SIAM J. SCI. COMPUT.
Vol. 15, No. 1, pp. 62–76, January 1994

004

THE DAVIDSON METHOD*

M. CROUZEIX[†], B. PHILIPPE[‡], AND M. SADKANE[§]

Abstract. This paper deals with the Davidson method that computes a few of the extreme eigenvalues of a symmetric matrix and corresponding eigenvectors. A general convergence result for methods based on projection techniques is given and can be applied to the Lanczos method as well. The efficiency of the preconditioner involved in the method is discussed. Finally, by means of numerical experiments, the Lanczos and Davidson methods are compared and a procedure for a dynamic restarting process is described.

Key words. Davidson method, Lanczos method, Krylov space, preconditioner, eigenvalue, sparse matrices, eigenvectors

AMS subject classification. 65F15

1. Introduction.

To compute a few of the extreme eigenvalues and the corresponding eigenvectors of a large, sparse, and symmetric matrix, two classes of iterative methods are usually considered. Their common characteristic is to build a sequence of subspaces that contains, in the limit, the desired eigenvectors. The subspaces of the first class are of constant dimension; this class includes simultaneous iteration [1] and the trace minimization method [11]. In the second class of methods, the sequence is increasing, at least piecewise, since there often exists a restarting process that limits the dimension of the subspaces to a feasible size; the class includes the well-known Lanczos method which is based on Krylov subspaces [6]. This paper deals with another method of the same class, namely, the Davidson method.

Davidson published his algorithm in the quantum chemistry field [2] as an efficient way to compute the lowest energy levels and the corresponding wave functions of the Schrödinger operator. The original algorithm that computes the largest (or the smallest) eigenvalue of the matrix A can be expressed by the following algorithm where D stands for the diagonal of the matrix A.

```
Choose an initial unit vector $v_1$; $V_1 := [v_1]$;
for $k = 1, \dots$ do
    Compute the interaction matrix $H_k := V_k^t A V_k$;
    Compute the largest (or the smallest) eigenpair $(\mu_k, y_k)$ of $H_k$;
    Compute the corresponding Ritz vector $x_k := V_k y_k$;
    Compute the residual $r_k := (\mu_k I - A)x_k$;
    if convergence then exit;
    Compute the new direction to be incorporated $t_{k+1} := (\mu_k I - D)^{-1} r_k$;
    Orthogonalize the system $[V_k; t_{k+1}]$ into $V_{k+1}$;
end for
```

This algorithm looks like an algorithm of the Lanczos type with a diagonal preconditioning. When the dimension of the basis V_k becomes too large, the process restarts with the last Ritz vector as initial vector. In this paper we consider a more general method in the sense that

*Received by the editors November 19, 1990; accepted for publication (in revised form) February 16, 1993.

[†] Institut de Recherche en Informatique et Systèmes Aleatoires/University of Rennes, Campus de Beaulieu, 35042 Rennes Cedex, France (crouzeix@irisa.fr).

[‡] Institut de Recherche en Informatique et Systèmes Aleatoires-Institut National de Recherche en Informatique et en Automatique, Campus de Beaulieu, 35042 Rennes Cedex, France (philippe@irisa.fr).

[§] Centre Européen de Recherche et de Formation Avancée en Calcul Scientifique, 42 Av. G. Coriolis, 31057 Toulouse Cedex, France. Present address: Institut de Recherche en Informatique et Systèmes Aleatoires-Institut National de Recherche en Informatique et en Automatique, Campus de Beaulieu, 35042 Rennes Cedex, France (sadkane@irisa.fr).

- several eigenpairs are sought at the same time;
- several vectors are incorporated in the basis at every step, leading to a block implementation;
- a general preconditioner is considered.

The block adaptation is important with supercomputers since it allows parallelism and efficient use of local memory.

Before analyzing the Davidson method, we formulate, in §2, a general convergence result for methods based on projection techniques; it can be applied to the Lanczos process as well. Consequences for the Davidson method are described in §3. Section 4 is devoted to a discussion on selecting the preconditioner and on the class of matrices on which the algorithm is the method of choice. In §5, numerical experiments illustrate the study and an improvement for the restarting process is proposed.

Notations and general assumptions. $A = (a_{ij})_{1\le i,\ j\le n}$ is a symmetric matrix supposedly large and sparse; $\lambda_1 \ge \cdots \ge \lambda_n$ are its eigenvalues and $u_1, \dots, u_n$ a corresponding set of eigenvectors such that $u_i^t u_j = \delta_{ij}$ (Kronecker's symbol) for $1 \le i,\ j \le n$. The goal consists in computing the l ($l \ll n$) largest (or smallest) eigenpairs of A.

Throughout this paper, the symbol $\|\ .\ \|$ denotes the Euclidean norm and MGS stands for the modified Gram– Schmidt procedure. The orthogonal complement of the subspace spanned by the vectors $x_1, \dots, x_k$ is denoted by $\{x_1, \dots, x_k\}^\perp$.

$\rho(x) = x^tAx/\|x\|^2$ is the Rayleigh quotient of the vector $x \ne 0$ and $R(x_1, \dots, x_k) = \max_{x\in \mathrm{Span}(x_1,\dots,x_k)} \rho(x)$ is the maximum of the Rayleigh quotient over the space spanned by the vectors $x_1, \dots, x_k$.

$\{\mathcal{V}_k\}$ is a sequence of subspaces of $\mathbf{R}^n$ of dimension $n_k \ge l$ and V_k is a matrix whose column set is an orthonormal basis of $\mathcal{V}_k$. The matrix $H_k = V_k^t A V_k$ is called the Rayleigh or interaction matrix; it is of order n_k and its l largest eigenvalues are $\lambda_{k,1} \ge \cdots \ge \lambda_{k,l}$ with the corresponding eigenvectors $y_{k,1}, \dots, y_{k,l}$, which constitute an orthonormal set of vectors in $\mathbf{R}^{n_k}$. The corresponding Ritz vectors $x_{k,1}, \dots, x_{k,l}$ are defined by $x_{k,i} = V_k y_{k,i}$ for $i = 1, \dots, l$. The reals $\lambda_{k,1}, \dots, \lambda_{k,l}$ are called the Ritz values of A over $\mathcal{V}_k$.

2. Proof of convergence.

THEOREM 2.1. *Under the assumption*

$$x_{k,i} \in \mathcal{V}_{k+1} \quad \text{for } i = 1, \dots l \quad \text{and} \quad k \in \mathbf{N},$$

the sequences $\{\lambda_{k,i}\}_{k\in\mathbf{N}}$ *are nondecreasing and convergent.*

Moreover, if, in addition, (i) *for any* $i = 1, \dots, l$ *the set of matrices* $\{C_{k,i}\}_{k\in\mathbf{N}}$ *satisfies the following assumption: there exist* $K_1,\ K_2 > 0$ *such that for any* $k \in \mathbf{N}$ *and for any vector* $v \in \mathcal{V}_k^\perp$: $K_1 \|\ v\ \|^2 \le v^t C_{k,i} v \le K_2 \|\ v\ \|^2$;

(ii) *for any* $i = 1, \dots, l$ *and* $k \in \mathbf{N}$, *the vector* $(I - V_kV_k^t)C_{k,i}(A - \lambda_{k,i}I)x_{k,i}$ *belongs to* $\mathcal{V}_{k+1}$, *then the limit* $\mu_i = \lim_{k\to\infty} \lambda_{k,i}$ *is an eigenvalue of* A *and the accumulation points of* $\{x_{k,i}\}_{k\in\mathbf{N}}$ *are corresponding eigenvectors.*

Proof. The first statement is a straight application of the well-known Courant–Fischer theorem [6] that characterizes the eigenvalues of a symmetric operator.

Let us prove the second statement. Let $r_{k,i} = (\lambda_{k,i}I - A)x_{k,i}$ and $w_{k,i} = (I - V_kV_k^t)C_{k,i}r_{k,i}$. Since the Ritz vectors are unit vectors and since $r_{k,i} = -(I - V_kV_k^t)Ax_{k,i}$, the residuals $r_{k,i}$ belong to $\mathcal{V}_k^\perp$ and are uniformly bounded by $\|A\|$; hence the vectors $w_{k,i}$ are uniformly bounded as well. Moreover, since

$$w_{k,i}^t A x_{k,i} = -r_{k,i}^t C_{k,i} r_{k,i}, \tag{1}$$

and since the matrix $C_{k,i}$ is assumed to be positive definite on $\mathcal{V}_k^\perp$, we may ensure that $w_{k,i} = 0$ if and only if $r_{k,i} = 0$.

When $w_{k,i} \neq 0$, let us denote $v_{k,i} = w_{k,i} / \parallel w_{k,i} \parallel$ and $\Pi_k = [x_{k,1}, \ldots, x_{k,i}, v_{k,i}]$. Π_k is an $n \times (i+1)$ matrix whose columns are orthonormal. Consequently, the matrix $\Pi_k \Pi_k^t$ corresponds to the orthogonal projection onto a subspace of $\mathcal{V}_{k+1}$.

The matrix $\mathcal{H}_{k,i} = \Pi_k^t A \Pi_k$ has the following pattern:

$$\begin{pmatrix} \mu_{k,1} & & & \alpha_{k,1} \\ & \ddots & & \vdots \\ & & \mu_{k,i} & \alpha_{k,i} \\ \alpha_{k,1} & \ldots & \alpha_{k,i} & \beta_k \end{pmatrix},$$

where $\alpha_{k,j} = x_{k,j}^t A v_{k,i}$ for $j = 1, \ldots, i$ and $\beta_k = v_{k,i}^t A v_{k,i}$.

Let $\lambda_{k,1} \geq \lambda_{k,2} \geq \cdots \geq \lambda_{k,i} \geq \lambda_{k,i+1}$ be the eigenvalues of $\mathcal{H}_{k,i}$. Cauchy's interlace theorem and the optimality of the Rayleigh–Ritz procedure [6] ensure that

$$\mu_{k,j} \leq \lambda_{k,j} \leq \mu_{k+1,j}, \qquad j = 1, \ldots, i\ .$$

The Frobenius norm of the matrix $\mathcal{H}_{k,i}$ is

$$\sum_{j=1}^{i} \lambda_{k,j}^2 + \lambda_{k,i+1}^2 = 2 \sum_{j=1}^{i} \alpha_{k,j}^2 + \beta_k^2 + \sum_{j=1}^{i} \mu_{k,j}^2;$$

therefore

$$2 \sum_{j=1}^{i} \alpha_{k,j}^2 = \sum_{j=1}^{i} (\lambda_{k,j} - \mu_{k,j})(\lambda_{k,j} + \mu_{k,j}) + (\lambda_{k,i+1} - \beta_k)(\lambda_{k,i+1} + \beta_k).$$

Evaluating the trace of the matrix $\mathcal{H}_{k,i}$ by $\sum_{j=1}^{i} \mu_{k,j} + \beta_k = \sum_{j=1}^{i+1} \lambda_{k,j}$, we obtain

$$\begin{aligned} 2 \sum_{j=1}^{i} \alpha_{k,j}^2 &= \sum_{j=1}^{i} (\lambda_{k,j} - \mu_{k,j})(\lambda_{k,j} + \mu_{k,j} - \lambda_{k,i+1} - \beta_k) \\ &\leq 4 \parallel A \parallel \sum_{j=1}^{i} (\lambda_{k,j} - \mu_{k,j}), \end{aligned}$$

which implies

$$\alpha_{k,p}^2 \leq 2 \parallel A \parallel \sum_{j=1}^{i} (\mu_{k+1,j} - \mu_{k,j}) \ \text{ for } \ p = 1, \ldots, i.$$

This last bound proves that $\lim_{k \to \infty} \alpha_{k,p} = 0$ for $p = 1, \ldots, i$. Therefore, since from (1), we have the relation

$$r_{k,i}^t C_{k,i} r_{k,i} = - \parallel w_{k,i} \parallel \alpha_{k,i}$$

so that $\lim_{k \to \infty} r_{k,i}^t C_{k,i} r_{k,i} = 0$. From the assumption of uniform positive definiteness of $C_{k,i}$ over $\mathcal{V}_k{}^{\perp}$, and since $r_{k,i} \in \mathcal{V}_k{}^{\perp}$, we may conclude that $\lim_{k \to \infty} r_{k,i} = 0$.

Let x_i be an accumulation point of the sequence $\{x_{k,i}\}$; then $\parallel x_i \parallel = 1$. From the definition of $r_{k,i}$, we obtain by continuity that $\mu_i x_i - A x_i = 0$. □

A straightforward application of the theorem may be obtained for a well-known version of the Lanczos method, namely, the block version with restarting process as defined in [8]. From an initial block S of l vectors that constitute an orthonormal set, the matrix V_k is recursively built in such a way that its columns form an orthonormal basis of the Krylov space which is spanned by the columns of $S, AS, \dots, A^{k-1}S$; this is done while $kl \leq m$, where m is a fixed maximum dimension. The Rayleigh matrix H_k, which is built from V_k, is a block tridiagonal matrix. When kl is larger than m, the process restarts with a new block S that corresponds to the Ritz vectors found with the last matrix V_k. Then, we claim the following corollary.

COROLLARY 2.2. *The block version of the Lanczos algorithm, used with restarting, converges.*

Proof. The Lanczos method corresponds to the situation where $C_{k,i}$ is the identity matrix and where $\mathcal{V}_k$ is the Krylov subspace generated from the block V_1. Therefore Theorem 2.1 may be applied. □

3. The generalized Davidson method.

3.1. Algorithm. The following algorithm computes the l largest (or smallest) eigenpairs of the matrix A; m is a given integer that limits the dimension of the basis.

```
Choose an initial orthonormal matrix $V_1 := [v_1, \dots, v_l] \in \mathbf{R}^{n\times l}$;
for $k = 1, \dots$ do
    1. Compute the matrix $W_k := AV_k$;
    2. Compute the Rayleigh matrix $H_k := V_k^t W_k$;
    3. Compute the $l$ largest (or smallest) eigenpairs $(\lambda_{k,i}, y_{k,i})_{1\leq i\leq l}$ of $H_k$;
    4. Compute the Ritz vectors $x_{k,i} := V_k y_{k,i}$ for $i = 1, \dots, l$;
    5. Compute the residuals $r_{k,i} := \lambda_{k,i} x_{k,i} - W_k y_{k,i}$ for $i = 1, \dots, l$;
           if convergence then exit;
    6. Compute the new directions $t_{k,i} := C_{k,i} r_{k,i}$ for $i = 1, \dots, l$;
    7. if $\dim(V_k) \leq m - l$
           then $V_{k+1} :=$ MGS$(V_k, t_{k,1}, \dots, t_{k,l})$;
           else $V_{k+1} :=$ MGS$(x_{k,1}, \dots, x_{k,l}, t_{k,1}, \dots, t_{k,l})$;
           end if
end for
```

Steps (1)–(5) correspond to the classical Rayleigh–Ritz procedure [6]. We point out that only the last columns of W_k and H_k have to be computed at iteration k. At each iteration, the vectors $t_{k,i}$ are incorporated into the previous subspace. Unlike the Lanczos method, the Rayleigh matrix is dense.

The block size can be greater than l, and this may give faster convergence [3]. Since orthogonalization is performed at every iteration, too large a dimension for the basis implies prohibitive complexity. This is the reason for setting a maximum size for the basis. In §6, a dynamic choice for the restart point is described based on an index of efficiency for the iteration.

The selection of efficient preconditioners $C_{k,i}$ is studied in §4. As remarked in Corollary 2.2, the method becomes equivalent to the Lanczos method when the matrices $C_{k,i}$ are proportional to the identity matrix I. However, since in the Davidson method it is necessary to compute the Ritz vectors explicitly at every iteration, this version of the Lanczos algorithm has a much more expensive complexity than the regular version.

In the classical Davidson method, the preconditioners are built from the diagonal D of the matrix A: $C_{k,i} = (\lambda_{k,i}I - D)^{-1}$, which exists when $\lambda_{k,i}$ is not a diagonal entry of A. This choice is efficient when D is a good approximation of the matrix A in the sense that the matrix of eigenvectors of A is close to the identity matrix. More general preconditioners $C_{k,i} = (\lambda_{k,i}I - M)^{-1}$ have already been studied [5]; as for any preconditioning process, the tradeoff consists in finding a matrix M that speeds up the convergence and keeps the complexity of the preconditioning step at a reasonable level. Finally, we mention that a block version of the Davidson method has already been introduced in [4].

Remark. It can be proved [10] that the accumulation points H of the sequence $\{H_k\}$ are of the form

$$H = \begin{pmatrix} \theta_1 & & & \mathbf{0} \\ & \ddots & & \\ & & \theta_l & \\ \mathbf{0} & & & E \end{pmatrix},$$

where $\theta_1 \geq \cdots \geq \theta_l$ are the l largest eigenvalues of H. Therefore, under the assumption that none of the θ_i, $i = 1, \ldots, l$ is an eigenvalue of the matrix E, the components of the corresponding eigenvectors of H are zero along the second block. As pointed out by Davidson [2], this fact can be used in practice to measure the convergence.

3.2. Convergence. In this section we assume that a diagonal preconditioner is used, i.e., $C_{k,i} = (\lambda_{k,i} - D)^{-1}$ for $i = 1, \ldots, l$, where D is the diagonal of A. We assume also that we require the largest eigenpair of A. The situation is analyzed in two different ways depending on the number of eigenpairs needed. The end of the section is devoted to an example of nonconvergence when the hypotheses of Theorem 2.1 are not satisfied.

3.2.1. Classical algorithm ($l = 1$). Theorem 2.1 ensures the convergence of the Davidson method when $(\lambda_{k,1} - D)^{-1}$ is positive definite. Since the sequence $\{\lambda_{k,1}\}$ is nondecreasing, it is sufficient to start with a vector v_1 such that $(\mu_{1,1} - D)^{-1}$ is positive definite. This can be ensured in the following way:

- Let i_o be the index of the largest diagonal entry of D. If the problem is not reducible into two smaller problems, there exists an index j_o such that $a_{i_o,j_o} \neq 0$.
- Let V_1 be the system $[e_{i_o}, e_{j_o}]$ built from the corresponding canonical vectors.

Since the matrix $H_1 = V_1^t A V_1$ is the matrix

$$\begin{pmatrix} a_{i_o,i_o} & a_{i_o,j_o} \\ a_{i_o,j_o} & a_{j_o,j_o} \end{pmatrix},$$

we have $\mu_{1,1} > a_{i_o,i_o} = \max_{1 \leq i \leq n} a_{i,i}$. In conclusion, the following bounds are obtained:

$$\|C_{k,1}\| \leq \frac{1}{\mu_{1,1} - a_{i_o,i_o}},$$
$$v^t C_{k,1} v \geq \alpha \|v\|^2 \text{ with } \alpha = \frac{1}{\max_{1 \leq i \leq n}(\mu_{1,1} - a_{i,i})}.$$

Hence Theorem 2.1 can be applied.

3.2.2. Block version ($l \neq 1$). The technique defined in the previous case can be used here to ensure that $(\lambda_{k,1} - D)^{-1}$ is positive definite; therefore the convergence is certain for the first eigenpair, but not for the others. However, it is possible to define another preconditioner $C_{k,i} = \mathrm{diag}(\mu_{k,i,1}, \ldots, \mu_{k,i,n})$ by $\mu_{k,i,j} = \min(|\lambda_{k,i} - a_{j,j}|^{-1}, M)$, where M is some large constant. With this preconditioning procedure, convergence is guaranteed for any initial system V_1.

Remark. The above choice of M is made only for the sake of the completion of the proof. In practice, the algorithm works regardless of this assumption.

3.2.3. Example of possible nonconvergence. The following example shows the importance of the assumption of positive definiteness for the preconditioning matrices. Let us assume that we look for the two largest eigenpairs ($l = 2$) of the matrix

$$A = \begin{pmatrix} 4 & 0 & 0 & 0 & 0 \\ 0 & -4 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix},$$

and that the process is initialized by $V_1 = [v_1, v_2]$, where

$$v_1 = \left(\sqrt{\tfrac{7}{8}}, \ \sqrt{\tfrac{1}{8}}, \ 0, \ 0, \ 0 \right)^t,$$

$$v_2 = \left(0, \ 0, \ \sqrt{\tfrac{3(5-\sqrt{5})}{20}}, \ \tfrac{\sqrt{3\sqrt{5}-5}}{2}, \ \sqrt{\tfrac{3(5-2\sqrt{5})}{10}} \ \right)^t.$$

We use diagonal preconditioner, i.e., $C_{k,i} = (\lambda_{k,1} - D)^{-1}$ and restart if $m \geq 4$. A straightforward computation shows that $(\lambda_{k,1}, x_{k,1}) = (3, v_1)$ and $(\mu_{k,2}, x_{k,2}) = (\frac{1}{2}, v_2)$ for all k, although neither 3 nor $\frac{1}{2}$ are eigenvalues of A. Of course, it is clear that the assumption of positive definiteness of the preconditioning matrices is violated.

Remarks. 1. Convergence of the Davidson method is automatic if restarting is not used, because eventually the subspace V_k fills up. The significance of Theorem 2.1 is that it proves convergence even when restarting is used.

2. Even when the sequence $\{r_{k,i}\}$ of the residuals converges to zero, it is not clear that the limit $\mu_i = \lim_{k\to\infty} \lambda_{k,i}$ is the ith eigenvalue of A, since we may create situations where the subspaces $\mathcal{V}_k$ remain orthogonal to a required eigenvector. However, it can be proved [10] that this situation would be unstable and is unlikely to happen in finite arithmetic; it may only increase the number of iterations significantly.

4. Quality of the preconditioner.

In this section, we restrict the study to the case $l = 1$. We assume also that $C_{k,i} = C(\lambda_{k,i})$, where $C(\mu)$ satisfies a Lipschitz condition in a neighbourhood of the first eigenvalue of A. This is the situation when $C(\mu) = (\mu I - M)^{-1}$ with M symmetric with eigenvalues smaller than the largest eigenvalue of A.

Since $l = 1$ we replace the index $(k, 1)$ by k in the algorithm. To simplify the notations, we denote by λ, λ', and $\lambda_{\min}$ the first, second, and last eigenvalue of A, respectively. Let x be the eigenvector corresponding to λ ($\lambda \geq \mu_k > \lambda'$ is assumed). Let θ_k be the angle $\angle(x, x_k)$. We may write $x_k = \alpha_k x + \beta_k y_k$, where $\alpha_k = \cos(\theta_k)$, $\beta_k = \sin(\theta_k)$, and where y_k is a unit vector orthogonal to x. The first lemma relates the convergence of the sequences $\{\mu_k\}$, $\{\theta_k\}$, and $\{\|r_k\|\}$.

LEMMA 4.1. *The following relations are true*:

$$\sqrt{\frac{\lambda - \mu_k}{\lambda - \lambda_{\min}}} \leq |\sin(\theta_k)| \leq \sqrt{\frac{\lambda - \mu_k}{\lambda - \lambda'}}, \tag{2}$$

$$\frac{2\,\|r_k\|}{\sqrt{5}(\lambda - \lambda_{\min})} \leq |\sin(\theta_k)| \leq \frac{\|r_k\|}{\mu_k - \lambda'}. \tag{3}$$

Proof. The proof is based on the sin Θ theorem [6]. □

Lemma 4.2 provides an estimate for the effect of the preconditioning process within one iteration. We may define a unit vector z_k such that the system (x, y_k, z_k) is orthonormal and such that $t_k = \gamma_k x + \delta_k y_k + \sigma_k z_k$ for some scalars γ_k, δ_k, and σ_k.

LEMMA 4.2. *The preconditioning process implies that*

$$t_k = \beta_k C(\lambda)(\lambda I - A)y_k + u_k \quad \textit{where} \quad \|u_k\| = O(\beta_k^2) \tag{4}$$

and

$$0 \le \lambda - \mu_{k+1} \le K_1 \, (\lambda - \mu_k), \tag{5}$$

$$|\sin\theta_{k+1}| \le K_2 |\sin\theta_k|, \tag{6}$$

where

$$K_1 = \frac{(\frac{\sigma_k}{\delta_k})^2}{(\alpha_k - \frac{\beta_k \gamma_k}{\delta_k})^2 + (\frac{\beta_k \sigma_k}{\delta_k})^2} \, \frac{\lambda - \lambda_{\min}}{\lambda - \lambda'}, \tag{7}$$

$$K_2 = \frac{\left|\frac{\sigma_k}{\delta_k}\right|}{\sqrt{(\alpha_k - \frac{\beta_k \gamma_k}{\delta_k})^2 + (\frac{\beta_k \sigma_k}{\delta_k})^2}} \, \frac{\lambda - \lambda_{\min}}{\lambda - \lambda'}. \tag{8}$$

Proof. Since $t_k = C(\mu_k)(\mu_k I - A)x_k$, we may write

$$t_k = \alpha_k(\mu_k - \lambda)C(\mu_k)x + \beta_k C(\mu_k)(\mu_k I - A)y_k,$$

and therefore (2) and the Lipschitz condition on $C(\lambda)$ imply (4).

By definition, $\mu_{k+1} \equiv \rho(x_{k+1}) = R(v_1, \ldots, v_k, t_k)$. Let us consider the vector $s_k = x_k - (\beta_k/\delta_k)t_k$, which belongs to the subspace spanned by V_{k+1}. From the optimality of the Rayleigh–Ritz procedure, we have the bounds

$$\rho(x_{k+1}) \ge R(x_k, t_k) \ge \rho(s_k). \tag{9}$$

Since

$$s_k = \left(\alpha_k - \frac{\beta_k \gamma_k}{\delta_k}\right)x - \frac{\beta_k \sigma_k}{\delta_k} z_k, \tag{10}$$

we obtain from (9)

$$\rho(x_{k+1}) \ge \frac{(\alpha_k - \frac{\beta_k \gamma_k}{\delta_k})^2 \lambda + (\frac{\beta_k \sigma_k}{\delta_k})^2 z_k^t A z_k}{(\alpha_k - \frac{\beta_k \gamma_k}{\delta_k})^2 + (\frac{\beta_k \sigma_k}{\delta_k})^2}$$

$$\ge \lambda - \frac{(\frac{\beta_k \sigma_k}{\delta_k})^2}{(\alpha_k - \frac{\beta_k \gamma_k}{\delta_k})^2 + (\frac{\beta_k \sigma_k}{\delta_k})^2}(\lambda - \lambda_{\min}),$$

which implies

$$\lambda - \mu_{k+1} \le \beta_k^2 K_1 (\lambda - \lambda').$$

Since $\rho(x_k) = \alpha_k^2 \, \lambda + \beta_k^2 \, y_k^t A y_k$, we also have

$$\rho(x_k) \le \lambda - \beta_k^2 \, (\lambda - \lambda').$$

The relation (5) with (7) is obtained from the last two bounds.

From (2) and (5), we obtain

$$\begin{aligned}\sin^2\theta_{k+1} &\le \frac{\lambda-\mu_{k+1}}{\lambda-\lambda'}\\ &\le K_1\frac{\lambda-\mu_k}{\lambda-\lambda'}\\ &\le K_1\frac{\lambda-\lambda_{\min}}{\lambda-\lambda'}\sin^2\theta_k,\end{aligned}$$

which proves the relation (6) with (8). □

The best situation, which cannot be obtained in practice, would be to find a $C(\lambda)$ that admits x as an eigenvector and therefore $\{x\}^\perp$ as an invariant subspace. If we assume

$$\|\,(C(\lambda)(\lambda I-A)-I)\,|_{\{x\}^\perp}\| = \epsilon < 1,$$

then

$$\|t_k-\beta_k y_k\| = O(\beta_k(\beta_k+\epsilon)),$$

which implies

$$\begin{aligned}\gamma_k &= O(\beta_k(\beta_k+\epsilon)),\\ \sigma_k &= O(\beta_k(\beta_k+\epsilon)),\\ \delta_k &= \beta_k + O(\beta_k(\beta_k+\epsilon)),\end{aligned}$$

and therefore

$$\begin{aligned}\frac{\gamma_k}{\delta_k} &= O(\beta_k+\epsilon),\\ \frac{\sigma_k}{\delta_k} &= O(\beta_k+\epsilon).\end{aligned}$$

From (7) and (8), we obtain the estimate

$$K_1 = \left(\frac{\sigma_k}{\delta_k}\right)^2\left(\frac{\lambda-\lambda_{\min}}{\lambda-\lambda'}\right)(1+O(\beta_k(\beta_k+\epsilon))),$$

$$K_2 = \left|\frac{\sigma_k}{\delta_k}\right|\left(\frac{\lambda-\lambda_{\min}}{\lambda-\lambda'}\right)(1+O(\beta_k(\beta_k+\epsilon))).$$

Note that if $\epsilon = 0$, convergence is obtained after one step, since in this case $\sigma_k = 0$ and thus x belongs to the subspace spanned by (x_k, t_k).

The usual way to define the preconditioning matrix is to consider a matrix M that approximates A and hence the matrix $C(\lambda) = (\lambda I - M)^{-1}$. Let us consider two extreme situations: $M = I$ or $M = A$. In the former case, the method becomes equivalent to the Lanczos method as has been pointed out, while in the latter case, the method fails since $t_k = x_k$ and $w_k = (I - V_k V_k^T)t_k = 0$. Therefore M has to be an approximation of A, but with its largest eigenvalue smaller than λ to ensure the positive definiteness of the matrix $(\mu_k I - M)^{-1}$ as discussed in Theorem 2.1. With such a matrix we have

$$t_k = (\mu_k I - M)^{-1}(\mu_k I - A)x_k \tag{11}$$

$$= \alpha_k(\mu_k-\lambda)(\mu_k I - M)^{-1}x + \beta_k(\mu_k I - M)^{-1}(\mu_k I - A)y_k \tag{12}$$

and

$$x_k - t_k = \alpha_k \left(I - (\mu_k - \lambda)(\mu_k I - M)^{-1}\right) x + \beta_k \left(I - (\mu_k I - M)^{-1}(\mu_k I - A)\right) y_k. \tag{13}$$

Expressions (12) and (13) show that when the Ritz pair $(\mu_k,\ x_k)$ starts to approximate the solution $(\lambda,\ x)$, then, provided $(\mu_k I - M)^{-1}$ is bounded and $\left(I - (\mu_k I - M)^{-1}(\mu_k I - A)\right)|_{\{x\}^\perp}$ is small in comparison with $\left(I - (\mu_k - \lambda)(\mu_k I - M)^{-1}\right) x$, the components of t_k in the direction of x are small, whereas the other components of t_k remain about the same as those of x_k. Therefore the vector $x_k - t_k$ has small components except those in the direction of x, and hence constitutes an improvement over x_k. We expect to have an efficient preconditioner when the angle $\angle(x, z_k)$ is smaller than the angle $\angle(x, x_k)$, where $z_k = x_k - t_k$. Table 2 illustrates this point in the context of Example 5.1.

To have an easy-to-invert matrix, the appropriate choice for M may be the main diagonal of A or its tridiagonal part, when A is strongly diagonally dominant in the sense that its eigenvectors are close to the vectors of the canonical basis.

5. Experimental results and implementation.

5.1. Efficiency of the preconditioner.

The usual experience is to consider that the better the preconditioner approximates the matrix, the faster is the convergence. The diagonal preconditioner is the easiest to use, but often a larger part of the matrix, as, for example, the tridiagonal part brings a better efficiency. The following example illustrates an extreme case of the benefit that may be obtained from a good preconditioner.

Example 5.1. A is the matrix of order $n = 1000$ such that

$$a_{i,j} = \begin{cases} i & \text{if} \quad i = j, \\ 0.5 & \text{if} \quad j = i + 1 \text{ or } j = i - 1, \\ 0.5 & \text{if} \quad (i, j) \in \{(1, n), (n, 1)\}, \\ 0 & \text{otherwise.} \end{cases}$$

Table 1 displays the sequence of the residuals corresponding to the largest eigenvalue for the Lanczos and Davidson methods with diagonal and tridiagonal preconditioning, where multA is the number of matrix-vector multiplications. The algorithm is started as discussed in §3.2.1, which means here that $V_1 = [e_n, e_1]$ is used as starting vectors for the Davidson method, and $v_1 = V_1 y_1$ where y_1 is the eigenvector corresponding to the largest eigenvalue of the matrix $V_1^T A V_1$ is the starting vector for the Lanczos method.

Table 2 shows, as claimed in §4, that the efficient preconditioner approximates much better $z_k = x_k - t_k$ to x than does x_k. The vectors x_k, t_k, and x have been defined in (12) and (13) in §4.

Unfortunately, this rule of thumb may fail when the evaluation of the quality of a preconditioner is limited to only the consideration of the norm of its difference with the original matrix. It is well known that when two matrices are close, their spectrum are also close, but not necessarily their eigenvectors. Example 5.2 illustrates such a situation.

Example 5.2. A is the matrix of order $n = 1000$ such that

$$a_{i,j} = \begin{cases} 4 & \text{if} \quad i = j, \\ -1 & \text{if} \quad j = i + 1 \text{ or } j = i - 1, \\ -1 & \text{if} \quad j = i + 2 \text{ or } j = i - 2, \\ 0 & \text{otherwise.} \end{cases}$$

TABLE 1
Sequence of residuals depending on the preconditioner (*Example* 5.1).

		Davidson	Davidson
multA	Lanczos	diagonal	tridiagonal
1	0.5000000e+00	0.5000000e+00	0.5000000e+00
2	0.2702456e+00	0.1913128e+00	0.2056694e+00
3	0.2697534e+00	0.4586425e−01	0.8539853e−04
4	0.6456297e−01	0.7378828e−02	0.3830080e−12
5	0.6436916e−01	0.8900376e−03	
6	0.1032884e−01	0.8615493e−04	
7	0.1028572e−01	0.6973576e−05	
8	0.1236296e−02	0.4852756e−06	
9	0.1229767e−02	0.2962304e−07	
10	0.1185097e−03	0.1610810e−08	
11	0.1177558e−03	0.7897330e−10	
12	0.9479829e−05	0.3541615e−11	

TABLE 2
Example of efficient preconditioner (*Example* 5.1).

iter	$\sin \angle(x_k, x)$	$\sin \angle(z_k, x)$
	(tridiagonal)	(tridiagonal)
1	0.4202633e+00	0.4526842e+00
2	0.1176235e−01	0.7403736e−03
3	0.7440150e−06	0.2458322e−10
4	0.1116130e−12	0.2096840e−12

The diagonal preconditioner is not considered since, as already stated, a constant diagonal is equivalent to no preconditioning, and therefore the Lanczos and Davidson methods become equivalent. Figure 1 plots the behaviour of the residuals corresponding to the largest eigenvalue for the Lanczos and Davidson methods with tridiagonal preconditioning. Both methods have poor convergence. For the Lanczos method, this may be explained by the smallness of the gap ratio of λ_1 [6]: $\lambda_1 - \lambda_2/\lambda_2 - \lambda_n \approx 2.33\ 10^{-8}$, whereas for the Davidson method, the poor performance of the preconditioner may be explained by the near orthogonality of the eigenvectors corresponding to the largest eigenvalue of A and the largest eigenvalue of its tridiagonal part (angle $\approx \frac{\pi}{2}$).

Example 5.3. We consider the matrix $LAP30$ from the Harwell–Boeing test collection. This matrix is of order 900, symmetric with 4322 nonzero elements in its lower part. It is originated from a nine-point discretisation of the Laplacian on the unit square with Dirichlet boundary conditions. The eigenvalues as computed by EISPACK are $\lambda_1 = \lambda_2 = 11.9590598 \geq \lambda_3 = \lambda_4 = 11.9286959 \geq \lambda_5 = \lambda_6 = 11.8784356 \geq \cdots \geq \lambda_{900} = 0.0614628$.

The diagonal of this matrix is constant. So the Davidson method with diagonal preconditioning is equivalent to no preconditioning. We computed its four largest eigenpairs using the Davidson method with tridiagonal preconditioning. In this version of Davidson's algorithm, we decided to restart whenever the maximum size of the basis reaches 40. And once an eigenvector is converged, we put it at the beginning of the basis so that all vectors are orthogonalized against it and continue with a reduced block size. Since eigenvalues/eigenvectors seldom converge at the same time, this strategy can prevent harm and additional work in the most slowly converging eigenvectors. The stopping criterion was satisfied when the residual norm of the sought eigenpair is less than 10^{-7}.

We also computed the four largest eigenpairs of LAP30 by the Lanczos method. The version of the Lanczos method used here is the Lanczos algorithm with selective orthogonalization

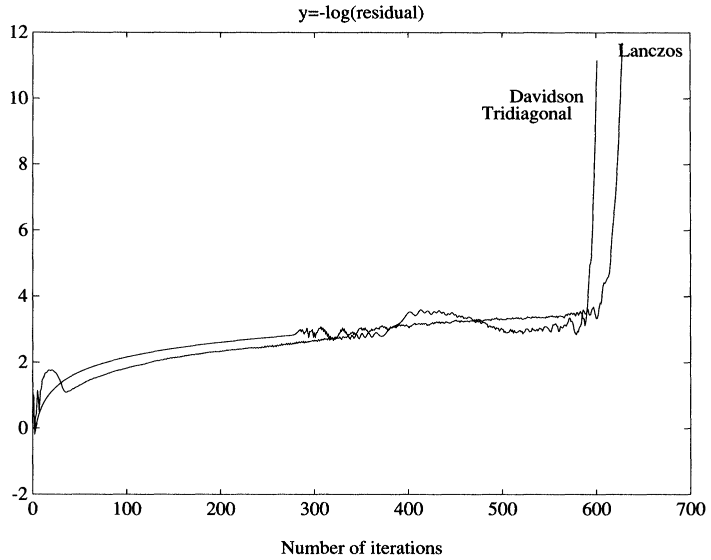

FIG. 1. *Example of a nonefficient preconditioner* (*Example* 5.2).

LASO2, developed by Parlett and Scott [8]. The blocksize, NBLOCK, for Lanczos was taken equal to four, and the number of decimal digits of accuracy desired in the eigenvalues, NFIG, was chosen equal to seven. The computations were carried out on one processor of a Cray 2 and the results are reported in Table 3. The error in eigenvalues and eigenvectors as given by Lanczos (LASO2) are also reported. Unfortunately, we do not have similar estimations for Davidson. But the residual norms and the execution times reveal that the Davidson method with tridiagonal preconditioning performs better on this example.

TABLE 3
Eigenpairs of the matrix LAP30 *by the Lanczos* (LASO2) *and Davidson methods.* (*Example* 5.3).

Method	Matrix–vector products	Time(sec)	Max residual norms	Max error in eigenvalues	Max error in eigenvectors
Lanczos	884	12.534	2.57 E−04	1.13 E−06	4.74 E−03
Davidson	607	11.487	9.36 E−08		

Example 5.4. In Table 4, we compare the Davidson method using diagonal preconditioning with the Lanczos method. (The version used is, again, the Lanczos algorithm with selective orthogonalization LASO2.) The matrix dealt with is of order 1000 and is generated randomly by setting its density of nonzero elements at 0.01. The nonzero off-diagonal entries are in the range $[-1, +1]$; the full diagonal entries are in the range [0, diagscal], where diagscal is a diagonal scaling factor to be varied. The four smallest eigenpairs are sought. The version of Davidson's algorithm uses the same technique concerning converged eigenvectors as described in Example 5.3. The stopping criterion was satisfied when the residual norm of the sought eigenpair is less than 10^{-7}. For the Lanczos algorithm, we chose the block size NBLOCK = 4, and the number of decimal digits of accuracy desired in the eigenvalues NFIG = 7. Experiments were run on a Cray 2. As expected the Davidson method becomes more efficient when the relative importance of the main diagonal increases. The efficiency of the Lanczos method is spectrum-dependent and is not sensitive to the dominance of the matrix.

TABLE 4
Davidson and Lanczos run time comparison (Example 5.4).

Diagonal factor	Time(sec)	
	Davidson	Lanczos
1	5.481	4.131
20	1.091	5.1531
40	0.844	11.574
60	0.657	9.959
80	0.537	16.062
100	0.400	11.171

5.2. Effect of the maximum size for the basis on the convergence. The easiest implementation for the restarting process consists in defining a fixed maximum size for the basis. The selection of an efficient value for m is difficult: too small a value increases the number of steps needed for convergence, whereas too large a value increases complexity and causes numerical problems.

Example 5.5 and Fig. 2 illustrate that the larger m is, the lower is the number of steps necessary to reach convergence.

Example 5.5. A is the matrix of order $n = 5000$ such that

$$a_{i,j} = \begin{cases} \text{if } i = j & \text{random in } [-10, +10], \\ \text{if } i \neq j & \begin{cases} \text{with probability } \alpha : & \text{random in } [-1, +1], \\ \text{ith probability } (1-\alpha) : & 0, \end{cases} \end{cases}$$

where $\alpha = 2 \times 10^{-3}$. There is an average of 11 nonzero entries per row. The eight largest eigenvalues ($l = 8$), that are sought, lie in the range [10.89, 11.57]. Convergence is obtained when the maximum of the L_1 norm of the residuals is smaller than $5. * 10^{-10}$. Experiments were run on an Alliant FX/80.

However, the value of m needs to be limited for three reasons.

1. The memory requirement is roughly proportional to nm, and this introduces a limit on m for large matrices.

2. The orthogonality of V_k is poorly maintained when the number of vectors in V_k is high (a loss of orthogonality plagues the convergence).

3. The complexity of the computation that is involved in one iteration increases with the number of vectors in V_k; therefore it may become too high compared to the benefit obtained from the decrease of the residual norms.

Actually, to be efficient, it is necessary to decide dynamically when to restart the process. The first reason implies a maximum for the size of the basis, but it can be more useful to restart before that limit. The second reason concerns a loss of orthogonality that is detected by an increasing sequence of the norms of the residuals; this may signal a necessary restarting. Let us now consider the detail of the computation involved in one iteration to define some index of efficiency that should indicate when it is worthwhile to restart.

The kth step since the last restart involves l multiplications by A and l applications of the preconditioning process that are of constant complexity. It involves also for the

computation of H_k:	kl^2n	flops;
diagonalization of H_k:	$O(k^3l^3)$	flops;
computation of the Ritz vectors:	kl^2n	flops;
computation of the residuals:	kl^2n	flops;
orthogonalization process:	$2kl^2n$	flops.

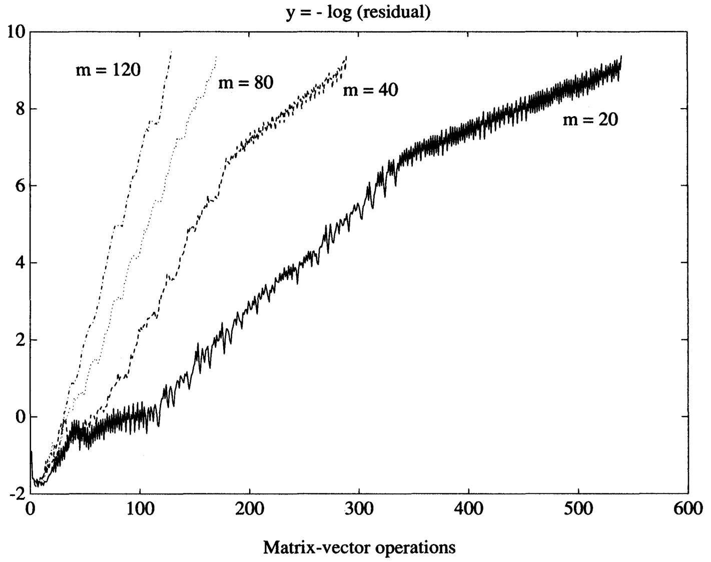

FIG. 2. *Influence of m on the convergence* (*Example* 5.5).

The diagonalization may be estimated as involving approximately $2k^3l^3$ flops. Let us denote by $\mathcal{C}(k)$ the complexity involved at each iteration and by $\tau_k = \|R_k\|/\|R_{k-1}\|$ the local rate of convergence, where R_k stands for the matrix $[r_{k,1}, \dots, r_{k,l}]$. The index of efficiency may be defined as

$$\mathcal{E}_k = \frac{1}{\mathcal{C}_k \tau_k}.$$

By incorporating within the code a procedure that checks the variation of $\mathcal{E}_k$, the process can be restarted as soon as the index decreases significantly.

Example 5.6. The matrix under consideration is the same as in Example 5.5, where its eight largest eigenvalues and their corresponding eigenvectors are sought. Figure 3 plots the variation of the maximum of the residuals with respect to the number of iterations using dynamic restarting. We note in Fig. 3 that the convergence is reached within 140 iterations and 13 irregular restarts, while in Fig. 2 the same convergence with restarting at exactly every 20th step is reached within 500 iterations. This may be explained by the fact that dynamic restarting allows the algorithm to restart in due time, taking into account the quality of the basis for approximating the eigenvectors and the complexity involved during the iterations. The ratio $\tau_k = \|R_k\|/\|R_{k-1}\|$ also measures the changes in rate of convergence. In Table 5, the run with this dynamic restarting procedure is compared to the runs with a static restarting procedure for six values of the maximum block size (20, 40, 60, 80, 100, 120).

The efficiency of the dynamic restarting process is clearly seen in this example, since it corresponds to obtaining the optimum size of the basis automatically.

6. Conclusion. The Davidson method can be regarded as a preconditioned version of the Lanczos method. It appears to be the preferred method for some special classes of matrices, especially those where the matrix of eigenvectors is close to the identity. Although when used with a poor preconditioner it converges slowly, the Davidson method may overcome the Lanczos method tremendously.

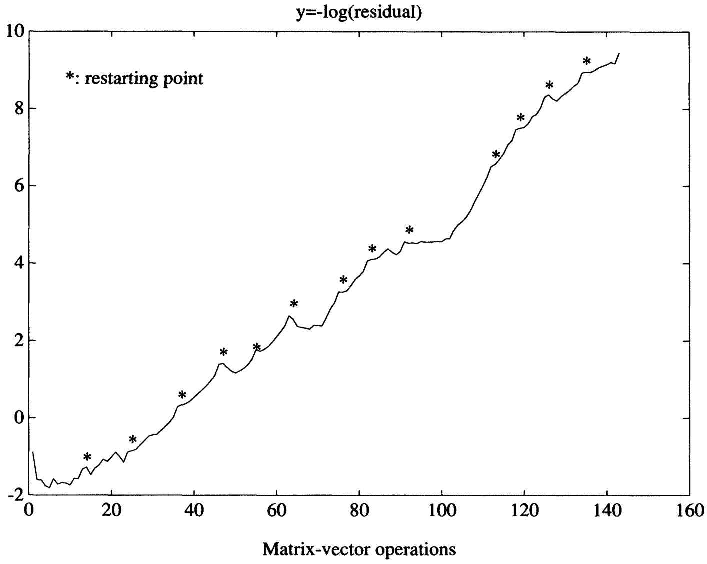

FIG. 3. *Dynamic restarting profile* (*Example* 5.6).

TABLE 5
Comparing static and dynamic restarting procedure (*Examples* 5.5 *and* 5.6).

Running times(s) with dynamic restarting	with fixed restarting	
	m	Times(s)
277.64	20	1015.94
	40	549.09
	60	334.07
	80	345.13
	100	277.83
	120	287.70

Acknowledgement. The authors would like to thank one of the referees for providing many instructive comments.

REFERENCES

[1] M. CLINT AND A. JENNINGS, *The evaluation of eigenvalues and eigenvectors of real symmetric matrices by simultaneous iteration*, Comput. J., 13 (1970), pp. 76–80.

[2] E. R. DAVIDSON, *The iterative calculation of a few of the lowest eigenvalues and corresponding eigenvectors of large real-symmetric matrices*, Comput. Phys., 17 (1975), pp. 87–94.

[3] N. KOSUGI, *Modification of the Liu-Davidson methodfor obtaining one or simultaneously several eigensolutions of a large real-symmetric matrix*, Comput. Phys., 55 (1984), pp. 426–436.

[4] B. LIU, *The simultaneous expansion for the solution of several of the lowest eigenvalues and corresponding eigenvectors of large real-symmetric eigenvectors of large real-symmetric matrices*, in Numerical Algorithms in Chemistry: Algebraic method, C. Moler and I. Shavitt, eds., LBL-8158 Lawrence Berkeley Lab., Univ. of California, 1978, pp. 49-53.

[5] R. B. MORGAN AND D. S. SCOTT, *Generalizations of Davidson's method for computing eigenvalues of sparse symmetric matrices*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 817–825.

[6] B. N. Parlett, *The symmetric eigenvalue problem*, Prentice–Hall, Englewood Cliffs, NJ, 1980.

[7] ———, *The software scene in the extraction of eigenvalues from sparse matrices*, SIAM J. Sci. Statist. Comput., 5 (1984), pp. 590–604.

[8] B. N. Parlett and D. S. Scott, *The Lanczos algorithm with selective orthogonalization*, Math. Comp., 33 (1979), pp. 217–238.

[9] B. Philippe and Y. Saad, *Solving large sparse eigenvalue problems on supercomputers*, in Proc. Internat. Workshop on Parallel Algorithms and Architectures, Oct. 3–6, 1988, Bonas, France, North-Holland, Amsterdam, 1989.

[10] M. Sadkane, *Analyse Numérique de la Méthode de Davidson*, Ph.D thesis, Université de Rennes, France, June 1989.

[11] A. H. Sameh and J. A. Wisniewski, *A trace minimization algorithm for the generalized eigenvalue problem*, SIAM J. Numer. Anal., 19 (1982), pp. 1243–1259.

[12] D. B. Szyld, *A Two-Level Iterative Method for Large Sparse Generalized Eigenvalue Calculations*, Ph.D thesis, Courant Institute, New York, Oct. 1983.