Serdica
Mathematical Journal
Сердика
Математическо списание

Serdica Math. J. **21** (1995), 335-344

Serdica
Mathematical Journal

Institute of Mathematics
Bulgarian Academy of Sciences

AN ITERATIVE METHOD FOR THE MATRIX PRINCIPAL n-th ROOT

Slobodan Lakić

Communicated by R. Van Keer

In this paper we give an iterative method to compute the principal n-th root and the principal inverse n-th root of a given matrix. As we shall show this method is locally convergent. This method is analyzed and its numerical stability is investigated.

1. Introduction. Computation methods for the n-th root of some matrices have been proposed in [1], [2], [3], etc. In Section 2 an iterative method with high convergence rates is developed. In Section 3 we shall show that this method is locally stable. In Section 4 we illustrate the performance of the method by numerical examples.

Let $a = re^{it} \in \mathbb{C}$, where $r, t \in \mathbb{R}$ and $r \geq 0$, $t \in (-\pi, \pi]$.

Definition 1.1. *The principal n-th root of a is defined as* $a^{1/n} = r^{1/n}e^{it/n}$, *where the number* $r^{1/n}$ *is the unique real and non-negative n-th root of r.*

Let $A \in \mathbb{C}^{m,m}$, $\sigma(A) = \{a_i,\ i = 1, \ldots, m\}$, $a_i \neq 0$, where a_i are the eigenvalues of A.

Definition 1.2. *The principal inverse n-th root of A is defined as* $X = A^{-1/n} \in \mathbb{C}^{m,m}$ *and* $AX^n = I$, *each eigenvalue of* $A^{-1/n}$ *is the principal n-th root of each* $1/a_i$.

1991 *Mathematics Subject Classification:* 65F30
Key words: iterative methods, matrix function

Definition 1.3. *The principal n-th root of A is defined as* $X = A^{1/n} \in \mathbb{C}^{m,m}$ *and* $X^n = A$*, each eigenvalue of* $A^{1/n}$ *is the principal n-th root of each* a_i.

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

Theorem 2.1. *Let* $f_k(z) = (1-z)^{-1/k}$*, where* $(1-z)^{1/k}$ *is the principal k-th root of* $1-z$*,* $k \in \mathbb{N}$*,* $k \geq 2$*,* $z \in \mathbb{C}$*,* $j \in \mathbb{N}$*,* $R_{j-1}(z) = \sum\limits_{i=0}^{j-1} b_i z^i$*,* $b_i = f_k^{(i)}(0)/i!$*. Then it holds*

$$1-(1-z)R_{j-1}^k(z) = z^j \sum_{i=0}^{(k-1)(j-1)} c_{i,k} z^i \tag{2.1}$$

for some positive constants $c_{i,k} = c_{i,k}(k,j)$*,*

$$i = 0, \ldots, (k-1)(j-1) \quad \text{and} \quad \sum_{i=0}^{(k-1)(j-1)} c_{i,k} = 1. \tag{2.2}$$

P r o o f. By mathematical induction for $j = 1$

$$1-(1-z)R_{j-1}^k(z) = 1-(1-z) = z = zc_0$$

where $c_0 = 1$.

We assume that (2.1) holds for $k \geq 2$. Then

$$\begin{aligned}
&1-(1-z)R_j^k(z) = 1-(1-z)(R_{j-1}(z)+b_j z^j)^k \\
&= 1-(1-z)\sum_{m=0}^{k}\binom{k}{m} R_{j-1}^m(z) b_j^{k-m} z^{(k-m)j} \\
&= 1-(1-z)R_{j-1}^k(z) - (1-z)b_j^k z^{kj} - k(1-z)b_j^{k-1} z^{(k-1)j} R_{j-1}(z) \\
&+ \sum_{m=2}^{k-1}\binom{k}{m} b_j^{k-m} z^{(k-m)j}\left(-1+z^j \sum_{i=0}^{(m-1)(j-1)} c_{i,m} z^i\right) \\
&= -z^j \sum_{m=0}^{k-1}\binom{k}{m} b_j^{k-m} z^{(k-m-1)j} - z^j k b_j^{k-1} \sum_{m=1}^{j-1} b_m z^{(k-2)j+m} \\
&+ z^j\left[b_j^k z^{1+j(k-1)} + k b_j^{1+j(k-2)} R_{j-1}(z) + \sum_{m=2}^{k}\binom{k}{m} b_j^{k-m} z^{(k-m)j} \sum_{i=0}^{(m-1)(j-1)} c_{i,m} z^i\right] \\
&= z^j\Bigg[b_j^k z^{1+j(k-1)} + b_j^{k-1} z^{(k-1)j}(kb_{j-1} - b_j)
\end{aligned}$$

$$+kb_j^{k-1}\sum_{i=1}^{j-1}(b_{i-1}-b_i)z^{i+j(k-2)}+\sum_{m=2}^{k-1}b_j^{k-m}z^{j(k-m)}\left(\binom{k}{m}c_{0,m}-\binom{k}{m-1}b_j\right)$$
$$\left.+(c_{0,k}-kb_j)+\sum_{m=2}^{k}\binom{k}{m}b_j^{k-m}z^{j(k-m)}\sum_{i=1}^{(m-1)(j-1)}c_{i,m}z^i\right]$$

Now we prove

$$c_{0,m}=\frac{mf_m^{(j)}(0)}{j!}. \tag{2.3}$$

From (2.1) it follows that

$$R_{j-1}^{(j)}(z)=(h_m(z)f_m(z))^{(j)} \tag{2.4}$$

where $h_m(z)=g_m(T(z))$, $g_m(T)=T^{1/m}$ and $T(z)=1-z^j\sum_{i=0}^{(m-1)(j-1)}c_{i,m}z^i$. From (2.4) it follows that

$$0=f_m^{(j)}(z)+\sum_{i=1}^{j}\binom{j}{i}h_m^{(i)}(z)f_m^{(j-i)}(z).$$

Since

$$h_m^{(i)}(z)=\sum_{n_1,\dots,n_i}\frac{i!}{n_1!n_2!\dots n_i!}\,g_m^{(s)}(T)\prod_{k=1}^{i}\left(\frac{T^{(k)}(z)}{k!}\right)^{n_k},$$
$$s=n_1+n_2+\ldots+n_i,$$

where $n_1,\dots,n_i\geq 0$ are the integer solutions of the equation

$$n_1+2n_2+\cdots+in_i=i,$$

and since $T^{(i)}(0)=0$ for $1\leq i\leq j-1$, we have $h_m^{(i)}(0)=0$ for $1\leq i\leq j-1$, and finally $h_m^{(j)}(0)=g_m'(1)T^{(j)}(0)$. Now $0=f_m^{(j)}(0)-\dfrac{j!c_{0,m}}{m}$ i.e. (2.3).

Since $kb_{j-1}-b_j=\dfrac{(k-1)(kj+1)\prod_{i=0}^{j-1}((i-1)k+1)}{j!k^j}\geq 0$ for $k\in\mathbb{N}$,
$b_{i-1}-b_i=\dfrac{k-1}{i!k^i}\geq 0$ for $1\leq i\leq j-1$ and $k\in\mathbb{N}$,

$$\binom{k}{m}c_{0,m}-\binom{k}{m-1}b_j=\frac{k!}{j!(m-1)!(k-m)!}\left(\frac{\prod_{i=1}^{j-i}\left(\frac{1}{m}+i\right)}{m}-\frac{\prod_{i=1}^{j-i}\left(\frac{1}{k}+i\right)}{k(k-m+1)}\right)>0$$

for $k > m$, and $c_{0,k} - kb = 0$ we have $1-(1-z)R_j^k(z) = z^{j+1}\sum_{i=0}^{(k-1)j}\overline{c}_i z^i$ where $\overline{c}_0,\dots\overline{c}_{(k-1)j}$ are the positive constants. Setting $z=1$ gives (2.2). □

Theorem 2.2. *Let w be a complex number such that $w \neq 0$. We define the sequence $\{z_n\}$ by*

$$z_{n+1} = z_n \sum_{i=0}^{j-1} b_i(1-wz_n^k)^i \tag{2.5}$$

where b_i, k are as in Theorem 2.1, $j \in \mathbb{N}$, $j \geq 2$ and $|1-wz_0^k| < 1$. Then

$$|1-wz_n^k| \leq |1-wz_0^k|^{j^n} \tag{2.6}$$

and

$$\lim_{n\to\infty} z_n = \frac{1}{w^{1/k}} \tag{2.7}$$

where $w^{1/k}$ is the k-th principal root of w.

Proof. Using Theorem 2.1 we have

$$1-wz_1^k = (1-wz_0^k)^j \sum_{i=0}^{(k-1)(j-1)} c_{i,k}(1-wz_0^k)^i$$

and $|1-wz_1^k| \leq |1-wz_0^k|^j$.

Repeating this argument we have (2.6).

From (2.6) it holds $\lim_{n\to\infty}|1-wz_n^k| = 0$ i.e. (2.7). □

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

$$V^{-1}AV = D \tag{2.8}$$

where D=diag$\{a_1,\dots,a_m\}$ and $a_1,\dots,a_m$ are the eigenvalues of A.

We define the sequences $\{X_n\}$ and $\{S_n\}$ as follows

$$\text{(I)}\quad \begin{cases} X_{n+1} = X_n \sum_{i=0}^{j-1} b_i(I-S_n)^i & X_0 \in \mathbb{C}^{n,n}, \\ S_{n+1} = S_n \left[\sum_{i=0}^{j-1} b_i(I-S_n)^i\right]^k, & S_0 = AX_0^k, \end{cases}$$

where X_0 is a function of A, and j, k, b_i are as in Theorem 2.2.

Theorem 2.3. *Let $A \in \mathbb{C}^{m,m}$ be nonsingular and diagonalizable. Let $\{X_n\}$, $\{S_n\}$ be the sequences defined by* (I) *and*

$$\|I - S_0\| < 1 \tag{2.9}$$

Then $\lim_{n\to\infty} X_n = A^{-1/k}$, $\lim_{n\to\infty} S_n = I$, $\|I - AX_n^k\| = O(\|I - AX_{n-1}^k\|^j)$, where $A^{-1/k}$ is the principal inverse k-th root of A.

Proof. Let

$$L_n = V^{-1}X_nV, \quad H_n = V^{-1}S_nV. \tag{2.10}$$

Now

$$\begin{cases} L_{n+1} = L_n + \sum_{i=0}^{j-1} b_i(I - H_n)^i, & L_0 = V^{-1}X_0V \\ H_{n+1} = H_n \left[\sum_{i=0}^{j-1} b_i(I - S_n)^i\right]^k, & H_0 = DL_0^k. \end{cases} \tag{2.11}$$

From the equations (2.11) it follows that L_n and H_n are diagonal matrices. Let

$$L_n = \operatorname{diag}\{l_1^{(n)}, \ldots, l_m^{(n)}\}, \quad H_n = \operatorname{diag}\{h_1^{(n)}, \ldots, h_m^{(n)}\}.$$

Equation (2.11) is equivalent to m sequence of equations.

$$\begin{cases} l_i^{(n+1)} = l_i^{(n)} \sum_{m=0}^{j-1} b_m(1 - h_i^{(n)})^m, & l_i^{(0)} \in \mathbb{C}, \\ h_i^{(n+1)} = h_i^{(n)} \left[\sum_{m=0}^{j-1} b_m(1 - h_i^{(n)})^m\right]^k, & h_i^{(0)} = a_i l_i^{(0)}. \end{cases} \tag{2.12}$$

From (2.12) one can show that

$$l_i^{(n+1)} = l_i^{(n)} \sum_{m=0}^{j-1} b_m \left(1 - a_i(l_i^{(n)})^k\right)^m, \quad l_i^{(0)} \in \mathbb{C}. \tag{2.13}$$

Since the matrix $I - AX_0^k$ is diagonalizable, its matrix norm satisfies

$$\|I - AX_0^k\| = \rho(I - AX_0^k) = \rho(I - DL_0^k) = \|I - DL_0^k\|.$$

So we have

$$\|I - DL_0^k\| < 1. \tag{2.14}$$

From (2.14) it follows that

$$|1 - a_i(l_i^{(0)})^k| < 1, \quad i = 1, \ldots, m. \tag{2.15}$$

From (2.13) and (2.15) using Theorem 2.2 it follows that

$$\lim_{n\to\infty} l_i^{(n)} = a_i^{-1/k}, \quad i = 1, \ldots, m. \tag{2.16}$$

From (2.12) and (2.16) it follows that

$$\lim_{n\to\infty} h_i^{(n)} = 1, \quad i = 1, \ldots, n.$$

So,

$$\lim_{n\to\infty} L_n = D^{-1/k}, \qquad \lim_{n\to\infty} H_n = I. \tag{2.17}$$

From (2.17), (2.10) and (2.8) it follows

$$\lim_{n\to\infty} X_n = A^{-1/k}, \qquad \lim_{n\to\infty} S_n = I.$$

From (2.13) using Theorem 2.1 it follows

$$\begin{aligned} 1 - a_i(l_i^{(n)})^k &= (1 - a_i(l_i^{(n-1)})^k)^j \sum_{m=0}^{(j-1)(k-1)} c_{m,k}(1 - a_i(l_i^{(n-1)})^k)^m, \\ I - DL_n^k &= (I - DL_{n-1}^k)^j \sum_{m=0}^{(j-1)(k-1)} c_{m,k}(I - DL_{n-1}^k)^m. \end{aligned}$$

So, $I - AX_n^k = (I - AX_{n-1}^k)^j \sum_{m=0}^{(j-1)(k-1)} c_{m,k}(I - AX_{n-1}^k)^m$. Taking the norm of the above equation, the bound in the theorem is established. □

Remark. If $S_0 = A^{-1}X_0^k$ then $\lim_{n\to\infty} X_n = A^{1/k}$.

Theorem 2.4. *Let $A \in \mathbb{C}^{n,n}$ be a hermitian positive definite matrix, $X_0 = sI$, $s \in \mathbb{R}$,*

$$0 < s < \left(\frac{2 \min\limits_{1\le i\le n} a_i}{\rho^2(A)} \right)^{1/k},$$

then $\lim_{n\to\infty} X_n = A^{-1/k}$, *where* $A^{-1/k}$ *is the principal inverse* k*-th root of* A.

Proof. It is known that each hermitian matrix is diagonalizable. Then the matrix norm of $I - s^kA$ satisfies $\|I - s^kA\| = \rho(I - s^kA) = \max_{1\le i\le n} |1 - s^k a_i| = \max_{1\le i\le n} \sqrt{1 - 2s^k a_i + s^{2k} a_i^2} \le \sqrt{1 - 2s^k \min_{1\le i\le n} a_i + s^{2k}\rho^2(A)} < 1$. □

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

We define $\tilde{P}_n = V^{-1}P_nV$, $\tilde{Q}_n = V^{-1}Q_nV$. Using the perturbation result in [4]

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

from $\tilde{X}_{n+1} = \tilde{X}_n \sum_{i=0}^{j-1} b_i(I - \tilde{S}_n)^i$ and $\tilde{S}_{n+1} = \tilde{S}_n \left[\sum_{i=0}^{j-1} b_i(I - \tilde{S}_n)^i\right]^k$ direct calculations give

$$\begin{aligned}
\tilde{P}_{n+1} &= -L_n \sum_{i=1}^{j-1} b_i \sum_{m=0}^{i-1} (I - H_n)^m \tilde{Q}_n (I - H_n)^{i-m-1} + \tilde{P}_n \sum_{i=0}^{j-1} b_i (I - h_n)^i + O(\varepsilon^2)\\
\tilde{Q}_{n+1} &= -H_n \left[\sum_{l=0}^{k-1} \left(\sum_{i=0}^{j-1} b_i (I - H_n)^i\right)^l\right] \left[\sum_{i=1}^{j-1} b_i \sum_{m=0}^{i-1} (I - H_n)^m \tilde{Q}_n (I - H_n)^{i-m-1}\right]\\
&\quad \times \left[\sum_{i=0}^{j-1} b_i (I - H_n)^i\right]^{k-l-1} + \tilde{Q}_n \left[\sum_{i=0}^{j-1} b_i (I - H_n)^i\right]^k + O(\varepsilon^2).
\end{aligned}$$

Writing the above equations element-wise we have, $r, s = 1, \ldots, n$,

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

where

$$\begin{aligned}
v_{rs}^{(n)} &= -l_r^{(n)} \sum_{i=1}^{j-1} b_i \sum_{m=0}^{i-1} \left(1 - h_r^{(n)}\right)^m \left(1 - h_s^{(n)}\right)^{i-m-1},\\
g_{rs}^{(n)} &= \sum_{i=0}^{j-1} b_i (1 - h_s^{(n)})^i,\\
d_{rs}^{(n)} &= -h_r^{(n)} \left[\sum_{l=0}^{k-1} \left(\sum_{i=0}^{j-1} b_i \left(1 - h_r^{(n)}\right)^i\right)^l\right] \left[\sum_{i=1}^{j-1} b_i \sum_{m=0}^{i-1} \left(1 - h_r^{(n)}\right)^m \left(1 - h_s^{(n)}\right)^{i-m-1}\right]
\end{aligned}$$

$$\times \left[\sum_{i=0}^{j-1} b_i \left(1-h_s^{(n)}\right)^i\right]^{k-l-1} + \left[\sum_{i=0}^{j-1} b_i \left(1-h_s^{(n)}\right)^i\right]^{k}.$$

Let

$$e_{rs}^{(n)} = \begin{bmatrix} \tilde{q}_{rs}^{(n)} \\ \tilde{p}_{rs}^{(n)} \end{bmatrix}.$$

Now we have

$$e_{rs}^{(n+1)} = W_{rs}^{(n)} e_{rs}^{(n)} + O(\varepsilon^2)$$

where

$$W_{rs}^{(n)} = \begin{bmatrix} d_{rs}^{(n)} & 0 \\ v_{rs}^{(n)} & g_{rs}^{(n)} \end{bmatrix}.$$

Since $\lim_{n\to\infty} d_{rs}^{(n)} = 1 - kb_1 = 0$, $\lim_{n\to\infty} g_{rs}^{(n)} = 1$, $\lim_{n\to\infty} v_{rs}^{(n)} = \dfrac{-1}{ka_i^{1/k}}$, we can write $W_{rs}^{(n)}$ as $W_{rs}^{(n)} = W_{rs} + O\left(\varepsilon^{(n)}\right)$

$$W_{rs} = \begin{bmatrix} 0 & 0 \\ \dfrac{-1}{ka_i^{1/k}} & 1 \end{bmatrix},$$

where $\varepsilon^{(n)}$ is sufficiently small for large n.

The matrix W_{rs} has eigenvalues 0 and 1, let z_0 and z_1 be the corresponding eigenvectors, so

$$e_{rs}^{(n)} = u_0^{(n)} z_0 + u_1^{(n)} z_1.$$

For sufficiently small ε and large n we have

$$e_{rs}^{(n+m)} \cong W_{rs}^m e_{rs}^{(n)} = u_1^{(n)} z_1 \qquad m = 1, 2, \dots.$$

Consequently $\|e_{rs}^{(n+m)}\| = \|e_{rs}^{(n+1)}\|$ and method (I) is locally stable.

The usual assumption that the multiplication of two $n \times n$ matrices requires n^3 flops.

For method (I) if the matrix A is general, the cost is approximatelly

$$(j - 1 + B_k + \lfloor \log_2 k \rfloor) n^3$$

flops per iteration, where B_k=number of ones in binary representation of k, $\lfloor \log_2 k \rfloor$ denotes the largest integer not exceeding $\log_2 k$, and the number of flops is determined as follows

(1) $(j-2)n^3$ flops to find $\sum_{i=0}^{j-1} b_i (I - S_n)^i$

(2) $(B_k + \lfloor \log_2 k \rfloor)n^3$ flops to find S_{n+1} [5]

(3) n^3 flops to find X_{n+1}.

If the matrix A is hermitian, the cost is approximately $\dfrac{(j-1+B_k+\lfloor \log_2 k \rfloor)\, n^3}{2}$ flops per iteration. If the condition $\|I - S_0\| < 1$ in Theorem 2.3 is not satisfied then the start method (I) must be used until $\|I_0 - S\| < 1$.

4. Numerical Examples. In this section we will use the Frobenius matrix norm $\|A\|_F = \sqrt{\sum_{i,j} |a_{i,j}|^2}$, the error $e_n = \|X_n - X_{n-1}\|_F$ and the following definition.

Definition 4.1. *The method* (I) *converges within* n *iterations if* $e_n \le \delta$, *where* δ *is a given error tollerance.*

Example 1.

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

It is desired to find $A^{1/3}$. We will use method (I) with 3-rd order convergence rate ($j = 3$). The matrix A is not diagonalizable. If $X_0 = I$ then $\|I - A^{-1}X_0^3\|_F = 1.26$. If $\delta = 10^{-7}$ then method (I) converges within 6 iterations.

This example illustrates that the conditions in Theorem 2.3 are not necessary conditions.

Example 2. In this example we compare method (I) with the quadratically convergent method in [3]. Let A be the 10×10 matrix defined by

$$a_{ij} = \begin{cases} 1 & \text{if} \quad i = j \\ -1 & \text{if} \quad i < j \\ 0 & \text{if} \quad i > j \end{cases}.$$

It is desired to find $A^{1/3}$.

For the quadratically convergent method in [3] the cost is approximately $(2 + k(3k+1)/2)n^3$ flops per iteration. Let $\delta = 10^{-5}$. The method in [3] converges within 5 iterations and the error $e_5 = 8.71\text{E}{-6}$. The costs (for 5 iterations) are approximately 85000 flops in total.

We shall use method (I) with 5-th order covergence rate and $X_0 = I$. The method (I) converges within 3 iterations and the error $e_3 < 1.0\text{E}{-8}$. The costs (for 3 iterations) are approximately 21000 flops in total.

We see that the method (I) converges 4 times faster than the method in [3].

Single precision calculations were used for the two examples.

REFERENCES

[1] E. D. Denman. Roots of real matrices. *Linear Algebra Appl.*, **36** (1981), 133-139.

[2] W. D. Hoskins, D. J. Walton. A faster more stable method for computing the n-th roots of positive definite matrices. *Linear Algebra Appl.*, **26** (1979), 139-164.

[3] Y. T. Tsay, L. S. Shieh, J. S. H. Tsai. A fast method for computing the principal n-th roots of complex matrices, *Linear Algebra Appl.*, **76** (1986), 205-221.

[4] G. W. Stewart. Introduction to matrix computation. New York, Academic Press, 1974.

[5] D. E. Knuth. The Art of Computer Programming, vol. 2. Addison-Wesley, Don Mills, 1969.

University of Novi Sad
Technical Faculty "Mihajlo Pupin"
23000 Zrenjanin
Yugoslavia

Received December 30, 1994
Revised July 7, 1995