The matrix square root from a new functional perspective: theoretical results and computational issues

Beatrice Meini*

Abstract

We give a new characterization of the matrix square root and a new algorithm for its computation. We show how the matrix square root is related to the constant block coefficient of the inverse of a suitable matrix Laurent polynomial. This fact, besides giving a new interpretation of the matrix square root, allows one to design an efficient algorithm for its computation. The algorithm, which is mathematically equivalent to Newton's method, is quadratically convergent, numerically insensitive to the ill-conditioning of the original matrix, and works also in the special case where the original matrix is singular and has a square root.

1 Introduction

Let us denote by $\lambda(H)$ and $\rho(H)$ the set of the eigenvalues and the spectral radius of the square matrix H, respectively, by $\mathcal{D}$ the open unit disk in the complex plane, and by $\mathbb{C}^+$ ($\mathbb{C}^-$) the open right (left) half complex plane.

Let $A \in \mathbb{C}^{n\times n}$ be a matrix having no nonpositive real eigenvalues. Under this assumption, the quadratic matrix equation

$$X^2 - A = 0 \qquad (1)$$

has a unique solution such that $\lambda(X) \subset \mathbb{C}^+$ (see [6, 9]). We will denote by $A^{1/2}$ that solution, which is usually called the principal square root of A.

The matrix square root has several properties and is closely related to the matrix sign function. A well known fact is that [10]

$$\mathrm{sign}(A) = A(A^2)^{-1/2}.$$

Based on this property, in [11] Higham has shown that the matrix square root can be characterized in terms of the sign of a suitable 2×2 block matrix:

$$\mathrm{sign}\left(\begin{bmatrix} 0 & A \\ I & 0 \end{bmatrix}\right) = \begin{bmatrix} 0 & A^{1/2} \\ A^{-1/2} & 0 \end{bmatrix}. \qquad (2)$$

*Dipartimento di Matematica, Università di Pisa, Italy, meini@dm.unipi.it

This is a fundamental relation, since it allows to compute the matrix square root by using any method for computing the matrix sign. Indeed, Newton's iteration applied directly to the matrix equation (1) suffers from instability problems (see [15, 8]) while methods which are based on the computation of the matrix sign of the matrix in (2) are generally more stable [11].

In this paper we derive a new characterization of the matrix square root. More specifically we prove the following result:

Theorem 1. *Let $A \in \mathbb{C}^{n\times n}$ be a matrix having no nonpositive real eigenvalues. Then the Laurent matrix polynomial*

$$R(z) = (I - A)z^{-1} + 2(I + A) + (I - A)z \tag{3}$$

is invertible for any $z \in \mathbb{C}$ such that $r < |z| < 1/r$ where

$$r = \rho\left((A^{1/2} - I)(A^{1/2} + I)^{-1}\right) < 1.$$

Moreover, $H(z) = R(z)^{-1} = H_0 + \sum_{i=1}^{\infty} H_i(z^i + z^{-i})$ is such that

$$H_0 = \frac{1}{4}A^{-1/2}.$$

The theorem states that the block constant coefficient of the inverse of $R(z)$ fully defines the matrix square root. This is a new characterization in terms of matrix functions.

At a first glance this property may appear to be of purely theoretical interest, since how to compute the constant coefficient of $H(z)$ is not immediately clear. However, we show that we can apply the cyclic reduction algorithm, which is quadratically convergent, has a low computational cost per iteration, and enjoys good numerical stability properties. In particular the cyclic reduction algorithm, in contrast to other methods, works fine also in the critical case where A is close to a singular matrix. If A is singular, convergence still occurs provided that the zero eigenvalues are semi-simple, i.e., the Jordan blocks corresponding to 0 have size 1. In this case the convergence is linear. One open question concerning cyclic reduction is to figure out if a scaling technique can be applied in order to speed up the convergence; in fact, for particular problems our algorithm can be less efficient than other methods which allow scaling.

The paper is organized as follows. In Section 2 we prove Theorem 1. In Section 3 we propose the algorithm, we prove its convergence properties, and we show some specific features of the algorithm for the singular case, the symmetric positive case, the case of M-matrices and the skew-Hamiltonian case. In Section 4 we report the numerical results and the comparisons with the available methods. Finally in Section 5 we draw conclusions.

2 Proof of Theorem 1

Let us introduce the Möbius function $w : \mathbb{C} \setminus \{-1\} \to \mathbb{C}$,

$$w(x) = (x - 1)/(x + 1),$$

which maps the right half plane $\mathbb{C}^+$ into the open unit disk $\mathcal{D}$. Since $\lambda(A^{1/2}) \subset \mathbb{C}^+$, the matrix

$$W = (A^{1/2} - I)(A^{1/2} + I)^{-1} \tag{4}$$

is well defined, and is such that $\lambda(W) \subset \mathcal{D}$. Moreover, it is easy to verify that

$$A^{1/2} = (I + W)(I - W)^{-1}. \tag{5}$$

By substituting X in (1) with the above expression for $A^{1/2}$, since all the matrices commute, we obtain that W solves the matrix equation

$$(I - A) + 2(I + A)Y + (I - A)Y^2 = 0. \tag{6}$$

Moreover, from the uniqueness of $A^{1/2}$, W is the unique solution such that $\rho(W) < 1$.

We associate with the matrix equation (6) the matrix Laurent polynomial $R(z)$ of (3). Since W solves the matrix equation (6) and since $R(z)$ is a symmetric matrix Laurent polynomial it follows that $R(z)$ can be factorized as

$$R(z) = G(z)G(z^{-1}) \tag{7}$$

where

$$G(z) = z(A^{1/2} + I) - (A^{1/2} - I)$$

(this latter property can be also verified by direct inspection). In particular, $G(z)$ is invertible for any $z \in \mathbb{C}$ such that $\rho(W) < |z|$, and we may write

$$G(z)^{-1} = z^{-1}(A^{1/2} + I)^{-1}(I - z^{-1}W)^{-1} = z^{-1}(A^{1/2} + I)^{-1} \sum_{i=0}^{\infty} z^{-i}W^i.$$

Hence, from the factorization (7), we obtain

$$H(z) = R(z)^{-1} = G(z^{-1})^{-1}G(z)^{-1} = (A^{1/2} + I)^{-2} \sum_{i=0}^{\infty} z^i W^i \sum_{i=0}^{\infty} z^{-i} W^i,$$

which is convergent for any $z \in \mathbb{C}$ such that $\rho(W) < |z| < 1/\rho(W)$. The constant matrix coefficient of $H(z)$ is

$$H_0 = (A^{1/2} + I)^{-2} \sum_{i=0}^{\infty} W^{2i} = (A^{1/2} + I)^{-2}(I - W^2)^{-1} = \frac{1}{4}A^{-1/2}.$$

3 The algorithm

According to the results of [1], the cyclic reduction algorithm [4] applied to the bi-infinite block tridiagonal block Toeplitz matrix

$$\mathcal{T} = \begin{bmatrix} \ddots & \ddots & & 0 \\ \ddots & 2(I+A) & (I-A) & \\ & (I-A) & 2(I+A) & \ddots \\ 0 & & \ddots & \ddots \end{bmatrix}$$

generates a sequence $\{\mathcal{T}_k\}_{k\geq 0}$ of bi-infinite block tridiagonal block Toeplitz matrices which quadratically converge to the block diagonal matrix having H_0^{-1} on the main diagonal, where H_0 is the constant term of the inverse of the matrix Laurent polynomial (3).

In this particular case, the cyclic reduction algorithm consists in generating the two sequences of matrices $\{Y_k\}_{k\geq 0}$, $\{Z_k\}_{k\geq 0}$ in the following way:

$$\begin{aligned} &Z_0 = 2(I + A), \quad Y_0 = I - A, \\ &Y_{k+1} = -Y_k Z_k^{-1} Y_k, \\ &Z_{k+1} = Z_k - 2Y_k Z_k^{-1} Y_k, \quad k = 0, 1, \ldots. \end{aligned} \tag{8}$$

At the k-th step, Z_k is the matrix defining the main block diagonal of $\mathcal{T}_k$ and Y_k is the matrix defining the lower and the upper diagonal blocks of $\mathcal{T}_k$. Observe that, since both Y_k and Z_k are rational functions in A, Y_k and Z_k commute.

We now show some properties of the sequences $\{Y_k\}_{k\geq 0}$, $\{Z_k\}_{k\geq 0}$, from which we can also directly derive that the sequence $\{Z_k\}_k$ converges quadratically to $4A^{1/2}$.

Proposition 2. *Let $A \in \mathbb{C}^{n\times n}$ be a matrix having no nonpositive real eigenvalues. Then the matrices Z_k generated by* (8) *are nonsingular for all $k \geq 0$ and*

$$Z_k + 2Y_k W^{2^k} = 4A^{1/2}, \qquad k = 0, 1, \ldots, \tag{9}$$

where W is defined in (4).

Proof. From the properties of cyclic reduction applied to matrix equations [2], the following identity holds

$$Y_k W^{2^{k+1}} + Z_k W^{2^k} + Y_k = 0, \tag{10}$$

for $k = 0, 1, 2, \ldots$.

Let us prove by induction on k that Z_k is nonsingular. Z_0 is nonsingular since A cannot have an eigenvalue equal to 1. Let us assume the thesis true for a fixed $k \geq 0$ and let us prove it for $k + 1$. From (10) we have

$$Z_k^{-1} Y_k = -W^{2^k} (I + W^{2\cdot 2^k})^{-1}. \tag{11}$$

Hence, from (8),

$$Z_{k+1} = Z_k \left(I - 2(Z_k^{-1}Y_k)^2\right) = Z_k \left(I - 2\left(W^{2^k}(I + W^{2\cdot 2^k})^{-1}\right)^2\right). \tag{12}$$

Since Z_k is nonsingular by inductive assumption, it is sufficient to show that $I - 2\left(W^{2^k}(I + W^{2\cdot 2^k})^{-1}\right)^2$ is nonsingular. Any eigenvalue of this matrix has the form $\mu = 1 - 2\lambda^{2^{k+1}}(1 + \lambda^{2\cdot 2^k})^{-2}$, where λ is an eigenvalue of W. Since $\lambda(W) \subset \mathcal{D}$, μ cannot be zero.

Now from (10) it is a simple matter to verify by induction that

$$Z_k + 2Y_k W^{2^k} = Z_0 + 2Y_0 W$$

for any k. If we replace Z_0, Y_0 and W with their expression in terms of A and $A^{1/2}$ we obtain that

$$
\begin{aligned}
Z_0 + 2Y_0W &= 2(I+A) + 2(I-A)(A^{1/2}-I)(I+A^{1/2})^{-1} \\
&= 2\left((I+A)(I+A^{1/2}) + (I-A)(A^{1/2}-I)\right)(I+A^{1/2})^{-1} \\
&= 4A^{1/2}(I+A^{1/2})(I+A^{1/2})^{-1} = 4A^{1/2}.
\end{aligned}
$$

□

From the above proposition and from the spectral properties of W, we derive the following convergence results, which describe the quadratic convergence of our algorithm:

Theorem 3. *Let $A \in \mathbb{C}^{n\times n}$ be a matrix having no nonpositive real eigenvalues. Then the sequences $\{Y_k\}_k$ and $\{Z_k\}_k$ of* (8) *are convergent and*

$$||Y_k|| = O(\sigma^{2^k}), \tag{13}$$

$$||Z_k - 4A^{1/2}|| = O(\sigma^{2\cdot 2^k}), \tag{14}$$

for any real number σ, $\rho(W) < \sigma < 1$, and for any matrix norm $||\cdot||$.

Proof. From (11), since $\rho(W) < 1$, it follows

$$||Z_k^{-1}Y_k|| = O(\sigma^{2^k}), \tag{15}$$

for any real number σ, $\rho(W) < \sigma < 1$, and for any matrix norm $||\cdot||$. Thus, from (12), it follows that the sequences $\{Z_k\}_k$ and $\{Z_k^{-1}\}_k$ have norm bounded above by a constant. Hence from (15) it follows (13). Finally, from (9) and (13) we obtain (14). □

The resulting algorithm for the computation of $A^{1/2}$ consists of computing the sequences $\{Y_k\}_k$ and $\{Z_k\}_k$ of (8), until $||Y_k|| < \epsilon$, for a fixed error bound ϵ and for a chosen matrix norm $||\cdot||$, and then approximating $A^{1/2}$ by $Z_k/4$. It is worth noting that each step of our algorithm requires only one matrix inversion and 2 matrix multiplications.

Is is also interesting to point out that the sequence $\{Z_k\}_k$ generated by our algorithm coincides, up to within the multiplicative constant 4, with the sequence generated by applying Newton's method to the matrix equation (1). More specifically, by denoting with $\{X_k\}_k$ the sequence generated by Newton's algorithm, i.e., $X_0 = A$, $X_{k+1} = (X_k + AX_k^{-1})/2$, $k \geq 0$, then we can easily verify by induction on k that

$$Y_k = (A - X_k^2)X_k^{-1},\ Z_k = 4X_{k+1},\ k = 0, 1, \ldots \tag{16}$$

Observe that Y_k represents a "normalized" residual matrix of the approximation X_k. Thus we have a different way to compute the sequence generated by Newton's method, which is a little bit more expensive, but which, unlike the latter method, has good stability properties, as we will show in Section 4. This analogy with Newton's method is shared also by the Denman-Beavers algorithm [5], which is generally more stable than Newton's method [8].

The singular case Suppose that $A \in \mathbb{C}^{n\times n}$ is a singular matrix having no negative real eigenvalues, and suppose that the null eigenvalues are semi-simple, i.e., their Jordan blocks have size 1. This hypothesis on the null eigenvalues is a necessary and sufficient condition for a singular complex matrix A to have a square root which is a function of A (see Theorem 6.4.12 in [12]).

We show that also in this case Z_k is nonsingular for any $k \geq 0$ and that $||Y_k|| = O(2^{-k})$, $||Z_k - 4A^{1/2}|| = O(2^{-k})$. Hence our algorithm can be still applied and converges linearly with rate $1/2$.

Under our assumption, if p is the number of eigenvalues equal to zero, the Jordan canonical form of A has the structure

$$J = PAP^{-1} = \left[\begin{array}{c|c} O & O \\ \hline O & H \end{array}\right]$$

where H is a nonsingular block diagonal matrix of size $(n-p)\times(n-p)$. In particular

$$J^{1/2} = PA^{1/2}P^{-1} = \left[\begin{array}{c|c} O & O \\ \hline O & H^{1/2} \end{array}\right].$$

Let us denote $\hat{W} = PWP^{-1}$, $\hat{Z}_k = PZ_kP^{-1}$, $\hat{Y}_k = PY_kP^{-1}$, $k \geq 0$. Then it is immediate to verify that $\hat{Y}_{k+1} = -\hat{Y}_k\hat{Z}_k^{-1}\hat{Y}_k$, $\hat{Z}_{k+1} = \hat{Z}_k - 2\hat{Y}_k\hat{Z}_k^{-1}\hat{Y}_k$, $k \geq 0$, and that $\hat{Z}_k + 2\hat{Y}_k\hat{W}^{2^k} = 4J^{1/2}$, $k \geq 0$. Moreover, $\hat{Y}_k$ and $\hat{Z}_k$, $k \geq 0$, have the structure

$$\hat{Y}_k = \left[\begin{array}{c|c} R_k^{(1)} & O \\ \hline O & R_k^{(2)} \end{array}\right], \quad \hat{Z}_k = \left[\begin{array}{c|c} S_k^{(1)} & O \\ \hline O & S_k^{(2)} \end{array}\right], \quad k \geq 0,$$

where $R_k^{(1)} = r_kI$, $S_k^{(1)} = s_kI$ are $p\times p$ diagonal matrices with equal diagonal entries. By following the same arguments used in Section 3 we can show that $S_k^{(2)}$ is nonsingular for any k and that $||R_k^{(2)}|| = O(\sigma^{2^k})$, $||S_k^{(2)} - 4H^{1/2}|| = O(\sigma^{2\cdot 2^k})$ where σ is any real number $0 < \sigma < r$, and $r = \rho\left((H^{1/2} - I)(H^{1/2} + I)^{-1}\right) < 1$. Moreover, we have that

$$\begin{array}{l} r_0 = 1, \;\; s_0 = 2, \\ r_{k+1} = -r_k^2/s_k, \;\; s_{k+1} = s_k - 2r_k^2/s_k, \quad k = 0, 1, \ldots \end{array}$$

Hence $r_k = -2^{-k}$, $s_k = 2^{-k+1}$, $k = 1, 2, \ldots$, from which the nonsingularity of $\{Z_k\}_k$ and the convergence properties of $\{Z_k\}_k$ and $\{Y_k\}_k$ follow.

The symmetric positive definite case In the case where A is real symmetric, then the matrices Y_k and Z_k are obviously symmetric for any k. If in addition A is positive definite (we will write $A > 0$), then it is possible to show that $-Y_k$ and Z_k are positive definite, and to give a bound on the spectral condition number of Z_k:

Theorem 4. *If A is symmetric positive definite, then the matrices $-Y_k$, $k \geq 1$, and Z_k, $k \geq 0$, generated by* (8)*, are positive definite, and*

$$\kappa_2(Z_k) = \frac{\max \lambda(Z_k)}{\min \lambda(Z_k)} \leq \frac{1 + \max \lambda(A)}{2\sqrt{\min \lambda(A)}}, \quad k = 0, 1, \ldots.$$

Proof. Since W and Z_k are rational functions in $A^{1/2}$, from (12) it follows that the eigenvalues of Z_{k+1} are the eigenvalues of Z_k times the eigenvalues of the matrix $H_k = I - \left(W^{2^k}(I + W^{2\cdot 2^k})^{-1}\right)^2$; since the eigenvalues of W belong to the interval $(0,1)$, we deduce that also the eigenvalues of H_k belong to the interval $(0,1)$. Thus $Z_{k+1} > 0$ if $Z_k > 0$; since $Z_0 > 0$, we conclude that Z_k is positive definite for any k. Hence $-Y_k$, for $k \geq 1$, is also positive definite, and $Z_k - Z_{k+1}$ is positive definite for any $k \geq 0$; in particular, since $\lim_k Z_k = 4A^{1/2}$ and $Z_0 = 2(I + A)$, also $Z_k - 4A^{1/2}$ and $2(I + A) - Z_k$ are positive definite, and thus the eigenvalues of Z_k belong to the interval $[\mu_1, \mu_2]$, where $\mu_1 = 4 \min \lambda(A^{1/2})$, $\mu_2 = 2(1 + \max \lambda(A))$. □

From the above theorem it follows that the matrices Z_k, $k \geq 0$, which must be inverted at each step, can be better conditioned than A. In particular, assuming that $\max \lambda(A) = 1$ and $\min \lambda(A) < 1$, we always have

$$\kappa_2(Z_k) \leq \frac{1 + \max \lambda(A)}{2\sqrt{\min \lambda(A)}} = \frac{1}{\sqrt{\min \lambda(A)}} < \frac{1}{\min \lambda(A)} = \kappa_2(A).$$

The case of M-matrices Also in the important case where A is an M-matrix, we can show some interesting properties of our algorithm. Up to within a scaling of A by its maximum diagonal entry, A may be written as $A = I - B$, where B is a nonnegative matrix such that $\rho(B) < 1$. Thus $Y_0 = I - A = B$ is nonnegative, and $Z_0 = 2(I + A) = 4I - 2B$ is an M-matrix. Moreover, observe that the matrices Y_k, Z_k, $k \geq 1$, generated by (8) do not change if we replace Y_0 with $-Y_0$, thus we may suppose $Y_0 = A - I = -B$. It is immediate to verify that, if $Z_0 = 4I - 2B$ and $Y_0 = -B$, then the $k \times k$ block matrix

$$H_k = \begin{bmatrix} Z_0 & Y_0 & & 0 \\ Y_0 & Z_0 & \ddots & \\ & \ddots & \ddots & Y_0 \\ 0 & & Y_0 & Z_0 \end{bmatrix}$$

is a nonsingular M-matrix for any $k \geq 1$. If we apply one step of block cyclic reduction to H_k, where $k = 2^q + 1$, $q \geq 1$, we obtain the permuted block LU factorization

$$\Pi H_k \Pi^T = \left[\begin{array}{cccc|cccc} Z_0 & & & 0 & Y_0 & & 0 \\ & \ddots & & & Y_0 & \ddots & \\ & & \ddots & & & \ddots & Y_0 \\ 0 & & & Z_0 & 0 & & Y_0 \\ \hline Y_0 & Y_0 & & 0 & Z_0 & & 0 \\ & \ddots & \ddots & & & \ddots & \\ 0 & & Y_0 & Y_0 & 0 & & Z_0 \end{array}\right] = \left[\begin{array}{c|c} V_1 & V_2 \\ \hline V_3 & V_4 \end{array}\right] = LU,$$

$$L = \left[\begin{array}{c|c} I & 0 \\ \hline V_3 V_1^{-1} & I \end{array}\right], \quad U = \left[\begin{array}{c|c} V_1 & V_2 \\ \hline 0 & S \end{array}\right]$$

where Π is the block even-odd premutation matrix and where the Schur complement $S = V_4 - V_3 V_1^{-1} V_2$ is the $(2^{q-1}+1) \times (2^{q-1}+1)$ block matrix

$$S = \left[\begin{array}{cccc} Z_1 & Y_1 & & 0 \\ Y_1 & Z_1 & \ddots & \\ & \ddots & \ddots & Y_1 \\ 0 & & Y_1 & Z_1 \end{array}\right].$$

Since $\Pi H_k \Pi^T$ is a nonsingular M-matrix, for the properties of Schur complements, it follows that S is a nonsingular M-matrix; in particular Z_1 is an M-matrix, and Y_1 has nonpositive entries. We can prove by induction on k that Z_k is a nonsingular M-matrix, and Y_k has nonpositive entries for all $k \geq 0$. In particular, our algorithm has the same stability properties as block Gaussian elimination applied to M-matrices.

The skew-Hamiltonian case Let us now consider the case where A is skew-Hamiltonian, i.e., $n = 2m$ and A has the structure

$$A = \begin{bmatrix} A_1 & A_2 \\ A_3 & A_1^T \end{bmatrix},$$

where A_1, A_2, A_3 are $m \times m$ matrices, and A_2, A_3 are skew-symmetric. This situation occurs in the numerical solution of algebraic Riccati equations, that is ultimately reduced to the computation of the square root of a matrix [18, 16].

Since A is skew-Hamiltonian, there exists an orthogonal matrix Q such that

$$B = Q^T A Q = \begin{bmatrix} B_1 & B_2 \\ 0 & B_1^T \end{bmatrix}, \tag{17}$$

where B_1 is upper Hessenberg and B_2 is skew-symmetric. Moreover, the reduced matrix $B = Q^T A Q$ can be computed in a numerically stable way by means of the Van Loan algorithm [17], and $A^{1/2} = Q B^{1/2} Q^T$. Indeed, the methods proposed in [18, 16] for computing the square root of a skew-Hamiltonian matrix A consist first in reducing A into the form (17), and then in approximating the square root of the reduced matrix B by exploiting the structure of B. Also our algorithm can be adapted for the computation of $B^{1/2}$, thus generating matrices which keep the structure of B, with a substantial reduction of the computational cost with respect to the direct application of the algorithm to the skew-Hamiltonian matrix A. Indeed, it is a simple matter to verify that the matrices $\{Y_k\}_k$ and $\{Z_k\}_k$, obtained by starting with $Y_0 = I - B$, $Z_0 = 2(I + B)$, have the structure:

$$Y_k = \begin{bmatrix} Y_{1,k} & Y_{2,k} \\ 0 & Y_{1,k}^T \end{bmatrix}, \quad Z_k = \begin{bmatrix} Z_{1,k} & Z_{2,k} \\ 0 & Z_{1,k}^T \end{bmatrix},$$

where $Y_{2,k}^T = -Y_{2,k}$, $Z_{2,k}^T = -Z_{2,k}$, and

$$\begin{aligned} Y_{1,k+1} &= -Y_{1,k} Z_{1,k}^{-1} Y_{1,k}, \\ Y_{2,k+1} &= -Y_{1,k} Z_{1,k}^{-1} Y_{2,k} + Y_{1,k} Z_{1,k}^{-1} Z_{2,k} Z_{1,k}^{-T} Y_{1,k}^T - Y_{2,k} Z_{1,k}^{-T} Y_{1,k}^T, \\ Z_{1,k+1} &= Z_{1,k} + 2Y_{1,k+1}, \\ Z_{2,k+1} &= Z_{2,k} + 2Y_{2,k+1}. \end{aligned}$$

4 Numerical experiments

We have performed several numerical experiments on an Athlon XP 2400, CPU at 2002 MHz, using MATLAB. We have compared our algorithm with:

1. The Denman-Beavers (DB) iteration [5]

$$\begin{aligned} &Y_0 = A, \ Z_0 = I, \\ &Y_{k+1} = (Y_k + Z_k^{-1})/2, \\ &Z_{k+1} = (Z_k + Y_k^{-1})/2, \ k = 0, 1, 2, \ldots \end{aligned}$$

   such that $Y_k \to A^{1/2}$, $Z_k \to A^{-1/2}$ as $k \to \infty$.

2. The scaled Denman-Beavers (DB) iteration [11]

$$\begin{aligned} &Y_0 = A, \ Z_0 = I, \\ &\gamma_k = |\det(Y_k)\det(Z_k)|^{1/(2n)}, \\ &Y_{k+1} = (\gamma_k Y_k + \gamma_k^{-1} Z_k^{-1})/2, \\ &Z_{k+1} = (\gamma_k Z_k + \gamma_k^{-1} Y_k^{-1})/2, \ k = 0, 1, 2, \ldots \end{aligned}$$

   such that $Y_k \to A^{1/2}$, $Z_k \to A^{-1/2}$ as $k \to \infty$.

3. The method based on Padè approximations [14, 13, 11]

$$\begin{aligned} &Y_0 = A, \ Z_0 = I, \\ &Y_{k+1} = \tfrac{1}{p} Y_k \textstyle\sum_{i=1}^{p} \tfrac{1}{\xi_i} (Z_k Y_k + \alpha_i^2 I)^{-1}, \\ &Z_{k+1} = \tfrac{1}{p} Z_k \textstyle\sum_{i=1}^{p} \tfrac{1}{\xi_i} (Y_k Z_k + \alpha_i^2 I)^{-1}, \ k = 0, 1, 2, \ldots \end{aligned}$$

where $p \geq 1$ is a chosen integer,

$$\xi_i = \frac{1}{2}\left(1 + \cos\frac{(2i-1)\pi}{2p}\right), \; \alpha_i^2 = \frac{1}{\xi_i} - 1, \quad i = 1, \ldots, p,$$

and $Y_k \to A^{1/2}$, $Z_k \to A^{-1/2}$ as $k \to \infty$.

4. The method based on scaled Padè approximations [11]

$$\begin{aligned}
&Y_0 = A, \; Z_0 = I, \\
&\gamma_k = |\det(Y_k)\det(Z_k)|^{1/(2n)}, \\
&Y_{k+1} = \tfrac{1}{p}\gamma_k Y_k \textstyle\sum_{i=1}^{p} \tfrac{1}{\xi_i}(\gamma_k^2 Z_k Y_k + \alpha_i^2 I)^{-1}, \\
&Z_{k+1} = \tfrac{1}{p}\gamma_k Z_k \textstyle\sum_{i=1}^{p} \tfrac{1}{\xi_i}(\gamma_k^2 Y_k Z_k + \alpha_i^2 I)^{-1}, \; k = 0, 1, 2, \ldots
\end{aligned}$$

5. The method based on polar decomposition (PD) [7], when A is real symmetric positive definite:
   - compute $A = R^T R$ the Cholesky decomposition;
   - compute $U = X_\infty$ from $X_0 = R$, $X_{k+1} = (X_k + X_k^{-T})/2$, $k = 0, 1, \ldots$.
   - set $A^{1/2} = U^T R$.

6. The method based on polar decomposition with scaling (PD scaled) [7], when A is real symmetric positive definite:
   - compute $A = R^T R$ the Cholesky decomposition;
   - compute $U = X_\infty$ from $X_0 = R$, $X_{k+1} = (\gamma_k X_k + \gamma_k^{-1} X_k^{-T})/2$, $\gamma_k = \left(\frac{\|X_k^{-1}\|_1 \, \|X_k^{-1}\|_\infty}{\|X_k\|_1 \|X_k\|_\infty}\right)^{1/4}$, $k = 0, 1, \ldots$.
   - set $A^{1/2} = U^T R$.

7. The `sqrtm` function of MATLAB, which uses the Schur algorithm [3, 9].

For each test matrix we have reported:

1. the condition number estimate $\mu(A^{1/2})$ of the matrix square root provided by the MATLAB function `sqrtm`;

2. the relative residual
$$\frac{||X^2 - A||}{||A||};$$

3. the number of iterations needed to reach that residual; we stopped the iterations when the residual error stopped to decrease significantly;

4. the relative error
$$\frac{||X - A^{1/2}||}{||A^{1/2}||},$$
   where $A^{1/2}$ has been computed by means of CR by using a precision floating point arithmetic with 40 decimal digit accuracy.

Method	Iters.	Residual	Error
CR	25	1.7e-15	9.3e-10
Padé unscaled $p=1$	17	4.6e-12	4.7e-7
$p=2$	9	2.8e-12	3.3e-7
$p=3$	7	2.6e-12	3.1e-7
$p=4$	6	2.0e-12	3.3e-7
Padé scaled $p=1$	12	3.7e-10	1.9e-10
$p=2$	8	2.9e-10	1.5e-10
$p=3$	6	1.4e-10	8.6e-11
$p=4$	5	8.6e-11	2.0e-10
DB unscaled	5	9.7e-04	2.4e-2
DB scaled	11	1.2e-06	5.7e-7
PD unscaled	17	5.8e-11	3.0e-6
PD scaled	8	1.6e-14	7.4e-12
`sqrtm`	*	1.8e-15	1.3e-10

Table 1: Prolate matrix, $n = 20$, $\mu(A^{1/2}) = 3.7 \cdot 10^6$

Here $||X||$ is the Frobenius norm of the matrix X.

Test 1 (Prolate matrix). We used the `prolate` function of the MATLAB `gallery` test matrices, which generates the Prolate matrix. It is a symmetric, ill-conditioned Toeplitz matrix, which depends on an input parameter w. If $0 < w < 0.5$ then the Prolate matrix is positive definite, its eigenvalues are distinct, lie in $(0,1)$, and tend to cluster around 0 and 1. We used the default value of w, i.e., $w = 0.25$ and the size $n = 20$; in this case the spectral condition number of A is $\kappa_2(A) = 5.7 \cdot 10^{13}$. The results are reported in Table 1. The smallest residuals are obtained by the CR algorithm, by PD scaled and by the `sqrtm` function. Also the errors of PD scaled are good. Since the condition number of $A^{1/2}$ is of the order of 10^6, we cannot expect a residual error better than 10^{-10}. Thus all the results, except the ones obtained with the DB iteration, are acceptable. The poor performance of the latter algorithm is due to the ill-conditioning of the matrix A, which must be inverted at the first step. The CR algorithm, in terms of accuracy, seems insensitive to the ill-conditioning of A. The large number of iterations is due to the fact that A is close to a singular matrix, and that CR converges linearly in the case where A is singular. For larger values of n the residual error of CR remains unchanged; the residual error of (scaled) DB and Padè methods grows with n. It is also interesting to point out that the residual error of CR, as a function of the number of iterations, reaches a minimum and after that it does not change in the subsequent iterations; for DB and Padè methods, the residual error can grow much in the subsequent iterations.

Test 2 (Frobenius matrix). For this test A is the 4×4 Frobenius matrix whose characteristic polynomial is $p(x) = \sum_{i=0}^{4} p_i x^i = (x-2)(x-5)\left((x+2)^2 + \epsilon\right)$,

Method	Iters.	Residual	Error
CR	19	7.5e-03	2.6e-5
Padé unscaled $p = 1$	24	7.7e-02	1.9e-5
$p = 2$	10	2.2e-03	5.0e-5
$p = 3$	12	6.1e-02	8.5e-6
$p = 4$	6	5.9e-03	4.2e-5
Padé scaled $p = 1$	15	3.2e-02	2.0e-5
$p = 2$	15	7.9e-02	1.8e-5
$p = 3$	12	2.9e-02	6.4e-7
$p = 4$	19	5.4e-03	1.5e-5
DB unscaled	20	2.0e-05	1.9e-9
DB scaled	13	1.4e-05	2.0e-9
`sqrtm`	*	6.3e-9	2.1e-8

Table 2: Frobenius matrix, $p(x) = (x-2)(x-5)\left((x+2)^2+\epsilon\right)$, $\epsilon = 10^{-8}$, $\mu(A^{1/2}) = 1.5 \cdot 10^{10}$

i.e.,

$$A = \begin{bmatrix} -p_3 & -p_2 & -p_1 & -p_0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix},$$

where ϵ is a small fixed real positive number, The matrix A has as eigenvalues the zeros of $p(x)$. In particular A has two complex conjugate eigenvalues $-2 \pm i\sqrt{\epsilon}$ close to the negative imaginary axis. The matrix A is well conditioned ($\mu(A) \approx 10$), while $A^{1/2}$ is ill-conditioned when ϵ is small. In this case the residual errors are large, due to the ill-conditioning of $A^{1/2}$, and the DB method performs better than the other algorithms. The results are reported in Table 2.

We tried also with the Frobenius matrix associated with the polynomial $p(x) = (x-2)(x-5)\left((x+1)^2+\epsilon\right)$. In this case A has two complex conjugate eigenvalues $-1 \pm i\sqrt{\epsilon}$ close to -1, thus the matrix $I + A$, which must be inverted at the first step of CR, is close to a singular matrix when ϵ is small. We tried with $\epsilon = 10^{-5}$. The results are reported in Table 3. In this case the DB and Padé methods performs better than the CR algorithm, which suffers from the ill-conditioning of the matrix $I + A$. This drawback could be overcome by computing the matrix square root of αA, where α is a positive real number such that $I + \alpha A$ is well conditioned, and then by scaling by $\alpha^{1/2}$ the obtained approximation.

Test 3 (Dorr matrix). This is the matrix generated by the `dorr` function of the MATLAB `gallery` test matrices. It is a row diagonally dominant, tridiagonal matrix that is ill-conditioned for small values of the input argument $\theta \geq 0$. We performed experiments with $\theta = 10^{-7}$ and size $n = 10$. The numerical results are reported in Table 4. We observe that the CR algorithm provides the lowest

Method	Iters.	Residual	Error
CR	13	1.4e-03	3.2e-7
Padé unscaled $p=1$	14	2.6e-07	1.2e-7
$p=2$	12	2.9e-09	7.2e-12
$p=3$	6	5.8e-08	1.1e-8
$p=4$	8	3.6e-10	6.3e-12
Padé scaled $p=1$	12	7.3e-09	5.6e-12
$p=2$	8	6.4e-09	6.3e-11
$p=3$	5	9.2e-09	3.8e-11
$p=4$	4	3.3e-09	4.5e-11
DB unscaled	15	5.8e-12	3.1e-12
DB scaled	12	2.4e-12	1.4e-12
`sqrtm`	*	2.4e-12	1.4e-12

Table 3: Frobenius matrix, $p(x) = (x-2)(x-5)\left((x+1)^2+\epsilon\right)$, $\epsilon = 10^{-5}$, $\mu(A^{1/2}) = 1.2 \cdot 10^6$

residual error, and the approximation provided by the DB method is very poor. Since the best approximations are provided by the CR algorithm and by the function `sqrtm`, we have compared the execution time of these two algorithms for increasing values of the size n, with the default value of θ. In Table 5 we have reported the CPU time in seconds and, into brackets, the number of iterations of CR. We observe that for small values of n the execution time of both algorithms is negligible, while for increasing values of n the advantage of the CR algorithm becomes larger.

Test 4 (Scaled matrix). For this test $A = \alpha R$, where α is a positive real number, and R is the randomly generated 5×5 matrix

$$R = \begin{bmatrix} 0.20277 & 0.015274 & 0.41865 & 0.83812 & 0.50281 \\ 0.19872 & 0.74679 & 0.84622 & 0.019640 & 0.70947 \\ 0.60379 & 0.44510 & 0.52515 & 0.68128 & 0.42889 \\ 0.27219 & 0.93181 & 0.20265 & 0.37948 & 0.30462 \\ 0.19881 & 0.46599 & 0.67214 & 0.83180 & 0.18965 \end{bmatrix}.$$

We have compared the number of iterations needed to reach the minimum residual error for small and large values of α. We have applied CR, Padè and scaled Padè (with $p=2$), DB and scaled DB. The number of iterations are reported in Figure 1. We observe that the scaling of DB and Padé methods keeps constant the number of iterations. For the nonscaled methods the number of iterations reaches a minimum for a certain $\bar{\alpha}$, and then it grows for larger and smaller values of $\bar{\alpha}$. The minimum number of iterations of CR coincides with the number of iterations of DB scaled.

Test 5 (Neumann matrix). We tested our algorithm on the matrix generated by the `neumann` function of the MATLAB `gallery` test matrices. It is the

Method	Iters.	Residual	Error
CR	25	7.0e-16	2.6e-8
Padé unscaled $p = 1$	6	1.7e-15	5.1e-9
$p = 2$	3	1.9e-15	5.1e-9
$p = 3$	2	2.5e-15	5.1e-9
$p = 4$	2	1.7e-15	5.1e-9
Padé scaled $p = 1$	11	6.4e-4	3.2e-4
$p = 2$	20	6.4e-8	5.3e-8
$p = 3$	9	6.3e-8	4.3e-8
$p = 4$	7	9.6e-9	8.6e-9
DB unscaled	1	1.1	4.5e-1
DB scaled	11	8.7e-3	9.3e-3
`sqrtm`	*	2.8e-15	1.2e-8

Table 4: Dorr matrix, $n = 10$, $\mu(A^{1/2}) = 3.0 \cdot 10^8$

n	CR	`sqrtm`
50	0.10 (12)	0.02
100	0.1 (12)	0.2
200	0.5 (11)	1.1
500	5.5 (10)	12.5
1000	40 (10)	114

Table 5: Dorr matrix, CPU time (sec.)

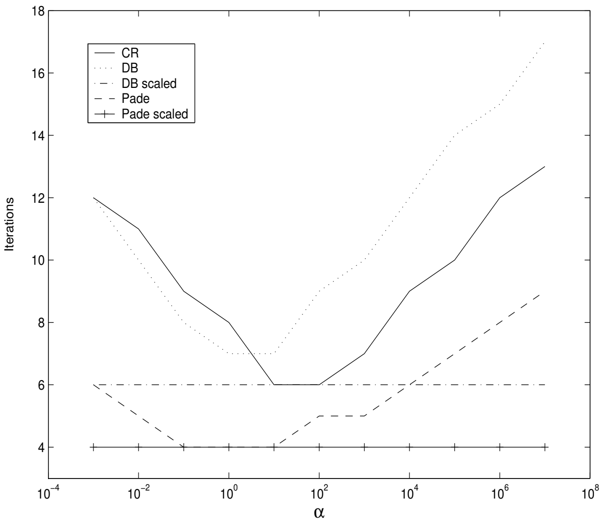

Figure 1: Scaled matrix: number of iterations

matrix resulting from discretizing the Neumann problem with the usual five point operator on a regular mesh. The matrix has a one-dimensional null space with null vector the vector of 1s. The CR algorithm shows a linear convergence, and reaches the residual error 2.9e-16 in 26 iterations. The matrix square root computed with the function `sqrtm` has a residual error 3.9e-15 and has complex entries, with imaginary part of magnitude 10^{-9}. The Padé method, with $p = 1$, reaches the residual error 2.4e-15 after 6 iterations. However it is not clear if the Padé method may have any break down for a general singular matrix.

5 Conclusions

We have given a new functional interpretation of the matrix square root of a matrix A, in terms of the inverse of a suitable matrix Laurent polynomial $R(z)$. This interpretation has allowed us to reduce the computation of the matrix square root to computing the constant block coefficient of $R(z)^{-1}$. For this purpose we have applied cyclic reduction, which, from the several numerical experiments, is the iterative method that provides the most accurate approximation of $A^{1/2}$. In particular, the cyclic reduction algorithm seems insensitive to the ill-conditioning of A, while other methods may fail to converge if A is ill-conditioned. In fact, cyclic reduction converges also in the important case where A is singular. An open question, which we are investigating, is to understand if a similar functional interpretation holds also for the p-th matrix square root, for $p > 2$. This would open the way to new algorithms for computing the matrix p-th root.

Acknowledgments. The author wishes to thank N. J. Higham for providing useful suggestions to improve the presentation of this paper, and D. A. Bini and L. Gemignani for the useful discussions. In particular, the relation (16) with Newton's method was observed by L. Gemignani.

References

[1] D. A. Bini, L. Gemignani, and B. Meini. Computations with infinite Toeplitz matrices and polynomials. *Linear Algebra Appl.*, 343/344:21–61, 2002. Special issue on structured and infinite systems of linear equations.

[2] D. A. Bini and B. Meini. On the solution of a nonlinear matrix equation arising in queueing problems. *SIAM J. Matrix Anal. Appl.*, 17:906–926, 1996.

[3] Åke Björck and Sven Hammarling. A Schur method for the square root of a matrix. *Linear Algebra Appl.*, 52/53:127–140, 1983.

[4] B. L. Buzbee, G. H. Golub, and C. W. Nielson. On direct methods for solving Poisson's equation. *SIAM J. Num. Anal.*, 7:627–656, 1970.

[5] Eugene D. Denman and Alex N. Beavers, Jr. The matrix sign function and computations in systems. *Appl. Math. Comput.*, 2(1):63–94, 1976.

[6] C. R. DePrima and C. R. Johnson. The range of $A^{-1}A^*$ in $\mathrm{GL}(n, C)$. *Linear Algebra and Appl.*, 9:209–222, 1974.

[7] Nicholas J. Higham. Computing the polar decomposition—with applications. *SIAM J. Sci. Statist. Comput.*, 7(4):1160–1174, 1986.

[8] Nicholas J. Higham. Newton's method for the matrix square root. *Math. Comp.*, 46(174):537–549, 1986.

[9] Nicholas J. Higham. Computing real square roots of a real matrix. *Linear Algebra Appl.*, 88/89:405–430, 1987.

[10] Nicholas J. Higham. The matrix sign decomposition and its relation to the polar decomposition. *Linear Algebra Appl.*, 212/213:3–20, 1994. Proceedings of the 3rd ILAS Conference (Pensacola, FL, 1993).

[11] Nicholas J. Higham. Stable iterations for the matrix square root. *Numer. Algorithms*, 15(2):227–242, 1997.

[12] Roger A. Horn and Charles R. Johnson. *Topics in matrix analysis*. Cambridge University Press, Cambridge, 1994. Corrected reprint of the 1991 original.

[13] C. S. Kenney and A. J. Laub. A hyperbolic tangent identity and the geometry of Padé sign function iterations. *Numer. Algorithms*, 7(2-4):111–128, 1994.

[14] Charles Kenney and Alan J. Laub. Rational iterative methods for the matrix sign function. *SIAM J. Matrix Anal. Appl.*, 12(2):273–291, 1991.

[15] Pentti Laasonen. On the iterative solution of the matrix equation $AX^2 - I = 0$. *Math. Tables Aids Comput.*, 12:109–116, 1958.

[16] L. Lu and C. E. M. Pearce. On the square-root method for continuous-time algebraic Riccati equations. *J. Austral. Math. Soc. Ser. B*, 40(4):459–468, 1999.

[17] C. F. Van Loan. A symplectic method for approximating all the eigenvalues of a Hamiltonian matrix. *Linear Algebra Appl.*, 61:233–251, 1984.

[18] Hong Guo Xu and Lin Zhang Lu. Properties of a quadratic matrix equation and the solution of the continuous-time algebraic Riccati equation. *Linear Algebra Appl.*, 222:127–145, 1995.