Mathematica Pannonica
9/1 (1998), 121–129

UPPER BOUNDS FOR THE BEST NORMAL APPROXIMATION

Lajos **László**

Departement of Numerical Analysis, Eötvös Loránd University, H-1088 Budapest, Múzeum krt. 6-8, Hungary

Received: September 1997

MSC 1991: 15 A 45, 15 A 60

Keywords: Normal matrix, departure from normality, Schur form, Kronecker product, singular value.

Abstract: Let A be a complex $n \times n$ matrix, and denote by $\mathcal{N}_n$ the set of normal $n \times n$ matrices. We conjecture a "strong" upper bound for the distance $\|A - \mathcal{N}_n\|_F$ in terms of the Schur form of A, a "weak" consequence of which can be formulated by the eigenvalues $\{\lambda_i\}$ of A only:

$$\|A - \mathcal{N}_n\|_F^2 \le \frac{n-1}{n}\left(\|A\|_F^2 - \sum_{i=1}^{n} |\lambda_i|^2\right).$$

Both the "strong" and the "weak" conjectures will be confirmed by some relevant, interesting in themselves results.

1. Introduction and notations

The following notation will be used.

1. $\mathcal{M}_{n,k}$: the set of complex $n \times k$ matrices; I_n is the identity of $\mathcal{M}_n \equiv \mathcal{M}_{n,n}$.
2. $\mathcal{N}_n, \mathcal{U}_n, \mathcal{T}_n \subset \mathcal{M}_n$: normal, unitary and upper triangular matrices.
3. $\|A\|_F = \left(\sum_{i=1}^n \sum_{j=1}^n |a_{i,j}|^2\right)^{1/2}$: the Frobenius norm of $A \in$ $\in \mathcal{M}_n$.
4. $\mathrm{dep}(A) = (\|A\|_F^2 - \|\Lambda\|_F^2)^{1/2}$: the departure from normality by Henrici with $\Lambda = \mathrm{diag}(\lambda_i)$ being the diagonal matrix of the

eigenvalues of $A \in \mathcal{M}_n$.

5. If $A = UTU^*$ for some $A, T \in \mathcal{M}_n$, $U \in \mathcal{U}_n$ (i.e. A is unitarily equivalent to T), and $T \in \mathcal{T}_n$, then T is said to be a Schur form of A.

We guess the following.

Conjecture 1 (strong version). *If $A \in M_n$ and T is a Schur form of A, then*

$$\|A - \mathcal{N}_n\|_F^2 \le \sum_{i=1}^{n-1} \frac{i}{i+1} \sum_{j=1}^{n-i} |t_{j,i+j}|^2. \tag{1}$$

By $\frac{i}{i+1} \le \frac{n-1}{n}$, $1 \le i \le n-1$ we get from this

Conjecture 2 (weak version). *If $A \in M_n$, then*

$$\|A - \mathcal{N}_n\|_F^2 \le \frac{n-1}{n} \operatorname{dep}^2(A). \tag{2}$$

In the next section we will give reason of why our conjectures can be guessed. To this we will discuss the "dual" problem of maximizing the main diagonal by using unitary equivalence. In Section 3 we prove that the weak conjecture holds in an infinitesimal sense, i.e. when a matrix is close enough to a normal one. In the last section we will derive Conjecture 1 from the more plausible conjecture of L. Elsner, which states that every upper triangular matrix can be completed (below the diagonal) to a normal matrix. There we also list the promising results related to this problem.

2. Motivation by the dual problem

Here we sketch the steps which prove the lower bound in [13] and also enable us to guess the conjectures on the upper bound. The treatment is similar for both cases, there is a difference only in Subsection 2.4.

2.1. Equivalence of two problems

As is known [3], for $A \in \mathcal{M}_n$ both problems

$$\delta(A) = \max\{\|\operatorname{diag} UAU^*\|_F : U \in \mathcal{U}_n\}, \tag{3}$$

$$\nu(A) = \|A - \mathcal{N}_n\|_F = \min\{\|A - N\|_F : N \in \mathcal{N}_n\} \tag{4}$$

are equivalent; in particular the "Pythagorean theorem"

$$\delta^2(A) + \nu^2(A) = \|A\|_F^2$$

holds showing that any bound for (3) yields an opposite bound for the dual problem (4), and vice versa. For instance, the lower bound [13]

$$\|A - \mathcal{N}_n\|_F^2 \geq \frac{1}{n} \operatorname{dep}^2(A) \tag{5}$$

has been proved in such a way.

2.2. Decomposition into diagonals

We wish to represent the diagonal of UAU^* as a matrix-vector product. However, the usual Kronecker product [8] takes the columns (or the rows [11]) of the matrix, while (1) is formulated in terms of the diagonals, therefore we make use of the MATLAB-function *diag*.

Definition. For $A \in \mathcal{M}_n$ let $\operatorname{diag}(A) = (a_{11}, \ldots, a_{nn})^T$, and more generally,

$$\operatorname{diag}(A, i) = (a_{j,i+j})_{j=1}^{n-i} \in \mathbf{C}^{n-i}, \quad 0 \leq i \leq n-1,$$

$$\operatorname{diag}(A, i) = (a_{j-i,j})_{j=1}^{n+i} \in \mathbf{C}^{n+i}, \quad 1-n \leq i \leq 0.$$

Now $\operatorname{diag}(UAU^*)$ can be written as a linear combination

$$\operatorname{diag}(UAU^*) = \sum_{i=1-n}^{n-1} Q_i \ \operatorname{diag}(A, i)$$

of the $\operatorname{diag}(A, i)$-s, where $Q_i \in \mathcal{M}_{n,n-|i|}$, $1-n \leq i \leq n-1$. For instance, $Q_0 = (|u_{ij}|^2) \in \mathcal{M}_n$ is a square, doubly stochastic matrix (used also in connection with majorization), while e.g. Q_{n-1} is a complex one-column matrix.

Observe that the $n \times n^2$ matrix $(Q_{1-n}, \ldots, Q_0, \ldots, Q_{n-1})$ consists of full rows (with appropriately permuted columns) of the Kronecker product $\overline{U} \otimes U$ (or $U \otimes \overline{U}$, if one uses the formalism [11]), which is known to be unitary, thus

$$\sum_{i=1-n}^{n-1} Q_i Q_i^* = I_n. \tag{6}$$

2.3. Passing over to the Schur form

In this section we replace A by its Schur form $T \in \mathcal{T}_n$. This is possible by the Schur decomposition theorem and the unitarily invariance of the Frobenius norm, giving $\|A - \mathcal{N}_n\| = \|T - \mathcal{N}_n\|$. However, for $T \in \mathcal{T}_n$ $\operatorname{diag}(T, i) = 0$, $i < 0$ holds, and the Q_i-s with negative indices disappear:

$$\mathrm{diag}(UTU^*) = \sum_{i=0}^{n-1} Q_i \ \mathrm{diag}(T, i).$$

For brevity, let $t_i = \mathrm{diag}(T, i)$, $i = 0, \ldots, n-1$, and write the above sum as a matrix-vector product Qt with

$$(Q_0, \ldots, Q_{n-1}), \quad t = (t_0^T, \ldots, t_{n-1}^T)^T.$$

Then the equivalent dual form of Conjecture 1 is

$$\sum_{i=0}^{n-1} \frac{1}{i+1} \|t_i\|^2 \le \delta^2(T). \tag{7}$$

2.4. Matrix inequalities via scaling

Let $s = \{s_i\}_{i=0}^{n-1}$ be a sequence of positive numbers, and use them to scale down the matrices $\{Q_i\}$ and the vectors $\{t_i\}$ as follows:

$$Q_s = (s_0 Q_0, \ldots, s_{n-1} Q_{n-1}), \quad t_s = (s_0^{-1} t_0^T, \ldots, s_{n-1}^{-1} t_{n-1}^T)^T.$$

Then $\mathrm{diag}(UTU^*) = Q_s t_s$ holds, independently on s.

In [12] we proved that the choice $s_i = (\frac{n+i-1}{n-1})^{1/2}$, $0 \le i \le n-1$ yields $Q_s Q_s^* \le I_n$, with $\le$ denoting the positive semidefinite ordering, which implies

$$\| \mathrm{diag}(UTU)^* \|^2 \le \|Q_s\|_2^2 \|t_s\|^2 = \sum_{I=0}^{n-1} \frac{n-1}{n+i-1} \|t_i\|^2, \tag{8}$$

$\|Q_s\|_2$ being the operator norm (or spectral norm) of Q_s. Taking the maximum for $U \in \mathcal{U}_n$ yields an upper bound for $\delta^2(T)$. Then reformulating it by using the equivalence of problems (3), (4), one obtains the attainable lower bound [13] for $\nu^2(T)$, see (5) as a special case.

As regards the Conjecture, below we prove that if $s'_i = (i+1)^{1/2}$, $0 \le i \le n-1$, and $s' = \{s'_i\}_{i=0}^{n-1}$, then $Q_{s'} Q_{s'}{}^* \ge I_n$ holds, giving a chance to show (1). Unfortunately, while formerly $Q_s Q_s^* \le I_n$ was enough to prove $\delta^2(T) \le \|t_s\|^2$, now — although the nonzero singular values of $Q_{s'}^* Q_{s'}$ are the same as those of $Q_{s'} Q_{s'}^*$ —, the inequality $Q_{s'} Q_{s'}^* \ge I_n$ does not imply $\delta^2(T) \ge \|t_{s'}\|^2$, i.e. (7). Nevertheless, the guess (7) still seems to be well-founded, since it must hold only for the maximizer in (3), contrary to (8), which holds for *all* $U \in \mathcal{U}_n$.

2.5. Two lemmas on unitaries

Lemma 1. *If $U \in \mathcal{U}_n$, then for the Q_i-s defined above we have*

$$\sum_{k=0}^{n-1} (k+1)\ Q_k Q_k^* \geq I_n.$$

Proof. We use the identities

$$I_n - Q_0 Q_0^* = 2\sum_{k=1}^{n-1} \Re(Q_k Q_k^*), \qquad \sum_{k=1}^{n-1} k\ \Im(Q_k Q_k^*) = 0,$$

and the notations

$$F_k = \Re(Q_k Q_k^*), \qquad G_k = \Im(Q_k Q_k^*), \quad 0 \leq k \leq n-1$$

from [12]. The statement to be proved is then equivalent to the relation

$$\sum_{k=1}^{n-1}(k+1)(F_k + iG_k) \geq 2\sum_{k=1}^{n-1} F_k, \quad \text{or}$$

$$\sum_{k=1}^{n-1}(k-1)F_k + i\sum_{k=1}^{n-1}(k+1)G_k \geq 0,$$

$i = \sqrt{-1}$. By $\sum_{k=1}^{n-1}(k+1)G_k = -\sum_{k=1}^{n-1}(k-1)G_k$ it remains to show that

$$\sum_{k=1}^{n-1}(k-1)(F_k - iG_k) = \sum_{k=1}^{n-1}(k-1)\ \overline{Q_k Q_k^*} \geq 0,$$

which is true, since the matrix at issue is a nonnegative linear combination of positive semidefinite matrices. ◊

The other Lemma is not used here, but is interesting, because it extends $Q_s Q_s^* \leq I_n$ to an equality, by involving the negative diagonals. (See (6) for the "original" unscaled equality.) Since it can be proved by the similar technique, we state it without proof.

Lemma 2. *For the coefficient matrices $\{Q_i\}_{i=1-n}^{n-1}$ generated by $U \in \in \mathcal{U}_n$ it holds that*

$$\sum_{i=1-n}^{n-1} \frac{n+i-1}{n-1} Q_i Q_i^* = I_n.$$

3. Near the closest normal matrix

Let $A \in \mathcal{M}_n$ and $N \in \mathcal{N}_n$ be its closest normal matrix. As a necessary condition, we have then $A = N + NH - HN$ with $H \in \mathcal{H}_n$. Moreover, A becomes in the coordinates of the eigenvectors of N a so-called ΔH matrix:

(9) $\quad B = D + DG - GD, \quad D:$ diagonal , $\quad G:$ Hermitian,
where $N = UDU^*$, $U \in \mathcal{U}_n$ is the spectral decomposition of N, $B = U^*AU$ and $G = U^*HU$. In addition,

(10) $$\|B - D\|_F = \|B - \mathcal{N}_n\|_F$$

also holds, due to the unitarily invariance of the Frobenius norm. Thus it can be assumed without loss of generality that B is a ΔH matrix with his diagonal as closest normal matrix.

Theorem. *Let $B \in \mathcal{M}_n$ be a ΔH matrix* (9) *with the best approximation property* (10). *Suppose that the diagonal elements of B are distinct, and introduce the family $B(\varepsilon) = D + \varepsilon(B - D)$, $0 \le \varepsilon \le 1$. Then we have*

(a) $\|B(\varepsilon) - D\|_F = \|B(\varepsilon) - \mathcal{N}_n\|_F, \quad 0 \le \varepsilon \le 1,$

(b) $\|B(\varepsilon) - D\|_F^2 \le \frac{n-1}{n} \operatorname{dep}^2(B(\varepsilon)) + O(\varepsilon^3), \quad \varepsilon \to 0,$

(c) $\lim_{\varepsilon\to 0} \|B(\varepsilon) - D\|_F^2 / \operatorname{dep}^2(B(\varepsilon)) = \frac{1}{2}.$

Proof. (a) This follows from the geometry of the Euclidean space $\mathcal{M}_n$ provided by the inner product $(X, Y) = \Re(\operatorname{trace}(Y^*X))$, $X, Y \in \mathcal{M}_n$.

(b) Denote by $\{\lambda_i(\varepsilon)\}_{i=1}^n$ the eigenvalues of $B(\varepsilon)$. Consider the formulas (5.5), (9.4) and (11.3) from [15], giving the eigenvalue expansion, the first and the second order terms for $\lambda_1(\varepsilon)$:

$$\lambda_1(\varepsilon) = \lambda_1 + k_1\varepsilon + k_2\varepsilon^2 + \ldots; \qquad k_1 = \frac{\beta_{11}}{s_1}; \qquad k_2 = \frac{1}{s_1}\sum_{i=2}^{n} \frac{\beta_{i1}\beta_{1i}}{s_i(\lambda_1 - \lambda_i)}.$$

Since $B(\varepsilon)$ is a nondiagonal perturbation of the diagonal matrix D, we have $k_1 = 0$ and $s_i = 1$ for all i, giving

$$\lambda_i(\varepsilon) = b_{i,i} + \sum_{j\ne i} \frac{b_{i,j}b_{j,i}}{b_{i,i} - b_{j,j}} + O(\varepsilon^3), \quad 1 \le i \le n.$$

Using the ΔH structure of $B(\varepsilon)$, $\lambda_i(\varepsilon)$ assumes

$$\lambda_i(\varepsilon) = d_{ii} - \sum_{j\ne i} |g_{i,j}|^2 (d_{ii} - d_{jj}) + O(\varepsilon^3), \quad 1 \le i \le n,$$

whence

(11) $$\|\Lambda(\varepsilon)\|_F^2 = \|D\|_F^2 - \varepsilon^2\|B - D\|_F^2 + O(\varepsilon^3)$$

can be derived. This implies

$$\|B(\varepsilon) - D\|_F^2 = \varepsilon^2\|B - D\|_F^2 = \|D\|_F^2 - \|\Lambda(\varepsilon)\|_F^2 + O(\varepsilon^3),$$

therefore (b) is equivalent with

$$\|D\|_F^2 - \|\Lambda(\varepsilon)\|_F^2 \le \frac{n-1}{n}(\|B(\varepsilon)\|_F^2 - \|\Lambda(\varepsilon)\|_F^2) + O(\varepsilon^3),$$

i.e. with

$$\frac{1}{n}(\|B(\varepsilon)\|_F^2 - \|\Lambda(\varepsilon)\|_F^2) \le \|B(\varepsilon) - D\|_F^2 + O(\varepsilon^3),$$

which is a consequence of (5).

(c) To this rewrite (11) into

$$\|B(\varepsilon)\|_F^2 - \|\Lambda(\varepsilon)\|_F^2 = 2(\|D\|_F^2 - \|\Lambda(\varepsilon)\|_F^2) + O(\varepsilon^3),$$

and use

$$\frac{\|B(\varepsilon) - D\|_F^2}{\|B(\varepsilon)\|_F^2 - \|\Lambda(\varepsilon)\|_F^2} = \frac{\|D\|_F^2 - \|\Lambda(\varepsilon)\|_F^2 + O(\varepsilon^3)}{2(\|D\|_F^2 - \|\Lambda(\varepsilon)\|_F^2 + O(\varepsilon^3)}. \diamond$$

4. Conjecture by conjecture

Here we show that the following conjecture by L. Elsner implies Conjecture 1.

Conjecture (on normal completion [4]). Any upper triangular matrix can be completed to a normal matrix by specifying appropriately the entries under the diagonal. In short: any $T \in \mathcal{T}_n$ has the property of normal completion.

For $B \in \mathcal{M}_n$ let $b_i = \mathrm{diag}(B, i)$ be the column vector formed from the elements of the i-th diagonal of B, $1 - n \le i \le n - 1$, as introduced in 2.2. We write $B = \{b_i\}_{1-n}^{n-1}$ for $B \in \mathcal{M}_n$ arbitrary, and $T = \{t_i\}_0^{n-1}$ for $T \in \mathcal{T}_n$.

Lemma 3. *Given an upper triangular $T = \{t_i\}_0^{n-1} \in \mathcal{T}_n$, create the scaled matrix $T_s = \{t_i/(i+1)\}_0^{n-1} \in \mathcal{T}_n$. If T_s has the property of normal completion, then T (and any $A \in \mathcal{M}_n$, unitarily equivalent to T) satisfies* (1).

Proof. By definition, $b_i = t_i/(i+1)$, $i \ge 0$, therefore we have

$$\|T - B\|_F^2 = \sum_{i=1-n}^{-1} \|b_i\|^2 + \sum_{i=0}^{n-1} \left(\frac{i}{i+1}\right)^2 \|t_i\|^2.$$

This is less than or equal to the right hand side of (1) if and only if

$$\sum_{i=1-n}^{-1} \|b_i\|^2 \le \sum_{i=0}^{n-1} \frac{i}{(i+1)^2} \|t_i\|^2$$

holds. However, Lemma 1 [13] implies

$$\sum_{i=1-n}^{-1} \|b_i\|^2 \le \sum_{i=1-n}^{-1} |i| \; \|b_i\|^2 = \sum_{i=1}^{n-1} i\|b_i\|^2,$$

hence, using again $b_i = t_i/(i+1)$, $i \ge 0$ completes the proof. $\diamond$

Remark. In order to prove the analogous implication (cf. [4]) for the weak version (2), it suffices to require that the matrix $T_D + T_U/n$ has the property on normal completion, where T_D is the main diagonal of $T \in \mathcal{T}_n$, and $T_U = T - T_D$.

Remark. Note that in case of $n = 2$ inequalities (1) and (2) coincide and are true [3], [7]. Further, since the conjecture on normal completion is true for $n = 3$ [9], [10], Conjecture 1 also holds in this case. For this reason we display the known bounds for $n = 3$ in detail. For $T \in \mathcal{T}_3$ we have

$$\frac{1}{3}(|t_{1,2}|^2+|t_{2,3}|^2)+\frac{1}{2}|t_{1,3}|^2 \le \|T-\mathcal{N}_3\|_F^2 \le \frac{1}{2}(|t_{1,2}|^2+|t_{2,3}|^2)+\frac{2}{3}|t_{1,3}|^2.$$

The right hand side is the result obtained from [9], [10] by Lemma 3, while the left hand inequality is known [13].

Remark. Recently considerable steps have been taken in proving Conjecture 2. A. Barrlund [1] proved an inequality with $(n-1/2)/n$ instead of $(n-1)/n$. He used an appropriate permutation of the upper triangular T followed by normal approximations of the 2×2 blocks along the diagonal. L. Elsner and Kh. D. Ikramov [6] obtained a further refinement by help of 3×3 normal matrices.

Finally, A. Barrlund [2] proved that (2) is correct for all even n and for $n = 3, 5, 7$. He also obtained a bound for odd n which converges to the bound given in the weak conjecture, when n tends to infinity. Moreover, he derived sharper bounds for $n = 3, 5, 6, 7$ by help of an interesting LP technique.

Acknowledgement. I am grateful to Prof. Anders Barrlund and Prof. Ludwig Elsner for sending me all informations about the recent development of the subject.

References

[1] BARRLUND, A.: The proof of a modified form of the weak bound by block 2×2 matrices, private communication.

[2] BARRLUND, A.: On a conjecture on the closest normal matrix, submitted to *Mathematical Inequalities and Applications.*

[3] CAUSEY, R. L.: On Closest Normal Matrices, Tech. Report CS-10, Dept. of Computer Science, Stanford University, Stanford, CA, 1964.

[4] ELSNER, L.: Proving the weak conjecture by help of the conjecture on normal completion, private communication.

[5] ELSNER, L. and PAARDEKOOPER, M.H.C.: On measures of nonnormality of matrices, *Linear Algebra Appl.* **92** (1987), 107–124.

[6] ELSNER, L. and IKRAMOV, Kh. D.: Improving Barrlunds bound, private communication.

[7] GABRIEL, R.: Matrizen mit maximaler Diagonale bei unitärer Similarität, *J. Reine Angew. Math.* **307/308** (1979), 31–52.

[8] HORN, R.A. and JOHNSON, C.R.: Topics in Matrix Analysis, Cambridge University Press, 1991.

[9] IKRAMOV, Kh.D.: On normal completions of triangular matrices, *Doklady Akademii Nauk* **351** (1996), 1–2 (in Russian).

[10] IKRAMOV, Kh.D.: On normal dilations of triangular matrices, *Mathematical notes*, **60** N6(1996) (in Russian).

[11] LANCASTER, P.: Theory of matrices, Academic Press, 1969.

[12] LÁSZLÓ, L.: Upper bounds for matrix diagonals, *Linear and Multilinear Algebra* **30** (1991), 283–301.

[13] LÁSZLÓ, L.: An attainable lower bound for the best normal approximation, *Siam J. Matrix Anal. Appl.* **15**/3 (1994), 1035–1043.

[14] RUHE, A.: Closest normal matrix finally found!, *BIT* **27** (1987), 585–598.

[15] WILKINSON, J.H.: The Algebraic Eigenvalue Problem, Clarendon Press, Oxford, 1965.