The structured distance to normality of an irreducible real tridiagonal matrix*

S. Noschese, L. Pasquini and L. Reichel

Abstract

The problem of computing the distance in the Frobenius norm of a given real irreducible tridiagonal matrix T to the set of real normal irreducible tridiagonal matrices is solved. Simple formulas for evaluating the distance and for computing the closest real normal irreducible matrix are presented. The special case of tridiagonal Toeplitz matrices is also considered.

*Submitted

- **Matrix nearness problems**

(symmetry, positive definiteness, orthogonality, normality, defectiveness, rank-deficiency, instability) [Henrici 62] [Wilkinson 65, 72, 84a, 84b, 86] [Ruhe 75, 87] [Demmel 87,90] [Higham 89].

- **Role played by the structure**

[Higham & Higham 92, 98] [Chaitin-Chatelin et al. 96, 00] [Tisseur 03] [Byers & Kressner 04] [Bini et al. 05] [N. & Pasquini 06, 07] [Rump 06].

- **Distance to normality**

[Henrici 62] [Causey 64] [Ruhe 70, 75, 87] [Gabriel 79, 87] [Elsner & Paardekooper 87] [László 94] [Lee 95, 96] [Ipsen 03].

- **Structured distance to normality**

Irreducible real tridiagonal case. Irreducible real tridiagonal Toeplitz case.

OUTLINE

- Upper bounds to $d_F(T,\mathcal{N})$ and to $d_F(T,\mathcal{N}_\mathcal{T})$
- Real normal tridiagonal matrices
- Monotype matrices
- Computation of $d_F(T,\mathcal{M})$
- The structured distance
- Toeplitz case
- Examples
- Report on the tests

- **Upper bounds to $d_F(T,\mathcal{N})$ and to $d_F(T,\mathcal{N}_{\mathcal{T}})$**

$$(A,B) = \mathrm{trace}\,(B^T A), \qquad A, B \in \mathbf{R}^{n\times n}$$

$$\Downarrow$$

Theorem *$\mathcal{S}$ and $\mathcal{A}$ are orthogonal.*

$$\Downarrow$$

$$d_F(A,\mathcal{N}) \le \min\{d_F(A,\mathcal{S}), d_F(A,\mathcal{A})\} \le \frac{\sqrt{2}}{2}\|A\|_F$$

$$d_F(T,\mathcal{N}_{\mathcal{T}}) \le \min\{d_F(T,\mathcal{S}_{\mathcal{T}}), d_F(T,\mathcal{A}_{\mathcal{T}})\} \le \frac{\sqrt{2}}{2}\|T\|_F$$

- The right-hand side inequality is achieved iff

$$\sum_{i=1}^{n} \delta_i^2 + 2\sum_{i=1}^{n-1} \sigma_i\tau_i = 0.$$

- Equality $d_F(T,\mathcal{N}_{\mathcal{T}}) = \frac{\sqrt{2}}{2}\|T\|_F$ can be achieved.

$T = (n; \sigma, \delta, \tau)$ denotes the real tridiagonal matrix

$$T = \begin{bmatrix} \delta_1 & \tau_1 & & & & \mathbf{0} \\ \sigma_1 & \delta_2 & \tau_2 & & & \\ & \sigma_2 & \delta_3 & & & \\ & & & \ddots & & \\ & & \ddots & \ddots & \tau_{n-2} & \\ & & & \sigma_{n-2} & \delta_{n-1} & \tau_{n-1} \\ \mathbf{0} & & & & \sigma_{n-1} & \delta_n \end{bmatrix} \in \mathbf{R}^{n \times n}.$$

- **Real normal tridiagonal matrices**

Theorem $T = (n; \sigma, \delta, \tau)$ *is normal iff it is block diagonal, with each block either a diagonal block or an irreducible tridiagonal block* $T^* = (\nu; \sigma^*, \delta^*, \tau^*)$, *whose entries satisfy one of the conditions:*
i) $\sigma_h^* = \ \tau_h^*, \ h = 1 : \nu - 1$
ii) $\sigma_h^* = -\tau_h^*, \ h = 1 : \nu - 1; \ \ \delta_1^* = \delta_2^* = \ldots = \delta_\nu^*.$

- **Monotype matrices**

Definition T *is a monotype matrix of type j) if its off-diagonal entries satisfy*

$$\sigma_h = \tau_h, \qquad h = 1 : n - 1.$$

T *is a monotype of type jj) if its entries satisfy*

$$\sigma_h = -\tau_h, \ h = 1 : n - 1, \ \ \delta_1 = \delta_2 = \ldots = \delta_n.$$

- $\mathcal{M}^{(j)}$ and $\mathcal{M}^{(jj)}$ are closed subsets of $\mathcal{N}_{\mathcal{T}}$.

- $\mathcal{M} = \mathcal{M}^{(j)} \cup \mathcal{M}^{(jj)}$.

- **Computation of** $d_F(T, \mathcal{M})$

$$d_F(T, \mathcal{M}^{(j)}) = \sqrt{\frac{1}{2}\sum_{i=1}^{n-1}(\sigma_i - \tau_i)^2}.$$

$$d_F(T, \mathcal{M}^{(jj)}) =$$

$$\sqrt{\frac{1}{2}\sum_{i=1}^{n-1}(\sigma_i + \tau_i)^2 + \sum_{i=1}^{n}\left(\delta_i - \frac{\sum_{j=1}^{n}\delta_j}{n}\right)^2}.$$

Theorem *The distance in the Frobenius norm* $d_F(T, \mathcal{M})$ *between* $T = (n; \sigma, \delta, \tau)$ *and* $\mathcal{M}$ *is*

$$d_F(T, \mathcal{M}) = \min\left\{d_F(T, \mathcal{M}^{(j)}), d_F(T, \mathcal{M}^{(jj)})\right\}.$$

- **The structured distance**

Theorem $d_F(T, \mathcal{I}) = d_F(T, \mathcal{M})$

Proof. The set of monotype matrices $\mathcal{M}$ is closed, because it is the union of the closed sets $\mathcal{M}^{(j)}$ and $\mathcal{M}^{(jj)}$.
A real irreducible normal tridiagonal matrix is a monotype matrix. Thus, $\mathcal{I} \subset \mathcal{M}$.
In every neighborhood of any matrix T in $\mathcal{M}^{(j)}$ [in $\mathcal{M}^{(jj)}$] there is a real normal irreducible tridiagonal matrix $T^* \neq T$ of type j) [of type jj)]. Thus, $\mathcal{M}$ is the closure of $\mathcal{I}$. $\diamondsuit$

Note that when $\sigma_i^2 \neq \tau_i^2$, $i = 1 : n-1$, the closest matrix to T in $\mathcal{M}$, $T^{(j)}$ or $T^{(jj)}$, is irreducible. Otherwise, at least one of the matrices $T^{(j)}$ and $T^{(jj)}$ is reducible and the closest matrix to T in $\mathcal{M}$ may be reducible.

• **Toeplitz case**

Theorem *A real tridiagonal Toeplitz matrix is normal if and only if its entries satisfy* $\sigma = \tau$ *or* $\sigma = -\tau$.

Theorem *Let* T *be a real tridiagonal Toeplitz matrix. The closest real tridiagonal Toeplitz matrix* $T^{(j)}$ *to* T *in the set* $\mathcal{M}^{(j)}$ *has diagonal entries* δ*, and sub- and super-diagonal entries* $\frac{1}{2}(\sigma + \tau)$*. The closest real tridiagonal Toeplitz matrix* $T^{(jj)}$ *to* T *in the set* $\mathcal{M}^{(jj)}$ *has sub-diagonal entries* $\frac{1}{2}(\sigma - \tau)$*, diagonal entries* δ*, and super-diagonal entries* $-\frac{1}{2}(\sigma - \tau)$*.*

$$d_F(T, \mathcal{M}) = \sqrt{\frac{n-1}{2}} \min\{|\sigma - \tau|, |\sigma + \tau|\}.$$

is the $\mathcal{T}$*-patterned* distance of real irreducible tridiagonal Toeplitz matrix from normality.

• **Example 1**

Consider the quasi-Jordan block

$$J_n = \begin{bmatrix} \lambda & \mu & 0 & \dots & \dots & 0 & 0 \\ 0 & \lambda & \mu & 0 & \dots & 0 & 0 \\ 0 & 0 & \lambda & \mu & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & \lambda & \mu & 0 \\ 0 & 0 & 0 & \dots & \dots & \lambda & \mu \\ 0 & 0 & 0 & \dots & \dots & \dots & \lambda \end{bmatrix} \in \mathbf{R}^{n\times n}.$$

Thus, J_n is an upper bidiagonal Toeplitz matrix. Assume that $\mu \neq 0$. Then the closest normal matrix to J_n is the circulant matrix

$$\begin{bmatrix} \lambda & \frac{n-1}{n}\mu & 0 & \dots & \dots & 0 & 0 \\ 0 & \lambda & \frac{n-1}{n}\mu & 0 & \dots & 0 & 0 \\ 0 & 0 & \lambda & \frac{n-1}{n}\mu & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & \lambda & \frac{n-1}{n}\mu & 0 \\ 0 & 0 & 0 & \dots & \dots & \lambda & \frac{n-1}{n}\mu \\ \frac{n-1}{n}\mu & 0 & 0 & \dots & \dots & \dots & \lambda \end{bmatrix}$$

On the other hand the closest matrices to J_n in $\mathcal{M}^{(j)}$ and $\mathcal{M}^{(jj)}$ respectively are

$$\begin{bmatrix} \lambda & \mu/2 & 0 & \dots & \dots & 0 & 0 \\ \mu/2 & \lambda & \mu/2 & 0 & \dots & 0 & 0 \\ 0 & \mu/2 & \lambda & \mu/2 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \mu/2 & \lambda & \mu/2 & 0 \\ 0 & 0 & 0 & \dots & \mu/2 & \lambda & \mu/2 \\ 0 & 0 & 0 & \dots & \dots & \mu/2 & \lambda \end{bmatrix}$$

and

$$\begin{bmatrix} \lambda & \mu/2 & 0 & \dots & \dots & 0 & 0 \\ -\mu/2 & \lambda & \mu/2 & 0 & \dots & 0 & 0 \\ 0 & -\mu/2 & \lambda & \mu/2 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & -\mu/2 & \lambda & \mu/2 & 0 \\ 0 & 0 & 0 & \dots & -\mu/2 & \lambda & \mu/2 \\ 0 & 0 & 0 & \dots & \dots & -\mu/2 & \lambda \end{bmatrix}$$

Both the matrices are irreducible.

$$\frac{d_F(J_n, \mathcal{N})}{\|J_n\|_F} = \frac{1}{\sqrt{n\left(1 + \frac{n}{n-1}\left(\frac{\lambda}{\mu}\right)^2\right)}}$$

is a poor indicator of the conditioning of the eigenvalue problem. It goes to zero when n increases for any real ratio λ/μ.

$$\frac{d_F(J_n, \mathcal{N}_{\mathcal{T}})}{\|J_n\|_F} = \frac{d_F(J_n, \mathcal{I})}{\|J_n\|_F} = \frac{1}{\sqrt{2\left(1 + \frac{n}{n-1}\left(\frac{\lambda}{\mu}\right)^2\right)}}$$

provides a better measure of the sensitivity of the eigenvalue problem. It converges to its maximum value $\sqrt{2}/2$, when the ratio $|\lambda|/|\mu|$ approaches 0.

• **Example 2**

$$T = \begin{bmatrix} 0 & 1 & 0 & 0 \\ -1 & 0 & \varepsilon & 0 \\ 0 & \varepsilon & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

with $\varepsilon > 0$ a tiny parameter.

$$\frac{d_F(T, \mathcal{I})}{||T||_F} = \frac{\sqrt{2}}{2} \frac{1}{\sqrt{1 + \frac{\varepsilon^2}{2}}}.$$

The normalized distance to $\mathcal{I}$ is close to maximal for $\varepsilon > 0$ small.

$$\frac{d_F(T, \mathcal{N}_{\mathcal{T}})}{||T||_F} \leq \frac{\varepsilon}{\sqrt{2 + \varepsilon^2}}.$$

The normalized distance to $\mathcal{N}_{\mathcal{T}}$ is small in this situation.

- **Report on the tests**

Distance $\Leftrightarrow$ Eigenvalue Conditioning?

Structured Distance *small*
$\Downarrow$
Structured Condition numbers *small*

$\mathcal{T}$-Patterned Distance *small*
$\Downarrow$
$\mathcal{T}$-Patterned Condition numbers *small*