SOLVING NONLINEAR MATRIX EQUATION $X + A^*X^{-1}A = Q$ VIA FEJÉR-RIESZ FACTORIZATION

DRAFT AS OF December 5, 2014

MOODY T. CHU*

Abstract. The nonlinear matrix equation $X + A^*X^{-1}A = Q$ can be cast as a linear Sylvester equation subject to unitary constraint. The Sylvester equation can be obtained by means of hermitian eigenvalue computation. The unitary constraint can be satisfied by means of either a straightforward alternating projection method or by a coordinate-free Newton iteration. The idea proposed in this paper originates from the operator-valued Fejér-Riesz theorem on an abstract factorization of some rational matrix-valued function over the unit disk. The work now makes the factorization realizable by numerical computation.

Key words. nonlinear matrix equation, Fejér-Riesz factorization, alternating projection method, coordinate-free Newton method

AMS subject classifications. 39B42, 15A23, 15A24, 47J25, 49M15, 65J15,

1. Introduction.

The nonlinear matrix equation

$$X + A^*X^{-1}A = Q, \tag{1.1}$$

where $A, Q \in \mathbb{C}^{n\times n}$ are given and Q is hermitian positive definite (HPD), has been extensively studied in the literature. Far from being complete, we mention [1, 6, 8, 10, 15] as a few general references on the subject of existence theory. A variety of numerical methods for actually computing the HPD solution has been proposed in [9, 12, 14, 18]. The majority of currently available algorithms takes on notions from fixed-point iteration, Newton-type iteration, or cyclic reduction. Often these methods are effective only for computing the so called extreme solutions, though the problem generally has multiple solutions.

With regard to the solvability of (1.1), perhaps the most complete analysis is developed in [6] by using an analytic factorization argument. The following result, in particular, characterizes a sufficient and necessary condition for the existence of an HPD solution [6, Theorem 2.1] .

THEOREM 1.1. *Corresponding to (1.1), define a rational matrix-valued function via*

$$\psi(\lambda) = \lambda A + Q + \frac{A^*}{\lambda}. \tag{1.2}$$

Then, (1.1) has an HPD solution if and only if $\det(\psi(\lambda))$ *is not identically zero and* $\psi(\lambda) \succeq 0$ *for all* λ *on the unit circle. In this case,*

a. The function $\psi(\lambda)$ *can be factorized as*

$$\psi(\lambda) = (C_0^* + \frac{C_1^*}{\lambda})(C_0 + \lambda C_1), \tag{1.3}$$

with $\det(C_0) \neq 0$*, and*

$$X = C_0^*C_0 \tag{1.4}$$

is a solution of (1.1).

b. Every positive definite solution of (1.1) is obtained in this way.

The factorization (1.3) in Theorem 1.1 is a special case of the more comprehensive operator-valued Fejér-Riesz theorem [16, Theorem 6.6] which is a generalization the classical Fejér-Riesz theorem for nonnegative trigonometric Laurent polynomials on the unit circle. It is interesting to note that the proof of the operator-valued Fejér-Riesz theorem in the monograph [16] itself is merely a one-line statement calling on another

*Department of Mathematics, North Carolina State University, Raleigh, NC 27695-8205, USA. (chu@math.ncsu.edu.) This research was supported in part by the National Science Foundation under grant DMS-1316779.

result concerning the factorization of pseudo-meromorphic functions. Over all, the disquisition in [16] employs machineries built upon the notion of Nevanlinna and Hardy classes. A more direct proof of the matrix-valued Fejér-Riesz factorization is given in [7]. See also [4] for an excellent survey of this topic.

While a factorization such as (1.3) might be obvious from operator theory point of view, the needed simultaneous decomposition

$$\begin{cases} Q &= C_0^* C_0 + C_1^* C_1 \\ A &= C_0^* C_1 \end{cases} \tag{1.5}$$

for a given pair of matrices $A, Q \in \mathbb{C}^{n\times n}$ is not obvious. Motivated by (1.4) and the fact that every positive definite solution is obtained in this way, we are curious to think that it might be reasonable to propose a feasible numerical procedure that does the factorization (1.5). Simultaneous decomposition of multiple matrices should be an interesting mathematical problem in itself.

In this paper, we offer a framework to tackle this particular decomposition problem (1.5). In turns, the decomposition can be adapted to formulate two new numerical methods for solving (1.1). The advantages of our approach are multi-fold:

1. It effectively reduced the problem via spectral decomposition to the problem of finding the intersection of the manifold $\mathscr{U}(n)$ of unitary matrices with a specific affine subspace.
2. Projections onto $\mathscr{U}(n)$ and the affine subspace can easily be done. Hence, a convenient, but global convergent alternating projection method is readily available.
3. Using merely the geometry, a coordinate-free Newton-type iteration can also be developed to gain high precision and quadratic rate of convergence.
4. All positive solutions can be parameterized by means of $\mathscr{U}(n)$ and, hence, all positive solutions are taken into account.

This paper is organized as follows. We begin in Section 2 by sampling the rational function $\psi(\lambda)$ at two points on the unit disk. Using spectral information of the samples, we rewrite the nonlinear equation (1.1) into a Sylvester equation which is linear. The cost for clearing out the nonlinearity by such a transformation is that the solution to the Sylvester equation must remain unitary. Our first main result is a one-to-one correspondence in representing the solution to the nonlinear problem by a unitary solution to the linear problem. In Section 3 we argue by using the theory of parameter continuation that generically there are only finitely many geometrically isolated solutions. More importantly, the geometry of unitary matrices is so well structured that we propose two simple numerical procedures by using projections. In Section 3.1 we employ the standard Euclidean projection, including the polar decomposition, to formulate an alternative projection algorithm. The method converges linearly, but globally. In Section 3.2 we employ a projection along the tangent direction which thus constitutes a Newton-type iteration. A nice feature of this quadratically convergent method is that it refers to no particular coordinate frame. In our actual implementation for numerical test, we combine both methods into a hybrid algorithm that starts with several steps of alternating projection to drive the iterates closer to a true solution before the Newton projection is activated for faster convergence and better precision. Finally, in Section 4, we present a new way to characterize the partial ordering among solutions.

2. Discrete Fejér-Riesz factorizations. As $\det(\psi(\lambda))$ is itself a rational (scalar) function, there are only a finite number of points for which $\det(\psi(\lambda)) = 0$. Using different values of λ on the unit circle if necessary, we may assume without loss of generality that the two discrete samples $\psi(1)$ and $\psi(-1)$ are positive definite. Thus, there exist matrices $\alpha, \beta \in \mathbb{C}^{n\times n}$ such that

$$\begin{cases} Q + A + A^* &= \alpha^*\alpha \\ Q - A - A^* &= \beta^*\beta. \end{cases} \tag{2.1}$$

Upon comparing with (1.3), we may take

$$\begin{cases} C_0 &= \frac{\alpha+\beta}{2} \\ C_1 &= \frac{\alpha-\beta}{2}, \end{cases} \tag{2.2}$$

once α, β are determined. The resulting C_0 and C_1 in (2.2) must satisfy the conditions specified in (1.5). It is easy to check by substitution that for any factorization in (2.1), the first constraint $Q = C_0^* C_0 + C_1^* C_1$ in (1.5) is always satisfied. It remains to require that the factors α, β must be such that

$$\beta^* \alpha - \alpha^* \beta = 2(A - A^*). \tag{2.3}$$

Being skew hermitian, the condition (2.3) entails n^2 real-valued equations which will be further detailed below. Our first goal for now is to find these suitable factors α and β.

2.1. Sylvester equation subject to unitary constraint. There also exist unitary matrices $U_1, U_2 \in \mathscr{U}(n)$, and positive diagonal matrices $\Sigma_1, \Sigma_2 \in \mathbb{R}^{n \times n}$ such that

$$\left\{ \begin{array}{rcl} Q + A + A^* & = & U_1 \Sigma_1 U_1^* \\ Q - A - A^* & = & U_2 \Sigma_2 U_2^*. \end{array} \right. \tag{2.4}$$

Define

$$\left\{ \begin{array}{rcl} \widehat{\alpha} & := & U_1^* \alpha U_1 \Sigma_1^{-\frac{1}{2}} \\ \widehat{\beta} & := & U_2^* \beta U_2 \Sigma_2^{-\frac{1}{2}}. \end{array} \right. \tag{2.5}$$

It can be seen trivially that $\widehat{\alpha}, \widehat{\beta} \in \mathscr{U}(n)$. We may rewrite the constraint (2.3) in terms of $\widehat{\alpha}, \widehat{\beta}$, which becomes

$$U_1^* U_2 \Sigma_2^{\frac{1}{2}} \widehat{\beta}^* U_2^* U_1 \widehat{\alpha} \Sigma_1^{\frac{1}{2}} - \Sigma_1^{\frac{1}{2}} \widehat{\alpha}^* U_1^* U_2 \widehat{\beta} \Sigma_2^{\frac{1}{2}} U_2^* U_1 = U_1^* (A - A^*) U_1. \tag{2.6}$$

As the spectral decompositions in (2.4) are readily available, the matrices

$$\left\{ \begin{array}{rcl} \Theta & := & U_1^* U_2 \\ S & := & \Theta \Sigma_2^{\frac{1}{2}} \\ K & := & U_1^* (A - A^*) U_1 \end{array} \right. \tag{2.7}$$

are known and K is skew-hermitian. We may conveniently condense the two unknown factors $\widehat{\alpha}, \widehat{\beta} \in \mathscr{U}(n)$ into a single unknown matrix $\Gamma \in \mathscr{U}(n)$ defined by

$$\Gamma := \widehat{\beta}^* \Theta^* \widehat{\alpha} \tag{2.8}$$

which must satisfy the linear Sylvester equation

$$S \Gamma \Sigma_1^{\frac{1}{2}} - \Sigma_1^{\frac{1}{2}} \Gamma^* S^* = K. \tag{2.9}$$

Solving the linear equation (2.9) is relatively simpler than solving the quadratic equation (2.3), except that the solution Γ must be a unitary matrix by the definition (2.8). We shall discuss two simple numerical schemes to accomplish this goal in Section 3.

2.2. Parametrization of solutions. Thus far, all the steps are reversible. No inversion is needed except for the inverses of diagonal matrices Σ_1 and Σ_2, which are trivial. Once a solution Γ is found, we may take

$$\widehat{\alpha} := \Theta \widehat{\beta} \Gamma, \tag{2.10}$$

where $\widehat{\beta} \in \mathscr{U}(n)$ can be arbitrary and can be used as a parameter. Using (2.5), we may recover the desirable α, β

$$\left\{ \begin{array}{rcl} \alpha & := & U_2 \widehat{\beta} \Gamma \Sigma_1^{\frac{1}{2}} U_1^* \\ \beta & := & U_2 \widehat{\beta} \Sigma_2^{\frac{1}{2}} U_2^*. \end{array} \right. \tag{2.11}$$

In particular, the set of positive definite solutions to (1.1) can be characterized via the set of intersection of the manifold $\mathscr{U}(n)$ and the affine subspace of solutions to (2.9) as follows.

THEOREM 2.1. *Every positive definite solution to (1.1) is of the form*

$$X = \frac{1}{2}\left(Q + \mathfrak{s}(\Gamma)\right), \tag{2.12}$$

where

$$\mathfrak{s}(\Gamma) := \frac{1}{2}\left(U_2\Sigma_2^{\frac{1}{2}}\Gamma\Sigma_1^{\frac{1}{2}}U_1^* + U_1\Sigma_1^{\frac{1}{2}}\Gamma^*\Sigma_2^{\frac{1}{2}}U_2^*\right) \tag{2.13}$$

and Γ is a unitary solution to the Sylvester equation (2.9).

Proof. Using the relationship (2.2), we may rewrite C_0, C_1 in terms of (2.11). By (1.4), we see further that

$$X = \frac{1}{4}\left(U_2\Sigma_2^{\frac{1}{2}} + U_1\Sigma_1^{\frac{1}{2}}\Gamma^*\right)\left(\Gamma\Sigma_1^{\frac{1}{2}}U_1^* + \Sigma_2^{\frac{1}{2}}U_2^*\right)$$

which may be simplified to (2.12) by using (2.1). □

Observe in (2.12) that only the unitary solution Γ is needed. The reference to $\widehat{\beta}$ as is required in (2.11) is immaterial. Observer also that X is related to Γ in a linear way, the difference being that X satisfies a nonlinear equation (1.1) while Γ satisfies a linear equation (2.9) and is unitary. After introducing more detailed notations, we shall argue further in Section 4 that such a correspondence is one-to-one. We think that such a natural parametrization through $\Gamma \in \mathscr{U}(n)$ satisfying (2.9) for each positive definite solution X of (1.1) is simple and interesting. We now characterize the parameter Γ.

3. Solving for Γ. The Sylvester equation (2.9) is under-determined due to the skew-hermitian structure. A general solution to the equation can be expressed as

$$\Gamma = \Gamma_0 + \sum \gamma_i \Omega_i, \tag{3.1}$$

where Γ_0 is a particular solution and the set $\{\Omega_i\}$ forms a basis for the solution subspace of the homogeneous problem

$$S\Gamma\Sigma_1^{\frac{1}{2}} - \Sigma_1^{\frac{1}{2}}\Gamma^*S^* = 0. \tag{3.2}$$

We need to clarify the meaning of summation in (3.1).

First, cares must be given when counting the dimensionality of the null space of (3.2). For a generic $\Gamma \in \mathbb{C}^{n\times n}$, there are $2n^2$ real-valued unknown entries. The skew hermitian structure, however, gives rise to only n^2 real-valued equations. The null space therefore should have real dimensionality n^2 for complex-valued solutions.

More specifically, let the real and the imaginary parts of a matrix M be expressed as

$$M = \Re(M) + \imath\Im(M). \tag{3.3}$$

We can rewrite the Sylvester equation as a system of equations for the pair $(\Re(\Gamma), \Im(\Gamma)) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times n}$

$$\left\{\begin{array}{rcl} \left(\Re(S)\Re(\Gamma) - \Im(S)\Im(\Gamma)\right)\Sigma_1^{\frac{1}{2}} - \Sigma^{\frac{1}{2}}(\Re(\Gamma)^\top\Re(S)^\top - \Im(\Gamma)^\top\Im(S)^\top) & = & \Re(K) \\ \left(\Im(S)\Re(\Gamma) + \Re(S)\Im(\Gamma)\right)\Sigma_1^{\frac{1}{2}} + \Sigma^{\frac{1}{2}}(\Im(\Gamma)^\top\Re(S)^\top + \Re(\Gamma)^\top\Im(S)^\top) & = & \Im(K). \end{array}\right. \tag{3.4}$$

Denote

$$\left\{\begin{array}{rcl} \mathcal{A} & := & \Sigma_1^{\frac{1}{2}} \otimes \Re(S) \\ \mathcal{B} & := & \Re(S) \otimes \Sigma_1^{\frac{1}{2}} \\ \mathcal{C} & := & \Sigma_1^{\frac{1}{2}} \otimes \Im(S) \\ \mathcal{D} & := & \Im(S) \otimes \Sigma_1^{\frac{1}{2}} \end{array}\right. \tag{3.5}$$

and let $\mathcal{P}$ denote the permutation matrix of indices that enumerates entries of a matrix row-wise. Then (3.4) is equivalent to

$$\begin{bmatrix} \mathcal{A}-\mathcal{B}\mathcal{P} & -\mathcal{C}+\mathcal{D}\mathcal{P} \\ \mathcal{C}+\mathcal{D}\mathcal{P} & \mathcal{A}+\mathcal{B}\mathcal{P} \end{bmatrix} \begin{bmatrix} \mathbf{vec}(\Re(\Gamma)) \\ \mathbf{vec}(\Im(\Gamma)) \end{bmatrix} = \begin{bmatrix} \mathbf{vec}(\Re(K)) \\ \mathbf{vec}(\Im(K)) \end{bmatrix}. \tag{3.6}$$

This square system (3.6) is rank deficient because, in truth, the first equation in (3.4) is skew-symmetric and the second equation is symmetric. Taking this structure into account, there are generically n^2 independent equations. It is in this context that we define the basis $\{\Omega_i\}$ and write (3.1) where γ_i are real-valued.

It is worth mentioning the special case when $A, Q \in \mathbb{R}^{n\times n}$ and Q is symmetric positive definite (SPD). The question of existence for this real-valued problem has been studied in [6, Section 8] and more detailed in [5]. In this case, the above discussion can be carried over, except that (skew-)hermitian matrices are replaced by (skew-)symmetric matrices and unitary matrices by orthogonal matrices. In particular, the system (3.6) is reduced to

$$(\mathcal{A}-\mathcal{B}\mathcal{P})\mathbf{vec}(\Gamma) = \mathbf{vec}(K), \tag{3.7}$$

whose coefficient matrix $\mathcal{A}-\mathcal{B}\mathcal{P}$ is of rank $\frac{n(n-1)}{2}$ generically and, hence, Γ is characterized by $\frac{n(n+1)}{2}$ basis solutions $\{\Omega_i\}$ to the real-valued homogeneous problem (3.2). Other than these modifications, the theory and algorithm described below can be applied without trouble.

Returning to the general problem of complex-valued problem, the representation of Γ in the form (3.1) provides a parametrization of solutions to (2.9). Our theory requires that Γ also be unitary. This will require $\{\gamma_i\}$ to satisfy the nonlinear system

$$\begin{bmatrix} I \\ 0 \end{bmatrix} = \begin{bmatrix} \Re(\Gamma_0^*\Gamma_0) \\ \Im(\Gamma_0^*\Gamma_0) \end{bmatrix} + \sum_{i=1}^{n^2} \gamma_i \begin{bmatrix} \Re(\Omega_i^*\Gamma_0 + \Gamma_0^*\Omega_i) \\ \Im(\Omega_i^*\Gamma_0 + \Gamma_0^*\Omega_i) \end{bmatrix} + \sum_{1\le i\le j\le n^2} \gamma_i\gamma_j \begin{bmatrix} \Re(\Omega_i^*\Omega_j + \Omega_j^*\Omega_i) \\ \Im(\Omega_i^*\Omega_j + \Omega_j^*\Omega_i) \end{bmatrix}. \tag{3.8}$$

The top n^2 equations in (3.8) are from entries of symmetric matrices, while the bottom n^2 equations are from skew-symmetric matrices. So, in total this is a square polynomial system with n^2 unknowns in n^2 equations. Regarding Γ_0, Ω_i and, correspondingly, the problem data A and Q, as the parameters of the polynomial system, the follow result is known from the theory of parameter continuation [17, Theorem 7.1.1].

THEOREM 3.1. *Let $\mathcal{N}(A,Q)$ denote the number of geometrically isolated solutions to the corresponding (3.8) over the algebraically closed complex space. Then*

1. *$\mathcal{N}(A,Q)$ is the same, say $\mathcal{N}$, for almost all $A, Q \in \mathbb{C}^{n\times n}$ and Q is HPD.*
2. *For all $A, Q \in \mathbb{C}^{n\times n}$, Q is HPD, $\mathcal{N}(A,Q) \le \mathcal{N}$.*
3. *The subset of A, Q where $\mathcal{N}(A,Q) = \mathcal{N}$ is a Zariski open set, that is, the exceptional subset of tensors $A, Q \in \mathbb{C}^{n\times n}$ where $\mathcal{N}(A,Q) < \mathcal{N}$ is an affine algebraic set[1] contained within an algebraic set of codimension one.*

Since the real space is Zariski dense in the complex space, the above statements hold for almost all $A, Q \in \mathbb{R}^{n\times n}$, Q is SPD, except that the number of real-valued isolated solutions varies as a function of A, Q and is no longer a constant. The latter is an interesting topic in the active and ongoing research area called real algebraic geometry. For our application, we only need the fact that generically solutions to (1.1) are geometrically isolated. See also [6, Corollary 6.6 and Theorem 8.2]. We portray the isolated intersection points of the surface $\mathscr{U}(n)$ of unitary matrices and the affine subspace $\mathscr{A}$ containing all Γ's in the form (3.1) in the drawings of Figure 3.1 and Figure 3.3.

We now describe two new methods to find Γ and, hence, a solution X to (1.1). The first approach alternates between the surface $\mathscr{U}(n)$ and the affine subspace $\mathscr{A}$. The computation involves a sequence of polar decompositions. No matrix inversion is needed. This alternating projection approach offers global convergence, but

[1] A subset of affine n-space $\mathbb{A}^n$ over an algebraically closed field k is called an affine algebraic set if it can be written as the zero locus of a set of polynomials. By the Hilbert basis theorem, this set of polynomials can be assumed to be finite. The Zariski topology on $\mathbb{A}^n$ is simply a topology where the closed sets are precisely the algebraic sets in $\mathbb{A}^n$.

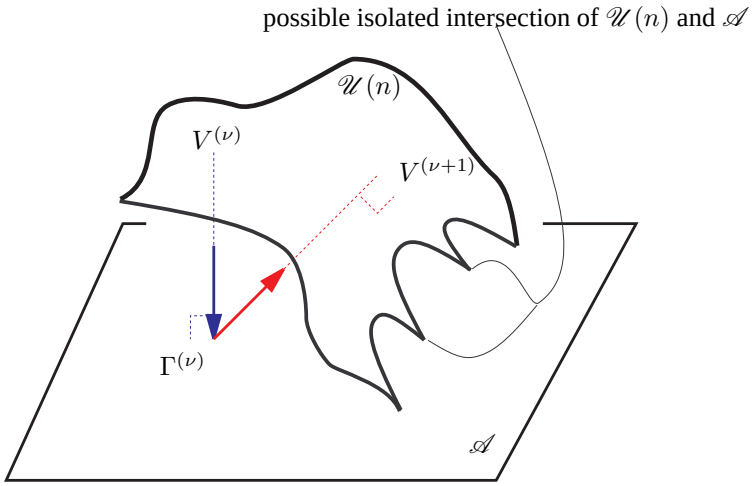

FIGURE 3.1. *Alternating projection between $\mathscr{U}(n)$ and $\mathscr{U}(n)$.*

at linear rate. The second approach is a convenient Newton-type procedure that adjusts the parameters $\{\gamma_i\}$ with the aim at moving the corresponding Γ toward the set $\mathscr{U}(n)$. As is typical, it offers local convergence, but at quadratic rate. Notably this Newton approach involves only geometric interpretation without referring to the nonlinear system (3.8).

3.1. Alternating projection method. The notion of alternating projections has been used in many disciplines. The mechanism of projections between $\mathscr{U}(n)$ and $\mathscr{A}$ is particularly easy. The idea is sketched in Figure 3.1. Specifically, by identifying a complex-valued matrix M as $(\Re(M), \Im(M))$, the bilinear functional

$$\langle M, N\rangle := \langle \Re(M), \Re(N)\rangle_F + \langle \Im(M), \Im(N)\rangle_F \tag{3.9}$$

defines an inner product over $\mathbb{C}^{n\times n}$. In this sense, we may assume that the basis $\{\Omega_i\}$ for the solution subspace of the homogeneous problem (3.2) are chosen to be mutually orthonormal to begin with[2]. Starting with $V^{(\mu)} \in \mathscr{U}(n)$, the matrix $\Gamma^{(\nu)}$ associated with the vector $\boldsymbol{\gamma}^{(\nu)} \in \mathbf{R}^{n^2}$ defined by

$$\boldsymbol{\gamma}^{(\nu)} := \left[\langle V^{(\nu)} - \Gamma_0, \Omega_i\rangle\right], \tag{3.10}$$

is the projection of $V^{(\nu)}$ onto to the affine subspace $\mathscr{A}$. In the meantime, the nearest point[3] on the manifold $\mathscr{U}(n)$ to a specified $\Gamma^{(\nu)}$ is given by the unitary matrix which occurs in the polar decomposition of $\Gamma^{(\nu)}$ [13]. Polar decomposition can be computed easily via the singular value decomposition [11].

By construction, the distance between $V^{(\nu)}$ and $\Gamma^{(\nu)}$ is being reduced per iteration. The sequence of values $\|V^{(\nu)} - \Gamma^{(\nu)}\|_F$ converges globally at linear rate. However, the set $\mathscr{U}(n)$ being non-convex, it is possible that the iterates will get stagnated at a local solution. If the stagnation does not occur, then the iterations should converge globally to an intersection of $\mathscr{U}(n)$ and $\mathscr{A}$ at linear rate.

Example 1. Consider the problem (1.1) with $n = 6$ and

$$A = \begin{bmatrix} -1.7043 & 0.6249 & -1.4613 & -0.0089 & -0.9208 & -0.2915 \\ 0.6892 & 0.8939 & 1.3194 & 0.8582 & 0.8131 & 1.6303 \\ -1.2885 & -1.0669 & 0.2346 & -0.4312 & 0.3086 & -0.7185 \\ -1.7242 & 0.3818 & 0.6473 & 1.6436 & 0.0197 & -1.3301 \\ 0.0695 & -0.0223 & 2.0685 & 0.1987 & -0.7067 & -1.5836 \\ 0.7089 & 0.0048 & -3.1480 & -0.4098 & 0.8164 & -0.9869 \end{bmatrix},$$

[2]That is, we identify $\Omega_i = (\Re(\Omega_i), \Im(\Omega_i)) \in \mathbb{R}^{n^2} \times \mathbb{R}^{n^2}$ and consider the corresponding linear system (3.16).

[3]Under the usual Euclidean norm over $\mathbb{C}^{n\times n}$, which is defined by $\|M\|_F := \sqrt{\mathrm{tr}(MM^*)}$ and is equivalent to the induced norm by the induced inner product (3.9).

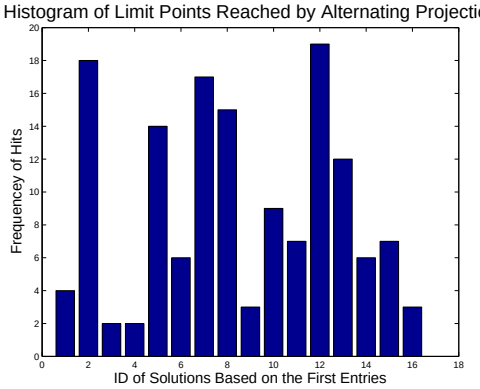

FIGURE 3.2. *Distribution of hit frequencies on solutions by random test of the alternating projection method.*

$$Q = \begin{bmatrix} 7.1618 & 1.8363 & -0.9226 & -3.3863 & -0.2137 & -1.3190 \\ 1.8363 & 12.1889 & 3.4528 & -4.8090 & 3.8368 & 0.9035 \\ -0.9226 & 3.4528 & 16.7803 & 3.3950 & 3.6227 & 1.0194 \\ -3.3863 & -4.8090 & 3.3950 & 10.0691 & -4.3373 & 0.9658 \\ -0.2137 & 3.8368 & 3.6227 & -4.3373 & 10.7882 & 4.6472 \\ -1.3190 & 0.9035 & 1.0194 & 0.9658 & 4.6472 & 9.3834 \end{bmatrix}.$$

For simplicity, we limit ourselves only to real-valued computation. So, in our theory, we replace unitary matrices by orthogonal matrices, HSD solutions by SPD solutions, adjoint operation by transpose operation, and so on. In particular, there are only 21 basis solutions $\{\Omega_i\} \in \mathbb{R}^{6\times 6}$ for Γ.

Out of 250 runs with randomly selected starting points, 144 runs converge respectively to a total of 16 distinct real-valued solutions. This count is agreeable with that estimated in [6, Proposition 8.2]. If we label these solutions based on the size their $(1, 1)$ entries, then plotted in Figure 3.2 is the frequency of each solution reached by the alternating projection method through this random test. We shall comment on the significance of"ordering" these solutions in Section 4. For now we only point out that there is a good probability that the alternating projection method with a random starting point will converge to a solution that is neither maximal nor minimal.

We point out in passing that the above projection mechanisms are so easy to perform that they can be readily extended to the more sophisticated schemes, such as Dykstra's projection algorithm [2]. While the unmodified alternating projection method described above leads to some arbitrary point in the intersection, Dykstra's algorithm is generally capable of relating the starting point to the nearest limit point through a few easily manageable intermediate steps. The only possible concern is that Dykstra's algorithm works best for finding the intersection of convex sets, whereas in our case the set $\mathscr{U}(n)$ is not convex. So, even though the projections can be performed at every step, the resulting iteration may not converge globally anymore. We choose not to implement Dykstra's algorithm in this study.

Another application of the alternating projection method is to perform the iteration only a few times with the hope that $V^{(\nu)}$ is brought close enough to the intersection for the Newton iteration to kicks in for faster convergence and better precision. Such a tactic is implemented in the coordinate-free Newton method which we now describe below.

3.2. Coordinate-free Newton iteration. Recall that a classical Newton step for a function $f : \mathbb{R} \longrightarrow \mathbb{R}$ is composed of two components: First, compute the iterate

$$x^{(\nu+1)} = x^{(\nu)} - (f'(x^{(\nu)}))^{-1} f(x^{(\nu)}) \tag{3.11}$$

which is precisely the x-intercept of the tangent line to the graph of f at the point $(x^{(\nu)}, f(x^{(\nu)}))$. Second, lift $x^{(\nu+1)}$ to the new point $(x^{(\nu+1)}, f(x^{(\nu+1)}))$ on the graph of f along the y-axis and repeat the iteration. The

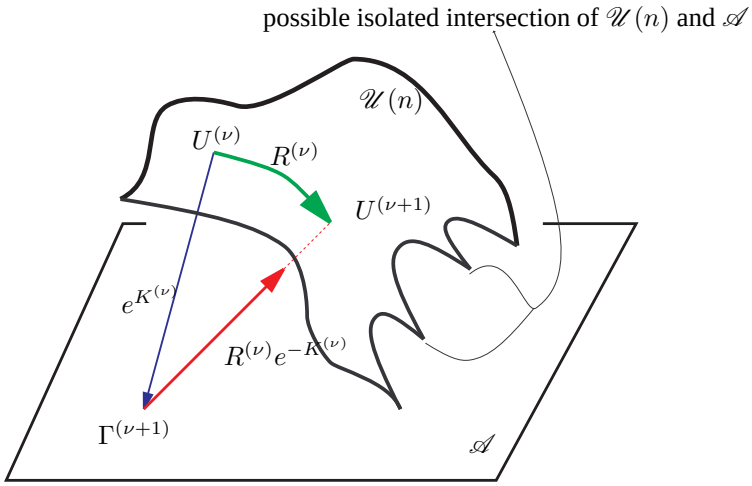

FIGURE 3.3. *One Newton step completed by finding an $\mathscr{A}$-intercept through a tangent vector and a lift back to $\mathscr{U}(n)$.*

set $\mathscr{U}(n)$ can be thought of as a smooth manifold[4] of real dimension n^2. If we think of the surface $\mathscr{U}(n)$ as playing the role of the graph of f and the affine subspace $\mathscr{A}$ containing all Γ's in the form (3.1) as playing the role of the x-axis, then an iterative process analogous to the Newton method might be developed. The idea is sketched in the drawing of Figure 3.3. The challenge is to formulate such an iteration without referring to any coordinate frames.

The set $\mathscr{U}(n)$ is a special Lie group with well understood Lie algebra structure. In particular, any tangent vector $T(U)$ to $\mathscr{U}(n)$ at a given point $U \in \mathscr{U}$ is of the form

$$T(U) = UK \tag{3.12}$$

for some skew-hermitian matrix $K \in C^{n\times n}$. Given $U^{(\nu)} \in \mathscr{U}(n)$, the "array" $U^{(\nu)} + U^{(\nu)}K$ with any skew-hermitian matrix K represents a tangent line to $\mathscr{U}(n)$ emanating from the point $U^{(\nu)}$. Mimicking the tangent step in the classical Newton method, we thus seek an $\mathscr{A}$-intercept of such an array with the affine subspace $\mathscr{A}$. In other words, the task demands to find both a skew-hermitian matrix $K^{(\nu)}$ and parameters $\{\gamma_i^{(\nu+1)}\}$ such that the equation

$$U^{(\nu)} + U^{(\nu)}K^{(\nu)} = \Gamma_0 + \sum_{i=1}^{n^2} \gamma_i^{(\nu+1)}\Omega_i \tag{3.13}$$

is satisfied. Equivalently, we work on solving the equation

$$K^{(\nu)} = {U^{(\nu)}}^*\Gamma_0 - I + \sum_{i=1}^{n^2} \gamma_i^{(\nu+1)}{U^{(\nu)}}^*\Omega_i, \tag{3.14}$$

where both $K^{(\nu)}$ and $\{\gamma_i^{(\nu+1)}\}$ are unknowns.

We may first solve for real-valued $\{\gamma_i^{(\nu+1)}\}$ independent of $K^{(\nu)}$ as follows. Denote

$$\begin{cases} \Phi_0^{(\nu)} & := \quad I - {U^{(\nu)}}^*\Gamma_0 \\ \Phi_i^{(\nu)} & := \quad {U^{(\nu)}}^*\Omega_i, \quad i = 1, \ldots, n^2. \end{cases} \tag{3.15}$$

[4]The Lie theory asserts that a unitary matrix U can be parameterized by a skew-hermitian matrix K through the exponential map, whereas K is skew-hermitian if and only if $\Re(K)$ is skew-symmetric and $\Im(U)$ is symmetric. The subspace of real symmetric matrices is of dimension $\frac{n(n+1)}{2}$ and that of real skew-symmetric matrices is of dimension $\frac{n(n-1)}{2}$.

For convenience, adopt also notations that M^{D} stands for the column vector of diagonal entries of M, M^{U} for column vector formed by vectorizing the strictly upper triangular part of M row-wise, and M^{L} the column vectors by vectorizing the strictly lower part matrix of M column-wise. Then, using the fact that $K^{(\nu)}$ is skew-hermitian, $\{\gamma_i^{(\nu+1)}\}$ must satisfy the linear equation

$$\begin{bmatrix} \Re(\Phi_1^{(\nu)})^{\mathrm{D}} & \dots & \Re(\Phi_{n^2}^{(\nu)})^{\mathrm{D}} \\ \Re(\Phi_1^{(\nu)})^{\mathrm{U}} + \Re(\Phi_1^{(\nu)})^{\mathrm{L}} & \dots & \Re(\Phi_{n^2}^{(\nu)})^{\mathrm{U}} + \Re(\Phi_{n^2}^{(\nu)})^{\mathrm{L}} \\ \Im(\Phi_1^{(\nu)})^{\mathrm{U}} - \Im(\Phi_1^{(\nu)})^{\mathrm{L}} & \dots & \Im(\Phi_{n^2}^{(\nu)})^{\mathrm{U}} - \Im(\Phi_{n^2}^{(\nu)})^{\mathrm{L}} \end{bmatrix} \begin{bmatrix} \gamma_1^{(\nu+1)} \\ \vdots \\ \gamma_{n^2}^{(\nu+1)} \end{bmatrix} = \begin{bmatrix} \Re(\Phi_0^{(\nu)})^{\mathrm{D}} \\ \Re(\Phi_0^{(\nu)})^{\mathrm{U}} + \Re(\Phi_0^{(\nu)})^{\mathrm{L}} \\ \Im(\Phi_0^{(\nu)})^{\mathrm{U}} - \Im(\Phi_0^{(\nu)})^{\mathrm{L}} \end{bmatrix}. \quad (3.16)$$

Note that the actual value of $K^{(\nu)}$ does not play any role at all in the linear system (3.16). Once the set $\{\gamma_i^{(\nu+1)}\}$ is determined, $K^{(\nu)}$ follows from (3.14) which is guaranteed to be skew-hermitian.

The above procedure amounts to only the tangent step in the classical Newton method. We now need a way to "lift up" the point $\Gamma^{(\nu+1)} \in \mathscr{A}$ back to a point $U^{(\nu+1)} \in \mathscr{U}(n)$. The challenge here is that, unlike the conventional Newton method in the Euclidean space, there is no obvious coordinate axis to follow. One possible way of this lifting can be motivated as follows. Since our goal is to find an intersection of the two sets $\mathscr{U}(n)$ and $\mathscr{A}$, we hope that all the iterations eventually cluster near a point of intersection. Thus we should expect

$$U^{(\nu+1)} \approx \Gamma^{(\nu+1)}. \quad (3.17)$$

On the other hand, the tangent equation (3.13) is merely a linearization of a nonlinear relationship in the sense that

$$\Gamma^{(\nu+1)}) \approx U^{(\nu)} e^{K^{(\nu)}}. \quad (3.18)$$

To evaluate the exponential matrix $e^{K^{(\nu)}}$ in (3.18) is not needed and is expensive. Instead, we define the Cayley transform

$$R^{(\nu)} := \left(I + \frac{K^{(\nu)}}{2}\right)\left(I - \frac{K^{(\nu)}}{2}\right)^{-1} \quad (3.19)$$

which happens to be the $(1,1)$ Padé approximation of the matrix $e^{K^{(\nu)}}$. It is well known that $R^{(\nu)} \in \mathscr{U}(n)$ and that

$$R^{(\nu)} \approx e^{K^{(\nu)}} \quad (3.20)$$

if $\|K^{(\nu)}\|$ is small. Combining (3.17) and (3.18), we now define

$$U^{(\nu+1)} := U^{(\nu)} R^{(\nu)} \quad (3.21)$$

and the next iteration is ready to begin. In this way, we have completed the lifting of the matrix $\Gamma^{(\nu+1)}$ from the affine subspace $\mathscr{A}$ to the surface $\mathscr{U}(n)$.

We shall not be bothered to provide a convergence proof of the iteration described above, because an argument following step by step of that given in [3, Theorem 4.2] can easily be laid out. We simply point out that since the scheme follows the geometry so closely comparable to the conventional Newton method, a rate of quadratic convergence is expected, as is evidenced in Figure 3.4 which represents merely one of our many numerical experiments. In this experiment, we start with the alternating projection method to until $\|U^{(\nu+1)} - U^{(\nu)}\|_F \leq 10^{-2}$ (which takes 19 iterations in this particular instance) and then let the coordinate-free Newton method described in Section 3.2 kick in (which takes only 4 iterations to reach the machine precision).

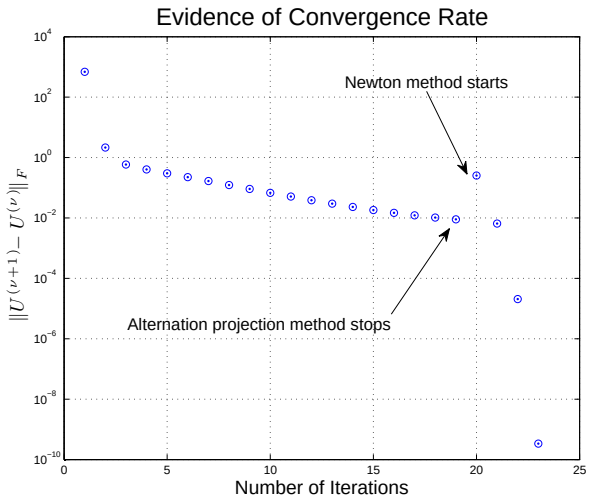

FIGURE 3.4. *Linear convergence of the alternating projection method (iterations 1 to 19) and quadratic convergence of the Newton method (iterations 20 to 23).*

4. Ordered solutions. Let $X^{[1]}$ and $X^{[2]}$ denote two distinct solutions to (1.1). We sometimes prefer to compare the solutions by the partial ordering in the sense that $X^{[1]} \succeq X^{[2]}$ if $X^{[1]} - X^{[2]}$ is positive semi-definite. Most of the algorithms proposed in the literature have the feature of computing the maximal solution X_+, that is, $X_+ \succeq X$ for any other HSD solution X of (1.1). See [1, 6, 9, 12, 14] for some specific schemes and the proofs of their convergence. A common feature of these methods is that the iterates $\{X_k\}$ generated are inherently monotone. Even when the conventional Newton's method is applied to (1.1), which can be characterized as simply solving the Stein's equation

$$X_k - L_k^* X_k L_k = Q - 2L_k^* A, \quad k = 1, \ldots \tag{4.1}$$

per step, where $X_0 = Q$ and $L_k := X_{k-1}^{-1} A$, it is still the case that $X_0 \succeq X_1 \succeq \ldots \succeq X_+$ [9, Theorem 5.3]. Our method described in Section 3.2, in contrast, can be used to find other solutions.

An analytic way for checking whether $X^{[1]} \succeq X^{[2]}$ is as follows [6, Theorem 2.2].

LEMMA 4.1. *Corresponding to $X^{[k]}$, $k = 1, 2$, let $C_0^{[k]} := X^{[k]^{\frac{1}{2}}}$ and $C_1^{[k]} := X^{[k]^{\frac{1}{2}}} A$. Form the Fejér-Riesz factor $C_0^{[k]} + \lambda C_1^{[k]}$ as is described in Theorem 1.1.*

1. *If the matrix-valued function $\left(C_0^{[2]} + \lambda C_1^{[2]}\right)\left(C_0^{[1]} + \lambda C_1^{[1]}\right)^{-1}$ is analytic in the open unit disk for λ, then $X^{[1]} \succeq X^{[2]}$.*
2. *In particular, $X^{[1]}$ is a maximal solution if $\det\left(C_0^{[1]} + \lambda C_1^{[1]}\right) \neq 0$ for $|\lambda| < 1$.*

In contrast, we offer the following criterion which might be computationally more feasible.

LEMMA 4.2. *With respect to $X^{[k]}$, $k = 1, 2$, let $\Gamma^{[k]}$ be the corresponding unitary solutions to (2.9) and $\{\gamma_i^{[k]}\}$ be the set of associated real-valued coefficients according to (3.1). Then*

1. *$X^{[1]} \succeq X^{[2]}$ if and only if $\mathfrak{s}(\Gamma^{[1]}) \succeq \mathfrak{s}(\Gamma^{[2]})$.*
2. *$X^{[1]} \succeq X^{[2]}$ if and only if*[5]

$$\sum_{i=1}^{n^2} \left(\gamma_i^{[1]} - \gamma_i^{[2]}\right) S\Omega_i \Sigma_1^{\frac{1}{2}} \succeq 0. \tag{4.2}$$

[5] As a by-product of this proof, we also see that the correspondence (2.12) between X and Γ is one-to-one. The reason is as follows. By definition, $\{\Omega_i\}$ and, consequently, $\{S\Omega_i\Sigma_1^{\frac{1}{2}}\}$ are linearly independent. It follows from (4.2) that $X^{[1]} = X^{[2]}$ if and only if $\gamma_i^{[1]} = \gamma_i^{[2]}$ for all $1 \leq i \leq n^2$.

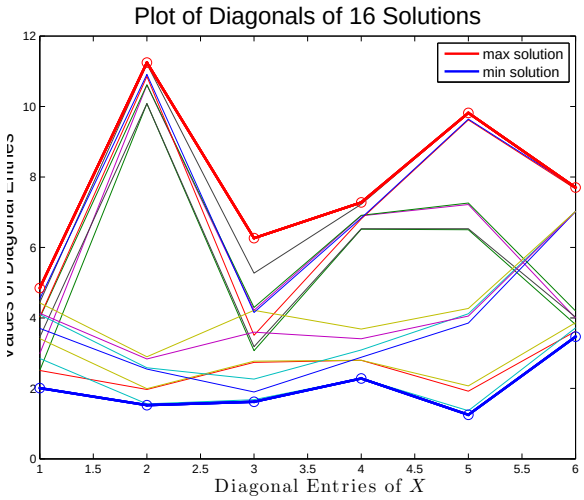

FIGURE 4.1. *Diagonal entries of 16 real-valued solutions to Example 1.*

Proof. The first claim follows from (2.12). Observe next that the difference

$$\delta\Gamma := \Gamma^{[1]} - \Gamma^{[2]}.$$

satisfies the homogeneous problem (3.2), implying that the matrix $S\delta\Gamma\Sigma_1^{\frac{1}{2}}$ must be hermitian. It follows that

$$U_1^*\mathfrak{s}(\delta\Gamma)U_1 = \frac{1}{2}U_1^*\left(U_2\Sigma_2^{\frac{1}{2}}\delta\Gamma\Sigma_1^{\frac{1}{2}}U_1^* + U_1\Sigma_1^{\frac{1}{2}}\delta\Gamma^*\Sigma_2^{\frac{1}{2}}U_2^*\right)U_1 = S\delta\Gamma\Sigma_1^{\frac{1}{2}}.$$

The second claim follows from the fact that $S\delta\Gamma\Sigma_1^{\frac{1}{2}} = \sum_{i=1}^{n^2}\left(\gamma_i^{[1]} - \gamma_i^{[2]}\right)S\Omega_i\Sigma_1^{\frac{1}{2}}$. □

Note that $S\Omega_i\Sigma_1^{\frac{1}{2}}$, $i = 1, \ldots, n^2$, are fixed hermitian matrices. What is interesting in the expression (4.2) is that it means a linear combination of these fixed matrices via coefficients $\delta\gamma_i := \gamma_i^{[1]} - \gamma_i^{[2]}$, $i = 1, \ldots, n^2$, should be in the cone of positive semi-definite matrices. Obviously, $\{\delta\gamma_i\}$ are restricted as there are only finitely many solutions $\{\gamma_i\}$ to (2.9).

By now, it should be clear on how to characterize the extreme solutions in terms of its corresponding $\{\gamma_i\}$. For the maximal solution in the complex problem, for example, we seek to solve this multi-objective optimization problem:

$$\max_{\{\gamma_i\}} \quad \sum_{i=1}^{n^2}\gamma_i\left(S\Omega_i\Sigma_1^{\frac{1}{2}}\right)^{\mathrm{D}}, \tag{4.3}$$

where recall that $\left(S\Omega_i\Sigma_1^{\frac{1}{2}}\right)^{\mathrm{D}}$ denotes the column vector of diagonal entries of the fixed hermitian matrix $S\Omega_i\Sigma_1^{\frac{1}{2}}$ and, therefore, is real-valued, subject to the equality constraint that $\{\gamma_i\}$ satisfies the system (3.8). We stress that, based on Theorem 3.1, there are only finitely many feasible solutions.

Example 2. Consider the problem (1.1) with data given in Example 1. We have mentioned that the alternating projection method can find a total of 16 real-valued solutions. With the Newton method, we gain precision and speed. It is not feasible to list all 16 solutions, so we simply plot their diagonals in Figure 4.1. The point to make is that the maximal (minimal) solution must be such that its diagonal entries are larger (smaller) than any other solution. The graph also clearly indicates that not all solutions can be partially ordered.

5. Conclusion. The linear matrix equation (1.1) has been studied extensively in the literature. This work revisits this problem from the Fejér-Riesz factorization point of view. The original Fejér-Riesz factorization theorem concerns an abstract factorization of a rational matrix-valued function over the unit disk. This approach offers a numerical procedure to realize such a factorization and makes it possible to find all solutions to the equation (1.1).

Specifically, it is shown that every HPD solution X to the nonlinear matrix equation can be expressed in a unique way as in (2.12) where Γ is a unitary solution to the linear Sylvester equation (2.9). The Sylvester equation might be easier to solve where the unitary constraint can be enforced via simple notion of projections. Two projection mechanisms are discussed — one is the usual Euclidean projection which gives rise to a minimal distance and the other is the Newton-type projection which does not refer to any coordinate frame.

REFERENCES

[1] W. N. ANDERSON, JR., T. D. MORLEY, AND G. E. TRAPP, *Positive solutions to $X = A - BX^{-1}B^*$*, Linear Algebra Appl., 134 (1990), pp. 53–62.

[2] J. P. BOYLE AND R. L. DYKSTRA, *A method for finding projections onto the intersection of convex sets in Hilbert spaces*, in Advances in order restricted statistical inference (Iowa City, Iowa, 1985), vol. 37 of Lecture Notes in Statist., Springer, Berlin, 1986, pp. 28–47.

[3] M. T. CHU, *Numerical methods for inverse singular value problems*, SIAM J. Numer. Anal., 29 (1992), pp. 885–903.

[4] M. A. DRITSCHEL AND J. ROVNYAK, *The operator Fejér-Riesz theorem*, in A glimpse at Hilbert space operators, vol. 207 of Oper. Theory Adv. Appl., Birkhäuser Verlag, Basel, 2010, pp. 223–254.

[5] J. C. ENGWERDA, *On the existence of a positive definite solution of the matrix equation $X + A^T X^{-1} A = I$*, Linear Algebra Appl., 194 (1993), pp. 91–108.

[6] J. C. ENGWERDA, A. C. M. RAN, AND A. L. RIJKEBOER, *Necessary and sufficient conditions for the existence of a positive definite solution of the matrix equation $X + A^* X^{-1} A = Q$*, Linear Algebra Appl., 186 (1993), pp. 255–275.

[7] L. EPHREMIDZE, G. JANASHIA, AND E. LAGVILAVA, *A simple proof of the matrix-valued Fejér-Riesz theorem*, J. Fourier Anal. Appl., 15 (2009), pp. 124–127.

[8] A. FERRANTE AND B. C. LEVY, *Hermitian solutions of the equation $X = Q + NX^{-1}N^*$*, Linear Algebra Appl., 247 (1996), pp. 359–373.

[9] C.-H. GUO AND P. LANCASTER, *Iterative solution of two matrix equations*, Math. Comp., 68 (1999), pp. 1589–1603.

[10] V. I. HASANOV AND S. M. EL-SAYED, *On the positive definite solutions of nonlinear matrix equation $X + A^\star X^{-\delta} A = Q$*, Linear Algebra Appl., 412 (2006), pp. 154–160.

[11] N. J. HIGHAM, *Computing the polar decomposition—with applications*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 1160–1174.

[12] I. G. IVANOV, V. I. HASANOV, AND F. UHLIG, *Improved methods and starting values to solve the matrix equations $X \pm A^* X^{-1} A = I$ iteratively*, Math. Comp., 74 (2005), pp. 263–278.

[13] J. B. KELLER, *Closest unitary, orthogonal and Hermitian operators to a given operator*, Math. Mag., 48 (1975), pp. 192–197.

[14] B. MEINI, *Efficient computation of the extreme solutions of $X + A^* X^{-1} A = Q$ and $X - A^* X^{-1} A = Q$*, Math. Comp., 71 (2002), pp. 1189–1204.

[15] A. C. M. RAN AND M. C. B. REURINGS, *On the nonlinear matrix equation $X + A^* \mathscr{F}(X) A = Q$: solutions and perturbation theory*, Linear Algebra Appl., 346 (2002), pp. 15–26.

[16] M. ROSENBLUM AND J. ROVNYAK, *Hardy classes and operator theory*, Dover Publications, Inc., Mineola, NY, 1997. Corrected reprint of the 1985 original.

[17] A. J. SOMMESE AND C. W. WAMPLER, II, *The numerical solution of systems of polynomials arising in engineering and science*, World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ, 2005.

[18] X. ZHAN, *Computing the extremal positive definite solutions of a matrix equation*, SIAM J. Sci. Comput., 17 (1996), pp. 1167–1174.