GALN09, Perugia, 16-17 febbraio 2009

STRUCTURED DISTANCE TO NORMALITY OF BANDED TOEPLITZ MATRICES

S. Noschese and L. Reichel

Outline

- Normal Toeplitz matrices [5, 6, 4]
- $(2k+1)$-banded Toeplitz matrices - Notation
- Normal banded Toeplitz matrices, with $k \leq \lfloor n/2 \rfloor$
  - Spectral properties
  - The real case [3, 1]
- Large banded nonnormal Toeplitz matrices [2]
- Distance to the algebraic variety $\mathcal{N}$ of normal matrices [7, 10]
- Structured distance to normality
  - The closest structured normal matrix
  - The real case
- The tridiagonal case. Structured distance to normality
  - Sensitivity of the eigenvalues
  - The $\varepsilon-$pseudospectrum [9]
- Unstructured versus structured. An example [8]

References

[1] D. Bini and M. Capovani, *Spectral and computational properties of band symmetric Toeplitz matrices*, Linear Algebra Appl., 52/53 (1983), pp. 99–126.

[2] A. Böttcher and S. Grudsky, *Spectral Properties of Banded Toeplitz Matrices*, SIAM, Philadelphia, 2005.

[3] A. Cantoni and P. Butler, *Eigenvalues and eigenvectors of symmetric centrosymmetric matrices*, Linear Algebra Appl., 13 (1976), pp. 275–288.

[4] P. J. Davis, *Circulant Matrices*, 2nd ed., Chelsea, New York, 1994.

[5] D. R. Farenick, M. Krupnik, N. Krupnik, and W. Y. Lee, *Normal Toeplitz matrices*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 1037–1043.

[6] C. Gu and L. Patton, *Commutation relations for Toeplitz and Hankel matrices*, SIAM J. Matrix Anal. Appl., 24 (2003), pp. 728–746.

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

[8] S. Noschese, L. Pasquini, and L. Reichel, *The structured distance to normality of an irreducible real tridiagonal matrix*, Electron. Trans. Numer. Anal., 28 (2007), pp. 65–77.

[9] L. Reichel and L. N. Trefethen, *Eigenvalues and pseudo-eigenvalues of Toeplitz matrices*, Linear Algebra Appl., 162–164 (1992), pp. 153–185.

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

1. Normal Toeplitz matrices

Normal Toeplitz matrices are characterized in, e.g., [5, 6].

$$T = \begin{bmatrix} \delta & \tau_1 & \tau_2 & & \cdots & & \tau_{n-1} \\ \sigma_1 & \delta & \tau_1 & \tau_2 & \cdots & & \tau_{n-2} \\ \sigma_2 & \sigma_1 & & \ddots & & & \vdots \\ \vdots & & \ddots & \ddots & & & \\ & & & & & \tau_1 & \tau_2 \\ \sigma_{n-2} & & \cdots & & \sigma_1 & \delta & \tau_1 \\ \sigma_{n-1} & & \cdots & & \sigma_2 & \sigma_1 & \delta \end{bmatrix} \in \mathbb{C}^{n\times n}$$

is normal if and only if $\sigma_\ell = e^{i\varphi}\bar{\tau}_\ell$, $1 \leq \ell < n$, or $\sigma_\ell = e^{i\varphi}\tau_{n-\ell}$, $1 \leq \ell < n$, for some $-\pi < \varphi \leq \pi$.

Thus, normal Toeplitz matrices are either modifications of Hermitian matrices or are so-called $\{e^{i\varphi}\}$-circulant matrices. Properties of the latter are discussed in, e.g., [4, Section 3.2.1].

2. $(2k+1)$-banded Toeplitz matrices - Notation

$$T_{(k)} = (n;k;\sigma,\delta,\tau) = \begin{bmatrix} \delta & \tau_1 & \tau_2 & \dots & \tau_k & & & 0 \\ \sigma_1 & \delta & \tau_1 & & & & & \\ \sigma_2 & \sigma_1 & \ddots & & & & \ddots & \\ \vdots & & & & \ddots & & & \tau_k \\ \sigma_k & & \ddots & \ddots & \ddots & & & \vdots \\ & & & & & & \tau_1 & \tau_2 \\ & & \ddots & & & \sigma_1 & \delta & \tau_1 \\ 0 & & & \sigma_k & \dots & \sigma_2 & \sigma_1 & \delta \end{bmatrix} \in \mathbb{C}^{n\times n}$$

$$\mathcal{N}_{\mathcal{T}_{(k)}} = \mathcal{N} \cap \mathcal{T}_{(k)}, \qquad \mathcal{N}^{\mathbb{R}}_{\mathcal{T}_{(k)}} = \mathcal{N}_{\mathcal{T}_{(k)}} \cap \mathbb{R}^{n\times n}$$

$$\begin{aligned} \Delta_F(A) &= \Delta_F(A, \mathcal{N}_{\mathcal{T}_{(k)}}) = \min\{\|E\|_F : A+E \in \mathcal{N}_{\mathcal{T}_{(k)}}\} \\ \Delta^{\mathbb{R}}_F(A) &= \Delta_F(A, \mathcal{N}^{\mathbb{R}}_{\mathcal{T}_{(k)}}) = \min\{\|E\|_F : A+E \in \mathcal{N}^{\mathbb{R}}_{\mathcal{T}_{(k)}}\} \end{aligned}$$

3. Normal banded Toeplitz matrices, with $k \leq \lfloor n/2 \rfloor$

Theorem 3.1. *The $(2k+1)$-banded Toeplitz matrix $T_{(k)} = (n; k; \sigma, \delta, \tau)$, with $k \leq \lfloor n/2 \rfloor$, is normal if and only if there is an angle θ, such that*

$$\sigma_h = \overline{\tau}_h e^{i\theta}, \qquad h = 1 : k.$$

Corollary 3.2. *Let the Toeplitz matrix $T_{(k)} = (n; k; \sigma, \delta, \tau)$, with $k \leq \lfloor n/2 \rfloor$, be normal. Then $T_{(k)} = \delta I + e^{i\theta/2} S_{(k)}$, where $S_{(k)} = (n; k; \sigma, 0, \overline{\sigma})$ is a $(2k+1)$-banded Hermitian Toeplitz matrix.*

The restriction on the bandwidth rules out non-Hermitian $\{e^{i\varphi}\}$-circulant Toeplitz matrices.

The eigenvalues of a normal banded Toeplitz matrix of suitably restricted bandwidth can be determined from the eigenvalues of the associated Hermitian Toeplitz matrix by an affine transformation.

Similarly, the eigenvectors of a normal banded Toeplitz matrix can be computed as eigenvectors of a Hermitian matrix.

3.1. **Spectral properties.** Introduce the symbol for the matrix $T_{(k)} - \delta I = (n; k; \sigma, 0, \tau)$,

$$f(t) = \sum_{j=1}^{k} \left(\sigma_j e^{-ijt} + \tau_j e^{ijt}\right).$$

Theorem 3.3. *Let the $(2k+1)$-banded Toeplitz matrix $T_{(k)} = (n; k; \sigma, \delta, \tau)$, with $k \leq \lfloor n/2 \rfloor$, be normal. Then its eigenvalues are collinear. More precisely, let*

$$M = \max_{t \in \mathbf{R}} e^{-i\theta/2} f(t), \qquad m = \min_{t \in \mathbf{R}} e^{-i\theta/2} f(t),$$

where f is the symbol. Then the spectrum lives in the interval $\delta + e^{i\theta/2}[m, M]$.

Faber and Manteuffel investigated necessary and sufficient conditions on the matrix $M \in \mathbb{C}^{n\times n}$ for the existence of an iterative method of conjugate gradient-type with a three-term recurrence formula for the solution of linear systems of equations $Mx = b$, with $x, b \in \mathbb{C}^n$.

We show that normal $(2k + 1)$-banded Toeplitz matrices $T_{(k)} = (n; k; \sigma, \delta, \tau)$, with $k \leq \lfloor n/2 \rfloor$, satisfy these conditions.

3.2. The real case.

Corollary 3.4. *A real $(2k+1)$-banded Toeplitz matrix is normal if and only if it is symmetric or shifted skew-symmetric.*

Corollary 3.5. *The eigenvalues of a real normal $(2k+1)$-banded Toeplitz matrix $T_{(k)} = (n; k; \sigma, \delta, \tau)$, with $k \leq \lfloor n/2 \rfloor$, live on the real axis or on the straight line parallel to the imaginary axis through δ.*

Many results on eigenvalues and eigenvectors of symmetric Toeplitz matrices follow from the persymmetry of these matrices; see [3]. Related properties are shown in [1].

Theorem 3.6. *Let T be a shifted skew-symmetric Toeplitz matrix of order $2m$, $T = \delta I + A$, where $\delta \in \mathbb{C}$ and A is a real skew-symmetric Toeplitz matrix. If $\delta + \lambda$ is an eigenvalue of T associated with the eigenvector v, then $\delta - \lambda$ is an eigenvalue associated with the eigenvector Jv, where J is the reversal matrix.*

4. Large banded nonnormal Toeplitz matrices

Banded Toeplitz matrices arise in many applications in signal processing, time-series analysis, and numerical methods for the solution of partial differential equations.

The eigenvalue problem for large banded Toeplitz matrices is unanimously considered “numerically unreliable.”

In particular, the sensitivity of the eigenvalues of a banded Toeplitz matrix grows exponentially with the dimension n, except when the boundary of the spectrum of the associated Toeplitz operator is a curve with no interior.

This curve is related to the ε-pseudospectrum of the matrix as $\varepsilon \longrightarrow 0$ and $n \longrightarrow \infty$; see [9, Theorem 3.2].

A recent treatment of asymptotic properties of the spectra of banded Toeplitz matrices is provided in [2].

5. Distance to the algebraic variety $\mathcal{N}$ of normal matrices

In the numerical analysis community, the papers by Henrici [7] and Ruhe [10] on the distance to normality

$$d_F(A) = d_F(A, \mathcal{N}) = \min\{\|E\|_F : A + E \in \mathcal{N}\}$$

and the computation of the closest normal matrix, respectively, have received particular attention.

The algorithm presented in [10] is iterative and computationally expensive for large matrices.

We illustrate that the closest normal matrix and the distance to normality easily can be determined if further structure is imposed.

6. Structured distance to normality

Many properties of banded Toeplitz matrices can be shown in a simpler way than for general Toeplitz matrices by exploiting the bandedness. Since the normal and nonnormal matrices are required to have the same band- and Toeplitz structure, we refer to the distance as the *structured distance*.

As for the structured distance of banded Toeplitz matrices to the algebraic variety of normal banded Toeplitz matrices of the same bandwidth.

The structured distance to normality of a $(2k+1)$-banded Toeplitz matrix $T_{(k)} = (n; k; \sigma, \delta, \tau)$, with $k \leq \lfloor n/2 \rfloor$, in the Frobenius norm, is equal to

$$\Delta_F(T_{(k)}) = \sqrt{\frac{1}{2}\sum_{h=1}^{k}(n-h)(|\sigma_h|^2 + |\tau_h|^2) - \left|\sum_{h=1}^{k}(n-h)\sigma_h\tau_h\right|}.$$

6.1. The closest structured normal matrix.

Theorem 6.1. *Let $T_{(k)}$, with $k \leq \lfloor n/2 \rfloor$, satisfy*

$$\sum_{h=1}^{k} (n-h)\sigma_h \tau_h \neq 0. \tag{6.1}$$

Then $T^*_{(k)} = (n; k; \sigma^*, \delta, \tau^*)$ with

$$\sigma^* = \frac{\sigma + \overline{\tau} e^{i\theta^*}}{2}, \qquad \tau^* = \frac{\tau + \overline{\sigma} e^{i\theta^*}}{2}, \qquad \theta^* = \arg(\sum_{h=1}^{k} (n-h)\sigma_h \tau_h),$$

is the unique closest matrix, in the Frobenius norm, to $T_{(k)}$ in $\mathcal{N}_{\mathcal{T}_{(k)}}$.

Moreover, if $(\sigma, \tau) \neq (0,0)$ and (6.1) is violated, then there are infinitely many matrices $T^*_{(k)}$, depending on an arbitrary angle $\theta \in \mathbb{R}$, in $\mathcal{N}_{\mathcal{T}_{(k)}}$ at the same minimal distance from $T_{(k)}$, namely

$$T^*_{(k)} = T^*_{(k)}(\theta) = (n; k; \frac{\sigma + \overline{\tau} e^{i\theta}}{2}, \delta, \frac{\tau + \overline{\sigma} e^{i\theta}}{2}).$$

Proof. It suffices to determine the closest matrix $T^*_{0,(k)} = (n; k; \sigma^*, 0, \tau^*) \in \mathcal{N}_{\mathcal{T}_{(k)}}$ to the matrix $T_{0,(k)} = (n; k; \sigma, 0, \tau)$. Thanks to Theorem 3.1 there is an angle θ, such that $\sigma^* = \overline{\tau}^* e^{i\theta}$. We seek to determine a vector τ^* and an angle θ^*, for which

$$D(\tau^*, \theta^*) = \min_{\substack{z\in\mathbb{C}^k \\ -\pi<\theta\leq\pi}} D(z, \theta) = \min_{\substack{z\in\mathbb{C}^k \\ -\pi<\theta\leq\pi}} \|T^*_{0,(k)} - T_{0,(k)}\|_F^2.$$

Since $z \longrightarrow D(z, \theta)$ is convex and $\nabla_z D(z, \theta) = 0$ for the vector $z = z(\theta) = \frac{\tau + \overline{\sigma} e^{i\theta}}{2}$, the angle $\theta^* \in \mathbb{R}$ is obtained by minimizing

$$d(\theta) = D(z(\theta), \theta) = \frac{1}{2}\sum_{h=1}^{k}(n-h)(|\sigma_h|^2 + |\tau_h|^2) - \mathrm{Re}(e^{-i\theta}\sum_{h=1}^{k}(n-h)\sigma_h\tau_h).$$

If (6.1) holds, one has $d'(\theta) = 0$ iff $\theta^* = \arg(\sum_{h=1}^{k}(n-h)\sigma_h\tau_h)$ and the minimum

$$d(\theta^*) = \frac{1}{2}\sum_{h=1}^{k}(n-h)(|\sigma_h|^2 + |\tau_h|^2) - \left|\sum_{h=1}^{k}(n-h)\sigma_h\tau_h\right|.$$

If (6.1) is violated, $d(\theta) = \frac{1}{2}\|T_{0,(k)}\|_F^2$ for all values of $\theta \in \mathbb{R}$. ∎

6.2. The real case.

Corollary 6.2. $T_{(k)} \in \mathbb{R}^{n\times n}$, $k \le \lfloor n/2 \rfloor$. *If* $\sum_{h=1}^{k}(n-h)\sigma_h\tau_h$ *is positive, then the projection of* $T_{(k)}$ *onto* $\mathcal{N}^{\mathbb{R}}_{\mathcal{T}_{(k)}}$ *is the real symmetric* $(2k+1)$*-banded Toeplitz matrix*

$$T^*_{1,(k)} = (n; k; \frac{\sigma+\tau}{2}, \delta, \frac{\sigma+\tau}{2}).$$

If the sum is negative, the projection is the real shifted skew-symmetric $(2k+1)$*-banded Toeplitz matrix*

$$T^*_{2,(k)} = (n; k; \frac{\sigma-\tau}{2}, \delta, \frac{\tau-\sigma}{2}).$$

If the sum vanishes, then both the matrices are closest matrices to $T_{(k)}$ *in* $\mathcal{N}^{\mathbb{R}}_{\mathcal{T}_{(k)}}$ *in the Frobenius norm.*

$$\Delta^{\mathbb{R}}_F(T_{(k)}) = \sqrt{\frac{1}{2}\min\left\{\sum_{j=1}^{k}(n-j)(\sigma_j-\tau_j)^2, \sum_{j=1}^{k}(n-j)(\sigma_j+\tau_j)^2\right\}}.$$

7. The tridiagonal case. Structured distance to normality

Theorem 7.1. $T_{(1)} = (n; 1; \sigma, \delta, \tau)$ *is normal if and only if*

$$|\sigma| = |\tau| \, .$$

Theorem 7.2. *Let* $T_{(1)}$ *be any matrix in* $\mathcal{T}_{(1)}$*. If* $\sigma\tau \neq 0$,

$$T^*_{(1)} = (n; 1; \frac{|\sigma| + |\tau|}{2} \, e^{i \arg(\sigma)}, \delta, \frac{|\sigma| + |\tau|}{2} \, e^{i \arg(\tau)})$$

is the unique closest matrix, in the Frobenius norm, to $T_{(1)}$ *in* $\mathcal{N}_{\mathcal{T}_{(1)}}$*. If* $(\sigma, \tau) \neq (0, 0)$ *and the condition above is violated (i.e.* $T_{(1)}$ *is defective) there are infinitely many matrices in* $\mathcal{N}_{\mathcal{T}_{(1)}}$ *at the same minimal distance from* $T_{(1)}$*.*

$$\Delta_F(T_{(1)}) = \sqrt{\frac{n-1}{2}} \, \big| |\sigma| - |\tau| \big| \, .$$

7.1. **Sensitivity of the eigenvalues.** Consider a non defective tridiagonal Toeplitz matrix $T_{(1)}$ and denote with r the (nonzero) ratio

$$\frac{\min\{|\sigma|, |\tau|\}}{\max\{|\sigma|, |\tau|\}}.$$

If $0 < r < 1$, as for the hth individual condition number, $h = 1 : n$, one has

$$\kappa(\lambda_h(T_{(1)})) = \frac{\|x_h\|_2 \|y_h\|_2}{\left|y_h^H x_h\right|} = \frac{(1 - r^{n+1})(1 + r)(1 - \cos\frac{2h\pi}{n+1})}{r^{(n-1)/2}(n + 1)(1 - r)(1 + r^2 - 2r\cos\frac{2h\pi}{n+1})},$$

which is exponentially large as a function of n and tends to ∞ as r approaches 0. Conversely, if $r = 1$, $\kappa(\lambda_h(T_{(1)})) = 1$.

An estimate for the optimal global condition number $\kappa_F(\lambda) = \sum_{h=1}^{n} \kappa(\lambda_h(T_{(1)}))$, that models high-risk situations (large n, small r), results to be

$$\kappa_F(\lambda) \approx (1/r)^{\frac{n-1}{2}}.$$

7.2. **The $\varepsilon-$pseudospectrum of $T_{(1)}$.** The symbol of $T_{(1)}$ is $f(z) = \tau z + \delta + \sigma z^{-1}$. The ellipse $f(S)$, where S is the unit circle, is the boundary of the spectrum of the Toeplitz operator associated to $T_{(1)}$, which is strictly related with the $\varepsilon-$pseudospectrum of $T_{(1)}$ for $\varepsilon \to \infty$ and $n \to \infty$ [9, Theorem 3.2].

The spectrum of $T_{(1)}$ lies on the segment connecting the foci of the ellipse

$$\delta + te^{i(\arg(\sigma)+\arg(\tau))/2}, \qquad t \in \mathbb{R}, |t| \leq 2\sqrt{|\sigma||\tau|}$$

whereas the spectrum of $T^*_{(1)}$ lives in the major axis of the ellipse

$$\delta + te^{i(\arg(\sigma)+\arg(\tau))/2}, \qquad t \in \mathbb{R}, |t| \leq |\sigma| + |\tau|.$$

Thus, the $\varepsilon-$pseudospectral radius of $T_{(1)}$ is strictly related to the spectral radius of $T^*_{(1)}$.

The eccentricity of the ellipse depends on the size of the minor axis $||\sigma| - |\tau||$, i.e. it depends on $\Delta_F(T_{(1)})$. As $\Delta_F(T_{(1)})$ approaches zero, the boundary of the spectrum of the Toeplitz operator associated to $T_{(1)}$ tends to the segment in the complex plane which constitutes the asymptotic spectrum of $T_{(1)}$ (the limiting set on which the eigenvalues aggregate as $n \to \infty$).

7.3. **Unstructured versus structured.** Whether the (unstructured) distance to normality is more meaningful than the structured distance to normality depends on the application.

The restriction on the (upper/lower) bandwidth - that cannot exceed half the matrix dimension - rules out $\{e^{i\varphi}\}$-circulant Toeplitz matrices, natural normal approximations to banded Toeplitz matrices.

PROS From the matrix entries one has access to the symbol of the operator, and hence the spectrum of the corresponding Toeplitz operator, which seems more likely to provide an approximation of the spectrum of a large Toeplitz matrix than the eigenvalues of a nearest structured approximation (which must fall on a line, and will thus often miss key aspects of the spectrum of the original matrix).

CONS Here is an example involving a Jordan block, in which the unstructured distance is a poor indicator of the conditioning of the problem [8, Example 9.1].

7.4. **An example.** Consider a Jordan block, and its closest circulant matrix

$$T_{(1)} = (n; 1; 0, 0, \mu) = \begin{bmatrix} 0 & \mu & 0 & \dots & \dots & 0 & 0 \\ 0 & 0 & \mu & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & \mu & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 0 & \mu & 0 \\ 0 & 0 & 0 & \dots & \dots & 0 & \mu \\ 0 & 0 & 0 & \dots & \dots & \dots & 0 \end{bmatrix} \in \mathbb{R}^{n\times n}$$

$$C = \begin{bmatrix} 0 & \frac{n-1}{n}\mu & 0 & \dots & \dots & 0 & 0 \\ 0 & 0 & \frac{n-1}{n}\mu & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & \frac{n-1}{n}\mu & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 0 & \frac{n-1}{n}\mu & 0 \\ 0 & 0 & 0 & \dots & \dots & 0 & \frac{n-1}{n}\mu \\ \frac{n-1}{n}\mu & 0 & 0 & \dots & \dots & \dots & 0 \end{bmatrix} \in \mathbb{R}^{n\times n}$$

Which measure of the distance to normality reflects the fact that the eigenvalues of a Jordan block are sensitive to perturbations?

- *The normalized unstructured distance decreases to zero as the size of the block increases:*

$$\frac{d_F(T_{(1)})}{\|T_{(1)}\|_F} = \frac{\|T_{(1)} - C\|_F}{\|T_{(1)}\|_F} = \frac{1}{\sqrt{n}}.$$

- *The normalized structured distance is maximal:*

$$T^*_{1,(1)} = (n; 1; \frac{\mu}{2}, 0, \frac{\mu}{2}), \qquad T^*_{2,(1)} = (n; 1; -\frac{\mu}{2}, 0, \frac{\mu}{2})$$

$$\frac{\Delta^{\mathbb{R}}_F(T_{(1)})}{\|T_{(1)}\|_F} = \frac{\|T_{(1)} - T^*_{1,(1)}\|_F}{\|T_{(1)}\|_F} = \frac{\|T_{(1)} - T^*_{2,(1)}\|_F}{\|T_{(1)}\|_F} = \frac{1}{\sqrt{2}}.$$