Hindawi Publishing Corporation
The Scientific World Journal
Volume 2015, Article ID 964257, 6 pages
http://dx.doi.org/10.1155/2015/964257

Research Article

On a Cubically Convergent Iterative Method for Matrix Sign

M. Sharifi,[1] S. Karimi Vanani,[1] F. Khaksar Haghani,[1] M. Arab,[1] and S. Shateyi[2]

[1]*Department of Mathematics, Islamic Azad University, Shahrekord Branch, Shahrekord, Iran*
[2]*Department of Mathematics and Applied Mathematics, University of Venda, Thohoyandou 0950, South Africa*

Correspondence should be addressed to S. Shateyi; stanford.shateyi@univen.ac.za

Received 20 July 2014; Revised 29 August 2014; Accepted 30 August 2014

Academic Editor: Predrag S. Stanimirovic

We propose an iterative method for finding matrix sign function. It is shown that the scheme has global behavior with cubical rate of convergence. Examples are included to show the applicability and efficiency of the proposed scheme and its reciprocal.

1. Introduction

It is known that the function of sign in the scalar case is defined for any $z \in \mathbb{C}$ not on the imaginary axis by

$$\operatorname{sign}(z) = \begin{cases} 1, & \operatorname{Re}(z) > 0, \\ -1, & \operatorname{Re}(z) < 0. \end{cases} \tag{1}$$

An extension of (1) for the matrix case was given firstly by Roberts in [1]. This extended matrix function is of clear importance in several applications (see, e.g., [2] and the references therein).

Assume that $A \in \mathbb{C}^{n\times n}$ is a matrix with no eigenvalues on the imaginary axis. To define this matrix function formally, let

$$A = TJT^{-1} \tag{2}$$

be a Jordan canonical form arranged so that $J = \operatorname{diag}(J_1, J_2)$, where the eigenvalues of $J_1 \in \mathbb{C}^{p\times p}$ lie in the open left half-plane and those of $J_2 \in \mathbb{C}^{q\times q}$ lie in the open right half-plane; then

$$S = \operatorname{sign}(A) = T \begin{pmatrix} -I_p & 0 \\ 0 & I_q \end{pmatrix} T^{-1}, \tag{3}$$

where $p + q = n$. A simplified definition of the matrix sign function for Hermitian case (eigenvalues are all real) is

$$S = U \operatorname{diag}(\operatorname{sign}(\lambda_1), \ldots, \operatorname{sign}(\lambda_n)) U^*, \tag{4}$$

where

$$U^* A U = \operatorname{diag}(\lambda_1, \ldots, \lambda_n) \tag{5}$$

is a diagonalization of A.

The importance of computing S is also due to the fact that the sign function plays a fundamental role in iterative methods for matrix roots and the polar decomposition [3].

Note that although $\operatorname{sign}(A)$ is a square root of the identity matrix, it is not equal to I or $-I$ unless the spectrum of A lies entirely in the open right half-plane or open left half-plane, respectively. Hence, in general, $\operatorname{sign}(A)$ is a nonprimary square root of I.

In this paper, we focus on iterative methods for finding S. In fact, such methods are Newton-type schemes which are in essence fixed-point-type methods by producing a convergent sequence of matrices via applying a suitable initial matrix.

The most famous method of this class is the quadratic Newton method defined by

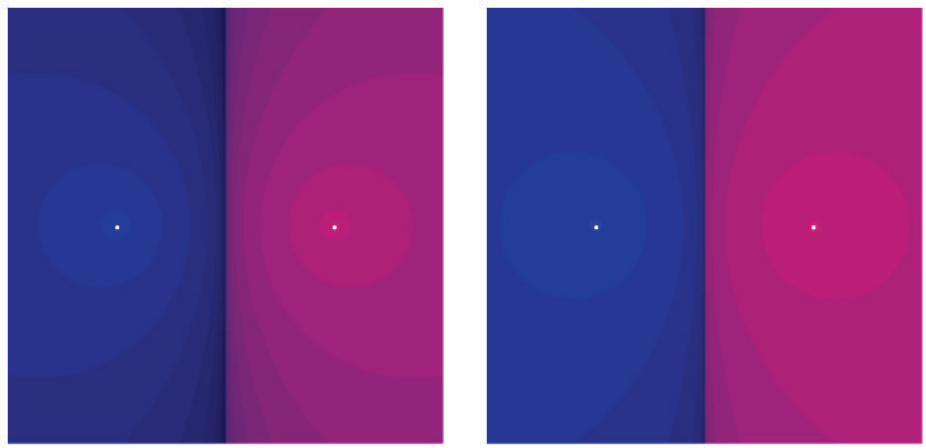
$$X_{k+1} = \frac{1}{2}\left(X_k + X_k^{-1}\right). \tag{6}$$

It should be remarked that iterative methods, such as (6), and the Newton-Schultz iteration

$$X_{k+1} = \frac{1}{2} X_k \left(3I - X_k^2\right) \tag{7}$$

or the cubically convergent Halley method

$$X_{k+1} = \left[I + 3X_k^2\right]\left[X_k\left(3I + X_k^2\right)\right]^{-1}, \tag{8}$$



(a) (b)

Figure 1: Attraction basins for (6) (a) and (8) (b) for the polynomial $g(x) = x^2 - 1$.

are all special cases of the Padé family proposed originally in [4]. The Padé approximation belongs to a broader category of rational approximations. Coincidentally, the best uniform approximation of the sign function on a pair of symmetric but disjoint intervals can be expressed as a rational function.

Note that although (7) does not possess a global convergence behavior, on state-of-the-art parallel computer architectures, matrix inversions scale less satisfactorily than matrix multiplications do, and subsequently (7) is useful in some problems. However, due to local convergence behavior, it is excluded from our numerical examples in this work.

The rest of this paper is organized as follows. In Section 2, we discuss how to construct a new iterative method for finding (3). It is also shown that the constructed method is convergent with cubical rate. It is noted that its reciprocal iteration obtained from our main method is also convergent. Numerical examples are furnished to show the higher numerical accuracy for the constructed solvers in Section 3. The paper ends in Section 4 with some concluding comments.

2. A New Method

The connection of matrix iteration methods with the sign function is not immediately obvious, but in fact such methods can be derived by applying a suitable root-finding method to the nonlinear matrix equation

$$X^2 = I \tag{9}$$

and when of course sign(A) is one solution of this equation (see for more [5]).

Here, we consider the following root-solver:

$$x_{k+1} = x_k - \frac{10 - 4L(x_k)}{10 - 9L(x_k)} \frac{f(x_k)}{f'(x_k)}, \tag{10}$$

with $L(x_k) = f''(x_k) f(x_k) / f'(x_k)^2$. In what follows, we observe that (10) possesses third order of convergence.

Theorem 1. *Let $\alpha \in D$ be a simple zero of a sufficiently differentiable function $f : D \subseteq \mathbb{C} \to \mathbb{C}$, which contains x_0 as an initial approximation. Then the iterative expression (10) satisfies*

$$e_{k+1} = \left(\frac{c_2^2}{5} - c_3 \right) e_k^3 + O\left(e_k^4\right), \tag{11}$$

where $c_j = f^{(j)}(\alpha)/j! f'(\alpha)$, $e_k = x_k - \alpha$.

Proof. The proof would be similar to the proofs given in [6]. □

Applying (10) on the matrix equation (9) will result in the following new matrix fixed-point-type iteration for finding (3):

$$X_{k+1} = \left(2I + 15X_k^2 + 3X_k^4\right)\left[9X_k + 11X_k^3\right]^{-1}, \tag{12}$$

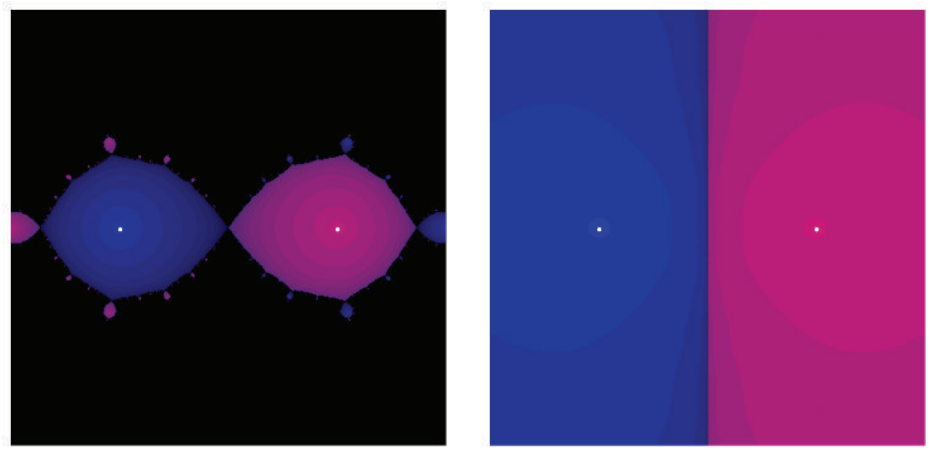
where $X_0 = A$. This is named PM1 from now on.

The proposed scheme (12) is not a member of Padé family [4]. Furthermore, applying (10) on the scalar equation $g(x) = x^2 - 1$ provides a global convergence in the complex plane (except the points lying on the imaginary axis). This global behavior, which is kept for matrix case, has been illustrated in Figure 1 by drawing the basins of attraction for (6) and (8). The attraction basins for (7) (local convergence) and (12) (global convergence) are also portrayed in Figure 2.

Theorem 2. *Let $A \in \mathbb{C}^{n\times n}$ have no pure imaginary eigenvalues. Then, the matrix sequence $\{X_k\}_{k=0}^{k=\infty}$ defined by (12) converges to S, choosing $X_0 = A$.*

Proof. We remark that all matrices, whether they are diagonalizable or not, have a Jordan normal form $A = TJT^{-1}$, where the matrix J consists of Jordan blocks. For this reason, let A have a Jordan canonical form arranged as

$$T^{-1}AT = \Lambda = \begin{bmatrix} C & 0 \\ 0 & N \end{bmatrix}, \tag{13}$$



(a) (b)

FIGURE 2: Attraction basins of (7) (a) and (12) (b) for the polynomial $g(x) = x^2 - 1$.

where T is a nonsingular matrix and C, N are square Jordan blocks corresponding to eigenvalues lying in $\mathbb{C}^-$ and $\mathbb{C}^+$, respectively. We have

$$\begin{aligned}
&\operatorname{sign}(\Lambda)\\
&\quad = \operatorname{sign}\left(T^{-1}AT\right) = T^{-1}\operatorname{sign}(A)\,T\\
&\quad = \operatorname{diag}\left(\operatorname{sign}(\lambda_1), \ldots, \operatorname{sign}\left(\lambda_p\right), \operatorname{sign}\left(\lambda_{p+1}\right), \ldots, \operatorname{sign}(\lambda_n)\right).
\end{aligned} \tag{14}$$

If we define $D_k = T^{-1}X_kT$, then, from the method (12), we obtain

$$D_{k+1} = \left(2I + 15D_k^2 + 3D_k^4\right)\left[9D_k + 11D_k^3\right]^{-1}. \tag{15}$$

Note that if D_0 is a diagonal matrix, then, based on an inductive proof, all successive D_k are diagonal too. From (15), it is enough to show that $\{D_k\}$ converges to $\operatorname{sign}(\Lambda)$. We remark that the case at which D_0 is not diagonal will be discussed later in the proof.

In the meantime, we can write (15) as n uncoupled scalar iterations to solve $g(x) = x^2 - 1 = 0$, given by

$$d_{k+1}^i = \left(2 + 15{d_k^i}^2 + 3{d_k^i}^4\right)\left[9d_k^i + 11{d_k^i}^3\right]^{-1}, \tag{16}$$

where $d_k^i = (D_k)_{i,i}$ and $1 \le i \le n$. From (15) and (16), it is enough to study the convergence of $\{d_k^i\}$ to $\operatorname{sign}(\lambda_i)$.

It is known that $\operatorname{sign}(\lambda_i) = s_i = \pm 1$. Thus, we attain

$$\frac{d_{k+1}^i - 1}{d_{k+1}^i + 1} = \frac{\left(-1 + d_k^i\right)^3\left(-2 + 3d_k^i\right)}{\left(1 + d_k^i\right)^3\left(2 + 3d_k^i\right)}. \tag{17}$$

Since $|d_0^i| = |\lambda_i| > 0$, we have

$$\lim_{k\to\infty}\left|\frac{d_{k+1}^i - 1}{d_{k+1}^i + 1}\right| = 0, \tag{18}$$

and $\lim_{k\to\infty}|d_k^i| = 1 = |\operatorname{sign}(\lambda_i)|$. This shows that $\{d_k^i\}$ is convergent.

In the convergence proof, D_0 may not be diagonal. Since the Jordan canonical form of some matrices may not be diagonal, thus, one cannot write (15) as n uncoupled scalar iterations (16). We comment that in this case our method is also convergent. To this goal, we must pursue the scalar relationship among the eigenvalues of the iterates for the studied rational matrix iteration.

In this case, the eigenvalues of X_k are mapped from the iterate k to the iterate $k+1$ by the following relation:

$$\lambda_{k+1}^i = \left(2 + 15{\lambda_k^i}^2 + 3{\lambda_k^i}^4\right)\left[9\lambda_k^i + 11{\lambda_k^i}^3\right]^{-1}. \tag{19}$$

So, (19) clearly shows that the eigenvalues in the general case are convergent to ± 1; that is to say,

$$\lim_{k\to\infty}\left|\frac{\lambda_{k+1}^i - 1}{\lambda_{k+1}^i + 1}\right| = 0. \tag{20}$$

Consequently, we have

$$\lim_{k\to\infty} X_k = T\left(\lim_{k\to\infty} D_k\right)T^{-1} = T\operatorname{sign}(\Lambda)\ T^{-1} = \operatorname{sign}(A). \tag{21}$$

The proof is ended. □

Theorem 3. *Let $A \in \mathbb{C}^{n\times n}$ have no pure imaginary eigenvalues. Then the proposed method (12) converges cubically to the sign matrix S.*

Proof. Clearly, X_k are rational functions of A and, hence, like A, commute with S. On the other hand, we know that $S^2 = I$,

Table 1: Results of comparisons for Example 5 using $X_0 = A$.

Methods	NM	HM	PM1	PM2
IT	14	9	8	8
R_{k+1}	1.41584×10^{-249}	1.0266×10^{-299}	2.5679×10^{-298}	1.45091×10^{-337}
ρ	1.99077	3	3	3

Table 2: Results of comparisons for Example 6 using $X_0 = A$.

Methods	NM	HM	PM1	PM2
IT	10	7	6	6
R_{k+1}	5.7266×10^{-155}	5.80819×10^{-203}	8.38265×10^{-153}	1.55387×10^{-143}
ρ	2.00228	3.00001	3.00015	3

$S^{-1} = S$, $S^{2j} = I$, and $S^{2j+1} = S$, $j \geq 1$. Using the replacement $B_k = 9X_k + 11X_k^3$, we have

$$\begin{aligned} X_{k+1} - S &= \left(2I + 15X_k^2 + 3X_k^4\right) B_k^{-1} - S \\ &= \left(2I + 15X_k^2 + 3X_k^4 - SB_k\right) B_k^{-1} \\ &= \left(2I + 15X_k^2 + 3X_k^4 - 9SX_k - 11SX_k^3\right) B_k^{-1} \\ &= -\left(-2S - 15SX_k^2 - 3SX_k^4 + 9X_k + 11X_k^3\right) \\ &\quad \times S^{-1} B_k^{-1} \\ &= \left(X_k - S\right)^3 \left(2I - 3SX_k\right) S^{-1} B_k^{-1}. \end{aligned} \tag{22}$$

Now, using any matrix norm from both sides of (22), we attain

$$\left\|X_{k+1} - S\right\| \leq \left(\left\|B_k^{-1}\right\| \left\|S^{-1}\right\| \left\|2I - 3SX_k\right\|\right) \left\|X_k - S\right\|^3. \tag{23}$$

This reveals the cubical rate of convergence for the new method (12). The proof is complete. □

It should be remarked that the reciprocal iteration obtained from (12) is also convergent to the sign matrix (3) as follows:

$$X_{k+1} = \left(9X_k + 11X_k^3\right) \left[2I + 15X_k^2 + 3X_k^4\right]^{-1}, \tag{24}$$

where $X_0 = A$. This is named PM2. Similar convergence results as the ones given in Theorems 2-3 hold for (24).

A scaling approach to accelerate the beginning phase of convergence is normally necessary since the convergence rate cannot be seen in the initial iterates. Such an idea was discussed fully in [7] for Newton's method. An effective way to enhance the initial speed of convergence is to scale the iterates prior to each iteration; that is, X_k is replaced by $\mu_k X_k$. Subsequently, we can present the accelerated forms of our proposed methods as follows:

$$\begin{gathered} X_0 = A, \\ \mu_k = \text{is the scaling parameter computed by (27)}, \\ X_{k+1} = \left(2I + 15\mu_k^2 X_k^2 + 3\mu_k^4 X_k^4\right) \left[9\mu_k X_k + 11\mu_k^3 X_k^3\right]^{-1}, \end{gathered} \tag{25}$$

or

$$\begin{gathered} X_0 = A, \\ \mu_k = \text{is the scaling parameter computed by (27)}, \\ X_{k+1} = \left(9\mu_k X_k + 11\mu_k^3 X_k^3\right) \left[2I + 15\mu_k^2 X_k^2 + 3\mu_k^4 X_k^4\right]^{-1}, \end{gathered} \tag{26}$$

$$\mu_k = \begin{cases} \sqrt{\dfrac{\left\|X_k^{-1}\right\|}{\left\|X_k\right\|}}, & \text{(norm scaling)}, \\ \sqrt{\dfrac{\rho\left(X_k^{-1}\right)}{\rho\left(X_k\right)}}, & \text{(spectral scaling)}, \\ \sqrt{\left|\det\left(X_k\right)\right|^{-1/n}}, & \text{(determinantal scaling)}, \end{cases} \tag{27}$$

where $\lim_{k\to\infty}\mu_k = 1$ and $\lim_{k\to\infty}X_k = S$. The different scaling factors for μ_k in (27) are borrowed from Newton's method. For this reason it is important to show the behavior of the accelerator methods (25)-(26) and this will be done in the next section.

3. Numerical Examples

In this section, the results of comparisons in terms of number of iterations and the residual norms have been reported for various matrix iterations. We compare PM1 and PM2 with (6) denoted by NM and (8) denoted by HM. The programming package Mathematica [8] is used throughout this section. In Tables 1 and 2, IT stands for the number of iterates.

Note that the computational order of convergence for matrix iterations in finding S can be estimated by [9]

$$\rho = \frac{\log\left(\left\|X_{k+1}^2 - I\right\| / \left\|X_k^2 - I\right\|\right)}{\log\left(\left\|X_k^2 - I\right\| / \left\|X_{k-1}^2 - I\right\|\right)}, \tag{28}$$

where X_{k-1}, X_k, and X_{k+1} are the last three approximations.

Example 4. In this example, we compare the methods for the following 500 × 500 complex matrix:

```
n = 500; SeedRandom[123];
A = RandomComplex[{-100 - I, 100 + I},{n,n}];
```

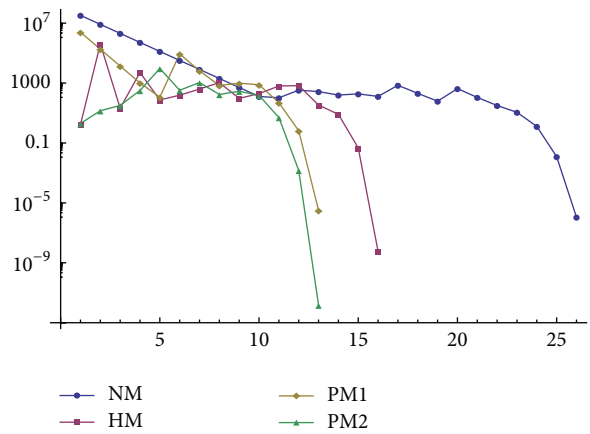


Figure 3: Convergence history versus number of iterations for different methods in Example 4.

We apply here double precision arithmetic with the stop termination $R_{k+1} = \|X_{k+1}^2 - I\|_\infty \leq 10^{-5}$. Results are given in Figure 3.

Example 5 (academic test). We compute the matrix sign for the following complex test problem:

$$A = \begin{pmatrix} 0 & 10 & i & 7+i \\ 7 & -5 & 6 & -5 \\ 0 & 60 & -2 & 9 \\ 0 & 5 & 9 & i \end{pmatrix}, \tag{29}$$

where

$$S = \begin{pmatrix} 0.882671 + 0.0118589i & 0.461061 - 0.0519363i & -0.167387 + 0.0215728i & 0.168184 - 0.0194164i \\ 0.219355 + 0.00464485i & 0.136809 - 0.00840032i & 0.313995 - 0.00196855i & -0.314977 - 0.00219388i \\ -0.566306 - 0.0184534i & 2.22878 + 0.0471091i & 0.189109 - 0.00416224i & 0.813305 + 0.0149399i \\ 0.145285 + 0.00157401i & -0.57165 + 0.000347003i & 0.207909 - 0.00345322i & 0.791412 + 0.000703638i \end{pmatrix}. \tag{30}$$

We apply here 600-digit fixed point arithmetic in our calculations with the stop termination $R_{k+1} = \|X_{k+1}^2 - I\|_\infty \leq 10^{-150}$. The results for this example are illustrated in Table 1. We report the COCs in l_∞.

Iterative schemes PM1 and PM2 are evidently believed to be more favorable than the other compared methods due to their fewer number of iterations and acceptable accuracy. Hence, the proposed methods with properly chosen initial matrix X_0 can be helpful in finding the sign of a nonsingular complex matrix.

Example 6. Here we rerun Example 5 using the scaling approaches (27) with the stop termination $R_{k+1} = \|X_{k+1}^2 - I\|_\infty \leq 10^{-100}$. The results for this example are illustrated in Table 2. We used the determinantal scaling for all compared methods. The numerical results uphold the theoretical discussions of Section 2.

A price paid for the high order convergence is the increased amount of matrix multiplications and inversions. This is a typical consequence. However the most important advantage of the presented methods in contrast to the methods of the same orders, such as (8), is their larger attraction basins. This superiority basically allows the new methods to converge to a required tolerance in one lower iteration than their same order methods. Hence, studying the thorough computational efficiency index of the proposed methods may not be an easy task and it must be pursued experimentally. In an experimental manner, if the costs of one matrix-matrix product and one matrix inversion are unity and 1.5 of unity, respectively, then we have the following efficiency indices for different methods: $E_{(6)} = 2^{1/(14(1)+14(1.5))} \simeq 1.020$, $E_{(8)} = 3^{1/(9(3)+9(1.5))} \simeq 1.027$, and $E_{(12)} = 3^{1/(8(4)+8(1.5))} \simeq 1.025$. Note that for Newton's method we have one matrix-matrix product per cycle due to the computation of stopping criterion. Other similar computations for efficiency indices for different examples show similar behaviors to the above mentioned one.

4. Summary

Matrix functions are used in many areas of linear algebra and arise in numerous applications in science and engineering. The function of a matrix can be defined in several ways, of which the following three are generally the most useful: Jordan canonical form, polynomial interpolation, and finally Cauchy integral.

In this paper, we have focus on iterative methods for this purpose. Hence, a third order nonlinear equation solver has been employed for constructing a new method for S. It was shown that the convergence is global via attraction basins in the complex plane and the rate of convergence is cubic. Furthermore, PM2 as the reciprocal of the method PM1 with the same convergence properties was proposed. The acceleration of PM1 and PM2 via scaling was also illustrated simply.

Finally some numerical examples in both double and multiple precisions were performed to show the efficiency of PM1 and PM2. Further researches must be forced to extend the obtained iterations for computing polar decompositions in future studies.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

The authors would like to thank the referees for their helpful corrections and suggestions.

References

[1] J. D. Roberts, "Linear model reduction and solution of the algebraic Riccati equation by use of the sign function," *International Journal of Control*, vol. 32, no. 4, pp. 677–687, 1980.

[2] C. S. Kenney and A. J. Laub, "The matrix sign function," *IEEE Transactions on Automatic Control*, vol. 40, no. 8, pp. 1330–1348, 1995.

[3] N. J. Higham, *Functions of Matrices: Theory and Computation*, Society for Industrial and Applied Mathematics, Philadelphia, Pa, USA, 2008.

[4] C. Kenney and A. J. Laub, "Rational iterative methods for the matrix sign function," *SIAM Journal on Matrix Analysis and Applications*, vol. 12, no. 2, pp. 273–291, 1991.

[5] F. Soleymani, P. S. Stanimirović, S. Shateyi, and F. K. Haghani, "Approximating the matrix sign function using a novel iterative method," *Abstract and Applied Analysis*, vol. 2014, Article ID 105301, 9 pages, 2014.

[6] F. Soleymani, "Some high-order iterative methods for finding all the real zeros," *Thai Journal of Mathematics*, vol. 12, no. 2, pp. 313–327, 2014.

[7] C. Kenney and A. J. Laub, "On scaling Newton's method for polar decomposition and the matrix sign function," *SIAM Journal on Matrix Analysis and Applications*, vol. 13, no. 3, pp. 698–706, 1992.

[8] S. Wagon, *Mathematica in Action*, Springer, New York, NY, USA, 3rd edition, 2010.

[9] F. Soleymani, E. Tohidi, S. Shateyi, and F. Haghani, "Some matrix iterations for computing matrix sign function," *Journal of Applied Mathematics*, vol. 2014, Article ID 425654, 9 pages, 2014.