The Davidson method

Michel Crouzeix, Bernard Philippe, Miloud Sadkane

UNITÉ DE RECHERCHE
INRIA-RENNES

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
B.P.105
78153 Le Chesnay Cedex
France
Tél.:(1) 39 63 55 11

Rapports de Recherche

N° 1353

Programme 2
Structures Nouvelles d'Ordinateurs

THE DAVIDSON METHOD

Michel CROUZEIX
Bernard PHILIPPE
Miloud SADKANE

Décembre 1990

RR-1353

IRISA

INSTITUT DE RECHERCHE EN INFORMATIQUE
ET SYSTÈMES ALÉATOIRES

Campus Universitaire de Beaulieu
Avenue du Général Leclerc
35042 - RENNES CÉDEX
FRANCE

THE DAVIDSON METHOD

LA METHODE DE DAVIDSON

Michel CROUZEIX*, Bernard PHILIPPE+ et Miloud SADKANE†

Novembre 1990

Programme 2

Publication interne n°558 - 22 pages.

Abstract

The present paper deals with the Davidson method which computes a few of the extreme eigenvalues of a symmetric matrix and their corresponding eigenvectors. A general convergence result for methods based on projection technics 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.

Keywords : method of Davidson, method of Lanczos, Krylov space, preconditioner.

Résumé

Cet article étudie la méthode de Davidson qui permet de calculer un petit nombre de valeurs propres extrémales d'une matrice symétrique ainsi que leurs vecteurs propres correspondants. La convergence qui est prouvée dans un cadre général de méthodes de sous - espaces s'applique aussi à la méthode de Lanczos. La qualité du préconditionnement de la méthode est étudiée. Enfin, les méthodes de Lanczos et de Davidson sont comparées par des expériences numériques ; une procédure dynamique de redémarrage est définie.

Mots clés: méthode de Davidson, méthode de Lanczos, espaces de Krylov, préconditionnements.

AMS (MOS) : 65F15, 65F10, 65F50.

* IRISA/Université de Rennes, Campus de Beaulieu, 35042 RENNES Cedex, FRANCE.
+ IRISA/INRIA, Campus de Beaulieu, 35042 RENNES Cedex, FRANCE.
† CERFACS, 42 avenue Gustave Coriolis, 31057 TOULOUSE Cedex, FRANCE.

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 which 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 [9]. In the second class of methods, the sequence is increasing, at least piecewise since there often exists a restarting process which limits the dimension of the subspaces to a feasible size; the class includes the well known Lanczos method which is based on Krylov subspaces [4]. 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 which computes the largest (or the smallest) eigenvalue of the matrix A can be expressed by:

Choose an initial unit vector v_1; $V_1 := [v_1]$;
for $k = 1, \ldots$ **do**
  Compute the interaction matrix $H_k := V_k^t A V_k$;
  Compute the largest (or the smallest) eigenpair (λ_k, y_k) of H_k;

Compute the corresponding Ritz vector $x_k := V_k y_k$;
Compute the residual $r_k := (\lambda_k I - A)x_k$;
if convergence **then** exit ;
Compute the new direction to be incorporated $t_{k+1} := (\lambda_k I - D)^{-1} r_k$;
Orthogonalize the system $[V_k; t_{k+1}]$ into V_{k+1};
end for

where D stands for the diagonal of the matrix A. 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. We consider in this paper a more general method in the sense that

- 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 Section 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 Section 3. Sections 4 and 5 are devoted to a discussion on selecting the preconditioner and on the class of matrices on which the algorithm is the method of choice. In Section 6, numerical experiments illustrate the study and an improvement for the restarting process is proposed.

Notations and general assumptions

$A = (a_{ij})_{1 \leq i,\ j \leq n}$ is a symmetric matrix supposedly large and sparse; $\mu_1 \geq \cdots \geq \mu_n$ are its eigenvalues and $u_1, \cdots, u_n$ a corresponding set of eigenvectors such that $u_i^t u_j = \delta_{ij}$ (Kronecker's symbol) for $1 \leq i,\ j \leq n$. 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 $x_1, \ldots, x_k$ is denoted by $\{x_1, \ldots, x_k\}^\perp$.
$\rho(x) = \frac{x^t A x}{\|x\|}$ is the Rayleigh quotient of the vector $x \neq 0$ and $R(x) = \max_{x \in Span(x_1, \cdots, x_k)} \rho(x)$ is the maximum of the Rayleigh quotient over the space spanned by the vectors $x_1, \ldots, x_k$.

$\{\mathcal{V}_k\}$ is a sequence of subspaces of $\mathbf{R}^n$ of dimension $n_k \geq l$ 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 $\lambda_{k,1} \geq \cdots \geq \lambda_{k,l}$ with the corresponding eigenvectors $y_{k,1}, \cdots, y_{k,l}$ which constitute an orthonormal set of vectors in $\mathbf{R}^{n_k}$. The corresponding Ritz vectors $x_{k,1}, \cdots, x_{k,l}$ are defined by $x_{k,i} = V_k y_{k,i}$ for $i = 1, \cdots, l$. The reals $\lambda_{k,1}, \cdots, \lambda_{k,l}$ are called the Ritz values of A over $\mathcal{V}_k$.

We assume that the preconditioning matrices $C_{k,i}$ for $i = 1, \cdots, l$, are l uniformly bounded symmetric matrices.

2 Proof of convergence

Theorem 2.1 *Under the assumption:*

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

the sequences $\{\lambda_{k,i}\}_{k \in \mathbf{N}}$ *are non decreasing and convergent.*
Moreover, if

1. *for any* $i = 1, \cdots, l$ *the set of matrices* $\{C_{k,i}\}_{k \in \mathbf{N}}$ *is uniformly positive definite on* $\mathcal{V}_k^{\perp}$ *(i.e. there exists a real* α *such that for any* $i = 1, \cdots, l$ *and* $k \in \mathbf{N}$ *and for any vector* $v \in \mathcal{V}_k^{\perp}$*:* $v^t C_{k,i} v \geq \alpha \| v \|^2$ *)*
2. *for any* $i = 1, \cdots, l$ *and* $k \in \mathbf{N}$*, the vector* $(I - V_k V_k^t) C_{k,i} (A - \lambda_{k,i} I) x_{k,i}$ *belongs to* $\mathcal{V}_{k+1}$

then the limit $\lambda_i = \lim_{k \to \infty} \lambda_{k,i}$ *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 [4] which characterizes the eigenvalues of a symmetric operator. Let $X_{k,i}$ be the subspace $Span(x_{k,1}, x_{k,2}, \ldots, x_{k,i})$. Then

$$\lambda_{k,i} = \min_{x \in X_{k,i}} \frac{x^t A x}{\| x \|^2} \ .$$

Therefore

$$\lambda_{k+1,i} = \max_{X \subset \mathcal{V}_{k+1}, \dim X = i} \min_{x \in X} \frac{x^t A x}{\| x \|^2} \geq \lambda_{k,i}$$

and from

$$\lambda_{k,i} = \max_{X \subset \mathcal{V}_k, \dim X = i} \min_{x \in X} \frac{x^t A x}{\| x \|^2} \leq \max_{X \subset \mathbf{R}^n, \dim X = i} \min_{x \in X} \frac{x^t A x}{\| x \|^2} = \mu_i$$

it becomes clear that the sequence $\{\lambda_{k,i}\}_{k \in \mathbf{N}}$ is non decreasing and bounded. Let λ_i be its limit.

The second statement is more difficult to prove. Let $r_{k,i} = (\lambda_{k,i} I - A) x_{k,i}$ and $w_{k,i} = (I - V_k V_k^t) C_{k,i} r_{k,i}$. Since the Ritz vectors are unit vectors and since $r_{k,i} = -(I - V_k V_k^t) A x_{k,i}$, 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 $v_{k,i} = \frac{w_{k,i}}{\|w_{k,i}\|}$ and $\Pi_k = [x_{k,1}, \ldots, x_{k,i}, v_{k,i}]$. Π_k is a $n \times (i+1)$ 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

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

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 $\mu_{k,1} \geq \mu_{k,2} \geq \ldots \geq \mu_{k,i} \geq \mu_{k,i+1}$ be the eigenvalues of $\mathcal{H}_{k,i}$. Cauchy's interlace theorem and the optimality of the Rayleigh-Ritz procedure [4] ensure that

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

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

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

therefore

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

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

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

which implies

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

This last bound proves that $\lim_{k \to \infty} \alpha_{k,p} = 0$ for $p = 1, \cdots, i$. 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 definitness of $C_{k,i}$ over $\mathcal{V}_{k+1}^{\perp}$ and since $r_{k,i} \in \mathcal{V}_{k+1}^{\perp}$, we may conclude that $\lim_{k\to\infty} r_{k,i} = 0$ and that λ_i is an eigenvalue of A.

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

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 [6]. From an initial block S of l vectors which 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, $\cdots$, $A^{k-1}S$; 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 which corresponds to the Ritz vectors found with the last matrix V_k. Then, we claim

Corollary 2.1 *The block version of the Lanczos method used with restarting satisfies the assumptions of Theorem 2.1.*

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 Generalized Davidson's method

3.1 Algorithm

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

Choose an initial orthonormal matrix $V_1 := [v_1, \cdots, v_l] \in \mathbf{R}^{n\times l}$;

for $k = 1, \cdots$ **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 $i = 1, \cdots, l$;
5. Compute the residuals $r_{k,i} := \lambda_{k,i} x_{k,i} - W_k y_{k,i}$, for $i = 1, \cdots, l$;
   if convergence **then** exit;
6. Compute the new directions $t_{k,i} := C_{k,i} r_{k,i}$, for $i = 1, \cdots, l$;

7. **if** $\dim(V_k) \leq m$
   then $V_{k+1} := MGS(V_k, t_{k,1}, \cdots, t_{k,l})$;
   else $V_{k+1} := MGS(x_{k,1}, \cdots, x_{k,l}, t_{k,1}, \cdots, t_{k,l})$;
   end if

end for

Steps (1) to (5) correspond to the classical Rayleigh Ritz procedure [4]. 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.

Since a full 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 Section 6, a dynamic choice for the restart point is described. It is based on an index of efficiency for the iteration.

The selection of efficient preconditioners $C_{k,i}$ is studied in Section 4. As remarked in Corollary 2.1, the method becomes equivalent to the Lanczos method when the matrices $C_{k,i}$ are proportional to the identity matrix I. However, since in Davidson's 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. As it will be seen in Section 5, 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 [3]; as for any preconditioning process, the tradeoff consists in finding a matrix M which speeds up the convergence and keeps the complexity of the preconditioning step at a reasonable level.

Remark
It can be proved [8] 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, $i = 1, \cdots, l$ is an eigenvalue of the matrix E, the components of the sought eigenvectors 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

We assume in this section that a diagonal preconditioner is used, i.e. $C_{k,i} = (\lambda_{k,i} - D)^{-1}$ for $i = 1, \cdots, l$ 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 non-convergence when the hypotheses of Theorem 2.1 are not satisfied.

3.2.1 Classical algorithm ($l = 1$)

Theorem 2.1 ensures the convergence of Davidson's method when $(\lambda_{k,1} - D)^{-1}$ is positive definite. Since the sequence $\{\lambda_{k,1}\}$ is non decreasing it is sufficient to start with a vector v_1 such that $(\lambda_{1,1} - D)^{-1}$ 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 $\lambda_{1,1} > a_{i_o,i_o} = \max_{1 \leq i \leq n} a_{i,i}$. In conclusion, the following bounds are obtained:

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

Hence, Theorem 2.1 can be applied.

3.2.2 Block version ($l \neq 1$)

The technique which has been 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 redefine another preconditioner $C_{k,i} = diag(\mu_{k,i,1}, \cdots, \mu_{k,i,n})$ 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.

3.2.3 Example of possible non-convergence

The following example shows the importance of the assumption of positive definitness 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

$$v_1 = \left(\sqrt{\tfrac{7}{8}},\ \sqrt{\tfrac{1}{8}},\ 0,\ 0,\ 0 \right)^t$$
$$v_2 = \left(0,\ 0,\ \sqrt{\tfrac{3(5-\sqrt{5})}{20}},\ \tfrac{\sqrt{3\sqrt{5}-5}}{2},\ \sqrt{\tfrac{3(5-2\sqrt{5})}{10}} \right)^t.$$

A straightforward computation shows that $(\lambda_{k,1}, x_{k,1}) = (3, v_1)$ and $(\lambda_{k,2}, x_{k,2}) = (\frac{1}{2}, v_2)$ for all k, although neither 3 nor $\frac{1}{2}$ are eigenvalues of A. Of course, it is clear that the assumption of positive definitness of the preconditioning matrices is violated.

Remark
Even when the sequence $\{r_{k,i}\}$ of the residuals converges to 0, it is not clear that the limit $\lambda_i = \lim_{k\to\infty} \lambda_{k,i}$ is the i-th 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 [8] that this situation would be unstable and hence cannot 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 the preconditioning matrix satisfies a Lipschitz condition with respect to a parameter λ at all the eigenvalues of A and that $C_{k,i} = C(\lambda_{k,i})$. This is the situation when $C(\lambda) = (\lambda I - M)^{-1}$ 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 λ_{min} the first, second and last eigenvalue of A respectively. Let x be the eigenvector corresponding to λ ($\lambda \geq \lambda_k > \lambda'$ 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 $\{\lambda_k\}$, $\{\theta_k\}$ and $\{\|r_k\|\}$.

Lemma 4.1 *The following relations are true*

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

Proof Since $Ax_k = \alpha_k \lambda x + \beta_k A y_k$ with $x \perp A y_k$, then $\lambda_k = x_k^t A x_k = \alpha_k^2 \lambda + \beta_k^2 \rho(y_k)$ with $\lambda_{min} \leq \rho(y_k) \leq \lambda'$; therefore the first part of the lemma is proved.

In the same way the residual may be expressed by $r_k = \alpha_k(\lambda - \lambda_k)x + \beta_k(\lambda_k I - A)y_k$. Then $\|r_k\|^2 = \alpha_k^2(\lambda - \lambda_k)^2 + \beta_k^2\|(\lambda_k I - A)y_k\|^2$. Since $(\lambda_k - \lambda') \leq \|(\lambda_k I - A)y_k\| \leq (\lambda_k - \lambda_{min}) \leq (\lambda - \lambda_{min})$, we have

$$\beta_k^2\|(\lambda_k I - A)y_k\|^2 \leq \|r_k\|^2 \leq \alpha_k^2(\lambda - \lambda_k)^2 + \beta_k^2(\lambda - \lambda_{min})^2$$

and from (2) and

$$1 + \cos^2\theta_k \sin^2\theta_k \leq \frac{5}{4},$$

the second assertion of the lemma is obtained. □

The second lemma 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

$$0 \leq \lambda - \lambda_{k+1} \leq K_1 \, (\lambda - \lambda_k) \tag{5}$$

$$|\sin\theta_{k+1}| \leq K_2 \; |\sin\theta_k| \tag{6}$$

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 $t_k = C(\lambda_k)(\lambda_k I - A)x_k$, we may write

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

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

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

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

Since

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

we obtain from (9)

$$\begin{aligned} \rho(x_{k+1}) &\geq \frac{(\alpha_k - \frac{\beta_k \gamma_k}{\delta_k})^2 \lambda + (\frac{\beta_k \sigma_k}{\delta_k})^2 z_k^t A z_k}{(\alpha_k - \frac{\beta_k \gamma_k}{\delta_k})^2 + (\frac{\beta_k \sigma_k}{\delta_k})^2} \\ &\geq \lambda - \frac{(\frac{\beta_k \sigma_k}{\delta_k})^2}{(\alpha_k - \frac{\beta_k \gamma_k}{\delta_k})^2 + (\frac{\beta_k \sigma_k}{\delta_k})^2} (\lambda - \lambda_{min}) \end{aligned}$$

which implies

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

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

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

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

From (2) and (5) we obtain

$$\begin{aligned} \sin^2 \theta_{k+1} &\leq \frac{\lambda - \lambda_{k+1}}{\lambda - \lambda'} \\ &\leq K_1 \frac{\lambda - \lambda_k}{\lambda - \lambda'} \\ &\leq K_1 \frac{\lambda - \lambda_{min}}{\lambda - \lambda'} \sin^2 \theta_k \end{aligned}$$

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

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

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

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

$$\begin{aligned} K_1 &= \left(\frac{\sigma_k}{\delta_k}\right)^2 \left(\frac{\lambda - \lambda_{min}}{\lambda - \lambda'}\right) (1 + O(\beta_k(\beta_k + \epsilon))) \\ K_2 &= \left|\frac{\sigma_k}{\delta_k}\right| \left(\frac{\lambda - \lambda_{min}}{\lambda - \lambda'}\right) (1 + O(\beta_k(\beta_k + \epsilon))). \end{aligned}$$

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 which 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 $w_k = 0$. Therefore, M has to be an approximation of A but with its largest eigenvalue smaller than λ to ensure the positive definitness of the matrix $(\lambda_k I - M)^{-1}$. With such a matrix we have

$$(\lambda_k I - M)^{-1}(\lambda_k I - A) = I + (\lambda_k I - M)^{-1}(M - A).$$

We expect to have an efficient preconditioner when the matrix

$$(I - xx^t)(\lambda_k I - M)^{-1}(M - A)\Big|_{\{x\}^\perp}$$

has a small norm.

In order 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 Diagonal preconditioning

We turn back to the general situation where more than one eigenvector is sought ($l > 1$) but for a strongly diagonally dominant matrix. We consider the eigendecomposition of the matrix $A = Q\Lambda Q^t$ with Q orthogonal and Λ diagonal, and the diagonal D of A. We assume that

$$a_{1,1} \geq \cdots \geq a_{l,l} > a_{l+1,l+1} \geq \cdots \geq a_{n,n}$$

Definition 5.1 *The gap between A and its diagonalization is the quantity*

$$\epsilon = \min\left(\|Q - I\| \ ; \ Q^tQ = I \ \text{ and } \ Q^tAQ = \Lambda\right).$$

The matrix A is strongly diagonally dominant when $\epsilon \ll 1$.

Lemma 5.1

$$\|\Lambda - D\| = O(\epsilon^2) \tag{11}$$

Proof For any $i = 1, \cdots, l$ we consider the eigenvector $x_i = e_i + z_i$ where e_i stands for the i-th canonical vector ; then $\|z_i\| = O(\epsilon)$. Since x_i is a unit vector, we obtain that $z_i^t x_i = O(\epsilon^2)$ and therefore $\mu_i - a_{i,i} = O(\epsilon^2)$. □

Lemma 5.2 *If we assume that, for a given iteration k,*

- $\lambda_{k,i} > a_{i+1,i+1}$, *for* $i = 1, \cdots, l$,
- $x_{k,i} = \sum_{j=1}^{l} \alpha_{k,j}^{(i)} e_j + z_{k,i}$ *for* $i = 1, \cdots, l$,
  with $\|z_{k,i}\| = O(\epsilon)$ *and* $z_{k,i} \in \{x_{k,1}, \cdots, x_{k,i}\}^{\perp}$,

then, for $i = 1, \cdots, l$, *the matrix* $C_{k,i} = (\lambda_{k,i} I - D)^{-1}$ *is positive definite on* $\{x_{k,1}, \cdots, x_{k,i}\}^{\perp}$.

Proof The vectors $\{x_{k,i}\}$ and $\{z_{k,i}\}$ are the columns of two matrices X_k and Z_k respectively. The assumptions of the lemma may be expressed by

$$X_k = \begin{pmatrix} B_k \\ 0 \end{pmatrix} + Z_k \quad \text{and} \quad Z_k^t \begin{pmatrix} B_k \\ 0 \end{pmatrix} = 0$$

where $B_k = \left(\alpha_{k,j}^{(i)}\right) \in \mathbf{R}^{l \times l}$ and $\|Z_k\| = O(\epsilon)$. From the orthonormality of X_k, we get

$$B_k^t B_k = I - Z_k^t Z_k \ .$$

Since this last matrix is positive definite when ϵ is small enough, the matrix

$$U = B_k (I - Z_k^t Z_k)^{-1/2}$$

is orthogonal. Therefore

$$B_k = U + O(\epsilon^2) \ . \tag{12}$$

Let us consider a unit vector $s = (\tau_1, \cdots, \tau_n)^t$ in the orthogonal complement of the subspace spanned by the vectors $\{x_{k,i}\}$; hence $X_k^t s = 0$. By denoting $s_1 = (\tau_1, \cdots, \tau_l)^t$ we have

$$s_1 = -B_k^{-t} Z_k^t s.$$

From (12), we obtain $\|s_1\| = O(\epsilon)$. Therefore

$$\begin{aligned} s^t(\lambda_{k,l} I - D)s &= \lambda_{k,l} - \sum_{i=l+1}^{n} \tau_i^2 a_{i,i} + O(\epsilon^2) \\ &\geq \lambda_{k,l} - a_{l+1,l+1} + O(\epsilon^2) \end{aligned}$$

which implies, for a sufficiently small ϵ and for $i = 1, \cdots, l$ that $s^t(\lambda_{k,i} I - D)s > 0$. □

Theorem 2.1 implies the convergence of the method when the assumptions of Lemma 5.2 are satisfied.

6 Experimental results and implementation

6.1 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 which may be obtain from a good preconditioner.

Example 6.1 *A is the matrix of order* $n = 30$ *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 Lanczos method and for Davidson with diagonal and tridiagonal preconditioning.

iter	Lanczos	Davidson Diagonal	Davidson Tridiagonal
1	0.5000000e+00	0.5000000e+00	0.5000000e+00
2	0.2587566e+00	0.1903375e+00	0.2066855e+00
3	0.2398753e+00	0.4548047e-01	0.8574510e-04
4	0.6751775e-01	0.7299598e-02	0.9895190e-10
5	0.4269169e-01	0.8790536e-03	0.3950563e-13
6	0.1620573e-01	0.8497183e-04	
7	0.4745619e-02	0.6870503e-05	
8	0.2306424e-02	0.4776255e-06	

Table 1: Sequence of residuals depending on the preconditioner (Example 6.1).

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. We illustrate such a situation by the following example.

Example 6.2 *A is the matrix of order* $n = 30$ *such that:*

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

The diagonal preconditioner is not considered since, as already stated, a constant diagonal is equivalent to no preconditioning and therefore Lanczos and Davidson's method become equivalent. Table 2 displays the sequence of the residuals corresponding to the largest eigenvalue for Lanczos method and for Davidson with tridiagonal preconditioning. 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 $\approx 0.4995\pi$*).*

iter	Lanczos	Davidson (Tridiagonal)
1	0.5416026e+00	0.5416026e+00
2	0.1499668e+01	0.1443112e+01
3	0.1069298e+01	0.8727821e+00
4	0.7652074e+00	0.3362850e+00
5	0.5635817e+00	0.3130116e+00
6	0.4348702e+00	0.2746764e+00
7	0.3389445e+00	0.2557626e+00
8	0.2351623e+00	0.2576515e+00
9	0.9712444e-01	0.3789814e+00
10	0.8007923e-01	0.2349359e+00
11	0.1171786e+00	0.2229828e+00
12	0.7952803e-01	0.2824065e+00
13	0.1069201e+00	0.8516665e-01
14	0.3614983e-01	0.1681754e-01
15	0.1843160e-08	0.3355165e-02

Table 2: Example of a non-efficient preconditioner (Example 6.2).

Example 6.3 *In the next example, Figure 1, we compare the Davidson method using diagonal preconditioning, with the Lanczos method. 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 range* $[0, diagscal]$*, where diagscal is a diagonal scaling factor to be varied. The five smallest eigenpairs are sought. Experiments were run on a CRAY X-MP. Note that the two methods have opposite behaviour with respect to diagonal dominance.*

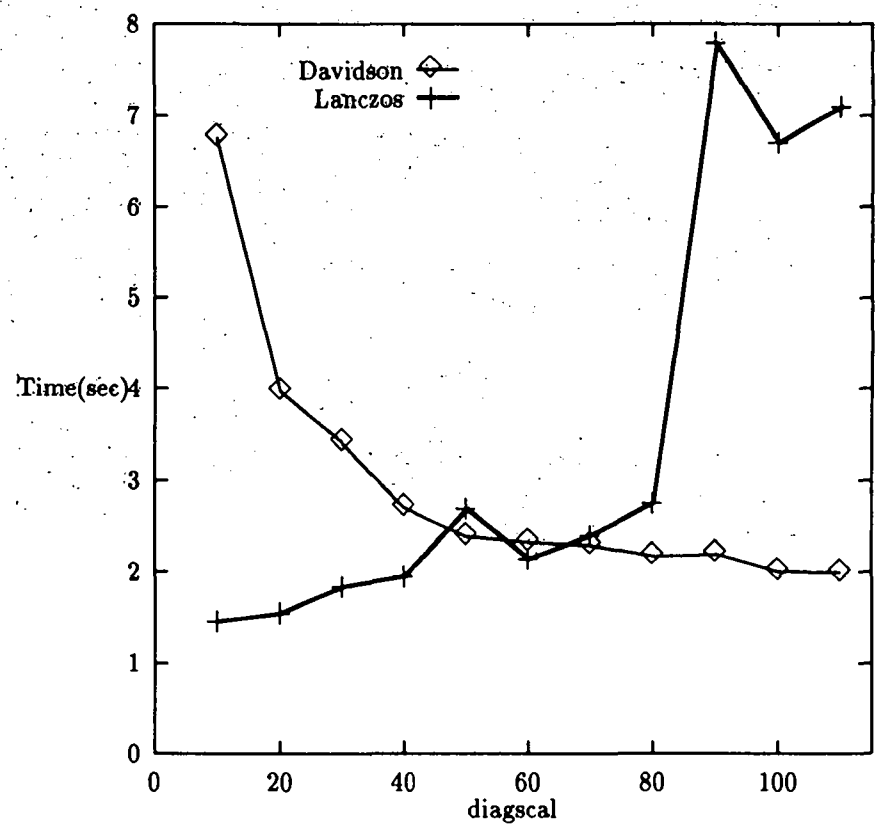

Figure 1: Davidson and Lanczos runtime comparison (Example 6.3).

6.2 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 problem also.

The next example illustrates that the larger m is, the lower is the number of steps necessary to reach convergence.

Example 6.4 *A is the matrix of order $n = 5000$ such that:*

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

where $\alpha = 2 \times 10^{-3}$. *There is an average of 11 nonzero entries per row. The eight largest eigenvalues which 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.*

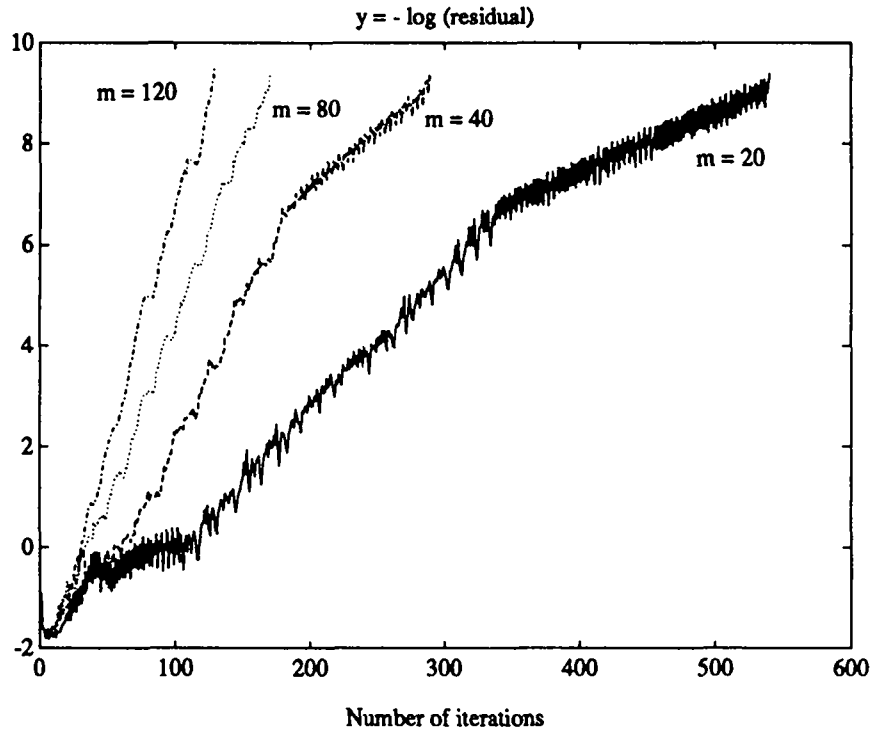

Figure 2: Influence of m on the convergence (Example 6.4).

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 which 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 which is detected by an increasing sequence of the norms of the residuals ; that may

signal a necessary restarting. Let us now consider the detail of the computation involved in one iteration in order to define some index of efficiency which should indicate when it is worthwhile to restart.

The k-th step since the last restart involves l multiplications by A and l applications of the preconditioning process which are of constant complexity. It involves also :

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

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 $[r_{k,1}, \ldots, r_{k,l}]$. 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 which checks the variation of $\mathcal{E}_k$, the process can be restarted as soon as the index decreases significantly.

Example 6.5 *The matrix under consideration is the same as in Example 6.4. In Table 3, the run with a 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 eight largest eigenvalues of A and their corresponding eigenvectors were sought.*

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

Table 3: Comparing static or dynamic restarting procedure (Example 6.4).

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

7 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.

References

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

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

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

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

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

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

[7] B. Philippe and Y. Saad, *"Solving large sparse eigenvalue problems on supercomputers"*, in Proceedings of the International Workshop on Parallel and Distributed Algorithms, M. Cosnard et al., ed., 3-6 October 1988, Bonas, France, 1989, North-Holland.

[8] M. Sadkane, *Analyse numérique de la méthode de Davidson*, PhD thesis, Université de Rennes, June 1989.

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

LISTE DES DERNIERES PUBLICATIONS INTERNES

PI 548 **LES PREDICATS COLLECTIFS : UN MOYEN D'EXPRESSION DU CONTROLE DU PARALLELISME "OU" EN PROLOG**
René QUINIOU, Laurent TRILLING
Septembre 1990, 34 Pages.

PI 549 **NORMALISATION SOUS HYPOTHESE D'ABSENCE DE LIEN APPLICATION AU CAS NOMINAL**
François DAUDE
Septembre 1990, 42 Pages.

PI 550 **MULTISCALE SIGNAL PROCESSING : FROM QMF TO WAVELETS**
Albert BENVENISTE
Septembre 1990, 28 Pages.

PI 551 **ON THE TRANSITION GRAPHS OF AUTOMATA AND GRAMMARS**
Didier CAUCAL, Roland MONFORT
Septembre 1990, 46 Pages.

PI 552 **ERREURS DE CALCUL DES ORDINATEURS**
Jocelyne ERHEL
Septembre 1990, 58 Pages.

PI 553 **SEQUENTIAL FUNCTIONS**
Boubakar GAMATIE, Octobre 1990, 16 Pages.

PI 554 **ANALYSE DE LA FORME D'UN COEFFICIENT D'ASSOCIATION ENTRE VARIABLES QUALITATIVES**
Mohamed OUALI ALLAH
Octobre 1990, 26 Pages.

PI 555 **APPROXIMATION BY NONLINEAR WAVELET NETWORKS**
Qinghua ZHANG, Albert BENVENISTE
Octobre 1990, 16 Pages.

PI 556 **CONCEPTION ET INTEGRATION D'UN CORRELATEUR SYSTOLIQUE**
Catherine DEZAN, Eric GAUTRIN, Patrice QUINTON
Novembre 1990, 16 Pages.

PI 557 **VARIATIONAL APPROACH OF A MAGNETIC SHAPING PROBLEM**
Michel CROUZEIX
Novembre 1990, 14 Pages.

PI 558 **THE DAVIDSON METHOD**
Michel CROUZEIX, Bernard PHILIPPE et Miloud SADKANE
Novembre 1990, 22 Pages.

ISSN 0249-6399