Aspects of nonnormality for iterative methods

Marko Huhtanen

Abstract

Recently new optimal Krylov subspace methods have been discovered for normal matrices. In light of this, novel ways to quantify nonnormality are considered in connection with various families of matrices. We use as a criterion how, for a given matrix, these iterative methods introduced can be employed via, e.g., inexpensive matrix factorizations. The unitary orbit of the set of binormal matrices provides a natural extension of normal matrices. Its elements yield polynomially normal matrices of moderate degree. In this context several matrix nearness problems arise.

© 2004 Elsevier Inc. All rights reserved.

AMS classification: 65F10; 65F15

Keywords: Nonnormal matrix; Binormal matrix; Polynomially normal operator; Unitary orbit; Involution; Iterative methods; Ritz values; Measure of nonnormality

1. Introduction

Extending iterative methods, optimal in some sense, beyond Hermitian matrices is a challenging problem; see, e.g., [11, Chapter 6] for an informative discussion.
Recently there has been progress in this regard as two different optimal methods have been discovered for normal matrices. One relies on a 3-term recurrence [20,22] and the other on a recurrence with a slowly growing length [23,24]. In this paper we study various aspects of nonnormality arising from the existence of these algorithms. To this end we consider the set of binormal matrices [4] as well as its unitary orbit. These two families are then associated with polynomially normal matrices of moderate degree. Related matrix nearness problems are posed.

Binormal matrices possess a 2-by-2 block structure with commuting normal matrices as blocks. These matrices are typically far from being normal with respect to the classical measures of nonnormality [9]. In spite of this, there are ways to relate them with normal matrices. This we achieve by splitting a binormal matrix as the sum $S_1 + PS_2$, where $S_1$ and $S_2$ are commuting block diagonal normal matrices and $P$ is an involution satisfying a relaxed commutativity relation with $S_1$ and $S_2$. The arising structure is an algebra so that the inverse can be split analogously. This splitting allows us, generically, to employ algorithms for normal matrices for solving linear systems involving binormal matrices. Another aspect of binormality, and the sum representation proposed, is that all the matrices encountered can be regarded as being polynomially normal of moderate degree.

Polynomial normality is originally an infinite dimensional operator theoretic concept; see [27,28] and references therein. To adapt this to matrices, we define $A \in \mathbb{C}^{n \times n}$ to be polynomially normal of degree $d$ if there exists a monic polynomial $p$ of degree $d$ such that $p(A)$ is normal and $q(A)$ is not normal for any monic polynomial $q$ with $\deg(q) < d$. Modulo a constant term, $p$ is unique. In particular, involutions are polynomially normal matrices of degree two. Binormal matrices are polynomially normal of degree at most half of the dimension of the underlying space. The magnitude of $d$ is critical for our purposes, motivated by computations, because otherwise the concept is vacuous for matrices.

Since polynomial normality of particular degree remains invariant under unitary similarity transformations, we consider the unitary orbit of binormal matrices which we denote by $\mathcal{BN}$. This set, studied in a completely different context [36], provides a natural extension of normal matrices. It arises in connection with $\mathbb{R}$-linear operators in $\mathbb{C}^n$ [8]. Elements of $\mathcal{BN}$ have also appeared in illustrating various aspects of iterative methods [31,13]. They can be linked with the discussions in [38]. For a large scale engineering problem, see [32]. Besides bringing up these connections, we show that for these matrices polynomial normality is well understood.

Aside from being an interesting matrix analytical concept, polynomial normality yields a way to iteratively solve linear systems with methods for normal matrices. To this end, assume a polynomially normal matrix $A \in \mathbb{C}^{n \times n}$ of degree $d$ is factored as $A = Ns(A)^{-1}$ for a normal matrix $N$ and a polynomial $s$ of degree $d - 1$. In practice the computation of the inverse is never realized since solving a linear system $Ax = b$, for $b \in \mathbb{C}^n$, can be accomplished by solving

$$N x = s(A)b$$ (1.1)
instead. Hence algorithms for normal matrices can be employed with this system obtained. Stated in the context of polynomial preconditioning, we are concerned with finding a polynomial with the aim at having a normal matrix when evaluated in A.

For recent attempts to extend “commutative spectral theory” of normal matrices to nonnormal matrices, see [1,30,26] and references therein. In our approach having a normal \( p(A) \) for a monic polynomial \( p \) means that with \( p(A) \) we can employ methods for normal matrices for locating eigenvalues. Consequently, sparse matrix algorithms relying on real analytic techniques recently introduced in [24] become available. It remains to convert the information computed to concern \( A \). This is achieved with two simple applications of the spectral mapping theorem.

In view of these remarks, for iterative methods it seems to be somewhat unsatisfactory to measure nonnormality of \( A \) exclusively. Since any application of an iterative method involves polynomials in \( A \), it appears to be more natural to inspect the least nonnormality of the polynomial family

\[
\{ p(A) \}_{p \text{ monic}, \deg(p) \leq k} \tag{1.2}
\]

for a fixed \( k \ll n \). If \( A \) is already normal, then these matrices remain normal. If \( A \) is not normal but some \( p(A) \) is, then we can associate a particular Schur decomposition with \( A \) and give a qualitative description of a related matrix Krylov subspace. If there are no normal matrices among (1.2), then we ask how far is this family from the set of normal matrices. Another option is to try to find nearly normal matrices with polynomials in \( A \) by simultaneously employing small rank perturbations.

The paper is organized as follows. In Section 2 we introduce binormal matrices and compute their dimension. We demonstrate that a linear system involving a binormal matrix can be solved by executing an optimal 3-term recurrence for normal matrices. A naturally arising circulant structure is presented. In Section 3 we study the unitary orbit of binormal matrices and polynomially normal matrices of moderate degree after showing how iterative methods for normal matrices can be employed with such matrices. In Section 4 we group together related measures of nonnormality arising in this context. We illustrate how “almost normality” in our sense allows us to compute Ritz values with modest storage requirements. In Section 5 we consider numerical algorithms for computing the polynomials introduced.

2. Binormal matrices and a related circulant structure

How to benefit from optimal methods for normal matrices while dealing with large nonnormal problems? Since every square matrix is the product of two normal matrices, any linear system can be solved by solving two consecutive linear system involving normal matrices. Presently this is an impractical alternative since finding such a factorization, like the polar decomposition, is too expensive with the existing
techniques. Hence, it is of interest to identify matrix structures for which there exists inexpensive factorizations with normal factors, or nearly so.

To this end, assume \( \mathcal{S} \subset \mathbb{C}^{n \times n} \) is an algebra of matrices containing \( I \), the identity matrix. Let \( P \in \mathbb{C}^{n \times n} \) be an involution such that \( P \mathcal{S} = \mathcal{S} P \), i.e., for any \( S \in \mathcal{S} \) there exists \( \widehat{S} \in \mathcal{S} \) such that \( PS = \widehat{S} P \). Recall that \( P \) is an involution if we have \( P^2 = I \). Considering matrices that can be represented as the sum

\[
S_1 + PS_2,
\]

with \( S_1, S_2 \in \mathcal{S} \), \( (2.1) \) gives us an algebra because of the relaxed commutativity relation.

This structure is preserved under similarity transformations, i.e., if \( X \) is invertible, then take \( X S X^{-1} \) to replace \( S \) while \( XPX^{-1} \) inherits the role of \( P \).

**Example 1.** The commutant of a set \( \mathcal{S} \subset \mathbb{C}^{n \times n} \), denoted by \( \mathcal{S}' \), consists of matrices commuting with every element of \( \mathcal{S} \). Hence, any involution belonging to \( \mathcal{S}' \) satisfies the prescribed condition. In particular, \( P = I \) is the most trivial and, of course, the least interesting choice.

If the elements of \( \mathcal{S} \) can be easily inverted (or preconditioned), it can be beneficial to represent the inverse of \( S_1 + PS_2 \) in the form

\[
\tilde{S}_1 = (S_1 - PS_2 S_1^{-1} S_2) - 1 \quad \text{and} \quad \tilde{S}_2 = (PS_2 P - S_1 S_2^{-1} P S_1 P)^{-1}.
\]

Generically these inverses exist. In case \( \mathcal{S} \) is commutative, these simplify to

\[
\tilde{S}_1 = \hat{S}_1 (S_1 \hat{S}_1 - S_2 \hat{S}_2)^{-1} \quad \text{and} \quad \tilde{S}_2 = S_2 (S_2 \hat{S}_2 - S_1 \hat{S}_1)^{-1}
\]

so that \( (S_1 + PS_2)^{-1} = (\tilde{S}_1 - PS_2) (\hat{S}_1 \hat{S}_1 - S_2 \hat{S}_2)^{-1} \) and hence only a single matrix inversion of an element of \( \mathcal{S} \) is needed to find the inverse of \( S_1 + PS_2 \).

Assume \( \mathcal{S}_0 \subset \mathbb{C}^{n \times n} \) is a set of commuting normal matrices. Let \( \mathcal{S}' \subset \mathbb{C}^{2n \times 2n} \) consist of block-diagonal matrices with two \( n \)-by-\( n \) main diagonal blocks from \( \mathcal{S}_0 \).

Hence \( \mathcal{S} \) is commutative. Let \( P = \begin{bmatrix} 0 & I \\ I & 0 \end{bmatrix} \). Then with \( (2.1) \) we obtain the set of so-called binormal matrices introduced by Brown [4]. Admitting a closed form solution to the problem of finding a nearest normal approximant, they have received a lot of attention; see [33,3,34] and references therein. The following canonical form is useful in practice.

**Theorem 2.1** [4]. Any binormal matrix is unitarily similar to a block upper triangular binormal matrix.

A binormal matrix can be very far from being normal, like the nilpotent matrix

\[
\begin{bmatrix} 0 & I \\ 0 & 0 \end{bmatrix} \in \mathbb{C}^{2n \times 2n}
\]

illustrates. In spite of this, due to \( (2.2) \), solving a linear system involving a binormal matrix is equivalent to solving a linear system involving the arising normal matrix \( S_1 \hat{S}_1 - S_2 \hat{S}_2 \). For this purpose one can employ, e.g., the optimal 3-term recurrence proposed in [22]. Hence in this regard binormal matrices possess a useful and inexpensive factorization with normal matrices as factors.
The dimension of binormal matrices is as follows.

**Theorem 2.2.** The set of binormal matrices is a stratified submanifold of $\mathbb{C}^{2n \times 2n}$ with the stratum of maximal real dimension $n^2 + 7n$.

**Proof.** We employ techniques from [20] adapted to our setting. Namely, considering the 2-by-2 block structure, let $A = \begin{bmatrix} N_1 & N_2 \\ N_3 & N_4 \end{bmatrix} \in \mathbb{C}^{2n \times 2n}$ be any, not necessarily a binormal, matrix. Then $A$ is binormal if and only if $\{N_j\}_{j=1}^4$ satisfy

\[ N_j N_k - N_k N_j = 0 \quad \text{and} \quad N_j N_j^* - N_j^* N_j = 0 \quad (2.3) \]

for all $1 \leq j, k \leq 4$. As these are polynomial equations for the entries of $A$ separated into the real and imaginary parts, the set of binormal matrices admits a stratification [10].

Recall that $N = H + iK$, with $H = \frac{1}{2}(N + N^*)$ and $K = \frac{1}{2i}(N - N^*)$, is normal if and only if $H$ and $K$ commute; see, e.g., [14]. Moreover, assume $N_1 = H_1 + iK_1$ and $N_2 = H_2 + iK_2$ are commuting normal matrices. Then, by the Fuglede–Putnam–Rosenblum theorem [35], $N_1^*$ commutes with $N_2$ and $N_2^*$ commutes with $N_1$.

Therefore $[H_1, H_2, K_1, K_2]$ is a commuting family of Hermitian matrices. Hence we can consider, instead of $\{N_j\}_{j=1}^4$, the commuting family $[H_j, K_j]_{j=1}^4$ of Hermitian matrices (the converse is clearly true as well).

Denoting by $\mathcal{H} \subset \mathbb{C}^{n \times n}$ the set of Hermitian matrices, consider the direct product $\mathcal{H}^8$. Fix $H_1 \in \mathcal{H}$ with distinct eigenvalues and consider those Hermitian matrices $H_2, \ldots, H_8$ that commute with $H_1$. Then, since $H_1$ is nonderogatory, $H_2 = p_2(H_1), \ldots, H_8 = p_8(H_1)$ with polynomials $p_j$, for $2 \leq j \leq 8$ [19]. Being Hermitian matrices, each $p_j$ is of degree $n - 1$ at most with real coefficients. In particular, varying $H_1$ and the polynomials $p_j$, the matrices of the form

\[ \begin{bmatrix} H_1 + ip_1(H_1) & p_2(H_1) + ip_3(H_1) \\ p_4(H_1) + ip_5(H_1) & p_6(H_1) + ip_7(H_1) \end{bmatrix} \quad (2.4) \]

give rise to an open dense subset of the set of binormal matrices. Since the set of nonderogatory Hermitian matrices is of dimension $n^2$, this sums up to $n^2 + 7n$ free real parameters as claimed. □

At first sight $n^2 + 7n$ may not impress compared with $8n^2$, the real dimension of $\mathbb{C}^{2n \times 2n}$. However, such a comparison is not quite reasonable since most practical problems give rise to matrices with structure. For instance, the real dimension of the set of Toeplitz matrices is even of different magnitude, that is, $8n - 2$ in $\mathbb{C}^{2n \times 2n}$.

An adaptation of the methods proposed in [20] yields a way to generate binormal approximations to a given matrix with sparse matrix techniques. This amounts to taking a Hermitian matrix $H_1$ and forming (2.4) with polynomials $p_j$ with real coefficients, for $j = 1, \ldots, 7$. These polynomials can be generated inexpensively with a modification of the Hermitian Lanczos algorithm.
As a natural example related to binormality, consider again the structure (2.1) but now with \( \mathcal{S} \) being the set of circulant matrices. Let \( P \in \mathbb{C}^{n \times n} \) denote the “backward identity” [19], i.e., the permutation matrix with ones on the diagonal joining the left lower corner with the right upper corner. Then for any circulant matrix \( S_2 \) both \( PS_2 \) and \( S_2P \) are Hankel matrices with cyclically appearing antidiagonals so that \( P \mathcal{S} = \mathcal{S} P \). We call such matrices circulant-Hankel matrices.

**Theorem 2.3.** For \( n \) even, assume \( S_1, S_2 \in \mathbb{C}^{n \times n} \) are circulant matrices and \( P \) is the backward identity. Then \( S_1 + PS_2 \) is unitarily similar to a binormal matrix.

**Proof.** With the Fourier matrix \( F_n \in \mathbb{C}^{n \times n} \) (see, e.g., [7, p. 32]) the matrix

\[
F_n^* S_1 F_n \quad \text{is diagonal while} \quad F_n^* P S_2 F_n = \begin{bmatrix}
d_1 & 0 & \cdots & 0 \\
0 & 0 & \cdots & d_2 \\
0 & 0 & \cdots & 0 \\
0 & d_n & \cdots & 0
\end{bmatrix},
\]

i.e., only the \((1, 1)\)-element and the first sub-antidiagonal are possibly nonzero which follows after standard manipulations with the Fourier matrix. Performing a similarity transformation with a permutation in an obvious way keeps \( F_n^* S_1 F_n \) diagonal and transforms \( F_n^* P S_2 F_n \) into a 2-by-2 block diagonal matrix. \( \square \)

The proof also implies that we are dealing with an extension of circulant matrices preserving the algebraic structure in a natural way as follows.

**Corollary 2.4.** The set of matrices with the structure \( S_1 + PS_2 \) is a noncommutative \( \mathbb{C}^n \)-sub-algebra of \( \mathbb{C}^{n \times n} \) of complex dimension \( 2n - 1 \) if \( n \) is odd, and \( 2n - 2 \) if \( n \) is even.

**Proof.** We have an algebra which is closed in taking the adjoint since \( \mathcal{S} \) is the set of circulant matrices and \( P \) is Hermitian. One readily constructs an example to see that this is not a commutative algebra. The claim concerning the dimensions follows from a straightforward counting. \( \square \)

In appropriate dimensions this is an interesting algebra by the fact that for an invertible \( S_1 + PS_2 \) the inverse can be computed in \( O(n \log n) \) operations with the FFT techniques since for the similarity transformed matrix \( F_n^* (S_1 + PS_2) F_n \) the inverse can be found in \( O(n) \) operations due to its simple structure. Analogously, for the spectrum of such a matrix we have a closed form solution through solving \( \left\lfloor \frac{n}{2} \right\rfloor \) eigenvalue problems of size 2-by-2.

Unlike with classical structured matrices, by directly inspecting the entries of a given matrix \( A \in \mathbb{C}^{n \times n} \) one can not tell whether it is presentable as \( S_1 + PS_2 \). This can be found out by considering the matrix nearness problem.
This is readily solvable, for instance, in the Frobenius norm after performing a similarity transformation with $F_n$ to have

$$\min_{A_1, A_2 \text{ diagonal}} \| F_n^* A F_n - A_1 - A_2 \Gamma \|_F,$$

where $\Gamma = U P$ with $U$ being the unitary forward shift. Hence $\Gamma$ is the permutation matrix with ones on the first sub-antidiagonal and at the position $(1, 1)$; see [7, Eq. (2.4.20)] for $\Gamma$. With the FFT, finding $A_1$ and $A_2$ realizing (2.6) is an $O(n^2 \log n)$ computation. The resulting approximation can be used, e.g., in preconditioning linear systems involving $A$ analogously to the way circulant matrices are used in preconditioning [5].

3. Polynomial normality for matrices

Involution has the property that a low degree polynomial evaluated at them yields the identity, i.e., a normal matrix. This interpretation can be used for classifying nonnormality more generally.

**Definition 3.1.** $A \in \mathbb{C}^{n \times n}$ is polynomially normal of degree $d$ if $p(A)$ is normal for a monic polynomial $p$ of the least possible degree $d$. Then $p$ is called a minimal normal polynomial of $A$.

In an analogous way $A$ is said to be polynomially Hermitian of degree $d$ if $p(A)$ is Hermitian for a monic polynomial $p$ of the least possible degree $d$. These are unitarily invariant concepts both so that, as opposed to binormality, polynomial normality is not confined to any particular block structure.

Initially polynomial normality was introduced for analyzing infinite dimensional operators; see [27,28] where a typical problem was, e.g., to characterize operators which are polynomially normal. This type of questions are vacuous for matrices simply because every $A \in \mathbb{C}^{n \times n}$ is polynomially normal of degree $n$ at most (employ the characteristic polynomial of $A$). Instead, in finite dimensions the size of $d$ is of interest. This is illustrated in the first two subsections that follow.

3.1. Solving nonnormal linear systems

If the coefficient matrix $A \in \mathbb{C}^{n \times n}$ of a linear system is polynomially normal of moderate degree, then the problem can be solved with algorithms for normal matrices. To see this, suppose $p(A)$ is normal for a monic polynomial $p$ of degree $d > 1$. Since normality is a translation invariant property, we can assume that $p(z) = z^d - zq(z)$ for a polynomial $q$ of degree $d - 2$ at most. Modulo translations,
a minimal normal polynomial is clearly unique. Hence \( A(A^{d-1} - q(A)) = N \) for a normal matrix \( N \) and a unique polynomial \( q \).

Assuming \( s(A) = A^{d-1} - q(A) \) to be invertible, we obtain a factorization
\[
A = Ns(A)^{-1}
\]
of \( A \) which can be employed implicitly. More precisely, solving a linear system \( Ax = b \), for a vector \( b \in \mathbb{C}^n \), is equivalent to solving \( Ns(A)b \). Since now the coefficient matrix is normal, the problem can be solved with the 3-term recurrence for normal matrices \([22]\), being realistic provided \( d \) is not large. This factorization can also be viewed in the context of polynomial preconditioning with the relaxed aim at having a normal matrix instead of the identity.

**Example 2.** For an illustration, let \( A \) (see \([13,25]\)) be of the form \( ZAZ^{-1} \) with
\[
Z = \begin{bmatrix}
1 & \sqrt{1-\delta} & 0 & \cdots & 0 \\
0 & \sqrt{\delta} & 0 & \cdots & 0 \\
0 & 0 & 1 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \cdots & 1
\end{bmatrix}
\]
in such a way that, besides 20 and 10, the remaining eigenvalues of \( A \) are uniformly distributed in the interval \([1, 5]\). The factor in the minimal polynomial of \( A \) corresponding to the 2-by-2 block is \((z - 20)(z - 10) = z^2 - 30z + 200\) which also yields the minimal normal polynomial of \( A \). Namely, taking \( p(z) = z^2 - 30z = z(z - 30) \) gives us a Hermitian matrix \( p(A) = As(A) \).

In the preceding example the degree of the minimal normal polynomial was 2 regardless of the size of the matrix described. By taking \( p \) to be the factor in the characteristic polynomial corresponding to the \( d \)-by-\( d \) block, this has an obvious generalization as follows.

**Proposition 3.2.** Assume \( A \in \mathbb{C}^{n \times n} \) is unitarily similar to \( M \oplus A \), with \( M \in \mathbb{C}^{d \times d} \) and a diagonal matrix \( A \in \mathbb{C}^{(n-d) \times (n-d)} \). Then \( A \) is polynomially normal of degree \( d \) at most.

### 3.2. Locating eigenvalues of nonnormal matrices

The standard way of converting a matrix \( A \in \mathbb{C}^{n \times n} \) into a normal matrix is to form \( A^*A \), i.e., to “symmetrize” \( A \). In this operation spectral information is lost since the eigenvalues of \( A \) and \( A^*A \) are generally not related in any reasonable way. For this, see \([37, \text{Section } 4]\). Polynomial normality preserves data better since by knowing the spectrum \( \sigma(A) \) of \( A \), the eigenvalues of \( p(A) \) becomes available. For the converse, the following two propositions are also direct consequences of the spectral mapping theorem.
Proposition 3.3. Let \( p(A) = M \in \mathbb{C}^{n \times n} \) for a polynomial \( p \) and assume \( \sigma(M) \subset \Omega \). Then \( \sigma(A) \subset \{ z \in \mathbb{C} : p(z) \in \Omega \} \).

In particular, if \( p(A) \) is normal, then the algorithms proposed in [24] can be employed for generating sets \( \Omega \) containing its spectrum. What then remains is to find the inverse image of \( \Omega \) to get an exclusion region for the eigenvalues of \( A \). This latter task is a simple one as opposed to solving large nonnormal eigenvalue problems.

Example 3. We consider \( A \) of Example 2, that is, \( N = p(A) \) is Hermitian with \( p(z) = z(z - 30) \). The extreme eigenvalues of \( N \) can be computed fast with the Hermitian Lanczos method. Thus, assume knowing that the spectrum of \( N \) belongs to the interval \([-200, -29]\) on the real axis. Finding the inverse image of this interval for \( p \) is straightforward; it consists of the intervals \([1, 10]\) and \([20, 29]\) on the real axis. Both of these intervals contain eigenvalues of \( A \).

Since iterative methods are often aimed at finding just a few eigenvalues of very large matrices, the following is useful.

Proposition 3.4. Let \( p(A) = M \in \mathbb{C}^{n \times n} \) for a polynomial \( p \) and assume \( \lambda \in \sigma(M) \). Then the set \( \{ z \in \mathbb{C} : p(z) = \lambda \} \) contains an eigenvalue of \( A \).

This yields us a circuitous way to use real analytic computational techniques for finding approximations to eigenvalues of a nonnormal matrix. Assuming \( p(A) \) to be normal, the idea is to generate Ritz values for \( p(A) \) with the methods proposed in [20,24] and then to find their inverse image with respect to \( p \).

Example 4. Consider the matrix of Example 2 again. To illustrate Proposition 3.4, assume having computed the rightmost eigenvalue \( \lambda_n = -29 \) of \( N \) with, e.g., the Hermitian Lanczos method. Then solving \( p(z) = z(z - 30) = -29 \) gives \( z = 29 \) and \( z = 1 \), the latter being an eigenvalue of \( A \).

Hence Proposition 3.4 can give us “shadow” eigenvalues. Their number depends on the degree of \( p \) such that the smaller its degree the fewer of them occur. For solving the arising polynomial equation accurately, the degree of \( p \) should be moderate.

There are other approaches to partially conserve the commutative spectral theory of normal matrices with nonnormal matrices. Our considerations can be related to certain hereditary classes of matrices; see [1,30,26]. We briefly describe the connection as follows. If \( A \in \mathbb{C}^{n \times n} \) is such that \( p(A) \) is Hermitian, then

\[
p(A) - p(A)^* = \sum_{0 \leq k,l \leq d} c_{k,l} A^k A^l = 0,
\]

(3.2)
for some $c_{k,l} \in \mathbb{C}$, with $c_{d,0} = c_{0,d} = 1$. In this case most of these coefficients equal zero. Because multiplications by $A^*$ precede multiplications by $A$, the matrix in question can be regarded as to belong to the hereditary class corresponding to $p$.

3.3. Polynomially normal matrices of low degree and the unitary orbit of binormal matrices

There are matrices for which the characteristic polynomial coincides with the minimal normal polynomial. In particular, its degree can equal the dimension of the underlying space.

**Example 5.** Let $A \in \mathbb{C}^{n \times n}$ be the nilpotent backward shift, that is, the matrix has ones on the first super-diagonal while the other elements are zero. Then any $p(A)$, for a monic polynomial $p$ of degree $d \leq n - 1$, has ones on the $d$th super-diagonal. Also, being upper triangular, $p(A)$ is already Schur decomposed and, consequently, $p(z) = z^n$ is the minimal normal polynomial of $A$.

In spite of this, let us first make a few remarks on the very low degree cases.

If the product of two Hermitian matrices is Hermitian, then the matrices must commute. This does not hold for normal matrices (take, for instance, two unitary matrices). First degree polynomial normality yields a sufficient condition as follows.

**Theorem 3.5.** For $A, B \in \mathbb{C}^{n \times n}$ assume $p(A)q(B)$ is normal for any polynomials $p$ and $q$ of degree at most 1. Then $A$ and $B$ are commuting normal matrices.

**Proof.** With obvious choices for $p$ and $q$ we can deduce that $A$ and $B$ are normal. For commutativity, take $p(z) = z + \alpha I$ and $q(z) = z + \beta I$ with $\alpha, \beta \in \mathbb{C}$. Then

$$p(A)p(B)(p(A)p(B))^* - (p(A)p(B))^*p(A)p(B) = \overline{\alpha} \beta (AB^* - B^*A) + \alpha \overline{\beta} (BA^* - A^*B) + M_1(\alpha) + M_2(\beta),$$

(3.3)

where $M_1$ and $M_2$ are first degree terms in the variables $\alpha$ and $\beta$, respectively. Hence, for (3.3) to be zero independently of $\alpha$ and $\beta$, necessarily $\overline{\alpha} \beta (AB^* - B^*A) + \alpha \overline{\beta} (BA^* - A^*B) = 0$. Choosing $\alpha$ and $\beta$ real and positive, it follows that the Hermitian part of $AB^* - B^*A$ equals zero. Choosing $\alpha = s$ and $\beta = is$ with $s$ real and positive, we can deduce that the skew-Hermitian part of $AB^* - B^*A$ is zero. Hence, $AB^* - B^*A = 0$. By the Putnam–Fuglede theorem, $AB = BA$. □

If $A$ is the square root of a normal matrix, like an involution, then $A$ is polynomially normal of degree two. These can be characterized completely; see also [6].
Theorem 3.6 [34]. If \( A \in \mathbb{C}^{n \times n} \) is the square root of a normal matrix, then \( A \) is unitarily similar to \( \begin{bmatrix} N_1 & N_2 \\ 0 & -N_1 \end{bmatrix} \oplus N \) with normal matrices \( N_1 \) and \( N \) and a positive definite matrix \( N_2 \) commuting with \( N_1 \).

Note that also \( N_2 \) can be chosen to be normal.

Assume \( A \) is a square root of a normal matrix. Then the converted linear system (1.1) reads \( A^2x = Ab \) which misleadingly looks like solving the normal equations. For a normal matrix we do have \( \kappa(A^2) = \kappa(AA^*) \) while with nonnormal matrices this need not hold. In fact, even \( \kappa(A^2) \ll \kappa(AA^*) \) can be expected. For this, consider involutions and see Example 6 below.

Commuting normal matrices are simultaneously unitarily diagonalizable; see, e.g., [19]. Using this with the canonical form of Brown, we can infer that any binormal matrix is unitarily similar to a binormal block upper triangular matrix with diagonal blocks. Therefore the unitary orbit of binormal matrices can be regarded as a natural extension of normal matrices.

Definition 3.7. The set \( \mathcal{BN} \subset \mathbb{C}^{2n \times 2n} \) consists of matrices unitarily similar to \( \begin{bmatrix} D_1 & D_2 \\ 0 & D_3 \end{bmatrix} \) for diagonal matrices \( D_1, D_2, D_3 \in \mathbb{C}^{n \times n} \).

To give an example, in Theorem 2.3 we had matrices belonging to \( \mathcal{BN} \).

This yields us a unitarily invariant family of matrices containing the set of normal as well as the set of binormal matrices. Since already the set of normal matrices is of real dimension \( 4n^2 + 2n \) in \( \mathbb{C}^{2n \times 2n} \), we have a significantly larger set than just binormal matrices. For these matrices the formula of Phillips [33] for finding a nearest normal approximant holds, after performing a unitary similarity transformation.

This is an interesting structure also because unitarily diagonalizable \( \mathbb{R} \)-linear operators in \( \mathbb{C}^n \) give rise to elements of \( \mathcal{BN} \) through their real form; see [8].

If \( D_1 \) and \( D_3 \) are real such that \( D_2(D_1 - D_3) = 0 \), then \( A \) is readily seen to be 3-selfadjoint, i.e., \( A \) belongs to a particular class of Hereditary matrices [30].

The Geršgorin region \( \mathcal{G}(A) \) of \( A \in \mathbb{C}^{n \times n} \) is the union of the Geršgorin disks

\[
\mathcal{G}_l(A) = \left\{ \lambda \in \mathbb{C} : |a_{ll} - \lambda| \leq \sum_{j \neq l} |a_{lj}| \right\}
\]

for \( l = 1, \ldots, n \). For locating eigenvalues with the Geršgorin regions of unitary orbits, see [38]. We denote by \( \mathcal{U} \) the set of unitary matrices.

Theorem 3.8. If \( A \in \mathcal{BN} \), then the spectrum of \( A \) equals \( \bigcap_{U \in \mathcal{U}} \mathcal{G}(U^*AU) \).
Proof. We can assume \( A \) to be in its canonical form of Definition 3.7. Then, for \( j = 1, \ldots, n \), each span\{\( e_j, e_{n+j} \)\} is invariant for both \( A \) and \( A^* \) and these subspaces are orthogonal. So apply the corresponding unitary similarity to have \( A \) as a direct sum of matrices of size 2-by-2. Then use [38, Theorem 1] block-wise. □

The following simple fact is useful.

**Proposition 3.9.** If \( A \in BN \) and \( p \) is a polynomial, then \( p(A) \in BN \).

An involution acting on an even dimensional space, although typically is far from being binormal, belongs to \( BN \). More generally, the following holds.

**Theorem 3.10.** If the degree of the minimal normal polynomial of \( A \in C^{2n \times 2n} \) is two, then \( A \in BN \).

**Proof.** Let \( p(z) = z^2 + \alpha z \), with \( \alpha \in C \), be the minimal normal polynomial of \( A \). Then consider \( q(z) = p(z) + \alpha^2/4 = (z + \alpha/2)^2 \), for which \( q(A) \) is also normal since the set of normal matrices is translation invariant.

By Theorem 3.6, \( A + \alpha/2I \) is unitarily similar to a matrix of the form
\[
\begin{bmatrix}
N_1 & N_2 \\
0 & -N_1
\end{bmatrix} \oplus N.
\]
Since the blocks \( N_1 \) and \( N_2 \) can be chosen to be commuting normal matrices, by being simultaneously unitarily diagonalizable, we can assume \( N_1 \) and \( N_2 \) to be diagonal. Also, we can assume \( N \) to be diagonal. Since the dimension of the space is even, decompose \( N = J_1 \oplus J_2 \) into two equally large diagonal blocks \( J_1 \) and \( J_2 \). Then with a similarity permutation arrange \( D_1 = N_1 \oplus J_1 \), \( D_2 = N_2 \oplus 0 \) and \( D_3 = (-N_1) \oplus J_2 \) to have a binormal matrix of type of Definition 3.7. Since by Proposition 3.9 the set \( BN \) is translation invariant, the claim follows. □

Combining this with Theorem 3.8 extends [38, Theorem 2] and its corollaries since the set of matrices \( A \in C^{2n \times 2n} \) whose minimal normal polynomial is of degree two obviously contains the matrices with a quadratic minimal polynomial. See also [2].

**Example 6 (For this large scale problem, see [18,32]).** Consider
\[
A = \begin{bmatrix} 0 & I \\ H & -dI \end{bmatrix},
\]
where \( H \) is a (tridiagonal) Hermitian matrix and \( d \in C \). Now, \( A^2 - (-d)A = H \oplus H \) is Hermitian, so that the matrix in question is polynomially Hermitian of degree two.

Using the notation of Definition 3.7, we have the following result.
Theorem 3.11. Assume \( A = U \begin{bmatrix} D_1 & D_2 \\ 0 & D_3 \end{bmatrix} U^* \) for a unitary matrix \( U \in \mathbb{C}^{2n \times 2n} \). Let \( M_k = \sum_{j=0}^{k-1} D_1^j D_3^{k-j-1} \), let \( d \) be the smallest integer such that the matrices \( \{ D_2 M_k \}_{k=1}^d \) are linearly dependent. Then \( A \) is polynomially normal of degree \( d \).

Proof. Since \( D_1, D_2 \) and \( D_3 \) commute, there holds
\[
\begin{bmatrix}
D_1 & D_2 \\
0 & D_3
\end{bmatrix}^2 = \begin{bmatrix}
D_1^2 & (D_1 + D_3)D_2 \\
0 & D_3^2
\end{bmatrix}
\]
Hence, for \( k = 1, 2, \ldots \), by using commutativity and induction we have
\[
\begin{bmatrix}
D_1 & D_2 \\
0 & D_3
\end{bmatrix}^k = \begin{bmatrix}
D_1^k & M_k D_2 \\
0 & D_3^k
\end{bmatrix}
\] (3.5)
with \( M_k = \sum_{j=0}^{k-1} D_1^j D_3^{k-j-1} \). Any linear combination of the matrices (3.5) has as its (1, 2)-block the corresponding linear combination of the matrices \( M_k D_2 \). Since this linear combination is already Schur decomposed, a monic polynomial in \( A \) is normal if and only if this (1, 2)-block is the zero matrix. □

Any matrix \( A \in \mathcal{B} V \) is thus polynomially normal of degree \( \text{rank}(D_2) + 1 \) at most. Then it is also readily verified that if the matrix has only real eigenvalues, its minimal normal polynomial in \( A \) yields a Hermitian matrix.

Regarding the speed of convergence to the set of normal matrices, the following generalization of the distance formula of Phillips [33] illustrates how polynomial normality is well understood for the elements of \( \mathcal{B} V \). By \( \mathcal{P}_j(\infty) \) we denote the set of monic polynomials of degree \( j \) and by \( \mathcal{N} \) the set of normal matrices, while \( \| \cdot \| \) is the spectral norm.

Corollary 3.12. For \( A \in \mathcal{B} V \) and \( j \geq 2 \) we have
\[
\min_{p \in \mathcal{P}_j(\infty)} \text{dist}(p(A), \mathcal{N}) = \frac{1}{2} \min_{a_1, \ldots, a_{j-1} \in \mathbb{C}} \left\| D_2 \left( M_j - \sum_{k=1}^{j-1} \alpha_k M_k \right) \right\|.
\]

Proof. By Proposition 3.9, any polynomial in \( A \) remains in \( \mathcal{B} V \). The claim follows by using the distance formula of Phillips [33]. □

The formula of Phillips [33] also gives a best normal approximant explicitly.

The canonical form of Definition 3.7 for a nonderogatory element \( A \in \mathcal{B} V \) can be found in a numerically stable way by computing a Schur decomposition \( A = V(D + T)V^* \) of \( A \), where \( D \) and \( T \) is a diagonal and a strictly upper triangular matrix, respectively.
Corollary 3.13. Let $A \in \mathcal{B}N \subset \mathbb{C}^{2n \times 2n}$ be nonderogatory with a Schur decomposition $A = V(D + T)V^*$. Then $T$ has at most $n$ nonzero entries such that every row and column of $T$ has at most 1 nonzero entry.

**Proof.** Let $\{e_1, \ldots, e_{2n}\}$ be the standard basis corresponding to $U^*AU$ in its canonical form of Definition 3.7 with a unitary matrix $U$. Then, for $j = 1, \ldots, n$, each span$[e_j, e_{n+j}]$ is invariant for both $A$ and $A^*$ and these subspaces are orthogonal. Moreover, for a Schur decomposition $A = V(D + T)V^*$, the diagonal matrix $D_1 \oplus D_3$ equals $D$ after a possible permutation of its diagonal entries. Two eigenvalues of $D$ are connected with a nonzero element in $T$ if and only there was a connection between $[e_j, e_{n+j}]$ through $D_2$ with the index $j$ corresponding to the same eigenvalue pair. □

For a generic element $A \in \mathcal{B}N$ its minimal normal polynomial is computable by employing the matrices $\{D_2M_k\}_{k=1}^{d}$ since only a permutation is needed for constructing the canonical form of Definition 3.7 by using a Schur decomposition. Being diagonal matrices, the algorithm has a low complexity since only finding a Schur decomposition is an $O(n^3)$ computation.

By now it is clear that the set $\mathcal{B}N$ can also be characterized as consisting of those square matrices acting on an even dimensional space which are unitarily similar to a block diagonal matrix with blocks of size two at most. For these matrices, see [36]. In particular, matrices illustrating different aspects of iterative methods often appear to be elements of $\mathcal{B}N$; see, e.g., [31, Section 8] where the matrices $B_1, B_{\pm 1},$ and $B_\kappa$ considered all belong to $\mathcal{B}N$. Also the matrix of Example 2 is from $\mathcal{B}N$ (when the dimension is even).

Corollary 3.14. If $A = XJX^{-1}$ is a Jordan canonical form of $A \in \mathcal{B}N$, then the Jordan blocks of $J$ are of size 2 at most.

**Proof.** Perform a similarity transformation for the 2-by-2 blocks corresponding to each invariant subspace span$[e_j, e_{n+j}]$. □

Using this structure we can compute the dimension of $\mathcal{B}N$.

**Lemma 3.15.** Let $\mathcal{S}_0 \subset \mathbb{C}^{2 \times 2}$ denote the set of upper triangular matrices with a nonnegative (1, 2)-entry. Then $\mathbb{C}^{2 \times 2}$ equals the image of the mapping $(S, U) \mapsto USU^*$ with $S \in \mathcal{S}_0$ and $U \in \mathcal{U}$.

**Proof.** If $M = \begin{bmatrix} \lambda_1 & \lambda_2 \\ 0 & \lambda_3 \end{bmatrix}$ is an upper triangular matrix with complex entries, then

$$M = \begin{bmatrix} 1 & 0 \\ 0 & e^{-i\theta} \end{bmatrix} \begin{bmatrix} \lambda_1 \\ 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & e^{i\theta} \end{bmatrix}$$
where \( \theta = \arg(\lambda_2) \) (if \( \lambda_2 = 0 \), then put \( \theta = 0 \)). Thus, \( M \) is unitarily similar to an element of \( \mathscr{F}_0 \). Using this with the Schur decomposition proves the claim. □

**Theorem 3.16.** \( \mathcal{BN} \subset \mathbb{C}^{2n\times 2n} \) is a stratified submanifold of \( \mathbb{C}^{2n\times 2n} \) with the stratum of maximal real dimension \( 4n^2 + 4n \).

**Proof.** In the proof of Corollary 3.13 we showed that any \( A \in \mathcal{BN} \) is unitarily similar to block-diagonal matrix with blocks of size 2-by-2 at most. Conversely, every matrix of this form is in \( \mathcal{BN} \).

Let the set \( \mathcal{S} \subset \mathbb{C}^{2n\times 2n} \) consist of block-diagonal matrices whose blocks are upper triangular matrices of size 2-by-2 each having a nonnegative (1, 2)-entry. Then the image of the mapping \( f \) defined by \( (U,S) \mapsto USU^* \) \( \) from \( \mathcal{U} \times \mathcal{S} \) to \( \mathbb{C}^{2n\times 2n} \) is \( \mathcal{BN} \). Since this mapping is real analytic and proper (compact sets have compact pre-images), its image admits a stratification [15]. Let us find the maximum dimension of the strata.

Denote by \( \mathcal{S}_0 \) the subset of \( \mathcal{S} \) consisting of matrices with nonnormal 2-by-2 blocks whose diagonal entries satisfy the ordering

(i) \( |s_{2j}^2| > |s_{2(j+1)}^2| \) and

(ii) \( |s_{2j}^2| > |s_{j+2j+1}^2| \),

for \( j = 1, \ldots, (n-1) \).

By Lemma 3.15, the image of \( f \) restricted to \( \mathcal{U} \times \mathcal{S}_0 \) is dense in \( \mathcal{BN} \). Furthermore, \( USU^* = VS_2V^* \) with \( U, V \in \mathcal{U} \) and \( S_2 \in \mathcal{S}_0 \) if and only if \( S_2 = S_1 \) and \( U^*V \) commutes with \( S_1 \). Since \( S_1 \) has distinct eigenvalues, this forces \( U^*V \) to be polynomial in \( S_1 \). Moreover, since each block of \( S_1 \) is nonnormal, \( U^*V \) must be a direct sum \( e^{i\theta_j}I_2 \oplus \cdots \oplus e^{i\theta_n}I_2 \), where \( I_2 \) is the 2-by-2 identity matrix and \( \theta_j \in \mathbb{R} \) for \( j = 1, \ldots, n \). Since the real dimension of \( \mathcal{U} \subset \mathbb{C}^{2n\times 2n} \) is \( (2n)^2 \), there are \( 4n^2 - n \) real degrees of freedom to choose \( U \) and \( 5n \) real parameters for choosing \( S_1 \). In all, this yields \( 4n^2 + 4n \) real parameters. □

To deal with general square matrices, let us define

\[
\mathcal{P}_N^j = \{ A \in \mathbb{C}^{n\times n} : p(A) \text{ is normal for a monic polynomial } p \text{ of degree } j \text{ at most} \}. \tag{3.6}
\]

Clearly, \( \mathcal{P}_N^1 \) equals the set of normal matrices while \( \mathcal{P}_N^n = \mathbb{C}^{n\times n} \).

**Proposition 3.17.** The sequence \( \mathcal{N}^0 \subset \mathcal{P}_N^1 \subset \cdots \subset \mathcal{P}_N^n \subset \mathbb{C}^{n\times n} \) is strictly increasing.

**Proof.** It is clear that the sequence is increasing. The strictness can be established with the help of Example 5. Namely, take \( A = M \oplus A \in \mathbb{C}^{n\times n} \), where \( M \) is the
nilpotent shift of size $j \leq n$ and $A$ is a diagonal matrix. Then $A \in \mathcal{P}_j$ but $A \not\in \mathcal{P}_{j-1}$. □

On the growth of the dimension of $\mathcal{P}_j$ we can give the following lower bound.

**Theorem 3.18.** $\mathcal{P}_j \subset \mathbb{C}^{n \times n}$ is a star-shaped set containing a star-shaped smooth manifold of real dimension $n^2 + n + j(j - 1)$, for $j = 1, \ldots, n$.

**Proof.** Let $p$ be a polynomial. If $p(A)$ is normal and $s \in \mathbb{R}$, then after scaling the coefficients of $p$ appropriately to have $p_s$, we have a normal $p_s(sA)$. So linearly connecting $A$ to the zero matrix, we can infer that $\mathcal{P}_j$ is star-shaped.

For the second claim, consider those upper triangular matrices $D + T$ of size $j$-by-$j$, where $D$ is diagonal and $T$ is strictly upper triangular such that the entries $t_{1,k}$ of $T$, for $k = 2, \ldots, j$, are restricted to be strictly positive. Furthermore, assume that the diagonal entries of $D$ satisfy $|d_k| > |d_{k+1}|$, for $k = 1, \ldots, j - 1$. Denote the upper triangular matrices satisfying these restrictions by $\mathcal{F}_j$. Assume $B \in \mathbb{C}^{j \times j}$ belongs to the unitary orbit of an element of $\mathcal{F}_j$. With $\theta_k \in \mathbb{R}$ assume $\theta_k x_1, \ldots, \theta_k x_j$ are the eigenvectors of $B$ arranged according to $|d_k| > |d_{k+1}|$. Choose $\theta_1$ such that the first component of $e^{\theta_1 x_1}$ is nonnegative. Thereafter it is easy to verify that to get a Schur decomposition of $B$ with a triangular part from $\mathcal{F}_j$, the remaining $\theta_k$ are uniquely determined.

Consider matrices $(D + T) \oplus A$ with $D + T \in \mathcal{F}_j$, where $A$ is a diagonal matrix of size $(n - j)$-by-$(n - j)$ with $|\lambda_k| > |\lambda_{k+1}|$, for $k = 1, \ldots, n - j$. Denote these matrices by $\mathcal{F}_0$. Let $\mathcal{U}_0 \subset \mathbb{C}^{n \times n}$ denote those unitary matrices whose $(1, 1)$-entry is nonnegative. Its real dimension is $n^2 - 1$. Consider the mapping $(U, S) \mapsto USU^*$ from $\mathcal{U}_0 \times \mathcal{F}_0$ to $\mathbb{C}^{n \times n}$. Since this mapping is real analytic and proper (compact sets have compact preimages), its image admits a stratification [15]. Let us find the maximum dimension of the strata.

Assume $U, V \in \mathcal{U}_0$. Then $US_1 U^* = VS_2 V^*$ with $S_1, S_2 \in \mathcal{F}_0$ if and only if $S_2 = S_1$ and $U^* V$ commutes with $S_1$. This forces $U^* V = e^{\theta_1 I_j} \oplus e^{\theta_2} \oplus \cdots \oplus e^{\theta_{j-1}}$, where $I_j$ is the $j$-by-$j$ identity matrix and $\theta_k \in \mathbb{R}$, for $k = 1, \ldots, n - j + 1$. Since $U, V \in \mathcal{U}_0$ we have $\theta_1 = 0$. In all this yields us $n^2 - 1 - (n - j)$ real parameters for choosing the unitary matrix. To choose an element of $\mathcal{F}_0$, we have $j(j - 1) + j + 1 + 2(n - j)$ free real parameters. This gives us $n^2 + n + j(j - 1)$ parameters as claimed. □

Biinormal matrices were defined via the polynomial equations (2.3). However, each $\mathcal{P}_j$, for $2 \leq j \leq n - 1$, being defined as a union of the solution set of an infinite number of polynomial equations, is a more complicated set than that of binormal matrices. For an illustration, consider $\mathcal{P}_n$. Then, for any fixed $\alpha \in \mathbb{C}$, the variety

$$\{ A \in \mathbb{C}^{n \times n} : (A^2 - \alpha A)(A^* - \alpha A^*) - (A^{*2} - \alpha A^*)(A^2 - \alpha A) = 0 \}$$

is a subset of $\mathcal{P}_n$. Letting $\alpha$ vary and taking the union yields $\mathcal{P}_n$.
3.4. A canonical Schur decomposition

For a general square matrix $A \in \mathbb{C}^{n \times n}$ it is not obvious how to compute its minimal normal polynomial with an algorithm of $O(n^3)$ complexity. A brute force method can be devised in the Frobenius norm $\| \cdot \|_F$ although then the distance formula of Corollary 3.12 does not hold.

**Algorithm 1** (for computing the minimal normal polynomial of $A \in \mathbb{C}^{n \times n}$).

2. For $j = 2, 3, \ldots$
   1. Compute the Schur decomposition $A^j = V(D^j + T_j)V^*$ of $A^j$.
   2. Compute $\min_{\alpha_1, \ldots, \alpha_{j-1} \in \mathbb{C}} \| T_j - \sum_{k=1}^{j-1} \alpha_k T_k \|_F$.

The implementation is illustrative albeit naive. In Section 5 we present numerically a more reliable method.

The intermediate steps give rise to a "vector" measure of nonnormality in accordance with Henrici’s measure [16] defined, for $j = 1, \ldots, n - 1$, via

$$
\text{He}_j(A) = \min_{\alpha_1, \ldots, \alpha_{j-1} \in \mathbb{C}} \| T_j - \sum_{k=1}^{j-1} \alpha_k T_k \|_F.
$$

Hence $\text{He}_1(A)$ is the original deviation due to Henrici. If $\text{He}_j(A) = 0$ with $j < n$, then a particular Schur decomposition can be associated with the matrix.

**Theorem 3.19.** Let $p$ be the minimal normal polynomial of $A \in \mathbb{C}^{n \times n}$ with $k = \deg(p(A))$ such that $k_j$ are the multiplicities of the eigenvalues of $p(A)$, for $j = 1, 2, \ldots, k$. Then there is a Schur decomposition $A = U \text{diag}(M_1, \ldots, M_k)U^*$ of $A$ with upper triangular blocks $M_j \in \mathbb{C}^{k_j \times k_j}$, for $j = 1, \ldots, k$.

**Proof.** Let $A = V(D + T)V^*$ be a Schur decomposition of $A$. Since $p(A)$ is normal, $p(D + T)$ is diagonal. We assume the Schur decomposition to be such that the equaling diagonal entries of $p(D + T)$ are arranged in blocks (this can be achieved since $p(D + T)$ is a diagonal matrix commuting with $D + T$). Since the degree of $p(A)$ is $k$, there are $k$ blocks of size $k_j$, for $j = 1, 2, \ldots, k$. For $p(A)$ to commute with $A$, the corresponding Schur decomposition of $A$ must have $k$ triangular blocks of the respective size.

Consequently, if $\text{He}_j(A) = 0$ with $j < n$, then $A$ is reducible, i.e., it can be represented, after performing a unitary similarity transformation, as a direct sum of smaller matrices. For reducibility, see [17].
Polynomial normality can also be used in characterizing matrix Krylov subspaces qualitatively. To this end, consider
\[ K(A; I) = \text{span}\{I, A, \ldots, A^{n-1}\} \] (3.8)
which is also called the double commutant of \( A \). It is well known that its dimension equals the degree of the minimal polynomial of \( A \) and is thereby bounded by \( n \). In this regard polynomial normality yields more insightful qualitative information. For example, the double commutant of the matrix of Example 5 does not contain any other normal matrices besides multiples of the identity.

**Corollary 3.20.** For \( A \in \mathbb{C}^{n \times n} \) the dimension of \( \mathcal{N} \cap K(A; I) \) equals
\[ \max_{p(A) \in \mathcal{P}} \deg(p(A)). \]

**Proof.** The dimension is well defined since \( \mathcal{N} \cap K(A; I) \) is a subspace of \( \mathbb{C}^{n \times n} \) consisting of those polynomials in \( A \) that give a normal matrix. These matrices are closed under addition and multiplication by a scalar. The claim follows from the Schur decomposition introduced. \( \square \)

Aside from being unitarily invariant, this number is also translation and (non-zero) scaling invariant of \( A \). It is also invariant under taking the adjoint because the minimal as well as the minimal normal polynomial have the same degree for \( A \) and \( A^* \).

In contrast to Example 5, the dimension of \( \mathcal{N} \cap K(A; I) \) for the matrix of Example 2 equals \( n - 1 \), the largest value one can have with a nonnormal matrix. Hence, in view of iterative methods, we are dealing with an almost normal matrix in this geometrical sense proposed. Remark also that the dimension of \( \mathcal{N} \cap K(A; I) \) is always at least \( n \) for a nonderogatory \( A \in \mathcal{B}_N \subset \mathbb{C}^{2n \times 2n} \).

### 4. Measures of nonnormality related to iterative methods

Instead of expecting to find a low degree monic polynomial yielding a normal matrix when evaluated at the matrix, a more realistic alternative in practice is to strive for decrease in nonnormality. This aim gives rise to measures of nonnormality differing from the classical ones [9] since there is now an element of discreteness through the increase of the degree of the polynomial. For more familiar polynomial approximation problems related to iterative methods, see [12].

Denote by \( \mathcal{P}_j(\infty) \) the set of monic polynomials of degree \( j \) at most. The first problem is to find, for \( A \in \mathbb{C}^{n \times n} \) and \( j = 1, \ldots, n \), the value of
\[ \min_{p(A) \in \mathcal{P}_j(\infty)} \text{dist}(p(A), \mathcal{N}) \] (4.1)
in the spectral norm. For attaining zero the degree can be \( n \) and, as was illustrated in Example 5, it cannot be improved in general. This is a difficult problem in the spectral norm; no explicit formula is known even in the case \( j = 1 \).

In [25] we introduced, for \( j = 1, \ldots, \lfloor \frac{n}{2} \rfloor \), the problem of finding
\[
\min_{\text{rank}(F) \leq j} \text{dist}(A - F, \mathcal{N})
\]
and in [21] it was shown that for attaining zero \( j = \lfloor \frac{n}{2} \rfloor \) suffices. With binormal matrices we can demonstrate that this cannot be improved in general.

**Theorem 4.1.** There exist a binormal matrix \( A \in \mathbb{C}^{2n \times 2n} \) such that \( A - F \) is nonnormal for every \( F \) with \( \text{rank}(F) < n \).

**Proof.** Take \( A = \begin{bmatrix} 0 & I \\ 0 & 0 \end{bmatrix} \in \mathbb{C}^{2n \times 2n} \). Let \( A - F = A - UV^* \) with \( U, V \in \mathbb{C}^{n \times k} \) for \( k < n \). We identify \( V \) with the subspace it spans. Clearly, the nullspace \( N(A) \) of \( A \) is spanned by the first \( n \) standard basis vectors. We assume \( A - UV^* \) is normal and show that it leads to a contradiction. To this end we employ the fact that a square matrix \( M \) is normal if and only if \( \| Mx \| = \| M^*x \| \) for every vector \( x \); see, e.g., [14, Condition 64].

We have \( \dim(V^\perp \cap N(A)) \geq 1 \) so that taking a nonzero \( x \in V^\perp \cap N(A) \) gives \((A - UV^*)x = 0\). Assuming \( A - UV^* \) to be normal, we have \((A^* - VU^*)x = 0\) as well. Note that \( A^*x \neq 0 \) since the nullspace of \( A \) equals the orthogonal complement of the nullspace of \( A^* \). Since the range of \( A^* \) equals the orthogonal complement of the nullspace of \( A \), the equality \( A^*x = VU^*x \) implies that there is a vector in \( V \) belonging to the orthogonal complement of the nullspace of \( A \). Consequently, \( \dim(V^\perp \cap N(A)) \geq 2 \). Continuing this argument inductively, we can deduce that \( \dim(V^\perp \cap N(A)) \geq n \). The same reasoning can be used to show that \( \dim(U^\perp \cap N(A^*)) \geq n \). Therefore
\[
A - UV^* = \begin{bmatrix} 0 & I - R \\ 0 & 0 \end{bmatrix}
\]
with a matrix \( R \) with rank strictly less than \( n \). Since by (4.3) the matrix \( A - UV^* \) is already Schur decomposed, this forces \( R = I \) for the matrix to be normal. This, however, is in contradiction with the assumption that \( \text{rank}(R) < n \) and the claim follows. \( \square \)

The measures (4.1) and (4.2) quantify nonnormality very differently. For example, the matrix \( A \) of Example 5 was polynomially normal of degree \( n \) although we attain zero in (4.2) with a rank-1 perturbation by replacing the \((n, 1)\)-element of \( A \) with 1.

Conversely, if we take \( A = \begin{bmatrix} 0 & I \\ 0 & 0 \end{bmatrix} \in \mathbb{C}^{2n \times 2n} \), then \( A \) is the square root of a normal matrix while \( A - F \) is nonnormal for every matrix \( F \) with rank less than \( n \). Hence,
for iterative methods it is quite a challenge to find a single measure of nonnormality that would tell everything.

Since these two measures are so dissimilar, let us combine them as

$$\min_{p \in \mathcal{P}} \text{dist}(p(A) - F, \mathcal{N})$$

for $j + l \leq \lfloor \frac{n}{2} \rfloor$. The motivation for this is as follows. If zero is attained, then a polynomial in $A$ is a small rank perturbation of a normal matrix. Since $\mathcal{N}$ is translation invariant, $A_s(A) - F = N$ is normal for a polynomial $p(z) = zs(z)$. Hence solving a linear system $Ax = b$, for $b \in \mathbb{C}^n$, is equivalent to solving $N(I + N^{-1}F)x = s(A)b$ as long as $N$ is invertible. Using the Sherman–Morrison formula, this latter problem amounts to solving $\text{rank}(F) + 1$ linear systems involving $N$. For this we can execute the 3-term recurrence for normal matrices [22].

In computing Ritz values it is not clear how to preserve the length of the recurrence because in a small rank perturbation of $A$, the spectrum changes typically drastically. Let us describe a way to circumvent this in case (4.2) is zero for $j \ll n$.

**Example 7.** Assume $A = N + F$, where $N$ is normal and $F = UV^*$ is of rank $j \ll n$. To compute Ritz values for $A$ with the method proposed in [24], store

$$U = [u_1 \ldots u_j] \quad \text{and} \quad V = [v_1 \ldots v_j]. \quad (4.5)$$

Denote by $Q_k \in \mathbb{C}^{n \times k}$ the matrix with orthonormal columns that [24, Algorithm 1] has generated with $N$ at the $k$th step. For Ritz values, consider

$$Q_k^* A Q_k = Q_k^* N Q_k + Q_k^* F Q_k = Q_k^* N Q_k + (U^* Q_k)^* V^* Q_k. \quad (4.6)$$

Treating the terms on the right separately, [24, Algorithm 1] yields $Q_k^* N Q_k$ with a recurrence whose length does not exceed $\sqrt{8k}$. To find $U^* Q_k$ and $V^* Q_k$ we do not need to preserve any of the columns of $Q_k$ while the computation proceeds. Hence the storage consumption for Ritz values with this approach is bounded by $2j + \sqrt{8k}$. The difference is more drastic if the spectrum of $N$ lies on low degree algebraic curve; see [24]. Depending on the degree, the maximum number of vectors that needs be stored is constant. For example, if $N$ is Hermitian, then only $2j + 3$ vectors needs to be saved.

Note that in the preceding example the actual iteration did not employ $F$. Only in the projection (4.6) was $F$ taken into account. A way to employ $F$ also during the iteration is to choose the (re-)starting vector(s) from the columns of $U$ and $V$.

**Example 8.** We demonstrate the idea of Example 7 with a small but illustrative example by using Matlab [29]. Assume $A = N + F$ with $n = 1000$, where $R = \text{randn}(n, n) + i \text{randn}(n, n)$ and $N = R + R^*$, and $F$ is a random matrix with rank $(F) = 5$. Rounding to five digits, we had $\|A\| = 125.92$, $\|N\| = 125.87$, and the largest and the smallest nonzero singular values of $F$ were $\sigma_1(F) = 45.755$ and $\sigma_5(F) = 41.666$. By using a random complex vector as a starting vector, we took
30 steps of the Arnoldi method and 50 steps of the method (4.6). See Figures 4.1 and 4.2, respectively. Note that with the latter alternative we needed to save only 13 vectors, independently of the number of steps. Regardless of that, to our mind the method (4.6) yields here better approximations to several extreme eigenvalues of $A$.

Sometimes in the numerical solution of a PDE, a splitting $A = N + F$ of $A$ can be obtained directly by discretizing the boundary conditions separately.

By the same arguments that led to (4.4), we are interested in finding

$$
\min_{p \in \mathcal{P}_j(\infty), \text{rank}(F) \leq l} \text{dist}(p(A - F), \mathcal{N}),
$$

(4.7)

for $j + l \leq \lceil \frac{n}{2} \rceil$. The following relation between (4.4) and (4.7) is obvious.

**Proposition 4.2.** Assume $A = M + F \in \mathbb{C}^{n \times n}$ with $p(M)$ normal. Then $p(A) = p(M) + G$ with $\text{rank}(G) \leq \deg(p) \text{rank}(F)$.

5. **Algorithms for computing the polynomials introduced**

The computational approach outlined in Section 3.4 is not numerically reliable. For a more stable algorithm, consider a Schur decomposition $A = V(D + T)V^*$ of

![Figure 4.1. The spectrum of $A$ (depicted by 'x') and the Ritz values with the Arnoldi method (depicted by 'o') after 30 steps from Example 8.](image-url)
A ∈ ℂ^{n×n}. Define a linear operator on ℂ^{n×n} via the matrix-matrix product

\[ X \mapsto (D + T)X \]  

for \( X \in ℂ^{n×n} \). Using the Arnoldi method, compute a Hessenberg form \( H = (h_{l,k}) \) for this operator by using \( Q_1 = (D + T)/\|A\|_F \) as a starting vector. We denote by \( Q_j \) the arising orthonormal matrices and set \( V_1 = T \) and \( \alpha_j = \prod_{l=2}^{j-1} h_{l,l-1} \), for \( j \geq 2 \).

**Algorithm 2** (for computing the minimal normal polynomial of \( A \in ℂ^{n×n} \)).

1. for \( j = 2, 3, \ldots \) compute the orthonormal matrices \( Q_j \)
2. set \( V_j = \alpha_j \|A\|_F (Q_j - \text{diag}(Q_j)) \)
3. compute \( \text{He}_j(A) = \min_{\gamma_1, \ldots, \gamma_{j-1} \in ℂ} \|V_j - \sum_{k=1}^{j-1} \gamma_k V_k\|_F \)
4. if \( \text{He}_j(A) = 0 \), end
5. form the polynomial corresponding to zero

In this manner, by computing an orthonormal basis of the matrix Krylov subspace

\[ \mathcal{K}(D + T; D + T) = \text{span}[D + T, (D + T)^2, \ldots, (D + T)^n] \]
we avoid generating the power basis \( \{ A^k \}_{k \geq 1} \). With this orthonormal basis the strictly upper triangular matrices \( T_k \) of Algorithm 1 can then be found in a more stable way.

For an illustration of Algorithm 2, consider the following example computed by using MATLAB.

**Example 9.** We take four matrices of size 20-by-20 each scaled to have the spectral norm equal 1. The matrix \( A_1 \) is a complex random matrix. The matrix \( A_2 \) is binormal with random complex diagonal blocks. The matrix \( A_3 = M_1 \otimes M_2 \) with a complex random \( M_1 \in C^{4 \times 4} \) and a Hermitian diagonal matrix \( M_2 \in C^{5 \times 5} \). Finally, \( A_4 \) is the matrix of Example 5, i.e., the nilpotent shift. See Figure 5.1 for the behavior of the \( \text{He}_j(A) \).

The algorithm proposed cannot be regarded as practical for large problems. The following method is "semi sparse" in the sense that we need a single Schur decomposition. Thereafter we compute only matrix-vector products. To this end, recall that the Arnoldi method with \( D + T \) and a starting vector \( \tilde{q}_0 \in C^n \) generates orthonormal vectors \( \tilde{q}_j \) which can be represented as \( \tilde{q}_j = p_j(D + T)\tilde{q}_0 \) with polynomials \( p_j \). These polynomials can be formed by using the entries of the Hessenberg matrix computed.

![Fig. 5.1. From Example 9 the behavior of \( \text{He}_j(A) \) for matrices \( A_1, A_2, A_3 \) and \( A_4 \in C^{20 \times 20} \) denoted by the solid line, ‘--’, ‘+-’ and ‘*--’, respectively.](image-url)
Algorithm 3 (for computing the minimal normal polynomial of $A \in \mathbb{C}^{n \times n}$).

compute a Schur decomposition $A = V(D + T)V^*$ of $A$

for $\tilde{q}_0 \in \mathbb{C}^n$

using the Arnoldi method with $D + T$ compute $\tilde{q}_j = p_j(D + T)\tilde{q}_0$

compute $\hat{q}_j = \tilde{q}_j - p_j(D)\tilde{q}_0$

orthogonalize $\hat{q}_j$ against $\text{span}\{q_1, q_2, \ldots, q_{j-1}\}$ to get $q_j$

if $q_j = 0$, end

form the polynomial corresponding to $q_j = 0$

end.

Hence the purpose of the step $\hat{q}_j = \tilde{q}_j - p_j(D)\tilde{q}_0$ is to “deflate the diagonal part” from the vector $\tilde{q}_j = p_j(D + T)\tilde{q}_0$.

It is critical to compute the coefficients of the polynomials $p_j$ accurately in order to generate $p_j(D)\tilde{q}_0$ accurately. Note that since $D$ is a diagonal matrix, the latter is a polynomial evaluation and not a matrix–vector multiplication problem.

6. Conclusions

We have considered aspects of nonnormality from the point of view of the Krylov subspace methods recently devised for normal matrices. A matrix is regarded as almost normal if there is a circuitous way of using these methods for solving linear systems or finding approximations to its eigenvalues. Binormal matrices, their unitary orbit and, as their natural extension, polynomially normal matrices of moderate degree were studied. Various matrix nearness problems were introduced. As a motivation, we showed how Ritz values can be computed with modest storage requirements in case we have an almost normal matrix in the sense proposed. Three algorithms were devised for computing the polynomials introduced.

References
