Annales Univ. Sci. Budapest., Sect. Comp. 21 (2004) 209-217

PERTURBED SCHUR DECOMPOSITION APPLIED FOR NORMAL HESSENBERG MATRICES

L. László (Budapest, Hungary)

Abstract. The scaled departure from normality, defined for triangular matrices, is a useful quantity for giving an upper bound for the best normal approximation. Here its definition will be extended to arbitrary matrices with the help of a perturbation analysis. This enables us to investigate the equality case in a recently verified upper bound.

1. Introduction

The Schur decomposition of a matrix is not unique, for instance the following upper triangular matrices given in Horn-Johnson [4]

$$\begin{pmatrix} 1 & 1 & 4 \\ 0 & 2 & 2 \\ 0 & 0 & 3 \end{pmatrix}, \qquad \begin{pmatrix} 2 & -1 & 3\sqrt{2} \\ 0 & 1 & \sqrt{2} \\ 0 & 0 & 3 \end{pmatrix}$$

are unitarily equivalent. Nevertheless, the departure from normality

$$\mathrm{dep}(A) = \left(\|A\|_F^2 - \sum_{i=1}^{n} |\lambda_i|^2 \right)^{1/2}$$

defined by Henrici for an arbitrary n-th order matrix A with eigenvalues $(\lambda_i)_1^n$ is invariant under unitary similarity. (For our 3×3 matrices this quantity is $\sqrt{21}$ – note that for upper triangular A, $\mathrm{dep}^2(A)$ simplifies to $\sum\limits_{i<j} |a_{i,j}|^2$.)

However, there is another quantity, the *scaled* departure from normality, defined originally only for upper triangular matrices as

$$\operatorname{sdep}(A) = \left(\sum_{i<j} \frac{j-i}{j-i+1} |a_{i,j}|^2 \right)^{1/2},$$

which is not unitarily invariant. It occurs in the recently proved bound

$$\nu_F(A) \leq \operatorname{sdep}(A), \tag{1}$$

where

$$\nu_F(A) = \inf\{\|A - Z\|_F : Z \text{ is normal}\}$$

is the distance of A from the normal matrices in Frobenius norm. (The squared scaled departures of the example are $13\frac{1}{6}$ and $13\frac{1}{2}$, resp.)

As regards the history of inequality (1), it was conjectured by the author [7] and became true when Friedland [3] proved the normal completion theorem guessed by Elsner [2]. The case $n = 3$ has been previously settled by Ikramov [5].

Our aim is to discuss the case of equality in (1). To this we need to define the matrix function *sdep* for a non-triangular matrix that is close to a triangular one. We derive the necessary formulae in Section 2 and give an illustrative example for the 2×2 case. In Section 3 then we apply the results for the problem of the best normal approximation.

2. Perturbation of the Schur decomposition

Let A be a complex upper triangular matrix and P be a matrix of the same size. Let us find the Schur form of the perturbation $A \to A + \varepsilon P$, where ε is small! Recall that the Schur decomposition theorem yields the factorisation $A = UTU^*$ with U unitary and T upper triangular for any square matrix A. Hence we take the formula $A + \varepsilon P = U(\varepsilon)T(\varepsilon)U(\varepsilon)^*$, where $U(\varepsilon)$ is unitary and $T(\varepsilon)$ is upper triangular for any ε, and write it into the form $(A + \varepsilon P)U(\varepsilon) = T(\varepsilon)U(\varepsilon)$ to get rid of the conjugate transpose. Our method is similar to that used e.g. in Wilkinson [10] for the Jordan decomposition except that here the unitarity can be utilized, as well. Expanding both $U(\varepsilon)$ and $T(\varepsilon)$ into Taylor series gives

$$(A + \varepsilon P)(I + \varepsilon U_1 + \varepsilon^2 U_2 + \ldots) = (I + \varepsilon U_1 + \varepsilon^2 U_2 + \ldots)(A + \varepsilon A_1 + \varepsilon^2 A_2 + \ldots),$$

where $\{A_i, i = 1, 2, \ldots\}$ are upper triangular. Equating the coefficients of the corresponding powers one obtains

$$A_1 = AU_1 - U_1A + P, \quad A_2 = AU_2 - U_2A + PU_1 - U_1A_1, \ldots$$

Similarly, we can make use of the unitary property of $U(\varepsilon)$, writing

$$(I + \varepsilon U_1 + \varepsilon^2 U_2 + \ldots)^*(I + \varepsilon U_1 + \varepsilon^2 U_2 + \ldots) = I$$

and also equating the corresponding coefficients. As a result we get the following

Theorem 1. *Let A and P be complex square matrices with A upper triangular. Let the eigenvalues (the main diagonal elements) of A be distinct. The locally unique Schur decomposition of $A + \varepsilon P$ is*

$$A + \varepsilon P = U(\varepsilon)T(\varepsilon)U(\varepsilon)^*$$

with convergent power series

$$T(\varepsilon) = \sum_{i=0}^{\infty} \varepsilon^i A_i \quad \textit{and} \quad U(\varepsilon) = \sum_{i=0}^{\infty} \varepsilon^i U_i,$$

where $U_0 = I$, $A_0 = A$, all the $\{A_i\}$ are upper triangular, and the matrices A_k and U_k can be calculated from

$$A_k = AU_k - U_kA + B_k, \quad U_k + U_k^* + C_k = 0$$

by help of matrices

$$B_k = PU_{k-1} - \sum_{j=1}^{k-1} U_j A_{k-j} \quad \textit{and} \quad C_k = \sum_{j=1}^{k-1} U_j^* U_{k-j}, \ k = 1, 2, \ldots$$

Proof. Calculating and equating the coefficients at equal powers of ε is straightforward, for, the matrix U_k on the k-th step can be always determined to give an upper triangular A_k, due to the assumption on the main diagonal elements. As for convergence we refer to Chapter 2 of [10]. Matrices B_k and C_k were introduced only for convenience.

Example 1. The case of 2×2 matrices. Let

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \qquad P = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.$$

For real ε the matrix $A + \varepsilon P$ is examined. (Note that this family is quite general if we allow transformations like adding a scalar multiple of the identity and multiplying by a scalar.) The matrix series converges for $|\varepsilon| < \frac{1}{2}$, and can be summed up to give

$$\begin{pmatrix} 1 & \varepsilon \\ -\varepsilon & 0 \end{pmatrix} = A + \varepsilon P \;\longrightarrow\; T(\varepsilon) = \begin{pmatrix} \frac{1+s}{2} & 2\varepsilon \\ 0 & \frac{1-s}{2} \end{pmatrix},$$

where $s = \sqrt{1-4\varepsilon^2}$. To this, calculate the A_i'-s and take into account the expansion

$$\sqrt{1-4\varepsilon^2} = 1 - 2\varepsilon^2 - 2\varepsilon^4 - 4\varepsilon^6 - 10\varepsilon^8 - 28\varepsilon^{10} - 84\varepsilon^{12} - \dots$$

that is convergent for $|\varepsilon| < \frac{1}{2}$. (For completeness notice that a Schur form for $|\varepsilon| \geq \frac{1}{2}$ also exists, and differs from the above $T(\varepsilon)$ only in the $(1,2)$ position: 2ε is to be replaced by $2\varepsilon/(1+s)$. Of course, this can no more be obtained by the perturbation method above; an independent consideration is needed.)

The situation is similar for P complex. In that case we need another parameter, say h – a complex number with module one –, and define

$$P = \begin{pmatrix} 0 & h \\ -\overline{h} & 0 \end{pmatrix}.$$

For real ε with $|\varepsilon| < \frac{1}{2}$ we have the correspondence

$$\begin{pmatrix} 1 & \varepsilon h \\ -\varepsilon\overline{h} & 0 \end{pmatrix} = A + \varepsilon P \;\longrightarrow\; T(\varepsilon) = \begin{pmatrix} \frac{1+\overline{s}}{2} & 2\varepsilon h \\ 0 & \frac{1-\overline{s}}{2} \end{pmatrix},$$

where $s = \sqrt{1-4\varepsilon^2}$. (The case of $|\varepsilon| \geq \frac{1}{2}$ is similar as before.)

3. Application to normal Hessenberg matrices

The following lemma will be useful for the subsequent considerations.

Lemma. *Let Z be a normal upper Hessenberg matrix of order n, and choose $H = \operatorname{diag}(1, \dots n)$. Then the matrix $A = Z + ZH - HZ$ is upper triangular and*

$$\|A - Z\|_F = \operatorname{sdep}(A).$$

Proof. The elements of A are calculated to be $a_{i,j} = (j-i+1)z_{i,j}$ showing that A is upper triangular. With these formulae both sides of the equality can be rewritten in terms of the $z_{i,j}$-s:

$$\mathrm{sdep}^2(A) = \sum_{i<j} \frac{j-i}{j-i+1}|a_{i,j}|^2 = \sum_{i<j}(j-i)(j-i+1)|z_{i,j}|^2,$$

$$\|A-Z\|_F^2 = \sum_{i<j}(j-i)^2|z_{i,j}|^2 + \sum_{i=1}^{n}|z_{i+1,i}|^2.$$

Hence the equality to be proved is equivalent to

$$\sum_{i=1}^{n}|z_{i+1,i}|^2 = \sum_{i<j}(j-i)|z_{i,j}|^2,$$

which is a consequence of László [6], Lemma 1 in case of Hessenberg matrices.

Example 2. The pair

$$Z = \begin{pmatrix} 1-i & 1+i & -2+2i \\ 1+3i & 0 & 1+2i \\ 0 & -3+2i & 2i \end{pmatrix}, \quad A = \begin{pmatrix} 1-i & 2+2i & -6+6i \\ 0 & 0 & 2+4i \\ 0 & 0 & 2i \end{pmatrix}$$

where Z is normal upper Hessenberg, A is triangular, meets the requirements of the Lemma. We have thus $\|A-Z\|_F^2 = \mathrm{sdep}^2(A) = 62$.

Example 3. The matrix

$$Z = \begin{pmatrix} s & s^2-1 & 2s \\ s^2+1 & s^3 & s^2-1 \\ 0 & s^2+1 & s \end{pmatrix}$$

is normal for any real s. Moreover, it is obviously upper Hessenberg, and for the associated upper triangular matrix $A = Z+ZH-HZ$ we have $\|A-Z\|_F^2 = \mathrm{sdep}^2(A) = 4(s^4+4s^2+1)$.

Remark. Our aim is to find an upper triangular matrix A, for which equality, i.e. $\nu_F(A) = \mathrm{sdep}(A)$ holds in (1). To this we will use the Lemma, another equality of form $\|A-Z\|_F = \mathrm{sdep}(A)$, hence what remains to prove is, that Z is the closest normal matrix to A in the Frobenius norm, or equivalently,

$$\|A-Z\|_F = \nu_F(A)$$

holds for the matrix pair at issue. Recall that the first order necessary condition for Z to be the closest normal matrix to A is nothing else than the relation $A = Z + ZH - HZ$ in the Lemma. As regards the second order condition, that is quite troublesome, cf. Ruhe [9]: "We have found no way to express this in terms of small $n \times n$ matrices." Hence we are now interested in the first order condition, used however for a *neighbourhood* of matrix A.

To do so, let A and Z be as in the Lemma, and define $P = A - Z$. Let $T(\varepsilon)$ be convergent for $|\varepsilon| < \varepsilon_0$. Then by virtue of Theorem 1, for such ε the quantity $\mathrm{sdep}(A + \varepsilon P)$, and at the same time, the function

$$\varepsilon \to \varphi(\varepsilon) = \frac{\|A + \varepsilon P - Z\|^2}{\mathrm{sdep}^2(A + \varepsilon P)}$$

is well-defined. Let us make a *geometric* consideration.

Observation. Assume that $\|A + \varepsilon P - Z\|_F = \nu_F(A + \varepsilon P)$, i.e. Z is a closest normal matrix to $A + \varepsilon P$ for $\varepsilon, |\varepsilon| \le \varepsilon_1 < \varepsilon_0$. Then Theorem 1 implies $\varphi(\varepsilon) \le 1$ for such ε, while $\varphi(0) = 1$ also holds by the Lemma. Hence φ has a local maximum at $\varepsilon = 0$, i.e. φ is locally concave. It turns out that the stationarity property is independent of our assumption!

Theorem 2. *Let Z be normal upper Hessenberg with distinct diagonal elements. By help of the upper triangular matrix $A = Z + ZH - HZ$ with $H = \mathrm{diag}(1, \dots n)$ define the function φ as above. Then $\varphi'(0) = 0$.*

Proof. On the analogy of the definitions

$$< A, B >= \mathrm{Real} \sum_{i,j} a_{i,j} \overline{b_{i,j}} \implies < A, A >= \|A\|_F^2,$$

where the usual scalar product $< \cdot, \cdot >$ induces the Frobenius norm, we introduce a quasi-scalar product $\{\cdot, \cdot\}$ and induced quasi-norm *sdep* by

$$\{A, B\} = \mathrm{Real} \sum_{i<j} \frac{j-i}{j-i+1} a_{i,j} \overline{b_{i,j}} \implies \{A, A\} = \mathrm{sdep}^2(A).$$

By expanding the numerator and denominator of our function up to second order terms we obtain

$$\|A + \varepsilon P - z\|_F^2 = \|A - Z\|_F^2 + 2\varepsilon < P, A - Z > + O(\varepsilon^2),$$

$$\mathrm{sdep}^2(A + \varepsilon P) = \mathrm{sdep}^2(A) + 2\varepsilon\{A, A_1\} + O(\varepsilon^2),$$

using the locally unique (truncated) Schur form

$$A + \varepsilon P \longrightarrow T(\varepsilon) = A + \varepsilon A_1 + O(\varepsilon^2)$$

given by Theorem 1. The statement $\varphi'(0) = 0$ is easily reformulated into

$$\|A - Z\|_F^2 \ \{A, A_1\} = < P, A - Z > \ \{A, A\}.$$

However, $< P, A - Z >= \|A - Z\|_F^2$ by the definition of P, hence it suffices to show that

$$\{A, A - A_1\} = 0.$$

For this we have an identity

$$\{A, A - A_1\} =< [Z, Z^*], V >$$

holding for *arbitrary* upper Hessenberg Z with $A = Z + ZH - HZ$, where V is the lower bidiagonal matrix with

$$V_{i,i} = n - i, \quad V_{i+1,i} = \frac{z_{i+1,i}}{z_{i,i} - z_{i+1,i+1}}.$$

The presence of the commutator of Z and Z^* will imply that the left hand side of the identity equals zero for normal Z indeed. To prove the identity we have to follow the first step of calculating the A_i-s given in Theorem 1. We have $U_1 + U_1^* = 0$, i.e. a skew Hermitian U_1 is determined so that $AU_1 - U_1A + P = A_1$ is upper triangular ($P = A - Z$) – a solvable linear problem. Observe that U_1 is tridiagonal with zero main diagonal elements, and that the subdiagonal, i.e. $(i + 1, i)$ entries of U_1 and V are the same.

Remark. Calculations with MATLAB and Maple show that more is true: the assumption $P = A - Z$ in the above theorem can be omitted! This means that if we consider the function φ defined for P (not only for ε), then the derivative of this function is zero, i.e. the equality

$$\|A - Z\|_F^2 \ \{A, A_1\} - < P, A - Z > \{A, A\} = 0$$

holds for *any* P. This is a homogeneous linear (in P) form, for A_1 also depends linearly on P. This shows how crucial is the normality of Z.

Summary. Our aim was to show that inequality (1) is attainable, i.e. there are triangular matrices A with equality in (1). For a special set of pairs $\{A, Z\}$ with A triangular, Z Hessenberg normal we could show (cf. the Lemma and Theorem 2) that $\varphi(0) = 1$ and $\varphi'(0) = 0$ hold for an appropriately defined function φ. One would guess that if φ is locally concave at zero then Z is

the closest normal matrix to A. Unfortunately this is not the case, therefore further investigation is necessary. Nevertheless we hope that present approach is a step in the good direction: to prove equality in (1) and, in a general sense, to discover more inner properties of normal matrices.

Finally we mention that there is another scaled departure,

$$\Big(\sum_{i<j} \frac{j-i}{n+j-i-1}|a_{i,j}|^2\Big)^{1/2},$$

which gives an attainable *lower* bound [6] for $\nu_F(A)$. Using the estimations $\frac{k}{k+1} \le \frac{n-1}{n}$ and $\frac{k}{n+k-1} \ge \frac{1}{n}$ for both scaled departures (with denoting $k = j-i$) we find simple bounds expressed in terms of the departure by Henrici:

$$\frac{1}{n}\ \mathrm{dep}^2(A)\ \le\ \nu_F^2(A) \le \left(1-\frac{1}{n}\right)\ \mathrm{dep}^2(A).$$

However, these bounds are no more attainable, cf. e.g. Barrlund [1] and László [8] for the right hand side.

References

[1] **Barrlund A.,** On a conjecture on the closest normal matrix, *MIA*, **1** (3) (1998), 305-318.

[2] **ElsnerL.,** *private communication*

[3] **Friedland Sh.,** Normal matrices and the completion problem, *SIAM Journal on Matrix Analysis,* **23** (3) (2002), 896-902.

[4] **Horn R.A. and Johnson C.R.,** *Matrix analysis,* Cambridge University Press, 1985.

[5] **Икрамов Х.Д.,** О нормальной дилатации треугольных матриц, *Мат. заметки,* **60** (6) (1996), 861-872.

[6] **László L.,** An attainable lower bound for the best normal approximation, *SIAM J. Matrix Anal. Appl.,* **15** (3) (1994), 1035-1043.

[7] **László L.,** Upper bounds for the best normal approximation, *Mathematica Pannonica,* **9** (1) (1998), 123-129.

[8] **László L.,** A remark on Barrlund's LP method, *MIA* (accepted)

[9] **Ruhe A.,** Closest normal matrix finally found! *BIT,* **27** (1987), 585-598.

[10] **Wilkinson J.H.,** *The algebraic eigenvalue problem,* Oxford University Press, 1965.

(Received July 9, 2002)

L. László
Department of Numerical Analysis
Eötvös Loránd University
Pázmány Péter sétány 1/C.
H-1117 Budapest, Hungary
laszlo@numanal.inf.elte.hu