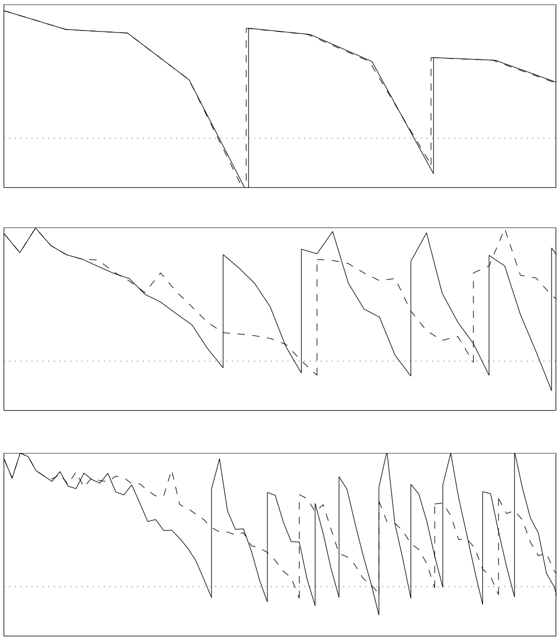

Universiteit Utrecht

Mathematical Institute

Alternative correction equations in the Jacobi-Davidson method

by

Menno Genseberger and Gerard L. G. Sleijpen

Preprint

No. 1073

June 1998, revised: March 1999

Alternative correction equations in the Jacobi-Davidson method

Menno Genseberger[*] Gerard L. G. Sleijpen[‡]

June 1998, revised: March 1999

Abstract

The correction equation in the Jacobi-Davidson method is effective in a subspace orthogonal to the current eigenvector approximation, whereas for the continuation of the process only vectors orthogonal to the search subspace are of importance. Such a vector is obtained by orthogonalizing the (approximate) solution of the correction equation against the search subspace. As an alternative, a variant of the correction equation can be formulated that is restricted to the subspace orthogonal to the current search subspace. In this paper, we discuss the effectiveness of this variant.

Our investigation is also motivated by the fact that the restricted correction equation can be used for avoiding stagnation in case of defective eigenvalues. Moreover, this equation plays a key role in the inexact TRQ method [18].

Keywords: Eigenvalues and eigenvectors, Jacobi-Davidson method

AMS subject classification: 65F15, 65N25

1 Introduction

For the computation of a few eigenvalues with associated eigenvectors of large n-dimensional linear eigenvalue problems

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x} \tag{1}$$

subspace methods have become very popular. The application of a subspace method is attractive when the method is able to calculate accurate solutions to (1) from relatively low dimensional subspaces, i.e. $m \ll n$ with m the dimension of the subspace. Keeping m small enables a reduction in computational time and memory usage.

There are many ways to construct a subspace and different options are possible for a subspace method. Globally three stages can be distinguished in such a method:

- Calculation of an approximation to the eigenpair inside the search subspace.
- Computation of new information about the behaviour of operator $\mathbf{A}$.
- Expansion of the search subspace with vector(s) containing this information.

[*]Mathematical Institute, Utrecht University and CWI, Amsterdam.
[‡]Mathematical Institute, Utrecht University.

P.O. Box 80.010, NL-3508 TA Utrecht, The Netherlands. E-mail: `genseber@math.uu.nl`, `sleijpen@math.uu.nl`.

In the Jacobi-Davidson method [15], Sleijpen and Van der Vorst propose to look for new information in the space orthogonal to the approximate eigenvector. A correction equation

$$(\mathbf{I}_n - \mathbf{u}_m\mathbf{u}_m^*)(\mathbf{A} - \theta_m\mathbf{I}_n)(\mathbf{I}_n - \mathbf{u}_m\mathbf{u}_m^*)\mathbf{t} = -\mathbf{r}_m, \tag{2}$$

is defined on this space. Here $(\theta_m, \mathbf{u}_m)$ is the current approximate eigenpair with residual $\mathbf{r}_m \equiv \mathbf{A}\mathbf{u}_m - \theta_m\mathbf{u}_m$. A correction $\mathbf{t}$ to the approximate eigenvector $\mathbf{u}_m$ is obtained by solving (2) approximately. Then the search subspace $\mathcal{V}_m$ is expanded to $\mathcal{V}_{m+1}$ with the component of $\mathbf{t}$ orthogonal to $\mathcal{V}_m$. One of the eigenvalues θ_{m+1} of the projection of matrix $\mathbf{A}$ on the new search subspace is selected. Inside $\mathcal{V}_{m+1}$ the so-called Ritz pair $(\theta_{m+1}, \mathbf{u}_{m+1})$ is considered to be an optimal approximation to the wanted eigenpair $(\lambda, \mathbf{x})$.

As the residual of a Ritz pair is orthogonal to the subspace this special choice of the approximation introduces some freedom for the projection of the correction equation. Another possibility is looking for a correction in the space orthogonal to the search subspace constructed so far. If the Ritz pair is indeed the "best" approximation inside the search subspace, then we should expect that really new information lies in the orthogonal complement of $\mathcal{V}_m$. This suggests a more restrictive correction equation

$$(\mathbf{I}_n - \mathbf{V}_m\mathbf{V}_m^*)(\mathbf{A} - \theta_m\mathbf{I}_n)(\mathbf{I}_n - \mathbf{V}_m\mathbf{V}_m^*)\mathbf{t} = -\mathbf{r}_m, \tag{3}$$

that will be investigated here. In equation (3), $\mathbf{V}_m$ is an n by m matrix of which the columns form an orthonormal basis of the current search subspace $\mathcal{V}_m$.

Although the approach in (3) does not seem to be unnatural, it is not clear whether it is more effective in practical computations. In general, the solutions of (2) and (3) do not lead to the same expansion of the search subspaces. Therefore, a different convergence behaviour of the Jacobi-Davidson process is to be expected.

The projection in (3) is more expensive, but the method for solving the correction equation may profit from projecting on a smaller subspace. To see this, note that $\mathbf{A} - \theta_m\mathbf{I}_n$ is nearly singular if $\theta_m \approx \lambda$. Restricting $\mathbf{A} - \theta_m\mathbf{I}_n$ to the space orthogonal to the approximate eigenvector $\mathbf{u}_m$ will give a well-conditioned operator in case λ is simple and fairly well isolated from the other eigenvalues. Projecting on the space orthogonal to $\mathcal{V}_m$ may further improve the conditioning. If eigenvalues cluster around the target eigenvalue λ then the associated eigenvectors should be removed as well. The search subspace may be expected to contain good approximations also of these eigenvectors [8, §3.4] and projecting on the space orthogonal to $\mathcal{V}_m$ may lead to a well-conditioned operator also in case of clustering eigenvectors. A reduction may be expected in the number of steps that are needed to solve the correction equation to a certain accuracy if an iterative linear solver is used. It also improves the stability of the linear solver. These effects may compensate for the more expensive steps. For precisely these reasons, a strategy is followed in [6, 4] where $\mathbf{u}_m$ in (2) is replaced by the matrix of all Ritz vectors that could be associated with eigenvalues in a cluster near the target eigenvalue.

GMRESR[1] [21] and GCRO[2] [3] are nested methods for solving linear systems $\mathbf{A}\mathbf{x} = \mathbf{b}$ iteratively. They both use GCR in the "outer loop" to update the approximate solution and GMRES in the "inner loop" to compute a new search direction from a correction equation. As argued in [7], Jacobi-Davidson with (2) can be viewed as the eigenvalue version of GMRESR, while Jacobi-Davidson with (3) is the analogue of GCRO. GCRO employs the search subspace to improve the convergence of GMRES for the solution of a correction equation (see also [2]). Experiments in [3, 1] for linear systems of equations show that GCRO can be more effective than GMRESR: for linear problems it appears to be worthwhile to use more expensive projections. Is this also the case for eigenvalue problems? If, for a linear system, the correction equation is solved exactly then both GMRESR and GCRO produce the exact solution of the linear system in the next step. However, eigenvalue problems are not linear and even if all correction equations are

[1] Generalized Minimum Residual Recursive
[2] Generalized Conjugate Residual with Orthogonalization in the inner iteration

solved exactly still a number of steps may be needed to find accurate approximations of an eigenpair. Replacing $\mathbf{u}_m$ in (2) by $\mathbf{V}_m$ may lead to an increase in the number of iteration steps. The loss in speed of convergence may not be compensated by the advantage of a better conditioned correction equation (3). In practical computations the situation is even more complicated since the correction equations will be solved only with a modest accuracy.

Jacobi-Davidson itself may also profit from projecting on a smaller subspace. If the Ritz value is a defective eigenvalue of the interaction matrix $\mathbf{V}_m^* \mathbf{A} \mathbf{V}_m$ then the correction equation (2) may have a solution in the current search subspace. In such a case the search subspace is not expanded and Jacobi-Davidson stagnates. Correction equation (3) will give a proper expansion vector and stagnation can be avoided [16]. In practical computations, where the correction equations are not solved exactly, it is observed that stagnation also can be avoided by a strategical and occasional use of (3).

Equation (3) also plays a key role in the inexact Truncated RQ iteration [18] of Sorensen and Yang (see also §§2.3 and 4.1). This provides another motivation for studying the effect of using (3) in Jacobi-Davidson.

This paper is organized as follows. First, in §2 we recall some facts about projecting the eigenvalue problem. An alternative derivation of a more general correction equation is given to motivate the correction equation (3). It appears that (3) and the original correction equation (2) are the extremal cases of this general correction equation. Next, in §3, an illustration is given in which the two correction equations can produce different results. We will show that, if the process is started with a Krylov subspace then the two exact solutions of the correction equations lead to mathematically equivalent results (§4). We will also argue that in other situations the correction equation (3) will lead to slower convergence. In §5 we conclude with some numerical experiments; partially as an illustration of the preceding, partially to observe what happens if things are not computed in high precision and whether round-off errors play a role of importance.

2 The framework: the Jacobi-Davidson method

We start with a brief summary of the Rayleigh-Ritz procedure. This procedure, where the large eigenvalue problem is projected on a small one, serves as a starting point for the derivation of a more general correction equation. We will consider the two extremal cases of this equation. One corresponds to the correction equation of the original Jacobi-Davidson method, the other one is employed in the inexact Truncated RQ iteration.

2.1 Interpolation: Rayleigh-Ritz procedure

Suppose some m-dimensional subspace $\mathcal{V}_m$ is available. Let $\mathbf{V}_m$ be an $n \times m$ dimensional matrix such that the column-vectors of $\mathbf{V}_m$ form an orthonormal basis of $\mathcal{V}_m$. The orthogonal projection of $\mathbf{A}$ on the subspace (the Rayleigh quotient or interaction matrix) will then be $H_m \equiv \mathbf{V}_m^* \mathbf{A} \mathbf{V}_m$.

Furthermore suppose that we selected a Ritz pair $(\theta_m, \mathbf{u_m})$ of $\mathbf{A}$ with respect to $\mathcal{V}_m$, i.e. a scalar θ_m and a vector $\mathbf{u}_m \in \mathcal{V}_m$ such that the residual $\mathbf{r}(\theta_m, \mathbf{u}_m) \equiv \mathbf{r}_m \equiv \mathbf{A}\mathbf{u}_m - \theta_m \mathbf{u}_m$ is orthogonal to $\mathcal{V}_m$. A Ritz pair can considered to be an optimal approximation inside the subspace to an eigenpair $(\lambda, \mathbf{x})$ of the matrix $\mathbf{A}$ in some sense (in [12, §11.4] this is argued for the real symmetric case).

The Ritz values are equal to the eigenvalues of H_m. Therefore they can be computed by solving the m-dimensional linear eigenvalue problem $H_m s = \theta s$. The Ritz vector associated with θ is $\mathbf{V}_m s$.

2.2 Extrapolation: correction equation

How well does the Ritz pair $(\theta_m, \mathbf{u}_m)$ approximate an eigenpair $(\lambda, \mathbf{x})$ of matrix $\mathbf{A}$? With a view restricted to the subspace there would be no better alternative. But outside $\mathcal{V}_m$ a remainder $\mathbf{r}_m$ is left. The norm of this residual gives an indication about the quality of the approximation. Let us try to minimize this norm.

For that purpose, consider $\mathbf{u}' = \mathbf{u}_m + \mathbf{t}$ and $\theta' = \theta_m + \varepsilon$. Define the residual $\mathbf{r}' \equiv \mathbf{A}\mathbf{u}' - \theta'\mathbf{u}' = \mathbf{r}_m + \mathbf{A}\mathbf{t} - \theta_m\mathbf{t} - \varepsilon\mathbf{u}_m - \varepsilon\mathbf{t}$. If we view ε and $\mathbf{t}$ as first order corrections then $\varepsilon\mathbf{t}$ represents some second order correction (cf. [11], [19]). Ignoring this contribution results in

$$\mathbf{r}' = \mathbf{r}_m + (\mathbf{A} - \theta_m\mathbf{I}_n)\mathbf{t} - \varepsilon\mathbf{u}_m. \tag{4}$$

Consider some subspace $\mathcal{W}$ such that $\mathbf{u}_m \in \mathcal{W} \subseteq \mathcal{V}_m$. With $\mathbf{W}$, a matrix of which the column-vectors form an orthonormal basis for $\mathcal{W}$, we decompose (4) (cf. [14]) in

$$\mathbf{W}\mathbf{W}^*\mathbf{r}' = \mathbf{W}\mathbf{W}^*(\mathbf{A} - \theta_m\mathbf{I}_n)\mathbf{t} - \varepsilon\mathbf{u}_m,$$

the component of $\mathbf{r}'$ in $\mathcal{W}$, and in

$$(\mathbf{I}_n - \mathbf{W}\mathbf{W}^*)\mathbf{r}' = (\mathbf{I}_n - \mathbf{W}\mathbf{W}^*)(\mathbf{A} - \theta_m\mathbf{I}_n)\mathbf{t} + \mathbf{r}_m, \tag{5}$$

the component of $\mathbf{r}'$ orthogonal to $\mathcal{W}$.

The new direction $\mathbf{t}$ will be used to expand the subspace $\mathcal{V}_m$ to $\mathcal{V}_{m+1}$. An approximation $(\theta_{m+1}, \mathbf{u}_{m+1})$ is computed with respect to $\mathcal{V}_{m+1}$. Because $\mathcal{W} \subseteq \mathcal{V}_m \subseteq \mathcal{V}_{m+1}$ the residual $\mathbf{r}_{m+1}$ of this Ritz pair is also orthogonal to $\mathcal{W}$. This means that if we write $(\theta_{m+1}, \mathbf{u}_{m+1}) = (\theta_m + \varepsilon, \mathbf{u}_m + \mathbf{t})$ then only (5) gives a contribution to the norm of $\mathbf{r}_{m+1}$:

$$\|\mathbf{r}_{m+1}\| = \|(\mathbf{I}_n - \mathbf{W}\mathbf{W}^*)(\mathbf{A} - \theta_m\mathbf{I}_n)\mathbf{t} + \mathbf{r}_m\|. \tag{6}$$

So to get a smaller norm in the next step we should calculate $\mathbf{t}$ such that

$$(\mathbf{I}_n - \mathbf{W}\mathbf{W}^*)(\mathbf{A} - \theta_m\mathbf{I}_n)\mathbf{t} = -\mathbf{r}_m. \tag{7}$$

Note that if $\mathbf{t} = \mathbf{u}_m$ then there is no expansion of the search space. So it can be assumed that $\mathbf{t} \neq \mathbf{u}_m$. As we are free to scale $\mathbf{u}_m$ to any length, we can require that $\mathbf{t} \perp \mathbf{u}_m$. From this it follows that if $\mathbf{t} \neq \mathbf{u}_m$ then equation (7) and

$$(\mathbf{I}_n - \mathbf{W}\mathbf{W}^*)(\mathbf{A} - \theta_m\mathbf{I}_n)(\mathbf{I}_n - \mathbf{u}_m\mathbf{u}_m^*)\mathbf{t} = -\mathbf{r}_m \tag{8}$$

can considered to be equivalent.

Drawback may be that the linear systems in (7) and (8) are underdetermined. The operators $(\mathbf{I}_n - \mathbf{W}\mathbf{W}^*)(\mathbf{A} - \theta_m\mathbf{I}_n)$ and $(\mathbf{I}_n - \mathbf{W}\mathbf{W}^*)(\mathbf{A} - \theta_m\mathbf{I}_n)(\mathbf{I}_n - \mathbf{u}_m\mathbf{u}_m^*)$ map $\mathbf{t}$ on a lower dimensional subspace $\mathcal{W}$. The operator $(\mathbf{I}_n - \mathbf{W}\mathbf{W}^*)(\mathbf{A} - \theta_m\mathbf{I}_n)(\mathbf{I}_n - \mathbf{W}\mathbf{W}^*)$ acts only inside the space orthogonal to $\mathcal{W}$. We expect this operator to have a more favourable distribution of eigenvalues for the iterative method. In that case the correction equation reads

$$(\mathbf{I}_n - \mathbf{W}\mathbf{W}^*)(\mathbf{A} - \theta_m\mathbf{I}_n)(\mathbf{I}_n - \mathbf{W}\mathbf{W}^*)\mathbf{t} = -\mathbf{r}_m. \tag{9}$$

If the correction equation is solved (approximately) by a Krylov subspace method where the initial guess is $\mathbf{0}$, then no difference will be observed between (7) and (9). The reason why is that $(\mathbf{I}_n - \mathbf{W}\mathbf{W}^*)^2 = \mathbf{I}_n - \mathbf{W}\mathbf{W}^*$.

2.3 Extremal cases

After m steps of the subspace method, $\mathcal{V}_m$ contains besides $\mathbf{u}_m$, $m-1$ other independent directions. Consequence: different subspaces $\mathcal{W}$ can be used in equation (7) provided that $\mathrm{span}(\mathbf{u}_m) \subseteq \mathcal{W} \subseteq \mathcal{V}_m$. Here we will consider the extremal cases $\mathcal{W} = \mathrm{span}(\mathbf{u}_m)$ and $\mathcal{W} = \mathcal{V}_m$.

The first case corresponds with the original Jacobi-Davidson method [15]:

$$(\mathbf{I}_n - \mathbf{u}_m\mathbf{u}_m^*)(\mathbf{A} - \theta_m\mathbf{I}_n)(\mathbf{I}_n - \mathbf{u}_m\mathbf{u}_m^*)\mathbf{t} = -\mathbf{r}_m.$$

The operator in this equation can be seen as a mapping in the orthogonal complement of $\mathbf{u}_m$.

Let us motivate the other case. Suppose $\mathcal{W}$ is a subspace contained in, but not equal to $\mathcal{V}_m$. Then $(\mathbf{I}_n - \mathbf{W}\mathbf{W}^*)$ projects still some components of $(\mathbf{A} - \theta_m\mathbf{I}_n)\mathbf{t}$ inside $\mathcal{V}_m$. These components will not contribute to a smaller norm in (6). To avoid this overhead of already known information it is tempting to take $\mathcal{W} = \mathcal{V}_m$:

$$(\mathbf{I}_n - \mathbf{V}_m\mathbf{V}_m^*)(\mathbf{A} - \theta_m\mathbf{I}_n)(\mathbf{I}_n - \mathbf{u}_m\mathbf{u}_m^*)\mathbf{t} = -\mathbf{r}_m. \tag{10}$$

Furthermore, if $\mathcal{W} = \mathcal{V}_m$ then equation (9) becomes

$$(\mathbf{I}_n - \mathbf{V}_m\mathbf{V}_m^*)(\mathbf{A} - \theta_m\mathbf{I}_n)(\mathbf{I}_n - \mathbf{V}_m\mathbf{V}_m^*)\mathbf{t} = -\mathbf{r}_m.$$

In the following with JD and JDV we will denote the Jacobi-Davidson method which uses (2) and (3) respectively as correction equation. The *exact* solution of (2) will be denoted by $\mathbf{t}_{\text{JD}}$, while $\mathbf{t}_{\text{JDV}}$ denotes the *exact* solution of (3). With an "*exact*" process we refer to a process in exact arithmetic in which all correction equations are solved exactly. Note that both $\mathbf{t}_{\text{JD}}$ and $\mathbf{t}_{\text{JDV}}$ are solutions of (10). As we will illustrate in an example in §3, the solution set of (10) may consist of more than two vectors. In fact, this set will be an affine space of dimension dim$(\mathcal{V}_m)$, while generally (2) and (3) will have unique solutions. For this reason, we will refer to equation (10) as the "in between" equation.

An equation similar to (3) appears in the truncated RQ-iteration of Sorensen and Yang [18]. In every step of this method the solution of the so-called TRQ equations is required. For the application of an iterative solver the authors recommend to use

$$(\mathbf{I}_n - \mathbf{V}_m\mathbf{V}_m^*)(\mathbf{A} - \mu\mathbf{I}_n)(\mathbf{I} - \mathbf{V}_m\mathbf{V}_m^*)\hat{\mathbf{w}} = \mathbf{f}_m \tag{11}$$

instead of the TRQ equations. Here μ is some shift which may be chosen to be fixed for some TRQ-iteration steps whereas in Jacobi-Davidson θ_m is an optimal shift which differs from step to step. Also here Sorensen and Yang expect (11) to give better results due to the fact that $(\mathbf{I}_n - \mathbf{V}_m\mathbf{V}_m^*)(\mathbf{A} - \mu\mathbf{I}_n)(\mathbf{I} - \mathbf{V}_m\mathbf{V}_m^*)$ has a more favourable eigenvalue distribution than $\mathbf{A} - \mu\mathbf{I}$ when μ is near an eigenvalue of $\mathbf{A}$ (see also the remark at the end of §4.1).

2.4 Convergence rate

The derivation in §2.2 of the alternative correction equations may suggest that expansion with an exact solution $\mathbf{t}$ of (10) would result in quadratic convergence (cf. [17]) like the original Jacobi-Davidson method ([15, §4.1], [14, Th.3.2]). Let us take a closer look.

As in §2.2, consider the residual $\mathbf{r}_{m+1}$ associated with $(\theta_{m+1}, \mathbf{u}_{m+1}) = (\theta_m + \varepsilon, \mathbf{u}_m + \mathbf{t})$.
If $\mathbf{t} \perp \mathbf{u}_m$ is the exact solution of (2) and ε is chosen such that $\mathbf{r}_{m+1}$ is orthogonal to $\mathbf{u}_m$ then it can be checked that $\mathbf{r}_{m+1}$ is equal to a quadratic term ($\mathbf{r}_{m+1} = -\varepsilon\mathbf{t}$), which virtually proves quadratic convergence. (Note: we are dealing not only with the directions $\mathbf{u}_m$ and $\mathbf{t}$ but with a search subspace from which the new approximation is computed, there could be an update for $\mathbf{u}_m$ that is even better than $\mathbf{t}$.)
If $\mathbf{t}$ solves (10) exactly then, by construction, the component of the residual orthogonal to $\mathcal{V}_m$ consists of a second order term. However, generally the component of $\mathbf{r}_{m+1}$ in the space $\mathcal{V}_m$ contains first order terms (see §3) and updating $\mathbf{u}_m$ with this exact solution $\mathbf{t}$ of (10) does not lead to quadratic convergence. One may hope for better updates in the space spanned by $\mathcal{V}_m$ and $\mathbf{t}$, but, as we will see in our numerical experiments in §5.1.1, equation (3), and therefore also (10), do not lead to quadratic convergence in general.

3 Two examples

The two following simple examples give some insight into the differences between the three correction equations (2), (10), and (3).

3.1 Different expansion of the subspace

Consider the following matrix

$$\mathbf{A} = \begin{pmatrix} 0 & \beta & \mathbf{c}_1^* \\ 0 & \alpha & \mathbf{c}_2^* \\ \mathbf{d}_1 & \mathbf{d}_2 & \mathbf{B} \end{pmatrix},$$

with α and β scalars, $\mathbf{c}_1, \mathbf{c}_2, \mathbf{d}_1$ and $\mathbf{d}_2$ vectors and $\mathbf{B}$ a non-singular matrix of appropriate size.

Suppose we already constructed the subspace $\mathcal{V}_2 = \mathrm{span}(\mathbf{e}_1, \mathbf{e}_2)$ and the selected Ritz vector $\mathbf{u}_2$ is $\mathbf{e}_1$. Then the associated Ritz value θ_2 equals 0,

$$\mathbf{r}_2 = \begin{pmatrix} 0 \\ 0 \\ \mathbf{d}_1 \end{pmatrix},$$

while $(\mathbf{I}-\mathbf{e}_1\mathbf{e}_1^*)\mathbf{A}(\mathbf{I}-\mathbf{e}_1\mathbf{e}_1^*)$, $(\mathbf{I}-\mathbf{V}_2\mathbf{V}_2^*)\mathbf{A}(\mathbf{I}-\mathbf{e}_1\mathbf{e}_1^*)$, and $(\mathbf{I}-\mathbf{V}_2\mathbf{V}_2^*)\mathbf{A}(\mathbf{I}-\mathbf{V}_2\mathbf{V}_2^*)$ are equal to

$$\begin{pmatrix} 0 & 0 & \mathbf{0}^* \\ 0 & \alpha & \mathbf{c}_2^* \\ \mathbf{0} & \mathbf{d}_2 & \mathbf{B} \end{pmatrix}, \quad \begin{pmatrix} 0 & 0 & \mathbf{0}^* \\ 0 & 0 & \mathbf{0}^* \\ \mathbf{0} & \mathbf{d}_2 & \mathbf{B} \end{pmatrix}, \quad \text{and} \quad \begin{pmatrix} 0 & 0 & \mathbf{0}^* \\ 0 & 0 & \mathbf{0}^* \\ \mathbf{0} & \mathbf{0} & \mathbf{B} \end{pmatrix},$$

respectively. From this it is seen that JD computes its correction from

$$\begin{pmatrix} \alpha & \mathbf{c}_2^* \\ \mathbf{d}_2 & \mathbf{B} \end{pmatrix} \begin{pmatrix} \gamma \\ \mathbf{t}' \end{pmatrix} = - \begin{pmatrix} 0 \\ \mathbf{d}_1 \end{pmatrix},$$

the “in between” from

$$\begin{pmatrix} \mathbf{d}_2 & \mathbf{B} \end{pmatrix} \begin{pmatrix} \gamma \\ \mathbf{t}' \end{pmatrix} = -\mathbf{d}_1,$$

and JDV from

$$\mathbf{B}\mathbf{t}' = -\mathbf{d}_1.$$

Let $\mathbf{t}_i'$ be the solution of $\mathbf{B}\mathbf{t}_i' = -\mathbf{d}_i$ $(i = 1, 2)$. Then the component of $\mathbf{t}_{\text{JDV}}$ for JDV orthogonal to $\mathcal{V}_2$ is represented by $\mathbf{t}_1'$ (to be more precise, $\mathbf{t}_{\text{JDV}} = (0, 0, \mathbf{t}_1'^{\mathrm{T}})^{\mathrm{T}}$), while the orthogonal component for JD is represented by a combination of $\mathbf{t}_1'$ and $\mathbf{t}_2'$: $\mathbf{t}_{\text{JD}} = (0, \gamma, (\mathbf{t}_1' + \gamma\mathbf{t}_2')^{\mathrm{T}})^{\mathrm{T}}$. So in general, when $\mathbf{d}_2$ is not a multiple of $\mathbf{d}_1$ and when $\gamma \neq 0$, JD and JDV will not produce the same expansion of $\mathcal{V}_2$. Note that $(\mathbf{I}-\mathbf{e}_1\mathbf{e}_1^*)\mathbf{A}(\mathbf{I}-\mathbf{e}_1\mathbf{e}_1^*)$ is non-singular on $\mathbf{e}_1^\perp$ if and only if $\alpha \neq -\mathbf{c}_2^*\mathbf{t}_2'$. The “in between” differs from JD and JDV in that it has no extra constraint for γ. Taking $\gamma = -\mathbf{c}_2^*\mathbf{t}_1'/(\alpha + \mathbf{c}_2^*\mathbf{t}_2')$ gives JD, taking $\gamma = 0$ gives JDV.

Finally, as an illustration of §2.4, we calculate the new residual associated with $\mathbf{u}_3 = \mathbf{u}_2 + \mathbf{t}$ and $\theta_3 = \theta_2 + \varepsilon$. We take $\beta = 0$. The new residual for the “in between” equals

$$\mathbf{r}_3 = \begin{pmatrix} \mathbf{c}_1^*\mathbf{t}' - \varepsilon \\ \alpha\gamma + \mathbf{c}_2^*\mathbf{t}' - \varepsilon\gamma \\ -\varepsilon\mathbf{t}' \end{pmatrix}.$$

If $\gamma = -\mathbf{c}_2^*\mathbf{t}_1'/(\alpha + \mathbf{c}_2^*\mathbf{t}_2')$ (as for JD) then the choice $\varepsilon = \mathbf{c}_1^*\mathbf{t}'$ reduces the terms in $\mathbf{r}_3$ to second order ones, while no clever choice for ε can achieve this if γ is not close to $-\mathbf{c}_2^*\mathbf{t}_1'/(\alpha + \mathbf{c}_2^*\mathbf{t}_2')$.

3.2 Stagnation

The example in this section shows that JD may stagnate where JDV expands.
Consider the matrix $\mathbf{A}$ of §3.1, but now take $\beta = 1$, $\alpha = 0$ and $\mathbf{d}_2 = \mathbf{d}_1$.

As initial space, we take $\mathcal{V}_1 = \text{span}\{e_1\}$. Then $\mathbf{u}_1 = e_1$ and $\mathbf{r}_1 = (0, 0, \mathbf{d}_1^{\mathrm{T}})^{\mathrm{T}}$. Any of the three approaches find $-e_2$ as expansion vector: $\mathcal{V}_2 = \text{span}\{e_1, e_2\}$. Now $\mathbf{u}_2$ is again e_1 and JD stagnates: $\mathbf{t}_{\text{JD}} = -e_2$ belongs already to $\mathcal{V}_2$ and does not lead to an expansion of $\mathcal{V}_2$. The JDV correction vector $\mathbf{t}_{\text{JDV}}$ is equal to $(0, 0, (\mathbf{B}^{-1}\mathbf{d}_1)^{\mathrm{T}})^{\mathrm{T}}$ and expands $\mathcal{V}_2$.

4 Exact solution of the correction equations

If, in the example in §3.1, $\mathbf{d}_1$ and $\mathbf{d}_2$ are in the same direction, or equivalently, if the residuals of the Ritz vectors are in the same direction, then exact JD and exact JDV calculate effectively the same expansion vector. One may wonder whether this also may happen in more general situations. Before we discuss this question, we characterize the situation in which all residuals are in the same direction.

All residuals of Ritz vectors with respect to some subspace $\mathcal{V}_m$ are in the same direction if and only if the components orthogonal to $\mathcal{V}_m$ of the vectors $\mathbf{A}\mathbf{v}$ are in the same direction for all $\mathbf{v} \in \mathcal{V}_m$. It is easy to see and well known that $\mathcal{V}_m$ has this last property if it is a Krylov subspace generated by $\mathbf{A}$ (i.e., $\mathcal{V}_m = \mathcal{K}_m(\mathbf{A}, \mathbf{v}_0) = \text{span}(\{\mathbf{A}^i\mathbf{v}_0 \mid i < m\})$ for some positive integer m and some vector $\mathbf{v}_0$). The converse is also true as stated in the following lemma. We will tacitly assume that all Krylov subspaces that we will consider in the remainder of this paper, are generated by $\mathbf{A}$.

LEMMA 1 *For a subspace $\mathcal{V}_m$ the following properties are equivalent.*
(a) *$\mathcal{V}_m$ is a Krylov subspace,*
(b) $\mathbf{A}\mathcal{V}_m \subset \text{span}(\mathcal{V}_m, \mathbf{v})$ *for some* $\mathbf{v} \in \mathbf{A}\mathcal{V}_m$.

Proof. We prove that (b) implies (a). The implication "(a) $\Rightarrow$ (b)" is obvious.
If the columns of the n by m matrix $\mathbf{V}_m$ form a basis of $\mathcal{V}_m$ then (b) implies that $\mathbf{A}\mathbf{V}_m = [\mathbf{V}_m, \mathbf{v}]H$ for some $m+1$ by m matrix H. There is an orthogonal m by m matrix Q such that $\tilde{H} := Q'^{*}HQ$ is upper Hessenberg. Here Q' is the $m+1$ by $m+1$ orthogonal matrix with m by m left upper block Q and $(m+1, m+1)$ entry equal to 1. Q can be constructed as product of Householder reflections.[3] Hence $\mathbf{A}\tilde{\mathbf{V}}_m = [\tilde{\mathbf{V}}_m, \mathbf{v}]\tilde{H}$, where $\tilde{\mathbf{V}}_m \equiv \mathbf{V}_m Q$. Since $\tilde{H}$ upper Hessenberg, this implies that $\mathcal{V}_m$ is a Krylov subspace (of order m) generated by the first column of $\tilde{\mathbf{V}}_m$. □

We will see in Cor. 4 that exact JD and exact JDV coincide after restart with a set of Ritz vectors taken from a Krylov subspace. The proof uses the fact, formulated in Cor. 1, that any collection of Ritz vectors of $\mathbf{A}$ with respect to a single Krylov subspace span a Krylov subspace themselves. This fact can be found in [9, §3] and is equivalent to the statement in [20, Th.3.4] that Implicit Restarted Arnoldi and unpreconditioned Davidson (i.e., Davidson with the trivial preconditioner $\mathbf{I}_n$) generate the same search subspaces. However, the proof below is more elementary.

COROLLARY 1 *If $\mathcal{V}_m$ is a Krylov subspace and if $\{(\theta_m^{(i)}, \mathbf{u}_m^{(i)}) \mid i \in J\}$ is a subset of Ritz pairs of $\mathbf{A}$ with respect to $\mathcal{V}_m$ then the Ritz vectors $\mathbf{u}_m^{(i)}$ ($i \in J$) span a Krylov subspace.*

Proof. Assume that $\mathcal{V}_m$ is a Krylov subspace. Then (b) of Lemma 1 holds and, in view of the Gram-Schmidt process, we may assume that the vector $\mathbf{v}$ in (b) is orthogonal to $\mathcal{V}_m$.

[3] Here the refections are defined from their right action on the $m+1$ by m matrix and work on the rows from bottom to top, whereas in the standard reduction to Hessenberg form of a square matrix they are defined from their left action and work on the columns from left to right.

Since $\mathbf{A}\mathbf{u}_m^{(i)} - \theta_m^{(i)}\mathbf{u}_m^{(i)} \perp \mathcal{V}_m$, (b) of Lemma 1 implies that $\mathbf{A}\mathbf{u}_m^{(i)} - \theta_m^{(i)}\mathbf{u}_m^{(i)} \in \text{span}(\mathbf{v})$. Hence $\mathbf{A}\mathbf{u}_m^{(i)} \in \text{span}(\mathcal{U}, \mathbf{v})$, where $\mathcal{U}$ is the space spanned by the Ritz vectors $\mathbf{u}_m^{(i)}$ ($i \in J$), and the corollary follows from Lemma 1. □

4.1 Expanding a Krylov subspace

In this section, $\mathcal{V}_m$ is a subspace, $\mathbf{V}_m$ a matrix of which the columns form an orthonormal basis of $\mathcal{V}_m$, $(\theta_m, \mathbf{u}_m)$ a Ritz pair of $\mathbf{A}$ with respect to $\mathcal{V}_m$, and $\mathbf{r}_m$ is the associated residual. Further, we assume that $(\mathbf{I}_n - \mathbf{V}_m\mathbf{V}_m^*)(\mathbf{A} - \theta_m\mathbf{I}_n)(\mathbf{I}_n - \mathbf{V}_m\mathbf{V}_m^*)$ is non-singular on $\mathcal{V}_m^\perp$, that is (3) has a unique solution, and we assume that $\mathbf{r}_m \neq \mathbf{0}$, that is $\mathbf{u}_m$ is not converged yet.
The assumption $\mathbf{r}_m \neq \mathbf{0}$ implies that $\mathbf{t}_{\text{JDV}} \neq \mathbf{0}$ and $\mathbf{A}\mathbf{u}_m \notin \mathcal{V}_m$.

Note that (cf. [15], [13])

$$\mathbf{t}_{\text{JD}} = -\mathbf{u}_m + \varepsilon(\mathbf{A} - \theta_m\mathbf{I}_n)^{-1}\mathbf{u}_m \quad \text{for} \quad \varepsilon = \frac{\mathbf{u}_m^*\mathbf{u}_m}{\mathbf{u}_m^*(\mathbf{A} - \theta_m\mathbf{I}_n)^{-1}\mathbf{u}_m}. \tag{12}$$

THEOREM 1 *Consider the following properties.*
(a) $\mathcal{V}_m$ *is a Krylov subspace.*
(b) span$(\mathcal{V}_m, \mathbf{t}) \subset$ span$(\mathcal{V}_m, \mathbf{t}_{\text{JDV}})$ *for all solutions* $\mathbf{t}$ *of* (10).
(c) span$(\mathcal{V}_m, \mathbf{t}_{\text{JD}})$ *is a Krylov subspace.*
Then (a) ⇔ (b) ⇒ (c).

Proof. Consider a solution $\mathbf{t}$ of (10). We first show the intermediate result that

$$\text{span}(\mathcal{V}_m, \mathbf{t}) = \text{span}(\mathcal{V}_m, \mathbf{t}_{\text{JDV}}) \quad \Leftrightarrow \quad \gamma\mathbf{A}\mathbf{u}_m + \mathbf{A}\mathbf{V}_m(\mathbf{V}_m^*\mathbf{t}) \in \mathcal{V}_m \quad \text{for some } \gamma \neq 1. \tag{13}$$

If we decompose $\mathbf{t}$ in

$$\mathbf{t} = \tilde{\mathbf{t}} + \mathbf{V}_m s \quad \text{with} \quad \tilde{\mathbf{t}} \equiv (\mathbf{I}_n - \mathbf{V}_m\mathbf{V}_m^*)\mathbf{t} \quad \text{and} \quad s \equiv \mathbf{V}_m^*\mathbf{t} \tag{14}$$

then we see that (10) is equivalent to

$$(\mathbf{I}_n - \mathbf{V}_m\mathbf{V}_m^*)(\mathbf{A} - \theta\mathbf{I}_n)(\mathbf{I}_n - \mathbf{V}_m\mathbf{V}_m^*)\tilde{\mathbf{t}} = -\mathbf{r}_m - (\mathbf{I}_n - \mathbf{V}_m\mathbf{V}_m^*)(\mathbf{A} - \theta\mathbf{I}_n)\mathbf{V}_m s. \tag{15}$$

The vectors $\tilde{\mathbf{t}}$ and $\mathbf{t}$ lead to the same expansion of $\mathcal{V}_m$. A combination of (3) and (15) shows that $\mathbf{t}_{\text{JDV}}$ and $\mathbf{t}$ lead to the same expansion of $\mathcal{V}_m$ if and only if

$$(1 - \gamma')\mathbf{r}_m + (\mathbf{I}_n - \mathbf{V}_m\mathbf{V}_m^*)(\mathbf{A} - \theta\mathbf{I}_n)\mathbf{V}_m s = \mathbf{0} \quad \text{for some scalar } \gamma' \neq 0; \tag{16}$$

use the non-singularity restriction for the "if-part". Since $(\mathbf{I}_n - \mathbf{V}_m\mathbf{V}_m^*)\mathbf{V}_m = \mathbf{0}$, (16) is equivalent to $(1 - \gamma')\mathbf{A}\mathbf{u}_m + \mathbf{A}\mathbf{V}_m s \in \mathcal{V}_m$, which proves (13).

"(a) ⇒ (b)": Since $\mathbf{r}_m \neq \mathbf{0}$, we see that $\mathbf{A}\mathbf{u}_m \notin \mathcal{V}_m$. Therefore, if (a) holds then (see Lemma 1) we have that $\mathbf{A}\mathbf{V}_m(\mathbf{V}_m^*\mathbf{t}) \in \text{span}(\mathcal{V}_m, \mathbf{A}\mathbf{u}_m)$ and (13) shows that (b) holds.

"(b) ⇒ (c)": Note that the kernel $\mathcal{N}$ of the operator in (10) consists of the vectors $\mathbf{s} \equiv \mathbf{t} - \mathbf{t}_{\text{JDV}}$ with $\mathbf{t}$ a solution of (10). Since (3) has a unique solution, we see that none of the non-trivial vectors in $\mathcal{N}$ is orthogonal to $\mathcal{V}_m$. Therefore, the space $\mathcal{N}$ and the space of all vectors $\mathbf{V}_m^*\mathbf{s}$ ($\mathbf{s} \in \mathcal{N}$) have the same dimension which is one less than the dimension of $\mathcal{V}_m$. From (13) we see that (b) implies that $\mathbf{A}\mathbf{V}_m(\mathbf{V}_m^*\mathbf{s}) \in \text{span}(\mathcal{V}_m, \mathbf{A}\mathbf{u}_m)$ for all $\mathbf{s} \in \mathcal{N}$. Since $\mathbf{s} = \mathbf{t} - \mathbf{t}_{\text{JDV}} \perp \mathbf{u}_m$, we see that $\mathbf{u}_m$ is independent of $\mathbf{A}\mathbf{V}_m(\mathbf{V}_m^*\mathbf{s})$ for all $\mathbf{s} \in \mathcal{N}$. Therefore, in view of the dimensions of the spaces involved we may conclude that $\mathbf{A}\mathcal{V}_m \in \text{span}(\mathcal{V}_m, \mathbf{A}\mathbf{u}_m)$, which, by Lemma 1, proves (a).

"(a) $\Rightarrow$ (c)": If $\mathcal{V}_m$ is a Krylov subspace of order m generated by $\mathbf{v}_0$, that is if (a) holds, then, also in view of (12), we have that

$$\text{span}(\mathcal{V}_m, \mathbf{t}_{\text{JD}}) = \text{span}(\mathcal{V}_m, (\mathbf{A} - \theta \mathbf{I}_n)^{-1}\mathbf{u}_m) \subset \{q(\mathbf{A})[(\mathbf{A} - \theta \mathbf{I}_n)^{-1}\mathbf{v}_0] \mid q \text{ pol. degree } \leq k\}.$$

The inclusion follows easily from the representation of $\mathcal{V}_m$ as $\mathcal{V}_m = \{p(A)\mathbf{v}_0 | p \text{ pol. degree } < k\}$. If $(\mathbf{A} - \theta \mathbf{I}_n)^{-1}\mathbf{u}_m \notin \mathcal{V}_m$ then a dimension argument shows that the subspaces coincide which proves that $\text{span}(\mathcal{V}_m, \mathbf{t}_{\text{JD}})$ is a Krylov subspace. If $(\mathbf{A} - \theta \mathbf{I}_n)^{-1}\mathbf{u}_m \in \mathcal{V}_m$ then there is no expansion and the Krylov structure is trivially preserved. □

Lemma 1 implies that any $n-1$ dimensional subspace is a Krylov subspace. In particular, $\text{span}(\mathcal{V}_m, \mathbf{t}_{\text{JD}})$ is a Krylov subspace if $\mathcal{V}_m$ is $n-2$-dimensional and it does not contain $\mathbf{t}_{\text{JD}}$. From this argument it can be seen that (c) does not imply (a).

Since $\mathbf{t}_{\text{JD}}$ is also a solution of (10), we have the following.

COROLLARY 2 *If $\mathcal{V}_m$ is a Krylov subspace then* $\text{span}(\mathcal{V}_m, \mathbf{t}_{\text{JD}}) \subset \text{span}(\mathcal{V}_m, \mathbf{t}_{\text{JDV}})$. □

If θ_m is simple then $\mathbf{t}_{\text{JD}} \notin \mathcal{V}_m$ and the expanded subspaces in Cor. 2 coincide. However, as the example in §3.2 shows, JD may not always expand the subspace. Note that, in accordance with (c) of Th. 1, the subspace $\mathcal{V}_2$ in this example is a Krylov subspace (generated by $\mathbf{A}$ and $\mathbf{v}_0 = e_2 - e_1$).

Cor. 2 does not answer the question whether $\mathbf{t}_{\text{JD}}$ and $\mathbf{t}_{\text{JDV}}$ lead to the same expansion of $\mathcal{V}_m$ only if $\mathcal{V}_m$ is a Krylov subspace. The example in §3 shows that the answer can be negative, namely if $\mathbf{t}_{\text{JD}} \perp \mathcal{V}_m$: then $\gamma = \mathbf{V}_m^* \mathbf{t}_{\text{JD}} = 0$. The answer can also be negative in cases where $\mathbf{t}_{\text{JD}} \not\perp \mathcal{V}_m$, provided that the dimension of the subspace $\mathcal{V}_m$ is larger than 2. The following theorem characterizes partially the situation where we obtain the same expansion. Note that $\mathcal{V}_m$ is a Krylov subspace if and only if the dimension of $\mathbf{A}\mathcal{V}_m \cap \mathcal{V}_m$ is at most one less than the dimension of $\mathcal{V}_m$ (see Lemma 1).

THEOREM 2 *If* $\text{span}(\mathcal{V}_m, \mathbf{t}_{\text{JD}}) \subset \text{span}(\mathcal{V}_m, \mathbf{t}_{\text{JDV}})$ *then* $\mathbf{A}\mathcal{V}_m \cap \mathcal{V}_m \neq \{\mathbf{0}\}$ *or* $\mathbf{t}_{\text{JD}} \perp \mathcal{V}_m$.

Proof. If $\mathbf{t}_{\text{JD}}$ and $\mathbf{t}_{\text{JDV}}$ give the same expansion then (13) shows that $\gamma \mathbf{A}\mathbf{u}_m + \mathbf{A}\mathbf{V}_m(\mathbf{V}_m^* \mathbf{t}_{\text{JD}}) \in \mathcal{V}_m$. Apparently, $\mathbf{A}\mathcal{V}_m \cap \mathcal{V}_m \neq \{\mathbf{0}\}$ or $\gamma = 0$ and $\mathbf{V}_m^* \mathbf{t}_{\text{JD}} = 0$. A similar argument applies to the case where $\mathbf{t}_{\text{JD}} \in \mathcal{V}_m$. □

In practical situations, where $\mathcal{V}_m$ is constructed from inexact solutions of the correction equations it will be unlikely that $\mathbf{A}\mathcal{V}_m$ will have a non-trivial intersection with $\mathcal{V}_m$ (unless the dimension of $\mathcal{V}_m$ is larger than $n/2$). Usually $\mathbf{t}_{\text{JD}} \not\perp \mathcal{V}_m$. Therefore, the exact expansion vectors $\mathbf{t}_{\text{JD}}$ and $\mathbf{t}_{\text{JDV}}$ will not lead to same expansions, and we may not expect that inexact expansion vectors will produce the same expansions.

The correction equation (11) in inexact TRQ is based on a Krylov subspace: the matrix $\mathbf{V}_m$ in this algorithm is produced by the Arnoldi procedure whenever equation (11) has to be solved.

4.2 Starting with one vector

As any one dimensional subspace is a Krylov subspace, one consequence of Theorem 1 is the following corollary. The proof follows by an inductive combination of Th. 1(c) and Cor. 2.

COROLLARY 3 *Exact JD and exact JDV started with the same vector $\mathbf{u}_1$ are mathematically equivalent as long as exact JD expands, i.e., they produce the same sequence of search subspaces in exact arithmetic.*

4.3 (Re-)Starting with several Ritz vectors

Once we start JD and JDV with one vector the dimension of the search subspace starts increasing. After a number of steps a restart strategy must be followed to keep the required storage limited and the amount of work related to the search subspace low. The question is which information should be thrown away and which should be kept in memory. A popular strategy is to select those Ritz pairs that are close to a specified shift/target. Cor. 1 and an inductive application of Theorem 1 imply that, with a one-dimensional initial start and restarts with the selected Ritz vectors, restarted exact JD and restarted exact JDV are mathematically equivalent.

COROLLARY 4 *Exact JD and exact JDV are mathematically equivalent as long as exact JD expands if they are both started with the same set of Ritz vectors of* $\mathbf{A}$ *with respect to one Krylov subspace.*

In practice, we have to deal with round off errors and the correction equations can only be solved with a modest accuracy. Therefore, even if we start with one vector or a Krylov subspace, the subsequent search subspaces will not be Krylov and the results in the above corollaries do not apply. If a search subspace is not Krylov, then from Th. 1 we learn that the “in between” variant may lead to expansions different from those of JDV. Th. 2 indicates that also JD will differ from JDV.

5 Numerical experiments

Here a few numerical experiments will be presented. We will see that JDV and JD show comparable speed of convergence also in finite precision arithmetic as long as the correction equations are solved in high precision (§5.1.1). JDV converges much slower than JD if the Krylov structure of the search subspace is seriously perturbed. We will test this by starting with a low dimensional random space (§5.1.1). We will also see this effect in our experiments where we solved the correction equations only in modest accuracy (§5.1.2). Moreover, we will be interested in the question whether the slower convergence of JDV in case of inaccurate solutions of the correction equations can be compensated by a better performance of the linear solver for the correction equation (§5.2.1). Further, some stability issues will be addressed (§5.1.3).

5.1 Example 1

In the experiments in this section 5.1, we apply the Jacobi-Davidson method on a tridiagonal matrix of order 100 with diagonal entries 2.4 and off-diagonal entries 1 ([15, Ex. 1]). Our aim is the largest eigenvalue $\lambda = 4.3990\ldots$. We start with a vector with all entries equal to 0.1.

5.1.1 Exact solution of the correction equation

When solving the correction equations exactly no difference between JD and JDV is observed (dash-dotted line in left plot in Fig. 1) which is in accordance with Cor. 3. The plots show the $\log_{10}$ of the error $|\theta_m - \lambda|$ in the Ritz value θ_m versus the iteration number m.

To see the effect of starting with an arbitrary subspace of dimension larger than 1 we added four random vectors to the start vector with all entries equal to 0.1. The right plot in Fig. 1 shows the convergence of exact JD (solid curve) and JDV (dashed curve). Here the results of `seed(253)` in our MATLAB-code are presented (other seeds showed similar convergence behaviour). The correction equations have been solved “exactly”, that is to machine precision. As anticipated in §4.1 (see Th. 2) the convergence behaviour of JDV now clearly differs from that of JD. Moreover, the speed of convergence of JDV seems to be much lower than of JD (linear rather than cubic? See §2.4). Apparently, expanding with $\mathbf{t}_{\text{JDV}}$ rather than with $\mathbf{t}_{\text{JD}}$ may slow down the convergence of Jacobi-Davidson considerably in case the initial subspace is not a Krylov subspace.

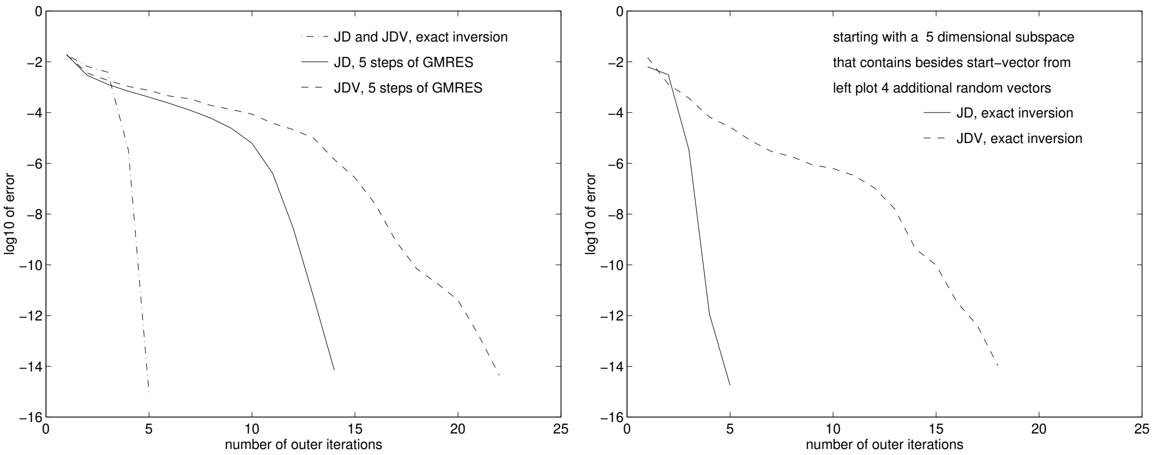

FIGURE 1. *Convergence plots for Example 1. Differences between* JD *and* JDV *when not solving the correction equation exactly* (*left plot*) *and when starting with an unstructured 5-dimensional subspace* (*right plot*)*. The plots show the* $\log_{10}$ *of the error* $|\theta_m - \lambda|$ *in the Ritz value* θ_m *versus the iteration number* m.

Note that JD performs slightly better with the five-dimensional start than with the one-dimensional start (compare the solid curve in the right plot with de dashed-dotted curve in the left plot). This may be caused by the extra (noisy) search directions.

5.1.2 Approximate solution of the correction equation

If the correction equations are not solved in high precision, we may not expect the constructed search subspaces $\mathcal{V}_m$ to be Krylov subspaces, even if the process is started with a Krylov subspace. Consequently $\mathbf{t}_{\text{JD}}$ and $\mathbf{t}_{\text{JDV}}$, and therefore their inexact approximations, will not lead to the same expansions of $\mathcal{V}_m$. In view of the experimental result in §5.1.1, we expect the inexact JDV to converge slower than inexact JD.

Again we start with one vector, but we use only 5 steps of GMRES to get an approximate solution of the correction equation in each outer iteration. The solid line (JD) and the dashed line (JDV) in the left plot of Fig. 1 show the results. JDV needs significantly more outer iterations for convergence than JD.

5.1.3 Loss of orthogonality

The (approximate) solution of (2) in JD will in general not be orthogonal to $\mathbf{V}_m$. Therefore, this solution is orthonormalized against $\mathbf{V}_m$ before it is used to expand $\mathbf{V}_m$ to $\mathbf{V}_{m+1}$. We refer to this step in the algorithm as *post*-orthogonalization (of the solution of the correction equation). In JDV, however, if the correction equation (3) is solved with, for instance, GMRES, then the (approximate) solution should be orthogonal to $\mathbf{V}_m$ and post-orthogonalization, i.e., the explicit orthogonalization before expanding $\mathbf{V}_m$, should be superfluous. This observation would offer a possibility of saving inner products. Here we investigate what the effect is of omitting the post-orthogonalization in JDV.

Again JDV is applied on the simple test matrix with the same starting vector as before and the correction equations are solved approximately with 5 steps of GMRES. As initial approximate solution for GMRES we take the zero vector.

From the experiment we learn that without post-orthogonalization the basis of the search subspace in JDV loses its orthogonality. As a measure for the orthonormality of $\mathbf{V}_m$ we took (see [12, §13.8]) $\kappa_m \equiv \|I_m - \mathbf{V}_m^*\mathbf{V}_m\|$. Table 1 lists the values of the error $|\lambda - \theta_m|$ and the quantity κ_m for the first 10 outer iterations. Column two and three ("with post-ortho.") show the results for the implementation of JDV where the approximate solution of the correction equation is explicitly orthogonalized against $\mathbf{V}_m$ before it is used to expand this matrix. In the columns four and five ("without post-ortho.") we see

m	with post-ortho. $\lvert\lambda-\theta_m\rvert$	κ_m	without post-ortho. $\lvert\lambda-\theta_m\rvert$	κ_m	with pre-ortho. $\lvert\lambda-\theta_m\rvert$	κ_m
1	1.903e−02	2.220e−16	1.903e−02	2.220e−16	1.903e−02	2.220e−16
2	3.611e−03	2.289e−15	3.611e−03	3.690e−14	3.611e−03	3.690e−14
3	1.856e−03	2.314e−15	1.856e−03	1.426e−11	1.856e−03	4.567e−14
4	1.076e−03	2.314e−15	1.076e−03	2.649e−09	1.076e−03	4.866e−14
5	7.480e−04	2.316e−15	7.480e−04	6.621e−07	7.480e−04	5.920e−14
6	4.464e−04	2.316e−15	4.423e−04	1.125e−04	4.464e−04	6.534e−14
7	3.454e−04	2.317e−15	4.135e−04	2.710e−02	3.454e−04	7.490e−14
8	1.909e−04	2.317e−15	3.135e+00	9.732e−01	1.909e−04	9.546e−14
9	1.317e−04	2.317e−15	7.004e+00	1.940e+00	1.317e−04	9.548e−14
10	8.747e−05	2.317e−15	1.094e+01	2.920e+00	8.747e−05	1.232e−13

TABLE 1. *The need of post-orthogonalization when using* JDV. *For the simple test, the* JDV *correction equation* (3) *is solved approximately with 5 steps of* GMRES. *The table shows the error* $|\lambda-\theta_m|$ *in the Ritz value* θ_m *and the "orthonormality" of the basis* $\mathbf{V}_m$ *of the search subspaces* ($\kappa_m = \|I_m - \mathbf{V}_m^*\mathbf{V}_m\|$) *for the implementation with post-orthogonalization of the solution of the correction equation* (*column two and three*)*, without post-orthogonalization* (*column four and five*)*, and without post-orthogonalization, but with pre-orthogonalization of the left-hand side vector of the correction equation* (*column six and seven*)*.*

that if the post-orthogonalization is omitted then the loss of orthonormality starts influencing the error significantly after just 5 outer iterations. After 8 iterations the orthonormality is completely lost. This phenomenon can be explained as follows.

The residual of the selected Ritz pair is computed as $\mathbf{r}_m = \mathbf{A}\mathbf{u}_m - \theta_m\mathbf{u}_m$. Therefore, in finite precision arithmetic, the residual will not be as orthogonal to the search subspace as intended even if $\mathbf{V}_m$ would have been orthonormal. For instance, at the second iteration of our experiment, we have an error $\|\mathbf{V}_2^*\mathbf{r}_2\|$ equal to $1.639\mathrm{e}{-}13$. With the norm of the residual equal to 0.02145 this results in a relative error of $7.640\mathrm{e}{-}12$. Note that, specifically at convergence, rounding errors in $\mathbf{r}_m$ may be expected to lead to relatively big errors. In each solve of the correction equation (3), GMRES is started with initial approximate $\mathbf{0}$ and the vector $\mathbf{r}_m$ is taken as the initial residual in the GMRES process.

Since $\mathbf{r}_m$ is supposed to be orthogonal against $\mathbf{V}_m$, this vector is not explicitly orthogonalized against $\mathbf{V}_m$, and the normalized $\mathbf{r}_m$ is simply taken as the first Arnoldi vector. In the subsequent GMRES steps the Arnoldi vectors are obtained by orthogonalization against $\mathbf{V}_m$ followed by orthogonalization against the preceding Arnoldi vectors. However, since the first Arnoldi vector will not be orthogonal against $\mathbf{V}_m$, the approximate GMRES solution will not be orthogonal against $\mathbf{V}_m$. Adding this "skew" vector to the basis of the search subspace will add to the non-orthogonality in the basis.

Columns six and seven ("with pre-ortho.") of Table 1 show that post-orthogonalization can be omitted as long as the residual $\mathbf{r}_m$ is sufficiently orthogonal with respect to $\mathbf{V}_m$: the post-orthogonalization is omitted here, but the right-hand side vector of the correction equation, the residual $\mathbf{r}_m$, is orthogonalized explicitly against $\mathbf{V}_m$ before solving the correction equation (pre-orthogonalization). Since pre- and post-orthogonalization are equally expensive and since pre-orthogonalization appears to be slightly less stable (compare the κ_m's in column 3 with those in column 7 of Table 1), pre-orthogonalization is not an attractive alternative, but the experimental results confirm the correctness of the above arguments.

Note that our test matrix here is only of order 100 and the effect of losing orthogonality may become even more important for matrices of higher order.

Also in JD the finite precision residual $\mathbf{r}_m$ of the Ritz pair will not be orthogonal to the search subspace. Since even in exact arithmetic you may not expect the solution of the JD correction equation (2) to be orthogonal to $\mathbf{V}_m$, post-orthogonalization is essential in the JD variant. In our experiment, using finite precision arithmetic, we did not observe any significant loss of orthogonality in the column vectors of $\mathbf{V}_m$. Nevertheless, we also checked whether pre-orthogonalization of $\mathbf{r}_m$ before solving the correction equation would enhance the convergence of JD. This was not the case: JD converged equally fast with and without pre-orthogonalization.

In the remaining experiments we used post-orthogonalization in JDV, too.

5.2 Example 2

In this section we consider a slightly more realistic eigenvalue problem. We are interested in the question whether the projections on the orthogonal complement of $\mathbf{V}_m$ in the JDV approach may significantly improve the performance of the linear solver for the correction equation.

For $\mathbf{A}$ we take the SHERMAN1 matrix from the Harwell-Boeing collection [5]. The matrix is real unsymmetric of order 1000. All eigenvalues appear to be real and in the interval [-5.0449,-0.0003]. About 300 eigenvalues are equal to -1. We want to find a few eigenvalues with associated eigenvectors that are closest to the target σ. Our target σ is set to -2.5. Note that the "target" eigenvalues are in the "interior" of the spectrum, which make them hard to find, no matter the numerical method employed.

In general, when started with a single vector, the Ritz values in the initial stage of the process will be relatively inaccurate approximations of the target eigenvalue λ, that is, if λ is the eigenvalue closest to σ then for the first few m we will have that $|\theta_m - \lambda|/|\sigma - \lambda| \gg 1$. Therefore, as argued in [14, §9.4] (see also [7, §4.0.1]), it is more effective to replace initially θ_m in the correction equation by σ (similar observations can be found in [10, §6] and [19, §3.1]). As the search subspace will not contain significant components of the target eigenvectors in this initial stage, the projections in (2) and (3) are not expected to be effective. Therefore, we expanded the search subspace in the first few steps of our process by approximate solutions of the equation

$$(\mathbf{A} - \sigma \mathbf{I}_n)\mathbf{t} = -\mathbf{r}_m, \tag{17}$$

which can be viewed as a generalized Davidson approach.

In the computations we did not use any preconditioning. We started JD and JDV with the same vector, the vector of norm one of which all entries are equal. The algorithms were coded in `C` and run on a Sun SPARCstation 4 using double precision.

5.2.1 Solving the correction equation in lower precision

Fig. 2 shows the $\log_{10}$ of the residual norm for JD (the solid curve) and for JDV (the dashed curve). In this example, all correction equations (including (17)) have been solved with 50 steps of GMRES except where GMRES reached a residual accuracy of 10^{-14} in an earlier stage. In the first 5 steps of the outer iteration we took the approximate solution of the Davidson correction equation (17) as the expansion vector. As the correction equations are not solved exactly, we expect that JD will need less outer iterations than JDV (see §§4.1 and 5.1.2), which is confirmed by the numerical results in the figure.

As argued in §1, the projections on the orthogonal complement of $\mathbf{V}_m$ in the JDV correction equation (3) may improve the conditioning (or more general, the spectral properties) of the operator in the correction equation. This may allow a more efficient or a more accurate way of solving the correction equation. Here we test numerically whether a better performance of the linear solver for the correction equations can compensate for a loss of speed of convergence in the outer iteration. In the figures in Fig. 3 we show how the performance of JD and JDV and the computational costs relate. As a measure for the costs we take the number of matrix-vector multiplications: we plot the $\log_{10}$ of the residual norm versus the number of matrix-vector multiplications by $\mathbf{A}$ (or by $\mathbf{A} - \theta_m \mathbf{I}_n$). Note that this way of measuring the costs favours JDV, since the projections in JDV are more costly than in JD. Nevertheless, we will see that JD outperforms JDV.

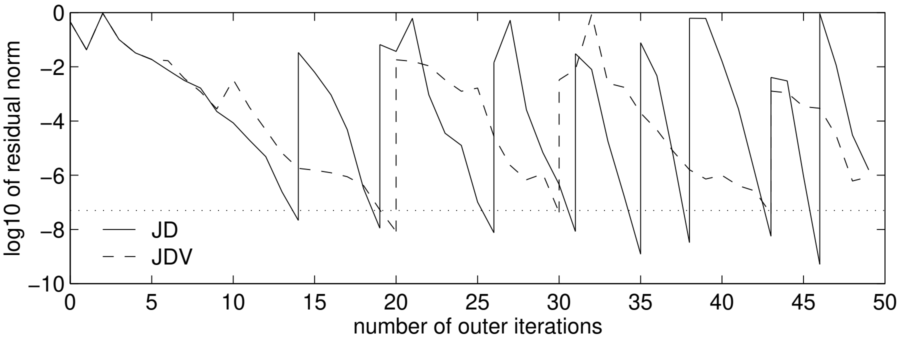

FIGURE 2. *The convergence history for the computation of eigenpairs with eigenvalue closest to* -2.5 *of the matrix* SHERMAN1. *The plot shows the* $\log_{10}$ *of the subsequent residual norms for* JD (*solid curve*) *and* JDV (*dashed curve*) *versus the iteration number* m. *A search for a next eigenpair is started when a Ritz pair is accepted as eigenpair* (*i.e., if* $\|\mathbf{r}_m\|_2 \leq 5\,10^{-8}$). *The correction equations are approximately solved with 50 steps of* GMRES.

method for the correction equation		number of outer iterations JD	JDV	number of matrix-vector multiplications JD	JDV	wallclock time in seconds JD	JDV
GMRES_{200}	(*a*)	4	4	798	790	64.1	64.3
	(*b*)	7	7	1401	1393	114.7	119.5
GMRES_{50}	(*a*)	14	20	715	1021	21.5	51.2
	(*b*)	19	30	970	1531	35.0	121.1
GMRES_{25}	(*a*)	26	37	677	963	41.3	143.0
	(*b*)	33	47	859	1223	83.2	301.4

TABLE 2. *Costs for the computation of two eigenpairs of* SHERMAN1 *with* JD *and* JDV. *The costs* (*b*) *for the computation of the second eigenpair* ($\lambda = -2.51545\ldots$) *include the costs* (*a*) *for the computation of the first eigenpair* ($\lambda = -2.49457\ldots$).

We solve all correction equations with GMRES_ℓ, that is with ℓ steps of GMRES, except where GMRES reaches a residual accuracy of 10^{-14} in an earlier stage. For ℓ we took 200 (top figure), 50 (middle figure), and 25 (bottom figure). In the first few outer iterations the Davidson correction equation (17) is solved approximately (2 outer iterations for $\ell = 200$ and 5 for $\ell = 50$ and for $\ell = 25$). When a Ritz pair is accepted as eigenpair (i.e., if $\|\mathbf{r}_m\| \leq 5\,10^{-8}$), a search is started for the next eigenpair. The accepted Ritz pairs are kept in the search subspace. Explicit deflation is used only in the correction equation (see [8]). Note that the correction equations (3) in JDV need no modification to accommodate the deflation, because accepted Ritz vectors are kept in the search space.
If GMRES would converge faster on JDV correction equations than on JD correction equations, then GMRES would need less steps for solving (3) in case the residual accuracy of 10^{-14} would be reached in less than ℓ GMRES steps, while in the other case it would produce more effective expansion vectors in JDV. With more effective expansion vectors the number of outer iterations may be expected to decrease. In both cases, there would be a positive effect on the number of matrix-vector multiplications needed in JDV.

In Table 2 the number of outer iterations, the number of matrix-vector multiplications and the amount of time needed for the computation for the first two eigenpairs ($\lambda = -2.49457\ldots$ and $\lambda = -2.51545\ldots$) are presented.

When solving the correction equation with 200 steps of GMRES no difference between JD and JDV is observed (upper plot in Fig. 3). Apparently with 200 steps of GMRES the correction equations are solved

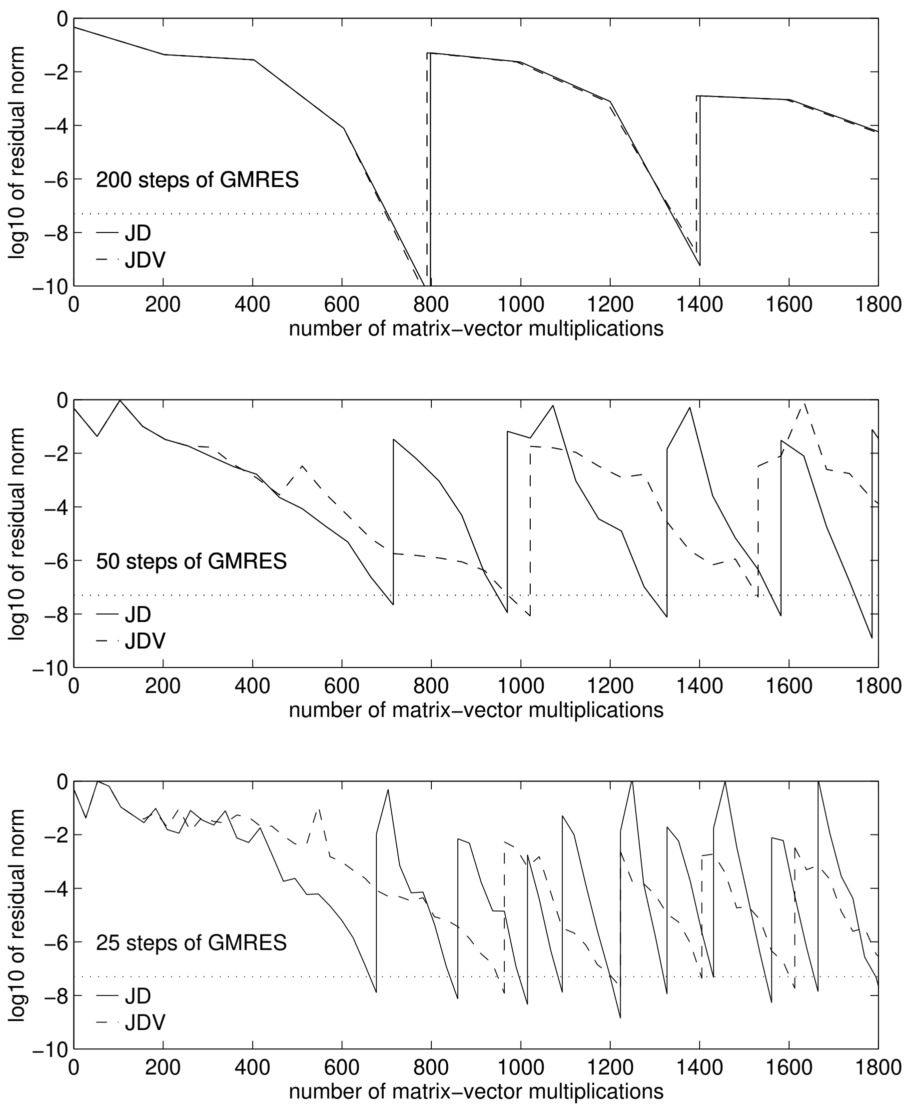

FIGURE 3. *The effect of reducing the precision of the solution method for the correction equation. The figures display the convergence history for the computation of eigenpairs with eigenvalue closest to* -2.5 *of the matrix* SHERMAN1. *Plotted are the* $\log_{10}$ *of the subsequent residual norms for* JD (*solid curve*) *and* JDV (*dashed curve*) *versus the number of matrix-vector multiplications. The correction equations are approximately solved with 200* (*top figure*)*, 50* (*center figure*) *and 25* (*bottom figure*) *steps of* GMRES.

in high precision and the results are in line with the theory and our previous experience. This can also be seen from Table 2. For the first eigenvalue JD uses 8 more matrix-vector multiplications than the 790 from JDV. On the other hand JDV takes a bit more time (about 0.2 seconds) than JD. From this we may conclude that, compared with the costs of the matrix-vector multiplications and the QR-algorithm for the computation of the eigenvalues of the projected matrix, the extra vector-vector operations involved in the correction equation of JDV are not very expensive.
Although JD and JDV need the same amount of time for convergence when using 200 steps of GMRES, the same eigenpairs can be computed in much less time. If 50 steps of GMRES are used, JD takes only 21.45 seconds for computing the first eigenpair whereas JDV takes 2.5 times that amount.
The differences between the two methods become more significant if we lower the precision of the solver for the correction equation by using only 25 steps of GMRES. With the same amount of matrix-vector multiplications the number of eigenpairs found by JD is much higher than JDV. Note, that the measured time for both JD and JDV in the case of GMRES_{25} is more than in the case of GMRES_{50} whereas the number of matrix-vector multiplications is less. The reason for this can only be the fact that in the case of GMRES_{25} more outer iterations are needed, every outer iteration the eigenvalues of the projected matrix are computed with a QR-algorithm.

6 Conclusions

In GMRESR, an iterative method for solving linear systems of equations, it pays to restrict the correction equations to the orthogonal complement of the space spanned by the search vectors. This approach, called GCRO, leads to new search directions that are automatically orthogonal with respect to the old ones. Although the restricted correction equations require more complicated projections with higher computational costs per matrix-vector multiplication, the number of matrix-vector multiplications may decrease tremendously leading to a better overall performance [3, 1]. In this paper, we investigated the question whether such an approach would be equally effective for the Jacobi-Davidson method for solving the eigenvalue problem. Note that eigenvalue problems are weakly non-linear.

When starting with a Krylov subspace and solving the correction equations exactly the standard approach (JD) of Jacobi-Davidson and its variant JDV with the more restricted correction equations, are mathematically equivalent (§4). However, in practical situations, where the correction equations are solved only in modest accuracy with finite precision arithmetic, the JDV variant appears to converge much more slowly than JD. Although the restricted correction equations in JDV may have spectral properties that are more favourable for linear solvers, a better performance of the linear solvers for the correction equation in JDV may not compensate for the slower convergence.

Acknowledgements

The first author's contribution was sponsored by the Netherlands Mathematics Research Foundation (SWON) under Grant No. 611-302-100.

References

[1] J. G. Blom and J. G. Verwer. VLUGR3: A vectorizable adaptive grid solver for PDEs in 3D. I. algorithmic aspects and applications. *Appl. Numer. Math.*, 16:129–156, 1994.

[2] E. de Sturler. Truncation strategies for optimal Krylov subspace methods. *SIAM J. Numer. Anal.*, To appear. Also Technical Report TR-96-38, Swiss Center for Scientific Computing (1996).

[3] E. de Sturler and D. R. Fokkema. Nested Krylov methods and preserving the orthogonality. In N. Duane Melson, T. A. Manteuffel, and S. F. McCormick, editors, *Sixth Copper Mountain Conference on Multigrid Methods*, pages 111–125. NASA Conference Publication 3224, Part 1, 1993.

[4] J. Descloux, J.-L. Fattebert, and F. Gygi. Rayleigh quotient iteration, an old recipe for solving modern large-scale eigenvalue problems. *Computers in Physics*, 12:22–27, 1998.

[5] I. S. Duff, R. G. Grimes, and J. G. Lewis. Users' guide for the Harwell-Boeing sparse matrix collection. Technical Report TR/PA/92/86, CERFACS, Toulouse, France, 1992. Also RAL Technical Report RAL 92-086.

[6] J.-L. Fattebert. *Une méthode numérique pour la résolution des problèmes aux valeurs propres liés au calcul de structure électronique moléculaire*. PhD thesis, Ecole Polytechnique Fédérale de Lausanne, 1997.

[7] D. R. Fokkema, G. L. G. Sleijpen, and H. A. van der Vorst. Accelerated inexact Newton schemes for large systems of nonlinear equations. *SIAM J. Sci. Comput.*, 19:657–674, 1998.

[8] D. R. Fokkema, G. L. 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.*, 20:94–125, 1998.

[9] R. B. Morgan. On restarting the Arnoldi method for large nonsymmetric eigenvalue problems. *Math. Comp.*, 65:1213–1230, 1996.

[10] R. B. Morgan and D. S. Scott. Generalizations of Davidson's method for computing eigenvalues of sparse symmetric matrices. *SIAM J. Sci. Stat. Comput.*, 7:817–825, 1986.

[11] J. Olsen, P. Jørgensen, and J. Simons. Passing the one-billion limit in full configuration-interaction (FCI) calculations. *Chem. Phys. Letters*, 169:463–472, 1990.

[12] B. N. Parlett. *The symmetric eigenvalue problem*. Prentice-Hall, Englewood Cliffs, N.J., 1980.

[13] A. Ruhe. Rational Krylov, a practical algorithm for large sparse nonsymmetric matrix pencils. *SIAM J. Sci. Comput.*, 19:1535–1551, 1998.

[14] G. L. G. Sleijpen, A. G. L. Booten, D. R. Fokkema, and H. A. van der Vorst. Jacobi-Davidson type methods for generalized eigenproblems and polynomial eigenproblems. *BIT*, 36:595–633, 1996.

[15] G. L. G. Sleijpen and H. A. van der Vorst. A Jacobi-Davidson iteration method for linear eigenvalue problems. *SIAM J. Matrix Anal. Appl.*, 17:401–425, 1996.

[16] G. L. G. Sleijpen and F. W. Wubs. In preparation.

[17] D. C. Sorensen. Truncated QZ methods for large scale generalized eigenvalue problems. *ETNA*, 7:141–162, 1998.

[18] D. C. Sorensen and C. Yang. A truncated RQ-iteration for large scale eigenvalue calculations. *SIAM J. Matrix Anal. Appl.*, 19:1045–1073, 1998.

[19] A. Stathopoulos, Y. Saad, and C. F. Fischer. Robust preconditioning of large sparse symmetric eigenvalue problems. *J. Comput. Appl. Math.*, 64:197–215, 1995.

[20] A. Stathopoulos, Y. Saad, and K. Wu. Dynamic thick restarting of the Davidson, and the implicitly restarted Arnoldi methods. *SIAM J. Sci. Comput.*, 19:227–245, 1998.

[21] H. A. van der Vorst and C. Vuik. GMRESR: A family of nested GMRES methods. *Num. Lin. Alg. Appl.*, 1:369–386, 1994.