JOURNAL OF COMPUTATIONAL PHYSICS 101, 287–291 (1992)

Generalizations of Davidson's Method for Computing Eigenvalues of Large Nonsymmetric Matrices

RONALD B. MORGAN

Department of Mathematics, Baylor University, Waco, Texas 76798

Received March 25, 1988; revised August 29, 1991

Davidson's method for nonsymmetric eigenvalue problems is examined. Some analysis is given for why Davidson's method is effective. An implementation is given that avoids use of complex arithmetic. This reduces the expense if complex eigenvalues are computed. Also discussed is a generalization of Davidson's method that applies the preconditioning techniques developed for systems of linear equations to nonsymmetric eigenvalue problems. Convergence can be rapid if there is an approximation to the matrix that is both factorable and fairly accurate.

I. INTRODUCTION

Finding eigenvalues of a large nonsymmetric matrix is often a difficult task. We examine the use of Davidson's method [1] and other preconditioning methods for this problem. This section reviews methods for the symmetric eigenvalue problem. Section 2 briefly discusses nonsymmetric Davidson's method and shows that the method is effective under certain conditions. Section 3 looks at an implementation of nonsymmetric Davidson's method that avoids complex vectors. Section 4 discusses generalizing Davidson's method for more powerful preconditioning.

For a large symmetric eigenvalue problem

$$Az = \lambda z,$$

the Lanczos algorithm [2] is a well known method. Given a starting vector x, this method generates a Krylov subspace, $\text{Span}\{x, Ax, A^2x, ..., A^{j-1}x\}$. Then the Rayleigh–Ritz procedure [2] is used to extract approximate eigenpairs from the subspace. The Rayleigh–Ritz procedure requires an orthonormal basis for the subspace, but the Lanczos algorithm uses a three-term recurrence that saves on orthogonalization costs. The convergence of this method is rapid if the desired eigenvalues are well separated from the rest of the spectrum, especially if they are on the exterior of the spectrum. Another popular method is subspace iteration [2], but it is generally not as powerful as the Lanczos algorithm [2, 3]. However, both of these methods have difficulty with problems that have poorly separated eigenvalues. Use of a shifted-and-inverted operator $(A-\sigma I)^{-1}$ changes the distribution of eigenvalues and improves convergence for eigenvalues near σ [4, 5]. The factorization required for implementing the inverted operator is often expensive, and we do not consider this option.

Another way to improve the distribution of eigenvalues is with Davidson's method. Davidson's method also uses the Rayleigh–Ritz procedure, but the subspace is generated by the operator $(D-\theta I)^{-1}(A-\theta I)$, where D is the diagonal of A and θ is the most recent approximate eigenvalue. To explain the effectiveness of Davidson's method, suppose that θ converges to the eigenvalue λ with associated eigenvector z. Then the operator in Davidson's method converges to $N \equiv (D-\lambda I)^{-1}(A-\lambda I)$. So we can view Davidson's method as being asymptotically related to the Lanczos method, but with N generating the subspace instead of A. Davidson's method is effective because N often has a better distribution of eigenvalues than A. Note N has one eigenvalue at 0 and the associated eigenvector is z. Ideally this zero eigenvalue tends to be well separated from the rest of the spectrum. The idea is that $(D-\lambda I)^{-1}$ is an approximate inverse for $(A-\lambda I)^{-1}$, so most of N's eigenvalues are pushed toward 1. The separation of the eigenvalue at 0 causes convergence to be rapid toward the eigenvector z. This can be viewed as using diagonal preconditioning. See [6] for more details.

Davidson's Method

Let $f_1, ..., f_k$ be vectors spanning the initial subspace and let Q be an n by k orthonormal matrix with columns spanning the subspace.

Iterate for $j = k, k+1, ...$

1. Form $H = Q^T A Q$.
2. Find the appropriate eigenpair of H, say (θ, g), and let $y = Qg$.

0021-9991/92 $5.00

3. Form the residual vector $r = (A - \theta I)\, y$, and check $\|r\|$ for convergence.

4. Let $f_{j+1} = (D - \theta I)^{-1}\, r$, where D is the diagonal of A.

5. Orthonormalize f_{j+1} against the previous columns of Q and append as the $(j+1)$th column.

The size of Q and H increases as the algorithm proceeds. The method can be restarted if the orthogonalization cost for Q becomes too great. The approximate eigenpair of A is (θ, y). See [7] for a discussion of how to choose θ from among the eigenvalues of H. See [8] and [9] for approaches if interior eigenvalues are desired.

Davidson's method converges much faster than the Lanczos algorithm for some problems [1]. The expense per iteration of the method is greater than for the Lanczos algorithm, because full orthogonalization is needed. But when the matrix-vector product with A is the main expense, any reduction in the number of iterations is important. Both methods can be implemented with one matrix-vector product per iteration.

Davidson's method is generally effective when the diagonal is a good approximation to the whole matrix. The method was generalized in [6] to allow for other preconditioners. Replace step 4, with $f_{j+1} = (M - \theta I)^{-1}\, r$, where M is any approximation to A. In this way, the preconditioners developed for solving systems of linear equations with the conjugate gradient method [10–12] can be applied to eigenvalue problems. This approach provides a compromise between standard Lanczos with possibly slow convergence and shift-and-invert Lanczos [4, 5] with great factorization expense. The approximate factorization used to implement $(M - \theta I)^{-1}$ can generally be less expensive, yet still give power to the method. This approach is referred to as the GD method (generalized Davidson's) or as a preconditioning method. It can also be applied to generalized eigenvalue problems [13]. For particularly sparse matrices, there is an approach that avoids the full orthogonalization [14].

II. DAVIDSON'S METHOD FOR NONSYMMETRIC MATRICES

Now we consider nonsymmetric problems. The symmetric Lanczos algorithm can be generalized in two ways. One way is the nonsymmetric Lanczos algorithm [15] which also has a three-term recurrence, but is not popular because of instability. The other way is the Arnoldi algorithm [16, 17]. Arnoldi also generates a Krylov subspace but it uses full orthogonalization, so the expense and storage requirements are greater than for symmetric Lanczos.

Meanwhile Davidson's method can be generalized to nonsymmetric problems relatively easily [18, 19]. It can be implemented with the same algorithm as given earlier. The only difference is that complex numbers may appear (see the next section). Since Davidson's method already required full orthogonalization, it is generally even more competitive with Krylov subspace methods in the nonsymmetric case.

The discussion in the previous section for why Davidson's method is effective also holds in the nonsymmetric case. Note that the operator $N = (D - \lambda I)^{-1}\, (A - \lambda I)$ is generally nonsymmetric even with a symmetric matrix A. Effective results have been reported for nonsymmetric Davidson's method [18, 19]. The following theorem shows why Davidson's method is effective when A has large and well separated diagonal elements. In this situation, N is close to a matrix with all of its eigenvalues at 1. For notation, the standard 2-norms is $\|\cdot\|$, and the infinity norm is $\|\cdot\|_\infty$.

THEOREM 1. *Let $A = D + F$, where D is the diagonal of A and F is the off-diagonal portion. Let λ be an eigenvalue of A, and a_{kk} be the diagonal element of A closest to λ. Let $\gamma = \min_{i \neq k} |a_{kk} - a_{ii}|$. Assume that $\|F\|_\infty < \gamma/2$. Then $N = P + E$, where P has all of its eigenvalues at 1, and where $\|E\| < 2\, \|F\|/\gamma$.*

Proof. Note $N = I + (D - \lambda I)^{-1}\, F$. Let P be the portion of N with its diagonal and its kth row, and let E have the rest of N. Interchanging the first and kth rows and columns of P is a similarity transformation, and it yields an upper triangular matrix with all 1's on the diagonal. So all of P's eigenvalues are 1. Meanwhile $E = \{(D - \lambda I)^{-1}\}' F$, where the prime indicates that the kth row is removed. The assumption that $\|F\|_\infty < \gamma/2$ says that the gap between a_{kk} and the other diagonal elements is twice as large as the maximum sum of absolute values of off-diagonal elements in a row. Using Gerschgorin bounds, λ is within $\gamma/2$ of a_{kk}. So for $i \neq k$, $|\lambda - a_{ii}| > \gamma/2$. Therefore

$$\begin{aligned} \|E\| &= \|\{(D - \lambda I)^{-1}\}' F\| \\ &\leqslant \|\{(D - \lambda I)^{-1}\}'\|\, \|F\| \\ &< \frac{2\, \|F\|}{\gamma}. \end{aligned}$$

As discussed earlier, asymptotic convergence of Davidson's method is controlled by the distribution of the eigenvalues of N. We want most of the eigenvalues of N to be clustered around 1. If the diagonal elements of A are extremely well spaced from a_{kk}, then the theorem says that $\|E\|$ is small, so N is close to a matrix with all of its eigenvalues at 1. This does not guarantee that all of N's eigenvalues are near 1. In fact, we know that one is at 0. But in some testing with matrices satisfying the conditions, all but two or three eigenvalues of N were close to 1. This indicates that convergence will be rapid toward the desired eigenvector.

A theorem on the convergence rate would probably be difficult. But if we assume that θ is close enough to λ so that

we can ignore the difference between them in generating the subspace, then it is possible to show that only a couple steps of Davidson's method are required to improve a fairly accurate approximate eigenvector. Decompose the approximate eigenvector as $y = \alpha z + t$, where t is assumed to be small. Note $P = I + e_k v^T$, where e_k is the kth coordinate vector and v is some vector. After one step of Davidson's method, the vector $y - Ny$ is in the subspace. We compute

$$\begin{aligned} Ny &= Nt, \qquad \text{since } (0, z) \text{ is an eigenpair of } N \\ &= Pt + Et \\ &= t + \beta e_k + Et, \qquad \text{for} \quad \beta = v^T t. \end{aligned}$$

Therefore

$$y - Ny = \alpha z - \beta e_k - Et.$$

Another step of Davidson's method will produce another vector of similar form. Then normally the Rayleigh–Ritz procedure will combine the vectors in such a way as to mostly eliminate the unwanted vector e_k. This produces a more accurate approximate eigenvector with only small error terms like Et.

Theorem 1 requires rather strong conditions to be meaningful. One possible improvement is to assume that a_{kk} is well separated from only $n - m$ of the other diagonal elements of A. Then the theorem can be restated with the conclusion that $n - m$ of the eigenvalues of P are equal to 1. This still indicates good convergence after a few steps, if m is small and $\|E\|$ is small. For this we need most of the diagonal elements of A to be well spaced from a_{kk}, relative to the size of the off-diagonal elements.

III. IMPLEMENTATION FOR COMPLEX EIGENVALUES

The major change in Davidson's method for the nonsymmetric case is that H may have complex eigenvalues. If θ is complex, then Q and H become complex. This complicates the implementation and increases the expense. The matrix-vector product takes twice as many real operations and the orthogonalization for Q requires four times as many. By comparison, the major computations in the Arnoldi method remain real [16, 17].

We investigate an approach that keeps Q real in Davidson's method. If f_{j+1} is complex, it is split into two real vectors from its real and imaginary parts. These two vectors are both used as new vectors for the subspace. They are orthonormalized and appended as new columns of Q. There is no disadvantage in separating the two parts, since the Rayleigh–Ritz procedure combines the columns of Q. The separation actually allows for more flexibility, because with more vectors, there are more ways to combine them. It is necessary to add both the real and imaginary parts. Otherwise the theory in the previous two sections for Davidson's method will not hold, and we can not expect good convergence.

If a combination of real and complex eigenvalues is computed, this splitting approach is definitely better, because Q, H, and f_{j+1} are real even after a complex value of θ has occurred. Then computing a real eigenvalue is less expensive since only one real vector is added per iteration.

EXAMPLE 1. As a test matrix, we choose the matrix of dimension 1000 that is tridiagonal except for the first and last rows and columns. The main diagonal has elements going from 1 to 1000, the superdiagonal has -1's, the subdiagonal has 1's, and the first and last rows and columns have 0.1's, except where they intersect the tridiagonal portion. We compute the five eigenvalues with smallest real parts. They are the complex conjugate pair $1.832 \pm 0.828i$ and the real values 3.527, 3.720, and 5.042. The convergence criteria is the residual norm dropping below 10^{-10}. The initial vectors are the first five coordinate vectors. To keep Q close to orthonormal in spite of roundoff error, a vector is reorthogonalized if its norm drops by over 90% during the orthogonalization (in step 5). The computations are performed on an IBM 4381-R14 using double precision.

Table I gives a comparison between the implementation with complex vectors and the implementation that splits the vectors. A complex vector is counted as adding 1 dimension and 1 iteration, but 2 matrix-vector products. Splitting the vector is counted as adding 2 in dimension, 1 iteration, and 2 matrix-vector products. The initial vectors are counted as 1 of each. We see that while the dimension of the subspace increases slightly with the splitting approach, the total number of matrix-vector products is reduced from 65 to 37.

If only complex eigenvalues are computed, the expense per iteration for the splitting approach is approximately the same as without splitting. Two matrix-vector products are needed for each iteration, and orthogonalization costs are four times as great as in the symmetric case, because two vectors are orthogonalized against twice as many previous vectors. However, it turns out that splitting can reduce the number of iterations and be cheaper.

TABLE I

Comparison of Two Implementations

	Example 1		Example 2	
	Complex	Split	Complex	Split
Dimension	35	37	31	39
Mvp's	65	37	57	39
Iterations	35	28	31	22
CPU time	35.5	14.2	26.8	13.9

EXAMPLE 2. The matrix is the same as in example 1 except that the superdiagonal elements are -2 instead of -1. The six eigenvalues with smallest real parts are computed. These are three complex conjugate pairs. Table I again has the results. Splitting the vectors reduces the number of iterations from 31 to 22, so less matrix-vector products are needed (39 instead of 57). It is perhaps surprising that there is such a difference in this case. This difference results from the Rayleigh–Ritz procedure having greater flexibility in combining the split vectors.

IV. GD FOR NONSYMMETRIC MATRICES

The matrices in Examples 1 and 2 are well suited to Davidson's method because the diagonal elements are relatively large and well spaced. But in many problems, the diagonal of the matrix is not such a good approximation to the whole matrix. If there is a better approximation, say M, then M should be used in place of D. This is the GD method [6] applied to nonsymmetric matrices. $M-\theta I$ is a preconditioner for $A-\theta I$. Solution of linear equations in $M-\theta I$ should be fairly inexpensive.

Theorem 1 can be generalized for this situation. Under the right conditions, N is again close to a matrix with all of its eigenvalues at 1.

THEOREM 2. *Let $A=M+F$, where M has spectral decomposition $M=U^{-1}\Delta U$. Let λ be an eigenvalue of A, and δ_k be the eigenvalue of M closest to λ. Let $\gamma=\min_{i\neq k}|\delta_k-\delta_i|$. Assume that $\|U^{-1}FU\|_\infty \leq \gamma/2$. Define $N\equiv(M-\lambda I)^{-1}(A-\lambda I)$. Then $N=P+E$, where P is a matrix with all of its eigenvalues at 1, and $\|E\|<2\|U\|\,\|U^{-1}\|\,\|F\|/\gamma$.*

Proof. Multiplying N on the front and back by U^{-1} and U, respectively, is a similarity transformation. This produces the matrix $(\Delta-\lambda I)^{-1}U^{-1}(M-\lambda I+F)U=I+(\Delta-\lambda I)^{-1}U^{-1}FU$. Separate the kth row of $(\Delta-\lambda I)^{-1}U^{-1}FU$ and append it to I and call this $U^{-1}PU$. Then the proof is similar to that for Theorem 1. All of P's eigenvalues are 1. Again using the prime to denote that the kth row has been deleted,

$$E=U\{(\Delta-\lambda I)^{-1}\}'U^{-1}F,$$

and

$$\|E\|<\frac{2\|U\|\,\|U^{-1}\|\,\|F\|}{\gamma}.$$

From this theorem, it appears that M will be a worthwhile choice as a preconditioner if it has eigenvalues that are well separated relative to the size of the rest of A. If possible, we would like for M to contain the portion of A with the larger elements. And hopefully there is some variety in the sizes of the eigenvalues of M.

TABLE II

Several Methods and Use of Real(θ)

Method	Use θ No. mvp's	Use Real(θ) No. mvp's
Arnoldi	350+	
Davidson	37	33
GD, $M=T$	16	22

EXAMPLE 3. Although the diagonal preconditioning in Davidson's method is effective for the matrix in Example 1, tridiagonal preconditioning is even better. Table II gives the number of matrix-vector products required to find the first five eigenvalues with diagonal and tridiagonal preconditioning. Tridiagonal preconditioning takes only 16 matrix-vector products. Results are also given for the Arnoldi method with restarting every 100 iterations, but only one eigenvalue is computed and this requires 350 matrix-vector products. Using preconditioning is very important for this matrix.

A complex value of θ can significantly increase the expense of factoring $M-\theta I$. In some cases, it is sufficient to use only the real part of θ in the factorization. Note $M-\theta I$ is only an approximation to $A-\theta I$ anyway. See Table II for results with only the real part of θ being used in the preconditioner (all of θ is used in forming the residual vector in step 3). Davidson's method actually improves with just the real part. However, GD with $M=T$ slows down a little. Because $M=T$ is such a good approximation to A, the difference between θ and Real(θ) is significant. Another worthwhile point is that in order to save on factorization expense, it may be desirable to only factor one time. We can use $M-\sigma I$, with a fixed σ, in place of $M-\theta I$ (see [6]).

IV. CONCLUSION

Nonsymmetric Davidson's method is particularly effective if the matrix has diagonal elements that are large and well separated relative to the size of the off-diagonal elements. Some theoretical justification is given for why the method is effective in this case. This is done by relating the asymptotic convergence of Davidson's method to the convergence of the Lanczos algorithm applied to the preconditioned matrix N.

An implementation of nonsymmetric Davidson's method that avoids using complex vectors is also given. New vectors are split into their real and imaginary parts and added separately to the subspace. This can reduce the number of iterations needed for convergence, because the Rayleigh–Ritz procedure has more freedom to combine the separated vectors.

The more powerful preconditioning of the GD method can improve the convergence for some matrices. This makes

these methods applicable to more types of problems. For GD to be successful, a preconditioner is needed that is fairly inexpensive to implement and yet is a good approximation to the matrix.

ACKNOWLEDGMENTS

The author thanks the referees for their suggestions. This project was partially supported by the University of Missouri Research Council.

REFERENCES

1. E. R. Davidson, *J. Comput. Phys.* **17** (1975), 87.
2. B. N. Parlett, *The Symmetric Eigenvalue Problem* (Prentice-Hall, Englewood Cliffs, NJ, 1980).
3. B. Nour-Omid, B. N. Parlett, and R. Taylor, *Int. J. Numer. Methods Eng.* **19**, 859 (1983).
4. T. Ericsson and A. Ruhe, *Math. Comput.* **35**, 1251 (1980).
5. D. S. Scott, *SIAM J. Sci. Stat. Comput.* **3**, 68 (1982).
6. R. B. Morgan and D. S. Scott, *SIAM J. Sci. Stat. Comput.* **7**, 817 (1986).
7. E. R. Davidson, *Comput. Phys. Commun.* **53**, 49 (1989).
8. W. Butscher and W. E. Kammer, *J. Comput. Phys.* **20**, 313 (1976).
9. R. B. Morgan, *Linear Algebra Appl.* **154-156**, 289 (1991).
10. P. Concus, G. H. Golub, and G. Meurant, *SIAM J. Sci. Stat. Comput.* **6**, 220 (1985).
11. G. H. Golub and D. P. O'Leary, *SIAM Rev.* **31**, 50 (1989).
12. J. A. Meijerink and H. A. van der Vorst, *Math. Comput.* **31**, 148 (1977).
13. R. B. Morgan, *J. Comput. Phys.* **89**, 241 (1990).
14. R. B. Morgan, Preconditioning the Lanczos Algorithm for Sparse Symmetric Eigenvalue Problems, to appear, *SIAM J. Sci. Stat. Comput.*
15. B. N. Parlett, D. R. Taylor and Z. A. Liu, *Math. Comput.* **44**, 105 (1985).
16. Y. Saad, *Linear Algebra Appl.* **34**, 269 (1980).
17. Y. Saad, *Comput. Phys. Commun.* **53**, 71 (1989).
18. H. Hirao and N. Nakatsuji, *J. Comput. Phys.* **45**, 246 (1982).
19. S. Rettrup, *J. Comput. Phys.* **45**, 100 (1982).