FASCICULI MATHEMATICI
Nr 26 1996

SLOBODAN LAKIĆ

# TWO ITERATIVE METHODS FOR THE MATRIX INVERSE SQUARE ROOT

Abstract: This paper presents two fast iterative methods for finding the matrix inverse square root. The two iterative methods with high convergence rates are analyzed and their numerical stability properties are investigated. Numerical examples show that this methods can be superior then some earlier methods.

Subject Classification 65F30
Key words: matrix function, symmetric orthogonalization.

## 1. Introduction

Given a nonsingular matrix $A \times C^{n,n}$, a matrix $X$ such that

$$AX^2 = 1$$

is called an inverse square root of $A$ and is denoted by $X = A^{-1/2}$. The inverse square root of a matrix has applications in the computation of an optimal symmetric orthogonalization of a set of vec tors [6], theory of oscillations [5], etc. It is known that if the matrix $A$ is real, then a real inverse square root which is a function of $A$ exists if $A$ has no negative real eigenvalues [4]. That $X$ is a function of $A$ means it is a polynomial in $A$, [5]. Numerical methods for finding the inverse square root have been proposed in many papers, for example [1], [2], [3], etc. In section 2 we give an iterative method to compute $A^{-1/2}$. This method we will show has $(2k+1)$th – order convergence rate and improves the rate of convergence than the methods in [1], [2], [3]. Under some restriction on the spectrum of $A$ this method is localy stable. In section 3, we proposed an alternative locally stable method to compute $A^{-1/2}$. In section 4, we illustrate the performance of the methods by numerical examples.

## 2. Comutation of $A^{-1/2}$

*Lemma. Let $w$ be a complex number such that $w \neq 0$ and $\arg(w) \neq \pi$. For some pre–chosen nathural number $k$, we define the sequence $(z_{n,k})$ by*

$$(2.1) \qquad \begin{cases} z_{n,k} = 1, \\ z_{n+1,k} = z_{n,k} \dfrac{\sum\limits_{m=0}^{k} \alpha_{m,k} (w z_{n,k}^2)^m}{\sum\limits_{m=0}^{k} \beta_{m,k} (w z_{n,k}^2)^m}, \end{cases}$$

*where* $\alpha_{m,k} = \binom{2k+1}{2m+1}$, $\beta_{m,k} = \binom{2k+1}{2m}$.

*Then* $\lim\limits_{n\to\infty} z_{n,k} = \dfrac{1}{\sqrt{w}}$, *where $\sqrt{w}$ is the principal square root of $w$.*

(Without loss of generality, we will denote $z_{n,k}$, $\alpha_{m,k}$, $\beta_{m,k}$ by $z_n$, $\alpha_m$, $\beta_m$ respectively.)

*Proof.* First we prove that the sequence $(z_n)$ is well defined. For some pre–chosen $k \in N$ we define the function

$$h_k(z) = \frac{(1+z)^{2k+1} - (1-z)^{2k+1}}{(1+z)^{2k+1} + (1-z)^{2k+1}}, \quad z \in C.$$

Let $S = C_+ \cup C_-$, where

$C_+ = \{z \in C: \operatorname{Re}(z) > 0\}$, $C_- = \{z \in C: \operatorname{Re}(z) < 0\}$.

If we suppose that $(1+z)^{2k+1} + (1-z)^{2k+1} = 0$, then $z \neq 1$ and $(1+z)/(1-z) = e^{i\phi}$ where is $\phi$ no an odd multiple of $\pi$, else $z = \infty$. Now we have $z = \dfrac{i \sin\phi}{1+\cos\phi}$ and the poles of $h_k$ lie on the imaginary axis. Now we prove that $h_k(S) \subseteq S$. Let $z \in S$ and we suppose that

$$(2.2) \qquad h_k(z) = ai, \qquad a \in R$$

From (2.2) it follows

$$\frac{(1+ai)}{(1-ai)}=\left(\frac{(1+z)}{(1-z)}\right)^{2k+1} \tag{2.3}$$

Since $z\in S$ hence $|(1+z)/(1-z)|\neq 1$ and $|(1+z)/(1-z)|^{2k+1}\neq 1$ which is a contradiction with (2.3) because $|(1+ai)/(1-ai)|=1$. So, $h_k(S)\subseteq S$. The function $h_k$ is continuous on the set $S$ and hence either $h_k(C_+)\subseteq C_+$ or $h_k(C_+)\subseteq C_-$ because the continuous image of a connected set is connected .

But $l\in C_+$ and $h_k(l)=l\in C_+$, so $h_k(C_+)\subset C_+$. Thus if $s_0\in C_+$ then $s_{n+1}=h_k(s_n)\in C_+$. Using (2.1) one can prove that

$$z_{n+1}\sqrt{w}=\frac{(1+z_n\sqrt{w})^{2k+1}-(1-z_n\sqrt{w})^{2k+1}}{(1+z_n\sqrt{w})^{2k+1}+(1-z_n\sqrt{w})^{2k+1}} \tag{2.4}$$

Let $s_n=z_n\sqrt{w}$. Now $s_0=\sqrt{w}\in C_+$ because $\arg(w)\neq\pi$. Consequently $s_n\in C_+$. If we suppose that for some $n$ it holds $(1+s_n)^{2k+1}+(1-s_n)^{2k+1}=0$, then $s_n$ lie on the imaginary axis which is a contradiction. So for each $n$ it holds

$$(1+z_n\sqrt{w})^{2k+1}+(1-z_n\sqrt{w})^{2k+1}\neq 0$$

and the sequence $(z_n)$ is well defined.

It is easy to show that

(2.5)

$$\frac{z_{n+1}\pm 1}{\sqrt{w}}=\frac{2}{\sqrt{w}\left(\left(z_n+\frac{1}{\sqrt{w}}\right)^{2k+1}+\left(z_n-\frac{1}{\sqrt{w}}\right)^{2k+1}\right)}\left(z_n\pm\frac{1}{\sqrt{w}}\right)^{2k+1}$$

$$\frac{z_n - \frac{1}{\sqrt{w}}}{z_n + \frac{1}{\sqrt{w}}} = \left( \frac{1-\sqrt{w}}{1+\sqrt{w}} \right)^{(2k+1)^n} \tag{2.6}$$

Consequently it follows that

$$z_n - \frac{1}{\sqrt{w}} = \frac{2d_n}{\sqrt{w}\,(1-d_n)} \tag{2.7}$$

where

$$d_n = \left( \frac{1-\sqrt{w}}{1+\sqrt{w}} \right)^{(2k+1)^n} .$$

Since $\sqrt{w}$ is the principal square root of $w$, hence $\operatorname{Re}\sqrt{w} > 0$ and $|(1-\sqrt{w})/(1+\sqrt{w})| < 1$.

So we have

$$\lim_{n\to\infty} d_n = 0 \qquad \text{and consequently} \qquad \lim_{n\to\infty} z_n = \frac{1}{\sqrt{w}} .$$

Now we define a matrix sequence $(X_n)$ to compute $A^{-1/2}$.

$$(1) \quad \begin{cases} X_0 = 1 \\ X_{n+1} = X_n \left( \sum_{m=0}^{k} \alpha_m (AX_n^2)^m \right) \left( \sum_{m=0}^{k} \beta_m (AX_n^2)^m \right)^{-1} \end{cases}$$

where $\alpha_m$, $\beta_m$ and $k$ are as in Lemma.

For our analysis we assume that $A$ is diagonalizable, that is there exists a nonsingular matrix $V$ such that

$$V^{-1} AV = D$$

where $D = \operatorname{diag}(a_1, \ldots, a_n)$ and $a_1, \ldots, a_n$ are the eigenvalues of $A$.

*Theorem 2.1. Let $A \in C^{n,n}$ be nonsingular and diagonalizable and assume A has no negative real eigenvalues. Then the matrix sequence (1) converges to X, where X is the principal inverse square root of A and*

$$\|X_{n+1} - X\| = 0\,(\|X_n - X\|^{2k+1}). \tag{2.9}$$

*Proof.* We define $L_n = V^{-1} X_n V$ where $V$ is as in (2.8). From iteration (1) we have

$$(2.10) \qquad \begin{cases} L_0 = 1 \\ L_{n+1} = L_n \left( \sum_{m=0}^{k} \alpha_m (DL_n^2)^m \right) \left( \sum_{m=0}^{k} \beta_m (DL_n^2)^m \right)^{-1} \end{cases}$$

The sequence $L_i$ is a sequence of diagonal matrices $L_i = diag\,(l_1^{(i)}, \ldots, l_n^{(i)})$. The equation (2.10) is equivalent to $n$ scalar equations

$$l_j^{(0)} = 1, \qquad l_j^{(n+1)} = l_j^{(n)} \frac{\sum_{m=0}^{k} \alpha_m \left( a_j (l_j^{(n)})^2 \right)^m}{\sum_{m=0}^{k} \beta_m \left( a_j (l_j^{(n)})^2 \right)^m}, \quad j = 1, \ldots, n.$$

Application of lemma yields

$$\lim_{n \to \infty} L_n = D^{-1/2} = diag \left( \frac{1}{\sqrt{a_1}}, \ldots, \frac{1}{\sqrt{a_n}} \right).$$

Consequently $\lim_{n \to \infty} X_n = A^{-1/2}$.

Now, $X_{n+1} - X = 2X((X + X_n)^{2k+1} + (X - X_n)^{2k+1})^{-1} (X_n - X)^{2k+1}$ follows upon using equation (2.5).

From the above equation by taking norms we have (2.9).

## 2. Stability Analysis

We now investigate the numerical local stability of the method (I). By local stability we mean that in a neigbourhood of the solution $A^{-1/2}$, a perturbation due to rounding error will notbe magnified in the succedings steps.

Let a rounding error $E_n = \mathrm{O}(\varepsilon)$ be introduced in the $n$–th step. Denote by $\tilde{X}_n$ the computed solution at this step, that is $\tilde{X}_n = X_n + E_n$. Using the perturbation result in [7]

$$(A+B)^{-1} = A^{-1} - A^{-1}BA^{-1} + O(\|B\|^2) \tag{2.11}$$

we have

$$\tilde{X}_{n+1} = \tilde{X}_n \left( \sum_{m=0}^{k} \alpha_m (A\tilde{X}_n^2)^m \right) \left( \sum_{m=0}^{k} \beta_m (A\tilde{X}_n^2)^m \right)^{-1},$$

$$\tilde{X}_{n+1} = (X_n + E_n) \times$$

$$\times \left( \sum_{m=0}^{k} \alpha_m (AX_n^2)^m + \sum_{m=0}^{k} \alpha_m \sum_{i=0}^{m-1} (AX_n^2)^i A(X_nE_n + E_nX_n)(AX_n^2)^{m-i-1} \right) \times$$

$$\times \left( \sum_{m=0}^{k} \beta_m (AX_n^2)^m + \sum_{m=0}^{k} \beta_m \sum_{i=0}^{m-1} (AX_n^2)^i A(X_nE_n + E_nX_n)(AX_n^2)^{m-i-1} \right)^{-1} +$$

$$+ \mathrm{O}(\varepsilon^2)$$

$$\tilde{X}_{n+1} = (X_n + E_n) \times$$

$$\times \left( \sum_{m=0}^{k} \alpha_m (AX_n^2)^m + \sum_{m=1}^{k} \alpha_m \sum_{i=0}^{m-1} (AX_n^2)^i A(X_nE_n + E_nX_n)(AX_n^2)^{m-i-1} \right) \times$$

$$\times \left( \left[ \sum_{m=0}^{k} \beta_m (AX_n^2)^m \right]^{-1} - \left[ \sum_{m=0}^{k} \beta_m (AX_n^2)^m \right]^{-1} \left[ \sum_{m=1}^{k} \beta_m \sum_{i=0}^{m-1} (AX_n^2)^i \times \right. \right.$$

$$\left. \left. \times A(X_nE_n + E_nX_n)(AX_n^2)^{m-i-1} \right] \left[ \sum_{m=0}^{k} \beta_m (AX_n^2)^m \right]^{-1} \right) + \mathrm{O}(\varepsilon^2).$$

We define $\tilde{E}_n = V^{-1}E_nV$.

Direct calculations give

$$\tilde{E}_{n+1} = L_n \left( \sum_{m=1}^{k} \alpha_m \sum_{i=0}^{m-1} (DL_n^2)^i D(L_n\tilde{E}_n + \tilde{E}_nL_n)(DL_n^2)^{m-i-1} \right) \times$$

$$\times\left(\left[\sum_{m=0}^{k}\beta_m(DL_n^2)^m\right]^{-1}-\left[\sum_{m=0}^{k}\alpha_m(DL_n^2)^m\right]\left[\sum_{m=0}^{k}\beta_m(DL_n^2)^m\right]^{-1}\right.$$

$$\left.\times\left[\sum_{m=1}^{k}\beta_m\sum_{i=0}^{m-1}(DL_n^2)^i\,A(L_n\widetilde{E}_n+\widetilde{E}_nL_n)\,(DL_n^2)^{m-i-1}\right]\left[\sum_{m=0}^{k}\beta_m(DL_n^2)^m\right]^{-1}\right)$$

$$\times\widetilde{E}_n\left[\sum_{m=0}^{k}\alpha_m(DL_n^2)^m\right]\left[\sum_{m=0}^{k}\beta_m(DL_n^2)^m\right]^{-1}+0(\varepsilon^2)\ .$$

Since $l_j^{(n)}=\varepsilon+\dfrac{1}{\sqrt{a_j}}$, $\alpha_m=\beta_{k-m}$ and $\sum_{m=0}^{k}\alpha_m=\sum_{m=0}^{k}\beta_m$ we obtain

$$\widetilde{E}_{n+1}=\left(1-\frac{\sum_{m=1}^{k}m(\beta_m-\beta_{k-m})}{\sum_{m=0}^{k}\beta_m}\right)\widetilde{E}_n-$$

$$-\frac{\sum_{m=1}^{k}m(\beta_m-\beta_{k-m})}{\sum_{m=0}^{k}\beta_m}D^{\frac{1}{2}}\widetilde{E}_nD^{\frac{1}{2}}+0(\varepsilon^2).$$

Now we show that

$$\frac{\sum_{m=1}^{k}m(\beta_m-\beta_{k-m})}{\sum_{m=0}^{k}\beta_m}=\frac{1}{2} \tag{2.12}$$

The equality (2.12) is equivalent to

$$\sum_{i=0}^{k}(2k+1)\binom{2k+1}{2i}=\sum_{i=1}^{k}4i\binom{2k+1}{2i}\ .$$

Using the facts that

$$\sum_{i=0}^{k}\binom{2k+1}{2i}=2^{2k}\,,\qquad \sum_{i=1}^{k}\binom{2k}{2i-1}=2^{2k-1}\,,$$

$$2i\binom{2k+1}{2i}=(2k+1)\binom{2k}{2i-1}\qquad \text{we have}$$

$$\sum_{i=0}^{k}(2k+1)\binom{2k+1}{2i}=(2k+1)2^{2k}$$

$$\sum_{i=1}^{k}4i\binom{2k+1}{2i}=2\sum_{i=1}^{k}(2k+1)\binom{2k}{2i-1}=(2k+1)2^{2k}\,.$$

So, it holds (2.12).

Finaly,

$$\tilde{E}_{n+1}=\frac{1}{2}(\tilde{E}_n-D^{\frac{1}{2}}\tilde{E}_n D^{-\frac{1}{2}})+0(\varepsilon^2)\,,$$

$$E_{n+1}=\frac{1}{2}(E_n-A^{\frac{1}{2}}E_n A^{-\frac{1}{2}})+0(\varepsilon^2)\,.$$

Assuming that $n$ is large enough and that $\varepsilon$ is small enough, we have

$$E_{n+1}\approx\frac{1}{2}(E_n-A^{\frac{1}{2}}E_n A^{-\frac{1}{2}})\,.$$

The above equality is equivalent to

$$\operatorname{Vec}E_{n+1}\approx\operatorname{Vec}E_n\qquad\text{where}\qquad G\in C^{n^2,n^2}\,,$$

$$G=I-A^{\frac{1}{2}}*A^{\frac{-1}{2}}\qquad (A*B \text{ denote the Kronecker product}).$$

Hence for $j=1,2,\ldots,\ \operatorname{Vec}E_{n+j}\approx G^j\operatorname{Vec}E_n$.

Consequently the errors in the succedings steps will not be magni fied if $\rho(G)<1$. The spectrum of $G$, $\sigma(G)$ can be shown to be [5]

$$\sigma(G)=\left(g_{i,j}:g_{i,j}=0.5\left(1-\sqrt{\frac{a_i}{a_j}}\right),\ 1\le i,\ j\le n\right)\,.$$

So $\rho(G)<1$ is equivalent to

(2.13) $$\left|1-\sqrt{\frac{a_i}{a_j}}\right|<2 .$$

The above inequality gives a sufficient condition for the method ( I ) to be locally stable.

## 3. An Alternative Method

We propose an alternative method to compute $A^{-1/2}$. This method will be shown to be locally stable. Let $A \in C^{n,n}$ be nonsingular, we define the sequences $(T_n)$ and $(S_n)$ as follows

$$(\mathrm{II})\quad \begin{cases} T_{n+1} = T_n\left(\sum_{m=0}^{k} \alpha_m S_n^m\right)\left(\sum_{m=0}^{k} \beta_m S_n^m\right)^{-1}, & T_0 = I \\ S_{n+1} = S_n\left(\left(\sum_{m=0}^{k} \alpha_m S_n^m\right)\left(\sum_{m=0}^{k} \beta_m S_n^m\right)^{-1}\right)^2, & S_0 = A \end{cases}$$

where $\alpha_m$, $\beta_m$ and $k$ are as in (2.1).

*Theorem 3.1. Let $A \in C^{n,n}$ be nonsingular and diagonalizable. Assume that A has no negative real eigenvalues. Then* $\lim_{n\to\infty} T_n = A^{-1/2}$, $\lim_{n\to\infty} S_n = 1$, *where $A^{-1/2}$ is the principal inverse square root of A. Further*

$$\|T_n - A^{-1/2}\| = 0(\|T_n - A^{-1/2}\|^{2k+1}) .$$

*Proof.* Let

(3.1) $$L_n = V^{-1}T_nV,\ H_n = V^{-1}S_nV$$

Now (II) reduces to

$$(3.2)\qquad \begin{cases} L_{n+1} = L_n \left( \sum_{m=0}^{k} \alpha_m H_n^m \right) \left( \sum_{m=0}^{k} \beta_m H_n^m \right)^{-1}, \quad L_0 = I \\ H_{n+1} = H_n \left( \left( \sum_{m=0}^{k} \alpha_m H_n^m \right) \left( \sum_{m=0}^{k} \beta_m H_n^m \right)^{-1} \right)^2, \quad H_0 = D \end{cases}$$

From equations (3.2) it follows that $L_j$ and $H_j$ are diagonal matrices. Let

$$(3.3)\qquad L_i = \operatorname{diag}(l_1^{(j)}, \ldots, l_n^{(j)}), \qquad H_j = \operatorname{diag}(h_1^{(j)}, \ldots, h_n^{(j)})$$

Equation (3.2) is equivaklent to $n$ scalar sequences

$$(3.4)\qquad \begin{cases} l_i^{(0)} = 1, & h_i^{(0)} = a_i, \\ l_i^{(n+1)} = l_i^{(n)} \dfrac{\sum_{m=0}^{k} \alpha_m (h_i^{(n)})^m}{\sum_{m=0}^{k} \beta_m (h_i^{(n)})^m}, & h_i^{(n+1)} = h_i^{(n)} \left( \dfrac{\sum_{m=0}^{k} \alpha_m (h_i^{(n)})^m}{\sum_{m=0}^{k} \beta_m (h_i^{(n)})^m} \right)^2 \end{cases}$$

Similarly as in Lemma One can show that

$$(3.5)\qquad l_i^{(n+1)} - \frac{1}{\sqrt{a_i}} = \frac{2\left( l_i^{(n)} - \frac{1}{\sqrt{a_i}} \right)^{2k+1}}{\sqrt{a_i}\left( l_i^{(n)} + \frac{1}{\sqrt{a_i}} \right)^{2k+1} + \left( -l_i^{(n)} + \frac{1}{\sqrt{a_i}} \right)^{2k+1}}$$

and

$$(3.6)\qquad \lim_{n\to\infty} l_i^{(n)} = \frac{1}{\sqrt{a_i}}$$

Since the sequences $(l_i^{(n)})$ converges and it holds (3.4), hence $\lim_{n\to\infty} h_i^{(n)} = 1$.

From (3.1), (3.3), (3.5), (3.6) and (3.7) it follows

$$\lim_{n\to\infty} T_n = A^{-1/2}, \qquad \lim_{n\to\infty} S_n = 1 \qquad \text{and}$$

$$T_{n+1} - A^{-1/2} = 2A^{-1/2}((A^{-1/2} + T_n)^{2k+1} + (A^{-1/2} - T_n)^{2k+1})^{-1}(T_n - A^{-1/2})^{2k+1}.$$

Taking the norm of the above equation, the bound in the theorem is established.

## Stability Analysis

Assume that at the $n$–th step errors $P_n$ and $Q_n$ are introduced in $T_n$ and $S_n$ respectively. Let $\tilde{T}_n$ and $\tilde{S}_n$ be the computed matrices of this step. Now $\tilde{T}_n = T_n + P_n$, $\tilde{S}_n = S_n + Q_n$ where $P_n = O(\varepsilon)$ and $Q_n = O(\varepsilon)$.

We define computations give

$$P_{n+1} = T_n\left(\left(\sum_{m=1}^{k}\alpha_m\sum_{i=0}^{m-1}S_n^i Q_n S_n^{m-i-1}\right)\left(\sum_{m=0}^{k}\beta_m S_n^m\right) - \left(\sum_{m=0}^{k}\alpha_m S_n^m\right)\right.$$

$$\left.\times\left(\sum_{m=0}^{k}\beta_m S_n^m\right)^{-1}\left(\sum_{m=1}^{k}\beta_m\sum_{i=0}^{m-1}S_n^i Q_n S_n^{m-i-1}\right)\left(\sum_{m=0}^{k}\beta_m S_n^m\right)^{-1}\right)$$

$$+P_n\left(\sum_{m=0}^{k}\alpha_m S_n^m\right)\left(\sum_{m=0}^{k}\beta_m S_n^m\right)^{-1} + O(\varepsilon^2),$$

$$\tilde{P}_{n+1} = L_n\left(\left(\sum_{m=1}^{k}\alpha_m\sum_{i=0}^{m-1}H_n^i\tilde{Q}_n H_n^{m-i-1}\right)\left(\sum_{m=0}^{k}\beta_m H_n^m\right)^{-1} - \left(\sum_{m=0}^{k}\alpha_m H_n^m\right)\right.$$

(3.8)

$$\left.\times\left(\sum_{m=0}^{k}\beta_m H_n^m\right)^{-1}\left(\sum_{m=1}^{k}\beta_m\sum_{i=0}^{m-1}H_n^i\tilde{Q}_n H_n^{m-i-1}\right)\left(\sum_{m=0}^{k}\beta_m H_n^m\right)^{-1}\right)$$

$$+\tilde{P}_n\left(\sum_{m=0}^{k}\alpha_m H_n^m\right)\left(\sum_{m=0}^{k}\beta_m H_n^m\right)^{-1} + O(\varepsilon^2)$$

$$Q_{n+1} = S_n\left(\left(\sum_{m=1}^{k}\alpha_m\sum_{i=0}^{m-1}S_n^i Q_n S_n^{m-i-1}\right)\left(\sum_{m=0}^{k}\beta_m S_n^m\right)^{-1}\left(\sum_{m=0}^{k}\alpha_m S_n^m\right)\right.$$

$$\times\left(\sum_{m=0}^{k}\beta_m S_n^m\right)^{-1}-\left(\sum_{m=0}^{k}\alpha_m S_n^m\right)\left(\sum_{m=0}^{k}\beta_m S_n^m\right)^{-1}\left(\sum_{m=1}^{k}\beta_m\sum_{i=0}^{m-1}S_n^i Q_n S_n^{m-i-1}\right)$$

$$\times\left(\sum_{m=0}^{k}\beta_m S_n^m\right)^{-1}\left(\sum_{m=0}^{k}\alpha_m S_n^m\right)\left(\sum_{m=0}^{k}\beta_m S_n^m\right)^{-1}+\left(\sum_{m=0}^{k}\alpha_m S_n^m\right)\left(\sum_{m=0}^{k}\beta_m S_n^m\right)^{-1}$$

$$\times\left(\sum_{m=1}^{k}\alpha_m\sum_{i=0}^{m-1}S_n^i Q_n S_n^{m-i-1}\right)\left(\sum_{m=0}^{k}\beta_m S_n^m\right)^{-1}-\left(\left(\sum_{m=0}^{k}\alpha_m S_n^m\right)\left(\sum_{m=0}^{k}\beta_m S_n^m\right)^{-1}\right)^2$$

$$\times\left(\sum_{m=1}^{k}\beta_m\sum_{i=0}^{m-1}S_n^i Q_n S_n^{m-i-1}\right)\left(\sum_{m=0}^{k}\beta_m S_n^m\right)^{-1}\Bigg)$$

$$+Q_n\left(\left(\sum_{m=0}^{k}\alpha_m S_n^m\right)\left(\sum_{m=0}^{k}\beta_m S_n^m\right)^{-1}\right)^2+O(\varepsilon^2)\ ,$$

$$\tilde{Q}_{n+1}=H_n\Bigg(\left(\sum_{m=1}^{k}\alpha_m\sum_{i=0}^{m-1}H_n^i\tilde{Q}_n H_n^{m-i-1}\right)\left(\sum_{m=0}^{k}\beta_m H_n^m\right)^{-1}\left(\sum_{m=0}^{k}\alpha_m H_n^m\right)$$

$$\times\left(\sum_{m=0}^{k}\beta_m H_n^m\right)^{-1}-\left(\sum_{m=0}^{k}\alpha_m H_n^m\right)\left(\sum_{m=0}^{k}\beta_m H_n^m\right)^{-1}\left(\sum_{m=1}^{k}\beta_m\sum_{i=0}^{m-1}H_n^i\tilde{Q}_n H_n^{m-i-1}\right)$$

$$\times\left(\sum_{m=0}^{k}\beta_m H_n^m\right)^{-1}\left(\sum_{m=0}^{k}\alpha_m H_n^m\right)\left(\sum_{m=0}^{k}\beta_m H_n^m\right)^{-1}+\left(\sum_{m=0}^{k}\alpha_m H_n^m\right)\left(\sum_{m=0}^{k}\beta_m H_n^m\right)^{-1}$$

$$\times\left(\sum_{m=1}^{k}\alpha_m\sum_{i=0}^{m-1}H_n^i\tilde{Q}_n H_n^{m-i-1}\right)\left(\sum_{m=0}^{k}\beta_m H_n^m\right)^{-1}-\left(\left(\sum_{m=0}^{k}\alpha_m H_n^m\right)\left(\sum_{m=0}^{k}\beta_m H_n^m\right)^{-1}\right)^2$$

(3.9)
$$\times\left(\sum_{m=1}^{k}\beta_m\sum_{i=0}^{m-1}H_n^i\tilde{Q}_n H_n^{m-i-1}\right)\left(\sum_{m=0}^{k}\beta_m H_n^m\right)^{-1}\Bigg)$$

$$+Q_n\left(\left(\sum_{m=0}^{k}\alpha_m H_n^m\right)\left(\sum_{m=0}^{k}\beta_m H_n^m\right)^{-1}\right)^2+O(\varepsilon^2)\ .$$

Writing (3.8) and (3.9) element–wise we have

$$\tilde{q}_{rs}^{(n+1)} = b_{rs}^{(n)}\,\tilde{q}_{rs}^{(n)},$$

$$\tilde{p}_{rs}^{(n+1)} = v_{rs}^{(n)}\,\tilde{q}_{rs}^{(n)} + c_{rs}^{(n)}\,\tilde{p}_{rs}^{(n)},$$

$$v_{rs}^{(n)} = l_r^{(n)}\left(\frac{\sum_{m=1}^{k}\alpha_m\sum_{i=0}^{m-1}(h_r^{(n)})^i(h_s^{(n)})^{m-i-1}}{\sum_{m=0}^{k}\beta_m(h_s^{(n)})^m} - \frac{\left(\sum_{m=0}^{k}\alpha_m(h_r^{(n)})^m\right)\left(\sum_{m=1}^{k}\beta_m\sum_{i=0}^{m-1}(h_r^{(n)})^i(h_s^{(n)})^{m-i-1}\right)}{\left(\sum_{m=0}^{k}\beta_m(h_r^{(n)})^m\right)\left(\sum_{m=0}^{k}\beta_m(h_s^{(n)})^m\right)}\right),$$

$$b_{rs}^{(n)} = b_r^{(n)}\left(\frac{\left(\sum_{m=1}^{k}\alpha_m\sum_{i=0}^{m-1}(h_r^{(n)})^i(h_s^{(n)})^{m-i-1}\right)\left(\sum_{m=0}^{k}\alpha_m(h_s^{(n)})^m\right)}{\left(\sum_{m=0}^{k}\beta_m(h_s^{(n)})^m\right)^2}\right.$$

$$-\frac{\left(\sum_{m=0}^{k}\alpha_m(h_r^{(n)})^m\right)\left(\sum_{m=1}^{k}\beta_m\sum_{i=0}^{m-1}(h_r^{(n)})^i(h_s^{(n)})^{m-i-1}\right)\left(\sum_{m=0}^{k}\alpha_m(h_s^{(n)})^m\right)}{\left(\sum_{m=0}^{k}\beta_m(h_r^{(n)})^m\right)\left(\sum_{m=0}^{k}\beta_m(h_s^{(n)})^m\right)^2}$$

$$\left.+\frac{\left(\sum_{m=0}^{k}\alpha_m(h_r^{(n)})^m\right)\left(\sum_{m=1}^{k}\alpha_m\sum_{i=0}^{m-1}(h_r^{(n)})^i(h_s^{(n)})^{m-i-1}\right)}{\left(\sum_{m=0}^{k}\beta_m(h_r^{(n)})^m\right)\left(\sum_{m=0}^{k}\beta_m(h_s^{(n)})^m\right)}\right)$$

$$-\frac{\left(\sum_{m=0}^{k}\alpha_m(h_r^{(n)})^m\right)\left(\sum_{m=1}^{k}\beta_m\sum_{i=0}^{m-1}(h_r^{(n)})^i(h_s^{(n)})^{m-i-1}\right)}{\left(\sum_{m=0}^{k}\beta_m(h_r^{(n)})^m\right)^2\left(\sum_{m=0}^{k}\beta_m(h_s^{(n)})^m\right)}+$$

$$+\left(\frac{\sum_{m=0}^{k}\alpha_m(h_s^{(n)})^m}{\sum_{m=0}^{k}\beta_m(h_s^{(n)})^m}\right)^2,$$

$$c_{rs}^{(n)} = \frac{\sum_{m=0}^{k}\alpha_m(h_s^{(n)})^m}{\sum_{m=0}^{k}\beta_m(h_s^{(n)})^m}.$$

Let

$$e_{rs}^{(n)} = \begin{bmatrix}\tilde{q}_{rs}^{(n)}\\ \tilde{p}_{rs}^{(n)}\end{bmatrix}, \qquad w_{rs}^{(n)} = \begin{bmatrix}b_{rs}^{(n)} & 0\\ v_{rs}^{(n)} & c_{rs}^{(n)}\end{bmatrix}.$$

Now we have $e_{rs}^{(n)} = w_{rs}^{(n)} e_{rs}^{(n)} + \mathrm{O}(\varepsilon^2)$.

From (3.6) and (3.7) we can write $w_{rs}^{(n)}$ as

$$w_{rs}^{(n)} = w_{rs} + \mathrm{O}(\varepsilon^{(n)}), \qquad w_{rs} = \begin{bmatrix}0 & 0\\ -1/2\sqrt{a_i} & 1\end{bmatrix}$$

where $\varepsilon^{(n)}$ is sufficiently small for large $n$. The matrix $w_{rs}$ has the eigenvalues 0 and 1, let $y_0$ and $y_1$ be the corresponding eigenvectors, so

$$e_{rs}^{(n)} = u_0^{(n)} y_0 + u_1^{(n)} y_1 .$$

For sufficiently small $\varepsilon$ and large $n$ we have

$$e_{rs}^{(n+k)} \approx w_{rs}^{k} e_{rs}^{(n)} = u_1^{(n)} y_1 , \qquad k = 1, 2, \ldots, .$$

Consequently $|e_{rs}^{(n+k)}| = |e_{rs}^{(n+1)}|$ the method (II) is locally stable.

It is easy to show that for iterations (I) and (II), the operation counts for one stage of each iterations, measured in flpos are as follows

| flops per stage | general A | symmetric positive definite A |
|---|---|---|
| method ( I ) | $(k+4)n^3$ | $(k+4)n^3/2$ |
| method ( II ) | $(k+4)n^3$ | $(k+4)n^3/2$ |

An alternative approach to compute $A^{-1/2}$ is via computing the eigenvectors of the matrix $A$. Let $A \in C^{n,n}$ be nonsingular, where $A$ is diagonalizable by $V$ and has eigenvalue matrix $D$. The inverse square root of $A$ can be defined as the matrix

$$A^{-1/2} = V D^{-1/2} V^{-1} \tag{3.11}$$

where $D^{-1/2} = \text{diag}(1/\sqrt{a_1}, \ldots, 1/\sqrt{a_n})$ and the matrix $V$ is eigenvector matrix [5].

In the general case the above approach to compute $A^{-1/2}$ is computationally expensive (for large $n$) and this is a major drawback when using (3.11) to compute $A^{-1/2}$. Namely, computation of $V$ requires $n^4/3$ flps (we need solved $n$ systems of linear equations of the form $(A - a_i C)x_i = 0$ by method Gaussian elimination).

The above approach to compute $A^{-1/2}$ can be effectiv, as we will see in the next section, if $A$ is hermitian matrix. In this case we can use the Jacobi iterative method to compute the eigenvalues and the eigenvectors of the matrix $A$. If the matrices $D$ and $V$ are computed after $k$ iterations, then computation of $A^{-1/2}$ reqires approximately

(3.12) $3kn^3/2$ flops.

## 4. Numerical Examples

In this section we will use the Frobenius matrix norm

$$|A|_F = \sqrt{\sum_{i,j} |a_{i,j}|^2} \quad \text{and the error} \quad e_n = \| A X_n^2 - 1 \|_F .$$

*Example* 1.

$$A = \begin{bmatrix} 5 & 4 & 1 & 1 \\ 4 & 5 & 1 & 1 \\ 1 & 1 & 4 & 2 \\ 1 & 1 & 2 & 4 \end{bmatrix}.$$

For $k = 1$ the method (I) converges in 3 iterations and $e_3 = 5.62E-7$.
For $k = 2$ the method (I) converges in 2 iterations and $e_2 = 1.12E-6$.
For $k = 3$ the method (I) converges in 2 iterations and $e_2 = 1.38E-5$.
For $k = 4$ the method (I) converges in 2 iterations and $e_2 = 2.65E-5$.
For $k = 5$ the method (I) converges in 2 iterations and $e_3 = 5.9E-5$.
For $k = 6$ the method (I) converges in 1 iterations and $e_1 = 1.99E-3$.
The eigenvalues of $A$ are 1, 2, 5, 10. This example illustrates that condition (2.13) is not a necessary condition for numerical stability of method (I). This example illustrates the coorespondence between the parameter $k$ and the computational complexity of the method (I).

*Example* 2.

$$A = \begin{bmatrix} 4 & 1 & 1 \\ 2 & 4 & 1 \\ 0 & 1 & 4 \end{bmatrix}.$$

The matrix $A$ is not diagonalizable. For $k = 4$ the method (I) converges in 2 iterations and $e_2 = 4.13E-7$.

*Example* 3.

$$A = \begin{bmatrix} 0 & 0.07 & 0.27 & -0.33 \\ 1.31 & -0.36 & 1.21 & 0.41 \\ 1.06 & 2.86 & 1.49 & -1.43 \\ -2.64 & -1.84 & -0.24 & -2.01 \end{bmatrix}.$$

The eigenvalues of $A$ are 0.03, 3.03, $-1.97 \pm i$. Method (I) diverges as predicated by stability analysis.

For $k = 1$ the method (II) converges in 4 iterations and $e_4 = 1.12E-4$.

For $k = 2$ the method (II) converges in 3 iterations and $e_3 = 9.64E-3$.

For $k = 3$ the method (II) converges in 2 iterations and $e_2 = 7.92E-4$.

This example illustrates the correspondence between the parameter $k$ and the computational complexity of the method (II).

In [2] the following quadratically convergent method was proposed to find $A^{-1/2}$.

$$\text{(III)} \quad \begin{cases} P_0 = A, \quad Q_0 = I \\ P_{n+1} = 0.5(P_n + Q_n^{-1}) \\ Q_{n+1} = 0.5(Q_n + P_n^{-1}) \\ \lim\limits_{n \to \infty} Q_n = A^{-1/2} \end{cases}$$

In [1] the following quadratically convergent methods were pro posed to find $A^{-1/2}$.

$$\text{(IV)} \quad \begin{cases} X_0 = I \\ X_{n+1} = 2X_n(I + AX_n^2)^{-1} \\ \lim\limits_{n \to \infty} X_n = A^{-1/2} \end{cases}$$

$$\text{(V)} \quad \begin{cases} T_{n+1} = T_n(I + S_n), & T_0 = I \\ S_{n+1} = S_n^2(2I - S_n^2)^{-1} & S_0 = (I - A)(I + A)^{-1} \\ \lim\limits_{n \to \infty} T_n = A^{-1/2} \end{cases}$$

In [3] the following quadratically convergent method was proposed to compute $A^{-1/2}$ for the special case when $A$ is hermitian positive definite.

$$\text{(VI)} \quad \begin{cases} X_0 = \nu I \\ X_{n+1} = X_n + \frac{1}{4} X_n (I - X_n A X_n) + \frac{1}{4} (I - X_n A X_n) X_n \end{cases}$$

where $0 < \nu < \sqrt{3/\rho(A)}$ .

For the method (V) the cost per iteration by [3] is

(4.1) $3n^3$ flops.

*Example* 4.

$$A = \begin{bmatrix} 0.2 & 100 & 150 & 50 \\ 0 & 0.4 & 50 & 50 \\ 0 & 0 & 0.4 & 100 \\ 0 & 0 & 0 & 0.4 \end{bmatrix}.$$

Both methods (I) and (I) converge in 2 iterations and $e_2 = 0$.

For the method (III) it holds $e_5 = e_6 = e_7 = \ldots = 1.05$,

For the method (V) it holds $e_5 = e_6 = e_7 = \ldots = 0.95$.

We see that in this example the methods (I) and (II) are more precisely than the methods (III ) and (V).

*Example* 5.

$$A = \begin{bmatrix} 0.002 & 1 & 1.5 & 0.5 \\ 0 & 0.003 & 0.5 & 0.5 \\ 0 & 0 & 0.003 & 1 \\ 0 & 0 & 0 & 0.005 \end{bmatrix}.$$

For the method (IV) it holds $e_{10} = e_{11} = e_{12} = \ldots = 1$.

For $k = 4$ the method (II ) converges in 3 iterations, $e_3 = 8.72E - 3$.

We see that in this example the method (II) is more precisely than the method (IV).

*Example* 6. Let $A$ be $10\times10$ symmetric positive definite matrix defined by

$$a_{ij} = \begin{cases} 100+ij & \text{if } i=j \\ ij & \text{if } i\neq j \end{cases}$$

For $k=3$ the method (I) converges in 3 iterations, $e_3 = 6.95E-6$.
Computation of $X_3$ requires by (3.10) approximately 10500 flops (for 3 iterations).
For the method (VI) the best results are obtened in the following cases:
for $\nu = 0.05$ the method (VI) converges in 6 iterations, $e_6 = 5.5E-7$
for $\nu = 0.06$ the method (VI) converges in 5 iterations, $e_5 = 6.03E-7$
for $\nu = 0.07$ the method (VI) converges in 6 iterations, $e_6 = 7.32E-7$.
For the method (VI), in the best case ($\nu = 0.06$), computation of $X_5$ requires by (4.1) approximately 15000 flops for 5 iteration. We see that in this example the method (I) converges approximate ly 1.5 times fasther than the method (VI).
If we use the Jacobi method and (3.11) then after 9 iterations we have $e_9 = 8.83E-7$, computation of $X_9$ requires by (3.12 approximately 13500 flops (for 9 iterations).
Single precision calculations were used for the 6 examples.

## References

[1] N. Sherif, On the comutation of a matrix inverse square root, *Computing* 46(1991), 295-305..
[2] N.J. Higham, Newtons method for matrix square root, *Math. Comp.* 46(1986), 537-549.
[3] B. Philippe, An algorithm to improve nearly orthonormal set of vectors on a vector processor, *SIAM Alg. Disc. Math.* 8(3)(1987), 396.
[4] H.J. Higham, Computing real square roots of a real matrix, *Linear Algebra Appl.* 88/89(1987), 405-430.

[ 5] P. Lancaster, Theory of matrices, *Academic Press*, New York 1969.
[ 6] KyFan, A.J. Hoffman, Some metric inequalities in the space of matrices, *Proc. Amer. Math. Soc.* 6,111(1955).
[ 7] G.W. Stewart, Introduction to matrix computation, Academic Press, New York 1974.

(University of Novi Sad, Technical Faculty „M. Pupin", 23000 Zrenjanin, Yugosławia).
Received on 4.11.1994 and, in revised form, on 3.2.1995.