Computing primary solutions of equations involving primary matrix functions

Fasi, Massimiliano and Iannazzo, Bruno

2018

ISSN 1749-9097

Computing primary solutions of equations involving primary matrix functions[*]

Massimiliano Fasi[†] and Bruno Iannazzo[‡]

Abstract

The matrix equation $f(X) = A$, where f is an analytic function and A is a square matrix, is considered. Some results on the classification of solutions are provided. When f is rational, a numerical algorithm is proposed to compute all solutions that can be written as a polynomial of A. For real data, the algorithm yields the real solutions using only real arithmetic. Numerical experiments show that the algorithm performs in a stable fashion when run in finite precision arithmetic.

Keywords: Schur normal form, block triangular matrices, substitution algorithm, matrix equation, matrix function.

2010 MSC: 15A16, 15A24, 65F60.

1 Introduction

We consider the matrix equation

$$f(X) = A, \tag{1}$$

where $A, X \in \mathbb{C}^{N\times N}$ and f is a complex function applied to a matrix (in the sense of primary matrix functions, see Section 2). Remarkable examples of (1) are the matrix equations $X^k = A$, $e^X = A$, and $Xe^X = A$, which define the matrix kth root [22, 16], the matrix logarithm [1], and the matrix Lambert W function [8], respectively. Existence and finiteness of real and complex solutions to (1) are discussed, along with other properties of this matrix equation, in the excellent treatise by Evard and Uhlig [7].

In order to better understand the computational properties of the matrices that satisfy (1), it is useful to distinguish the solutions that can be written as a polynomial of A, or *primary* solutions, from those that cannot, called *nonprimary*. A useful characterization of primary solutions in terms of their eigenvalues is provided in [7].

After discussing some further properties of primary solutions, we focus our attention on *isolated* solutions, that is, solutions that are unique in a neighborhood. We show that nonprimary solutions are not isolated, characterize isolated solutions in terms of their eigenvalues, and show that they are in fact primary solutions with some additional properties.

[*] Version of November 22, 2018. Funding: This work was supported by the Istituto Nazionale di Alta Matematica, INdAM–GNCS Project 2017; and MathWorks. The opinions and views expressed in this publication are those of the authors, and not necessarily those of the funding bodies.

[†] School of Mathematics, The University of Manchester, Oxford Road, Manchester M13 9PL, UK (massimiliano.fasi@manchester.ac.uk).

[‡] Dipartimento di Matematica e Informatica, Università di Perugia, Via Vanvitelli 1, 06123 Perugia, Italy (bruno.iannazzo@unipg.it).

Turning to numerical computation, we restrict our attention to the equation

$$r(X) = A, \tag{2}$$

where $r = p/q$, and p and q are polynomials. The algorithm we propose is designed in the spirit of and generalizes the method developed by Björk and Hammarling [3] for the square root of a matrix, tailored for the real case by Higham [13] and extended to the kth root by Smith [22].

First, we consider the case of block upper triangular A and develop an algorithm that, using a sequence of substitutions, computes a primary solution to (2) given its diagonal blocks. Next we discuss how the Schur decomposition, which reduces any matrix to block upper triangular form with a similarity transformation, can be exploited to extend our approach to general matrices, and show that the algorithm, if no breakdown occurs, computes a primary solution, given its eigenvalues. Finally, we show that the algorithm is applicable with no breakdown if and only if there exists a unique solution with given diagonal blocks (which correspond to a given set of eigenvalues), which, moreover, is proved to be equivalent to requiring that the solution is isolated.

Being restricted to isolated solutions is not a severe limitation, since solutions that are not isolated are typically of little or no computational interest. Indeed a solution $\widetilde{X}$ that is not isolated is either nonprimary or *ill-posed*, in the sense that there exists a neighborhood $\mathcal{U}_{\widetilde{X}}$ of $\widetilde{X}$ and a matrix E, such that the perturbed equation $r(X) = A + tE$ has no solution in $\mathcal{U}_{\widetilde{X}}$ for any sufficiently small $t > 0$. For instance, when computing the square root of a matrix A with the algorithm of Björk and Hammarling, one requires that, if A is singular, then the eigenvalue zero is simple [3], which is a necessary and sufficient condition for a primary solution to $X^2 = A$ to be isolated. Primary square roots can exist when the zero eigenvalue has multiplicity larger than one, but in this case they are not isolated, and there exist arbitrarily small perturbations of A having no square root.

In the next section, we provide some background material, and in the following we give some theoretical results regarding the solutions of matrix equations of the type (1). In Section 4, we consider (2) and present our algorithm for block upper triangular matrices, discussing both the complex and the real Schur form. Section 5 is devoted to numerical experiments that illustrate the numerical behavior of our algorithm, and in Section 6 we draw some conclusions and discuss lines of future research.

2 Background and notation

Polynomials and rational functions. By convention, a summation is equal to zero if the starting index exceeds the ending one. We denote by $\mathbb{C}[x]$ the polynomials of the complex variable x with complex coefficients, and by $\mathbb{C}_k[x] \subset \mathbb{C}[x]$ the complex polynomials of degree at most k. Let $p(x) := \sum_{k=0}^{m} c_k x^k \in \mathbb{C}_m[x]$ and $q(x) := \sum_{k=0}^{n} d_k x^k \in \mathbb{C}_n[x]$ be coprime polynomials with nonzero leading coefficients. The quotient $r(x) := p(x)q(x)^{-1}$ is a rational function of type $[m, n]$. In the following sections, when using p, q or r, we will always refer to the functions defined above, and in particular, $c_0, \dots, c_m$ will denote the coefficients of p and $d_0, \dots, d_n$ those of q.

In order to evaluate a polynomial p at a point x_0, we make use of Horner's evaluation scheme [10, Alg. 9.2.1], that is, we define the polynomials $p^{[j]}(x) = \sum_{i=0}^{m-j} c_{i+j} x^i$, for $j = 0, \dots, m$, and evaluate $p^{[0]}(x_0) = p(x_0)$ by means of the recursion

$$\begin{aligned} p^{[m]}(x_0) &= c_m, \\ p^{[j]}(x_0) &= x_0 p^{[j+1]}(x_0) + c_j, \qquad \text{for } j = 0, \dots, m-1. \end{aligned}$$

Let $f : \Omega \to \mathbb{C}$, where $\Omega \subset \mathbb{C}$, and let $x, y \in \Omega$. We denote by $f[x, y]$ the divided difference operator, defined by

$$f[x,y] = \begin{cases} f'(x), & x = y, \\ \dfrac{f(x) - f(y)}{x - y}, & x \neq y, \end{cases}$$

which implicitly requires f to be differentiable at x, when $x = y$. The divided differences over $k + 1$ numbers, ordered so that equal numbers are contiguous, are

$$f[x_0, \dots, x_k] = \begin{cases} \dfrac{f^{(k)}(x_0)}{k!}, & x_0 = x_1 = \cdots = x_k, \\ \dfrac{f[x_1, \dots, x_k] - f[x_0, \dots, x_{k-1}]}{x_k - x_0}, & \text{otherwise.} \end{cases}$$

This definition can be extended to any set of $k + 1$ numbers by assuming that the divided differences are symmetric functions of their arguments. For the construction above to make sense, the function has to be differentiable t times at any point repeated $t + 1$ times.

Primary matrix functions. Let $A \in \mathbb{C}^{N\times N}$ and let $Z \in \mathbb{C}^{N\times N}$ be such that $Z^{-1}AZ = J = \operatorname{diag}(J(\lambda_1, \tau_1), \dots, J(\lambda_\nu, \tau_\nu))$ is the Jordan canonical form of A, with

$$J(\lambda, m) := \begin{bmatrix} \lambda & 1 & & \\ & \lambda & \ddots & \\ & & \ddots & 1 \\ & & & \lambda \end{bmatrix} \in \mathbb{C}^{m\times m},$$

where missing entries should be understood as zeros. In order to simplify the notation, we will often omit the diagonal element and the size of the Jordan block and write J_i for $J(\lambda_i, m_i)$. The index of the eigenvalue λ, denoted by $\iota(\lambda)$, is the size of the largest Jordan block where λ appears. An eigenvalue with index one is said to be *semisimple*, otherwise it is said to be *defective*; a semisimple eigenvalue appearing in only one block is said to be *simple*.

Let the complex function f and its derivatives up to the order $\iota(\lambda_k) - 1$ be defined at λ_k for $k = 1, 2, \dots, \nu$. Then we can define the primary matrix function

$$f(A) := Zf(J)Z^{-1} = Z \operatorname{diag}(f(J_1), f(J_2), \dots, f(J_\nu))Z^{-1}, \tag{3}$$

where

$$f(J_k) = \begin{bmatrix} f(\lambda_k) & f'(\lambda_k) & \dots & \frac{f^{(m_k-1)}(\lambda_k)}{(m_k-1)!} \\ & f(\lambda_k) & \ddots & \vdots \\ & & \ddots & f'(\lambda_k) \\ & & & f(\lambda_k) \end{bmatrix}.$$

This definition does not depend on the matrix Z, and it can be shown that if f is a primary matrix function then $f(M^{-1}AM) = M^{-1}f(A)M$ for any M invertible and A such that $f(A)$ is well-defined. We will refer to this fundamental property as *commutativity with similarities*, and it will be used throughout the paper. A consequence is that if $A = \operatorname{diag}(A_1, \dots, A_\nu)$ is block diagonal, then $f(A) = \operatorname{diag}(f(A_1), \dots, f(A_\nu))$.

Moreover, it is easy to show that $f(A)$ as defined in (3) coincides with a polynomial that interpolates f in the Hermite sense on the spectrum of A [14, Rem. 1.10]. Therefore, if $T \in \mathbb{C}^{N\times N}$ is block upper triangular, then $f(T)$ has the same block structure as T, and if $T_{11}, \dots, T_{\nu\nu}$ are the diagonal blocks of T, then the diagonal blocks of $f(T)$ are $f(T_{11}),\dots,f(T_{\nu\nu})$. An explicit formula for the function of an upper triangular matrix is [10, Thm. 9.1.4]

$$\begin{aligned} (f(T))_{ii} &= f(t_{ii}), && 1 \le i \le n, \\ (f(T))_{ij} &= \sum_{i_1=i<i_2<\dots<i_\ell=j} t_{i_1 i_2} t_{i_2 i_3} \cdots t_{i_{\ell-1} i_\ell} f[t_{i_1 i_1},\dots,t_{i_\ell i_\ell}], && 1 \le i < j \le n, \end{aligned} \tag{4}$$

where the sum is over all increasing sequences of integers starting with i and ending with j.

Let J be a nontrivial Jordan block in which the eigenvalue λ appears. The Jordan canonical form of $f(J)$ consists of:

1. only one Jordan block associated with $f(\lambda)$, if $f'(\lambda) \neq 0$;
2. two or more Jordan blocks associated with $f(\lambda)$, if $f'(\lambda) = 0$.

In the latter case, we say that the function f *splits the Jordan block* J. A complete description of the Jordan canonical form of $f(A)$ in terms of that of A is given in [15, sect. 6.2.25].

The Frechét derivative of a matrix function $f : \Omega \to \mathbb{C}^{N\times N}$ at a point $A \in \Omega \subset \mathbb{C}^{N\times N}$ is the linear functional $Df(A) : \mathbb{C}^{N\times N} \to \mathbb{C}^{N\times N}$ that satisfies

$$f(A+E) = f(A) + Df(A)[E] + o(\|E\|),$$

for any $E \in \mathbb{C}^{N\times N}$ with sufficiently small norm.

A measure of the sensitivity of matrix function, with respect to perturbation of the argument A, is given by the relative condition number, which, for any subordinate norm $\|\cdot\|$, is defined as [14, eq. (3.2)]

$$\kappa_f(A) = \lim_{\varepsilon\to 0} \sup_{\|E\|\le\varepsilon\|A\|} \frac{\|f(A+E)-f(A)\|}{\varepsilon\|f(A)\|}. \tag{5}$$

We conclude this section with a lemma and a corollary that will be useful later on.

Lemma 1. *Let $A \in \mathbb{C}^{N\times N}$ be upper bidiagonal, let e_i, for $i = 1,\dots,N$, be the standard basis of $\mathbb{C}^N$, and let $f(A)$ and $f[a_{11},a_{NN}]$ be well-defined. Then for any $\delta \in \mathbb{C}$ we have that*

$$f(A + \delta e_1 e_N^T) = f(A) + \delta f[a_{11},a_{NN}] e_1 e_N^T. \tag{6}$$

Proof. Let $T := A + \delta e_1 e_N^T$ and $F = f(T)$. Partitioning $T = \left[\begin{smallmatrix} T_1 & v \\ 0 & t_{NN} \end{smallmatrix}\right]$, with $T_1 \in \mathbb{C}^{(N-1)\times(N-1)}$, from the properties of primary matrix functions, we have that $F = \left[\begin{smallmatrix} f(T_1) & \widetilde{v} \\ 0 & f(t_{NN}) \end{smallmatrix}\right]$ and thus that $(F)_{ij} = (f(A))_{ij}$, for $j < N$. Using the partition $T = \left[\begin{smallmatrix} t_{11} & w \\ 0 & T_2 \end{smallmatrix}\right]$, with $T_2 \in \mathbb{C}^{(N-1)\times(N-1)}$, we get that $(F)_{ij} = (f(A))_{ij}$ for $i < N$. By using (4), for the top right element of the matrix we have

$$(F)_{1N} = \sum_{i_1=1<i_2<\dots<i_\ell=N} t_{i_1 i_2} t_{i_2 i_3} \cdots t_{i_{\ell-1} i_\ell} f[t_{i_1 i_1}, t_{i_2 i_2},\dots,t_{i_\ell i_\ell}].$$

Since $t_{ij} = 0$ for $i < j-1$ and $(i,j) \neq (1,N)$, the sum can be restricted to the two sequences $i_1 = 1, i_2 = 2, \dots, i_N = N$ and $i_1 = 1, i_2 = N$, giving

$$(F)_{1N} = t_{12} \dots t_{N-1,N} f[t_{11},\dots,t_{NN}] + t_{1N} f[t_{11},t_{NN}] = (f(A))_{1N} + \delta f[a_{11},a_{NN}],$$

which concludes the proof of the identity (6). □

Corollary 2. *We have the following relations, with $\delta \in \mathbb{C}$:*

(a) *if $A = \lambda I \in \mathbb{C}^{N\times N}$, and f is differentiable at λ, then $f(A + \delta e_1 e_N^T) = f(A) + \delta f'(\lambda) e_1 e_N^T$;*

(b) *if $A = J(\lambda, N)$, and f is differentiable at λ, then $f(A + \delta e_1 e_N^T) = f(A) + \delta f'(\lambda) e_1 e_N^T$;*

(c) *if $A = \operatorname{diag}(J(\lambda,k), J(\mu, N-k))$, with $\lambda \neq \mu$ and $1 \le k < N$, and f is well-defined at A, then $f(A + \delta e_1 e_N^T) = f(A) + \delta f[\lambda,\mu] e_1 e_N^T$.*

3 Classification of the solutions

The matrix equation $f(X) = A$, with f analytic, may have zero, finitely many, or infinitely many solutions. All these scenarios are possible, and here we are concerned with a classification of the solutions in terms of properties that are relevant from a computational viewpoint. In Section 3.1 we relate the notion of primary solution to that of primary matrix function, in Section 3.2, we consider isolated solutions and characterize them in several ways, and we conclude by briefly discussing critical solutions in Section 3.3.

3.1 Primary solutions

The matrices that satisfy (1) may define a function of the matrix A, but in general solutions to (1) need not be primary functions of A, in the sense of Section 2. The matrix $\left[\begin{smallmatrix}0&1\\0&0\end{smallmatrix}\right]$, for instance, satisfies the 2×2 matrix equation $X^2 = 0$, but is not a primary function of the zero matrix. Broadly speaking, solutions to a matrix equation can be divided into two classes, those that are primary functions of A and those that are not. In this section, we give some clarifications on this topic.

Let $A \in \mathbb{C}^{N\times N}$, let $f : \Omega \to \mathbb{C}$ be a function analytic on the open set $\Omega \subseteq \mathbb{C}$ and let $X \in \mathbb{C}^{N\times N}$ be a solution to $f(X) = A$ such that f is defined on the spectrum of X. A solution is *primary* if it can be written as a polynomial of A, and *nonprimary* otherwise.

A necessary and sufficient condition for a solution to be primary is provided by the following result, where an eigenvalue ξ of the solution X is said to be *critical* if $f'(\xi) = 0$.

Theorem 3 (Evard and Uhlig [7, Thm. 6.1]). *A solution $X \in \mathbb{C}^{N\times N}$ to the equation $f(X) = A$ is primary if and only if the following two conditions are true:*

1. *for any two distinct eigenvalues ξ_1 and ξ_2 of X, we have $f(\xi_1) \neq f(\xi_2)$;*
2. *all critical eigenvalues of X (if any) are semisimple.*

The definition of primary solution as a polynomial of A is related to the concept of primary function of a matrix. Informally, we could say that any primary solution is obtained as "an inverse of f applied to the matrix A". We now make this notion precise.

Let $\lambda_1, \dots, \lambda_s$ be the distinct eigenvalues of A, ordered so that the first t are semisimple and the remaining are not. We say that a solution X is *primary in the sense of functions* if $X = \widehat{f}^{-1}(A)$ where $\widehat{f}^{-1} : \{\lambda_1, \dots, \lambda_t\} \cup \mathcal{U} \to \mathbb{C}$ is analytic on an open set $\mathcal{U} \supseteq \{\lambda_{t+1}, \dots, \lambda_s\}$ and is such that $(f \circ \widehat{f}^{-1})(z) = \mathrm{id}(z)$ for any $z \in \{\lambda_1, \dots, \lambda_t\} \cup \mathcal{U}$.

Requiring that $\widehat{f}^{-1}$ is analytic on the eigenvalues that correspond to nontrivial Jordan blocks of A guarantees that $\widehat{f}^{-1}$ is defined on the spectrum of A and thus that $\widehat{f}^{-1}(A)$ is well-defined in the sense of (3). These two definitions are in fact the same, as the following proposition shows.

Proposition 4 (Equivalence of definitions of primary solution). *Let f be a complex function analytic on $\Omega \subset \mathbb{C}$ and let $A \in \mathbb{C}^{N\times N}$. A solution $X \in \mathbb{C}^{N\times N}$ to $f(X) = A$ with eigenvalues in Ω can be written as a polynomial of A if and only if it is primary in the sense of functions, i.e., if and only if $X = f^{-1}(A)$, where f^{-1} is an inverse of f defined on the spectrum of A and analytic at the defective eigenvalues of A.*

Proof. Assume that $X = f^{-1}(A)$ for some inverse of f. Since $f^{-1}(A)$ is a primary function of A, there exists a polynomial $p \in \mathbb{C}[x]$ such that $X = f^{-1}(A) = p(A)$, which implies that X is a primary solution to $f(X) = A$.

Conversely, suppose that $X = p(A)$ for some $p \in \mathbb{C}[x]$. From $f(X) = f(p(A)) = A$, it follows that $f(p(\lambda)) = \lambda$ for any eigenvalue λ of A. By taking $\widehat{f}^{-1}(\lambda) = p(\lambda)$ for any λ, it is enough to show that if λ is not semisimple, then $\widehat{f}^{-1}(\lambda)$ can be extended analytically in a neighborhood of λ to an inverse of f, and that $X = \widehat{f}^{-1}(A)$.

Let J be a nontrivial Jordan block of A in which the eigenvalue λ appears. From $f(p(A)) = A$ it follows that $f(p(J)) = J$, which entails that $(f \circ p)'(\lambda) \neq 0$, as $f \circ p$ would otherwise split the Jordan block. The latter inequality implies, in turn, that $f'(p(\lambda)) \neq 0$ and thus that f is invertible in a neighborhood of $p(\lambda) = \widehat{f}^{-1}(\lambda)$ with analytic inverse [9, sect. 4.6]. Thus, we can extend $\widehat{f}^{-1}$ in an open neighborhood of λ to a function such that $f \circ \widehat{f}^{-1} = \mathrm{id}$.

In order to prove that $X = \widehat{f}^{-1}(A)$, it suffices to show that $(\widehat{f}^{-1})^{(k)}(\lambda) = p^{(k)}(\lambda)$ for $k = 1, \dots, \ell - 1$, where ℓ is the size of $\widetilde{J}$, the largest Jordan block in which λ appears. First observe that $f(p(\widetilde{J})) = \widetilde{J}$ implies that $(f \circ p)^{(k)}(\lambda) = \mathrm{id}^{(k)}(\lambda)$ and thus that $(f \circ p)^{(k)}(\lambda) = (f \circ \widehat{f}^{-1})^{(k)}(\lambda)$, for $k = 1, \dots, \ell - 1$, since $(f \circ \widehat{f}^{-1})^{(k)}(\lambda) = \mathrm{id}^{(k)}(\lambda)$.

Next, we show by induction that for any $k > 0$ and any function g such that $f \circ g$ is analytic in a neighborhood of λ, one has that $(f \circ g)^{(k)}(\lambda) = f'(g(\lambda))g^{(k)}(\lambda) + h_k(g; \lambda)$, where h_k is a polynomial in $f''(g(\lambda)), \dots, f^{(k)}(g(\lambda)), g(\lambda), g'(\lambda), \dots, g^{(k-1)}(\lambda)$. Choosing $h_1(g; \lambda) = 0$ verifies the equality for $k = 1$, whereas for the inductive step we have

$$\begin{aligned}(f \circ g)^{(k+1)}(\lambda) &= f'(g(\lambda))g^{(k+1)}(\lambda) + f''(g(\lambda))g'(\lambda)g^{(k)}(\lambda) + h_k'(g; \lambda)\\ &=: f'(g(\lambda))g^{(k+1)}(\lambda) + h_{k+1}(g; \lambda),\end{aligned}$$

where $h_{k+1}(g; \lambda)$ is a polynomial in $f''(g(\lambda)), \dots, f^{(k+1)}(g(\lambda)), g(\lambda), \dots, g^{(k)}(\lambda)$.

Finally, we can prove that $(\widehat{f}^{-1})^{(k)}(\lambda) = p^{(k)}(\lambda)$ for $k = 0, \dots, \ell - 1$. For $k = 0$, this holds by definition of $\widehat{f}^{-1}$, while for $k < \ell - 1$, from $(f \circ p)^{(k+1)}(\lambda) = (f \circ \widehat{f}^{-1})^{(k+1)}(\lambda)$ we have that $f'(p(\lambda))p^{(k+1)}(\lambda) + h_{k+1}(p; \lambda) = f'(\widehat{f}^{-1}(\lambda))(\widehat{f}^{-1})^{(k+1)}(\lambda) + h_{k+1}(\widehat{f}^{-1}; \lambda)$. By the inductive hypothesis $h_{k+1}(g; \lambda) = h_{k+1}(\widehat{f}^{-1}; \lambda)$, and since $f'(\widehat{f}^{-1}(\lambda)) = f'(p(\lambda)) \neq 0$, we can conclude that $p^{(k+1)}(\lambda) = (\widehat{f}^{-1})^{(k+1)}(\lambda)$. □

Another property of nonprimary solutions is that they are not isolated, as we show in the next section.

3.2 Isolated solutions

A solution X to $f(X) = A$ is *isolated* if there exists a neighborhood $\mathcal{U}$ of X where the matrix equation has a unique solution. We will characterize isolated solution in several ways, and will start by showing that nonprimary solutions are not isolated.

Theorem 5. *Let $A \in \mathbb{C}^{N \times N}$, and let $X \in \mathbb{C}^{N \times N}$ be a nonprimary solution to the matrix equation $f(X) = A$ where f is a complex function analytic at the spectrum of X. Then X is not isolated. Moreover, the set of solutions is unbounded and there are infinitely many solutions having the same spectrum as X.*

Proof. In view of Theorem 3, if X is nonprimary, then necessarily either one of its critical eigenvalues, ξ say, is defective, or f takes the same value at two distinct eigenvalues $\xi_i \neq \xi_j$.

If ξ is defective, then there exists an invertible matrix M such that $M^{-1}XM = \begin{bmatrix} J_1 & 0 \\ 0 & J_2 \end{bmatrix}$, where $J_2 = J(\xi, k)$ is a Jordan block of size $k > 1$ associated with ξ. Using the notation of Corollary 2, define the parametrized matrix

$$X(\delta) := M \begin{bmatrix} J_1 & 0 \\ 0 & J_2 + \delta e_1 e_k^T \end{bmatrix} M^{-1},$$

with $\delta \in \mathbb{C}$. Noticing that

$$f(X(\delta)) = M \begin{bmatrix} f(J_1) & 0 \\ 0 & f(J_2 + \delta e_1 e_k^T) \end{bmatrix} M^{-1} = M \begin{bmatrix} f(J_1) & 0 \\ 0 & f(J_2) \end{bmatrix} M^{-1} = f(X),$$

where the second equality follows from Corollary 2(*b*) with $f'(\xi) = 0$, shows that $X(\delta)$ is a solution to $f(X) = A$ for any δ, and since $\lim_{\delta \to 0} X(\delta) = X$, we conclude that X is not isolated.

If the matrix has two distinct eigenvalues $\xi_i \neq \xi_j$ such that $f(\xi_i) = f(\xi_j)$, the proof is similar, and it suffices to consider the block $J_2 = \left[\begin{smallmatrix} J(\xi_1,k_1) & 0 \\ 0 & J(\xi_2,k_2) \end{smallmatrix}\right]$, with $k_1, k_2 \geq 1$, and use Corollary 2(*c*) with $f[\xi_i, \xi_j] = 0$.

In both cases, for any $\delta \in \mathbb{C}$ the matrix $X(\delta)$ has the same spectrum as X by construction, and the set $\{X(\delta) : \delta \in \mathbb{C}\}$ is infinite and unbounded. □

The converse of the Theorem 5 is not true. Indeed, the set of isolated solutions may be a strict subset of primary solution. The next results provides several interesting characterizations of the isolated solutions of $f(X) = A$. As we will see, the algorithm we introduce in Section 4 to solve (2), can compute a solution if and only if it is isolated.

Theorem 6. *Let $f : \Omega \to \mathbb{C}$ be an analytic non-constant function in the domain $\Omega \subset \mathbb{C}$. Let $A \in \mathbb{C}^{N \times N}$, and let $X \in \mathbb{C}^{N \times N}$ be a solution to $f(X) = A$, with eigenvalues $\xi_1, \ldots, \xi_N$ in Ω. The following are equivalent:*

(*a*) *X is isolated;*

(*b*) *X is primary with simple or no critical eigenvalues, that is,*

1. *for any two distinct eigenvalues ξ_i and ξ_j of X, we have $f(\xi_i) \neq f(\xi_j)$;*
2. *all critical eigenvalues of X (if any) are simple;*

(*c*) *X is the unique solution with eigenvalues $\xi_1, \ldots, \xi_N$;*

(*d*) *$f[\xi_i, \xi_j] \neq 0$ for $i, j = 1, \ldots, N$, with $i \neq j$.*

Proof. (*a*) ⇒ (*b*). By Theorem 5, if X is isolated, then it is primary, and we need to prove only that all its critical eigenvalues are simple (we know that they are semisimple by Theorem 3). By contradiction, assume that ξ is a semisimple critical eigenvalue of X with multiplicty at least 2. Then there exists an invertible matrix M such that $M^{-1}XM = \left[\begin{smallmatrix} J_1 & 0 \\ 0 & J_2 \end{smallmatrix}\right]$, where $J_2 = \xi I$, where I has size $\ell > 1$. With the notation of Corollary 2, the matrix

$$X(\delta) = M \begin{bmatrix} J_1 & 0 \\ 0 & J_2 + \delta e_1 e_\ell^T \end{bmatrix} M^{-1}$$

is a solution to $f(X) = A$ for any $\delta \in \mathbb{C}$, since

$$f(X(\delta)) = M \begin{bmatrix} f(J_1) & 0 \\ 0 & f(J_2 + \delta e_1 e_\ell^T) \end{bmatrix} M^{-1} = M \begin{bmatrix} f(J_1) & 0 \\ 0 & f(J_2) \end{bmatrix} M^{-1} = f(X)$$

where the second equality follows from Corollary 2(*a*) with $f'(\xi) = 0$. Since $\lim_{\delta \to 0} X(\delta) = X$, X is not isolated.

(*b*) ⇒ (*c*). The eigenvalues of A are the image under f of the eigenvalues of any solution, in particular, they are $f(\xi_1), \ldots, f(\xi_N)$. Assume that X is primary with simple critical eigenvalues, and let Y be a solution with the same eigenvalues as X. This implies that Y has simple critical eigenvalues, and that $f(\xi_i) \neq f(\xi_j)$ for any pair of distinct eigenvalues $\xi_i \neq \xi_j$. Therefore, by

Theorem 3, Y must be primary with simple critical eigenvalues; moreover, the images of these critical eigenvalues are simple eigenvalues of A as well. In particular, the defective eigenvalues of A (if any) are image of noncritical eigenvalues of X. By Proposition 4, $Y = \widehat{f}^{-1}(A)$ and $X = f^{-1}(A)$, where both $\widehat{f}^{-1}$ and f^{-1} are inverses of f and are analytic at the images of noncritical eigenvalues of X (and Y), and thus are analytic at the defective eigenvalues of A. Since $\widehat{f}^{-1}(\lambda) = f^{-1}(\lambda)$ for any eigenvalue λ of A, and the two functions are analytic and coincide in a neighborhood of λ if the eigenvalue is defective (the inverse of an analytic function is unique), we have that $Y = \widehat{f}^{-1}(A) = f^{-1}(A) = X$.

$(c) \Rightarrow (a)$. Without loss of generality, we may assume that if $\xi_i \neq \xi_k$ then $f(\xi_i) \neq f(\xi_k)$, since otherwise the solution X would be nonprimary and, by Theorem 5, there would be solutions other than X, but with the same spectrum as X.

Since the eigenvalues of A are $f(\xi_1), \dots, f(\xi_N)$, any solution to $f(X) = A$ has eigenvalues $\zeta_1, \dots, \zeta_N$ such that $f(\zeta_1) = f(\xi_1), \dots, f(\zeta_N) = f(\xi_N)$. Let $\{\tau_j^{(i)}\}_{j \in \mathcal{J}_i}$ be the (possibly empty) set of solutions to $f(x) = f(\xi_i)$ other than ξ_i. If $\mathcal{J}_i$ is empty for each i, then the eigenvalues of any solution must be $\xi_1, \dots, \xi_N$ and X is the unique solution, hence it is isolated.

Let us now assume that some of the $\mathcal{J}_i$ are nonempty. If $\mathcal{J}_i$ is nonempty, then $\tau_j^{(i)} \neq \xi_k$ for each j and k: this is true by definition when $\xi_i = \xi_k$, and when $\xi_i \neq \xi_k$, by the assumption above, since $f(\tau_j^{(i)}) = f(\xi_i) \neq f(\xi_k)$. Moreover, since the zeros of a non-constant analytic function cannot have accumulation points in the domain of analyticity [4, sect. 143], ξ_k cannot be an accumulation point of the set $\{\tau_j^{(i)}\}_{j \in \mathcal{J}_i}$ and hence $\varepsilon_{i,k} := \inf_{j \in \mathcal{J}_i} |\tau_j^{(i)} - \xi_k|$ must be positive for each k. Set $\varepsilon := \min_{i\,:\,\mathcal{J}_i \neq \emptyset} \min_{k=1,\dots,N} \varepsilon_{i,k} > 0$.

A solution $Y \neq X$, must have at least one eigenvalue of the type $\widehat{\tau} := \tau_j^{(i)}$ for some i and j. If that is the case, then $\min_{k=1,\dots,N} |\widehat{\tau} - \xi_k| \geq \varepsilon$ or, in other words, at least one eigenvalue of Y has distance at least ε from any eigenvalue of X. On the other hand, since the eigenvalues are continuous functions of the entries of a matrix, there exists a neighborhood $\mathcal{U}$ of X, such that for any $Z \in \mathcal{U}$, we have $\max_{\eta \in \sigma(Z)} \min_{k=1,\dots,N} |\eta - \xi_k| < \varepsilon/2$, where $\sigma(Z)$ is the spectrum of Z. Therefore Y does not belong to $\mathcal{U}$, and X is isolated.

$(b) \Leftrightarrow (d)$. A necessary and sufficient condition for $f[\xi_i, \xi_j] = 0$ for $i \neq j$, is that either $\xi_i \neq \xi_j$, with $f(\xi_i) = f(\xi_j)$ or $\xi_i = \xi_j$ and $f'(\xi_i) = 0$. These two conditions are equivalent to X being either nonprimary or primary with multiple critical eigenvalues. □

We observe that, when a primary solution X of $f(X) = A$ is not isolated, the corresponding solution X is *ill-posed*, that is, a small perturbation of A may produce an equation that has no solutions near X.

By Theorem 6, a primary solution X that is not isolated has at least one semisimple eigenvalue ξ with multiplicity $k > 1$ and such that $f'(\xi) = 0$. Hence $\lambda = f(\xi)$ is a semisimple eigenvalue of A, with the same multiplicity as ξ since X is primary. There exists a nonsingular matrix M such that $M^{-1}AM = \left[\begin{smallmatrix} J & 0 \\ 0 & \lambda I_k \end{smallmatrix}\right]$, where λ is not an eigenvalues of J. For $\varepsilon > 0$, the perturbed equation $f(X) = A(\varepsilon)$ where

$$A(\varepsilon) = M \operatorname{diag}\left(J, \begin{bmatrix} \lambda & \varepsilon & & \\ & \ddots & \ddots & \\ & & \lambda & \varepsilon \\ & & & \lambda \end{bmatrix} \right) M^{-1},$$

has no solutions with eigenvalue ξ. Indeed, since primary matrix functions split Jordan blocks in presence of critical eigenvalues, if there exists $X(\varepsilon)$ such that $f(X(\varepsilon)) = A(\varepsilon)$, it must have an eigenvalue μ, such that $f(\mu) = \lambda$ and $f'(\mu) \neq 0$, which in turn implies that $\mu \neq \xi$. Therefore, the solution $X(\varepsilon)$ can be ruled out from a sufficiently small neighborhood of X.

3.3 Critical solutions

Let f be an analytic complex function and let $Df(M) : \mathbb{C}^{N\times N} \to \mathbb{C}^{N\times N}$ be the Frechét derivative of f at the matrix $M \in \mathbb{C}^{N\times N}$. A solution X to the equation $f(X) = A$ is said to be *critical* if $Df(X)$ is singular, and *noncritical* otherwise. We may easily characterize critical solutions.

Proposition 7. *Let f be a complex function, let $A \in \mathbb{C}^{N\times N}$, and let $X \in \mathbb{C}^{N\times N}$ be a solution to the matrix equation $f(X) = A$. If f is differentiable at X, then the derivative $Df(X)$ is nonsingular if and only if the following two conditions are fulfilled:*

1. *for any two distinct eigenvalues ξ_i and ξ_j of X, we have $f(\xi_i) \neq f(\xi_j)$;*
2. *none of the eigenvalues of X is critical for f.*

Moreover, these conditions are equivalent to the condition that $f[\xi_i, \xi_j] \neq 0$, for $i, j = 1, \dots, N$, where $\xi_1, \dots \xi_N$ are the eigenvalues of X.

Proof. Observe that the two conditions hold if and only if the divided differences of any two eigenvalues of X is not zero. Since the the eigenvalues of $Df(X)$ are the divided differences of eigenvalues of X [14, Thm. 3.9], this is equivalent to requiring that $Df(X)$ is nonsingular. □

A further property of nonprimary solutions is that of being critical.

Proposition 8. *Let f be an analytic complex function, let $A \in \mathbb{C}^{N\times N}$, and let $X \in \mathbb{C}^{N\times N}$ be a nonprimary solution to the matrix equation $f(X) = A$. Then $Df(X)$ is singular.*

Proof. In view of Theorem 3, if X in not primary, then X has either two distinct eigenvalues ξ_i and ξ_j such that $f(\xi_i) = f(\xi_j)$ and thus $f[\xi_i, \xi_j] = 0$, or a defective eigenvalue ξ such that $f[\xi, \xi] = f'(\xi) = 0$. Since the eigenvalues of $Df(X)$ are the divided differences of two eigenvalues of X [14, Thm. 3.9], both cases yield a singular derivative. □

Notice that there might be primary or even isolated solutions that are critical: those with critical eigenvalues.

4 A substitution algorithm for rational equations

Given $A \in \mathbb{C}^{N\times N}$, we want to find the primary solutions $X \in \mathbb{C}^{N\times N}$ to (2). To this end, we first reduce this equation to

$$p(X) = Aq(X), \tag{7}$$

then consider a (block) triangular form of A, such as the Schur form, and devise an algorithm to compute the entries of X. We begin by showing that (2) and (7) are equivalent.

In the scalar case, if p and q are coprime, then a root of p cannot be a root of q and vice versa, and thus that the scalar equation $\frac{p(x)}{q(x)} = a$ has a solution if and only if $p(x) = aq(x)$ does. The matrix version of this implication is also true, as the following result shows.

Proposition 9. *Let $p \in \mathbb{C}_m[x], q \in \mathbb{C}_n[x]$ be coprime. Then $X \in \mathbb{C}^{N\times N}$ is a solution to $p(X)q(X)^{-1} = A$ if and only if it satisfies $p(X) = A\,q(X)$.*

Proof. If X is such that $p(X)q(X)^{-1} = A$, then $p(X) = Aq(X)$. For the other implication, first note that if X is such that $p(X) = Aq(X)$ and $q(X)$ is nonsingular, then $p(X)q(X)^{-1} = A$, hence it is enough to show that $q(X)$ is nonsingular.

For the sake of contradiction, assume that $q(X)$ is singular. Then there exists a nonzero vector $b \in \mathbb{C}^N$, such that $q(X)b = 0$, and thus $p(X)b = Aq(X)b = 0$. Since the set $\mathcal{I} = \{s \in \mathbb{C}[x] \, : \, s(X)b = 0\}$ is an ideal in a principal ideal domain, it is generated by a minimal polynomial $s(x) \in \mathbb{C}[x]$, that is not constant since $b \neq 0$ and thus $\mathcal{I} \neq \mathbb{C}[x]$. Hence, $s(x)|q(x)$ and $s(x)|p(x)$, which leads to a contradiction since $p(x)$ and $q(x)$ are coprime. □

Let us consider a similarity transformation that reduces A to a block upper triangular matrix $U^{-1}AU =: T = [T_{ij}]_{i,j=1,\ldots,\nu} \in \mathbb{C}^{N \times N}$, where $T_{ij} \in \mathbb{C}^{\tau_i \times \tau_j}$, with $\sum_{i=1}^{\nu} \tau_i = N$ and $T_{ij} = 0$ for $i > j$.

We are mostly interested in the Schur decomposition, where U is unitary and T is upper triangular, and, for $A \in \mathbb{R}^{N \times N}$, in the the real Schur decomposition, where U is real orthogonal and T is upper quasi-triangular. Nevertheless, we prefer to work in greater generality, as a different blocking strategy may allow for more efficient implementations of the algorithms (for instance, in order to exploit caching and parallelism in modern computer architectures).

Since matrix polynomials commute with similarities, X is a solution to (2) if and only if $Y := U^{-1}XU$ satisfies $r(Y) = T$, and in view of Proposition 9, in order to solve (2) we can work with the simpler matrix equation $p(Y) = Tq(Y)$. By exploiting Horner's scheme for polynomial evaluation [10, Alg. 9.2.1], we can rewrite the latter equation as $P^{[0]} = TQ^{[0]}$, where $P^{[0]} = p(Y)$ and $Q^{[0]} = q(Y)$, are defined recursively by

$$\begin{aligned} P^{[0]} &= c_0 I + YP^{[1]}, & Q^{[0]} &= d_0 I + YQ^{[1]}, \\ P^{[1]} &= c_1 I + YP^{[2]}, & Q^{[1]} &= d_1 I + YQ^{[2]}, \\ &\vdots & &\vdots \\ P^{[m-1]} &= c_{m-1} I + YP^{[m]}, & Q^{[n-1]} &= d_{n-1} I + YQ^{[n]}, \\ P^{[m]} &= c_m I, & Q^{[n]} &= d_n I. \end{aligned} \tag{8}$$

If we look for primary solutions only, we may assume that Y is block upper triangular with the same block structure as T, which implies in turn that all $P^{[u]}$s and $Q^{[v]}$s have the same block upper triangular structure. We adopt the following notation: for a matrix M with the same block structure as T, we denote by M_{ij} the block in position (i, j) of M.

We assume that the ν blocks along the diagonal of Y are known, for instance they can be deduced by a direct formula when the size is 1 or 2. Note that in most cases the diagonal blocks can be chosen in several ways, and that this choice determines what solution the algorithm will compute among all those that are primary. We discuss these points in details in the next section.

The blocks along the diagonal of $P^{[u]}$, for $u = 0, \ldots, m-1$, and $Q^{[v]}$, for $v = 0, \ldots, n-1$, can be uniquely determined by means of (8), and in order to compute the blocks in the upper triangular part of Y, $P^{[u]}$ and $Q^{[v]}$, note that for $1 \leq i < j \leq \nu$, we have

$$\begin{aligned} P_{ij}^{[u]} &= \sum_{k=i}^{j} Y_{ik} P_{kj}^{[u+1]} = Y_{ii} P_{ij}^{[u+1]} + Y_{ij} P_{jj}^{[u+1]} + \sum_{k=i+1}^{j-1} Y_{ik} P_{kj}^{[u+1]}, & u &= 0, \ldots, m-1, \\ Q_{ij}^{[v]} &= \sum_{k=i}^{j} Y_{ik} Q_{kj}^{[v+1]} = Y_{ii} Q_{ij}^{[v+1]} + Y_{ij} Q_{jj}^{[v+1]} + \sum_{k=i+1}^{j-1} Y_{ik} Q_{kj}^{[v+1]}, & v &= 0, \ldots, n-1. \end{aligned} \tag{9}$$

By substituting (9) for $P_{ij}^{[u+1]}$ and $Q_{ij}^{[v+1]}$ into those for $P_{ij}^{[u]}$ and $Q_{ij}^{[v]}$, respectively, and recursively repeating this procedure, we get, as shown in the following proposition, an expression where Y_{ij} appears together with blocks of Y, $P^{[u]}$, and $Q^{[v]}$ lying to the left of the block in position (i, j)

or below it. This discussion translates immediately into a two-step algorithm for computing Y: first compute the diagonal blocks and then compute the off-diagonal blocks a superdiagonal at a time.

Proposition 10. *Let $p \in \mathbb{C}_m[x]$ and $q \in \mathbb{C}_n[x]$ be coprime, let $T \in \mathbb{C}^{N\times N}$ be block upper triangular, let $Y \in \mathbb{C}^{N\times N}$ be a solution to the matrix equation $p(Y) = Tq(Y)$ with the same block structure as T, and let $P^{[u]}, Q^{[v]} \in \mathbb{C}^{N\times N}$, for $u = 0,\dots,m$ and $v = 0,\dots,n$, be as in (8). Then $P^{[u]}$ and $Q^{[v]}$ have the same block structure as T, and their off-diagonal blocks, for $1 \le i < j \le \nu$, are given by the formulae*

$$\begin{aligned} P_{ij}^{[u]} &= \sum_{e=1}^{m-u} Y_{ii}^{e-1} Y_{ij} P_{jj}^{[u+e]} + \sum_{f=1}^{m-u-1} Y_{ii}^{f-1} C_{ij}^{[u+f]}, \qquad u = 0,\dots,m-1, \\ Q_{ij}^{[v]} &= \sum_{g=1}^{n-v} Y_{ii}^{g-1} Y_{ij} Q_{jj}^{[v+g]} + \sum_{h=1}^{n-v-1} Y_{ii}^{h-1} D_{ij}^{[v+h]}, \qquad v = 0,\dots,n-1, \end{aligned} \tag{10}$$

where

$$C_{ij}^{[u]} = \sum_{k=i+1}^{j-1} Y_{ik} P_{kj}^{[u]}, \qquad D_{ij}^{[v]} = \sum_{k=i+1}^{j-1} Y_{ik} Q_{kj}^{[v]}.$$

Moreover, one has the following

$$\sum_{e=1}^{m} Y_{ii}^{e-1} Y_{ij} P_{jj}^{[e]} - T_{ii} \sum_{g=1}^{n} Y_{ii}^{g-1} Y_{ij} Q_{jj}^{[g]} = \sum_{k=i+1}^{j} T_{ik} Q_{kj}^{[0]} - \sum_{f=1}^{m-1} Y_{ii}^{f-1} C_{ij}^{[f]} + T_{ii} \sum_{h=1}^{n-1} Y_{ii}^{h-1} D_{ij}^{[h]}. \tag{11}$$

Proof. The two claims in (10) can be proved by induction on an auxiliary variable k. We limit ourselves to the recurrence for $P^{[u]}$, the proof for $Q^{[v]}$ being analogous. For $u = m-1$, equation (10) reduces to $P_{ij}^{[m-1]} = c_m Y_{ij}$, which follows directly from the definition of $P^{[m-1]}$ in (8). For the inductive step, we have, for $1 < k \le m$,

$$\begin{aligned} P_{ij}^{[m-k]} &= Y_{ii} P_{ij}^{[m-k+1]} + Y_{ij} P_{jj}^{[m-k+1]} + \sum_{k=i+1}^{j-1} Y_{ik} P_{kj}^{[m-k+1]} \\ &= \sum_{e=2}^{k} Y_{ii}^{e-1} Y_{ij} P_{jj}^{[m-k+e]} + \sum_{f=2}^{k-1} Y_{ii}^{f-1} C_{ij}^{[m-k+f]} + Y_{ii}^{0} Y_{ij} P_{jj}^{[m-k+1]} + Y_{ii}^{0} C_{ij}^{[m-k+1]} \\ &= \sum_{e=1}^{k} Y_{ii}^{e-1} Y_{ij} P_{jj}^{[m-k+e]} + \sum_{f=1}^{k-1} Y_{ii}^{f-1} C_{ij}^{[m-k+f]}. \end{aligned}$$

In order to establish (11), note that one can rewrite $P^{[0]} = TQ^{[0]}$ as

$$P_{ij}^{[0]} - T_{ii} Q_{ij}^{[0]} = \sum_{k=i+1}^{j} T_{ik} Q_{kj}^{[0]}.$$

Substituting (10) for $P_{ij}^{[0]}$ and $Q_{ij}^{[0]}$ and rearranging the terms concludes the proof. □

4.1 Complex Schur form

When $T \in \mathbb{C}^{N\times N}$ is upper triangular, the blocks along the diagonal of T are of size 1×1 and $\nu = N$. Equation (11) involves just scalars and can be written as $\psi_{ij}y_{ij} = \varphi_{ij}$, where

$$\psi_{ij} := \sum_{e=1}^{m} y_{ii}^{e-1} p_{jj}^{[e]} - t_{ii} \sum_{g=1}^{n} y_{ii}^{g-1} q_{jj}^{[g]}, \tag{12}$$

and

$$\varphi_{ij} := \sum_{k=i+1}^{j} t_{ik} q_{kj}^{[0]} - \sum_{f=1}^{m-1} y_{ii}^{f-1} C_{ij}^{[f]} + t_{ii} \sum_{h=1}^{n-1} y_{ii}^{h-1} D_{ij}^{[h]}. \tag{13}$$

If t_{ii} is a diagonal element of T, then for $i = 1,\dots,N$, y_{ii} will be any of the at most $\max(m,n)$ distinct roots of the polynomial $p(x) - t_{ii}q(x) = 0$. In order to compute the off-diagonal elements of Y, we can see the relation $\psi_{ij}y_{ij} = \varphi_{ij}$ as an equation

$$\psi_{ij}x = \varphi_{ij}, \tag{14}$$

whose unique solution is y_{ij} when $\psi_{ij} \neq 0$ and the values y_{hk} with $h-k < i-j$ are known quantities.

We give necessary and sufficient conditions for (14) to have unique solution, and relate them to the characterization of isolated solutions given in Section 3. We start with a couple of technical lemmas, then we give the main theorem.

Lemma 11. *Let $p(x) = \sum_{i=0}^{m} c_i x^i$, let $a, b \in \mathbb{C}$ and let $p^{[k]}(x) = \sum_{i=0}^{m-k} c_{k+i}x^i$, for $k = 0,\dots,m$, be the sequence of stages of Horner's rule applied to p. Then*

$$\chi := \sum_{k=1}^{m} a^{k-1}p^{[k]}(b) = p[a,b]. \tag{15}$$

Proof. By definition of $p[a,b]$, we have to prove that $\chi = p'(a)$ if $a = b$ and $\chi = \frac{p(a)-p(b)}{a-b}$ if $a \neq b$. In both cases we have

$$\sum_{k=1}^{m} a^{k-1}p^{[k]}(b) = \sum_{k=1}^{m} a^{k-1}\Bigl(\sum_{i=0}^{m-k} c_{k+i}b^i\Bigr) = \sum_{\ell=1}^{m} c_\ell \Bigl(\sum_{k=0}^{\ell-1} a^k b^{\ell-k-1}\Bigr).$$

If $a = b$, then we get

$$\sum_{k=1}^{m} a^{k-1}p^{[k]}(b) = \sum_{\ell=1}^{m} c_\ell \ell a^{\ell-1} = p'(a),$$

whereas, for $a \neq b$ we have

$$\sum_{k=1}^{m} a^{k-1}p^{[k]}(b) = \sum_{\ell=1}^{m} c_\ell \frac{a^\ell - b^\ell}{a-b} = \frac{1}{a-b}\Bigl(\sum_{\ell=0}^{m} c_\ell a^\ell - \sum_{\ell=0}^{m} c_\ell b^\ell\Bigr) = \frac{p(a)-p(b)}{a-b}.$$

□

Lemma 12. *Let $p(x) = \sum_{i=0}^{m} c_i x^i$ and $q(x) = \sum_{j=0}^{n} d_j x^j$, let $r = p/q$, and let $a, b \in \mathbb{C}$ be such that $q(a) \neq 0$ and $q(b) \neq 0$. Then*

$$\psi := \sum_{i=1}^{m} a^{i-1}p^{[i]}(b) - r(a)\sum_{j=1}^{n} a^{j-1}q^{[j]}(b) \neq 0$$

if and only if either $a \neq b$ and $r(a) \neq r(b)$ or $a = b$ and $r'(a) \neq 0$.

Proof. By Lemma 11, when $a \neq b$ we have that

$$\psi = \frac{p(a) - p(b) - r(a)(q(a) - q(b))}{a - b}, \tag{16}$$

which is nonzero if and only if

$$p(a) - p(b) - \frac{p(a)}{q(a)}(q(a) - q(b)) \neq 0, \tag{17}$$

or equivalently

$$r(a) \neq r(b).$$

On the other hand, if $a = b$, then

$$\psi = p'(a) - r(a)q'(a) = \frac{p'(a)q(a) - p(a)q'(a)}{q(a)} = r'(a)q(a), \tag{18}$$

which is nonzero if and only if $r'(a) \neq 0$. □

Theorem 13. *Let $T \in \mathbb{C}^{N\times N}$ be upper triangular, let p, q, Y, $P^{[u]}$, for $u = 0, \dots, m$, and $Q^{[v]}$, for $v = 0, \dots, n$, be as in Proposition 10, and let $r(x) = p(x)q(x)^{-1}$. Then equation (14) has a unique solution y_{ij} for all $1 \leq i < j \leq N$ if and only if $r[y_{ii}, y_{jj}]q(y_{jj}) \neq 0$.*

Proof. It is enough to show that for ψ_{ij} defined in (12), we have that $\psi_{ij} = r[y_{ii}, y_{jj}]q(y_{jj})$. If $y_{ii} = y_{jj}$, then the proof is the same as in (18). When $y_{ii} \neq y_{jj}$, by using (16), we get that

$$\psi_{ij} = \frac{-p(y_{jj}) + p(y_{ii})q(y_{jj})/q(y_{ii})}{y_{ii} - y_{jj}} = \frac{r(y_{ii}) - r(y_{jj})}{y_{ii} - y_{jj}} q(y_{jj}) = r[y_{ii}, y_{jj}]q(y_{jj}).$$ □

Corollary 14 (Applicability of the Schur algorithm for isolated solutions)**.** *Let $r = p/q$ be a rational function, with $p \in \mathbb{C}_m[x]$ and $q \in \mathbb{C}_n[x]$ coprime, and let $Y \in \mathbb{C}^{N\times N}$ be a solution to $r(Y) = T$, with $T \in \mathbb{C}^{N\times N}$ upper triangular. Let $P^{[u]}$ for $u = 0, \dots, m$, and $Q^{[v]}$ for $v = 0, \dots, n$, be as in (9). Then the following two conditions are equivalent:*

(*a*) *Y is an isolated solution;*

(*b*) *the Schur algorithm is applicable and computes Y, if we choose y_{ii} as solution of the equation $p(x) - t_{ii}q(x) = 0$, for $i = 1, \dots, N$, that is, equation (14) has y_{ij} as unique solution, for $1 \leq i < j \leq N$.*

Proof. By Theorem 13, (14) has unique solution if and only if $r[y_{ii}, y_{jj}]q(y_{jj}) \neq 0$, for $1 \leq i < j \leq N$. Proposition 9 ensures that $q(y_{jj}) \neq 0$, for $j = 1, \dots, N$, since $q(Y)$ is nonsingular (recall that the eigenvalues of $q(Y)$ are $q(y_{11}), \dots, q(y_{NN})$). Thus, equation (14) has a unique solution if and only if $r[y_{ii}, y_{jj}] \neq 0$ for $1 \leq i < j \leq N$, which in turn, by the symmetry of divided differences, is equivalent Theorem 6(*d*), that is equivalent to requiring that Y is isolated. □

These results show that if we focus on a primary solution with simple critical eigenvalues, then we can compute the solution to the triangular equation $r(Y) = T$, by first computing the diagonal elements of Y, taking care of choosing the same branch for the same eigenvalue of T, and then computing the elements y_{ij}, for $i < j$, by means of (14), one superdiagonal at a time. This is the basis of Algorithm 1, which we call the Schur algorithm.

We now discuss the cost of the algorithm. Computing the Schur decomposition of a square matrix of size N and recovering the result require $25N^3$ and $3N^3$ flops, respectively. The for loop at line 2 requires $O\big((m + n)N\big)$ flops, those on line 13 and 15 require $(m - 1)N^3/3$ and $(n-1)N^3/3$, respectively, and evaluating the expression on line 17 requires $N^3/3$ flops. All the other operations within the loop on line 10 require $O\big((m + n)N^2\big)$. Therefore the asymptotic cost of the algorithm is $\big(28 + \frac{m+n-1}{3}\big)N^3$.

Algorithm 1: Schur algorithm for rational matrix equations.

```
Input  : $A \in \mathbb{C}^{N\times N}$, $c \in \mathbb{C}^{m+1}$ coefficients of $p$, $d \in \mathbb{C}^{n+1}$ coefficients of $q$.
Output: $X \in \mathbb{C}^{N\times N}$ such that $p(X)q^{-1}(X) \approx A$.
Compute the complex Schur decomposition $A := UTU^*$.
for $i = 1$ to $N$ do
    $y_{ii} \leftarrow$ a solution to $p(x) - t_{ii}q(x) = 0$
    $p_{ii}^{[m-1]} \leftarrow c_{m-1} + c_m y_{ii}$
    for $u = m-2$ down to $0$ do
        $p_{ii}^{[u]} \leftarrow c_u + y_{ii} p_{ii}^{[u+1]}$
    $q_{ii}^{[n-1]} \leftarrow d_{n-1} + d_n y_{ii}$
    for $v = n-2$ down to $0$ do
        $q_{ii}^{[v]} \leftarrow d_v + y_{ii} q_{ii}^{[v+1]}$
for $\ell = 1$ to $N-1$ do
    for $i = 1$ to $N-\ell$ do
        $j \leftarrow i + \ell$
        for $f = 1$ to $m-1$ do
            $C_{ij}^{[f]} = \sum_{k=i+1}^{j-1} y_{ik} p_{kj}^{[f]}$
        for $h = 1$ to $n-1$ do
            $D_{ij}^{[h]} = \sum_{k=i+1}^{j-1} y_{ik} q_{kj}^{[h]}$
        $rhs \leftarrow \sum_{k=i+1}^{j} t_{ik} q_{kj}^{[0]} - \sum_{f=1}^{m-1} y_{ii}^{f-1} C_{ij}^{[f]} + t_{ii} \sum_{h=1}^{n-1} y_{ii}^{h-1} D_{ij}^{[h]}$
        $lhs \leftarrow \sum_{e=1}^{m} y_{ii}^{e-1} p_{jj}^{[e]} - t_{ii} \sum_{g=1}^{n} y_{ii}^{g-1} q_{jj}^{[g]}$
        $y_{ij} \leftarrow rhs/lhs$
        $p_{ij}^{[m-1]} \leftarrow c_m y_{ij}$
        for $u = m-2$ down to $1$ do
            $p_{ij}^{[u]} \leftarrow y_{ii} p_{ij}^{[u+1]} + y_{ij} p_{jj}^{[u+1]} + C_{ij}^{[u+1]}$
        $q_{ij}^{[n-1]} \leftarrow d_n y_{ij}$
        for $v = n-2$ down to $0$ do
            $q_{ij}^{[v]} \leftarrow y_{ii} q_{ij}^{[v+1]} + y_{ij} q_{jj}^{[v+1]} + D_{ij}^{[v+1]}$
$X \leftarrow UYU^*$
```

Remark. Corollary 14 shows that our algorithm cannot compute primary solutions with semisimple critical eigenvalues with multiplicity greater than one. We now describe how the algorithm can be modified in order to compute these ill-posed solutions.

Let Y be a primary solution to $r(Y) = T$ and let $\xi_1, \dots, \xi_s$, with $s > 0$, be its critical, and thus semisimple, eigenvalues with multiplicities $\nu_1, \dots, \nu_s$, greater than one. We have that $\lambda_\ell = r(\xi_\ell)$, for $\ell = 1, \dots, s$, is a semisimple eigenvalue of T with the same multiplicity as ξ_ℓ (the multiplicity cannot be larger since Y is primary).

Using the procedure described in [2], it is possible to reorder the matrix T so that, for $\ell = 1, \dots, s$, the occurrences of λ_ℓ are adjacent along the diagonal of T. By doing so, we get a new matrix $\widetilde{T} = Q^*TQ$, where Q is the unitary matrix that performs the reordering. Since λ_ℓ

is semisimple, the diagonal block of $\widetilde{T}$ corresponding to λ_ℓ is $\lambda_\ell I$, and we get

$$\widetilde{T} = \begin{bmatrix} \widetilde{T}_{11} & * & \cdots & * \\ & \lambda_1 I & \ddots & \vdots \\ & & \ddots & * \\ & & & \lambda_s I \end{bmatrix},$$

where the asterisks represent possibly nonzero blocks and $\widetilde{T}_{11}$ is a triangular block collecting all the eigenvalues other than $\lambda_1, \dots, \lambda_s$.

Any solution $\widetilde{Y}$ to $r(\widetilde{Y}) = \widetilde{T}$ yields the solution $Y = Q\widetilde{Y}Q^*$ of $r(Y) = T$, with the same eigenvalues. Moreover, since $\widetilde{Y}$ is a primary function of $\widetilde{T}$, it has the structure

$$\widetilde{Y} = \begin{bmatrix} \widetilde{Y}_{11} & * & \cdots & * \\ & \xi_1 I & \ddots & \vdots \\ & & \ddots & * \\ & & & \xi_s I \end{bmatrix},$$

where, $\xi_i \neq \xi_j$ for $i \neq j$, and $\widetilde{Y}_{11}$ collects all the eigenvalues not in the set $\{\xi_1, \dots, \xi_s\}$.

This implies that $\widetilde{y}_{ij} = 0$ when $\widetilde{y}_{ii} = \widetilde{y}_{jj} = \xi_\ell$ for some ℓ, and thus we can determine $\widetilde{y}_{ij}$, without solving (11), while (11) can be used for all other entries of the upper triangular part of $\widetilde{Y}$, for which the solution is unique. Therefore, in principle, any primary solution could be computed using (a variation) of Algorithm 1, but in practice, the problem is ill-posed and we focus our attention on solutions with simple critical eigenvalues.

4.2 Real Schur form

When $A \in \mathbb{R}^{N\times N}$ and one is interested in real solutions to (2), in order to use real arithmetic only, we consider the real Schur decomposition $A := UTU^T$, where $U \in \mathbb{R}^{N\times N}$ is orthogonal, and $T \in \mathbb{R}^{N\times N}$ is upper quasi-triangular and has $\nu \leq N$ diagonal blocks of size either 1×1 or 2×2. In the former case, the diagonal block Y_{ii} can be computed as discussed in the previous section. Otherwise, we can rely on the following result.

Proposition 15. *Let $M \in \mathbb{R}^{2\times 2}$, and let $V \in \mathbb{R}^{2\times 2}$ be such that $V^{-1}MV = \operatorname{diag}(\mu, \overline{\mu})$, for some $\mu = a + ib$, with $b \neq 0$. Let $f : \{\mu, \overline{\mu}\} \to \mathbb{C}$ be a function such that $f(\overline{\mu}) = \overline{f(\mu)}$, and let $f(\mu) = c + id$. Then*

$$f(M) = \frac{d}{b}M + \Big(c - \frac{ad}{b}\Big)I. \tag{19}$$

Proof. It is well known [14, Thm. 1.12] that $f(M)$ coincides with the interpolating polynomial of f at the eigenvalues of M, that is

$$p(x) = f(\mu)\frac{x - \overline{\mu}}{\mu - \overline{\mu}} + f(\overline{\mu})\frac{x - \mu}{\overline{\mu} - \mu} = \frac{f(\mu) - f(\overline{\mu})}{\mu - \overline{\mu}}x + \frac{\mu f(\overline{\mu}) - \overline{\mu} f(\mu)}{\mu - \overline{\mu}}. \tag{20}$$

By replacing the definitions of μ and $f(\mu)$ and simplifying, one obtains (19). □

In order to compute the off-diagonal blocks of Y, we need to solve for the block Y_{ij} the matrix equation (11), which, by using the vec operator, can be rewritten as the linear system

$$M_{ij} \operatorname{vec}(Y_{ij}) = \operatorname{vec}\left(\sum_{k=i+1}^{j} T_{ik} Q_{kj}^{[0]} - \sum_{f=1}^{m-1} Y_{ii}^{f-1} C_{ij}^{[f]} + T_{ii} \sum_{h=1}^{n-1} Y_{ii}^{h-1} D_{ij}^{[h]} \right),$$

Algorithm 2: Real Schur algorithm for rational matrix equations.

```
Input  : $A \in \mathbb{C}^{N \times N}$, $c \in \mathbb{C}^{m+1}$ coefficients of $p$, $d \in \mathbb{C}^{n+1}$ coefficients of $q$.
Output: $X \in \mathbb{C}^{N \times N}$ such that $p(X)q^{-1}(X) \approx A$.
Compute the real Schur decomposition $A := UTU^*$.
for $i = 1$ to $\nu$ do
    $Y_{ii} \leftarrow$ a solution to $p(X) - T_{ii}q(X) = 0$
    $P_{ii}^{[m-1]} \leftarrow c_{m-1}I_{\tau_i} + c_m Y_{ii}$
    for $u = m-2$ down to $0$ do
        $P_{ii}^{[u]} \leftarrow c_u I_{\tau_i} + Y_{ii}P_{ii}^{[u+1]}$
    $Q_{ii}^{[n-1]} \leftarrow d_{n-1}I_{\tau_i} + d_n Y_{ii}$
    for $v = n-2$ down to $0$ do
        $Q_{ii}^{[v]} \leftarrow d_v I_{\tau_i} + Y_{ii}Q_{ii}^{[v+1]}$
for $\ell = 1$ to $N-1$ do
    for $i = 1$ to $N-\ell$ do
        $j \leftarrow i + \ell$
        for $f = 1$ to $m-1$ do
            $C_{ij}^{[f]} = \sum_{k=i+1}^{j-1} Y_{ik}P_{kj}^{[f]}$
        for $h = 1$ to $n-1$ do
            $D_{ij}^{[h]} = \sum_{k=i+1}^{j-1} Y_{ik}Q_{kj}^{[h]}$
        $s_{ij} \leftarrow \mathrm{vec}\left(\sum_{k=i+1}^{j} T_{ik}Q_{kj}^{[0]} - \sum_{f=1}^{m-1} Y_{ii}^{f-1}C_{ij}^{[f]} + T_{ii}\sum_{h=1}^{n-1} Y_{ii}^{h-1}D_{ij}^{[h]}\right)$
        $M_{ij} \leftarrow \sum_{e=1}^{m} \left(P_{jj}^{[e]}\right)^T \otimes Y_{ii}^{e-1} - \sum_{g=1}^{n} \left(Q_{jj}^{[g]}\right)^T \otimes \left(T_{ii}Y_{ii}^{g-1}\right)$
        $\mathrm{vec}(Y_{ij}) \leftarrow M_{ij}^{-1}s_{ij}$
        $P_{ij}^{[m-1]} \leftarrow c_m Y_{ij}$
        for $u = m-2$ down to $1$ do
            $P_{ij}^{[u]} \leftarrow Y_{ii}P_{ij}^{[u+1]} + Y_{ij}P_{jj}^{[u+1]} + C_{ij}^{[u+1]}$
        $Q_{ij}^{[n-1]} \leftarrow d_n Y_{ij}$
        for $v = n-2$ down to $0$ do
            $Q_{ij}^{[v]} \leftarrow Y_{ii}Q_{ij}^{[v+1]} + Y_{ij}Q_{jj}^{[v+1]} + D_{ij}^{[v+1]}$
$X \leftarrow UYU^T$
```

where the coefficient matrix

$$M_{ij} = \sum_{e=1}^{m} \left(P_{jj}^{[e]}\right)^T \otimes Y_{ii}^{e-1} - \sum_{g=1}^{n} \left(Q_{jj}^{[g]}\right)^T \otimes \left(T_{ii}Y_{ii}^{g-1}\right) \tag{21}$$

can be of size 1, 2, or 4, depending on the size of the blocks Y_{ii} and Y_{jj}. In the following, we give necessary and sufficient conditions for M to be nonsingular.

Theorem 16. *Let $T \in \mathbb{C}^{N \times N}$ be upper quasi-triangular, let p, q, Y, $P^{[u]}$, for $u = 0, \ldots, m$, and $Q^{[v]}$, for $v = 0, \ldots, n$, be as in Proposition 10, and let $r(x) = p(x)q(x)^{-1}$. Then M_{ij} in (21) is nonsingular for all $1 \le i < j \le \nu$ if and only if Y is a primary solution to (2) with simple critical eigenvalues (if any).*

Proof. Let (ξ_i, u_i) be an eigenpair of Y_{ii} and let (ξ_j, u_j) be an eigenpair of Y_{jj}. Then by using the properties of the Kronecker product, we observe that

$$M_{ij}(u_j \otimes u_i) = \left(\sum_{e=1}^{m} \left(P_{jj}^{[e]}\right)^T \otimes Y_{ii}^{e-1} - \sum_{g=1}^{n} \left(Q_{jj}^{[g]}\right)^T \otimes \left(T_{ii} Y_{ii}^{g-1}\right) \right)(u_j \otimes u_i)$$
$$= \left(\sum_{e=1}^{m} p^{[e]}(\xi_j)\, \xi_i^{e-1} - r(\xi_i) \sum_{g=1}^{n} q^{[g]}(\xi_j)\, \xi_i^{g-1} \right)(u_j \otimes u_i) =: \zeta(u_j \otimes u_i),$$

and conclude that $(\zeta, u_j \otimes u_i)$ is an eigenpair of M_{ij}. Since the eigenpairs of Y_{ii} and Y_{jj} are chosen arbitrarily and everything is diagonalizable, all the eigenvalues of M_{ij} have this form, and we can conclude that the matrix M_{ij} is nonsingular if and only if $\zeta \neq 0$, which is guaranteed by Lemma 12, since Y is a primary solution to (2) with simple or no critical eigenvalues.

Conversely, let ξ_i, ξ_j be eigenvalues of different diagonal blocks of Y, Y_{ii} and Y_{jj} say, then there exist (ξ_i, u_i) and (ξ_j, v_j) eigenpairs of Y_{ii} and Y_{jj}, respectively. Since M_{ij} is nonsingular, its eigenvalue $\sum_{e=1}^{m} p^{[e]}(\xi_j)\, \xi_i^{e-1} - r(\xi_i) \sum_{g=1}^{n} q^{[g]}(\xi_j)\, \xi_i^{g-1}$ is nonzero, thus by Lemma 12 either $\xi_i \neq \xi_j$ and $r(\xi_i) \neq r(\xi_j)$ or $\xi_i = \xi_j$ and $r'(\xi_i) \neq 0$. If ξ_i and ξ_j belong to the same block, then either the block is of size 1×1 or ξ_i is the complex conjugate of ξ_j, and again, $\xi_i \neq \xi_j$ and $r(\xi_i) \neq r(\xi_j)$. Since ξ_i and ξ_j were chosen arbitrarily, the same relation is true for any chosen pair of eigenvalues, and Y is thus a primary solution to (2). □

5 Numerical experiments

To the best of our knowledge, no algorithm exists for the solution of the general matrix equation $r(X) = A$, thus we compare our approach with well-established techniques for the computation of primary matrix functions. We consider the (approximate) diagonalization method [5] and the Schur–Parlett algorithm [6, 20], applied to the function $r^{-1}(z)$, that is, the chosen inverse of $r(z)$ in a neighborhood of the eigenvalues of A.

If A is a normal, then its Schur form $T = U^*AU = \operatorname{diag}(\lambda_1, \dots, \lambda_N)$ is diagonal, and the solution to $r(X) = A$ is $X = U \operatorname{diag}\big(r^{-1}(\lambda_1), \dots, r^{-1}(\lambda_N)\big)U^*$, and in this case our algorithm coincides with the diagonalization. If A is nonnormal, then the diagonalization algorithm cannot be applied if A does not have a basis of eigenvectors. In principle, this is not a severe restriction, since a small perturbation can make it diagonalizable, but the eigenvectors can still be severely ill-conditioned, and this may lead to a significant loss of accuracy, as shown in Test 1.

On the other hand, the Schur–Parlett algorithm is a suitable choice for entire functions, but none of the branches of $r^{-1}(z)$ is. This algorithm, reduces the computation of a primary matrix function to the evaluation of the same function on matrices whose eigenvalues lie in a small ball, and the latter evaluation is performed by using a truncation of the Taylor series expansion of f. This is a severe restriction, as the Taylor series of $r^{-1}(x)$ in a neighborhood of the eigenvalue λ_i of A need not converge to $r^{-1}(\lambda_j)$, where λ_j is another eigenvalue of A near λ_i. For instance, the Taylor series expansion of the square root $z^{1/2}$ at $z_0 = -10 - i$, when evaluated at $z = -10 + i$, converges to $-(-10 + i)^{1/2}$ rather than to $(-10 + i)^{1/2}$. Moreover, if $r^{-1}(\lambda_i)$ is a critical point of r, then there exists no differentiable inverse of r extending $r^{-1}(\lambda)$ in a neighborhood of λ_i. For these reasons, we cannot consider the Schur–Parlett method in our experiments, and instead we focus our attention on the following algorithms.

- `invrat`: an implementation of Algorithm 1.
- `diag`: an implementation of the diagonalization approach to the evaluation of matrix functions. In order to evaluate $f(A)$, this algorithm exploits the eigendecomposition $A =: UDU^{-1}$, with $U \in \mathbb{C}^{N \times N}$ nonsingular and $D \in \mathbb{C}^{N \times N}$ diagonal, and approximates $f(A)$ as $Uf(D)U^{-1}$. This algorithm works for diagonalizable matrices only.

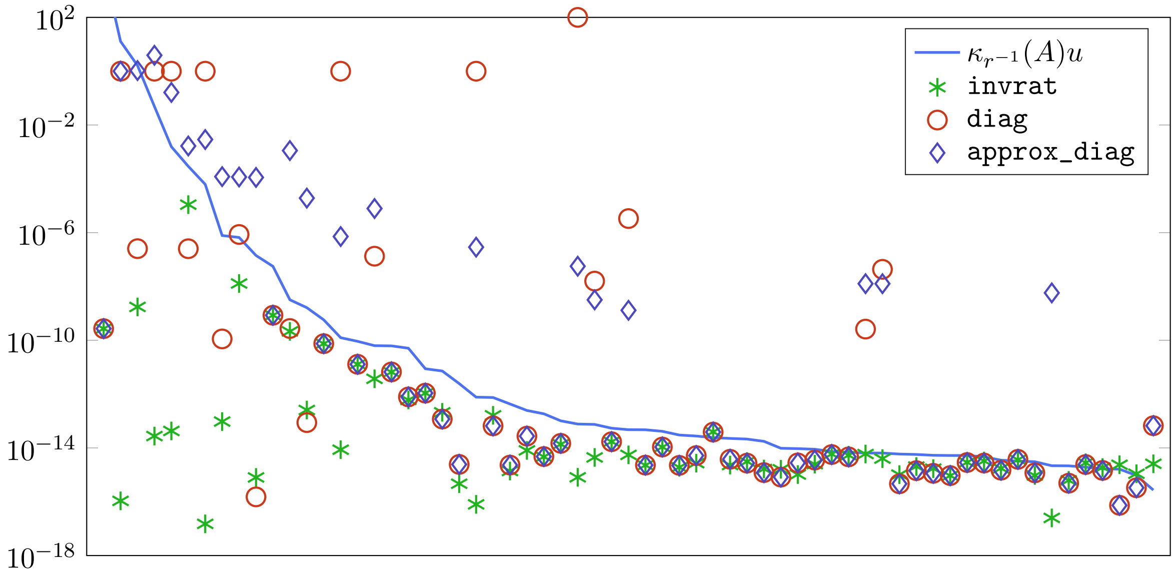

Figure 1: Relative forward errors of `invrat`, `diag`, and `approx_diag` on the test set.

- `approx_diag`: the variant of `diag` discussed by Davies [5]. In order to improve the stability of the diagonalization approach, this algorithm computes the eigendecomposition of a nearby matrix $A + \varepsilon I = \widetilde{U}\widetilde{D}\widetilde{U}^{-1}$ and then approximates $f(A)$ as $\widetilde{U} f(\widetilde{D})\widetilde{U}^{-1}$.

The experiments were performed using the 64-bit version of MATLAB 2017b on a machine equipped with an Intel I5-5287U processor, running at 2.90GHz, and 8GiB of RAM. The accuracy of the algorithms is measured by the relative error, in the spectral norm, with respect to a reference solution computed by running `invrat` with about 512 digits of accuracy using the Advanpix Multiprecision Computing Toolbox [19]. We will denote the machine precision by u.

Test 1 (Forward stability)**.** In this test, we investigate experimentally the forward stability of `invrat`, `diag`, and `approx_diag`. We consider the matrix equation $r(X) = A$, where

$$r(z) = \frac{\frac{z^3}{120} + \frac{z^2}{10} + \frac{z}{2} + 1}{-\frac{z^3}{120} + \frac{z^2}{10} - \frac{z}{2} + 1}$$

is the [3/3] Padé approximant to the exponential at 0. For A, we consider a test set including 63 real and complex nonnormal matrices, of size between 2×2 and 10×10, from the MATLAB `gallery` function and from the literature of the matrix logarithm.

Figure 1 compares the relative forward error of the three algorithms with the quantity $\kappa_{r^{-1}}(A)u$, the 1-norm condition number of a branch of r^{-1} that extends a real branch that contains 0 to the whole complex plane, estimated by means of the `funm_condest1` function from Higham's Matrix Function Toolbox [12].

Out of the three algorithms we consider, `diag` appears to be the most unreliable, as the relative forward error is of the order of 1 on more than 10% of the data set, and often several orders of magnitude larger than $\kappa_{r^{-1}}(A)u$. The forward error of `approx_diag` is larger than $\kappa_{r^{-1}}(A)u$ on almost 30% of the data set, but is of the order of 1 for four of the most ill-conditioned matrices only. Finally, the forward error of `invrat` is approximately bounded by $\kappa_{r^{-1}}(A)u$, which seems to indicate that the algorithm behaves in a forward stable manner.

Test 2. Critical solutions to the scalar equation $f(x) = y$ are ill-conditioned, and the effects of the ill-conditioning become obvious as the derivative of f approaches zero. This is the

Table 1: Solutions of the equation $r(X) = A$ in Test 2. The three columns contain the spectrum of X, the magnitude of the smallest eigenvalue of $Dr(X)$, and the relative error of the solution computed by `invrat`.

eigenvalues of X	$\lambda_{\min}(Dr(X))$	$\Vert\widetilde{X} - X\Vert_2/\Vert X\Vert_2$
$\{r_1^{-1}(\lambda_1), r_1^{-1}(\lambda_2), r_1^{-1}(\lambda_3)\}$	1.29×10^{-10}	3.42×10^{-06}
$\{r_1^{-1}(\lambda_1), r_1^{-1}(\lambda_2), r_2^{-1}(\lambda_3)\}$	5.77×10^{-11}	9.51×10^{-06}
$\{r_1^{-1}(\lambda_1), r_2^{-1}(\lambda_2), r_1^{-1}(\lambda_3)\}$	5.77×10^{-11}	3.25×10^{-06}
$\{r_1^{-1}(\lambda_1), r_2^{-1}(\lambda_2), r_2^{-1}(\lambda_3)\}$	$1.73 \times 10^{+00}$	2.80×10^{-16}
$\{r_2^{-1}(\lambda_1), r_1^{-1}(\lambda_2), r_1^{-1}(\lambda_3)\}$	$1.73 \times 10^{+00}$	3.93×10^{-16}
$\{r_2^{-1}(\lambda_1), r_1^{-1}(\lambda_2), r_2^{-1}(\lambda_3)\}$	5.77×10^{-11}	3.86×10^{-06}
$\{r_2^{-1}(\lambda_1), r_2^{-1}(\lambda_2), r_1^{-1}(\lambda_3)\}$	5.77×10^{-11}	6.98×10^{-06}
$\{r_2^{-1}(\lambda_1), r_2^{-1}(\lambda_2), r_2^{-1}(\lambda_3)\}$	1.29×10^{-10}	2.07×10^{-06}

case for matrices as well, thus the accuracy of our algorithm, as that of any stable algorithm, will be affected by what solution is being computed. Since an isolated solution is uniquely determined by its eigenvalues, choosing a solution of the scalar equation $r(x) = \lambda_i$, for each distinct eigenvalue λ_i of A is enough to fix what solution to $r(X) = A$ will be computed. This is equivalent to choosing an inverse r^{-1} of r and computing $X = r^{-1}(A)$, as discussed in Proposition 4.

In order to illustrate the numerical behavior of the Schur recurrence algorithm in computing different solutions of a matrix equation, we consider the equation $r(X) = A$, where $r(z) = -z/(z^2 + 1)$. This matrix equation is equivalent to $AX^2 + X + A = 0$, which was considered for theoretical purposes in [17] and [18].

It is easy to show [17, Lem. 3] that the equation $r(z) = \lambda$, with $\lambda \in \mathbb{C}$ has two distinct solutions if and only if $\lambda \notin \{0, \pm 1/2\}$, while

- if $\lambda \in (-\infty, -1/2] \cup [1/2, +\infty)$ then the solutions have modulus 1;
- if $\lambda \in \mathcal{D} := (\mathbb{C} \setminus \mathbb{R}) \cup (-1/2, 0) \cup (0, 1/2)$, then one solution lies inside the unit disc, while the other lies outside.

This allows one to identify two analytic branches for the inverse: $r_1^{-1} : \mathcal{D} \to \{z \in \mathbb{C} : |z| > 1\}$ and $r_2^{-1} : \mathcal{D} \cup \{0\} \to \{z \in \mathbb{C} : |z| < 1\}$, with branch cuts $(-\infty, -1/2] \cup [1/2, +\infty)$. The points $z = \pm 1$ are critical points for $r(z)$, indeed $r(\pm 1) = \mp 1/2$.

We show how the accuracy of a solution $\widetilde{X}$ to $r(X) = A$ degrades as the derivative of the function r at $\widetilde{X}$ approaches a singular matrix. This can occur in two cases: when two eigenvalues of A are close to each other but the corresponding eigenvalues of X are far apart (this may happen also when we choose the same branch for two nearby eigenvalues, if there is a branch cut in the middle); or when an eigenvalue of A is close to the image of a critical value of r and the corresponding eigenvalue of X is close to a critical point of r. We will examine one example for each situation.

Let us first consider the matrix $A = M \operatorname{diag}(1 - \varepsilon i, 1 + 2\varepsilon + \varepsilon i, 1 + 3\varepsilon + \varepsilon i)M^{-1}$, where $\varepsilon = 10^{-10}$, and $M \in \mathbb{R}^{3\times 3}$ is a matrix with entries drawn from a standard normal distribution. As one can choose two branches of the inverse of r for each of the eigenvalues of A, there exist eight isolated primary solutions X. For each of them, we report in Table 1 the magnitude of the smallest eigenvalue of $Dr(X)$ and the forward error of the solution $\widetilde{X}$ computed by `invrat`. The solutions that select a different branch of the inverse of r for the eigenvalues on the opposite sides of the branch cut lead to a better conditioned Frechét derivative, and the solution computed by `invrat` in this case has almost perfect accuracy.

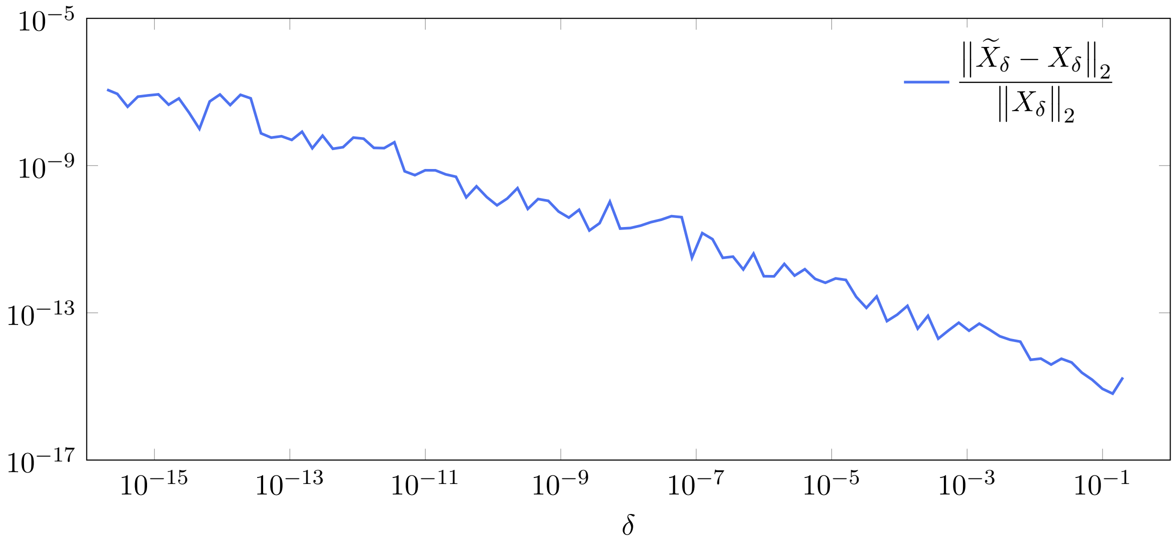

Figure 2: Relative error of the Schur algorithm for computing the solution of the matrix equation $r(X_\delta) = A_\delta$ in Test 2 with spectrum $\{r_1^{-1}(\lambda_1), r_1^{-1}(\lambda_2), r_1^{-1}(\lambda_3)\}$.

To show that the accuracy of the solution computed by `invrat` are influenced by the distance of the eigenvalues of A from the images of critical points, we investigate the behavior of the algorithm when trying to compute solutions with almost critical eigenvalues. We consider the matrix $A_\delta = M \operatorname{diag}(1/2 - \delta i, 1/2 - \delta, 1 + \delta i) M^{-1}$, where $M \in \mathbb{R}^{3\times 3}$ is a random matrix as in the previous test. Note that the eigenvalues of A tend to the image of the branch point of r as $\delta > 0$ tends to zero. Figure 2 shows the relative error of the primary solution to $r(X_\delta) = A_\delta$ computed by `invrat`, as δ varies between 2×10^{-16} and 2×10^{-1}. As expected, the accuracy of the solution is adversely affected by the proximity of the eigenvalues of A to the image of a critical point of r.

6 Conclusions

After discussing some properties of the solutions to the matrix equation $f(X) = A$, with f analytic, we developed an algorithm for computing primary solutions to the matrix equation $r(X) = A$, where r is a rational function. Our approach relies on a substitution algorithm based on Horner's scheme for the evaluation of numerator and denominator of r.

In previous work [11, 16], we have shown that, for the kth root, the computational cost of the straightforward algorithm [22] can be reduced by considering substitution algorithms that exploit more efficient matrix powering schemes. However, a fraction can be evaluated in several different ways, and some approaches require fewer matrix multiplications than applying Horner's method twice. One such example is the Paterson–Stockmeyer method [21], which can require considerably fewer matrix multiplications for polynomials of high degree.

In principle, any of these alternative schemes could produce a substitution algorithm for the solution of the matrix equation $r(X) = A$. The computational cost of the substitution algorithm induced by a given evaluation scheme would be the same as the cost of the evaluation scheme itself, since the number of intermediate matrices to be computed depends on the number of matrix multiplications needed to evaluate numerator and denominator. Therefore, starting with a cheaper evaluation scheme for rational functions, it might be possible to develop cheaper algorithms for the solution of matrix functions of the form $r(X) = A$: this will be the subject of future investigation.

Acknowledgements

The authors are grateful to Nicholas J. Higham and an anonymous referee for providing feedback on the manuscript and useful comments which improved the presentation of the paper.

References

[1] Awad H. Al-Mohy and Nicholas J. Higham, *Improved inverse scaling and squaring algorithms for the matrix logarithm*, SIAM J. Sci. Comput., 34 (2012), pp. C153–C169.

[2] Zhaojun Bai and James W. Demmel, *On swapping diagonal blocks in real Schur form*, Linear Algebra Appl., 186 (1993), pp. 75–95.

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

[4] Constantin Carathèodory, *Theory of Functions of a Complex Variable*, vol. 1, Chelsea Publishing, New York, NY, USA, 2nd ed., 1958.

[5] Edward Brian Davies, *Approximate diagonalization*, SIAM J. Matrix Anal. Appl., 29 (2008), pp. 1051–1064.

[6] Philip I. Davies and Nicholas J. Higham, *A Schur–Parlett algorithm for computing matrix functions*, SIAM J. Matrix Anal. Appl., 25 (2003), pp. 464–485.

[7] Jean-Claude Evard and Frank Uhlig, *On the matrix equation $f(X) = A$*, Linear Algebra Appl., 162-164 (1992), pp. 447–519.

[8] Massimiliano Fasi, Nicholas J. Higham, and Bruno Iannazzo, *An algorithm for the matrix Lambert W function*, SIAM J. Matrix Anal. Appl., 36 (2015), pp. 669–685.

[9] Wendell H. Fleming, *Functions of Several Variables*, Springer-Verlag, New York, NY, USA, 2nd ed., 1977.

[10] Gene H. Golub and Charles F. Van Loan, *Matrix Computations*, Johns Hopkins University Press, Baltimore, MD, USA, 4th ed., 2013.

[11] Federico Greco and Bruno Iannazzo, *A binary powering Schur algorithm for computing primary matrix roots*, Numer. Algorithms, 55 (2010), pp. 59–78.

[12] Nicholas J. Higham, *The Matrix Function Toolbox*. `http://www.maths.manchester.ac.uk/~higham/mftoolbox`.

[13] ——, *Computing real square roots of a real matrix*, Linear Algebra Appl., 88/89 (1987), pp. 405–430.

[14] ——, *Functions of Matrices: Theory and Computation*, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2008.

[15] Roger A. Horn and Charles R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1991.

[16] Bruno Iannazzo and Carlo Manasse, *A Schur logarithmic algorithm for fractional powers of matrices*, SIAM J. Matrix Anal. Appl., 34 (2013), pp. 794–813.

[17] Bruno Iannazzo and Beatrice Meini, *Palindromic matrix polynomials, matrix functions and integral representations*, Linear Algebra Appl., 434 (2011), pp. 174–184.

[18] ——, *The palindromic cyclic reduction and related algorithms*, Calcolo, 52 (2015), pp. 25–43.

[19] *Multiprecision Computing Toolbox*. Advanpix, Tokyo. 4.4.7.12739.

[20] Beresford N. Parlett, *A recurrence among the elements of functions of triangular matrices*, Linear Algebra Appl., 14 (1976), pp. 117–121.

[21] Michael S. Paterson and Larry J. Stockmeyer, *On the number of nonscalar multiplications necessary to evaluate polynomials*, SIAM J. Comput., 2 (1973), pp. 60–66.

[22] Matthew I. Smith, *A Schur algorithm for computing matrix pth roots*, SIAM J. Matrix Anal. Appl., 24 (2003), pp. 971–989.