Computation of the singular subspace associated with the smallest singular values of large matrices

Bernard Philippe, Miloud Sadkane

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Computation of the singular subspace associated with the smallest singular values of large matrices

Bernard Philippe and Miloud Sadkane

N° 2064

septembre 1993

PROGRAMME 6

Calcul scientifique,
modélisation
et logiciels numériques

ISSN 0249-6399

INRIA
RENNES

Computation of the singular subspace associated with the smallest singular values of large matrices

Bernard Philippe and Miloud Sadkane

Programme 6 — Calcul scientifique, modélisation et logiciel numérique
Projet ALADIN

Rapport de recherche n ° 2064 — septembre 1993 — 16 pages

Abstract: We compare the block-Lanczos and the Davidson methods for computing a basis of a singular subspace associated with the smallest singular values of large matrices. We introduce a simple modification on the preconditioning step of Davidson's method which appears to be efficient on a range of large sparse matrices.

Key-words: Block-Lanczos method, Davidson's method, SVD, preconditioning, sparse matrices.

(Résumé : tsvp)

Unité de recherche INRIA Rennes
IRISA, Campus universitaire de Beaulieu, 35042 RENNES Cedex (France)
Téléphone : (33) 99 84 71 00 – Télécopie : (33) 99 38 38 32

Calcul du sous espace singulier associé aux plus petites valeurs singulières de matrices creuses de grande taille

Résumé : Nous comparons la méthode de Lanczos par blocs et la méthode de Davidson pour calculer le sous-espace singulier associé aux plus petites valeurs singulières de matrices creuses de grande taille. Nous introduisons une modification sur la méthode de Davidson qui se révèle efficace par rapport aux deux méthodes précédentes.

Mots-clé : méthode de Lanczos par blocs, méthode de Davidson, SVD, préconditionnement, matrices creuses.

1 Introduction

In this paper, we consider the problem of computing a few smallest singular values and the associated singular vectors of a large sparse $m \times n \quad (m \geq n)$ rectangular matrix. This problem has attracted a great deal of interest from a variety of perspectives [1, 2, 4, 7, 8]. Among the examples mentioned in these references, one can cite seismic tomography where the smallest singular values and their corresponding singular vectors are required. In total least square applications, one is interested in solving the linear system $Ax = b$ by transforming it to the linear homogeneous system $[A, b] \begin{pmatrix} x \\ -1 \end{pmatrix} = 0$, where only the computation of the right singular vectors of the appended matrix $[A, b]$ associated with its zero singular value is required . However, the algorithms discussed there are either not suitable for large matrices or give efficient approximations only for the largest singular values. If the size of the matrix A is extremely large, it is not possible to rely upon the Singular Value Decomposition algorithm [7] due to the expense of storage requirements and the high computational cost. An alternative way to proceed for solving our problem is to apply the Lanczos method [10] to the matrix $B = \begin{pmatrix} 0 & A \\ A^T & 0 \end{pmatrix}$. The eigenvalues of B are $\pm\sigma_i(A), i = 1, \ldots, n$ the singular values of A, with $m-n$ additional zeros. This approach might not be efficient since the smallest singular values of A lie in the interior of the spectrum of B, and this part of the spectrum is usually the most difficult to compute for Lanczos' method. The second approach is to apply the Lanczos method to find the smallest eigenpairs of the matrix $A^T A$. The rate of convergence for the smallest eigenvalue σ_1^2 of $A^T A$ is governed by the gap ratio $\gamma = \frac{\sigma_2^2 - \sigma_1^2}{\sigma_n^2 - \sigma_2^2}$ [10] where σ_2 and σ_n are respectively the second and the largest singular values of A. The smallest is this gap ratio, the slower is the convergence. If σ_1 and σ_2 are very small compared to σ_n, then γ is much smaller than $\frac{\sigma_2 - \sigma_1}{\sigma_n - \sigma_2}$ which would be the ideal gap ratio. Therefore the smallest eigenvalues of $A^T A$ becomes smaller and closer and in this case one cannot expect fast convergence.

An other approach which might be of interest is Davidson's method [5], [9], [3] applied to $A^T A$. Davidson's method does exhibit the same behaviour as Lanczos' method, but it is less sensitive to the distribution of eigenvalues. However, to be efficient, Davidson's method needs a good preconditioner. The main difference between Lanczos and Davidson methods is that the former finds several eigenvalues from one Krylov subspace whereas the latter

adapts separately the vectors from which the eigenvectors are computed. The aim of this paper is to compare these two methods, namely the block Lanczos and the block Davidson methods applied to A^TA for computing the right singular vectors corresponding to the smallest singular values of A. One advantage of these two methods over the classical SVD method is that the (large) original matrix is not altered and that little storage is required since only block matrix-vector multiplications are computed.

Since the smallest eigenvalues of A^TA can become closer, we choose to use a block version for the two algorithms. The block strategy although much more expensive than the standard one, is accepted because of the following reasons:

- It may improve the numerical efficiency: Lanczos with block size l can compute close eigenvalues and eigenvalues of multiplicity less than l whereas the standard Lanczos algorithm may not. For Davidson's method, only the block version allows the computation of several eigenpairs at the same time.

- It involves BLAS 3 primitives which are more efficient for memory management and for parallelism.

Throughout this paper $A = U\Sigma W^T$ denotes the singular value decomposition of the matrix A where

$U = [u_1, \ldots, u_m]$ are the orthonormal left singular vectors,

$\Sigma = diag(\sigma_i)$ the diagonal matrix of singular values $\sigma_1 \leq \sigma_2 \leq \ldots \leq \sigma_n$

and $W = [w_1, \ldots, w_n]$ denotes the orthonormal right singular vectors.

We will be concerned only with the computation of right singular vectors since the left ones can be obtained for example by noticing that they satisfy $u_i = \frac{1}{\sigma_i} A w_i$. This formula may, however, not give a satisfactory accuracy on the left singular vectors if w_i is not well approximated. One remedy discussed in [2] consists in using the obtained left approximation as starting point and refining the approximation of the left singular vector using inverse iteration on AA^T or on $\begin{pmatrix} 0 & A \\ A^T & 0 \end{pmatrix}$.

This paper is organized as follows. In §2 we briefly recall the block Lanczos algorithm, in §3 we recall Davidson's method and give some convergence analysis for computing the smallest singular values. We also discuss several preconditioning techniques for Davidson's method. §4 is devoted to numerical experiments and comparison between the two methods for computing the smallest singular values.

2 The block Lanczos algorithm for $A^TAx = \sigma^2 x$

A block version of Lanczos algorithm with block size l for the $m \times n$ matrix A^TA can be written in the following

2.0.1 Algorithm 1

- Set $V_0 = 0, \quad B_1 = 0$
  Choose $V_1 \in \mathbf{R}^{n \times l}$ with $V_1^T V_1 = I$
  $A_1 = V_1^T A^T A V_1$
- for $k = 1, 2, \ldots, maxiter$
  $S_j = A^T A V_j - V_{j-1} B_j^T$
  $A_j = V_j^T S_j$
  $R_{j+1} = S_j - V_j A_j$
  $V_{j+1} B_{j+1} = R_{j+1} \quad QR\ decomposition$

The matrix A^TA is of course never formed explicitly, only successive computation of the form $A^T(A(V))$ are needed. The matrices V_k, S_k, R_k for $k = 1, 2, \ldots$ are $n \times l$, A_j and B_j are $l \times l$, with A_j symmetric. The matrices V_{j+1} and B_{j+1} are defined by the QR factorization of R_{j+1}, so that B_{j+1} is upper triangular and the column of V_{j+1} are orthonormal. The block Lanczos vectors can be grouped together as the columns of an $n \times kl$ matrix $\mathcal{V}_k$ where $\mathcal{V}_k = [V_1, V_2, \ldots, V_k]$; it is easy to show that the columns of $\mathcal{V}_k$ remain orthonormal provided none of the upper triangular matrices B_j are rank deficient. Furthermore the columns of $\mathcal{V}_k$ form an orthonormal basis of the Krylov subspace $K_k(A^TA, V_1) = Span\{V_1, A^TAV_1, \ldots, (A^TA)^{k-1}V_1\}$. The restriction T_k of the matrix A^TA to $K_k(A^TA, V_1)$ is the $kl \times kl$ band matrix

$$T_k = \mathcal{V}_k^t A^T A \mathcal{V}_k = \begin{pmatrix} A_1 & B_2^t & 0 & \ldots & 0 \\ B_2 & A_2 & B_3^t & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & B_k^t \\ 0 & \ldots & 0 & B_k & A_k \end{pmatrix} \tag{1}$$

with half band $l + 1$.

The convergence analysis of the block Lanczos algorithm, that is the convergence of some eigenvalues of T_k towards some eigenvalues of A^TA has

already been studied in [13] [12]. We recall one of the main convergence result.

Theorem 2.1 *Let $\theta_1 \leq \theta_2 \leq \ldots \leq \theta_{kl}$ be the eigenvalues of T_k labelled in increasing order, and assume that k steps of block Lanczos algorithm have been carried out, then for $j = 1, \ldots, l$*

$$0 \leq \frac{\theta_j^2 - \sigma_j^2}{\sigma_n^2 - \sigma_j^2} \leq \left[\frac{tan(\Theta_l)}{T_{k-1}\left(\frac{1+\gamma_j}{1-\gamma_j}\right)} \right]^2 \quad where \quad \gamma_j = \frac{\sigma_{l+1}^2 - \sigma_j^2}{\sigma_n^2 - \sigma_j^2} \tag{2}$$

and

$$0 \leq \frac{\sigma_{n-j+1}^2 - \theta_{kl-j+1}^2}{\sigma_{n-j+1}^2 - \sigma_1^2} \leq \left[\frac{tan(\Phi_l)}{T_{k-1}\left(\frac{1+\tau_j}{1-\tau_j}\right)} \right]^2 \tag{3}$$

where $\gamma_j = \frac{\sigma_{l+1}^2 - \sigma_j^2}{\sigma_n^2 - \sigma_j^2}$, $\tau_j = \frac{\sigma_{n-j+1}^2 - \sigma_{n-l}^2}{\sigma_{n-j+1}^2 - \sigma_1^2}$, Θ_l $(resp. \Phi_l)$ denotes the principle angle between $span\{V_1\}$ and the invariant subspace associated with the l smallest (resp. largest) singular values of A and T_{k-1} is the Chebyshev polynomial of order $k-1$.

Proof See [13], and especially [12] for an improved result. □

It is easy to see that the bounds (2) and (3) are quite satisfactory in the case of convergence to an extremal singular value of A. However, in practice, it is not always true that both ends of the spectrum are equally approximated.

In the case of the smallest singular value, bound (2) shows that the convergence rate of block Lanczos method depends on the gap between the square of the smallest singular value σ_{min} and the square of the next l^{th} singular value σ_{l+1}, and on the spread $\sigma_{max}^2 - \sigma_{l+1}^2$of the unwanted singular values. The larger this gap and this spread the larger the gain in speed. Note that the gap and the spread would be $\sigma_2^2 - \sigma_{min}^2$ and $\sigma_{max}^2 - \sigma_2^2$ if we had used the standard Lanczos method. We conclude that if the smallest singular values are close, block Lanczos with block size as large as the number of singular values in a given cluster can be helpful for accelerating the convergence.

In Theorem (2.1), it is not assumed that Algorithm 1 restarts periodically, thing that is very often used as remedy to the growth in storage [11]. There is, however, no difficulty in proving convergence of Algorithm 1 when restarting is used, but bounds similar to (2) and (3) remain to be done in this case.

We do not want to develop any further the properties of the block Lanczos and prefer to refer to the literature [10] for the details.

3 The generalized Davidson method for $A^TAx = \sigma^2 x$

Davidson published his algorithm in quantum chemistry field [5] as an efficient way to compute the lowest energy levels and the corresponding wave functions of the Schrödinger operator. The matrix dealt with must be strongly diagonally dominant in the sense that its eigenvectors are close to the canonical vectors. The algorithm uses then the diagonal as preconditioner. In [3, 9] Davidson's method has been used with general preconditioner.

From now on, C_k stands for a set of $n \times n$ preconditioning matrices whose choice will be discussed later.

3.1 Algorithm 2

The following algorithm computes the l smallest eigenpairs of the matrix A^TA; m is a given integer which limits the dimension of the basis. MGS stands for Modified Gram Schmidt Procedure

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 $U_k := AV_k$;
2. Compute the matrix $W_k := A^TU_k$;
3. Compute the Rayleigh matrix $H_k := V_k^TW_k$;
4. Compute the l smallest eigenpairs $(\nu_{k,i}^2, y_{ki})_{1 \le i \le l}$ of H_k;
5. Compute the vectors $x_{k,i} := V_k y_{k,i}$, for $i = 1, \cdots, l$;
6. Compute the residuals $r_{k,i} := W_k y_{k,i} - \nu_{k,i}^2 x_{k,i}$ for $i = 1, \cdots, l$;
   if convergence **then** exit;
7. Compute the new directions $t_{k,i} := C_{ki} r_{k,i}$, for $i = 1, \cdots, l$;
8. **if** $\dim(V_k) \le m - l$
   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

Here again, an important characteristic of the algorithm is that the matrix A^TA is not required explicitly. All that it is required is two subroutines that compute $A * u$ and $A^t * v$ for given vectors u and v. At step k, the basis V_{k+1} is obtained from V_k by incorporating the vectors $t_{k,i} = C_{k,i}r_{k,i}, \quad i = 1, \ldots, l$ after orthonormalization. The subspace spanned by V_k is not a Krylov subspace and if the matrices $C_{k,i}$ are not diagonal, then l linear systems must be solved at each iteration. The hope is to reach the convergence very quickly with a small value of m, thus rewarding the extra cost involved by these system resolutions. A detailed convergence analysis of the above algorithm can be found in [3]. We give here a simplified convergence result for the smallest singular value.

3.2 Rate of convergence of the smallest singular value

For $i = 1, 2, \ldots, l$, it is clear that the sequence $\{\nu_{k,i}^2\}$ is decreasing and bounded below by σ_i^2, hence it converges.

For the sake of simplicity we restrict the study to the case $l = 1$ where only the smallest singular value is sought. The numerical experiments deal, however, with the case where more than one singular value is computed. We simplify the notation in Algorithm 2 and drop the subscript i. We denote by (σ^2, x) the smallest eigenpair of A^TA, σ' and σ_{max} denote the second and the largest singular values of A, by $\{(\nu_k^2, x_k)\}_k$ the sequence of Ritz value/Ritz vector obtained at step 5 and by $t_k = C_k r_k$ where $r_k = (A^TA - \nu_k^2 I)x_k$ the step 7. We finally denote by s_k the vector $x_k - t_k$ and by ψ_k the angle $\angle(V_k, s_k)$ between the subspace spanned by V_k and the vector s_k.

We have the following result:

Lemma 3.1

$$|r_k^T C_k r_k| \leq \| s_k \|_2 \sqrt{\| Av_{k+1} \|_2^2 - \nu_{k+1}^2} \sqrt{\nu_k^2 - \nu_{k+1}^2} |\sin(\psi_k)| \qquad (4)$$

Proof The new vector v_{k+1} of the basis constructed at step k of Algorithm 2 is such that $v_{k+1} = \frac{z_{k+1}}{\|z_{k+1}\|_2}$ where $z_{k+1} = (I - V_k V_k^T)C_k r_k$.

Since ν_{k+1}^2 is the smallest eigenvalue obtained from V_{k+1}, the optimality of Raleigh-Ritz procedure [10] ensures that

$$\nu_{k+1}^2 \quad \leq \quad \frac{(x_k - \alpha v_{k+1})^T A^T A (x_k - \alpha v_{k+1})}{1 + \alpha^2} \quad \forall \alpha \in \mathbf{R}$$

$$= \frac{\nu_k^2 - 2\alpha v_{k+1}^T A^T A x_k + \alpha^2 v_{k+1}^T A^T A v_{k+1}}{1+\alpha^2} \quad \forall \alpha \in \mathbf{R}.$$

Hence, for any $\alpha \neq 0$, such that $sgn(\alpha) = sgn(v_{k+1}^T A^T A x_k)$

$$|2 v_{k+1}^T A^T A x_k| \leq \frac{1}{|\alpha|}(\nu_k^2 - \nu_{k+1}^2) + |\alpha|(v_{k+1}^T A^T A v_{k+1} - \nu_{k+1}^2)$$

By choosing $|\alpha| = \left(\frac{\nu_k^2 - \nu_{k+1}^2}{v_{k+1}^T A^T A v_{k+1} - \nu_{k+1}^2} \right)^{\frac{1}{2}}$ we have

$$|v_{k+1}^T A^T A x_k| \leq (\nu_k^2 - \nu_{k+1}^2)^{\frac{1}{2}} (v_{k+1}^T A^T A v_{k+1} - \nu_{k+1}^2)^{\frac{1}{2}}.$$

The proof follows by noticing that

$$v_{k+1}^T A^T A x_k = \frac{r_k^T C_k r_k}{\| z_{k+1} \|_2}$$

and that

$$\| z_{k+1} \|_2 = \| (I - V_k V_k^T) s_k \|_2 = \| s_k \|_2 |\sin(\psi_k)|$$

□

Theorem 3.1 *If the preconditioning matrix C_k is positive (or negative) definite, then Algorithm 2 converges.*

Proof If the preconditioning matrix C_k is positive definite, then from (4) and the convergence of $\{\nu_k^2\}$, we have $\lim_{k\to\infty} r_k^T C_k r_r = 0$ and thus $\lim_{k\to\infty} r_k = 0$. We conclude that $\lim_{k\to\infty} \nu_k$ is a singular value of A. □

4 Choice of preconditioning

We now discuss some choices of the preconditioning matrix C_k. Note that if no preconditioning is used, that is if $C_k = \gamma\ I, \gamma \neq 0$, then the block Lanczos method is recovered under an costly implementation. On the other hand, as was already pointed out a diagonal preconditioning of the form $C_k = (D - \nu_k^2 I)^{-1}$ with $D = diag(A^T A)$ is the simplest choice. It is effective only if the eigenvectors of $A^T A$ are close to the canonical vectors. If we take a good approximation to $A^T A$, then we are faced to the problem of solving (several) linear systems at each iteration and the amount of work could be untolerably high.

In order to overcome this difficulty, we propose to take $C_k = M^{-1}$ as a preconditioner in Algorithm2, where M is a good approximation to A^TA. An advantage is that the cost remains low since the matrix M is constant during the iterations. Furthermore it is clear that the convergence result of Theorem 3.1 still holds as soon as M is positive (or negative) definite. With this modification $s_k \equiv x_k - t_k = (I - M^{-1}A^TA)x_k + \nu_k^2 M^{-1}x_k$ which shows that if M is a good approximation to A^TA, and if we write $x_k = \alpha_k x + \beta_k y_k$, where y_k is of norm 1 and $y_k \perp x$, the components of s_k will be dominated by those of $(A^TA)^{-1}x_k = \frac{\alpha_k}{\sigma^2}\, x + \beta_k (A^TA)^{-1}y_k$. Now if σ is small compared to the other singular values, then the components of x becomes increasingly dominant.

Theorem 4.1 *Assume that M is such that at some step k, $\| \left(M^{-1} - (A^TA)^{-1}\right) r_k \|_2 = O(\epsilon)$, then*

$$0 \leq \frac{\nu_{k+1}^2 - \sigma^2}{\nu_k^2 - \sigma^2} \leq \frac{\sigma^4}{\sigma'^2(\sigma'^2 - \sigma^2)}\, \frac{\sigma_{max}^4 - \sigma^2\sigma'^2}{\sigma_{max}^4 - \sin^2\theta_k(\sigma_{max}^4 - \sigma^4)} + O(\epsilon) \qquad (5)$$

where $\theta_k = \angle(x_k, x)$ is the angle between x_k and x.

Proof $t_k = M^{-1}r_k$ whence $s_k \equiv x_k - t_k = \nu_k^2(A^TA)^{-1}x_k - u_k$ with $u_k = (M^{-1} - (A^TA)^{-1})r_k$

Now if we decompose the vector x_k as $x_k = cos\theta_k x + \sin\theta_k y_k$ where $y_k \in \{x\}^\perp$, then $s_k = \nu_k^2 \left[\frac{1}{\sigma^2} cos\theta_k x + \sin\theta_k (A^TA)^{-1}y_k\right] + u_k$.
The optimality of Rayleigh-Ritz procedure ensures that

$$\nu_{k+1}^2 \leq \frac{s_k^T A^T A s_k}{\| s_k \|^2} \qquad (6)$$

so that

$$\begin{aligned}
\nu_{k+1}^2 - \sigma^2 &\leq \frac{s_k^T(A^TA - \sigma^2 I)s_k}{\| s_k \|^2} && (7)\\
&= \frac{\nu_k^4 \sin^2\theta_k\, y_k^T \left((A^TA)^{-1} - \sigma^2(A^TA)^{-2}\right) y_k + O(\epsilon^2)}{\nu_k^2 \left(\frac{1}{\sigma^4} cos^2\theta_k + sin^2\theta_k \| (A^TA)^{-1}y_k \|^2\right) + O(\epsilon)} && (8)\\
&= \frac{\nu_k^4 \sin^2\theta_k\, y_k^T \left((A^TA)^{-1} - \sigma^2(A^TA)^{-2}\right) y_k}{\nu_k^2 \left(\frac{1}{\sigma^4} cos^2\theta_k + sin^2\theta_k \| (A^TA)^{-1}y_k \|^2\right)} + O(\epsilon) && (9)
\end{aligned}$$

$$\leq \frac{\frac{1}{\sigma'^2} - \frac{\sigma^2}{\sigma_{max}^4}}{\frac{1}{\sigma^4}cos^2\theta_k + \frac{1}{\sigma_{max}^4}sin^2\theta_k} sin^2\theta_k + O(\epsilon) \quad (10)$$

$$= \frac{\sigma^4}{\sigma'^2} \frac{\sigma_{max}^4 - \sigma^2\sigma'^2}{\sigma_{max}^4 - \sin^2\theta_k(\sigma_{max}^4 - \sigma^4)} \sin^2\theta_k + O(\epsilon) \quad (11)$$

The proof follows by using the classical inequality [10] $\sin^2\theta_k \leq \frac{\nu_k^2 - \sigma^2}{\sigma'^2 - \sigma^2}$. □

It is important to stress that, unlike theorem 2.1, theorem 4.1 provides estimates on the singular values within one iteration. These estimates are often pessimistic and ν_k may converge towards σ much better than this theorem predicts. However, they are acceptable in the important case where the smallest singular value σ is small compared to the others. Consider for example a matrix in which $\sigma = 10^{-1}\ \sigma'$, then $\frac{\sigma_{max}^4 - \sigma^2\sigma'^2}{\sigma_{max}^4 - \sin^2\theta_k(\sigma_{max}^4 - \sigma^4)}$ is bounded by $1 + tg^2\theta_k$ and is expected to settle down around 1 while $\frac{\sigma^4}{\sigma'^2(\sigma'^2 - \sigma^2)} \approx 10^{-4}$. The choice of the preconditioning matrix M plays a crucial role for the success of the method. We have seen that the better the approximation to $(A^TA)^{-1}$, the closer the vector s_k to x. The extreme case happens when $M = (A^TA)^{-1}$. This case, although usually impossible in practice, may be used to enlighten the theory.

Proposition 4.1 *Assume that $M = (A^TA)^{-1}$, then the columns of V_k constructed by Algorithm 2 form an orthonormal basis of the Krylov subspace* $K_k((A^TA)^{-1}, V_1) = Span\{V_1, (A^TA)^{-1}V_1, \ldots, (A^TA)^{-k+1}V_1\}$ *where V_1 denotes the rectangular matrix used at each restart.*

Proof Follows from the fact that $V_k = MGS(V_{k-1}, t_{k-1,1}, \ldots, t_{k-1,l})$ and in this case for $i = 1, \ldots, l$, $t_{k-1,i} = (A^TA)^{-1}r_{k-1,i} = x_{k-1,i} - \sigma_{k-1,i}^2(A^TA)^{-1}x_{k-1,i} = V_{k-1,i}y_{k-1,i} - \sigma_{k-1,i}^2(A^TA)^{-1}V_{k-1,i}y_{k-1,i}$ □

We can thus, in this case, derive optimal bounds for Davidson's method. The smallest singular value of the matrix H_k satisfies

$$\nu_k^2 = \min_{v \in K} \frac{v^TA^TAv}{v^Tv} \qquad \text{where} \quad K = K_k\left((A^TA)^{-1}, V_1\right).$$

Let $w = (A^TA)^{\frac{1}{2}}v$ then

$$\frac{1}{\nu_k^2} = \max_{v \in K1} \frac{w^T(A^TA)^{-1}w}{w^Tw} \quad \text{where} \quad K1 = K_k\left((A^TA)^{-1}, (A^TA)^{\frac{1}{2}}V_1\right).$$

This means that Davidson's method applied to A^TA for computing the smallest singular values amounts, in this case, to Lanczos method applied to $(A^TA)^{-1}$ for computing the largest singular values. Unlike Lanczos method, Davidson's method does not require the explicit use of $(A^TA)^{-1}$, one can be happy with any good approximation to $(A^TA)^{-1}$ in step 7 of Algorithm 2. The disadvantages over Lanczos method are the cost and the storage which are high and the fact that the projected matrix H_k is not band as was the matrix T_k in (1). A straightforward application of theorem 2.1, bound (3), leads us to the following theorem

Theorem 4.2 *Assume that $M = (A^TA)^{-1}$, and that k steps of Algorithm 2 have been carried out, then*

$$0 \leq \frac{\frac{1}{\sigma^2} - \frac{1}{\nu_k^2}}{\frac{1}{\sigma^2} - \frac{1}{\sigma_{max}^2}} \leq \left[\frac{tan(\phi_l)}{T_{k-1}\left(\frac{1+\mu}{1-\mu}\right)} \right]^2 \tag{12}$$

where $\mu = \frac{\frac{1}{\sigma^2} - \frac{1}{\sigma_{l+1}^2}}{\frac{1}{\sigma^2} - \frac{1}{\sigma_{max}^2}}$ and ϕ_l denotes the principle angle between $span\{(A^TA)^{\frac{1}{2}}V_1\}$ and the invariant subspace associated with the l smallest singular values of A and T_{k-1} is the Chebyshev polynomial of order $k-1$.

Let us consider the favourable singular value distribution in which $\sigma = 10^{-1}\sigma_{l+1}$ and $\sigma = 10^{-2}\sigma_{max}$, then, after 10 steps of Algorithm 2 and if nothing goes wrong, (12) will be approximated by $1 - \left(\frac{\sigma}{\nu_k}\right)^2 \leq \left(tan(\phi_l) \times 7.2949 \times 10^{-24}\right)^2$.

5 Numerical experiments

Our main concern in this section is to illustrate the behaviour of Davidson's method for computing the smallest singular values, and to give comparisons with Lanczos' method. We choose a set of realistic test matrices coming all but one, from the Harwell-Boeing set of sparse matrices [6]. The experiments have been performed on an IBM/RISC 6000-550 using double precision.
Before considering our test examples, let us recall that steps 1 to 5 in Algorithm 2 are nothing but the Rayleigh-Ritz procedure [10] applied to A^TA where only the last columns of U_k, W_k and H_k are computed at each iteration. We point out that only the matrix A is stored and that we access the elements of A^T using the data structure of A. The algorithm involves

intensive use of matrix-matrix operations (BLAS 3 level) and the desired eigenpairs of H_k, step 4 in Algorithm 2, are computed with EISPACK thus the portability of the algorithm is preserved on a wide class of computers. In order to keep the complexity of the algorithm at a reasonable level and since some of the l singular values may converge before others, we choose to restart the algorithm not only when the maximum size of the basis is reached, but also whenever a singular vector $x_{k,i}$ is converged. In this latter case we put the converged singular vector at the beginning of the basis so that all the vectors are orthogonalized against it. We then restart with a reduced block size. This may be considered as a cheap deflation technique.

Concerning the preconditioner $C_{k,i}$, we choose first the most commonly used one $C_{k,i} = (D - \nu_{k,i}^2 I)^{-1}$ where D is the diagonal of $A^T A$, we call DAVID this version of Davidson's method. The second choice is of the form $C_{k,i} = (M^T M)^{-1}$ where M is an incomplete LU factorization of A, the so called $ILUTH$ which consists in removing any entry of A which is less that some prespecified drop tolerance. We call DAVIDLU this version. The reason of this choice is essentially due to its simplicity and the reasonable good results derived from it. More interesting preconditioners based on an incomplete factorization of $A^T A$ rather than A (incomplete Cholesky of $A^T A$ or incomplete QR factorization of A etc...) should worthwhile and will be investigated later.

The Lanczos method we choose to compare with is a sparse SVD via a hybrid block Lanczos procedure for eigensystems of the form $A^T A$ called BLSVD, developed by Berry [2] and available from netlib. BLSVD is normally designed to approximate the largest singular values and only a small modification, indicated by the author, is made for obtaining the smallest ones.

Both algorithms (Davidson and Lanczos) needs initial starting vectors, a maximum size m_{max} for the basis, an initial block size *nblock* and the number of desired singular values *nvalues*. In order to make a fair comparison between the different methods, we use the same above parameters.
For each test example we give a table summarizing the obtained results. We list in these tables the l singular values as computed by EISPACK and those computed by BLSVD, DAVID and DAVIDLU. The numbers in parentheses indicate the residual norm $\| A^T A\tilde{x} - \tilde{\sigma}^2\tilde{x} \|_2$, where $\tilde{\sigma}$ and $\tilde{x}$ ($\| \tilde{x} \|_2 = 1$) denote the computed singular value and the corresponding right singular vector. The quantity mat-vec is equal to the number of multiplications by A plus the number of multiplications by A^T. The execution time is also reported. The parameter tol is the drop tolerance for ILUTH preconditioning

and fill-in is the fill-in produced during the incomplete factorization of A. As a matter of fact we have tested different values of the parameter tol, but we report only the one which maintains the fill-in in the same order as the number of nonzero elements of A.

For the three methods, we set $itmax = 150$ as an upper bound on the number of outer iterations. The algorithms terminate when $itmax$ is exceeded.

Matrix ADI. This matrix comes from information retrieval and seismic tomography applications [2] and is available from netlib. It is rectangular with 374 rows and 82 columns and 1343 nonzero elements. This example does not present any special difficulty, we only use it for comparison purposes. We compare BLSVD and DAVID for computing the 8 smallest singular values and the corresponding right singular vectors. For the three methods we used $m_{max} = 40$, $nblock = nvalues = 8$ and a stopping criterion such that the residual norm is less than 10^{-6}. The results are listed in Table 1. Both methods perform well, especially BLSVD where the precision in the obtained residuals is higher than required. This is because Lanczos' method checks the stopping criterion only periodically and hence convergence may be obtained before the iterations terminate.

singular values EISPACK	singular values(Residuals) BLSVD	singular values(Residuals) DAVID
1.8842753757029	1.8842753757029 (3.94E-07)	1.8842753757029(6.21E-7)
2.1044515491758	2.1044515491757 (2.74E-07)	2.1044515491757(8.56E-7)
2.2646138010928	2.2646138010928 (6.27E-08)	2.2646138010927(4.45E-7)
2.4128073672232	2.4128073672232 (1.52E-07)	2.4128073672232(9.06E-7)
2.6076737719950	2.6076737719950 (1.77E-08)	2.6076737719950(8.985E-7)
2.6589475772586	2.6589475772586 (9.93E-08)	2.6589475772586(9.84E-7)
2.6631702788999	2.6631702788999 (3.89E-09)	2.6631702788999(8.56E-7)
2.7492829533297	2.7492829533297 (1.11E-08)	2.7492829533297(5.60E-7)
	mat- vec:1688	mat-vec:1234
	Time(sec): 1.16	Time(sec): 1.15

Table 1. Computation of the 8 smallest singular pairs
Matrix ADI

Matrix PORES3. This matrix comes from the Harwell-Boeing set of test matrices. It arises from reservoir simulation. It is square of order 532 and has 3474 nonzero elements. The smallest singular values are not small in magnitude, but they are small in comparison with the largest ones. For example $\sigma_{min} = 0.26733863526883$ and $\sigma_{max} = 149922.80169575$ which means, for the three methods that $\sigma_{min}^2 = 3.12\ 10^{-12}\sigma_{max}^2$. Here again we used $m_{max} = 40$, $nblock = nvalues = 8$ and a the stopping criterion such that the residual norm is less than 10^{-6}. Neither BLSVD nor DAVID were capable of computing the 8 wanted singular values. In Table 2 we list the results obtained by DAVIDLU.

singular values EISPACK	singular values(Residuals) DAVIDLU
0.2673386352688	0.2673386353738 (2.99E-7)
0.3189119825820	0.3189119826714 (9.92E-7)
1.0274629898236	1.0274629898151 (9.79E-7)
1.1551923630159	1.1551923630113 (9.77E-7)
2.6543014616552	2.6543014616463 (7.75E-7)
3.1991610201392	3.1991610201750 (6.01E-7)
4.2009362777080	4.2009362777098 (7.50E-7)
5.0281947778858	5.0281947778886 (8.48E-7)
	mat- vec:1186 Time(sec): 4.13
	fill-in 3474 $\rightarrow$7580 tol $=10^{-1}$

Table 2. Computation of the 8 smallest singular pairs
Matrix PORES3

Matrix SHERMAN1. This matrix comes also from the Harwell-Boeing set of test matrices and arises from the three dimensional simulation of black oil. It is symmetric but we can treat it as an unsymmetric matrix. The order is 1000 with 3750 nonzero elements. It has 1 singular value of order 10^{-4}, 47 of order 10^{-3}, 206 of order 10^{-2}, 251 of order 10^{-1}, 315 equal to 1 and the last 180 singular values lie between 1.1325183438792 and 5.0448693671654. We used the same parameters m_{max}, $nblock$, $nvalues$ and the stopping criterion as in the previous examples. Here again neither BLSVD nor DAVID were capable of computing the 8 wanted singular values. In Table 3 we list the results obtained by DAVIDLU.

singular values EISPACK	singular values(Residuals) DAVIDLU
3.234871204E-4	3.234883881E-4 (7.60E-7)
1.0178313925E-3	1.0178315577E-3 (6.01E-7)
1.1131469008E-3	1.1131470276E-3 (5.51E-7)
1.5108868655E-3	1.5108869337E-3 (2.89E-7)
1.9207831764E-3	1.9207832211E-3 (1.91E-7)
2.0521929282E-3	2.0521932271E-3 (5.41E-7)
2.0928681061E-3	2.0928683807E-3 (2.74E-7)
2.1038958599E-3	2.1038962048E-3 (6.92E-7)
	mat- vec:270 Time(sec): 1.15
	fill-in:3750→ 8304 tol = 10^{-3}

Table 3. Computation of the 8 smallest singular pairs
Matrix SHERMAN1

Matrix HOR131. This matrix also comes from the Harwell-Boeing set of test matrices. It arises in the flow network problem. It is square of order 434 with 4710 nonzero elements and has 17 singular values of order 10^{-5}, 43 of order 10^{-4}, 169 of order 10^{-3}, 173 of order 10^{-2} and 32 of order 10^{-1}. With $m_{max} = 80$, $nblock = nvalues = 20$ and a stopping criterion such that the residual norm is less than 10^{-8}. Table 4 reports the results given by DAVIDLU. After 148 outer iterations and 23794 matrix-vector multiplications 1.7924031832664E-5 and 2.4045394072133E-5 where declared good singular values with corresponding residuals 9.53E-9 and 9.50E-9 by DAVID. A look on Table 4 reveals that the two numbers 1.7924031832664E-5 and 2.4045394072133E-5 lie respectively between the smallest and the second and between the fourth and the fifth singular values of A. BLSVD did not converge.

In the conclusion, we can say that the numerical results confirm that Lanczos and the standard Davidson method are, in general, not suitable for computing the smallest singular values. The results given by the modification we introduced in Davidson's method are effective and by far superior to those obtained by the two previous methods. There are, however, some still problems regarding the choice of the preconditioner. The ideal would be to find a reliable approximation to A^TA using the data structure of A. We will investigate this later.

singular values EISPACK	singular values(Residuals) DAVIDLU
1.5338130942225D-05	1.53384826275538998E-5(7.41E-9)
2.1363547226369D-05	2.1363966304760514 1E-5 (8.35E-9)
2.2736655865418D-05	2.2737460581733088E-5 (9.42E-9)
2.3557411304847D-05	2.3557659815166145 7E-5 (7.88E-9)
2.5659788515646D-05	2.5660061127580043 4E-5 (7.29E-9)
2.7036649455056D-05	2.7036779980413491 7E-5 (4.03E-9)
3.0659915472795D-05	3.0660099540738222 4E-5(9.40E-9)
3.5933337181559D-05	3.5933514556299949 4E-5(6.10E-9)
4.0190136698560D-05	4.0191181999565675 2E-5(9.84E-9)
4.1630845269902D-05	4.1631003850727355 9E-5 (5.39E-9)
4.6249600876053D-05	4.6249846192020361 4E-5 (7.59E-9)
5.1509531011866D-05	5.1509766107810183 8E-5 (8.83E-9)
5.5485971489133D-05	5.5486255391052127 9E-5(9.75E-9)
5.8906973927180D-05	5.8907403736813753 3E-5 (6.58E-9)
6.3030823663830D-05	6.3031121301669528 3E-5(6.18E-9)
7.1026655095114D-05	7.1026688160043632 3E-5(3.94E-9)
8.5762312230985D-05	8.5762495406863003 5E-5(9.96E-9)
1.2385189669020D-04	1.2385197725459004 0E-4(7.83E-9)
1.2549423295339D-04	1.2549436976595724 8E-4(8.93E-9)
1.5598303222124D-04	1.5598402416672636 2E-4(8.31E-9)
	mat-vec:1110 Time(sec):4.31
	fill-in: 4710 → 7932 tol=10^{-4}

Table 4. Computation of the 20 smallest singular pairs
Matrix HOR131

References

[1] M. Berry and G. Golub, *Estimating the largest singular values of large sparse matrices via modified moments*, Numerical Algorithms, Vol. 1 (1991), pp. 353–374.

[2] M. W. Berry, *Multiprocessor sparse SVD algorithms and applications*, PhD thesis, CSRD- University of Illinois, November 1990.

[3] M. Crouzeix, B. Philippe, and M. Sadkane, *The Davidson method*, Tech. Rep., INRIA Report No. 1353. 1990. To appear in SIAM J. Sci. Stat. Comput.

[4] J. K. Cullum and R. A. Willoughby, *Lanczos algorithm for large symmetric eigenvalue computations. Volume 1 Theory*, Birkhauser,, Boston, 1985.

[5] 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.

[6] I. S. Duff, R. G. Grimes, and J. G. Lewis, *Spase Matrix Test Problems*, ACM Trans. Math. Softw., (1989), pp. 1–14.

[7] G. H. Golub, F. T. Luk, and M. L. Overton, *A block Lanczos method for computing the singular values and corresponding singular vectors of a matrix*, ACM Trans. on Math. Soft., Vol. 7 (1981), pp. 149–169.

[8] S. V. Huffel, *Iterative algorithms for computing the singular subspace of a matrix associated with its smallest singular values*, Lin. Alg. and its Appl., Vol. 154-156 (1980), pp. 675–709.

[9] 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.

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

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

[12] Y. Saad, *On the rates of convergence of the Lanczos and block-Lanczos methods*, SIAM J. Numer. Anal., Vol. 17 (1980), pp. 687–706.

[13] R. R. Underwood, *An iterative block Lanczos method for the solution of large sparse symmetric eigenproblems*, PhD thesis, Stanford University, May 1975.

Unité de recherche INRIA Lorraine, Technôpole de Nancy-Brabois, Campus scientifique,
615 rue de Jardin Botanique, BP 101, 54600 VILLERS LÈS NANCY
Unité de recherche INRIA Rennes, IRISA, Campus universitaire de Beaulieu, 35042 RENNES Cedex
Unité de recherche INRIA Rhône-Alpes, 46 avenue Félix Viallet, 38031 GRENOBLE Cedex 1
Unité de recherche INRIA Rocquencourt, Domaine de Voluceau, Rocquencourt, BP 105,
78153 LE CHESNAY Cedex
Unité de recherche INRIA Sophia-Antipolis, 2004 route des Lucioles, BP 93, 06902 SOPHIA-ANTIPOLIS
Cedex

Éditeur
INRIA, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex
(France)
ISSN 0249-6399