Helsinki University of Technology Institute of Mathematics Research Reports
Teknillisen korkeakoulun matematiikan laitoksen tutkimusraporttisarja
Espoo 2002 A453

ASPECTS OF NONNORMALITY FOR ITERATIVE METHODS

Marko Huhtanen

TEKNILLINEN KORKEAKOULU
TEKNISKA HÖGSKOLAN
HELSINKI UNIVERSITY OF TECHNOLOGY
TECHNISCHE UNIVERSITÄT HELSINKI
UNIVERSITE DE TECHNOLOGIE D'HELSINKI

Helsinki University of Technology Institute of Mathematics Research Reports
Teknillisen korkeakoulun matematiikan laitoksen tutkimusraporttisarja
Espoo 2002 A453

ASPECTS OF NONNORMALITY FOR ITERATIVE METHODS

Marko Huhtanen

Helsinki University of Technology
Department of Engineering Physics and Mathematics
Institute of Mathematics

Marko Huhtanen: *Aspects of nonnormality for iterative methods*; Helsinki University of Technology Institute of Mathematics Research Reports A453 (2002).

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.*

AMS subject classifications: 65F10, 65F15

Keywords: nonnormal matrix, binormal matrix, polynomially normal operator, unitary orbit, involution, iterative methods, Ritz values, measure of nonnormality

Marko.Huhtanen@hut.fi

ISBN 951-22-6155-3
ISSN 0784-3143
Institute of Mathematics, HUT, 2002

Helsinki University of Technology
Department of Engineering Physics and Mathematics
Institute of Mathematics
P.O. Box 1100, 02015 HUT, Finland
email:math@hut.fi http://www.math.hut.fi/

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 [7, 23, 24]. In this paper we study various aspects of nonnormality that arise from the existence of these algorithms. To this end we consider the set of binormal matrices [3] as well as its unitary orbit. These two families of matrices 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 [8]. However, there are ways to relate these matrices to normal matrices. For this purpose we decompose any invertible binormal matrix as the product of a normal matrix and one or two (nontrivial) involutions. A matrix $P \in \mathbb{C}^{n\times n}$ is an involution if $P^2 = I$. This factorization can be achieved inexpensively with a modification of the Schur complement. Moreover, all the factors can be regarded as polynomially normal matrices of very low degree.

Polynomial normality is originally an infinite dimensional operator theoretic concept; see [28, 29] 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 q with $\deg(q) < d$. Modulo a constant term, p is unique. In particular, involutions are polynomially normal matrices of degree 2. Binormal matrices are polynomially normal of degree at most half of the dimension of the underlying space. The size of d is critical for our purposes, motivated by computations, since the concept is otherwise vacuous for matrices.

Since polynomial normality of particular degree remains invariant under unitary similarity transformations, we consider the unitary orbit of binormal matrices denoted by $\mathcal{BN}$. This set, studied in a completely different context [38], provides a natural extension of normal matrices. It arises in connection with $\mathbb{R}$–linear operators in $\mathbb{C}^n$ [6]. Elements of $\mathcal{BN}$ have also appeared in illustrating various aspects of iterative methods [31, 13]. They can be linked with [41]. 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 $b \in \mathbb{C}^{n\times n}$, can be

accomplished by solving

$$Nx = s(A)b \tag{1}$$

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, 27] 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 the preceding, 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

$$\{p(A)\}_{p \text{ monic, } \deg(p)\leq k} \tag{2}$$

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 (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 also demonstrate that any invertible binormal matrix can be factored inexpensively as the product of involutions and a normal matrix. 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 them. 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

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 any such a factorization, like the polar decomposition,

is too expensive with the existing techniques. Therefore it is of interest to identify matrices for which there are inexpensive factorizations with nearly normal factors.

The members of the following set of matrices introduced by Brown [3] admit a closed form solution to the problem of finding a nearest normal approximant; see [33, 2, 34] and references therein.

Definition 1 $A = \left[\begin{smallmatrix} N_1 & N_2 \\ N_3 & N_4 \end{smallmatrix}\right] \in \mathbb{C}^{2n\times 2n}$ *is binormal if* $\{N_j\}_{j=1}^4 \subset \mathbb{C}^{n\times n}$ *are commuting normal matrices.*

The following canonical form of Brown is useful in practice; see [3, 2].

Theorem 1 *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 $\left[\begin{smallmatrix} 0 & I \\ 0 & 0 \end{smallmatrix}\right] \in \mathbb{C}^{2n\times 2n}$ and the example that follow illustrate.

Example 1 Take 4 circulant matrices $\{N_j\}_{j=1}^4$ and form a 2-by-2 block matrix out of them to have a binormal matrix A. For block circulant matrices arising in practice, see [4].

A square matrix P is an involution if $P^2 = I$.

Proposition 1 *Let* $A \in \mathbb{C}^{2n\times 2n}$ *be an invertible binormal matrix. Then* $A = NP$ *for a normal matrix* N *and an involution* P.

Proof. Let $A = UTU^*$, with a unitary U and an upper triangular binormal matrix $T = \left[\begin{smallmatrix} \widehat{N}_1 & \widehat{N}_2 \\ 0 & \widehat{N}_4 \end{smallmatrix}\right]$, be in the canonical form of Brown. Then factoring $T = T_1T_2 = \left[\begin{smallmatrix} -\widehat{N}_1 & 0 \\ 0 & \widehat{N}_4 \end{smallmatrix}\right] \left[\begin{smallmatrix} -I & -\widehat{N}_1^{-1}\widehat{N}_2 \\ 0 & I \end{smallmatrix}\right]$ gives $A = UT_1U^*UT_2U^* = NP$. □

Classically P is regarded as very nonnormal. However, from the point of view of iterative methods an involution is almost normal; see section 3. The converse also yields an interesting question: Characterize those matrices which can be represented as the product of a normal matrix and an involution. (Recall that every invertible matrix is the product of a complex symmetric matrix and an involution [10].)

For the canonical form of Brown one needs to compute an eigendecomposition which is costly. To avoid this, we proceed as follows. Multiplying a binormal matrix A by the involution $\Pi = \left[\begin{smallmatrix} 0 & I \\ I & 0 \end{smallmatrix}\right]$ from the left and/or right, we can have any block as the $(1,1)$-block. Binormality is preserved in this operation so we can assume that, after a possible permutation, the matrix N_1 is the easiest to solve linear systems with. In an ideal case N_1 would be close to the identity. Then

$$A = \left[\begin{smallmatrix} -I & 0 \\ -N_1^{-1}N_3 & I \end{smallmatrix}\right] \left[\begin{smallmatrix} -N_1 & -N_2 \\ 0 & N \end{smallmatrix}\right] \tag{3}$$

with $N = N_4 - N_1^{-1}N_3N_2$. These factors are still binormal.

Proposition 2 *Let $\{N_j\}_{j=1}^k \subset \mathbb{C}^{n\times n}$ be commuting normal matrices. Then $r(N_1,\ldots,N_k)$ is normal for any rational function r for which $r(N_1,...,N_k)$ is well defined.*

Proof. As commuting normal matrices are simultaneously diagonalizable by a unitary matrix [19, Theorem 2.5.5], we have the claim. □

Since $\left[\begin{smallmatrix} -N_1 & -N_2 \\ 0 & N \end{smallmatrix}\right] = \left[\begin{smallmatrix} N_1 & 0 \\ 0 & N \end{smallmatrix}\right] \left[\begin{smallmatrix} -I & -N_1^{-1}N_2 \\ 0 & I \end{smallmatrix}\right]$, we have the following.

Theorem 2 *If $A \in \mathbb{C}^{2n\times 2n}$ is invertible and binormal, then $A = P_1(M_1 \oplus M_2)P_2P_3$ with normal $M_1, M_2 \in \mathbb{C}^{n\times n}$ and involutions P_1, P_2 and P_3, where P_3 is either I or Π; this factorization does not use the canonical form of Brown.*

Proof. We have already proved the case of N_1 being invertible. If N_1 is not invertible but N_4 is, then $\Pi A\Pi$ reduces to the case where the $(1,1)$-block is invertible. Then multiplying with Π from the left and right yields

$$A = (\Pi P_1\Pi)\big(\Pi(M_1 \oplus M_2)\Pi\big)(\Pi P_2\Pi) = \widehat{P}_1(M_2 \oplus M_1)\widehat{P}_2. \tag{4}$$

It remains to consider the case where both N_1 and N_4 are singular. Then $A\Pi$ must yield an invertible $(1,1)$-block since A is invertible. Thus, proceeding as in the first case gives us $A\Pi = P_1(M_1 \oplus M_2)P_2$ proving the claim. □

Clearly only P_1 and P_2 involve computations.

If linear systems with N_1 can be solved very fast, then $Ax = b$, for $b \in \mathbb{C}^{2n}$, can be iteratively solved by solving 4 linear systems involving normal matrices. Each of these problems is half of the size of the original system. Since only matrix–vector products are performed, this factorization can be employed implicitly, i.e., the factors need not be explicitly constructed.

The dimension of binormal matrices in is as follows.

Theorem 3 *The set of binormal matrices in $\mathbb{C}^{2n\times 2n}$ is a stratified submanifold 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 $A = \left[\begin{smallmatrix} N_1 & N_2 \\ N_3 & N_4 \end{smallmatrix}\right] \in \mathbb{C}^{2n\times 2n}$ be any, not necessarily a binormal, matrix. Then A is binormal if and only if $\{N_j\}_{j=1}^4$ satisfy

$$N_jN_k - N_kN_j = 0 \text{ and } N_jN_j^* - N_j^*N_j = 0 \tag{5}$$

for all $1 \le j,k \le 4$. 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 [9].

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 $2 \le j \le 8$ [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

$$\begin{bmatrix} H_1+ip_1(H_1) & p_2(H_1)+ip_3(H_1) \\ p_4(H_1)+ip_5(H_1) & p_6(H_1)+ip_7(H_1) \end{bmatrix} \tag{6}$$

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 sheer comparison is not 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 (6) 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.

3 Polynomial normality for matrices

Involutions have 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 2 *$A \in \mathbb{C}^{n\times n}$ 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 we define $A \in \mathbb{C}^{n\times n}$ 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 [28, 29] 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. We illustrate this by extending iterative methods to nonnormal problems, both for solving linear systems and locating eigenvalues, when d is moderate.

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 readily seen to be 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

$$A = Ns(A)^{-1} \tag{7}$$

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

$$Nx = s(A)b \tag{8}$$

under the assumption that both A and N are invertible. Since in the latter system the coefficient matrix is normal, this seemingly nonnormal problem can be solved with techniques for normal matrices [22, 26], 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 $Z\Lambda Z^{-1}$ with

$$Z = \begin{bmatrix} 1 & \sqrt{1-\delta} & 0 & \dots & 0 \\ 0 & \sqrt{\delta} & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix} \quad \text{and} \quad \Lambda = \operatorname{diag}(20, 10, 5, \dots, 1),$$

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 *Assume $A \in \mathbb{C}^{n\times n}$ is unitarily similar to $M \oplus \Lambda$, with $M \in \mathbb{C}^{d\times d}$ and a diagonal matrix $\Lambda \in \mathbb{C}^{(n-d)\times(n-d)}$. 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 [39, Section 4]. Polynomial normality preserves data better since by knowing the eigenvalues of A, the spectrum of $p(A)$ available. For the converse, the following two propositions are also direct consequences of the spectral mapping theorem.

Proposition 4 *Let $p(A) = M \in \mathbb{C}^{n\times n}$ for a polynomial p and assume $\sigma(M) \subset \mathcal{S}$. Then $\sigma(A) \subset \{z \in \mathbb{C} : p(z) \in \mathcal{S}\}$.*

In particular, if $p(A)$ is normal, then algorithms proposed in [24] can be employed for generating sets $\mathcal{S}$ containing its spectrum. What then remains is to find the inverse image of $\mathcal{S}$ 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 5 *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 a circuitous way to using real analytic computational techniques for finding approximations to eigenvalues of a nonnormal matrix. The idea is to generate Ritz values for a normal $p(A)$ 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 5, 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 5 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, 27]. We briefly describe the connection as follows. If $A \in \mathbb{C}^{n\times n}$ is such that $p(A)$ is Hermitian, then

$$p(A) - p(A)^* = \sum_{0\leq k,l\leq d} c_{k,l} A^{*k} A^l = 0, \tag{9}$$

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 to belong to the hereditary class corresponding to p.

3.3 The unitary orbit of binormal matrices and polynomially normal matrices of low degree

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 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, $p(\lambda) = \lambda^n$ is the minimal normal polynomial of A.

If A is the square root of a normal matrix, like an involution, then A is polynomially normal of degree 2. These can be characterized completely; see also [5].

Theorem 4 *[34] If $A \in \mathbb{C}^{n\times n}$ is the square root of a normal matrix, then A is unitarily similar to $\left[\begin{smallmatrix} N_1 & N_2 \\ 0 & -N_1 \end{smallmatrix}\right] \oplus N$ 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 system (8) reads $A^2x = Ab$ which falsely resembles 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, $\kappa(A^2) \ll \kappa(AA^*)$ is quite realistic which can be illustrated, e.g., with involutions. See also Example 6.

Commuting normal matrices are simultaneously unitarily diagonalizable; see [19]. Hence by employing the canonical form of Brown, any binormal matrix is unitarily similar to a binormal upper triangular matrix with diagonal blocks. The unitary orbit of binormal matrices can be regarded as a natural extension of normal matrices.

Definition 3 *The set* $\mathcal{BN} \subset \mathbb{C}^{2n\times 2n}$ *consists of matrices* $A \in \mathbb{C}^{2n\times 2n}$ *with*

$$A = U \begin{bmatrix} D_1 & D_2 \\ 0 & D_3 \end{bmatrix} U^*$$

for diagonal matrices $D_1, D_2, D_3 \in \mathbb{C}^{n\times n}$ *and a unitary matrix* $U \in \mathbb{C}^{2n\times 2n}$.

This yields 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 [6].

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

$$\mathcal{G}_l(A) = \{\lambda \in \mathbb{C} : |a_{ll} - \lambda| \leq \sum_{j\neq l} |a_{lj}|\}, \tag{10}$$

for $l = 1, \ldots, n$. For locating eigenvalues with the Geršgorin regions of unitary orbits, see [41]. We denote by $\mathcal{U}$ the set of unitary matrices.

Theorem 5 *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. Then, for $j = 1, \ldots, n$, each $\operatorname{span}\{e_j, e_{n+j}\}$ 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 [41, Theorem 1] block–wise. □

The following simple fact is useful.

Proposition 6 *Let* $A \in \mathcal{BN}$ *and* p *be a polynomial. Then* $p(A) \in \mathcal{BN}$.

An involution acting in an even dimensional space, although typically is far from being binormal, belongs to $\mathcal{BN}$. More generally, the following holds.

Theorem 6 *If the degree of the minimal normal polynomial of* $A \in \mathbb{C}^{2n\times 2n}$ *is* 2, *then* $A \in \mathcal{BN}$.

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

By Theorem 4, $A + \alpha/2I$ is unitarily similar to a matrix of the form $\left[\begin{smallmatrix} N_1 & N_2 \\ 0 & -N_1 \end{smallmatrix}\right] \oplus N$. 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. Since by Proposition 6 the set $\mathcal{BN}$ is translation invariant, the claim follows. □

Combining this with Theorem 5 extends [41, Theorem 2] and its corollaries since the set of matrices $A \in \mathbb{C}^{2n\times 2n}$ whose minimal normal polynomial is of degree 2 obviously contains the matrices with a quadratic minimal polynomial.

Example 6 (For this large scale problem, see [18, 32].) Consider

$$A = \left[\begin{smallmatrix} 0 & I \\ H & -dI \end{smallmatrix}\right],$$

where H is a (tridiagonal) Hermitian matrix and $d \in \mathbb{C}$. Now, $A^2 - (-d)A = H \oplus H$ is Hermitian, so that the matrix in question is polynomially Hermitian of degree 2.

Using the notation of Definition 3, we have the following.

Theorem 7 *For $A = U \left[\begin{smallmatrix} D_1 & D_2 \\ 0 & D_3 \end{smallmatrix}\right] U^* \in \mathcal{BN}$ and $M_k = \sum_{j=0}^{k-1} D_1^j D_3^{k-j-1}$ let d be the smallest integer such that the diagonal 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, we have

$$\left[\begin{smallmatrix} D_1 & D_2 \\ 0 & D_3 \end{smallmatrix}\right]^2 = \left[\begin{smallmatrix} D_1^2 & (D_1+D_3)D_2 \\ 0 & D_3^2 \end{smallmatrix}\right].$$

Consequently, by using commutativity and by induction we have

$$\left[\begin{smallmatrix} D_1 & D_2 \\ 0 & D_3 \end{smallmatrix}\right]^k = \left[\begin{smallmatrix} D_1^k & M_k D_2 \\ 0 & D_3^k \end{smallmatrix}\right] \tag{11}$$

with $M_k = \sum_{j=0}^{k-1} D_1^j D_3^{k-j-1}$, for $k = 1, 2, \ldots$. Any linear combination of the matrices (11) 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 $A \in \mathcal{BN}$ is thus polynomially normal of degree $\operatorname{rank}(D_2) + 1$ at most. Moreover, it is readily verified that if the matrix has real eigenvalues, then the minimal normal polynomial 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 completely understood for the elements of $\mathcal{BN}$. By $\mathcal{P}_j(\infty)$ we denote the set of monic polynomials of exact degree j and by $\mathcal{N}$ the set of normal matrices while $||\cdot||$ is the spectral norm.

Corollary 1 *For $A \in \mathcal{BN}$ and $j \geq 2$ we have*

$$\min_{p \in \mathcal{P}_j(\infty)} \operatorname{dist}(p(A), \mathcal{N}) = \frac{1}{2} \min_{\alpha_1, \ldots, \alpha_{j-1} \in \mathbb{C}} \left\| D_2 (M_j - \sum_{k=1}^{j-1} \alpha_k M_k) \right\| .$$

Proof. By Proposition 6, any polynomial in A remains in $\mathcal{BN}$. 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 for a nonderogatory element $A \in \mathcal{BN}$ 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 2 *Let $A \in \mathcal{BN} \subset \mathbb{C}^{2n \times 2n}$ 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 with a unitary matrix U. Then, for $j = 1, \ldots, n$, each $\operatorname{span}\{e_j, e_{n+j}\}$ 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 generic element $A \in \mathcal{BN}$ the minimal normal polynomial is computable by employing the matrices $\{M_k D_2\}_{k=1}^{d}$ since only a permutation is needed for constructing the canonical form of Definition 3 by using a Schur decomposition. Being diagonal matrices, the algorithm has a low complexity; only finding a Schur decomposition is an $O(n^3)$ computation.

By now it is clear that the set $\mathcal{BN}$ can also be characterized as consisting of those square matrices acting in an even dimensional space which are unitarily similar to a block diagonal matrix with blocks of size two at most. For these matrices, see [38]. In particular, matrices illustrating different aspects of iterative methods often appear to be elements of $\mathcal{BN}$; see, e.g., [31, Section 8] where the matrices B_1, $B_{\pm 1}$, and B_κ considered all belong to $\mathcal{BN}$. Also the matrix of Example 2 is from $\mathcal{BN}$ (when the dimension is even).

Corollary 3 *If $A = XJX^{-1}$ is a Jordan canonical form of $A \in \mathcal{BN}$, 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 $\operatorname{span}\{e_j, e_{n+j}\}$. □

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

Lemma 1 *Let $\mathcal{S}_0 \subset \mathbb{C}^{2\times 2}$ denote the set of upper triangular matrices with a non–negative $(1,2)$-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

$$M = \begin{bmatrix} 1 & 0 \\ 0 & e^{-i\theta} \end{bmatrix} \begin{bmatrix} \lambda_1 & |\lambda_2| \\ 0 & \lambda_3 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & e^{i\theta} \end{bmatrix},$$

where $\theta = \arg(\lambda_2)$ (if $\lambda_2 = 0$, then put $\theta = 0$). Thus, M is unitarily similar to an element of $\mathcal{S}_0$. Using this with the Schur decomposition proves the claim. □

Theorem 8 *$\mathcal{BN} \subset \mathbb{C}^{2n\times 2n}$ 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 2 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 non–negative $(1,2)$-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 preimages), 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) $|s_{2j\,2j}| > |s_{2(j+1)\,2(j+1)}|$ and
(ii) $|s_{2j\,2j}| > |s_{2j+1\,2j+1}|$,
for $j = 1, \ldots, (n-1)$.

By Lemma 1, the image of f restricted to $\mathcal{U} \times \mathcal{S}_0$ is dense in $\mathcal{BN}$. Furthermore, $US_1U^* = VS_2V^*$ with $U, V \in \mathcal{U}$ and $S_1, S_2 \in \mathcal{S}_0$ 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 $e^{i\theta_1}I_2 \oplus \cdots \oplus e^{i\theta_n}I_2$, 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 $\mathcal{PN}_j = \{A \in \mathbb{C}^{n\times n} : p(A)$ is normal for a monic polynomial p of degree j at most$\}$.

Clearly, $\mathcal{PN}_1$ equals the set of normal matrices while $\mathcal{PN}_n = \mathbb{C}^{n\times n}$.

Proposition 7 *The sequence $\mathcal{N} \subset \mathcal{PN}_2 \subset \cdots \subset \mathcal{PN}_{n-1} \subset \mathbb{C}^{n\times n}$ 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 $A = M \oplus \Lambda \in \mathbb{C}^{n\times n}$,

where M is the nilpotent shift of size $j \leq n$ and Λ is a diagonal matrix. Then $A \in \mathcal{PN}_j$ but $A \notin \mathcal{PN}_{j-1}$. □

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

Theorem 9 *$\mathcal{PN}_j \subset \mathbb{C}^{n\times n}$ is a star–shaped set containing a star–shaped smooth manifold of 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 $\mathcal{PN}_j$ 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 $\mathcal{S}_j$. Assume $B \in \mathbb{C}^{j\times j}$ belongs to the unitary orbit of an element of $\mathcal{S}_j$. With $\theta_k \in \mathbb{R}$ assume $e^{i\theta_1}x_1, \ldots, e^{i\theta_j}x_j$ are the eigenvectors of B arranged according to $|d_k| > |d_{k+1}|$. Choose θ_1 such that the first component of $e^{i\theta_1}x_1$ is non–negative. Thereafter it is easy to verify that to get a Schur decomposition of B with a triangular part from $\mathcal{S}_j$, the remaining θ_k are uniquely determined.

Consider matrices $(D + T) \oplus \Lambda$ with $D + T \in \mathcal{S}_j$, where Λ 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 $\mathcal{S}_0$. Let $\mathcal{U}_0 \subset \mathbb{C}^{n\times n}$ denote those unitary matrices whose $(1,1)$-entry is non–negative. Its real dimension is $n^2 - 1$. Consider the mapping $(U, S) \mapsto USU^*$ from $\mathcal{U}_0 \times \mathcal{S}_0$ 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 $US_1U^* = VS_2V^*$ with $S_1, S_2 \in \mathcal{S}_0$ if and only if $S_2 = S_1$ and U^*V commutes with S_1. This forces $U^*V = e^{i\theta_1}I_j \oplus e^{i\theta_2} \oplus \cdots \oplus e^{i\theta_{n-j+1}}$, 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 $\mathcal{S}_0$, 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. □

Binormal matrices were defined via the polynomial equations (5). However, each $\mathcal{PN}_j$, 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 $\mathcal{PN}_2$. Then, for any fixed $\alpha \in \mathbb{C}$, the variety

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

is a subset of $\mathcal{PN}_2$. Letting α vary and taking the union yields $\mathcal{PN}_2$.

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 $\|\cdot\|_{\mathcal{F}}$ although then the distance formula of Corollary 1 does not hold.

Algorithm 1. "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 $j = 2, 3, \ldots$
  compute the Schur decomposition $A^j = V(D_j + T_j)V^*$ *of* A^j
  compute $\min_{\alpha_1,\ldots,\alpha_{j-1}\in\mathbb{C}} \left|\left| T_j - \sum_{k=1}^{j-1} \alpha_k T_k \right|\right|_{\mathcal{F}}$
end.

The implementation is illustrative albeit naive. In section 5 we present a numerically 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

$$\mathrm{He}_j(A) = \min_{\alpha_1,\ldots,\alpha_{j-1}\in\mathbb{C}} \left|\left| T_j - \sum_{k=1}^{j-1} \alpha_k T_k \right|\right|_{\mathcal{F}}. \tag{12}$$

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

Theorem 10 *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 $p(A)$. Then there is a Schur decomposition $A = U\mathrm{diag}(M_1, \ldots, M_k)U^*$ 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 $D+T$). Then the k blocks are of size k_j, for $j = 1, \ldots, k$. For $p(A)$ to commute with A, the corresponding Schur decomposition of A must have k triangular blocks of respective size. □

If $\mathrm{He}_j(A) = 0$ 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

$$\mathcal{K}(A; I) = \mathrm{span}\{I, A, \ldots, A^{n-1}\} \tag{13}$$

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 4 *For $A \in \mathbb{C}^{n\times n}$ the dimension of $\mathcal{N} \cap \mathcal{K}(A; I)$ equals*

$$\max_{p(A)\in\mathcal{N}} \deg(p(A)).$$

Proof. The dimension is well defined since $\mathcal{N} \cap \mathcal{K}(A; I)$ 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. □

Aside from being unitary invariant, this number is also translation and (nonzero) 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 $\mathcal{N} \cap \mathcal{K}(A; I)$ for the matrix of Example 2 equals $n - 1$, the largest value one can have with a nonnormal matrix. Hence, regarding iterative methods, we are dealing with an almost normal matrix in this geometrical sense proposed. Remark also that the dimension of $\mathcal{N} \cap \mathcal{K}(A; I)$ is always at least n for a nonderogatory $A \in \mathcal{BN} \subset \mathbb{C}^{2n\times 2n}$.

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 [8] 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 $\widehat{\mathcal{P}}_j(\infty)$ 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

$$\min_{p\in\widehat{\mathcal{P}}_j(\infty)} \operatorname{dist}(p(A), \mathcal{N}) \tag{14}$$

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

$$\min_{\operatorname{rank}(F)\le j} \operatorname{dist}(A - F, \mathcal{N}) \tag{15}$$

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

Proof. Take $A = \left[\begin{smallmatrix} 0 & I \\ 0 & 0 \end{smallmatrix}\right] \in \mathbb{C}^{2n\times 2n}$. 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 null space $N(A)$ 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 $\dim\big(V^\perp \cap N(A)\big) \geq 1$ 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 null space of A equals the orthogonal complement of the null space of A^*. Since the range of A^* equals the orthogonal complement of the null space of A, the equality $A^*x = VU^*x$ implies that there is a vector in V belonging to the orthogonal complement of the null space of A. Consequently, $\dim\big(V^\perp \cap N(A)\big) \geq 2$. Continuing this argument inductively, we can deduce that $\dim\big(V^\perp \cap N(A)\big) \geq n$. The same reasoning can be used to show that $\dim\big(U^\perp \cap N(A^*)\big) \geq n$. Therefore

$$A - UV^* = \left[\begin{smallmatrix} 0 & I-R \\ 0 & 0 \end{smallmatrix}\right] \tag{16}$$

with a matrix R with rank strictly less than n. Since by (16) 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 $\mathrm{rank}(R) < n$ and the claim follows. □

The measures (14) and (15) quantify nonnormality very differently. The matrix A of Example 5 was polynomially normal of degree n although we attain zero in (15) with a rank-1 perturbation by replacing the $(n, 1)$-entry of A with 1. Conversely, $A = \left[\begin{smallmatrix} 0 & I \\ 0 & 0 \end{smallmatrix}\right] \in \mathbb{C}^{2n\times 2n}$ is the square root of a normal matrix while $A - F$ is nonnormal for every matrix F with rank less than n.

Since these two prescribed measures are so dissimilar, we combine them as

$$\min_{p\in\widehat{\mathcal{P}}_j(\infty),\,\mathrm{rank}(F)\leq l} \mathrm{dist}(p(A) - F, \mathcal{N}), \tag{17}$$

for $j + l \leq \lfloor \frac{n}{2} \rfloor$. The motivation for solving linear systems 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 a translation invariant set, $As(A) - F = N$ is normal for a polynomial s. Hence solving a linear system $Ax = b$ 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 $\mathrm{rank}(F) + 1$ linear systems involving N. See also [21].

For Ritz values it is not clear how to preserve the length of recurrence since in a small rank perturbation the spectrum changes typically drastically. Let us describe a way to circumvent this in case (15) 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

$$U = [u_1 \ \ldots \ u_j] \text{ and } V = [v_1 \ \ldots \ v_j]. \tag{18}$$

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

$$Q_k^* A Q_k = Q_k^* N Q_k + Q_k^* F Q_k = Q_k^* N Q_k + (U^* Q_k)^* \, V^* Q_k. \tag{19}$$

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.

REMARK. In the preceding example the actual iteration did not employ F. Only in the projection (19) 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`. 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 $\text{rank}\,(F) = 5$. Rounding to five digits, we had $||A|| = 125.92$, $||N|| = 125.87$, and the largest and the smallest non–zero 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 (19). See Figure 1 and Figure 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 (19) 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 (17), we are interested in finding

$$\min_{p\in\widehat{\mathcal{P}}_j(\infty),\,\text{rank}(F)\le l} \text{dist}(p(A - F), \mathcal{N}), \tag{20}$$

for $j + l \le \lfloor \frac{n}{2} \rfloor$. The following relation between (17) and (20) is obvious.

Proposition 8 *Assume* $A = M + F \in \mathbb{C}^{n\times n}$ *with* $p(M)$ *normal. Then* $p(A) = p(M) + G$ *with* $\text{rank}(G) \le \deg(p)\,\text{rank}(F)$.

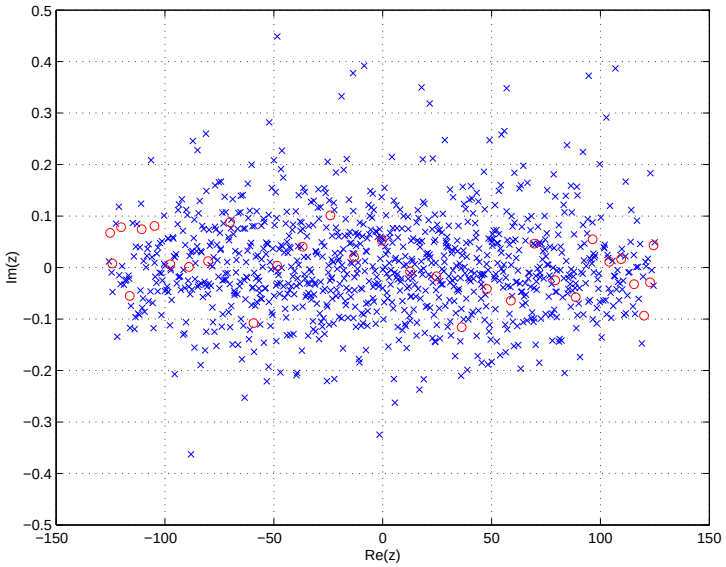

Figure 1: The spectrum of A (depicted by 'x') and the Ritz values with the Arnoldi method (depicted by'o') after 30 steps from Example 8.

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 $A \in \mathbb{C}^{n\times n}$. Define a linear operator on $\mathbb{C}^{n\times n}$ via the matrix–matrix product

$$X \mapsto (D+T)X \tag{21}$$

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

Algorithm 2. "for computing the minimal normal polynomial of $A \in \mathbb{C}^{n\times n}$."

for $j = 2, 3, \ldots$ *compute the orthonormal matrices* Q_j
set $V_j = \alpha_j \left|\left|A\right|\right|_{\mathcal{F}} \left(Q_j - \mathrm{diag}(Q_j)\right)$
compute $\mathrm{He}_j(A) = \min_{\gamma_1,\ldots,\gamma_{j-1}\in\mathbb{C}} \left|\left| V_j - \sum_{k=1}^{j-1} \gamma_k V_k \right|\right|_{\mathcal{F}}$
if $\mathrm{He}_j(A) = 0$, *end*
form the polynomial corresponding to zero
end.

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

$$\mathcal{K}(D+T; D+T) = \mathrm{span}\left\{D+T, (D+T)^2, \ldots, (D+T)^n\right\}$$

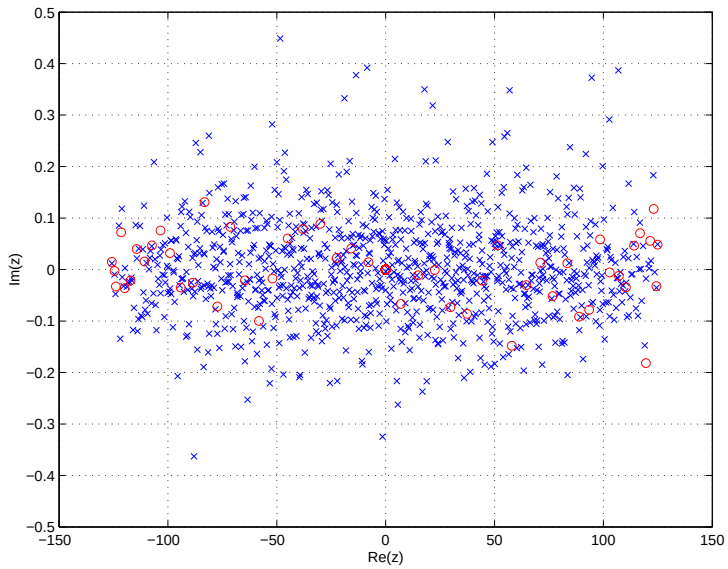

Figure 2: The spectrum of A (depicted by 'x') and the Ritz values with the method (19) (depicted by 'o') after 50 steps from Example 8.

we avoid generating the power basis $\{A^k\}_{k\geq 1}^n$. 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 $M_1 \in \mathbb{C}^{4\times 4}$ and a Hermitian diagonal matrix $M_2 \in \mathbb{C}^{5\times 5}$. Finally, A_4 is the matrix of Example 5, i.e., the nilpotent shift. See Figure 3 for the behavior of the $\mathrm{He}_j(A)$.

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 $\widehat{q}_0 \in \mathbb{C}^n$ generates orthonormal vectors $\widehat{q}_j$ which can be represented as $\widehat{q}_j = p_j(D+T)\widehat{q}_0$ with polynomials p_j. These polynomials can be formed by using the entries of the Hessenberg matrix computed.

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 $\widehat{q}_0 \in \mathbb{C}^n$
  using the Arnoldi method with $D+T$ *compute* $\widehat{q}_j = p_j(D+T)\widehat{q}_0$
  compute $\widetilde{q}_j = \widehat{q}_j - p_j(D)\widehat{q}_0$
  orthogonalize $\widetilde{q}_j$ *against* $\mathrm{span}\{q_1, q_2, \ldots, q_{j-1}\}$ *to get* q_j

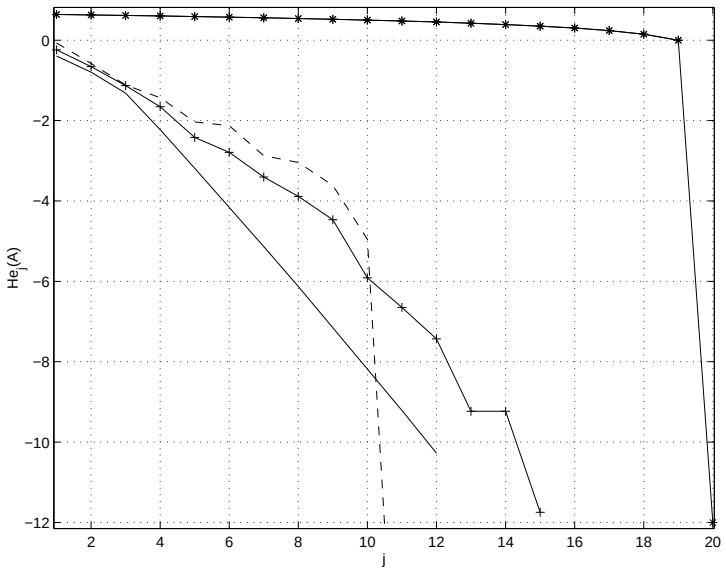

Figure 3: From Example 9 the behavior of $\mathrm{He}_j(A)$ for matrices A_1, A_2, A_3 and $A_4 \in \mathbb{C}^{20\times 20}$ denoted by the solid line, '$--$', '$+-$' and '$*-$', respectively.

if $q_j = 0$, *end*
form the polynomial corresponding to $q_j = 0$
end.

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

It is critical to compute the coefficients of the polynomials p_j accurately in order to generate $p_j(D)\widehat{q}_0$ 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 for iterative methods. Our point of view is weighted by the Krylov subspace methods recently introduced for normal matrices such that a matrix regarded as almost normal if there is a circuitous way to employ these methods for solving linear systems or finding approximations to eigenvalues of the matrix. To this end we have studied binormal matrices, their unitary orbit and, as their natural extension, polynomially normal matrices of moderate degree. We have collected various matrix nearness problems and shown how, e.g., 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

[1] J. Agler, W. Helton and M. Stankus, *Classification of hereditary matrices,* Lin. Alg. Appl., 274 (1998), pp. 125–160.

[2] R. Bhatia, R.A. Horn and F. Kittaneh, *Normal approximants of binormal operators,* Lin. Alg. Appl. 147 (1991), pp. 169–179.

[3] A. Brown, *The unitary equivalence of binormal operators,* Amer. J. Math., 76 (1954), pp. 414–439.

[4] R.N. Chan and M.K. Ng, *Conjugate gradient methods for Toeplitz systems*, SIAM Rev. 38 (1996), pp. 427–482.

[5] J. Conway and B. Morrel, *Roots and logarithms of bounded operators on Hilbert space*, J. Funct. Anal., 70 (1987), pp. 171–193.

[6] T. Eirola, M. Huhtanen and J. von Pfaler, *Solution methods for* $\mathbb{R}$*–linear problems in* $\mathbb{C}^n$*,* manuscript, 2002.

[7] L. Elsner and Kh.D. Ikramov, *On a condensed form for normal matrices under finite sequence of elementary similarities,* Lin. Alg. Appl., 254 (1997), pp. 79–98.

[8] L. Elsner and M.H.C. Paardekooper, *On measures of nonnormality of matrices,* Lin. Alg. Appl., 92 (1987), pp. 107–124.

[9] C.G. Gibson, K Withmüller, A.A. du Plessis and E.J.N Looijenga, *Topological Stability of Smooth Mappings,* Springer-Verlag, New York, 1975.

[10] R. Gow, *The equivalence of an invertible matrix to its transpose,* Lin. Multilin. Alg. 8 (1979/80), pp. 329–336.

[11] A. Greenbaum, Iterative Methods for Solving Linear Systems, *Frontiers in Applied Mathematics, SIAM, Philadelphia,* 1997.

[12] A. Greenbaum and L.N. Trefethen, *GMRES/CR and Arnoldi/Lanczos as matrix approximation problems.* SIAM J. Sci. Comp., 15 (1994), pp. 348–358.

[13] A. Greenbaum and Z. Strakos, *Matrices that generate same Krylov residual spaces,* Recent Advances in Iterative Methods, G. Golub, A. Greenbaum and M. Luskin eds. IMA volumes in mathematics and its applications, Vol 60, Springer-Verlag, New York, 1994.

[14] R. Grone, C.R. Johnson, E.M. Sa and H. Wolkowicz, *Normal matrices,* Lin. Alg. Appl., 87 (1987), pp. 213–225.

[15] R. Hardt, *Stratification of real analytic sets and images,* Inventiones Math., 28 (1975), pp. 193–208.

[16] P. Henrici, *Bounds for iterates, inverses, spectral variation and field of values of nonnormal matrices,* Numer. Math., 4 (1962), pp. 24–40.

[17] D. Herrero and S. Szarek, *How well can an $n \times n$ matrix be approximated by reducible ones?,* Duke Math. J., 53 (1986), pp. 233–248.

[18] A. Hodel, K. Poolla and B. Tenison, *Numerical solution of the Lyapunov equation by approximate power iteration,* Lin. Alg. Appl., 236 (1996), pp. 205–230.

[19] R.A. Horn and C.R. Johnson, Topics in Matrix Analysis, *Cambridge Univ. Press* 1991.

[20] M. Huhtanen, *A stratification of the set of normal matrices,* SIAM J. Matrix Anal. Appl., 23 (2001), pp. 349–367.

[21] M. Huhtanen, *A matrix nearness problem related to iterative methods*, SIAM J. Numer. Anal., 39 (2001), pp. 407–422.

[22] M. Huhtanen, *A Hermitian Lanczos method for normal matrices*, SIAM J. Matrix Anal. Appl., 23 (2002), pp. 1092-1108.

[23] M. Huhtanen, *Orthogonal polyanalytic polynomials and normal matrices,* Math. Comp., to appear.

[24] M. Huhtanen and R.M. Larsen, *Exclusion and inclusion regions for the eigenvalues of a normal matrix*, SIAM J. Matrix Anal. Appl., 23 (2002), pp. 1070-1091.

[25] M. Huhtanen and O. Nevanlinna, *Minimal decompositions and iterative methods,* Numer. Math., 86 (2000), pp. 257–281.

[26] M. Huhtanen, *Combining normality with the FFT techniques*, manuscript, 2002.

[27] Kh.D. Ikramov and L. Elsner, *On matrices that admit a unitary reduction to band form,* Math. Notes, 64 (1998), 753–760.

[28] F. Kittaneh, *On the structure of polynomially normal operators,* Bull. Austral. Math. Soc., 30 (1984), pp. 11–18.

[29] F. Kittaneh, *On normality of operators,* Rev. Roum. Math. Pures Appl., 29 (1984), pp. 703–705.

[30] S.A Mccullough and L. Rodman, *Hereditary classes of operators and matrices,* Amer. Math. Monthly, 104 (1997), pp. 415-430.

[31] N.M. Nachtigal , S. Reddy and L.N. Trefethen, *"How fast are nonsymmetric matrix iterations?"*, SIAM J. Matrix Anal. Appl., 3 (1992), pp. 778–795.

[32] T. Penzl, *A cyclic low-rank Smith method for large sparse Lyapunov equations,* SIAM J. Sci. Comp., 21 (2000), pp. 1401–1418.

[33] J. Phillips, *Nearest normal approximation for certain operators,* Proc. Amer. Math. Soc. 67 (1977), pp. 236–240.

[34] H. Radjavi and P. Rosenthal, *On roots of normal operators,* J. Math. Anal. Appl., 34 (1971), pp. 653–664.

[35] W. Rudin, Functional Analysis, *McGraw-Hill, New York,* 1973.

[36] Y. Saad and M.H. Schultz, *GMRES: A generalized minimum residual algorithm for solving nonsymmetric linear systems,* SIAM J. Sci. Stat. Comp., 7 (1986), pp. 856–869.

[37] Y. Saad and H. van der Vorst, *Iterative solution of linear systems in the 20th century,* J. Comput. Appl. Math., 123 (2000), pp. 1–33.

[38] H. Shapiro, *A survey of canonical forms and invariants for unitary similarity,* Lin. Alg. Appl., 147 (1991), pp. 101–167.

[39] O. Taussky, *On the variation of the characteristic roots of a finite matrix under various changes of its elements,* 1964 Recent Advances in Matrix Theory (Proc. Advanced Seminar, Math. Res. Center, U.S. Army, Univ. Wisconsin, Madison, Wis., 1963), pp. 125–138.

[40] P.Y. Wu, *The operator factorization problems,* Lin. Alg. Appl., 117 (1989), pp. 35–63.

[41] A. Zalewska-Mitura and J. Zemanek, *The Gerschgorin discs under unitary similarity,* Linear operators (Warsaw, 1994), 427–441, Banach Center Publ., 38, Polish Acad. Sci., Warsaw, (1997), pp. 427–441

(continued from the back cover)

A446 Tuomas Hytönen
R-Boundedness is Necessary for Multipliers on H^1
February 2002

A445 Philippe Clment , Stig-Olof Londen , Gieri Simonett
Quasilinear Evolutionary Equations and Continuous Interpolation Spaces
March 2002

A444 Tuomas Hytönen
Convolutions, Multipliers and Maximal Regularity on Vector-Valued Hardy Spaces
December 2001

A443 Tuomas Hytönen
Existence and Regularity of Solutions of the Korteweg - de Vries Equations and Generalizations
December 2001

A442 Ville Havu
Analysis of Reduced Finite Element Schemes in Parameter Dependent Elliptic Problems
December 2001

A441 Jukka Liukkonen
Data Reduction and Domain Truncation in Electromagnetic Obstacle Scattering
October 2001

A440 Ville Turunen
Pseudodifferential calculus on compact Lie groups and homogeneous spaces
September 2001

A439 Jyrki Piila , Juhani Pitkäranta
On corner irregularities that arise in hyperbolic shell membrane theory
July 2001

A438 Teijo Arponen
The complete form of a differential algebraic equation
July 2001

A437 Viking Högnäs
Nonnegative operators and the method of sums
June 2001

A436 Ville Turunen
Pseudodifferential calculus on compact homogeneous spaces
June 2001

HELSINKI UNIVERSITY OF TECHNOLOGY INSTITUTE OF MATHEMATICS
RESEARCH REPORTS

The list of reports is continued inside. Electronical versions of the reports are available at *http://www.math.hut.fi/reports/* .

A451 Marko Huhtanen
Combining normality with the FFT techniques
September 2002

A450 Nikolai Yu. Bakaev
Resolvent estimates of ellpitic differential and finite element operators in pairs of function spaces
August 2002

A449 Juhani Pitkäranta
Mathematical and historical reflections on the lowest order finite element models for thin structures
May 2002

A448 Teijo Arponen
Numerical solution and structural analysis of differential-algebraic equations
May 2002

A447 Timo Salin
Quencing extimate for a reaction diffusion equation with weakly singular reaction term
April 2002

ISBN 951-22-6155-3
ISSN 0784-3143
Institute of Mathematics, HUT, 2002