A minimization method for the solution of large symmetric eigenproblems

E. Galligani, V. Ruggiero, L. Zanni
1Department of Mathematics, University of Modena, Modena, Italy
2Department of Mathematics, University of Ferrara, Ferrara, Italy

Abstract. This paper concerns with the solution of a special eigenvalue problem for a large sparse symmetric matrix by a fast convergent minimization method. A theoretical analysis of the method is developed; it is proved that is convergent with a convergence rate of fourth order. This minimization method requires to solve a sequence of equality-constrained least squares problems that become increasingly ill-conditioned, as the solution of eigenvalue problem is approached. A particular attention has been addressed to this question of ill-conditioning for the practical application of the method.

Computational experiments carried out on Cray C90 show the behaviour of this minimization method as accelerating technique of the inverse iteration method. Also a comparison with the scaled Newton method has been done.

Keywords: special eigenvalue problem, symmetric matrix, minimization method, LSE problem, inverse iteration, parallel computing.


1 Introduction

Minimization methods are not uncommon to recent efforts for the numerical solution of the “special” eigenvalue problem. In many branches of science and technology, it is often required to evaluate some extreme eigenvalues and corresponding eigenvectors (special eigenvalue problem) of the generalized eigenproblem:

\[ A x = \lambda B x \]  

where \( A \) and \( B \) are \( n \times n \) real symmetric matrices and \( B \) is positive definite.

When \( A \) and \( B \) are large and sparse, it is not convenient to use similarity transformations or to factorize either \( A \) or \( B \) or a linear combination of \( A \) and \( B \) into a product of simple matrices. On the other hand, the large size of the problem suggests to recognize methods that are well suited for vector and parallel processing, such as, for example, many of the minimization methods developed in literature (e.g. [3], [11], [21], [24], [26]). When \( A \) and \( B \) are symmetric positive definite matrices, the minimization of Rayleigh quotient (RQM) by preconditioned conjugate gradient (PCG) method is an effective means for solving (1) in the case of very ill-conditioned problems. This is shown by the numerical results of a vast experimentation carried out on Cray Y–MP ([22]), where we solved by RQM–PCG problems arising in the discretization by finite difference method of the diffusion equation with piecewise continuous coefficients and in the finite integration of flow and structural mechanics equations ([20]). For this class of problems,
the choice of a “good” preconditioner $M$ is crucial for a fast convergence of PCG method. Widely used preconditioners are the incomplete Cholesky factorization ([1]), the generalized SSOR polynomial preconditioner derived from the SSOR preconditioner ([7]) and the generalized additive polynomial preconditioner derived from the arithmetic mean preconditioner ([8]) (see also [16]). From the computational experiments in [22], a definite advantage for the additive polynomial preconditioner with respect to the SSOR polynomial preconditioner can be deduced; furthermore the additive polynomial preconditioner appears more efficient with respect to the incomplete Cholesky factorization preconditioner when the matrices $A$ and $B$ arise from the finite difference method with multicolour ordering of the mesh–points. With regard to RQM–PCG method, recently in [24] it has been suggested to replace the computation of $p = M^{-1}g$ in the PCG scheme with few steps of the conjugate gradient method applied to $Ap = g$. As “emergency exit” a super–quadratic convergent projected Newton method ([19]) gives the necessary security and robustness properties. This paper concerns with the solution of the special eigenvalue problem (with $B = I$) for a large sparse symmetric matrix $A$ by a fast convergent minimization method. In the next section we present a theoretical analysis of the method, showing that it is convergent with a convergence rate of fourth order. This minimization method requires to solve a sequence of equality–constrained least squares problems that become increasingly ill–conditioned, as the solution of eigenvalue problem is approached. A particular attention has been addressed to this question of ill–conditioning for the practical application of the method. In the last section computational experiments carried out on Cray C90 show the effectiveness of the minimization method when it is used as accelerating technique of the inverse iteration method for middle–ill conditioned symmetric eigenproblems. Also a comparison with the scaled Newton method ([10]) has been done.

2 A fast convergent minimization method

The problem of finding an eigenpair $(\rho^*, x^*)$ of the symmetric matrix $A$ can be formulated as the following constrained least squares problem:

$$\min_{(x, \rho)} \phi(x, \rho) \equiv \| (A - \rho I)x \|^2$$

subject to $d^T x - 1 = 0$

where $\|\cdot\|$ denotes the Euclidean norm and $d$ is a non null vector, not orthogonal to the eigenspace of $A$ corresponding to $\rho^*$. This problem can be solved with the following iterative scheme, called Minimization Method (MM):

1. Let $x^{(0)}$ be a starting $n$–vector, such that $d^T x^{(0)} = 1$.
2. For $k = 0, 1, 2, ...$ until convergence
   - compute the solution $\rho_k$ of the problem
     $$\min_{\rho} \phi(x^{(k)}, \rho) \equiv \| (A - \rho I)x^{(k)}\|^2$$

$$\text{(2)}$$
Let denote by $x^{(k+1)}$ the solution of the least squares problem with one equality constraint (LSE problem)
\[
\min_x \phi(x, \rho_k) \equiv \| (A - \rho_k I)x \|^2
\]
subject to $d^T x - 1 = 0$

(3)

It is immediate that, for any starting vector $x^{(0)}$, the MM generates two sequences $\{\rho_k\}$ and $\{x^{(k)}\}$ univocally determined. Indeed the solution $\rho_k$ of the problem (2) is the Rayleigh quotient of $x^{(k)}$:
\[
\rho_k = \frac{x^{(k)^T} A x^{(k)}}{x^{(k)^T} x^{(k)}}
\]

Furthermore, the Lagrange optimality conditions for the problem (3) are:
\[
2(A - \rho_k I)^2 x + \mu d = 0
\]
\[
d^T x - 1 = 0
\]

(4)

where $\mu$ is the Lagrange multiplier. If $\rho_k$ is not an eigenvalue of $A$, we can solve the first equation of (4) for $x = x^{(k+1)}$:
\[
x^{(k+1)} = -\frac{\mu}{2} (A - \rho_k I)^{-1} (A - \rho_k I)^{-1} d
\]

(5)

and, substituting (5) in the equality constraint, we have
\[
\mu_{k+1} = -\frac{2}{\| (A - \rho_k I)^{-1} d \|^2} \sum_{i=1}^{n} \frac{\gamma_i u_i}{(\lambda_i - \rho)^2}
\]
\[
x^{(k+1)} = \frac{(A - \rho_k I)^{-1} (A - \rho_k I)^{-1} d}{\| (A - \rho_k I)^{-1} d \|^2}
\]

(6)

Let denote by $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$ the $n$ real eigenvalues of $A$ and by $u_1, u_2, \ldots, u_n$ the corresponding orthonormal eigenvectors.

If we write $d$ as linear combination of the orthonormal eigenvectors of $A$
\[
d = \sum_{i=1}^{n} \gamma_i u_i
\]

the $n$–vector $x^{(k+1)}$ in (6) is a particular value of the following function $x(\rho)$, defined for $\rho \neq \lambda_i$ ($i = 1, ..., n$):
\[
x(\rho) = \frac{\sum_{i=1}^{n} \frac{\gamma_i u_i}{(\lambda_i - \rho)^2}}{\sum_{i=1}^{n} \frac{\gamma_i^2}{(\lambda_i - \rho)^2}}
\]

(7)

The function $x(\rho)$ can be extended for $\rho = \lambda_r$, for any $r$, to be continuous on $R$:

\[
x(\lambda_r) = \lim_{\rho \to \lambda_r} x(\rho) = \begin{cases} 
\lim_{\rho \to \lambda_r} \frac{\sum_{i=1}^{n} \frac{\gamma_i u_i}{(\lambda_i - \rho)^2}}{\sum_{i=1}^{n} \frac{\gamma_i^2}{(\lambda_i - \rho)^2}} & \text{if } d \text{ is orthogonal to the eigenspace corresponding to } \lambda_r \\
\lim_{\rho \to \lambda_r} \frac{\sum_{i=1}^{n} \gamma_i u_i + (\lambda_r - \rho)^2 \sum_{i=1}^{n} \gamma_i u_i}{\sum_{i=1}^{n} \gamma_i^2 + (\lambda_r - \rho)^2 \sum_{i=1}^{n} \gamma_i^2} & \text{if there exists at least one } \gamma_i \neq 0 \text{ corresponding to } \lambda_i = \lambda_r
\end{cases}
\]

The expression $\sum_{i=1}^{n} \gamma_i u_i$ indicates the sum over the indices of the $s$ eigenvectors $u_{i_1}, u_{i_2}, \ldots, u_{i_s}$ associated to $\lambda_r$ of multiplicity $s$, and $\sum_{i=1}^{n} \gamma_i u_i$ denotes the sum over the indices of the remaining $(n - s)$ eigenvectors.
Then, if the sequence \( \{\rho_k\} \) generated by the MM converges to an eigenvalue \( \rho^* \) of \( A \), as \( k \to \infty \), from the continuity of \( x(\rho) \), we can conclude that the sequence \( \{x(\rho_k)\} \) (or \( \{x^{(c+1)}\} \)) converges, as \( k \to \infty \), to the vector \( x(\rho^*) \). Besides if \( d \) is not orthogonal to the eigenspace of \( A \) corresponding to \( \rho^* \), \( x^* = x(\rho^*) \) is an eigenvector of \( A \) corresponding to \( \rho^* \):

\[
x^* = x(\rho^*) = \lim_{k \to \infty} x(\rho_k) = \frac{\sum_{\lambda_i = \rho^*} \gamma_i u_i}{\sum_{\lambda_i = \rho^*} \gamma_i^2}
\]

Under the same hypotheses (\( \{\rho_k\} \) converges to an eigenvalue \( \rho^* \) as \( k \to \infty \) and there exists at least one \( \gamma_i \neq 0 \) with \( \lambda_i = \rho^* \)), it is easy to verify that the convergence rate of the sequences \( \{\rho_k\} \) and \( \{x(\rho_k)\} \) is of the fourth order. Indeed, from (7) we can write (6) as

\[
x(\rho_k) = \frac{\sum_{\lambda_i = \rho^*} \gamma_i u_i + (\rho^* - \rho_k)^2 \sum_{\lambda_i \neq \rho^*} \frac{\gamma_i u_i}{(\lambda_i - \rho_k)^2}}{\sum_{\lambda_i = \rho^*} \gamma_i^2 + (\rho^* - \rho_k)^2 \sum_{\lambda_i \neq \rho^*} \frac{\gamma_i^2}{(\lambda_i - \rho_k)^2}}
\]

Since \( d^T x(\rho_k) = 1 \), \( x(\rho_k) \neq 0 \) for any \( k \); then \( \rho_{k+1} \) is well-defined and from (9) we obtain

\[
\rho_{k+1} = \frac{x(\rho_k)^T A x(\rho_k)}{x(\rho_k)^T x(\rho_k)} = \frac{\rho^* \sum_{\lambda_i = \rho^*} \gamma_i^2 + (\rho^* - \rho_k)^4 \sum_{\lambda_i \neq \rho^*} \frac{\gamma_i^2 \lambda_i}{(\lambda_i - \rho_k)^2}}{\sum_{\lambda_i = \rho^*} \gamma_i^2 + (\rho^* - \rho_k)^4 \sum_{\lambda_i \neq \rho^*} \frac{\gamma_i^2}{(\lambda_i - \rho_k)^2}}
\]

and, consequently,

\[
\rho_{k+1} - \rho^* = (\rho^* - \rho_k)^4 \frac{\sum_{\lambda_i \neq \rho^*} \frac{\gamma_i^2 (\lambda_i - \rho^*)}{(\lambda_i - \rho_k)^2}}{\sum_{\lambda_i = \rho^*} \gamma_i^2 + (\rho^* - \rho_k)^4 \sum_{\lambda_i \neq \rho^*} \frac{\gamma_i^2}{(\lambda_i - \rho_k)^2}}
\]

Then (if we exclude the trivial case of \( d \) eigenvector of \( A \) corresponding to \( \rho^* \), i.e. \( \gamma_i = 0 \) for any \( i \) such that \( \lambda_i \neq \rho^* \)), the following limit exists and it is equal to a positive constant \( \theta^* \):

\[
\lim_{k \to \infty} \frac{\rho_{k+1} - \rho^*}{|\rho_k - \rho^*|^4} = 0 \quad \frac{1}{\sum_{\lambda_i = \rho^*} \gamma_i^2} \left| \sum_{\lambda_i \neq \rho^*} \frac{\gamma_i^2}{(\lambda_i - \rho^*)^2} \right| = \theta^*
\]

Furthermore, if we write \( x(\rho_k) \) of (9) as

\[
x(\rho_k) = \frac{x^* + (\rho^* - \rho_k)^2 w^{(k)}}{1 + (\rho^* - \rho_k)^2 c_k}
\]

with \( w^{(k)} = \frac{\sum_{\lambda_i \neq \rho^*} \frac{\gamma_i u_i}{(\lambda_i - \rho_k)^2}}{\sum_{\lambda_i = \rho^*} \gamma_i^2} \) and \( c_k = \frac{\sum_{\lambda_i \neq \rho^*} \frac{\gamma_i^2}{(\lambda_i - \rho_k)^2}}{\sum_{\lambda_i = \rho^*} \gamma_i^2} \), we have

\[
x(\rho_k) - x^* = \frac{(\rho^* - \rho_k)^2 (w^{(k)} - c_k x^*)}{1 + (\rho^* - \rho_k)^2 c_k}
\]
We observe that \( w^{(k)} - c_k x^* \neq 0 \) for any \( k \). In fact, if \( c_k = 0 \), we have \( \gamma_i = 0 \) for any \( i \) such that \( \lambda_i \neq \rho^* \); then \( w^{(k)} = 0 \) and \( x(\rho_k) = x^* \) for any \( k \). This is the trivial case of \( \text{d} \) eigenvector of \( A \) corresponding to \( \rho^* \). If \( c_k \neq 0 \), we denote \( w = \lim_{k \to \infty} w^{(k)} \) and \( c = \lim_{k \to \infty} c_k \); since \( w^{(k)} \) is orthogonal to \( x^* \) for any \( k \), we have \( w^{(k)} - c_k x^* \neq 0 \) and \( w - cx^* \neq 0 \).

Thus, the following limit exists and it is equal to a positive constant:

\[
\lim_{k \to \infty} \frac{\|x(\rho_{k+1}) - x^*\|}{\|x(\rho_k) - x^*\|^4} = \lim_{k \to \infty} \left( \frac{|\rho_{k+1} - \rho^*|}{|\rho_k - \rho^*|^4} \right)^2 \frac{\|w^{(k+1)} - c_{k+1} x^*\|}{\|w^{(k)} - c_k x^*\|^4} \frac{|1 + (\rho^* - \rho_k)^2 c_k|^4}{|1 + (\rho^* - \rho_{k+1})^2 c_{k+1}|}
\]

\[= \frac{\vartheta^2}{\|w - cx^*\|^3}\]

(11)

For the conclusions (8), (10) and (11), it is essential to prove that the sequence \{\( \rho_k \)\} generated by the MM converges to an eigenvalue \( \rho^* \) of \( A \) as \( k \to \infty \). This assertion is the main result of the following theorem, summarizing a set of propositions stated by Di Lena, Peluso, Piazza in [5].

**Theorem.** Let \( A \) be a symmetric matrix of order \( n \). Given a non null vector \( \text{d} \neq 0 \) and a non null vector \( x^{(0)} \) such that \( \text{d}^T x^{(0)} = 1 \), let \{\( \rho_k \)\} and \{\( x(\rho_k) \)\} be the sequences generated by the MM. If \( \text{d} \) is not orthogonal to the eigenspace of \( A \) corresponding to an eigenvalue \( \rho^* \) and if \( \rho_k \) is not an eigenvalue of \( A \) for any \( k \), then the sequences \{\( \rho_k \)\} and \{\( x(\rho_k) \)\} converge as \( k \to \infty \) to \( \rho^* \) and to a corresponding eigenvector \( x^* \) respectively.

The proof that \{\( \rho_k \)\} is a convergent sequence can be deduced by the following lemma ([15], see also [6], [4]):

**Lemma.** Let \( \Psi \) be a continuous function on \( R \). For a given \( \rho_0 \in R \), let \{\( \rho_k \)\} be a sequence generated by \( \rho_{k+1} = \Psi(\rho_k) \). If the following condition holds

\[\Psi(\Psi(\rho)) = \rho \implies \Psi(\rho) = \rho\]

(12)

then any bounded sequence \{\( \rho_k \)\} is convergent as \( k \to \infty \).

In the case of MM, \( \Psi(\rho) = \frac{x(\rho)^T A x(\rho)}{x(\rho)^T x(\rho)} \); \( \Psi(\rho) \) is a continuous function on \( R \) and the sequence \{\( \rho_k \)\} is bounded (\( |\rho_k| \leq \|A\| \) for any \( k \)).

In order to verify (12), we observe that, since \( \rho_k \) is the solution of (2) for \( x^{(k)} = x(\rho_{k-1}) \) given, we have

\[\phi(x(\rho_{k-1}), \rho_{k-1}) \geq \phi(x(\rho_{k-1}), \rho_k)\]

(13)

Analogously for \( \rho_{k+1} \)

\[\phi(x(\rho_k), \rho_k) \geq \phi(x(\rho_k), \rho_{k+1})\]

(14)

Since \( x(\rho_k) \) is solution of (3) for \( \rho_k \) given, we have

\[\phi(x(\rho_{k-1}), \rho_k) \geq \phi(x(\rho_k), \rho_k)\]

(15)

Analogously for \( x(\rho_{k+1}) \)

\[\phi(x(\rho_k), \rho_{k+1}) \geq \phi(x(\rho_{k-1}), \rho_{k-1})\]

(16)

If we impose the condition \( \rho_{k+1} = \rho_{k-1} \), the relations (13)–(16) give

\[\phi(x(\rho_{k-1}), \rho_{k-1}) \geq \phi(x(\rho_{k+1}), \rho_{k}) \geq \phi(x(\rho_k), \rho_k) \geq \phi(x(\rho_{k-1}), \rho_{k-1})\]

(17)
By (17), we obtain
\[ \phi(x(\rho_{k-1}), \rho_{k-1}) = \phi(x(\rho_k), \rho_k) \] (18)

Now,
\[ \phi(x(\rho_{k-1}), \rho_{k-1}) = \| (A - \rho_{k-1}I) x(\rho_{k-1}) \|^2 = \| (A - \rho_k I) x(\rho_{k-1}) + (\rho_k - \rho_{k-1}) x(\rho_{k-1}) \|^2 = \]
\[ = \| (A - \rho_k I) x(\rho_{k-1}) \|^2 + (\rho_k - \rho_{k-1})^2 \| x(\rho_{k-1}) \|^2 + 2(\rho_k - \rho_{k-1}) x(\rho_{k-1})^T (A - \rho_k I) x(\rho_{k-1}) = \]
\[ = \phi(x(\rho_{k-1}), \rho_k) + (\rho_k - \rho_{k-1})^2 \| x(\rho_{k-1}) \|^2 \]
Then, since \( \| x(\rho_{k-1}) \|^2 > 0 \), from (18) we obtain
\[ \rho_{k+1} = \rho_{k-1} \rightarrow \rho_k = \rho_{k-1} \]

Finally, theorem 3.5 in [5] shows that, for any fixed point \( \bar{\rho} \) of \( \Psi(\rho) \) such that \( (A - \bar{\rho} I)x(\bar{\rho}) \neq 0 \), \( \Psi'(\bar{\rho}) = 4 \). Then the sequences \{\( \rho_k \)\} and \{\( x(\rho_k) \)\} must converge as \( k \rightarrow \infty \) to an eigenpair \((\rho^*, x(\rho^*))\) of \( A \).

### 3 Practical application of the Minimization Method

An efficient implementation of the MM requires a “good” inner solver for the LSE problem (3). There exist two different approaches for obtaining a numerical approximation of the solution \( x^{(k+1)} \) of (3).

In the first approach, we compute the analytical expression (6) of \( x^{(k+1)} \) by solving for \( z \) the system
\[ (A - \rho_k I)^2 z = d \] (19)
Then, \( x^{(k+1)} = \frac{z}{d' z} \). The system (19) is positive definite, but it tends to become “nearly” positive semidefinite when \( \rho_k \) is close to \( \rho^* \). The solution of (19) can be obtained by the conjugate gradient (CG) method, implemented with slight modifications with respect to the version for normal equations. Because of the increasing condition number of (19), it is essential to use a preconditioner. The most simple preconditioner is a diagonal matrix \( D \) with diagonal entries equal to the euclidean norm of the columns of the matrix \( (A - \rho_k I)^2 \). Such preconditioner takes the effect of scaling the original problem.

An efficient stopping rule for the CG method is suggested by Ruhe and Wiberg in [23]; the rule permits to avoid unnecessary CG steps that could not give a smaller residual:
\[ \frac{\| r^{(l)} \|}{\| z^{(l)} \|} \leq \min(\varepsilon_0, \max(\gamma \| z^{(l)} \|^{-1}, \varepsilon_1)) \]
where \( z^{(l)} \) and \( r^{(l)} \) are the approximate solution of (19) and the residual at the \( l \)-step respectively, \( \varepsilon_0 \) and \( \varepsilon_1 \) are appropriate tolerance and \( \gamma > 1 \).

The numerical solution of (19) can be obtained also by the following two–steps procedure:
\[ (A - \rho_k I)y = d \]
\[ (A - \rho_k I)z = y \] (20)
At each step, we can use a solver for large sparse symmetric indefinite systems ([13], [17]). The second approach to obtain \( x^{(k+1)} \) consists in solving the large sparse LSE problem (3) by a numerical method. Also in this case it is convenient to scale the original problem by a diagonal matrix \( D \) of order \( n \) with diagonal entries equal to the Euclidean norm of the columns of the matrix \( (d^T A - \rho_k I) \).

The scaled LSE problem (3) has the following form

\[
\min_y \| (A - \rho_k I) D^{-1} y \|^2
\]

subject to \( d^T D^{-1} y - 1 = 0 \)

with \( y = Dx \).

It is well known that the LSE problem (3) is equivalent to solve the following augmented system:

\[
\begin{pmatrix}
0 & 0 & d^T \\
0 & -I & (A - \rho_k I) \\
d & (A - \rho_k I) & 0
\end{pmatrix}
\begin{pmatrix}
\mu \\
r \\
x
\end{pmatrix} =
\begin{pmatrix}
1 \\
0 \\
0
\end{pmatrix}
\] (21)

Again, for the solution of this sparse symmetric indefinite system see [13], [17].

Another way for solving (3) is the weighting method ([12]) which consists in solving a sequence of unconstrained least squares (LS) problems

\[
\min_x \left\| \left( \sqrt{\sigma^*} d^T \right) x - \left( \sqrt{\sigma^*} \right) \right\|^2
\]

(22)

The weighting method is equivalent to the penalty method ([14]) applied to (3) with a penalty parameter \( \sigma^* \). Then, when the positive weight \( \sqrt{\sigma^*} \to \infty \), the solution of (22) converges to the solution of the LSE problem.

In practice, we solve only one LS problem (22) with \( \sqrt{\sigma^*} \) sufficiently large (Lawson and Hanson suggest to use \( \sqrt{\sigma^*} = \text{macheps}^{-1} \)).

As solver of sparse problem (22), we can use the preconditioned CG method applied to normal equations or LSQR routine of Paige and Saunders ([18]).

An iterative scheme improving the solution of the weighting method has been proposed by Van Loan in [25]; see also [25] for the choice of a convenient weight and for the stopping criteria of the scheme and [2] for an error analysis.

We point out that the iterative refinement scheme of Van Loan is equivalent to the Hestenes method formulated for LSE problem. This is the subject of the next section.

### 4 The Hestenes method for LSE problem

The Hestenes method is devised for solving linearly equality constrained quadratic programming problems. An extensive analysis of this method is reported in [9]. In the case of the general LSE

\[
\min_x \frac{1}{2} \| Ax - b \|^2
\]

subject to \( Ex = f \)
where $A \in \mathbb{R}^{m \times n}$ ($m \geq n$), $b \in \mathbb{R}^m$, $E \in \mathbb{R}^{p \times n}$ ($p \leq n$) and $f \in \mathbb{R}^p$, with $E$ having full row–rank and with rank $\left( \begin{array}{c} A \\ E \end{array} \right) = n$, the method of Hestenes consists in generating the sequences $\{\mu^{(k)}\}$ and $\{x^{(k)}\}$, starting from an initial value $\mu^{(0)}$, by the following formulas $(k = 0, 1, 2, \ldots)$:

\begin{align}
(A^T A + \sigma_k E^T E)x^{(k)} &= A^T b + \sigma_k E^T f - E^T \mu^{(k)} \\
\mu^{(k+1)} &= \mu^{(k)} + \sigma_k(E x^{(k)} - f)
\end{align}

where $\{\sigma_k\}$ is a non decreasing sequence of positive numbers. For a convenient sequence $\{\sigma_k\}$, the sequence $\{x^{(k)}\}$ converges to the unique solution $x^*$ of (23) and $\{\mu^{(k)}\}$ converges to the Lagrange multiplier $\mu^*$ associated with $x^*$. Usually, in the implementation of the method, we select $\sigma_k = \sigma^*$ for all $k$, where $\sigma^*$ is a moderately large positive constant much greater than the value $\bar{\sigma} > 0$ for which $A^T A + \sigma E^T E$ is positive definite.

We observe that, if we choose $\mu^{(0)} = 0$, the formula (24) at the first step $k = 0$ of the Hestenes method, may be interpreted as the normal equations of the unconstrained least squares problem:

$$\min_{x} \frac{1}{2} \left\| \left( \begin{array}{c} \sqrt{\sigma^*} E \\ A \end{array} \right) x - \left( \begin{array}{c} \sqrt{\sigma^*} f \\ b \end{array} \right) \right\|^2$$

The solution $x^{(0)}$ of this last problem is equal to the solution of the first step of the weighting method.

Furthermore, for $\sigma_k = \sigma^*$ for any $k$, if we rewrite the formula (25) as follows:

$$\mu^{(k+1)} - \mu^{(k)} = \Delta \mu^{(k)} = -\sigma^*(f - E x^{(k)}) = -\sigma^* r^{(k)}$$

we observe that, at step $k + 1$,

$$(A^T A + \sigma^* E^T E)x^{(k+1)} = A^T b + \sigma^* E^T f - E^T \mu^{(k+1)}$$

or, equivalently,

$$(A^T A + \sigma^* E^T E)(x^{(k)} + \Delta x^{(k)}) = A^T b + \sigma^* E^T f - E^T \mu^{(k)} + \sigma^* E^T r^{(k)}$$

Then, we can obtain $\Delta x^{(k)}$ as the unique solution of the following system

$$(A^T A + \sigma^* E^T E)y = \sigma^* E^T r^{(k)}$$

This system may be interpreted as the normal equations of the LS problem:

$$\min_{y} \frac{1}{2} \left\| \left( \begin{array}{c} \sqrt{\sigma^*} E \\ A \end{array} \right) y - \left( \begin{array}{c} \sqrt{\sigma^*} r^{(k)} \\ 0 \end{array} \right) \right\|^2$$

In the case of problem (3) the Hestenes procedure can be stated as follows:

- Let $\mu^{(0)} = 0$; denote by $x^{(0)}$ the solution of

$$\min_{x} \frac{1}{2} \left\| \left( \begin{array}{c} \sqrt{\sigma^*} d^T \\ (A - \rho_k I) \end{array} \right) x - \left( \begin{array}{c} \sqrt{\sigma^*} d^T x^{(0)} \\ 0 \end{array} \right) \right\|^2$$

- Compute

$$r^{(0)} = 1 - d^T x^{(0)}$$

$$\Delta \mu^{(0)} = -\sigma^* r^{(0)}$$

$$\mu^{(1)} = \mu^{(0)} + \Delta \mu^{(0)}$$

8
• For \( k = 0, 1, \ldots \) until convergence
  
  Denote by \( \Delta x^{(k)} \) the solution of
  \[
  \min_y \frac{1}{2} \left\| \left( \begin{array}{c} \sqrt{\sigma^*} d^T \\ (A - \rho_k I) \end{array} \right) y - \left( \begin{array}{c} \sqrt{\sigma^*} r^{(k)} \\ 0 \end{array} \right) \right\|^2
  \]

  Compute
  \[
  \begin{align*}
  x^{(k+1)} &= x^{(k)} + \Delta x^{(k)} \\
  r^{(k+1)} &= 1 - d^T x^{(k+1)} \\
  \Delta \mu^{(k+1)} &= -\sigma^* r^{(k+1)} \\
  \mu^{(k+2)} &= \mu^{(k+1)} + \Delta \mu^{(k+1)}
  \end{align*}
  \]

This procedure is equal to the algorithm 4.2 of [25].

5 Computational experiments

The MM converges to an eigenpair \((\rho^*, x^*)\) that depends on the choice of the vectors \(d\) and \(x^{(0)}\). When we want to determine the eigenvalue of \(A\) having minimum absolute value (or the eigenvalue of \(A\) nearest to a given approximation \(\rho_{in}\)) and the corresponding eigenvector, we can use the MM as an accelerating technique for the inverse iteration (II) method: after few steps of II, we start the MM with \(d\) and \(x^{(0)}\) equal to the last vector iterate of II, normalized with respect to the Euclidean norm.

In order to describe the behaviour of the MM as acceleration of II, Table 1 shows the results of some computational experiments carried out on Cray C90.

Test problem TP1 is concerning with a matrix \(A = T^2\), where \(T = \{t_{i,j}\}\) is a tridiagonal matrix of order 300 with \(t_{i,i} = 2, t_{i,i+1} = t_{i,i-1} = 1\), having known eigenvalues and eigenvectors (see e.g. [23]). In test problem TP2, we consider a sparse symmetric matrix \(A\) of order 1000 without a particular structure. The sparsity of \(A\) is equal to 95% and the eigenvalues are uniformly random generated in the interval \([1, 1000]\).

In Table 1, \(k^*\) indicates the number of iterations for the convergence of II and MM (the number in brackets represents the steps of II executed before MM); \(t^*\) is the computer time in seconds for determining \(\rho^*\) with the relative error

\[
err = \frac{|\rho_k^* - \rho^*|}{|\rho^*|}
\]

and the Euclidean norm of residual

\[
res = \left\| (A - \rho_k^* I) x^{(\rho_k^*)} \right\|
\]

The convergence rate of II depends on the ratio

\[
R = \frac{|\rho^* - \rho_{in}|}{\min_{\lambda_i \neq \rho^*} |\lambda_i - \rho_{in}|}
\]

where \(\rho_{in} = 0\) when we want to determine the eigenvalue of minimum absolute value.

The two–steps procedure (20) is used as inner solver for the MM and the solution of the indefinite symmetric systems arising in MM and II is obtained by the SYMMLQ routine ([17]).
Table 1: MM as acceleration of inverse iteration method

<table>
<thead>
<tr>
<th>n</th>
<th>ε</th>
<th>(\rho_{in})</th>
<th>(R)</th>
<th>(err)</th>
<th>(res)</th>
<th>(k^*)</th>
<th>(k_{tot})</th>
<th>(t^*)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TP1</td>
<td>300</td>
<td>0.63 (-1)</td>
<td>5.40 (-7)</td>
<td>1.43 (-12)</td>
<td>6</td>
<td>5183</td>
<td>0.35</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>5.19 (-7)</td>
<td>7.81 (-13)</td>
<td>1 (1)</td>
<td>1311 (1281)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TP2</td>
<td>1000</td>
<td>0.57</td>
<td>8.63 (-4)</td>
<td>1.36 (-9)</td>
<td>37</td>
<td>2009</td>
<td>21.4</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>4.96 (-7)</td>
<td>5.43 (-11)</td>
<td>2 (3)</td>
<td>578 (569)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>511</td>
<td>0.51</td>
<td>4.96 (-7)</td>
<td>1.23 (-10)</td>
<td>23</td>
<td>9175 (1991)</td>
<td>304</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>4.96 (-7)</td>
<td>9.67 (-11)</td>
<td>3 (1)</td>
<td>409 (574)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>0.14</td>
<td>1.88 (-4)</td>
<td>7.47 (-11)</td>
<td>12</td>
<td>2071</td>
<td>21.2</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.88 (-4)</td>
<td>9.96 (-11)</td>
<td>1 (2)</td>
<td>409 (574)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>8.63</td>
<td>0.95</td>
<td>7.85 (-7)</td>
<td>6.93 (-7)</td>
<td>274</td>
<td>20953</td>
<td>273</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>7.85 (-7)</td>
<td>6.24 (-11)</td>
<td>4 (2)</td>
<td>2067 (662)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Comparison among different inner solvers for the inner LSE problem (3).

<table>
<thead>
<tr>
<th>(\rho_{in})</th>
<th>(R)</th>
<th>(err)</th>
<th>(res)</th>
<th>(k^*)</th>
<th>(k_{tot})</th>
<th>(t^*)</th>
</tr>
</thead>
<tbody>
<tr>
<td>II</td>
<td>0.63 (-1)</td>
<td>8.63 (-4)</td>
<td>1.36 (-9)</td>
<td>37</td>
<td>2009</td>
<td>21.4</td>
</tr>
<tr>
<td>MM</td>
<td>0.57</td>
<td>4.96 (-7)</td>
<td>5.43 (-11)</td>
<td>23</td>
<td>9175 (1991)</td>
<td>304</td>
</tr>
<tr>
<td>II</td>
<td>0.14</td>
<td>1.88 (-4)</td>
<td>7.47 (-11)</td>
<td>12</td>
<td>2071</td>
<td>21.2</td>
</tr>
<tr>
<td>MM</td>
<td>0.95</td>
<td>7.85 (-7)</td>
<td>6.93 (-7)</td>
<td>274</td>
<td>20953</td>
<td>273</td>
</tr>
</tbody>
</table>

number of steps performed by SYMMLQ is denoted by \(k_{tot}\) (in brackets is denoted the total number of steps of SYMMLQ performed in II before MM).

In both methods, the \(x\)-iterate is normalized with respect to the Euclidean norm at each step \(k\) and the convergence test is

\[
|\rho_k - \rho_{k-1}| \leq \varepsilon |\rho_k| \quad \text{or} \quad \left\| (A - \rho_k I) \frac{x(\rho_k)}{\|x(\rho_k)\|} \right\| \leq \varepsilon \|A\|
\]

The notation 8.53 (-4), for instance, means \(8.53 \cdot 10^{-4}\).

In Table 2 is reported a comparison among different inner solvers for the inner LSE problem (3). In this case \(k_{tot}\) indicates the total number of steps executed by the inner solver.

We indicate the inner solvers in the following way:

- CG denotes the solution of the system (19) by the conjugate gradient method with a diagonal preconditioner;
- 2–SYMMLQ indicates the two–steps procedure (20); each system is solved by the SYMMLQ routine;
- SYMMLQ represents the solution of the system (21) by the SYMMLQ routine;
- LSQR indicates the weighting method (22) implemented by the LSQR routine.

The time \(t^*\) does not include the time for the initial steps of II.

Table 2 shows that the choice of a particular inner solver does not influence the accuracy of the computed eigenvalue. Nevertheless the two inner solvers CG and 2–SYMMLQ appear convenient from the viewpoint of the computational complexity. Furthermore, if we apply a refinement
Table 2: Different inner solvers for MM

<table>
<thead>
<tr>
<th>Solver</th>
<th>$k_{tot}$</th>
<th>$err$</th>
<th>$res$</th>
<th>$t^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rho_{in} = 0$</td>
<td>CG</td>
<td>1678</td>
<td>8.63 (-4)</td>
<td>3.04 (-11)</td>
</tr>
<tr>
<td>$R = 0.57$</td>
<td>2-SYMMLQ</td>
<td>578</td>
<td>8.63 (-4)</td>
<td>5.43 (-11)</td>
</tr>
<tr>
<td>$k^* = 2$ (3)</td>
<td>SYMMLQ</td>
<td>3504</td>
<td>8.63 (-4)</td>
<td>2.13 (-9)</td>
</tr>
<tr>
<td></td>
<td>LSQR</td>
<td>2492</td>
<td>8.63 (-4)</td>
<td>1.74 (-10)</td>
</tr>
<tr>
<td>$\rho_{in} = 511$</td>
<td>CG</td>
<td>634</td>
<td>4.96 (-7)</td>
<td>6.97 (-11)</td>
</tr>
<tr>
<td>$R = 0.51$</td>
<td>2-SYMMLQ</td>
<td>9175</td>
<td>4.96 (-7)</td>
<td>1.23 (-10)</td>
</tr>
<tr>
<td>$k^* = 3$ (1)</td>
<td>SYMMLQ</td>
<td>5831</td>
<td>4.96 (-7)</td>
<td>3.68 (-12)</td>
</tr>
<tr>
<td></td>
<td>LSQR</td>
<td>4288</td>
<td>4.96 (-7)</td>
<td>1.63 (-10)</td>
</tr>
<tr>
<td>$\rho_{in} = 6$</td>
<td>CG</td>
<td>1095</td>
<td>1.88 (-4)</td>
<td>1.27 (-10)</td>
</tr>
<tr>
<td>$R = 0.14$</td>
<td>2-SYMMLQ</td>
<td>409</td>
<td>1.88 (-4)</td>
<td>9.96 (-11)</td>
</tr>
<tr>
<td>$k^* = 1$ (2)</td>
<td>SYMMLQ</td>
<td>2224</td>
<td>1.88 (-4)</td>
<td>9.06 (-10)</td>
</tr>
<tr>
<td></td>
<td>LSQR</td>
<td>1391</td>
<td>1.88 (-4)</td>
<td>1.69 (-10)</td>
</tr>
<tr>
<td>$\rho_{in} = 8.63$</td>
<td>CG</td>
<td>1487</td>
<td>7.85 (-7)</td>
<td>3.90 (-14)</td>
</tr>
<tr>
<td>$R = 0.95$</td>
<td>2-SYMMLQ</td>
<td>2067</td>
<td>7.85 (-7)</td>
<td>6.24 (-11)</td>
</tr>
<tr>
<td>$k^* = 4$ (2)</td>
<td>SYMMLQ</td>
<td>9266</td>
<td>7.85 (-7)</td>
<td>1.53 (-12)</td>
</tr>
<tr>
<td></td>
<td>LSQR</td>
<td>6782</td>
<td>7.85 (-7)</td>
<td>4.89 (-16)</td>
</tr>
</tbody>
</table>

scheme to the inner solver LSQR, the accuracy of the computed eigenvalue does not change; the value of $res$ decreases, but at the cost of an augmented computational complexity.

From Table 1 and Table 2 it is evident the effectiveness of the MM combined with the 2–SYMMLQ or CG inner solvers as acceleration of the II.

Table 3 shows a comparison between the MM (with CG and 2–SYMMLQ inner solvers) and the scaled Newton method (see [10]) with SYMMLQ routine as inner solver. Such version of Newton method has convergence rate of the third order.

Both the methods are used as acceleration of the II; as in Table 2, $t^*$ does not include the time for the initial steps of II.

From this table we can observe the faster convergence rate of MM in comparison with the Newton method. Moreover, in terms of computational complexity, the MM appears generally more convenient than Newton method provided that an appropriate solver for the inner LSE problem is used.

Finally, Table 4 shows that MM and Newton method can converge to different eigenpairs even if they start with the same approximation. In the table, $\rho_{II}$ is the approximation of $\rho^*$ obtained after $k_{II}$ steps of II, $\rho_k$ and $t_k$ denote the approximation of $\rho^*$ and the computer time at the iteration $k$ respectively.

We observe that MM gives an approximation of the eigenvalue $\rho^* = 510.61458315519$ with a relative error $err = 7.18 (-6)$ while Newton method gives an approximation of the eigenvalue $\rho^* = 510.77974687413$ with a relative error $err = 4.96 (-7)$.

When the two methods must approximate the same eigenvalue (as in Table 3), it would be
necessary to increase the number of steps of II.

The most important conclusion that emerges by all the numerical studies carried out on the above and other (more ill–conditioned) test–problems is that the minimization method (MM) can be used efficiently in a *polyalgorithm* for solving large middle–ill–conditioned symmetric eigenproblems as a “correction algorithm” for an approximate eigenpair obtained by the inverse iteration method.

<table>
<thead>
<tr>
<th>$\rho_m$</th>
<th>$R$</th>
<th>$k^*$</th>
<th>$\text{err}$</th>
<th>$\text{res}$</th>
<th>$t^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.57</td>
<td>2 (3)</td>
<td>8.63 (-4)</td>
<td>3.04 (-11)</td>
<td>34.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 (3)</td>
<td>8.63 (-4)</td>
<td>5.43 (-11)</td>
<td>5.99</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 (3)</td>
<td>8.63 (-4)</td>
<td>8.93 (-11)</td>
<td>6.1</td>
</tr>
</tbody>
</table>

**Table 3:** MM and Newton method as acceleration of inverse iteration method

<table>
<thead>
<tr>
<th>$\rho_m$</th>
<th>$R$</th>
<th>$k^*$</th>
<th>$\text{err}$</th>
<th>$\text{res}$</th>
<th>$t^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>511</td>
<td>0.51</td>
<td>3 (1)</td>
<td>4.96 (-7)</td>
<td>6.97 (-11)</td>
<td>13.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 (1)</td>
<td>4.96 (-7)</td>
<td>1.23 (-10)</td>
<td>95.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4 (1)</td>
<td>4.96 (-7)</td>
<td>3.19 (-8)</td>
<td>77.1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\rho_m$</th>
<th>$R$</th>
<th>$k^*$</th>
<th>$\text{err}$</th>
<th>$\text{res}$</th>
<th>$t^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>0.14</td>
<td>1 (2)</td>
<td>1.88 (-4)</td>
<td>1.27 (-10)</td>
<td>22.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 (2)</td>
<td>1.88 (-4)</td>
<td>9.96 (-11)</td>
<td>4.26</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 (2)</td>
<td>1.88 (-4)</td>
<td>3.80 (-12)</td>
<td>6.13</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\rho_m$</th>
<th>$R$</th>
<th>$k^*$</th>
<th>$\text{err}$</th>
<th>$\text{res}$</th>
<th>$t^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.63</td>
<td>0.95</td>
<td>4 (2)</td>
<td>7.85 (-7)</td>
<td>3.90 (-14)</td>
<td>30.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4 (2)</td>
<td>7.85 (-7)</td>
<td>6.24 (-11)</td>
<td>21.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5 (2)</td>
<td>7.85 (-7)</td>
<td>5.68 (-14)</td>
<td>15.2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\rho_m$</th>
<th>$R$</th>
<th>$k^*$</th>
<th>$\text{err}$</th>
<th>$\text{res}$</th>
<th>$t^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>0.22</td>
<td>2 (2)</td>
<td>7.82 (-6)</td>
<td>8.56 (-11)</td>
<td>41.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 (2)</td>
<td>7.82 (-6)</td>
<td>1.43 (-11)</td>
<td>7.95</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 (2)</td>
<td>7.82 (-6)</td>
<td>2.17 (-13)</td>
<td>9.30</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\rho_m$</th>
<th>$R$</th>
<th>$k^*$</th>
<th>$\text{err}$</th>
<th>$\text{res}$</th>
<th>$t^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>510</td>
<td>0.71</td>
<td>3 (3)</td>
<td>7.18 (-6)</td>
<td>1.44 (-11)</td>
<td>49.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 (3)</td>
<td>7.18 (-6)</td>
<td>1.19 (-10)</td>
<td>104.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5 (3)</td>
<td>7.18 (-6)</td>
<td>1.46 (-12)</td>
<td>92.3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\rho_m$</th>
<th>$R$</th>
<th>$k^*$</th>
<th>$\text{err}$</th>
<th>$\text{res}$</th>
<th>$t^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>992.83</td>
<td>0.98</td>
<td>4 (6)</td>
<td>1.76 (-6)</td>
<td>2.91 (-9)</td>
<td>3.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4 (6)</td>
<td>1.76 (-6)</td>
<td>7.97 (-11)</td>
<td>18.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4 (6)</td>
<td>1.76 (-6)</td>
<td>6.95 (-10)</td>
<td>11.1</td>
</tr>
</tbody>
</table>
Table 4: MM and Newton method

<table>
<thead>
<tr>
<th></th>
<th>MM (CG)</th>
<th>NEWTON</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k$</td>
<td>$\rho_k$</td>
<td>$\rho^*$</td>
</tr>
<tr>
<td>1</td>
<td>510.45412343452</td>
<td>0.99</td>
</tr>
<tr>
<td>2</td>
<td>510.62521841044</td>
<td>4.61 (-2)</td>
</tr>
<tr>
<td>3</td>
<td>510.61824974027</td>
<td>2.85 (-4)</td>
</tr>
<tr>
<td>4</td>
<td>510.61824925811</td>
<td>1.28 (-11)</td>
</tr>
<tr>
<td>5</td>
<td>510.77999999983</td>
<td>2.10 (-6)</td>
</tr>
<tr>
<td>6</td>
<td>510.78000000000</td>
<td>9.29 (-14)</td>
</tr>
</tbody>
</table>

References


