Received: 14 November 2018 | Revised: 17 May 2019 | Accepted: 23 July 2019

DOI: 10.1002/nla.2266

RESEARCH ARTICLE

WILEY

When is a matrix unitary or Hermitian plus low rank?

Gianna M. Del Corso[1] | Federico Poloni[1] | Leonardo Robol[2] | Raf Vandebril[3]

[1]Department of Computer Science, University of Pisa, Pisa, Italy

[2]Department of Mathematics, Institute of Information Science and Technologies, University of Pisa, Pisa, Italy

[3]Department of Computer Science, University of Leuven (KU Leuven), Leuven, Belgium

Correspondence
Gianna M. Del Corso, Department of Computer Science, University of Pisa, 56100 Pisa, Italy.
Email: gianna.delcorso@unipi.it

Funding information
Projects "Analisi di matrici sparse e data-sparse: metodi numerici ed applicazioni" and "Metodi di proiezione per equazioni di matrici e sistemi lineari con operatori definiti tramite somme di prodotti di Kronecker, e soluzioni con struttura di rango"; University of Pisa, Grant/Award Number: PRA-2017-05; Inverse free Rational Krylov Methods: Theory and Applications, Grant/Award Number: C14/16/056

Summary

Hermitian and unitary matrices are two representatives of the class of normal matrices whose full eigenvalue decomposition can be stably computed in quadratic computing complexity once the matrix has been reduced, for instance, to tridiagonal or Hessenberg form. Recently, fast and reliable eigensolvers dealing with low-rank perturbations of unitary and Hermitian matrices have been proposed. These structured eigenvalue problems appear naturally when computing roots, via confederate linearizations, of polynomials expressed in, for example, the monomial or Chebyshev basis. Often, however, it is not known beforehand whether or not a matrix can be written as the sum of a Hermitian or unitary matrix plus a low-rank perturbation. In this paper, we give necessary and sufficient conditions characterizing the class of Hermitian or unitary plus low-rank matrices. The number of singular values deviating from 1 determines the rank of a perturbation to bring a matrix to unitary form. A similar condition holds for Hermitian matrices; the eigenvalues of the skew-Hermitian part differing from 0 dictate the rank of the perturbation. We prove that these relations are linked via the Cayley transform. Then, based on these conditions, we identify the closest Hermitian or unitary plus rank k matrix to a given matrix A, in Frobenius and spectral norm, and give a formula for their distance from A. Finally, we present a practical iteration to detect the low-rank perturbation. Numerical tests prove that this straightforward algorithm is effective.

KEYWORDS

Hermitian plus low rank, rank-structured matrices, unitary plus low rank

1 | INTRODUCTION

Normal matrices are computationally among the most pleasant matrices to work with. The fact that their eigenvectors form a full orthogonal set is the basic ingredient for developing many stable algorithms. Even though generic normal matrices are less common in practice, the unitary and (skew-)Hermitian matrices are prominent members. The eigenvalue and system solvers for Hermitian[1,2] and unitary matrices[3–5] have been examined thoroughly, and well-tuned implementations are available in, for example, `eiscor`* and LAPACK.† Some matrices are not exactly Hermitian or unitary, yet a low-rank perturbation of those. These matrices are the subject of our study: Our aim is to provide a theoretical characterization of these matrices in terms of their singular- or eigenvalues. It has been noted that several linearizations of (matrix) polynomials are low-rank perturbations of unitary or Hermitian matrices (see, e.g., other works[6–10] and the

*EISCOR: eigenvalue solvers based on unitary core transformations. `https://github.com/eiscor/eiscor`
†LAPACK: linear algebra package. `https://www.netlib.org/lapack/`

https://doi.org/10.1002/nla.2266

references therein). Computing the roots of these polynomials hence coincides with computing the eigenvalues of the associated matrices. For instance, when computing roots of polynomials expressed in bases that admit a three-term recurrence such as the Chebyshev one, one ends up with the Comrade matrix, which is symmetric plus rank 1. Algorithms to solve this problem were developed by Chandrasekaran et al.,[11] Delvaux,[12] Vandebril et al.,[13] and the fastest and most reliable ones were due to Eidelman et al.[14] When studying orthogonal polynomials on the unit circle,[15,16] we end up with unitary-plus-low-rank matrices. The companion matrix, whose eigenvalues coincide with the roots of a polynomial in the monomial basis, is the most popular case. Various algorithms differing with respect to storage scheme, compression, explicit or implicit QR algorithms were proposed. We cite few and refer to the references therein for a full overview: Bini et al.,[17] Van Barel et al.,[18] Chandrasekaran et al.,[19] Boito et al.,[20] Bevilacqua et al.,[21] and a fast and provably stable version is presented in the book of Aurentz et al.[22] Extensions for efficiently handling corrections with larger rank, necessary to deal with block companion matrices, have been presented by Aurentz et al.,[23] Bini et al.,[24] Gemignani et al.,[25] Bevilacqua et al.,[26] and Delvaux.[12] In the context of Krylov methods, the structure of unitary and Hermitian plus low-rank matrices can be exploited as well. Not only it provides a fast matrix-vector multiplication but also the structure of the Galerkin projection profits from to design faster algorithms. Huckle[27] and Huhtanen (see the work of Huhtanen[28] and the references therein) analyzed the Arnoldi and Lanczos method for normal matrices. Barth et al.[29] examined classes of matrices resulting in short recurrences and proved that matrices whose adjoint is a low-rank perturbation of the original matrix allowed short CG recursions; a more detailed analysis can also be found in the overview of Liesen et al.[30] It is interesting to note that both unitary and Hermitian plus low-rank matrices allow for such recursions. Liesen examined in more detail the relation between a matrix and a rational function of its adjoint.[31] An alternative manner to devise the short recurrences was proposed by Beckermann et al.[32] An algorithm to exploit the short recurrences to develop efficient solvers for Hermitian plus low-rank case is the progressiver GMRES method, proposed by Beckermann et al.,[33] this method was tuned later on and stabilized by Embree et al.[34]

In this article, we characterize unitary and Hermitian plus low-rank matrices by examining their singular- and eigenvalues. We prove that a matrix having at most k singular values less than 1 and at most k greater than 1 is unitary plus rank k. Similarly, by examining the eigenvalues of the skew-Hermitian part of a matrix we show that if at most k of these eigenvalues are greater than 0 and at most k are smaller than zero, the matrix is Hermitian plus rank k. These characterizations enable us to determine the closest unitary or Hermitian plus rank k matrices in the spectral and Frobenius norms by setting some well-chosen singular- or eigenvalues to 1 or 0. We also show that the Cayley transform bridges between the Hermitian and unitary case. As a proof of concept, we designed and tested a straightforward Lanczos based algorithm to detect the low-rank part in some test cases.

The article is organized as follows. In Section 2, we revisit some preliminary results. Section 3 discusses necessary and sufficient conditions for a matrix to be unitary plus low rank. In Section 4, constructive proofs furnish the closest unitary plus low-rank matrix in spectral and Frobenius norm. Sections 5 and 6 discuss the analog of these results for the Hermitian plus low-rank structure. The Cayley transform, Section 7, allows us to transform the unitary into the Hermitian problem. Algorithms to extract the low-rank part from a Hermitian (unitary) plus low-rank matrix and some experiments are proposed in Sections 8 and 9. We conclude in Section 10.

2 | PRELIMINARIES

In this text, we make use of the following conventions. The symbols I and 0 denote the identity and zero matrix and may have subscripts denoting their size whenever that is not clear from the context. We use $\sigma_1(M) \geq \sigma_2(M) \geq \dots \geq \sigma_n(M)$ to denote the singular values of a matrix $M \in \mathbb{C}^{n\times n}$, and $\lambda_1(H) \geq \lambda_2(H) \geq \dots \geq \lambda_n(H)$ stand for the eigenvalues of a Hermitian matrix $H \in \mathbb{C}^{n\times n}$. We use the diag operator, which stacks its arguments, which could be scalars, matrices, or vectors, in a block diagonal matrix (possibly with nonsquare blocks on its diagonal).

The following results are classical. We rely on them in the forthcoming proofs, and for completeness, we have included them.

Theorem 1 (Weyl's inequalities [see Theorem 4.3.16 and Exercise 16 in Horn et al.[35p. 423]]). *For every pair of matrices $M, N \in \mathbb{C}^{n\times n}$ and for every i, j such that $i + j \leq n + 1$,*

$$\sigma_{i+j-1}(M \pm N) \leq \sigma_i(M) + \sigma_j(N).$$

If M, N are Hermitian, then the same inequality holds for their eigenvalues.

$$\lambda_{i+j-1}(M\pm N) \leq \lambda_i(M) + \lambda_j(N).$$

Theorem 2 (Interlacing inequalities [see Theorem 4.3.4 in the work of Horn et al.,[35] and the work of Thompson[36]]). *Let $M \in \mathbb{C}^{n\times n}$ and $N \in \mathbb{C}^{n\times(n-k)}$ be a submatrix of M obtained by removing k columns from it. Then,*

$$\sigma_{i+k}(M) \leq \sigma_i(N) \leq \sigma_i(M).$$

In the Hermitian case, we get similar inequalities. Let $M \in \mathbb{C}^{n\times n}$ be Hermitian and $N \in \mathbb{C}^{(n-k)\times(n-k)}$ be a (Hermitian) principal submatrix of M. Then,

$$\lambda_{i+k}(M) \leq \lambda_i(N) \leq \lambda_i(M).$$

Moreover, recall that $M \in \mathbb{C}^{n\times n}$ is unitary if and only if $\sigma_i(M) = 1, \forall i = 1, \dots, n$.

3 | DETECTING UNITARY-PLUS-RANK-k MATRICES

We call $\mathcal{U}_k$ the set of unitary-plus-rank-k matrices, that is, $A \in \mathcal{U}_k$ if and only if there exists a unitary matrix Q and two skinny matrices $G, B \in \mathbb{C}^{n\times k}$ such that $A = Q + GB^*$. This implies that $\mathcal{U}_k \subseteq \mathcal{U}_{k+1}$ for any k.

Theorem 3. *Let $A \in \mathbb{C}^{n\times n}$ and $0 \leq k \leq n$. Then, $A \in \mathcal{U}_k$ if and only if A has at most k singular values strictly greater than 1 and at most k singular values strictly smaller than 1.*

Before proving this result, we point out that looking at the singular values is a good guess, because being unitary-plus-rank-k is invariant under unitary equivalence transformations.

Remark 1. $A = Q + GB^* \in \mathcal{U}_k$ if and only if $U^*AV \in \mathcal{U}_k$ for any unitary matrices U, V. Indeed, we have that

$$U^*AV = \underbrace{U^*QV}_{\hat{Q}} + \underbrace{U^*G}_{\hat{G}}\underbrace{B^*V}_{\hat{B}^*} \in \mathcal{U}_k.$$

We start proving a simple case ($n = 2, k = 1$) of Theorem 3, which will act as a building block for the general proof.

Lemma 1. *For every pair of real numbers σ_1 and σ_2 such that $\sigma_1 \geq 1 \geq \sigma_2 \geq 0$, we have*

$$\begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix} \in \mathcal{U}_1.$$

Proof. We prove that the diagonal 2×2 matrix can be decomposed as a plane rotation plus a rank 1 correction. In particular, we look for $c, s, a, b \geq 0$ such that

$$\begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix} = \begin{bmatrix} c & s \\ -s & c \end{bmatrix} + \begin{bmatrix} a & -s \\ s & -b \end{bmatrix},$$

that is, $\sigma_1 = c + a$ and $\sigma_2 = c - b$. In addition, we impose that the first summand is unitary (i.e., $c^2 + s^2 = 1$), and the second has rank 1 (i.e., $s^2 = ab$). A simple computation shows that

$$c = \frac{\sigma_1\sigma_2 + 1}{\sigma_1 + \sigma_2}, \quad a = \frac{\sigma_1^2 - 1}{\sigma_1 + \sigma_2}, \quad b = \frac{1 - \sigma_2^2}{\sigma_1 + \sigma_2}, \quad \text{and} \quad s = \sqrt{ab}$$

satisfy these conditions. □

We are now ready to prove Theorem 3.

Proof of Theorem. First, note that the case $k = n$ is trivial $\mathcal{U}_n = \mathbb{C}^{n\times n}$. Therefore, we assume $k < n$. The conditions on the singular values can be written as two inequalities

$$1 \geq \sigma_{k+1}(A) \quad \text{and} \quad \sigma_{n-k}(A) \geq 1. \tag{1}$$

- We first show that if $A \in \mathcal{U}_k$, then the inequalities (1) hold. Suppose that $A = Q + GB^*$. Then, by Theorem 1,

  $$\sigma_{k+1}(A) \leq \sigma_1(Q) + \sigma_{k+1}(GB^*) = 1 + 0 = 1.$$

  Hence, again, following from Theorem 1,

  $$1 = \sigma_n(Q) \leq \sigma_{n-k}(A) + \sigma_{k+1}(GB^*) = \sigma_{n-k}(A).$$

- We now prove the reverse implication, that is, if the two inequalities (1) hold, then $A \in \mathcal{U}_k$. To simplify things, we introduce k_- denoting the number of singular values smaller than 1 and k_+ standing for the number of singular values larger than 1. The conditions state that $\ell = \max\{k_-, k_+\} \leq k$. We will prove that $A \in \mathcal{U}_\ell$. Note that $\mathcal{U}_\ell \subseteq \mathcal{U}_k$. Let $h = \min\{k_-, k_+\}$.

  We reorder the diagonal elements of Σ to group the singular values into separate diagonal blocks of three types:

  - Diagonal blocks $\Sigma_1, \Sigma_2, \dots, \Sigma_h$ of size 2×2, containing each one singular value larger than 1 and one smaller than 1. Because $h = \min\{k_-, k_+\}$, either all singular values smaller than 1 or all singular values larger than 1 are incorporated in these blocks.
  - Diagonal blocks $\Sigma_{h+1}, \dots, \Sigma_\ell$ of size 1×1, containing the remaining singular values different from 1. Note that all these blocks will contain either singular values that are larger than 1 or smaller than 1, depending on whether $h = k_-$ or $h = k_+$. In case $k_- = k_+ = h = \ell$, there will be no blocks of this type.
  - One final block equal to the identity matrix of size $m = n - h - \ell = n - k_- - k_+$, containing all the singular values equal to 1.

  For example, for A having singular values $(3, 2, 2, 1.3, 1.2, 1, 1, 1, 0.6, 0.2)$, we can take $\Sigma_1 = \operatorname{diag}(3, 0.2)$, $\Sigma_2 = \operatorname{diag}(2, 0.6)$, $\Sigma_3 = 2$, $\Sigma_4 = 1.3$, $\Sigma_5 = 1.2$, and $\Sigma_6 = I_3$.

  Clearly, this decomposition always exists, and although it is not unique, the number and types of blocks are. For each $i = 1, 2, \dots, \ell$, the matrix Σ_i is unitary plus rank 1. This follows from Lemma 1 for 2×2 blocks, and is trivial for the 1×1 blocks. Hence, for each i, we can write $\Sigma_i = Q_i + g_i b_i^*$, where Q_i is unitary and g_i, b_i are vectors. Therefore, we have

  $$\operatorname{diag}(\Sigma_1, \Sigma_2, \dots, \Sigma_\ell, I_m) = \operatorname{diag}(Q_1, Q_2, \dots, Q_\ell, I_m) + GB^*,$$

  with

  $$G = \begin{bmatrix} \operatorname{diag}(g_1, g_2, \dots, g_\ell) \\ 0_{m\times\ell} \end{bmatrix}, \quad \text{and} \quad B = \begin{bmatrix} \operatorname{diag}(b_1, b_2, \dots, b_\ell) \\ 0_{m\times\ell} \end{bmatrix}.$$

  Therefore $\operatorname{diag}(\Sigma_1, \Sigma_2, \dots, \Sigma_\ell, I) \in \mathcal{U}_\ell$. Because this matrix can be obtained from a unitary equivalence on A (singular value decomposition and a permutation), we have $A \in \mathcal{U}_\ell$.

□

Example 1. The matrix $U \operatorname{diag}(3, 2, 1, 1, 1, 0.5)V^*$ belongs to $\mathcal{U}_2$ (but not to $\mathcal{U}_1$). The matrix $U \operatorname{diag}(5, 0.4, 0.3, 0.2)V^*$ belongs to $\mathcal{U}_3$ (but not to $\mathcal{U}_2$). The matrix $5I_4$ belongs to $\mathcal{U}_4$ (but not to $\mathcal{U}_3$).

4 | DISTANCE FROM UNITARY PLUS RANK k

Theorem 3 provides an effective criterion to characterize matrices in $\mathcal{U}_k$ based on the singular values. One of the main features of the singular value decomposition is that it automatically provides the optimal rank k approximation of any matrix, in the sense of the 2- and the Frobenius norm. In this section, we show that the criterion of Theorem 3 can be used to compute the best unitary-plus-rank-k approximant for any value of k. The problem is thus to find a matrix in $\mathcal{U}_k$ that minimizes the distance to a given matrix A. This can be achieved by setting the supernumerary singular values preventing the inequalities (1) from being satisfied to 1.

Theorem 4. *Let the matrix $A \in \mathbb{C}^{n\times n}$ have singular value decomposition $U\Sigma V^*$, and denote by $\hat{A} = U\hat{\Sigma}V^*$, where $\hat{\Sigma}$ is the diagonal matrix with elements*

$$\hat{\sigma}_i = \begin{cases} 1, & \text{if } k < i \leq k_+, \text{ or } n - k_- < i \leq n - k, \\ \sigma_i, & \text{otherwise} \end{cases}$$

with k_+ (respectively k_-) the number of singular values of A strictly greater (respectively smaller) than 1, we have that

$$\min_{X\in\mathcal{U}_k} \|A - X\|_2 = \|A - \hat{A}\|_2 = \max\{0, \sigma_{k+1} - 1, 1 - \sigma_{n-k}\}.$$

Proof. The matrix $\hat{A} = U\hat{\Sigma}V^*$ clearly belongs to $\mathcal{U}_k$ by Theorem 3; hence, $\|A - U\hat{\Sigma}V^*\|_2 = \|\Sigma - \hat{\Sigma}\|_2 = \max\{\sigma_{k+1} - 1, 1 - \sigma_{n-k}, 0\}$. It remains thus to prove that for every $X \in \mathcal{U}_k$, one has $\|A - X\|_2 \geq \|A - \hat{A}\|_2$.

By Weyl's inequality (Theorem 1), we have that $\sigma_{k+1}(A) \leq \sigma_1(A - X) + \sigma_{k+1}(X)$; therefore,

$$\|A - X\|_2 \geq \sigma_{k+1}(A) - \sigma_{k+1}(X) \geq \sigma_{k+1}(A) - 1.$$

Similarly, from $\sigma_{n-k}(X) \leq \sigma_1(A - X) + \sigma_{n-k}(A)$, we have

$$\|A - X\|_2 \geq \sigma_{n-k}(X) - \sigma_{n-k}(A) \geq 1 - \sigma_{n-k}(A).$$

We obtain

$$\|A - X\|_2 \geq \max\{0, \sigma_{k+1} - 1, 1 - \sigma_{n-k}\} = \|A - \hat{A}\|_2.$$ □

Theorem 4 provides a deterministic construction for a minimizer of $\|A - X\|_2$, but this minimizer is not unique. This is demonstrated in the following example.

Example 2. Let us consider, for arbitrary unitary matrices U, V, the matrix A defined as

$$A = U\begin{bmatrix} 2 & & & & \\ & 1.5 & & & \\ & & 1 & & \\ & & & 1 & \\ & & & & .5 \end{bmatrix}V^*.$$

We know from Theorem 3 that $A \in \mathcal{U}_2$. We want to determine the distance from A to $\mathcal{U}_1$. Theorem 4 yields the approximant $\hat{A}$ defined as follows:

$$\hat{A} := U\begin{bmatrix} 2 & & & & \\ & 1 & & & \\ & & 1 & & \\ & & & 1 & \\ & & & & .5 \end{bmatrix}V^*, \qquad \|A - \hat{A}\|_2 = 0.5.$$

However, the solution is not unique. For instance, the family of matrices determined as

$$\hat{A}(t,s) := U\begin{bmatrix} 2+t & & & & \\ & 1 & & & \\ & & 1 & & \\ & & & 1 & \\ & & & & .5+s \end{bmatrix}V^*$$

still lead to $\|A - \hat{A}(t,s)\|_2 = 0.5$ for $t, s \in [-0.5, 0.5]$.

The same matrix is a minimizer also in the Frobenius norm.

Theorem 5. *Let the matrix $A \in \mathbb{C}^{n\times n}$ have singular value decomposition $U\Sigma V^*$ and denote by $\hat{A} = U\hat{\Sigma}V^*$, where $\hat{\Sigma}$ is the diagonal matrix with elements*

$$\hat{\sigma}_i = \begin{cases} 1, & \text{if } k < i \le k_+, \text{or } n - k_- < i \le n - k, \\ \sigma_i, & \text{otherwise}, \end{cases}$$

where k_+ (respectively k_-) is the number of singular values of A, which are strictly greater (respectively smaller) than 1. Then, we have

$$\min_{X\in\mathcal{U}_k} \|A - X\|_F = \|A - \hat{A}\|_F,$$

and moreover,

$$\|A - \hat{A}\|_F^2 = \sum_{i=k+1}^{k_+} (\sigma_i - 1)^2 + \sum_{i=n-k_-+1}^{n-k} (\sigma_i - 1)^2. \tag{2}$$

Proof. Because the Frobenius norm is unitarily invariant, we immediately have (2). To complete the proof, we show that for an arbitrary $X \in \mathcal{U}_k$, we have $\|A - X\|_F \ge \|A - \hat{A}\|_F$. Let Δ be the matrix such that $X = A + \Delta \in \mathcal{U}_k$. We will prove that $\|\Delta\|_F^2 \ge \|A - \hat{A}\|_F^2$. Consider $\tilde{\Delta} = U^*\Delta V$ and partition

$$U^*XV = U^*(A + \Delta)V = \left[\, \Sigma_1 + \tilde{\Delta}_1 \;\; \Sigma_2 + \tilde{\Delta}_2 \;\; \Sigma_3 + \tilde{\Delta}_3 \,\right],$$

where $\Sigma_1 \in \mathbb{C}^{n\times k_+}$ contains the singular values of A, which are larger than 1, Σ_2 the singular values equal to 1, and $\Sigma_3 \in \mathbb{C}^{n\times k_-}$ the singular values smaller than 1.

Because $U^*(A + \Delta)V \in \mathcal{U}_k$, we have

$$\sigma_{k+1}(U^*(A + \Delta)V) \le 1, \quad \sigma_{n-k}(U^*(A + \Delta)V) \ge 1.$$

- Assume first that $k_+ > k$. By the interlacing inequalities (Theorem 2),

  $$\sigma_{k+1}(\Sigma_1 + \tilde{\Delta}_1) \le \sigma_{k+1}(U^*(A + \Delta)V) \le 1,$$

  and then, by Weyl's inequalities (Theorem 1) for each $k < i \le k_+$,

  $$\sigma_i = \sigma_i(\Sigma_1) \le \sigma_{k+1}(\Sigma_1 + \tilde{\Delta}_1) + \sigma_{i-k}(\tilde{\Delta}_1) \le 1 + \sigma_{i-k}(\tilde{\Delta}_1),$$

  from which we obtain $\sigma_{i-k}(\tilde{\Delta}_1) \ge \sigma_i - 1$; hence,

  $$\|\tilde{\Delta}_1\|_F^2 \ge \sum_{i=1}^{k_+-k} \sigma_i(\tilde{\Delta}_1)^2 \ge \sum_{i=k+1}^{k_+} (\sigma_i - 1)^2. \tag{3}$$

  Note that (3) holds trivially also when $k_+ \le k$.
- Similarly, assume that $k_- > k$ (the case $k_- \le k$ is trivial), and we use interlacing inequalities (Theorem 2) to get

  $$\sigma_{k_--k}(\Sigma_3 + \tilde{\Delta}_3) \ge \sigma_{n-k}(U^*(A + \Delta)V) \ge 1,$$

  and Weyl's inequalities (Theorem 1) for each $k < i \le k_-$ to show

  $$1 \le \sigma_{k_--k}(\Sigma_3 + \tilde{\Delta}_3) \le \sigma_{k_-+1-i}(\Sigma_3) + \sigma_{i-k}(\tilde{\Delta}_3) = \sigma_{n+1-i} + \sigma_{i-k}(\tilde{\Delta}_3),$$

  from which we obtain $\sigma_{i-k}(\tilde{\Delta}_3) \ge 1 - \sigma_{n+1-i}$. Hence,

  $$\|\tilde{\Delta}_3\|_F^2 \ge \sum_{j=n+1-k_-}^{n-k} (1 - \sigma_j)^2. \tag{4}$$

Putting together (3) and (4), we have

$$\|\Delta\|_F^2 \geq \|\tilde{\Delta}_1\|_F^2 + \|\tilde{\Delta}_3\|_F^2 \geq \sum_{i=k+1}^{k_+} (\sigma_i - 1)^2 + \sum_{j=n+1-k_-}^{n-k} (1-\sigma_j)^2 = \|A - \hat{A}\|_F^2,$$

which is precisely what we wanted to prove. □

In this case, unlike in the 2-norm setting, this minimizer is unique when all singular values are different.[‡]

5 | DETECTING HERMITIAN-PLUS-RANK-k MATRICES

We call $\mathcal{H}_k$ the set of Hermitian-plus-rank-k matrices, that is, $A \in \mathcal{H}_k$ if and only if there exists a Hermitian matrix H and two matrices $G, B \in \mathbb{C}^{n\times k}$ such that $A = H + GB^*$.

In this and the next section, we will answer similar questions: How can we tell if $A \in \mathcal{H}_k$? How do we find the distance from a matrix A to the closed set $\mathcal{H}_k$? To answer these questions, we first need a Hermitian equivalent of Lemma 1.

Lemma 2. *For any pair of real numbers σ_1 and σ_2, such that $\sigma_1 \geq 0, \sigma_2 \leq 0$, there are two real vectors c and b such that*

$$\begin{bmatrix} \sigma_1 & \\ & \sigma_2 \end{bmatrix} = bc^T + cb^T. \tag{5}$$

Proof. Define the vectors c and b as follows:

$$b := \frac{1}{2}\begin{bmatrix} \sqrt{\sigma_1} \\ -\sqrt{-\sigma_2} \end{bmatrix}, \qquad c := \begin{bmatrix} \sqrt{\sigma_1} \\ \sqrt{-\sigma_2} \end{bmatrix}.$$

The result follows by a direct computation. □

The next Theorem is the analog of Theorem 3, where we look at eigenvalues of the skew-Hermitian part of a matrix instead of at the singular values. We rely on the following lemma.

Lemma 3. *Let B, C be any $n \times k$ full rank matrices, and let $S = BC^* + CB^*$. Then, S has at most k positive and at most k negative eigenvalues.*

Proof. Up to a change of basis, we can assume $C = \begin{bmatrix} I_k \\ 0 \end{bmatrix}$. Then, S has a trailing $(n-k)\times(n-k)$ zero submatrix T. Let $\lambda_1(S) \geq \ldots \geq \lambda_n(S)$ be the eigenvalues of S. By the interlacing inequalities, $\lambda_{k+1}(S) \leq \lambda_1(T) = 0$, and $\lambda_{n-k}(S) \geq \lambda_{n-k}(T) = 0$. □

Theorem 6. *Let $A \in \mathbb{C}^{n\times n}$, and $0 \leq k \leq n$. Then, $A \in \mathcal{H}_k$ if and only if the Hermitian matrix $S(A) := \frac{1}{2i}(A - A^*)$ has at most k positive eigenvalues and at most k negative eigenvalues.*

Proof. Let $\lambda_1, \ldots, \lambda_n$ be the eigenvalues of $\frac{1}{2i}(A - A^*)$, sorted by decreasing order, that is, $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$. Then, the stated condition is equivalent to demanding $\lambda_{k+1} \leq 0$, $\lambda_{n-k} \geq 0$. We first show that if $A \in \mathcal{H}_k$, these inequalities hold. If $A \in \mathcal{H}_k$, then there exists a Hermitian matrix H and two matrices $G, B \in \mathbb{C}^{n\times k}$ such that $A = H + GB^*$. Then,

$$\frac{1}{2i}(A - A^*) = \frac{1}{2i}(GB^* - BG^*) = CB^* + BC^*,$$

with $C = G/(2i)$. The result follows from Lemma 3.

[‡] In fact, the constraint of all singular values differing can be relaxed: One can construct examples in which there are several identical singular values and all (or none) of them should be changed to 1.

We now prove the converse, that is, each matrix satisfying the inequalities belongs to $\mathcal{H}_k$. We first prove that each Hermitian matrix S that has k_+ strictly positive eigenvalues and k_- strictly negative eigenvalues, where $\ell = \max\{k_-, k_+\} \leq k$ can be written as $CB^* + BC^*$ with $B, C \in \mathbb{C}^{n\times\ell}$.

Let us assume that the eigenvalues of S are λ_j with

$$\lambda_1, \ldots, \lambda_{k_+} > 0, \qquad \lambda_{k_++1}, \ldots, \lambda_{k_++k_-} < 0, \qquad \lambda_{k_++k_-+1}, \ldots, \lambda_n = 0.$$

Because S is normal, we can diagonalize it by an orthogonal transformation and obtain

$$U^*SU = \mathrm{diag}(\Lambda_1, \ldots, \Lambda_h, \Lambda_{h+1}, \ldots, \Lambda_\ell, 0_m),$$

where we get, as in the proof of the unitary case, three types of blocks as follows.

- Diagonal blocks $\Lambda_1, \Lambda_2, \ldots, \Lambda_h$ of size 2×2 containing one eigenvalue larger and one eigenvalue smaller than 0.
- 1×1 matrices $\Lambda_{h+1}, \ldots, \Lambda_\ell$ containing the remaining eigenvalues differing from 0. Because either all positive or negative eigenvalues are used already to form the blocks Λ_1 up to Λ_h, we end up with scalar blocks that all have the same sign.
- A final zero matrix of size $m = n - k_- - k_+ = n - 2h - \ell$.

Lemma 2 tells us that each of the 2×2 blocks Λ_j, for $j = 1, \ldots, h$ can be written as $b_jc_j^* + c_jb_j^*$, for appropriate choices of b_j, c_j, because the eigenvalues on the diagonal are real and have opposite sign. Moreover, the remaining diagonal entries Λ_j for $j = h+1, \ldots, \ell$ can be written choosing $b_j = \lambda_j$ and $c_j = 1/2$. Therefore, we conclude that Q^*SQ is of the form $\tilde{C}\tilde{B}^T + \tilde{B}\tilde{C}^T$ for some $\tilde{C}, \tilde{B}$ with ℓ columns. Setting $G := (-2i)Q\tilde{C}$ and $B := Q\tilde{B}$ proves our claim. It is immediate to verify that the matrix $A - GB^*$ is Hermitian, because $(A - GB^*)^* - (A - GB^*) = 0$. □

Theorem 6 has alternative formulations as well. We can for instance look at $A - A^*$ and count the number of eigenvalues with positive and negative imaginary parts, because when A is Hermitian, iA will be skew-Hermitian.

We will not go into the details, but it is obvious that Theorem 6 admits an equivalent formulation to check whether a matrix is a rank k perturbation of a skew-Hermitian matrix. To this end one considers $A + A^*$ and counts the number of positive and negative eigenvalues.

6 | DISTANCE FROM HERMITIAN PLUS RANK k

In this section, we will construct, given an arbitrary matrix A, the closest Hermitian-plus-rank-k matrix in both the 2- and Frobenius norm. The following lemma comes in handy.

Lemma 4. *Let $\|\cdot\|$ be any unitarily invariant norm, and $X \in \mathbb{C}^{n\times n}$, and $S(X) = \frac{X - X^*}{2i}$. Then, $\|S(X)\| \leq \|X\|$.*

Proof. Let $X = U\Sigma V^*$ be the SVD of X. Because $\|\cdot\|$ is unitarily invariant, we have that $\|X^*\| = \|V\Sigma^*U^*\| = \|\Sigma\| = \|U\Sigma V^*\| = \|X\|$. Then,

$$\left\|\frac{X - X^*}{2i}\right\| \leq \frac{\|X\| + \|X^*\|}{2} = \|X\|.$$

□

We formulate again two Theorems, one for the closest approximation in the 2-norm and another one for the closest approximation in the Frobenius norm. The approximants that we construct are identical, but the proofs differ significantly. Again, like in the unitary case, we will change particular eigenvalues of the skew-Hermitian part to find the best approximant.

Theorem 7. *Let $A \in \mathbb{C}^{n\times n}$ and $S(A) = \frac{1}{2i}(A - A^*)$, having eigendecomposition $S(A) = UDU^*$. The eigenvalues are ordered: $\lambda_1 \geq \ldots \geq \lambda_{k_+} > 0$, $\lambda_{k_++1} = \ldots = \lambda_{n-k_-} = 0$ and $0 > \lambda_{n-k_-+1} \geq \ldots \geq \lambda_n$, where k_+ stands for the number of eigenvalues strictly greater than 0 and k_- for the eigenvalues strictly smaller than 0. Then, we have that*

$$\min_{X\in\mathcal{H}_k} \|A - X\|_2 = \|A - \hat{A}\|_2,$$

where $\hat{A} = A - iU(D - \hat{D})U^*$ and $\hat{D}$ is a diagonal matrix with diagonal elements

$$\hat{d}_i = \begin{cases} 0, & \text{if } k < i \le k_+ \text{ or } n - k_- < i \le n - k \\ \lambda_i, & \text{otherwise.} \end{cases}$$

Proof. Theorem 6 implies that $\hat{A} = A - iU(D - \hat{D})U^*$ belongs to $\mathcal{H}_k$ because

$$S(\hat{A}) = \frac{\hat{A} - \hat{A}^*}{2i} = \frac{A - A^*}{2i} - U(D - \hat{D})U^* = U\hat{D}U^*$$

has at most k positive eigenvalues and at most k negative eigenvalues.

To prove that $\hat{A}$ is the minimizer of $\|A - X\|_2$, we have to prove that for every $X \in \mathcal{H}_k$ we have that $\|A - X\|_2 \ge \|A - \hat{A}\|_2 = \|D - \hat{D}\|_2 = \max\{\lambda_{k+1}(S(A)), -\lambda_{n-k}(S(A)), 0\}$.

Assume that $X = A + \Delta$, then $S(X) = S(A) + S(\Delta)$. Using Weyl's inequality (Theorem 1), we have

$$\lambda_{k+1}(S(X)) = \lambda_{k+1}(S(A) + S(\Delta)) \ge \lambda_{k+1}(S(A)) - \lambda_1(S(\Delta));$$

therefore,

$$\|A - X\|_2 = \|\Delta\|_2 \ge \|S(\Delta)\|_2 \ge \lambda_1(S(\Delta)) \ge \lambda_{k+1}(S(A)) - \lambda_{k+1}(S(X)) \ge \lambda_{k+1}(S(A))$$

because $X \in \mathcal{H}_k$ implies $\lambda_{k+1}(S(X)) \le 0$. Similarly from

$$\lambda_{n-k}(S(A + \Delta)) \le \lambda_1(S(\Delta)) + \lambda_{n-k}(S(A)),$$

we get

$$\|A - X\|_2 \ge \lambda_1(S(\Delta)) \ge \lambda_{n-k}(S(X)) - \lambda_{n-k}(S(A)) \ge -\lambda_{n-k}(S(A))$$

because $X \in \mathcal{H}_k$ implies $\lambda_{n-k}(S(X)) \ge 0$. Combining the two inequalities and using the nonnegativeness of the norm, we have

$$\|A - X\|_2 \ge \max\{\lambda_{k+1}(S(A)), -\lambda_{n-k}(S(A)), 0\}.$$

□

We remark that, comparable to the unitary case, the minimizer in the 2-norm is not unique. We have constructed a solution $\hat{A}$ such that $\|A - \hat{A}\|_2 = \|D - \hat{D}\|_2 = \max\{\lambda_{k+1}(S(A)), -\lambda_{n-k}(S(A)), 0\}$. It is, however, easy to find a concrete example and a matrix Ã different from $\hat{A}$ such that $\|A - \hat{A}\|_2 = \|D - \hat{D}\|_2 = \|D - \tilde{D}\|_2 = \|A - \tilde{A}\|_2$.

Theorem 8. *Let $A \in \mathbb{C}^{n\times n}$ and $S(A) = \frac{1}{2i}(A - A^*)$, having eigendecomposition $S(A) = UDU^*$. The eigenvalues are ordered: $\lambda_1 \ge \ldots \ge \lambda_{k_+} > 0$, $\lambda_{k_+ +1} = \ldots = \lambda_{n-k_-} = 0$, and $0 > \lambda_{n-k_- +1} \ge \ldots \ge \lambda_n$, where k_+ stands for the number of eigenvalues strictly greater than 0 and k_- stands for the eigenvalues strictly smaller than 0. Then, we have that*

$$\min_{X \in \mathcal{H}_k} \|A - X\|_F = \|A - \hat{A}\|_F,$$

where $\hat{A} = A - iU(D - \hat{D})U^*$, and $\hat{D}$ is a diagonal matrix with diagonal elements

$$\hat{d}_i = \begin{cases} 0, & \text{if } k < i \le k_+ \text{ or } n - k_- < i \le n - k \\ \lambda_i, & \text{otherwise.} \end{cases}$$

Proof. We know that $\hat{A} \in \mathcal{H}_k$. To prove that this is the minimizer, we have to show that for any $X \in \mathcal{H}_k$, we have that $\|A - X\|_F \ge \|A - \hat{A}\|_F$.

Consider $\Delta \in \mathbb{C}^{n\times n}$ such that $X = A + \Delta \in \mathcal{H}_k$. We know from Theorem 6 that this implies that $\lambda_{k+1}(S(A + \Delta)) \le 0$ and that $\lambda_{n-k}(S(A + \Delta)) \ge 0$. Consider $\tilde{\Delta} = U^* S(\Delta) U$ and partition it as follows:

$$U^* S(A + \Delta) U = \begin{bmatrix} D_1 + \tilde{\Delta}_{11} & \tilde{\Delta}_{12} & \tilde{\Delta}_{13} \\ \tilde{\Delta}_{21} & D_2 + \tilde{\Delta}_{22} & \tilde{\Delta}_{23} \\ \tilde{\Delta}_{31} & \tilde{\Delta}_{32} & D_3 + \tilde{\Delta}_{33} \end{bmatrix},$$

where $D = \operatorname{diag}(D_1, D_2, D_3)$ is the diagonal of the eigendecomposition $S(A) = UDU^*$. In particular, $D_1 \in \mathbb{R}^{k_+ \times k_+}$ contains the eigenvalues of $S(A)$, which are strictly greater than 0, and $D_3 \in \mathbb{R}^{k_- \times k_-}$ contains the eigenvalues of $S(A)$ strictly smaller than 0.

- Assume that $k_+ > k$. The matrix $U^* S(A + \Delta)U$ is Hermitian; hence, by the interlacing inequalities, we get

  $$\lambda_{k+1}(D_1 + \tilde{\Delta}_{11}) \leq \lambda_{k+1}(S(A + \Delta)) \leq 0,$$

  and then, by Weyl's inequalities for each $k < i \leq k_+$,

  $$\lambda_i = \lambda_i(D_1) \leq \lambda_{k+1}(D_1 + \tilde{\Delta}_{11}) + \lambda_{i-k}(\tilde{\Delta}_{11}) \leq \lambda_{i-k}(\tilde{\Delta}_{11}),$$

  from which we obtain $\lambda_{i-k}(\tilde{\Delta}_{11}) \geq \lambda_i$; hence,

  $$\|\tilde{\Delta}_{11}\|_F^2 = \sum_{i=1}^{k_+} \lambda_i(\tilde{\Delta}_{11})^2 \geq \sum_{i=k+1}^{k_+} \lambda_i^2. \tag{6}$$

  Note that (6) holds trivially also when $k_+ \leq k$.
- Similarly, assume that $k_- > k$, and use again the interlacing inequalities to get

  $$\lambda_{k_- - k}(D_3 + \tilde{\Delta}_{33}) \geq \lambda_{n-k}(S(A + \Delta)) \geq 0.$$

  Using Weyl's inequalities for each $k < i \leq k_-$, we can show that

  $$0 \leq \lambda_{k_- - k}(D_3 + \tilde{\Delta}_{33}) \leq \lambda_{k_- + 1 - i}(D_3) + \lambda_{i-k}(\tilde{\Delta}_{33}) = \lambda_{n+1-i} + \lambda_{i-k}(\tilde{\Delta}_{33}),$$

  from which we obtain $\lambda_{i-k}(\tilde{\Delta}_{33}) \geq -\lambda_{n+1-i} \geq 0$; hence,

  $$\|\tilde{\Delta}\|_{33F}^{2} \geq \sum_{i=k+1}^{k_-} \lambda_{i-k}(\tilde{\Delta}_{33})^2 \geq \sum_{i=k+1}^{k_-} (-\lambda_{n+1-i})^2 = \sum_{j=n+1-k_-}^{n-k} \lambda_j^2. \tag{7}$$

  We note that this equation holds trivially when $k_- \leq k$.

Combining the inequalities (6) and (7), and by Lemma 4 stating that $\|\Delta\|_F^2 \geq \|S(\Delta)\|_F^2$, we get

$$\|A - X\|_F^2 = \|\Delta\|_F^2 \geq \|\tilde{\Delta}_{11}\|_F^2 + \|\tilde{\Delta}_{33}\|_F^2 \geq \sum_{i=k+1}^{k_+} \lambda_i^2 + \sum_{j=n+1-k_-}^{n-k} \lambda_j^2,$$

which is precisely $\|D - \hat{D}\|_F = \|A - \hat{A}\|_F$. This concludes the proof. □

7 | THE CAYLEY TRANSFORM

Unitary and Hermitian structures are both special cases of normal matrices. Even more interestingly, it is known that they can be mapped one into the other through the use of the Cayley transform, defined as follows:

$$\mathcal{C}(z) := \frac{z - i}{z + i}, \qquad z \in \mathbb{C} \setminus \{-i\}.$$

The Cayley transform is a particular case of a Möbius transform, which permutes projective lines of the Riemann sphere. In particular, we have that $\mathcal{C}(\mathbb{R}) = \mathbb{S}^1$. The inverse transform can be readily expressed as

$$\mathcal{C}^{-1}(z) = i \cdot \frac{1 + z}{1 - z}, \qquad z \in \mathbb{C} \setminus \{-1\}.$$

The fact that $C(z)$ maps Hermitian matrices into unitary ones has been known for a long time.[37] More recently, the observation that one can switch between low-rank perturbations of these structures has been exploited for develop fast algorithms for unitary-plus-low-rank and Hermitian-plus-low-rank matrices.[38,39]

Lemma 5. *Let A be an $n \times n$ matrix. Then, we have the following.*

- *If A does not have the eigenvalue $-i$ and A is a rank k perturbation of a Hermitian matrix, then $C(A)$ will be a rank k perturbation of a unitary matrix. Moreover, $C(A)$ does not possess the eigenvalue 1.*
- *If A does not have the eigenvalue 1 and is a rank k perturbation of a unitary matrix, then $C^{-1}(A)$ will be a rank k perturbation of a Hermitian matrix, and $C^{-1}(A)$ does not possess eigenvalue $-i$.*

Proof. We show that the Cayley transform (and its inverse) preserve the rank of the perturbation. Note that both $C(z)$ and $C^{-1}(z)$ are degree $(1, 1)$ rational functions. For a rational function $r(z)$ of degree (at most) (d, d), we know that, for any matrix A and rank k perturbation E, $r(A + E) - r(A)$ has rank at most dk. It remains to prove that perturbation stays *exactly* of rank k, and not less. Let A be Hermitian plus rank (exactly) k, and by contradiction, assume we can write $C(A) = Q + E$, with Q unitary and $\text{rank}(E) = k' < k$. Then, we would have that $C^{-1}(Q + E) = A$ is a Hermitian plus rank k'' matrix where $k'' \leq k'$, leading to a contradiction. To prove that $C(A)$ does not have 1 in the spectrum, it suffices to note that $C(z)$ is a bijection of the Riemann sphere, and maps the point at ∞ to 1. Because the eigenvalues of $C(A)$ are $C(\lambda)$, with λ the eigenvalues of A, we see the eigenvalue 1 must be excluded. The same argument applies for the second case. □

Creating this bridge between low-rank perturbations of unitary and Hermitian matrices enables to use the criterion that we have developed for detecting matrices in $\mathcal{H}_k$ to matrices in $\mathcal{U}_k$, and the opposite direction as well. In fact, in the next lemma, we show that we can obtain alternative proofs for the characterizations of $\mathcal{U}_k$ and $\mathcal{H}_k$ by simply applying the Cayley transform.

Lemma 6. *The Cayley transformation implies that Theorem 3 and Theorem 6 are equivalent.*

Proof. We start by proving that Theorem 6 implies Theorem 3. Let A be an arbitrary matrix, and assume that 1 is not an eigenvalue. We know by Lemma 5 that A is in $\mathcal{U}_k$ if and only if $C^{-1}(A) \in \mathcal{H}_k$. Now, $C^{-1}(A)$ will be a rank k perturbation of a Hermitian matrix if and only if the Hermitian matrix $\frac{1}{2i}(C^{-1}(A) - C^{-1}(A)^*)$ has at most k positive eigenvalues and k negative ones. We can write

$$\frac{1}{2i}\left(C^{-1}(A) - C^{-1}(A)^*\right) = \frac{1}{2} \cdot \left[(A + I)^{-1}(A - I) + (A^* + I)^{-1}(A^* - I)\right].$$

Let us do a congruence by left-multiplying by $(A + I)$ and right multiplying by $(A^* + I)$. This does not change the sign characteristic and yields

$$\begin{aligned}\frac{1}{2i} \cdot (A + I)\left[C^{-1}(A) - C^{-1}(A)^*\right](A^* + I) &= \frac{1}{2}\left[(A - I)(A^* + I) + (A + I)(A^* - I)\right] \\ &= AA^* - I.\end{aligned}$$

The matrix above has eigenvalues $\lambda_i := \sigma_i^2(A) - 1$, where $\sigma_i(A)$ are the singular values of A. Therefore, the positive eigenvalues of $\frac{1}{2i}\left(C^{-1}(A) - C^{-1}(A)^*\right)$ correspond to singular values of A larger than 1, whereas the negative eigenvalues link to the singular values smaller than 1. In particular, A is unitary plus rank k, if and only if the characterization given in Theorem 3 is satisfied. Let us now consider the case where A has 1 as an eigenvalue. Then, we can multiply it by a unimodular scalar ξ to get $A' = \xi \cdot A$, where A' does not have 1 as an eigenvalue. Clearly $A' \in \mathcal{U}_k \iff A \in \mathcal{U}_k$, and $\sigma_i(A') = \sigma_i(A)$. Applying the previous steps to A' yields the characterization for A as well, completing the proof. The other implication (that is, Theorem 3 implies Theorem 6) can be obtained following the same steps backwards. □

Remark 2. We emphasize that, although in principle the Cayley transform enables to study unitary matrices looking at Hermitian ones (and the other way around), it cannot be used to answer questions about the closest unitary or Hermitian matrix. In fact, this transformation does not preserve the distances.

8 | CONSTRUCTION OF THE REPRESENTATIONS

We present in this section a proof-of-concept algorithm to show how one can use the results in this paper to construct, given a matrix $A \in \mathcal{H}_k$ (respectively $\mathcal{U}_k$), matrices $G, B \in \mathbb{C}^{n\times \ell}$, with $\ell \leq k$, such that $A - GB^*$ is Hermitian (respectively unitary). The procedure identifies the minimum possible rank ℓ automatically given the matrix A only, and requires $\mathcal{O}(n^2\ell)$ flops for a full matrix.

8.1 | Constructing a representation $A = H + GB^*$

If $A = H + GB^*$, with $H = H^*$, the Hermitian matrix $S(A) = \frac{1}{2i}(A - A^*)$ has rank $k_+ + k_- \leq 2k$ (where k_+, k_- are as in the proof of Theorem 6), hence the Lanczos algorithm (with a random starting vector b) applied to $S(A)$ will break down after at most $k_+ + k_-$ steps, in exact arithmetic, giving an approximation $S(A) = \frac{1}{2i}(A - A^*) \approx WTW^*$.

We apply to T the procedure described in the proof of Theorem 6, which is fully constructive, to recover a decomposition $T \approx \hat{B}\hat{C}^* + \hat{C}\hat{B}^*$, neglecting the eigenvalues smaller than a prescribed truncation threshold. This implies that we can write $\frac{1}{2i}(A-A^*) \approx BC^* + CB^*$, where $C := W\hat{C}, B := W\hat{B}$. Then, we construct the final decomposition by setting $H = A - 2iCB^*$. The procedure is sketched in the pseudocode of Algorithm 1.

Algorithm 1 Lanczos-based scheme to recover the Hermitian-plus-low-rank decomposition. A truncation threshold ε is given

```
**Procedure** HK_FIND($A, \varepsilon$)
    $S \leftarrow \frac{1}{2i}(A - A^*)$
    $W, T \leftarrow$ LANCZOS($S, \varepsilon$)                          ▷ $S = WTW^* + \mathcal{O}(\varepsilon)$
    $\hat{B}, \hat{C} \leftarrow$ COMPUTE FACTORS($T, \varepsilon$)      ▷ $T = \hat{B}\hat{C}^* + \hat{C}\hat{B}^* + \mathcal{O}(\varepsilon)$, using Lemma 2
    $B, C \leftarrow W\hat{B}, W\hat{C}$
    $G \leftarrow 2iC$
    $H \leftarrow A - GB^*$
    **Return** $\frac{1}{2}(H + H^*), G, B.$
**End procedure**
```

Note that in the unlikely case in which the process terminates early, WTW^* is not equal to S, but only to its restriction on the maximal Krylov subspace. Hence, when we detect that $WTW^* - S$ is still too large, we can continue the Lanczos iteration but with a randomly chosen vector.

Remark 3. A reconstruction procedure can be obtained from any method to approximate the range of S in $O(n^2k)$ flops. Indeed, once one obtains an orthogonal basis W for ImS, one can compute $W^*SW = T$ and continue as above.

8.2 | Constructing a representation $A = Q + GB^*$

The case of a unitary-plus-rank-k matrix can be solved with the same ideas, relying on the Golub–Kahan bidiagonalization.

If $A \in \mathcal{U}_k$, then Theorem 3 shows that the matrices A^*A and AA^* have (at most) $m := k_+ + k_- + 1$ distinct eigenvalues: k_+ greater than 1, k_- smaller than 1, and the eigenvalue 1, possibly with high multiplicity. Hence, the Lanczos process on each of them breaks down after at most m steps. Indeed, if $v_1, v_2, \ldots, v_n$ is an eigenvector basis for A^*A, ordered so that $v_m, v_{m+1}, \ldots, v_n$ are eigenvectors with eigenvalue 1, then we can write the initial vector for the Lanczos process as $b = \alpha_1 v_1 + \ldots + \alpha_{m-1}v_{m-1} + w$ where $w = \alpha_m v_m + \alpha_{m+1}v_{m+1} + \ldots + \alpha_n v_n$; the vector w is also an eigenvector w with eigenvalue 1. Thus b belongs to the invariant subspace span$(v_1, v_2, \ldots, v_{m-1}, w)$, which is also (generically) its maximal Krylov subspace.

It is well established that the Golub–Kahan bidiagonalization is equivalent to running the Lanczos process to A^*A and AA^*; hence, it will also break down after m steps (generically, when there is no earlier breakdown), returning matrices U_1, M, V_1 such that the decomposition

$$A \approx \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} M & \\ & I \end{bmatrix} \begin{bmatrix} V_1 & V_2 \end{bmatrix}^*$$

holds, with $M \in \mathbb{C}^{m\times m}$ upper bidiagonal, and $\begin{bmatrix} U_1 & U_2 \end{bmatrix}$ and $\begin{bmatrix} V_1 & V_2 \end{bmatrix}$ orthogonal.

We can use the constructive argument in the proof of Theorem 3 to decompose $M \approx Q + \hat{G}\hat{B}^*$, neglecting the singular values such that $|\sigma_i - 1|$ is below a certain truncation threshold. Then, $G = U_1\hat{G}, B = V_1\hat{B}$ give the required decomposition. The procedure is sketched in the pseudocode of Algorithm 2.

Algorithm 2 Golub–Kahan-based scheme to recover the unitary-plus-low-rank decomposition. A truncation threshold ε is given

```
Procedure UK_FIND($A, \varepsilon$)
    $U_1, M, V_1 \leftarrow$ GOLUB–KAHAN($A, \varepsilon$)        ▷ $A = [U_1, U_2]\,\mathrm{diag}(M, I)[V_1, V_2]^* + O(\varepsilon)$
    $\hat{G}, \hat{B} \leftarrow$ COMPUTE FACTORS($M, \varepsilon$)        ▷ $M = Q + \hat{G}\hat{B}^* + \mathcal{O}(\varepsilon)$ using Lemma 1
    $G, B \leftarrow U_1\hat{G}, V_1\hat{B}$
    $Q \leftarrow A - GB^*$
    Return $Q, G, B$.
End procedure
```

9 | NUMERICAL EXPERIMENTS

9.1 | Accuracy of the reconstruction procedure

We have implemented Algorithm 1 and Algorithm 2 in Matlab, and ran some tests with randomly generated matrices to validate the procedures. The algorithms appear to be quite robust and succeeded in all our tests in retrieving a decomposition with the correct rank and a small error. In each test, we generated a random $n \times n$ Hermitian or unitary plus rank-k matrix. For the Hermitian case, this is achieved with the Matlab commands

```
rng('default');
H = randn(n, n) + 1i*randn(n, n);
H = H + H';
[U,~] = qr(randn(n, k) + 1i*randn(n, k));
[V,~] = qr(randn(n, k) + 1i*randn(n, k));
A = H + U*diag(sv)*V';
```

Here, sv are logarithmically distributed singular values between 1 and a parameter σ. In the unitary case, the matrix H is replaced by the Q factor of a QR factorization of a random $n \times n$ matrix with entries distributed as $N(0, 1)$, that is, `[Q,~] = qr(randn(n))`.

We ran experiments with varying values of n, k, and σ; we report the convergence history by plotting the values of the subdiagonal entries obtained in the Lanczos process (in the Hermitian case), or in the Golub–Kahan bidiagonalization (for the unitary case). Experiments revealed that the parameter n has no visible effect on the convergence or the accuracy, so we only plotted the behavior with different values of k and σ, which are reported in Figure 1. The graphs show that there is a sharp drop in their magnitude after $2k$ steps, which is what is expected because $rank(S) = 2k$, generically.

The magnitude of the subdiagonal entries $T_{i+1,i}$ or superdiagonal ones $M_{i,i+1}$ reflects the decay in the singular values in the rank correction.

For the Hermitian case, the relative residual $\|\frac{1}{2}(H + H^*) + GB^* - A\|_2/\|A\|_2$ has been verified to be around $6 \cdot 10^{-17}$ in all the tests. Note that there is a difference in Algorithm 1 and Algorithm 2: in the former, the matrix H is guaranteed to be Hermitian, because it is symmetrized explicitly at the end of the algorithm, so it makes sense to measure the reconstruction error as $\|\frac{1}{2}(H + H^*) + GB^* - A\|_2/\|A\|_2$. In the latter, the unitary factor Q is obtained by the difference $Q = A - GB^*$: hence, we expect $\|Q + GB^* - A\|_2/\|A_2\|$ to be of the order of the machine precision (the error is given just by the subtraction), but Q will only be approximately unitary. Therefore, in this case we measure the error as $\max_j |\sigma_j(Q) - 1|$, which is the distance of Q to a unitary matrix in the Euclidean norm. In the experiments reported in Figure 1, the errors are enclosed in $[3u, 4u]$, where u is the machine precision $u \approx 2.22 \cdot 10^{-16}$.

9.2 | Rank structure in linearizations

A classical example of matrices with Hermitian or unitary plus low-rank structure are linearizations of polynomials and matrix polynomials. The most well-known linearization is probably the classical Frobenius companion matrix, obtained

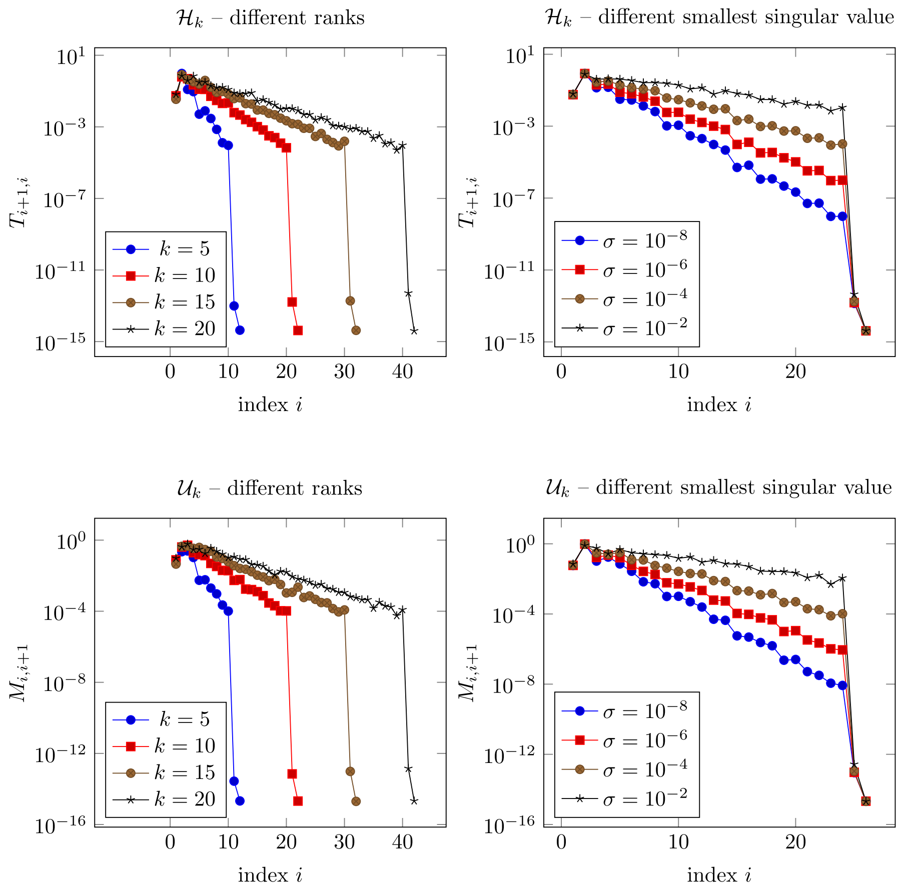

FIGURE 1 (Top) Magnitude of the subdiagonal entries $T_{i+1,i}$ (in the Lanczos process for the Hermitian case), or (bottom) the superdiagonal ones $M_{i,i+1}$ (in the Golub–Kahan bidiagonalization scheme), computed in the recovery of the approximate factors H, G, B in a Hermitian plus-low-rank matrix $A = H + GB^*$, or Q, G, B of the unitary plus-low-rank matrix $A = Q + GB^*$. The tests have been performed for different smallest singular values σ of GB^* and different ranks k

in MATLAB by the command `compan(p)`, where `p` is a vector with the coefficients of a polynomial. For many linearizations, one can find, by direct inspection, a value k such that they are $\mathcal{H}_k$ or $\mathcal{U}_k$; for instance the Frobenius companion form is seen to be $\mathcal{U}_1$, as one can turn it into a cyclic shift matrix by changing only the top row. In this section, we use some of these examples, to test the recovery procedure, and to confirm that the computed values k are optimal.

9.2.1 | The Fiedler pentadiagonal linearization

A classical variant of the scalar companion matrix is the pentadiagonal Fiedler matrix obtained by permuting the elementary Fiedler factors according to the odd–even permutation. Given a monic polynomial of degree n, $p(x) = x^n - \sum_{i=0}^{n-1} p_i x^i$, define

$$F_0 := \begin{bmatrix} p_0 & \\ & I_{n-1} \end{bmatrix},$$

$$F_i := \begin{bmatrix} I_{i-1} & & \\ & G(p_i) & \\ & & I_{n-i-1} \end{bmatrix}, \quad G(p_i) := \begin{bmatrix} 0 & 1 \\ 1 & p_i \end{bmatrix}, \quad i = 1, \dots, n-1.$$

Then, the matrix $F = F_1 F_3 \dots F_{2\lfloor \frac{n}{2} \rfloor - 1} F_0 F_2 \dots F_{2\lfloor \frac{n-1}{2} \rfloor}$ is a linearization of $p(x)$ and has the pentadiagonal structure depicted in Figure 2. From the other works,[7,8] we know that these matrices are unitary-plus-rank-k with k at most $\lceil \frac{n}{2} \rceil$. In particular, one can observe that

$$F = Q + GB^*, \tag{8}$$

where Q is an orthogonal matrix (depicted in red in Figure 2) whose nonzeros are precisely in position $(2;1)$, $(n-1;n)$ and those of the form $(2j-1;2j+1)$ and $(2j;2j+2)$ for $j \geq 1$, it is the matrix constructed analogously to F but starting from the polynomial $p(x) = x^n - 1$. Indeed, the nonzero entries of $F - Q$ (depicted in blue in Figure 2, plus an additional one in $(2;1)$) appear only in its odd-indexed columns, plus the last one, hence, the rank of this correction is clearly bounded by $\lceil \frac{n}{2} \rceil$.

Applying Algorithm 2 to a pentadiagonal Fiedler of size $n = 512$ generated from a monic random scalar polynomial, we obtain the subdiagonal entries $T_{i+1,i}$ plotted in Figure 3. As expected, we get that the first subdiagonal entry below the

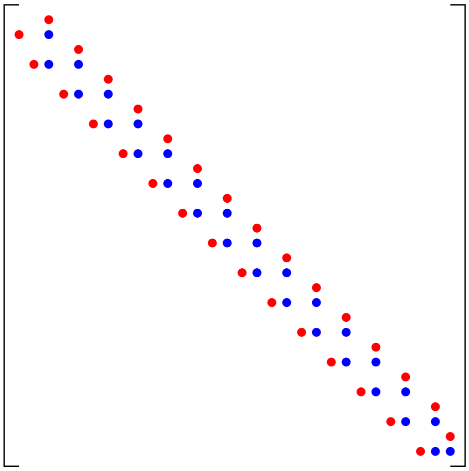

FIGURE 2 Nonzero pattern of a Fiedler pentadiagonal matrix of size 30. The entries displayed in red are all 1 except for $F_{2,1} = p_0$. In particular, when $p(x) = x^n - 1$, the matrix F is a (unitary) permutation matrix that has ±1 in the entries in red and 0 elsewhere

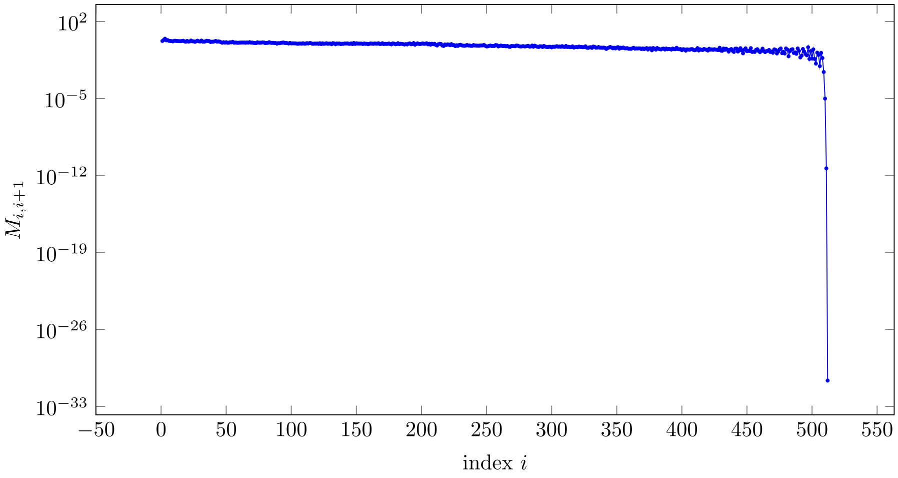

FIGURE 3 Superdiagonal entries obtained from applying the Golub–Kahan bidiagonalization scheme to the pentadiagonal Fiedler matrix of a random polynomial with $n = 513$

threshold $\epsilon = 1.0e - 14$ is $T_{513,512}$, so that $m = 512$ and $k_+ = k_- = 256$. The algorithm computes correctly a representation $F = \hat{Q} + \hat{G}\hat{B}^*$ with $\hat{G}, \hat{B} \in \mathbb{C}^{512\times 256}$. Thus, the experiments reveal that the bound $k \leq \lceil \frac{n}{2} \rceil$ is tight. However, the computed unitary term $\hat{Q}$ does not coincide with the one obtained theoretically in (8). This fact should not be surprising, because the representation is not necessarily unique.

9.2.2 | The colleague linearization

Consider an $m \times m$ matrix polynomial $P(\lambda)$ expressed in the Chebyshev basis

$$P(\lambda) = P_0T_0(\lambda) + P_1T_1(\lambda) + \ \dots\ + P_dT_d(\lambda), \qquad T_{j+1}(\lambda) = 2\lambda T_j(\lambda) - T_{j-1}(\lambda),$$

and $T_0(\lambda) = 1, T_1(\lambda) = \lambda$. Such a polynomial can be linearized using the *colleague matrix*

$$C = \begin{bmatrix} P_1 & P_2 & P_3 & \dots & P_d \\ \frac{1}{2}I & & \frac{1}{2}I & & \\ & \ddots & & \ddots & \\ & & \frac{1}{2}I & & \frac{1}{2}I \\ & & & I & \end{bmatrix} \in \mathbb{R}^{md\times md},$$

partitioned into blocks of size $m \times m$ each. It is simple to see that this matrix belongs to $\mathcal{H}_{2m}$, as it is sufficient to modify the first and last block row to obtain the Hermitian matrix $\text{tridiag}(\frac{1}{2}, 0, \frac{1}{2}) \otimes I_m$. However, $D^{-1}CD \in \mathcal{H}_m$ for a suitable diagonal scaling matrix D, so one may wonder if the rank $2m$ of this correction is optimal or if it can be lowered with a more clever construction.

We apply Algorithm 1 to a large-scale (sparse) matrix C, with $d = m = 100$, in which the first block row contains random coefficients (obtained with `randn(m)`). We plot the subdiagonal entries $T_{i+1,i}$ obtained in Figure 4. The first negligible

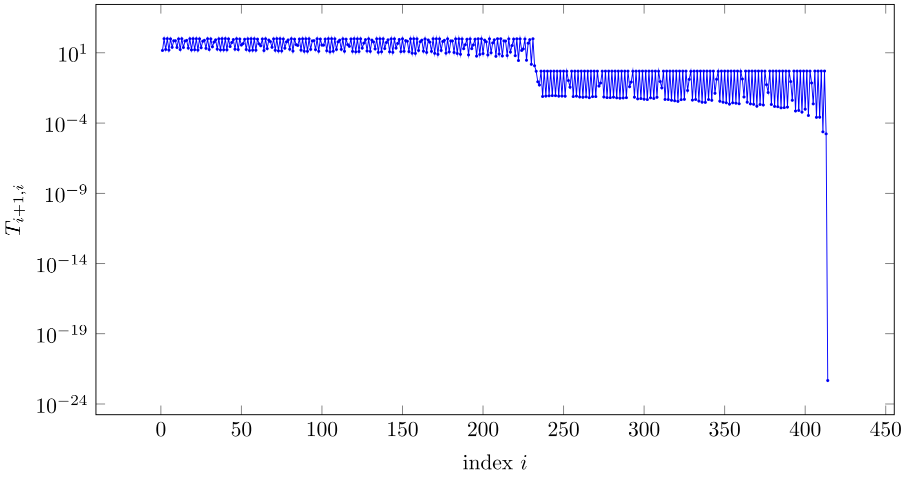

FIGURE 4 Subdiagonal entries obtained for the colleague matrix of a random polynomial with $d = 100, m = 100$

subdiagonal entry is $T_{415,414}$, giving an invariant subspace of dimension 414. Continuing the computation on T reveals only 400 nonzero eigenvalues, and returns a representation $C = A + UV \in \mathcal{H}_{200}$. Hence, our experiment confirms that in this example the minimum rank of the correction is $2m = 200$.

9.3 | Structure loss in computing the Schur form

Given a companion matrix C, in its Schur form $C = UTU^*$ the upper triangular factor T is unitary-plus-rank-1, in exact arithmetic (because C is so). As described in the introduction, several numerical methods in the literature try to exploit this structure, using special representations to enforce exactly the unitary plus rank 1 structure. If instead an approximation $\tilde{T}$ is computed using the standard QR algorithm (Matlab's `schur(C)`), can we measure the loss of structure in $\tilde{T}$, that is, the distance between $\tilde{T}$ and the closest matrix, which is unitary-plus-rank-1?

We have run some experiments in which this distance is computed using the formula in Theorem 4, in two different cases:

- The companion matrix of a polynomial whose roots are random numbers generated from a normal distribution with mean 0 and variance 1, that is, Matlab's `compan(poly(randn(n, 1)));`
- The companion matrix of Wilkinson's polynomial, that is, the polynomial with roots $1, 2, \dots, n$.

The singular values of $\tilde{T}$ have been computed using extended precision arithmetic, to get a more accurate result. Figure 5 displays the (relative) distance from structure

$$\frac{\|\tilde{T} - X\|_2}{\|\tilde{T}\|_2}, \quad X = \arg\min_{X \in \mathcal{U}_k} \|\tilde{T} - X\|_2$$

for different matrix sizes n. This distance is always within a moderate multiple of the machine precision, which is to be expected because the Schur form is computed with a backward stable algorithm. It appears that the loss of structure is less pronounced for the Wilkinson polynomial; note, though, that $\|T\| = \|C\|$ is much larger (for $n = 60$, $\|C\| \approx 2 \times 10^{83}$ for the Wilkinson polynomial vs. $\|C\| \approx 3 \times 10^{7}$ for the random polynomial).

Another interesting quantity is

$$\frac{\|\tilde{T} - X\|_2}{\|\tilde{T} - T\|_2}, \quad X = \arg\min_{X \in \mathcal{U}_k} \|\tilde{T} - X\|_2,$$

that is, the relative amount (measured as a fraction in $[0, 1]$) of the total error on $\tilde{T}$ that can be attributed to the loss of structure. If this ratio is close to 0, then it means that the main effect of the error is perturbing T to another matrix *inside* $\mathcal{U}_k$, whereas if it is close to 1, then its main effect is perturbing it to a matrix *outside* $\mathcal{U}_k$; hence, it would make sense to consider a projection procedure to map it back to $\mathcal{U}_k$.

We display this quantity in Figure 6. It is again smaller for the Wilkinson polynomial, and in both cases, it seems to decrease slowly as the dimension n increases.

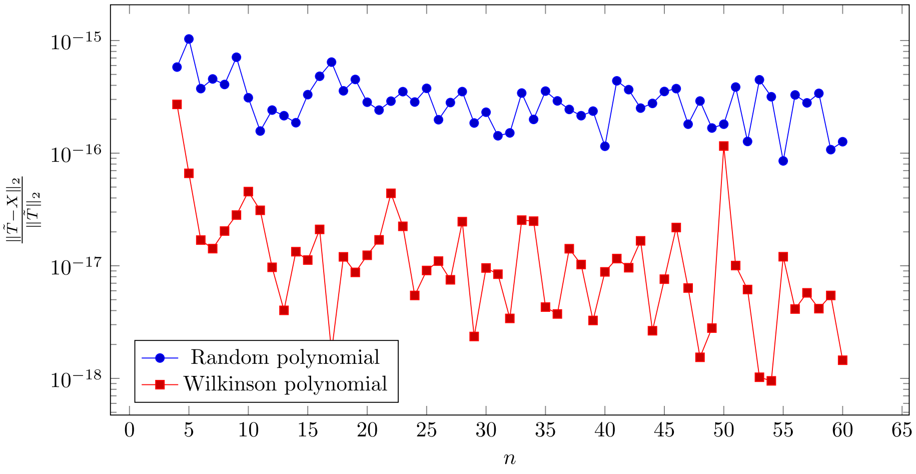

FIGURE 5 Relative distance from the structure

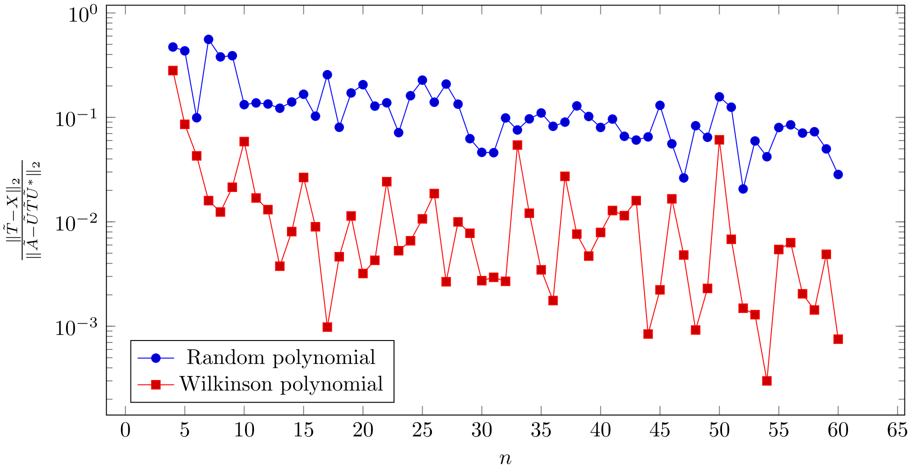

FIGURE 6 Ratio of error due to structure loss

10 | CONCLUSIONS

We have provided explicit conditions under which a matrix is unitary (respectively Hermitian) plus low rank, and have given a construction for the closest unitary (respectively Hermitian) plus rank k to a given matrix A, in both the spectral and the Frobenius norm.

We have presented an algorithm based on the Lanczos iteration to construct explicitly, given a matrix $A \in \mathcal{H}_k$, (where k is not known a priori), a representation of the form $A = H + GB^*$, where H is Hermitian and GB^* is a rank-k correction. A variant for unitary-plus-low-rank matrices based on the Golub–Kahan bidiagonalization scheme has been presented as well. We tested these two algorithms on various examples coming from applications in numerical linear algebra, including a large-scale example.

ACKNOWLEDGEMENTS

The research of the first two authors was partially supported by the INdAM-GNCS project “Analisi di matrici sparse e data-sparse: metodi numerici ed applicazioni” and by the University of Pisa under grant PRA-2017-05. The research of the third author was partially supported by the INdAM-GNCS project “Metodi di proiezione per equazioni di matrici e sistemi lineari con operatori definiti tramite somme di prodotti di Kronecker, e soluzioni con struttura di rango.” The research of the fourth author was supported by the Research Council KU Leuven, project C14/16/056 (Inverse-free Rational Krylov Methods: Theory and Applications).

ORCID

Gianna M. Del Corso https://orcid.org/0000-0002-5651-9368
Leonardo Robol https://orcid.org/0000-0002-6545-1748

REFERENCES

1. Parlett BN. The symmetric eigenvalue problem. Philadelphia, PA: SIAM; 1998.
2. Saad Y. Iterative methods for sparse linear systems. 2nd ed. Philadelphia, PA: SIAM; 2003.
3. Aurentz JL, Mach T, Vandebril R, Watkins DS. Fast and stable unitary QR algorithm. Electron Trans Numer Anal. 2015;44:327–341.
4. Bunse-Gerstner A, Elsner L. Schur parameter pencils for the solution of the unitary eigenproblem. Linear Algebra Appl. 1991;154–156:741–778.
5. Gragg WB. The QR algorithm for unitary Hessenberg matrices. J Computat Appl Math. 1986;16:1–8.
6. Bini DA, Robol L. On a class of matrix pencils and ℓ-ifications equivalent to a given matrix polynomial. Linear Algebra Appl. 2016;502:275–298.
7. De Terán F, Dopico FM, Pérez J. Condition numbers for inversion of Fiedler companion matrices. Linear Algebra Appl. 2013;439(4):944–981.
8. Del Corso GM, Poloni F, Robol L, Vandebril R. Factoring block Fiedler companion matrices. In: Structured Matrices in Numerical Linear Algebra: Analysis, Algorithms and Applications. Cham, Switzerland: Springer, 2019; p. 129–155.
9. Robol L. Exploiting rank structures for the numerical treatment of matrix polynomials [PhD thesis]. Pisa, Italy: Scuola Normale Superiore di Pisa; 2015.
10. Robol L, Vandebril R, Dooren PV. A framework for structured linearizations of matrix polynomials in various bases. SIAM J Matrix Anal Appl. 2017;38(1):188–216.
11. Chandrasekaran S, Gu M. Fast and stable eigendecomposition of symmetric banded plus semi-separable matrices. Linear Algebra Appl. 2000;313:107–114.
12. Delvaux S. Rank structured matrices [PhD thesis]. Leuven, Belgium: Department of Computer Science, Katholieke Universiteit Leuven; 2007.
13. Vandebril R, Del Corso GM. An implicit multishift *QR*-algorithm for Hermitian plus low rank matrices. SIAM J Sci Comput. 2010;32(4):2190–2212.
14. Eidelman Y, Gemignani L, Gohberg IC. Efficient eigenvalue computation for quasiseparable Hermitian matrices under low rank perturbation. Numerical Algorithms. 2008;47(3):253–273.
15. Barnett S. Polynomials and linear control systems. New York, NY: Marcel Dekker; 1983.
16. Vandebril R, Van Barel M, Mastronardi N. Matrix computations and semiseparable matrices: eigenvalue and singular value methods. Vol. II. Baltimore, MD: Johns Hopkins University Press; 2008.
17. Bini DA, Eidelman Y, Gemignani L, Gohberg I. Fast QR eigenvalue algorithms for Hessenberg matrices which are rank-one perturbations of unitary matrices. SIAM J Matrix Anal Appl. 2007;29(2):566–585.
18. Van Barel M, Vandebril R, Van Dooren P, Frederix K. Implicit double shift *QR*-algorithm for companion matrices. Numerische Mathematik. 2010;116(2):177–212.
19. Chandrasekaran S, Gu M, Xia J, Zhu J. A fast QR algorithm for companion matrices. Recent Adv Matrix Oper Theory. 2007;179:111–143.
20. Boito P, Eidelman Y, Gemignani L. Implicit QR for rank-structured matrix pencils. BIT Numer Math. 2013;54:85–111.
21. Bevilacqua R, Del Corso GM, Gemignani L. A CMV-based eigensolver for companion matrices. SIAM J Matrix Anal Appl. 2015;36(3):1046–1068.
22. Aurentz JL, Mach T, Robol L, Vandebril R, Watkins DS. Core-chasing algorithms for the eigenvalue problem. Philadelphia, PA: SIAM; 2018.
23. Aurentz JL, Mach T, Robol L, Vandebril R, Watkins DS. Fast and backward stable computation of the eigenvalues of matrix polynomials. Math Comput. 2019;88:313–347.
24. Bini DA, Robol L. Quasiseparable Hessenberg reduction of real diagonal plus low rank matrices and applications. Linear Algebra Appl. 2016;502:186–213.
25. Gemignani L, Robol L. Fast Hessenberg reduction of some rank structured matrices. SIAM J Matrix Anal Appl. 2017;38(2):574–598.
26. Bevilacqua R, Del Corso GM, Gemignani L. Fast QR iterations for unitary plus low rank matrices. arXiv:1810.02708. 2018.
27. Huckle T. The Arnoldi method for normal matrices. SIAM J Matrix Anal Appl. 1994;15(2):479–489.
28. Huhtanen M. A stratification of the set of normal matrices. SIAM J Matrix Anal Appl. 2001;23(2):349–367.
29. Barth TL, Manteuffel TA. Multiple recursion conjugate gradient algorithms part I: sufficient conditions. SIAM J Matrix Anal Appl. 2000;21(3):768–796.
30. Liesen J, Strakoš Z. On optimal short recurrences for generating orthogonal Krylov subspace bases. SIAM Review. 2008;50(3):485–503.
31. Liesen J. When is the adjoint of a matrix a low degree rational function in the matrix. SIAM J Matrix Anal Appl. 2007;29(4):1171–1180.
32. Beckermann B, Mertens C, Vandebril R. On an economic Arnoldi method for BML-matrices. SIAM J Matrix Anal Appl. 2018;39(2):737–768.
33. Beckermann B, Reichel L. The Arnoldi process and GMRES for nearly symmetric matrices. SIAM J Matrix Anal Appl. 2008;30(1):102–120.
34. Embree M, Sifuentes JA, Soodhalter KM, Szyld DB, Xue F. Short-term recurrence Krylov subspace methods for nearly Hermitian matrices. SIAM J Matrix Anal Appl. 2012;33(2):480–500.
35. Horn RA, Johnson CR. Matrix analysis. 2nd ed. Cambridge, UK: Cambridge University Press; 2013.
36. Thompson RC. Principal submatrices. IX: interlacing inequalities for singular values of submatrices. Linear Algebra Appl. 1972;5:1–12.
37. von Neumann J. Allgemeine eigenwerttheorie hermitescher funktionaloperatoren. Mathematische Annalen. 1930;102:49–131.

38. Aurentz JL, Mach T, Vandebril R, Watkins DS. Computing the eigenvalues of symmetric tridiagonal matrices via a Cayley transformation. Electron Trans Numer Anal. 2017;46:447–459.

39. Gemignani L. A unitary Hessenberg QR-based algorithm via semiseparable matrices. J Computat Appl Math. 2005;184:505–517.

How to cite this article: Del Corso GM, Poloni F, Robol L, Vandebril R. When is a matrix unitary or Hermitian plus low rank? *Numer Linear Algebra Appl*. 2019;e2266. https://doi.org/10.1002/nla.2266