Applications of Mathematics

Slobodan Lakić

A one parameter method for the matrix inverse square root

Applications of Mathematics, Vol. 42 (1997), No. 6, 401–410

Persistent URL: http://dml.cz/dmlcz/134366

A ONE PARAMETER METHOD FOR THE MATRIX INVERSE SQUARE ROOT

Slobodan Lakić, Zrenjanin

(Received August 23, 1996)

Abstract. This paper is motivated by the paper [3], where an iterative method for the computation of a matrix inverse square root was considered. We suggest a generalization of the method in [3]. We give some sufficient conditions for the convergence of this method, and its numerical stabillity property is investigated. Numerical examples showing that sometimes our generalization converges faster than the methods in [3] are presented.

Keywords: Newton method, matrix inverse square root

MSC 2000: 65F30

1. Introduction

An inverse square root of a matrix $A \in C^{m,m}$ is a solution $X \in C^{m,m}$ of the matrix equation $AX^2 = I$, and is denoted by $X = A^{-1/2}$. As was shown in [2] the inverse square root X of A always exists for a nonsingular matrix A. The inverse square root of a matrix has applications in the computation of an optimal symmetric orthogonalization of a set of vectors [1], theory of oscillations [2], etc.

We consider the scalar equations

$$az^2 = 1, \tag{1}$$

where $a \in C$ and $a \neq 0$. The equation (1) is equivalent to the equation

$$f(z) \equiv \frac{1}{az} - z = 0. \tag{2}$$

Applying Newton's method

$$z_{n+1} = z_n - \frac{f(z)}{f'(z)}$$

to the equation (2) we obtain

$$z_{n+1} = \frac{z_n}{\frac{1}{2} + \frac{1}{2}az_n^2}. \tag{3}$$

The denominator in (3), it is the easy to see, is the arithmetic mean of the line segment with the endpoints 1 and az_n^2. Having in mind this fact we can define the following iterative method with a paremeter $\alpha \in (0,1)$:

$$z_{n+1} = \frac{z_n}{1 - \alpha + \alpha a z_n^2}.$$

For simplicity of further analysis let $\alpha = \frac{1}{1+\varphi},\ \varphi \in \mathbb{R}$. Then

$$z_{n+1} = \frac{(1+\varphi)z_n}{\varphi + az_n^2}. \tag{4}$$

2. Computation of $A^{-1/2}$

We define the following matrix sequence $\{X_n\}$ based on the scalar sequence defined by (4).

Method (I):

$$\begin{aligned} X_0 &= I, \\ X_{n+1} &= (1+r)X_n(rI + AX_n^2)^{-1}, \quad r \in \mathbb{R}. \end{aligned}$$

First we give a theorem.

Theorem 1. *Let $w \in C$, $w \neq 0$, $\arg(w) \neq \pi$. We define the scalar sequence $\{z_n\}$ by*

$$\begin{aligned} z_0 &= 1, \\ z_{n+1} &= \frac{(1+r)z_n}{r + wz_n^2}, \quad r > 0. \end{aligned}$$

Then $\lim\limits_{n\to\infty} z_n = 1/\sqrt{w}$, where $\sqrt{w}$ is the principal square root of w.

P r o o f. First we prove that the sequence $\{z_n\}$ is well defined, i.e.

$$r + wz_n^2 \neq 0, \quad n = 0, 1, 2, \ldots.$$

For $n=0$ if $r+wz_0=0$, i.e. $w=-r$, we obtain $\arg(w)=\pi$ which is a contradiction. If we set $t_n=wz_n^2$ and assume that $r+wz_n^2=r+t_n\neq 0$, then

$$r+wz_{n+1}^2=r+\frac{(1+r)^2t_n}{(r+t_n)^2}=\frac{rt_n^2+(3r^2+2r+1)t_n+r^3}{(r+t_n)^2}.$$

We assume that $r+wz_{n+1}^2=0$, i.e.

$$rt_n^2+(3r^2+2r+1)t_n+r^3=0.$$

The solutions of the preceding equation are given as follows:

$$t_n=wz_n^2=-\frac{3r^2+2r+1}{2r}\left(1\pm\sqrt{1-\left(\frac{2r^2}{3r^2+2r+1}\right)^2}\right).$$

So $wz_n^2<0$ and $\arg(wz_n^2)=\pi$, a contradiction. Finally, $r+wz_n^2\neq 0$ and the sequence $\{z_n\}$ is well defined.

Now, we prove that $Re(\sqrt{w}z_n)>0$, i.e. $\sqrt{w}z_n$ lies in the right halfplane. For $n=0$, $\sqrt{w}z_0=\sqrt{w}$ lies in the right halfplane because $\sqrt{w}$ is the principal value. We assume that $\sqrt{w}z_n$ lies in the right halfplane. Then $\sqrt{w}z_n\neq 0$ and $r/\sqrt{w}z_n$ lies in the right halfplane. By the definition of the sequence $\{z_n\}$ we have

$$\sqrt{w}z_{n+1}=\frac{1+r}{r/\sqrt{w}z_n+\sqrt{w}z_n},$$

which means that $\sqrt{w}z_{n+1}$ lies in the right halfplane. Since $\sqrt{w}z_n$ lies in the right halfplane, hence

$$\sqrt{w}z_n\neq -r,\quad z_n+\frac{r}{\sqrt{w}}\neq 0.$$

Now we have

$$\begin{aligned}\frac{z_n-\frac{1}{\sqrt{w}}}{z_n+\frac{1}{\sqrt{w}}}&=-\frac{\left(z_{n-1}-\frac{1}{\sqrt{w}}\right)\left(z_{n-1}-\frac{r}{\sqrt{w}}\right)}{\left(z_{n-1}+\frac{1}{\sqrt{w}}\right)\left(z_{n-1}+\frac{r}{\sqrt{w}}\right)}\\&=(-1)^n\frac{\sqrt{w}-1}{\sqrt{w}+1}\prod_{i=0}^{n-1}\frac{z_i-\frac{r}{\sqrt{w}}}{z_i+\frac{r}{\sqrt{w}}}.\end{aligned}$$

Let

$$\left|\frac{z_k-\frac{r}{\sqrt{w}}}{z_k+\frac{r}{\sqrt{w}}}\right|=\max_{0\leqslant i\leqslant n-1}\left|\frac{z_i-\frac{r}{\sqrt{w}}}{z_i+\frac{r}{\sqrt{w}}}\right|\quad(k\in\{0,1,\ldots,n-1\}).$$

Then

$$\left|\frac{z_n - \frac{1}{\sqrt{w}}}{z_n + \frac{1}{\sqrt{w}}}\right| \leqslant \left|\frac{\sqrt{w}-1}{\sqrt{w}+1}\right| \left|\frac{z_k\sqrt{w}-r}{z_k\sqrt{w}+r}\right|^n.$$

Since $r > 0$ and $z_k\sqrt{w}$ lies in the right halfplane, hence

$$\left|\frac{\sqrt{w}z_k - r}{\sqrt{w}z_k + r}\right| < 1$$

and

$$\lim_{n\to\infty} \left|\frac{\sqrt{w}z_k - r}{\sqrt{w}z_k + r}\right|^n = 0.$$

Finally,

$$\lim_{n\to\infty} \left|\frac{z_n - \frac{1}{\sqrt{w}}}{z_n + \frac{1}{\sqrt{w}}}\right| = 0, \quad \lim_{n\to\infty} z_n = \frac{1}{\sqrt{w}}.$$

□

In the sequel we need the following definition: The matrix A is diagonalizable if there exists a nonsingular matrix V such that $V^{-1}AV = D$ where $D = \operatorname{diag}\{a_1, \ldots, a_n\}$.

Theorem 2. *Let $A \in C^{m,m}$ be nonsingular and diagonalizable. We assume that A has no negative real eigenvalues and $r > 0$. Then $\lim_{n\to\infty} X_n = A^{-1/2}$, where $A^{-1/2}$ is the matrix principal inverse square root of A.*

P r o o f. We define $K_n = V^{-1}X_nV$. From (I) it follows that

$$\begin{aligned} K_0 &= I, \\ K_{n+1} &= (1+r)K_n(rI + DK_n^2)^{-1}. \end{aligned} \tag{5}$$

The sequence (5) is a sequence of diagonal matrices $K_n = \operatorname{diag}\{k_1^{(n)}, \ldots, k_m^{(n)}\}$. The equation (5) is equivalent to m scalar equations

$$\begin{aligned} k_i^{(0)} &= 1, \\ k_i^{(n+1)} &= \frac{(1+r)k_i^{(n)}}{r + a_i(k_i^{(n)})^2}. \end{aligned}$$

Application of Theorem 1 yields $\lim_{n\to\infty} k_i^{(n)} = 1/\sqrt{a_i}$. Now,

$$\lim_{n\to\infty} K_n = D^{-1/2} \quad \text{and} \quad \lim_{n\to\infty} X_n = A^{-1/2}.$$

□

Stability analysis.

We apply a technique presented in [3] and assume that in the iterative step n of method (I), we compute an approximation Y_n of the exact matrix X_n. From the proof of Theorem 2 we conclude that there exists ε_i such that

$$k_i^{(n)} = \varepsilon_i + 1/\sqrt{a_i}.$$

We will assume that

$$E_n = Y_n - X_n = O(\varepsilon),$$

where $\varepsilon \geqslant \max\limits_i |\varepsilon_i|$.

Using the following result in [4]:

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

we have

$$\begin{aligned} Y_{n+1} &= (1+r)(X_n+E_n)(rI + A(X_n+E_n)^2)^{-1} \\ &= (1+r)(X_n+E_n)((rI+AX_n^2)^{-1} - (rI+AX_n^2)^{-1}A(X_nE_n+E_nX_n) \\ &\quad \times (rI+AX_n^2)^{-1}) + O(\varepsilon^2). \end{aligned}$$

If we define $F_n = V^{-1}E_nV$, then

$$\begin{aligned} F_{n+1} &= (1+r)F_n(rI+DK_n^2)^{-1} - (1+r)K_n(rI+DK_n^2)^{-1}D(K_nF_n+F_nK_n) \\ &\quad \times (rI+DK_n^2)^{-1} + O(\varepsilon^2). \end{aligned}$$

Writing the above equation elementwise we get

$$f_{ij}^{(n+1)} = \frac{1+r}{r+a_j(k_j^{(n)})^2} f_{ij}^{(n)} - \frac{(1+r)k_i^{(n)}a_i(k_i^{(n)}+k_j^{(n)})}{(r+a_i(k_i^{(n)})^2)(r+a_j(k_j^{(n)})^2)} f_{ij}^{(n)} + O(\varepsilon^2).$$

Moreover,

$$f_{ij}^{(n+1)} = c_{ij}^{(n)} f_{ij}^{(n)} + O(\varepsilon^2), \quad \text{where} \quad c_{ij}^{(n)} = \frac{r - \sqrt{a_i/a_j}}{1+r}.$$

To ensure the numerical stability of the method (I) we require

$$\left| r - \sqrt{a_i/a_j} \right| \leqslant 1 + r. \tag{6}$$

If the eigenvalues of A are real and positive then (6) is equivalent to

$$0 < \frac{a_i}{a_j} \leqslant (2r+1)^2. \tag{7}$$

The method (I), it is easy to see, is one parameter generalization of the method in [3]. Namely, for $r = 1$ we obtain the following method considered in [3].

Method SH1:

$$\begin{aligned} X_0 &= I, \\ X_{n+1} &= 2X_n(I + AX_n^2)^{-1}. \end{aligned}$$

Since the method SH1 is not stable in [3] the following alternative locally stable method is proposed.

Method SH2:

$$\begin{aligned} T_0 &= I, \\ S_0 &= (I - A)(I + A)^{-1}, \\ T_{n+1} &= T_n(I + S_n), \\ S_{n+1} &= S_n^2(2I - S_n^2)^{-1}, \\ T_n &\to A^{-1/2}. \end{aligned}$$

In the sequel we are concerned with a choice of r ($r \neq 1$) to ensure better stability and faster convergence than for the method SH1.

Choice of r. Let $A \in C^{m,m}$ be diagonalizable and let us assume that the eigenvalues of A are real and $a_i \geqslant 1$, ($a_i \in \sigma(A)$).

For the scalar sequence $\{k_i^{(n)}\}$ in Theorem 2 we have

$$\left| k_i^{(n)} - \frac{1}{\sqrt{a_i}} \right| = \frac{\left| \sqrt{a_i} k_i^{(n-1)} - r \right|}{r + a_i (k_i^{(n-1)})^2} \left| k_i^{(n-1)} - \frac{1}{\sqrt{a_i}} \right|. \tag{8}$$

Now, we consider the case $r \geqslant \sqrt{\varrho(A)}$. First we prove that

$$\frac{1}{\sqrt{a_i}} \leqslant k_i^{(n)} \leqslant 1.$$

For $n = 0$ it is obvious. From the induction hypothesis and the inequalities

$$r \geqslant \sqrt{\varrho(A)} \geqslant \sqrt{a_i} \geqslant 1$$

we have

$$\frac{1}{\sqrt{a_i}} \leqslant k_i^{(n)} \leqslant 1 \leqslant \frac{r}{\sqrt{a_i}}, \quad \text{i.e.} \quad \frac{1}{\sqrt{a_i}} \leqslant k_i^{(n)} \leqslant \frac{r}{\sqrt{a_i}}.$$

Now

$$k_i^{(n+1)} \geqslant \frac{1}{\sqrt{a_i}}.$$

Since $a_i(k_i^{(n)})^2 \geqslant 1$ and $k_i^{(n)} \leqslant 1$, we have

$$k_i^{(n+1)} \leqslant \frac{1+r}{r + a_i(k_i^{(n)})^2}, \quad k_i^{(n)} - \frac{r}{\sqrt{a_i}} \leqslant 0.$$

Finally,

$$\frac{r - \sqrt{a_i}k_i^{(n)}}{r + a_i(k_i^{(n)})^2} \leqslant \frac{r-1}{r+1} = f(r). \tag{9}$$

Since the matrix A is diagonalizable, there exists a matrix norm such that

$$\|X_n - A^{-1/2}\| = \varrho(X_n - A^{-1/2}) = \varrho(K_n - D^{-1/2}) = \max_i \left| k_i^{(n)} - \frac{1}{\sqrt{a_i}} \right|.$$

Using (8) and (9) we obtain

$$\|X_n - A^{-1/2}\| \leqslant f(r)\|X_{n-1} - A^{-1/2}\| \leqslant \ldots \leqslant (f(r))^n \|I - A^{-1/2}\|.$$

From the last inequality we see that the method (I) has the best rate of convergence if we minimize $f(r)$. Since f is increasing on the interval $\left[\sqrt{\varrho(A)}, \infty\right)$, hence f attains its minimum on this interval for $r = \sqrt{\varrho(A)}$. It is easy to see that for $r = \sqrt{\varrho(A)}$, $a_i \geqslant 1$ the inequality (7) is also valid because $a_i/a_j \leqslant \varrho(A)$. So, this choice of r ensures the local stability of the method (I), which is not the case for $r = 1$. In the next section we shall show some other adwantages of this choice of r.

The operation counts for one stage of each iteration, measured in flops are given in the following table:

flops per stage	general	hermitian pos. definite
method (I)	$3m^3$	$3m^3/2$

From [3], if the matrix A is hermitian positive definite, then the costs for the methods SH1 and SH2 are approximately $3m^3/2$ and $2m^3$ flops per iteration, respectively. If the matrix A is general the costs for the methods SH1 and SH2 are $3m^3$ and $4m^3$ flops per iteration, respectively. We see that the costs per iteration for the method (I) are equal to the costs per iteration for the method SH1, and the costs per iteration for the method (I) are less than the costs per iteration for the method SH2.

R e m a r k. If the eigenvalues of $A \in C^{m,m}$ are real and

$$0 < a_1 \leqslant a_2 \leqslant \ldots \leqslant a_m,$$

where $a_1 \neq 1$, then for the eigenvalues $b_1, \ldots, b_m$ of the matrix $B = \frac{1}{a_1}A$ we have

$$1 = b_1 \leqslant b_2 \leqslant \ldots \leqslant b_m.$$

We define the following method:

Method (II):

$$\begin{aligned} Y_0 &= I, \\ Y_{n+1} &= (1+r)Y_n(rI + BY_n^2)^{-1}, \\ X_n &= \frac{1}{\sqrt{a_1}}Y_n, \end{aligned}$$

where $a_1 = \min\limits_i a_i$, $a_i \in \sigma(A)$. Then the matrix sequence $\{X_n\}$ converges to $A^{-1/2}$ and

$$\|I - BY_n^2\| = \|I - AX_n^2\|.$$

3. Numerical examples

In this section we have used the Frobenius matrix norm

$$\|A\| = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{m}|a_{ij}|^2},$$

and the error $e_n = \|I - AX_n^2\|$.

E x a m p l e 1. Let A be the inverse Hilbert matrix of order 4,

$$A = invhilb(4) = \begin{bmatrix} 16 & -120 & 240 & -140 \\ -120 & 1200 & -2700 & 1680 \\ 240 & -2700 & 6480 & -4200 \\ -140 & 1680 & -4200 & 2800 \end{bmatrix},$$

$$\sigma(A) = \{0.66657, 5.9122, 148.40596, 10341.0524\},$$

In this example the method SH1 is not stable and diverges.

For the method SH2 we have obtained $e_9 = e_{10} = e_{11} = \ldots = 0.5$ and the error cannot be decreased by further iterating.

For the method (II) where $B = \frac{1}{0.66657}A$, $r = \sqrt{\varrho(B)} = 124.55$, after 450 iterations we have $e_{450} = 9.8E - 4$. In this example the method (II) is more precise than the method SH2, while the method SH1 diverges.

E x a m p l e 2. Let A be the Pascal matrix of order 6,

$$A = pascal(6) = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 & 5 & 6 \\ 1 & 3 & 6 & 10 & 15 & 21 \\ 1 & 4 & 10 & 20 & 35 & 56 \\ 1 & 5 & 15 & 35 & 70 & 126 \\ 1 & 6 & 21 & 56 & 126 & 252 \end{bmatrix},$$

$$\sigma(A) = \{0.003, 0.064, 0.489, 2.044, 15.55, 324.4\}.$$

In this example the method SH1 is not stable and diverges.

For the method SH2 we obtain $e_6 = e_7 = e_8 = \ldots = 0.168$ and the error cannot be decreased by further iterating.

For the method (II) where $B = \frac{1}{0.003}A$, $r = \sqrt{\varrho(B)} = 332.868$, after 1000 iterations we have $e_{1000} = 4.84E - 3$.

E x a m p l e 3.

$$A = 3I + hadamard(4) = \begin{bmatrix} 4 & 1 & 1 & 1 \\ 1 & 2 & 1 & -1 \\ 1 & 1 & 2 & -1 \\ 1 & -1 & -1 & 4 \end{bmatrix},$$

$$\sigma(A) = \{1, 5\}$$

where hadamard (4) is the Hadamard matrix of order 4. The method (I) where $r = \sqrt{\varrho(A)} = \sqrt{5}$ converges within 1 iteration and $e_1 = 5.41E - 7$, while the method SH1 converges within 5 iterations and $e_5 = 5.27E - 7$. In this example the method (I) converges faster than the method SH1.

E x a m p l e 4.

$$A = \begin{bmatrix} 0.003 & 0.01 & 1.5 & 0.5 \\ 0 & 0.003 & 0.5 & 0.5 \\ 0 & 0 & 0.003 & 1 \\ 0 & 0 & 0 & 0.0033 \end{bmatrix},$$

$$\sigma(A) = \{0.003, 0.0033\}.$$

The method (II) where $B = \frac{1}{0.003}A$, $r = \sqrt{\varrho(B)} = 1.0488$ converges within 6 iterations and $e_6 = 4.26E - 3$, while the method SH1 converges within 9 iterations and $e_9 = 8.96E - 3$. In this example the method (II) converges faster than the method SH1.

Single precision calculations were used for all the four examples.

Acknowledgment. I thank the referee for his suggestions.

References

[1] *A.J. Hoffman, K. Fan*: Some metric inequalities in the space of matrices. Proc. Amer. Math. Soc. *6* (1955), 111–116.
[2] *P. Lancaster*: Theory of Matrices. Academic Pres, New York, 1969.
[3] *N. Sherif*: On the computation of a matrix inverse square root. Computing *46* (1991), 295–305.
[4] *G.W. Stewart*: Introduction to Matrix Computation. Academic Pres, New York, 1974.

Author's address: *Slobodan Lakić*, University of Novi Sad, Technical Faculty "Mihajlo Pupin", 23000 Zrenjanin, Yugoslavia.