Short course on structured matrices

D.A. Bini, Università di Pisa
bini@dm.unipi.it

Journées Nationales de Calcul Formel (JNCF) 2014
CIRM, Luminy
November 3–7, 2014

Outline

1. Preliminaries
2. Toeplitz matrices
   - Applications
   - Asymptotic spectral properties
   - Computational issues
     - Some matrix algebras
     - displacement operators, fast and superfast algorithms
     - preconditioning
     - Wiener-Hopf factorization and matrix equations
   - A recent application
3. Rank structured matrices
   - Basic properties
   - Companion matrices
   - Application: Linearizng matrix polynomials

Preliminaries

Structured matrices are encountered almost everywhere

The structure of a matrix reflects the peculiarity of the mathematical model that the matrix describes

Exploiting matrix structures is a mandatory step for designing highly efficient *ad hoc* algorithms for solving computational problems

Structure analysis often reveals rich and interesting theoretical properties

Linear models lead naturally to matrices

Some structures are evident, some other structures are more hidden

Nonlinear model are usually linearized or approximated by means of linear models

Preliminaries

Some examples:

Band matrices: locality properties, functions with compact support. Spline interpolation, finite differences

Toeplitz matrices: shift invariance properties. Polynomial computations, queueing models, image restoration

Displacement structures, Toeplitz-like matrices: Vandermonde, Cauchy, Hankel, Bezout, Pick

Semi-separable and quasi-separable matrices: inverse of band matrices, polynomial and matrix polynomial computations, integral equations

Sparse matrices: Web, Page Rank, social networks, complex networks

Preliminaries

In this short course we will limit ourselves to describe some computational aspect of

- Toeplitz matrices
- Rank-structured matrices

and show some applications

The spirit is to give the flavour of the available results with pointers to the literature

Notations:
$\mathbb{F}$ is a number field, for our purpose $\mathbb{F} \in \{\mathbb{R},\ \mathbb{C}\}$
$\mathbb{N} = \{0, 1, 2, 3, \ldots\}, \quad \mathbb{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\}$
$\mathbb{T}$ is the unit circle in the complex plane
$\underline{i}$ imaginary unit such that $\underline{i}^2 = -1$
$\mathbb{F}^{m \times n}$ set of $m \times n$ matrices with entries in $\mathbb{F}$

Toeplitz matrices [Otto Toeplitz 1881-1940]

Let $\mathbb{F}$ be a field ($\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$)

Given a bi-infinite sequence $\{a_i\}_{i\in\mathbb{Z}} \in \mathbb{F}^{\mathbb{Z}}$ and an integer n, the $n \times n$ matrix $T_n = (t_{i,j})_{i,j=1,n}$ such that $t_{i,j} = a_{j-i}$ is called *Toeplitz matrix*

$$T_5 = \begin{bmatrix} a_0 & a_1 & a_2 & a_3 & a_4 \\ a_{-1} & a_0 & a_1 & a_2 & a_3 \\ a_{-2} & a_{-1} & a_0 & a_1 & a_2 \\ a_{-3} & a_{-2} & a_{-1} & a_0 & a_1 \\ a_{-4} & a_{-3} & a_{-2} & a_{-1} & a_0 \end{bmatrix}$$

T_n is a leading principal submatrix of the (semi) infinite Toeplitz matrix $T_\infty = (t_{i,j})_{i,j\in\mathbb{N}}$, $t_{i,j} = a_{j-i}$

$$T_\infty = \begin{bmatrix} a_0 & a_1 & a_2 & \cdots \\ a_{-1} & a_0 & a_1 & \ddots \\ a_{-2} & a_{-1} & \ddots & \ddots \\ \vdots & \ddots & \ddots & \ddots \end{bmatrix}$$

Toeplitz matrices

Theorem (Otto Toeplitz) The matrix T_∞ defines a bounded linear operator in $\ell^2(\mathbb{N})$, $x \to y = T_\infty x$, $y_i = \sum_{j=0}^{+\infty} a_{j-i} x_j$ if and only if a_i are the Fourier coefficients of a function $a(z) \in L^\infty(\mathbb{T})$

$$a(z) = \sum_{n=-\infty}^{+\infty} a_n z^n, \quad a_n = \frac{1}{2\pi} \int_0^{2\pi} a(e^{i\theta}) e^{-in\theta} d\theta$$

In this case

$$\|T\| = \text{ess sup}_{z \in \mathbb{T}} |a(z)|, \quad \text{where } \|T\| := \sup_{\|x\|=1} \|Tx\|$$

The function $a(z)$ is called *symbol* associated with T_∞

Example
If $a(z) = \sum_{i=-k}^{k} a_i z^i$ is a Laurent polynomial, then T_∞ is a banded Toeplitz matrix which defines a bounded linear operator

Block Toeplitz matrices

Let $\mathbb{F}$ be a field ($\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$)

Given a bi-infinite sequence $\{A_i\}_{i\in\mathbb{Z}}$, $A_i \in \mathbb{F}^{m\times m}$ and an integer n, the $mn \times mn$ matrix $T_n = (t_{i,j})_{i,j=1,n}$ such that $t_{i,j} = A_{j-i}$ is called *block Toeplitz* matrix

$$T_5 = \begin{bmatrix} A_0 & A_1 & A_2 & A_3 & A_4 \\ A_{-1} & A_0 & A_1 & A_2 & A_3 \\ A_{-2} & A_{-1} & A_0 & A_1 & A_2 \\ A_{-3} & A_{-2} & A_{-1} & A_0 & A_1 \\ A_{-4} & A_{-3} & A_{-2} & A_{-1} & A_0 \end{bmatrix}$$

T_n is a leading principal submatrix of the (semi) infinite block Toeplitz matrix $T_\infty = (t_{i,j})_{i,j\in\mathbb{N}}$, $t_{i,j} = A_{j-i}$

$$T_\infty = \begin{bmatrix} A_0 & A_1 & A_2 & \dots \\ A_{-1} & A_0 & A_1 & \ddots \\ A_{-2} & A_{-1} & \ddots & \ddots \\ \vdots & \ddots & \ddots & \ddots \end{bmatrix}$$

Block Toeplitz matrices with Toeplitz blocks

The infinite block Toeplitz matrix T_∞ defines a bounded linear operator in $\ell^2(\mathbb{N})$ iff the blocks $A_k = (a_{i,j}^{(k)})$ are the Fourier coefficients of a matrix-valued function $A(z) : \mathbb{T} \to \mathbb{C}^{m\times m}$,
$A(z) = \sum_{k=-\infty}^{+\infty} x^k A_k = (a_{i,j}(z))_{i,j=1,m}$ such that $a_{i,j}(x) \in L^\infty(\mathbb{T})$

If the blocks A_i are Toeplitz themselves we have a block Toeplitz matrix with Toeplitz blocks

A function $a(z, w) : \mathbb{T} \times \mathbb{T} \to \mathbb{C}$ having the Fourier series
$a(z, w) = \sum_{i,j=-\infty}^{+\infty} a_{i,j} z^i w^j$ defines an infinite block Toeplitz matrix
$T_\infty = (A_{j-i})$ with infinite Toeplitz blocks $A_k = (a_{k,j-i})$.
T_∞ defines a bounded operator iff $a(z, w) \in L_\infty$

For any pair of integers n, m we may construct an $n \times n$ Toeplitz matrix
$T_{m,n} = (A_{j-i})_{i,j=1,n}$ with $m \times m$ Toeplitz blocks $A_{j-i} = (a_{k,j-i})_{i,j=1,m}$

Multilevel Toeplitz matrices

A function $a : \mathbb{T}^d \to \mathbb{C}$ having the Fourier expansion

$$a(z_1, z_2, \ldots, z_d) = \sum_{i_1,\ldots,i_d=-\infty}^{+\infty} a_{i_1,i_2,\ldots,i_d} z_{i_1}^{i_1} z_{i_2}^{i_2} \cdots z_{i_d}^{i_d}$$

defines a *d-multilevel Toeplitz* matrix: that is a block Toeplitz matrix with blocks that are themselves $(d-1)$-multilevel Toeplitz matrices

Generalization: Toeplitz-like matrices

Let L_i and U_i be lower triangular and upper triangular $n \times n$ Toeplitz matrices, respectively, where $i = 1, \ldots, k$ and k is independent of n

$$A = \sum_{i=1}^{k} L_i U_i$$

is called a *Toeplitz-like* matrix

If $k = 2$, $L_1 = U_2 = I$ then A is a Toeplitz matrix.

If A is an invertible Toeplitz matrix then there exist L_i, U_i, $i = 1, 2$ such that

$$A^{-1} = L_1 U_1 + L_2 U_2$$

that is, A^{-1} is Toeplitz-like

Applications: polynomial arithmetic

Polynomial multiplication

$a(x) = \sum_{i=0}^{n} a_i x^i, \quad b(x) = \sum_{i=0}^{m} b_i x^i,$

$c(x) := a(x)b(x),\ c(x) = \sum_{i=0}^{m+n} c_i x^i$

$c_0 = a_0 b_0$

$c_1 = a_0 b_1 + a_1 b_0$

$\ldots$

$$\begin{bmatrix} c_0 \\ c_1 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ c_{m+n} \end{bmatrix} = \begin{bmatrix} a_0 & & & \\ a_1 & a_0 & & \\ \vdots & \ddots & \ddots & \\ a_n & \ddots & \ddots & a_0 \\ & \ddots & \ddots & a_1 \\ & & \ddots & \vdots \\ & & & a_n \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_m \end{bmatrix}$$

Applications: polynomial arithmetic

Polynomial division

$a(x) = \sum_{i=0}^{n} a_i x^i$, $b(x) = \sum_{i=0}^{m} b_i x^i$, $b_m \neq 0$

$a(x) = b(x)q(x) + r(x), \quad \deg\, r(x) < m$

$q(x)$ quotient, $r(x)$ remainder of the division of $a(x)$ by $b(x)$

$$\begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_m \\ \vdots \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} b_0 & & & \\ b_1 & b_0 & & \\ \vdots & \ddots & \ddots & \\ b_m & \ddots & \ddots & b_0 \\ & \ddots & \ddots & b_1 \\ & & \ddots & \vdots \\ & & & b_m \end{bmatrix} \begin{bmatrix} q_0 \\ q_1 \\ \vdots \\ q_{n-m} \end{bmatrix} + \begin{bmatrix} r_0 \\ \vdots \\ r_{m-1} \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

The last $n - m + 1$ equations form a triangular Toeplitz system

Applications: polynomial arithmetic

Polynomial division

$$\begin{bmatrix} b_m & b_{m-1} & \cdots & b_{2m-n} \\ & b_m & \ddots & \vdots \\ & & \ddots & b_{m-1} \\ & & & b_m \end{bmatrix} \begin{bmatrix} q_0 \\ q_1 \\ \vdots \\ q_{n-m} \end{bmatrix} = \begin{bmatrix} a_m \\ a_{m+1} \\ \vdots \\ a_n \end{bmatrix}$$

Its solution provides the coefficients of the quotient.
The remainder can be computed as a difference.

$$\begin{bmatrix} r_0 \\ \vdots \\ r_{m-1} \end{bmatrix} = \begin{bmatrix} a_0 \\ \vdots \\ a_{m-1} \end{bmatrix} - \begin{bmatrix} b_0 & & \\ \vdots & \ddots & \\ b_{m-1} & \cdots & b_0 \end{bmatrix} \begin{bmatrix} q_0 \\ \vdots \\ q_{n-m} \end{bmatrix}$$

(in the picture $n - m = m - 1$)

Applications: polynomial arithmetic

Polynomial gcd

If $g(x) = \gcd(a(x), b(x))$, $\deg(g(x)) = k$, $\deg(a(x)) = n$, $\deg(b(x)) = m$. Then there exist polynomials $r(x)$, $s(x)$ of degree at most $m - k - 1$, $n - k - 1$, respectively, such that (Bézout identity)

$$g(x) = a(x)r(x) + b(x)s(x)$$

In matrix form one has the $(m + n - k) \times (m + n - 2k)$ system

$$\left[\begin{array}{cccc|cccc}
a_0 & & & & b_0 & & & \\
a_1 & a_0 & & & b_1 & b_0 & & \\
\vdots & \ddots & \ddots & & \vdots & \ddots & \ddots & \\
a_n & \ddots & \ddots & a_0 & b_m & \ddots & \ddots & b_0 \\
 & \ddots & \ddots & a_1 & & \ddots & \ddots & b_1 \\
 & & \ddots & \vdots & & & \ddots & \vdots \\
 & & & a_n & & & & b_m
\end{array}\right]
\left[\begin{array}{c}
r_0 \\ r_1 \\ \vdots \\ r_{m-k-1} \\ \hline s_0 \\ s_1 \\ \vdots \\ s_{n-k-1}
\end{array}\right]
=
\left[\begin{array}{c}
g_0 \\ \vdots \\ g_k \\ 0 \\ \vdots \\ \vdots \\ 0
\end{array}\right]$$

Sylvester matrix

Applications: polynomial arithmetic

Polynomial gcd

The last $m + n - 2k$ equations provide a linear system of the kind

$$S \begin{bmatrix} r \\ s \end{bmatrix} = \begin{bmatrix} g_k \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

where S is the $(m + n - 2k) \times (m + n - 2k)$ submatrix of the Sylvester matrix in the previous slide formed by two Toeplitz matrices.

Applications: polynomial arithmetic

Infinite Toeplitz matrices

Let $a(x), b(x)$ be polynomials of degree n, m with coefficients a_i, b_j, define the Laurent polynomial

$$c(x) = a(x)b(x^{-1}) = \sum_{i=-m}^{n} c_i x^i$$

Then the following infinite UL factorization holds

$$\begin{bmatrix} c_0 & \dots & c_n & & \\ \vdots & c_0 & \ddots & \ddots & \\ c_{-m} & \ddots & \ddots & \ddots & \ddots \\ & \ddots & \ddots & \ddots & \end{bmatrix} = \begin{bmatrix} a_0 & \dots & a_n & & \\ & a_0 & \ddots & a_n & \\ & & \ddots & \ddots & \ddots \end{bmatrix} \begin{bmatrix} b_m & & & & \\ b_{m-1} & b_0 & & & \\ \vdots & \ddots & \ddots & & \\ b_0 & \ddots & \ddots & \ddots & \\ & \ddots & \ddots & \ddots & \ddots \end{bmatrix}$$

If the zeros of $a(x)$ and $b(x)$ lie in the unit disk, this factorization is called Wiener-Hopf factorization. This factorization is encountered in many applications.

Applications: polynomial arithmetic

Infinite Toeplitz matrices

The Wiener-Hopf factorization can be defined for matrix-valued functions $C(x) = \sum_{i=-\infty}^{+\infty} C_i x^i$, $C_i \in \mathbb{C}^{m\times m}$, in the Wiener class, i.e, such that $\sum_{i=-\infty}^{+\infty} \|C_i\| < \infty$, provided that $\det C(x) \neq 0$ for $|x| = 1$.

A canonical Wiener-Hopf factorization takes the form

$$C(x) = A(x)B(x^{-1}), \quad A(x) = \sum_{i=0}^{\infty} x^i A_i, \ B(x) = \sum_{i=0}^{\infty} B_i x^i$$

where $A(x)$ and $B(x)$ have zeros in the open unit disk.

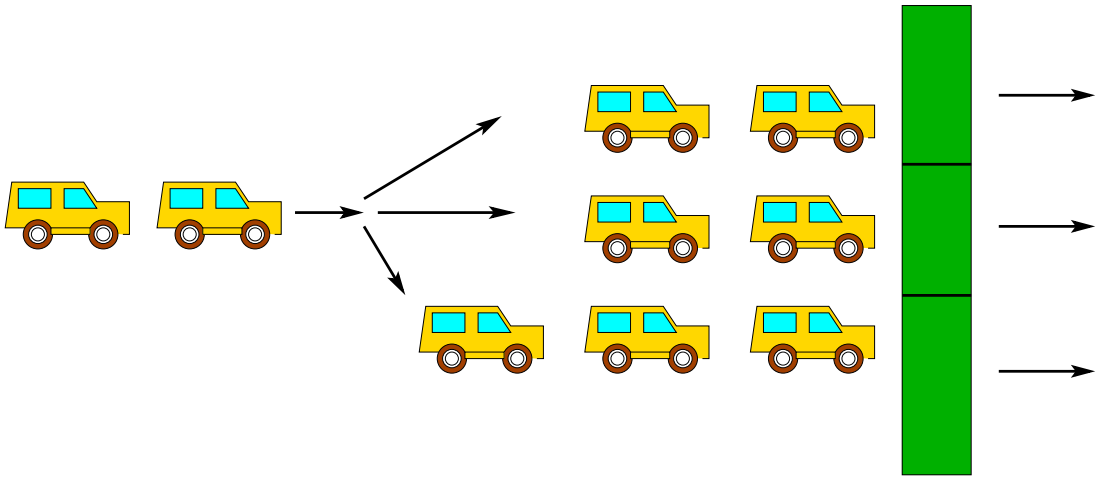
Its matrix representation provides a block UL factorization of the infinite block Toeplitz matrix (C_{j-i})

Applications: Polynomial arithmetic

$$\begin{bmatrix} C_0 & C_1 & \cdots & \\ C_{-1} & C_0 & C_1 & \ddots \\ \vdots & \ddots & \ddots & \ddots \end{bmatrix} = \begin{bmatrix} A_0 & A_1 & \cdots & \\ & A_0 & A_1 & \ddots \\ & & \ddots & \ddots \\ & & & \ddots \end{bmatrix} \begin{bmatrix} B_0 & & \\ B_{-1} & B_0 & \\ \vdots & \ddots & \ddots \end{bmatrix}$$

Applications: Queueing models

The shortest queue problem

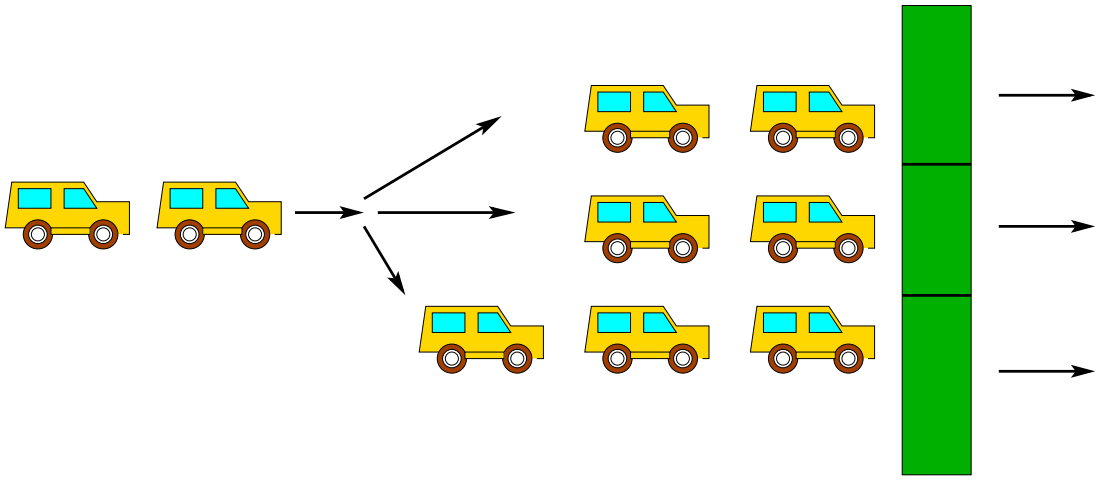


The problem: There are m gates at the highway:

- at each instant k cars arrive with a known probability
- each car follows the shortest line
- at each instant a car leaves its gate

Applications: Queueing models

The shortest queue problem



The problem: There are m gates at the highway:

- at each instant k cars arrive with a known probability
- each car follows the shortest line
- at each instant a car leaves its gate

what is the probability that there are ℓ cars in the lines waiting to be served?

Similar model: the wireless IEEE 802.11 protocol

Applications: Queueing models

The shortest queue problem

Denoting $p_{i,j}$ the probability that after one instant of time the length of the queue changes from i to j then $p_{i,j} = a_{j-i}$, if $i \geq m$, where $a_k \geq 0$ is the probability that $m + k$ cars arrive, $\sum_{k=-m}^{\infty} a_k = 1$, $a_k = 0$ for $k < -m$

The problem turns into an infinite eigenvalue problem of the kind

$$\pi^T P = \pi^T,$$

$\pi \in \mathbb{R}$ is a probability vector, i.e., $\sum \pi_i = 1$, $\pi_i \geq 0$, and $P = (p_{i,j})$ is almost Toeplitz in generalized upper Hessenberg form

Applications: Queueing models

$$P = \begin{bmatrix} b_{1,1} & b_{1,2} & \dots & \dots \\ \vdots & \vdots & \vdots & \vdots \\ b_{m,1} & b_{m,2} & \dots & \dots \\ a_0 & a_1 & a_2 & \dots \\ & a_0 & a_1 & \ddots \\ & & \ddots & \ddots \end{bmatrix}$$

where $b_{i,j}$ are suitable boundary probabilities. This matrix can be partitioned into $m \times m$ blocks as follows

$$P = \begin{bmatrix} B_0 & B_1 & B_2 & \dots & \\ A_{-1} & A_0 & A_1 & \ddots & \ddots \\ 0 & A_{-1} & A_0 & \ddots & \ddots \\ \vdots & \ddots & \ddots & \ddots & \ddots \end{bmatrix}$$

Applications: Queueing models

Removing the first block row and the first block column of the above matrix yields the block Hessenberg block Toeplitz matrix

$$\widehat{P} = \begin{bmatrix} A_0 & A_1 & A_2 & \cdots & \\ A_{-1} & A_0 & A_1 & \ddots & \ddots \\ 0 & A_{-1} & \ddots & \ddots & \ddots \\ \vdots & \ddots & \ddots & \ddots & \ddots \end{bmatrix}$$

The Wiener-Hopf factorization of $\widehat{P} - I$ allows to solve easily the problem $\pi(P - I) = 0$

Applications: Queueing models

The Wiener-Hopf factorization of $\widehat{P} - I$ takes the following form

$$\widehat{P} - I = \begin{bmatrix} U_0 & U_1 & \dots & \\ & U_0 & U_1 & \ddots \\ & & \ddots & \ddots \end{bmatrix} \begin{bmatrix} I & & & \\ -G & I & & \\ & -G & I & \\ & & \ddots & \ddots \end{bmatrix}$$

where G is the solution of the following matrix equation

$$X = \sum_{i=-1}^{+\infty} A_i X^i$$

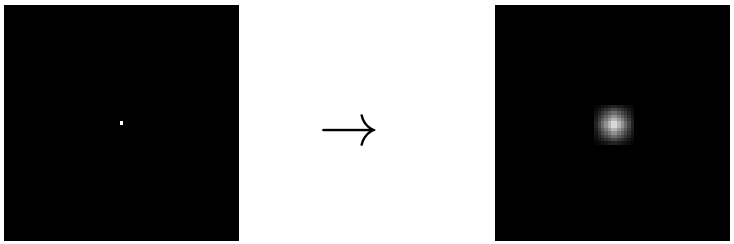
having nonnegative entries and spectral radius $\rho(X) = 1$.

A way for solving this equation is to reduce it to the following infinite linear block Toeplitz system

$$\begin{bmatrix} A_0 - I & A_1 & A_2 & \dots \\ A_{-1} & A_0 - I & A_1 & \dots \\ & A_{-1} & A_0 - I & \dots \\ & & \ddots & \ddots \end{bmatrix} \begin{bmatrix} X \\ X^2 \\ X^3 \\ \vdots \end{bmatrix} = \begin{bmatrix} -A_{-1} \\ 0 \\ 0 \\ \vdots \end{bmatrix}.$$

Applications: Image restoration

In the image restoration models, the blur of a single point of an image is independent of the location of the point and is defined by the Point-Spread Function (PSF)



The relation between the blurred and noisy image, stored as a vector b and the real image, represented by a vector x has the form

$$Ax = b - \text{noise}$$

Shift invariance of the PSF $\Rightarrow$ A is block Toeplitz with Toeplitz blocks

Due to the local effect of the blur, the PSF has compact support so that A is block banded with banded blocks

Typically, A is ill-conditioned so that solving the system $Ax = b$ obtained by ignoring the noise provides a highly perturbed solution

For instance the PSF which transforms a unit point of light into the 3×3 square $\frac{1}{15}\begin{bmatrix} 1 & 2 & 1 \\ 2 & 3 & 2 \\ 1 & 2 & 1 \end{bmatrix}$ leads to the following block Toeplitz matrix

$$T = \frac{1}{15}\begin{bmatrix} B & A & & & \\ A & B & A & & \\ & \ddots & \ddots & \ddots & \\ & & A & B & A \\ & & & A & B \end{bmatrix}$$

where

$$A = \begin{bmatrix} 2 & 1 & & & \\ 1 & 2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & 2 & 1 \\ & & & 1 & 2 \end{bmatrix}, \quad B = \begin{bmatrix} 3 & 2 & & & \\ 2 & 3 & 2 & & \\ & \ddots & \ddots & \ddots & \\ & & 2 & 3 & 2 \\ & & & 2 & 3 \end{bmatrix}$$

This way, restoring a blurred image is reduced to solving a block banded block Toeplitz systems with banded Toeplitz blocks. According to the boundary conditions assumed in the blurring model, the matrix can take additional specific structures.

Applications: Differential equations

The numerical treatment of linear partial differential equations with constant coefficients by means of the finite difference technique leads to linear systems where the matrix is block Toeplitz with Toeplitz blocks

For instance the discretization of the Laplace operator $\Delta u(x, y)$ applied to a function $u(x, y)$ defined over $[0, 1] \times [0, 1]$

$$-\Delta u(x, y) = -(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}) = \frac{1}{h^2}(4u_{i,j} - u_{i+1,j} - u_{i-1,j} - u_{i,j+1}, -u_{i,j-1}) + O(h^2)$$

for $x_i = ih$, $y_j = jh$, $i, j = 1, n$, $h = 1/(n + 1)$, $u_{i,j} = u(x_i, y_j)$ leads to the matrix

$$L = -\frac{1}{h^2}\begin{bmatrix} A & -I & & \\ -I & A & \ddots & \\ & \ddots & \ddots & -I \\ & & -I & A \end{bmatrix}, \quad A = \begin{bmatrix} 4 & -1 & & \\ -1 & 4 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 4 \end{bmatrix}.$$

The symbol associated with L is $a(x, y) = 4 - x - x^{-1} - y - y^{-1}$

Asymptotic spectral properties and preconditioning

Definition: Let $f(x) : [0, 2\pi] \to \mathbb{R}$ be a Lebesgue integrable function. A sequence $\{\lambda_i^{(n)}\}_{i=1,n}$, $n \in \mathbb{N}$, $\lambda_i^{(n)} \in \mathbb{R}$ is distributed as $f(x)$ if

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} F(\lambda_i^{(n)}) = \frac{1}{2\pi} \int_0^{2\pi} F(f(x))dx$$

for any continuous $F(x)$ with bounded support.

Example $\lambda_i^{(n)} = f(2i\pi/n)$, $i = 1, \ldots, n$, $n \in \mathbb{N}$ is distributed as $f(x)$.

With abuse of notation, given $a(z) : \mathbb{T} \to \mathbb{R}$ we write $a(\theta)$ in place of $a(x(\theta))$, $x(\theta) = \cos\theta + \underline{i} \sin\theta \in \mathbb{T}$

Asymptotic spectral properties and preconditioning

Assume that

- the symbol $a(\theta) : [0 : 2\pi] \to \mathbb{R}$ is a real valued function so that $a(\theta) = a_0 + 2\sum_{k=1}^{\infty} a_k \cos k\theta$
- T_n is the sequence of Toeplitz matrices associated with $a(\theta)$, i.e., $T_n = (a_{|j-i|})_{i,j=1,n}$; observe that $T^{(n)}$ is symmetric
- $m_a = \operatorname{ess\,inf}_{\theta\in[0,2\pi]} a(\theta)$, $M_a = \operatorname{ess\,sup}_{\theta\in[0,2\pi]} a(\theta)$ are the essential infimum and the essential supremum
- $\lambda_1^{(n)} \le \lambda_2^{(n)} \le \cdots \le \lambda_n^{(n)}$ are the eigenvalues of T_n sorted in nondecreasing order (observe that T_n is real symmetric).

Then

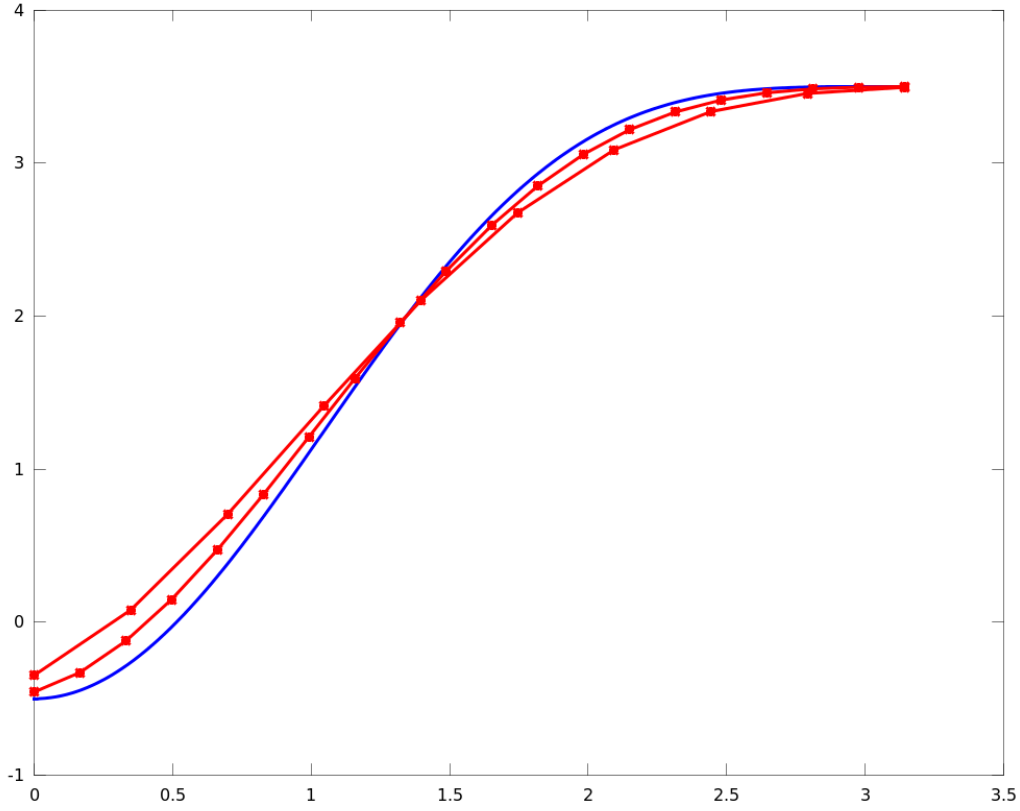
Asymptotic spectral properties and preconditioning

1. if $m_a < M_a$ then $\lambda_i^{(n)} \in (m_a, M_a)$ for any n and $i = 1, \ldots, n$; if $m_a = M_a$ then $a(\theta)$ is constant and $T_n(a) = m_a I_n$;
2. $\lim_{n\to\infty} \lambda_1^{(n)} = m_a$, $\lim_{n\to\infty} \lambda_n^{(n)} = M_a$;
3. the eigenvalues sequence $\{\lambda_1^{(n)}, \ldots, \lambda_n^{(n)}\}$ are distributed as $a(\theta)$

Moreover

- if $a(x) \geq 0$ the condition number $\mu^{(n)} = \|T^{(n)}\|_2 \|(T^{(n)})^{-1}\|_2$ of $T^{(n)}$ is such that $\lim_{n\to\infty} \mu^{(n)} = M_a/m_a$
- $a(\theta) > 0$ implies that $T^{(n)}$ is uniformly well conditioned
- $a(\theta) = 0$ for some θ implies that $\lim_{n\to\infty} \mu_n = \infty$

Asymptotic spectral properties and preconditioning



In red: eigenvalues of the Toeplitz matrix T_n associated with the symbol $f(\theta) = 2 - 2\cos\theta - \frac{1}{2}\cos(2\theta)$ for $n = 10$, $n = 20$

In blue: graph of the symbol. As n grows, the values $\lambda_i^{(n)}$ for $i = 1, \ldots, n$ tend to be shaped as the graph of the symbol

Asymptotic spectral properties and preconditioning

The same asymptotic property holds true for

- block Toeplitz matrices generated by a matrix valued symbol $A(x)$
- block Toeplitz matrices with Toeplitz blocks generated by a bivariate symbol $a(x, y)$
- multilevel block Toeplitz matrices generated by a multivariate symbol $a(x_1, x_2, \ldots, x_d)$
- singular values of any of the above matrix classes

The same results hold for the product $P_n^{-1} T_n$ where T_n and P_n are associated with symbols $a(\theta)$, $p(\theta)$, respectively

- eigenvalues are distributed as $a(\theta)/p(\theta)$
- (preconditioning) given $a(\theta) \geq 0$ such that $a(\theta_0) = 0$ for some θ_0; if there exists a trigonometric polynomial $p(\theta) = \sum_{i=-k}^{k} p_k \cos(k\theta)$ such that $p(\theta_0) = 0$, $\lim_{\theta \to \theta_0} a(\theta)/p(\theta) \neq 0$ then $P_n^{-1} T_n$ has condition number uniformly bounded by a constant

Trigonometric matrix algebras and Fast multiplication

Let $\omega_n = \cos\frac{2\pi}{n} + \underline{i}\sin\frac{2\pi}{n}$ be a primitive nth root of 1, that is, such that $\omega_n^n = 1$ and $\{1, \omega_n, \ldots, \omega_n^{n-1}\}$ has cardinality n.

Define the $n \times n$ matrix $\Omega_n = (\omega_n^{ij})_{i,j=0,n-1}$, $F_n = \frac{1}{\sqrt{n}}\Omega_n$.

One can easily verify that $F_n^* F_n = I$ that is, F_n is a unitary matrix.

For $x \in \mathbb{C}^n$ define

- $y = \mathrm{DFT}(x) = \frac{1}{n}\Omega_n^* x$ the Discrete Fourier Transform (DFT) of x
- $x = \mathrm{IDFT}(y) = \Omega y$ the Inverse DFT (IDFT) of y

Remark: $\mathrm{cond}_2(F_n) = \|F_n\|_2 \|F_n^{-1}\|_2 = 1, \quad \mathrm{cond}_2(\Omega_n) = 1$

This shows that the DFT and IDFT are numerically well conditioned when the perturbation errors are measured in the 2-norm.

Trigonometric matrix algebras and Fast multiplication

If n is an integer power of 2 then the IDFT of a vector can be computed with the cost of $\frac{3}{2} n \log_2 n$ arithmetic operations by means of FFT

FFT is backward numerically stable in the 2-norm. That is, if $\widetilde{x}$ is the value computed in floating point arithmetic with precision μ in place of $x = \mathrm{IDFT}(y)$ then

$$\|x - \widetilde{x}\|_2 \leq \mu\gamma\|x\|_2 \log_2 n$$

for a moderate constant γ

norm-wise well conditioning of DFT and the norm-wise stability of FFT make this tool very effective for **most numerical computations.**

Unfortunately, the norm-wise stability of FFT does not imply the component-wise stability. That is, the inequality

$$|x_i - \widetilde{x}_i| \leq \mu\gamma|x_i| \log_2 n$$

is **not generally true** for all the components x_i.

Trigonometric matrix algebras and Fast multiplication

This is a drawback of DFT and of FFT when numerically used for symbolic computations since, in order to guarantee a sufficiently accurate relative precision in the result, one has to choose a suitable value of the machine precision of the floating point arithmetic whose value depends on the ratio between the maximum and the minimum absolute value of the output.

This fact implies that the complexity bounds are depending on this ratio. When using FFT in this framework one should be aware of this fact.

There are algorithms for computing the DFT in $O(n \log n)$ ops whatever is the value of n.

DFT and FFT can be defined over finite fields where there exists a primitive root of 1. For instance, $\mathbb{Z}_{17}$ is a finite field and 3 is a primitive 16th root of 1. DFT and FFT can be defined over certain rings.

An example: Graeffe iteration

Let $p(x) = \sum_{i=0}^{n} p_i x^i$ be a polynomial of degree n such that $p(x)$ has zeros x_i, $i = 1, \ldots, n$ such that

$$|x_1| < \cdots < |x_m| < 1 < |x_{m+1}| < \cdots < |x_n|$$

With $p_0(x) := p(x)$ define the sequence (Graeffe iteration)

$$q(x^2) = p_k(x)p_k(-x), \quad p_{k+1}(x) = q(x)/q_m, \quad \text{for } k = 0, 1, 2, , \ldots$$

The zeros of $p_k(x)$ are $x_i^{2^k}$, so that $\lim_{k\to\infty} p_k(x) = x^m$

If $p_k(x) = \sum_{i=0}^{n} p_i^{(k)} x^i$ then

$$\lim_{k\to\infty} |p_{n-1}^{(k)}/p_n^{(k)}|^{1/2^k} = |x_n|$$

moreover, convergence is very fast. Similar equations hold for $|x_i|$

An example: Graeffe iteration

$$\lim_{k\to\infty} |p_{n-1}^{(k)}/p_n^{(k)}|^{1/2^k} = |x_n|$$

On the other hand (if $m < n-1$)

$$\lim_{k\to\infty} |p_{n-1}^{(k)}| = \lim_{k\to\infty} |p_n^{(k)}| = 0$$

with double exponential convergence
Computing $p_k(x)$ given $p_{k-1}(x)$ by using FFT (evaluation interpolation at the roots of unity) costs $O(n \log n)$ ops.

But as soon as $|p_n^{(k)}|$ and $|p_{n-1}^{(k)}|$ are below the machine precision the relative error in these two coefficients is greater than 1. That is **no digit is correct** in the computed estimate of $|x_n|$.

An example: Graeffe iteration

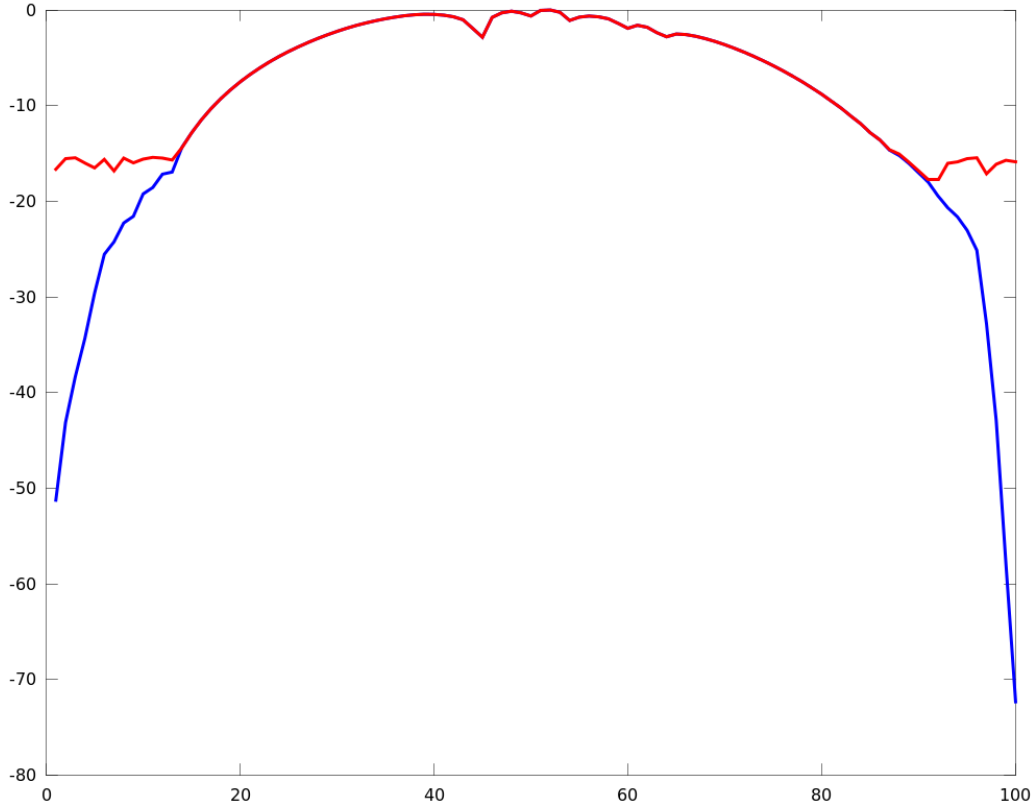


Figure: The values of $\log_{10} |p_i^{(6)}|$ for $i = 0, \ldots, n$ for the polynomial obtained after 6 Graeffe steps starting from a random polynomial of degree 100. In red the case where the coefficients are computed with FFT, in blue the coefficients computed with the customary algorithm

An example: Graeffe iteration

step	custom	FFT
1	1.40235695	1.40235695
2	2.07798429	2.07798429
3	2.01615072	2.01615072
4	2.01971626	2.01857621
5	2.01971854	1.00375471
6	2.01971854	0.99877589

An example: Graeffe iteration

A specific analysis shows that in order to have d correct digits in the computed approximation, one must use a floating point arithmetic with c digits, where

$$c = d * \left(1 + \gamma \frac{\log(|x_n|/|x_1|)}{\log(|x_n|/|x_{n-1}|)}\right), \quad \gamma > 1$$

Problems are encountered if $|x_n| \approx |x_{n-1}|$ or $|x_n/x_1|$ is large.

In the situation where the separation from two consecutive zeros is uniform, i.e., $|x_{i+1}/x_i| = |x_n/x_1|^{1/n}$ then the number of digits is

$$c = d * (1 + \gamma n)$$

$O(n \log n)$ ops with $O(nd)$ digits more expensive than $O(n^2)$ ops with d digits

Trigonometric matrix algebras and Fast multiplication

There are many other useful trigonometric transforms that can be computed fast

1. Sine transforms (8 different types), example: $S = (\sqrt{\frac{2}{n+1}} \sin \frac{\pi ij}{n+1})$
2. Cosine transforms (8 different types), example $C = (\sqrt{\frac{2}{n}} \cos \frac{\pi(2i+1)(2j+1)}{4n})$
3. Hartley transform $H = \sqrt{\frac{1}{n}}(\cos \frac{2\pi ij}{n} + \sin \frac{2\pi ij}{n})$

Trigonometric matrix algebras and Fast multiplication

Given the row vector $[a_0, a_1, \ldots, a_{n-1}]$, the $n \times n$ matrix

$$A = (a_{j-i \mod n})_{i,j=1,n} = \begin{bmatrix} a_0 & a_1 & \ldots & a_{n-1} \\ a_{n-1} & a_0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & a_1 \\ a_1 & \ldots & a_{n-1} & a_0 \end{bmatrix}$$

is called the *circulant* matrix associated with $[a_0, a_1, \ldots, a_{n-1}]$ and is denoted by $\mathrm{Circ}(a_0, a_1, \ldots, a_{n-1})$.

If $a_i = A_i$ are $m \times m$ matrices we have a block circulant matrix

Any circulant matrix A can be viewed as a polynomial with coefficients a_i in the unit circulant matrix S defined by its first row $(0, 1, 0, \ldots, 0)$

$$A = \sum_{i=0}^{n-1} a_i S^i, \quad S = \begin{bmatrix} 0 & 1 & & \\ \vdots & \ddots & \ddots & \\ & & \ddots & 1 \\ 0 & & & \\ 1 & 0 & \ldots & 0 \end{bmatrix}$$

Clearly, $S^n - I = 0$ so that circulant matrices form a matrix algebra isomorphic to the algebra of polynomials with the product modulo $x^n - 1$

Trigonometric matrix algebras and Fast multiplication

If A is a circulant matrix with first row r^T and first column c, then

$$A = \frac{1}{n}\Omega_n^* \,\mathrm{Diag}(w)\Omega_n \;=\; F^* \mathrm{Diag}(w) F$$

where $w = \Omega_n c = \Omega_n^* r$.

Consequences

$$Ax = \mathrm{DFT}_n(\mathrm{IDFT}_n(c) * \mathrm{IDFT}_n(x))$$

where "$*$" denotes the Hadamard, or component-wise product of vectors.

The product Ax of an $n \times n$ circulant matrix A and a vector x, as well as the product of two circulant matrices can be computed by means of two IDFTs and a DFT of length n in $O(n \log n)$ ops

$$A^{-1} = \frac{1}{n}\Omega_n^* \,\mathrm{Diag}(w^{-1})\Omega_n,$$

The inverse of a circulant matrix can be computed in $O(n \log n)$ ops

Trigonometric matrix algebras and Fast multiplication

The definition of circulant matrix is naturally extended to block matrices where $a_i = A_i$ are $m \times m$ matrices.

The inverse of a block circulant matrix can be computed by means of $2m^2$ IDFTs of length n and n inversions of $m \times m$ matrices for the cost of $O(m^2 n \log n + nm^3)$

The product of two block circulant matrices can be computed by means of $2m^2$ IDFTs, m^2 DFT of length n and n multiplications of $m \times m$ matrices for the cost of $O(m^2 n \log n + nm^3)$.

z-circulant matrices

A generalization of circulant matrices is provided by the class of z-circulant matrices.

Given a scalar $z \neq 0$ and the row vector $[a_0, a_1, \ldots, a_{n-1}]$, the $n \times n$ matrix

$$A = \begin{bmatrix} a_0 & a_1 & \ldots & a_{n-1} \\ za_{n-1} & a_0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & a_1 \\ za_1 & \ldots & za_{n-1} & a_0 \end{bmatrix}$$

is called the *z-circulant* matrix associated with $[a_0, a_1, \ldots, a_{n-1}]$.

Denote by S_z the z-circulant matrix whose first row is $[0, 1, 0, \ldots, 0]$, i.e.,

$$S_z = \begin{bmatrix} 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \ddots & \ddots & 0 & 1 \\ z & 0 & \ldots & 0 & 0 \end{bmatrix},$$

z-circulant matrices

- Any z-circulant matrix can be viewed as a polynomial in S_z.

$$A = \sum_{i=0}^{n-1} a_i S_z^i.$$

- $S_{z^n} = zD_z S D_z^{-1}, \qquad D_z = \mathrm{Diag}(1, z, z^2, \ldots, z^{n-1})$, where S is the unit circulant matrix.
- If A is the z^n-circulant matrix with first row r^T and first column c then

$$A = \frac{1}{n} D_z \Omega_n^* \,\mathrm{Diag}(w) \Omega_n D_z^{-1},$$

  with $w = \Omega_n^* D_z r = \Omega_n D_z^{-1} c$.
- Multiplication of z-circulants costs 2 IDFTs, 1 DFT and a scaling
- Inversion of a z-circulant costs 1 IDFT, 1 DFT, n inversions and a scaling
- The extension to block matrices trivially applies to z-circulant matrices.

Embedding Toeplitz matrices into circulants

An $n \times n$ Toeplitz matrix $A = (t_{i,j})$, $t_{i,j} = a_{j-i}$, can be embedded into the $2n \times 2n$ circulant matrix B whose first row is $[a_0, a_1, \ldots, a_{n-1}, *, a_{-n+1}, \ldots, a_{-1}]$, where $*$ denotes any number.

$$B = \left[\begin{array}{ccc|ccc} a_0 & a_1 & a_2 & * & a_{-2} & a_{-1} \\ a_{-1} & a_0 & a_1 & a_2 & * & a_{-2} \\ a_{-2} & a_{-1} & a_0 & a_1 & a_2 & * \\ \hline * & a_{-2} & a_{-1} & a_0 & a_1 & a_2 \\ a_2 & * & a_{-2} & a_{-1} & a_0 & a_1 \\ a_1 & a_2 & * & a_{-2} & a_{-1} & a_0 \end{array}\right].$$

More generally, an $n \times n$ Toeplitz matrix can be embedded into a $q \times q$ circulant matrix for any $q \geq 2n - 1$.

Consequence: the product $y = Ax$ of an $n \times n$ Toeplitz matrix A and a vector x can be computed in $O(n \log n)$ ops.

Embedding Toeplitz matrices into circulants

$y = Ax$,

$$\begin{bmatrix} y \\ w \end{bmatrix} = B \begin{bmatrix} x \\ 0 \end{bmatrix} = \begin{bmatrix} A & H \\ H & A \end{bmatrix} \begin{bmatrix} x \\ 0 \end{bmatrix} = \begin{bmatrix} Ax \\ Hx \end{bmatrix}$$

- embed the Toeplitz matrix A into the circulant matrix B
- embed the vector x into the vector $v = \left[\begin{smallmatrix} x \\ 0 \end{smallmatrix}\right]$
- compute the product $u = Bv$
- set $y = (u_1, \ldots, u_n)^T$

Cost: 3 FFTs of order $2n$, that is $O(n \log n)$ ops medskip
Similarly, the product $y = Ax$ of an $n \times n$ block Toeplitz matrix with $m \times m$ blocks and a vector $x \in \mathbb{C}^{mn}$ can be computed in $O(m^2 n \log n)$ ops.

Triangular Toeplitz matrices

Let $Z = (z_{i,j})_{i,j=1,n}$ be the $n \times n$ matrix

$$Z = \begin{bmatrix} 0 & & & 0 \\ 1 & \ddots & & \\ & \ddots & \ddots & \\ 0 & & 1 & 0 \end{bmatrix},$$

Clearly $Z^n = 0$, moreover, given the polynomial $a(x) = \sum_{i=0}^{n-1} a_i x^i$, the matrix $a(Z) = \sum_{i=0}^{n-1} a_i Z^i$ is a lower triangular Toeplitz matrix defined by its first column $(a_0, a_1, \ldots, a_{n-1})^T$

$$a(Z) = \begin{bmatrix} a_0 & & & 0 \\ a_1 & a_0 & & \\ \vdots & \ddots & \ddots & \\ a_{n-1} & \ldots & a_1 & a_0 \end{bmatrix}.$$

The set of lower triangular Toeplitz matrices forms an algebra isomorphic to the algebra of polynomials with the product modulo x^n.

Inverting a triangular Toeplitz matrix

The inverse matrix T_n^{-1} is still a lower triangular Toeplitz matrix defined by its first column v_n. It can be computed by solving the system $T_n v_n = e_1$

Let $n = 2h$, h a positive integer, and partition T_n into $h \times h$ blocks

$$T_n = \left[\begin{array}{c|c} T_h & 0 \\ \hline W_h & T_h \end{array}\right],$$

where T_h, W_h are $h \times h$ Toeplitz matrices and T_h is lower triangular.

$$T_n^{-1} = \left[\begin{array}{c|c} T_h^{-1} & 0 \\ \hline -T_h^{-1} W_h T_h^{-1} & T_h^{-1} \end{array}\right].$$

The first column v_n of T_n^{-1} is given by

$$v_n = T_n^{-1} e_1 = \begin{bmatrix} v_h \\ -T_h^{-1} W_h v_h \end{bmatrix} = \begin{bmatrix} v_h \\ -L(v_h) W_h v_h \end{bmatrix},$$

where $L(v_h) = T_h^{-1}$ is the lower triangular Toeplitz matrix whose first column is v_h.

Inverting a triangular Toeplitz matrix

The same relation holds if T_n is block triangular Toeplitz. In this case, the elements $a_0, \ldots, a_{n-1}$ are replaced with the $m \times m$ blocks $A_0, \ldots, A_{n-1}$ and v_n denotes the first block column of T_n^{-1}.

Recursive algorithm for computing v_n (block case)

INPUT: $n = 2^k$, $A_0, \ldots, A_{n-1}$

OUTPUT: v_n

COMPUTATION:

1. Set $v_1 = A_0^{-1}$
2. For $i = 0, \ldots, k-1$, given v_h, $h = 2^i$:
   1. Compute the block Toeplitz matrix-vector products $w = W_h v_h$ and $u = -L(v_h)w$.
   2. Set
$$v_{2h} = \begin{bmatrix} v_h \\ u \end{bmatrix}.$$

Cost: $O(n \log n)$ ops

z-circulant and triangular Toeplitz matrices

If $\epsilon = |z|$ is "small" then a z-circulant approximates a triangular Toeplitz

$$\begin{bmatrix} a_0 & a_1 & \dots & a_{n-1} \\ za_{n-1} & a_0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & a_1 \\ za_1 & \dots & za_{n-1} & a_0 \end{bmatrix} \approx \begin{bmatrix} a_0 & a_1 & \dots & a_{n-1} \\ & a_0 & \ddots & \vdots \\ & & \ddots & a_1 \\ & & & a_0 \end{bmatrix}$$

Inverting a z-circulant is less expensive than inverting a triangular Toeplitz (roughly by a factor of 10/3)

The advantage is appreciated in a parallel model of computation, over multithreading architectures

z-circulant and triangular Toeplitz matrices

Numerical algorithms for approximating the inverse of (block) triangular Toeplitz matrices. Main features:

- Total error=approximation error + rounding errors
- Rounding errors grow as $\mu\epsilon^{-1}$, approximation errors are polynomials in z
- the smaller ϵ the better the approximation, but the larger the rounding errors
- good compromise: choose ϵ such that $\epsilon = \mu\epsilon^{-1}$. This implies that the total error is $O(\mu^{1/2})$: half digits are lost

Different strategies have been designed to overcome this drawback

z-circulant and triangular Toeplitz matrices

Assume to work over $\mathbb{R}$

- (interpolation) The approximation error is a polynomial in z. Approximating twice the inverse with, say $z = \epsilon$ and $z = -\epsilon$ and taking the arithmetic mean of the results the approximation error becomes a polynomial in ϵ^2.
  $\Rightarrow$ total error$= O(\mu^{2/3})$
- (generalization) Approximate k times the inverse with values $z_1 = \epsilon\omega_k^i$, $i = 0, \ldots, k-1$. Take the arithmetic mean of the results and get the error $O(\epsilon^k)$.
  $\Rightarrow$ total error$= O(\mu^{k/(k+1)})$.
  Remark: for $k = n$ the approximation error is zero

z-circulant and triangular Toeplitz matrices

- (Higham trick) Choose $z = \underline{i}\epsilon$ then the approximation error affecting the real part of the computed approximation is $O(\epsilon^2)$.
  $\Rightarrow$ total error$= O(\mu^{2/3})$, i.e., only $1/3$ of digits are lost
- (combination) Choose $z_1 = \epsilon(1 + \underline{i})/\sqrt{2}$ and $z_2 = -z_1$; apply the algorithm with $z = z_1$ and $z = z_2$; take the arithmetic mean of the results. The approximation error on the real part turns out to be $O(\epsilon^4)$. The total error is $O(\mu^{4/5})$. Only $1/5$ of digits are lost.
- (replicating the computation) In general choosing as z_j the kth roots of $\underline{i}$ and performing k inversions the error becomes $O(\mu^{2k/(2k+1)})$, i.e., only $1/2h$ of digits are lost

Other matrix algebras

With any trigonometric transform G we may associate the matrix algebra $\{A = GDG^{-1},\ D \text{ diagonal}\}$. These classes are closely related to Toeplitz matrices

- Sine transform $G = \sqrt{\frac{2}{n+1}}(\sin(ij\frac{\pi}{n+1}))$
  τ-algebra generated by $S = \text{tridiag}_n(1, 0, 1)$
- Sine transform $G = \sqrt{\frac{4}{2n+1}}(\sin(i(2j-1)\frac{\pi}{2n+1}))$
  algebra generated by $S = \text{tridiag}_n(1, 0, 1) + e_1 e_1^T$
- There are 8 cosine transforms. For instance the DCT-IV is
  $G = \sqrt{\frac{2}{n}}(\cos\frac{\pi}{n}(i + 1/2)(j + 1/2))$
- The Hartley transform $G = \sqrt{\frac{1}{n}}(\cos(ij\frac{\pi}{n}) + \sin(ij\frac{\pi}{n}))$
  $\Rightarrow$ Hartley algebra which contains symmetric circulants

Displacement operators

Recall that $S_z = \begin{bmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ z & & & 0 \end{bmatrix}$ and let $T = \begin{bmatrix} a & b & c & d \\ e & a & b & c \\ f & e & a & b \\ g & f & e & a \end{bmatrix}$

Then

$$S_{z_1} T - T S_{z_2} = \begin{bmatrix} & \uparrow & \\ & & \end{bmatrix} - \begin{bmatrix} & \rightarrow & \\ & & \end{bmatrix}$$

$$= \begin{bmatrix} e & a & b & c \\ f & e & a & b \\ g & f & e & a \\ z_1 a & z_1 b & z_1 c & z_1 d \end{bmatrix} - \begin{bmatrix} z_2 d & a & b & c \\ z_2 c & e & a & b \\ z_2 b & f & e & a \\ z_2 a & g & f & e \end{bmatrix}$$

$$= \left[\begin{array}{c|ccc} * & & & \\ \vdots & & 0 & \\ \hline * & \multicolumn{3}{c}{\ldots \quad *} \end{array}\right] = e_n u^T + v e_1^T \quad \text{(rank at most 2)}$$

$T \to S_{z_1} T - T S_{z_2}$ displacement operator of Sylvester type

$T \to T - S_{z_1} T S_{z_2}^T$ displacement operator of Stein type

If the eigenvalues of S_{z_1} are disjoint from those of S_{z_2} then the operator of Sylvester type is invertible. Tis holds if $z_1 \neq z_2$

If the eigenvalues of S_{z_1} are different from the reciprocal of those of S_{z_2} then the operator of Sylvester type is invertible. This holds if $z_1 z_2 \neq 1$

Displacement operators: Some properties

For simplicity, here we consider $Z := S_0^T = \begin{bmatrix} 0 & & & \\ 1 & \ddots & & \\ & \ddots & \ddots & \\ & & 1 & 0 \end{bmatrix}$

If A is Toeplitz then $\Delta(A) = AZ - ZA$ is such that

$$\Delta(A) = \begin{bmatrix} & \leftarrow & \end{bmatrix} - \begin{bmatrix} & \downarrow & \end{bmatrix} = \left[\begin{array}{cccc|c} a_1 & a_2 & \dots & a_{n-1} & 0 \\ \hline & & & & -a_{n-1} \\ & & 0 & & \vdots \\ & & & & -a_2 \\ & & & & -a_1 \end{array}\right] = VW^T,$$

$$V = \left[\begin{array}{c|c} 1 & 0 \\ 0 & a_{n-1} \\ \vdots & \vdots \\ 0 & a_1 \end{array}\right], \quad W = \left[\begin{array}{c|c} a_1 & 0 \\ \vdots & \vdots \\ a_{n-1} & 0 \\ 0 & -1 \end{array}\right]$$

Any pair $V, W \in \mathbb{F}^{n\times k}$ such that $\Delta(A) = VW^T$ is called *displacement generator* of rank k.

Displacement operators: Some properties

Proposition.

If $A \in \mathbb{F}^{n\times n}$ has first column a and $\Delta(A) = VW^T$, $V, W \in \mathbb{F}^{n\times k}$ then

$$A = L(a) + \sum_{i=1}^{k} L(v_i)L^T(Zw_i), \quad L(a) = \begin{bmatrix} a_1 & & \\ \vdots & \ddots & \\ a_n & \dots & a_1 \end{bmatrix}$$

Proposition.

For $\Delta(A) = AZ - ZA$ it holds that $\Delta(AB) = A\Delta(B) + \Delta(A)B$ and

$$\Delta(A^{-1}) = -A^{-1}\Delta(A)A^{-1}$$

Therefore

$$A^{-1} = L(A^{-1}e_1) - \sum_{i=1}^{k} L(A^{-1}v_i)L^T(ZA^{-T}w_i)$$

In particular, the inverse of a Toeplitz matrix is Toeplitz-like

Displacement operators: Some properties

The Gohberg-Semencul-Trench formula

$$T^{-1} = \frac{1}{x_1}\left(L(x)L^T(Jy) - L(Zy)L^T(ZJx)\right),$$

$$x = T^{-1}e_1, \quad y = T^{-1}e_n, \quad J = \begin{bmatrix} & & 1 \\ & \iddots & \\ 1 & & \end{bmatrix}$$

- The first and the last column of the inverse define all the entries
- Multiplying a vector by the inverse costs $O(n \log n)$

Other operators

Define $\Delta(X) = D_1 X - X D_2$, $D_1 = \mathrm{diag}(d_1^{(1)}, \ldots, d_n^{(1)})$, $D_2 = \mathrm{diag}(d_1^{(2)}, \ldots, d_n^{(2)})$, where $d_i^{(1)} \neq d_j^{(2)}$ for $i \neq j$.

It holds that

$$\Delta(A) = uv^T \quad \Leftrightarrow \quad a_{i,j} = \frac{u_i v_j}{d_i^{(1)} - d_j^{(2)}}$$

Similarly, given $n \times k$ matrices U, V, one finds that

$$\Delta(B) = UV^T \quad \Leftrightarrow \quad b_{i,j} = \frac{\sum_{r=1}^{k} u_{i,r} v_{j,r}}{d_i^{(1)} - d_j^{(2)}}$$

A is said *Cauchy* matrix, B is said *Cauchy-like* matrix

Other operators: Some properties

A nice feature of Cauchy-like matrices is that their Schur complement is still a Cauchy-like matrix

Consider the case $k = 1$: partition the Cauchy-like matrix C as

$$C = \left[\begin{array}{c|ccc} \frac{u_1 v_1}{d_1^{(1)} - d_1^{(2)}} & \frac{u_1 v_2}{d_1^{(1)} - d_2^{(2)}} & \cdots & \frac{u_1 v_n}{d_1^{(1)} - d_n^{(2)}} \\ \hline \frac{u_2 v_1}{d_2^{(1)} - d_1^{(2)}} & & & \\ \vdots & & \widehat{C} & \\ \frac{u_n v_1}{d_n^{(1)} - d_1^{(2)}} & & & \end{array}\right]$$

where $\widehat{C}$ is still a Cauchy-like matrix. The Schur complement is given by

$$\widehat{C} - \left[\begin{array}{c} \frac{u_2 v_1}{d_2^{(1)} - d_1^{(2)}} \\ \vdots \\ \frac{u_n v_1}{d_n^{(1)} - d_1^{(2)}} \end{array}\right] \frac{d_1^{(1)} - d_1^{(2)}}{u_1 v_1} \left[\begin{array}{ccc} \frac{u_1 v_2}{d_1^{(1)} - d_2^{(2)}} & \cdots & \frac{u_1 v_n}{d_1^{(1)} - d_n^{(2)}} \end{array}\right]$$

Other operators: Some properties

The entries of the Schur complement can be written in the form

$$\frac{\widehat{u}_i \widehat{v}_j}{d_i^{(1)} - d_j^{(2)}}, \quad \widehat{u}_i = u_i \frac{d_1^{(1)} - d_i^{(1)}}{d_i^{(1)} - d_1^{(2)}}, \quad \widehat{v}_j = v_j \frac{d_j^{(2)} - d_1^{(2)}}{d_1^{(1)} - d_j^{(2)}}.$$

The values $\widehat{u}_i$ and $\widehat{v}_j$ can be computed in $O(n)$ ops.

The computation can be repeated until the LU decomposition of C is obtained

The algorithm is known as Gohberg-Kailath-Olshevsky (GKO) algorithm

Its overall cost is $O(n^2)$ ops

There are variants which allow pivoting

Algorithms for Toeplitz inversion

Consider $\Delta(A) = S_1 A - A S_{-1}$ where S_1 is the unit circulant matrix and S_{-1} is the unit (-1)-circulant matrix.

We have observed that the matrix $\Delta(A)$ has rank at most 2

Now, recall that $S_1 = F^* D_1 F$, $S_{-1} = D F^* D_{-1} F D^{-1}$, where $D_1 = \mathrm{Diag}(1, \bar{\omega}, \bar{\omega}^2, \ldots, \bar{\omega}^{n-1})$, $D_{-1} = \delta D_1$, $D = \mathrm{Diag}(1, \delta, \delta^2, \ldots, \delta^{n-1})$, $\delta = \omega_n^{1/2} = \omega_{2n}$ so that

$$\Delta(A) = F^* D_1 F A - A D F^* D_{-1} F D^{-1}$$

multiply to the left by F, and to the right by DF^* and discover that

$$D_1 B - B D_{-1} \quad \text{has rank at most 2, where } B = FADF^*$$

That is, B is Cauchy like of rank at most 2.

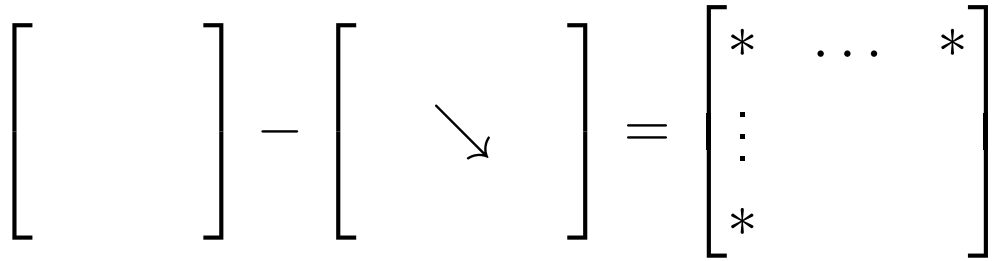
Toeplitz systems can be solved in $O(n^2)$ ops by means of the GKO algorithm

Super fast Toeplitz solvers

The term "fast Toeplitz solvers" denotes algorithms for solving $n \times n$ Toeplitz systems in $O(n^2)$ ops.

The term "super-fast Toeplitz solvers" denotes algorithms for solving $n \times n$ Toeplitz systems in $O(n \log^2 n)$ ops.

Idea of the Bitmead-Anderson superfast solver

Operator $F(A) = A - ZAZ^T =$  $=$.

Partition the matrix as

$$A = \begin{bmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \end{bmatrix}$$

Super fast Toeplitz solver

$$A = \begin{bmatrix} I & 0 \\ A_{2,1}A_{1,1}^{-1} & I \end{bmatrix} \begin{bmatrix} A_{1,1} & A_{1,2} \\ 0 & B \end{bmatrix}, \quad B = A_{2,2} - A_{2,1}A_{1,1}^{-1}A_{1,2}$$

Fundamental property

The Schur complement B is such that $\text{rank}F(A) = \text{rank}F(B)$; the other blocks of the LU factorization have almost the same displacement rank of the matrix A

Solving two systems with the matrix A (for computing the displacement representation of A^{-1}) is reduced to solving two systems with the matrix $A_{1,1}$ for computing $A_{1,1}^{-1}$ and two systems with the matrix B which has displacement rank 2, plus performing some Toeplitz-vector products

Cost: $\quad C(n) = 2C(n/2) + O(n \log n) \quad \Rightarrow \quad C(n) = O(n \log^2 n)$

Trigonometric matrix algebras and preconditioning

The solution of a positive definite $n \times n$ Toeplitz system $A_n x = b$ can be approximated with the Preconditioned Conjugate Gradient (PCG) method

Some features of the Conjugate Gradient (CG) iteration:

- it applies to positive definite systems $Ax = b$
- CG generates a sequence of vectors $\{x_k\}_{k=0,1,2,\ldots}$ converging to the solution in n steps
- each step requires a matrix-vector product plus some scalar products. Cost for Toeplitz systems $O(n \log n)$
- residual error: $\|Ax_k - b\| \leq \gamma\theta^k$, where $\theta = (\sqrt{\mu} - 1)/(\sqrt{\mu} + 1)$, $\mu = \lambda_{\max}/\lambda_{\min}$ is the condition number of A
- convergence is fast for well-conditioned systems, slow otherwise. However:
- (Axelsson-Lindskog) Informal statement: if A has all the eigenvalues in the interval $[\alpha, \beta]$ where $0 < \alpha < 1 < \beta$ except for q outliers which stay outside, then the residual error is bounded by $\gamma_1 \theta_1^{k-q}$ for $\theta_1 = (\sqrt{\mu_1} - 1)/(\sqrt{\mu_1} + 1)$, where $\mu_1 = \beta/\alpha$.

Trigonometric matrix algebras and preconditioning

Features of the Preconditioned Conjugate Gradient (PCG) iteration:

- it consists of the Conjugate Gradient method applied to the system $P^{-1}A_n x = P^{-1}b$, the matrix P is the preconditioner
- The preconditioner P must be choosen so that:
  - solving the system with matrix P is cheap
  - P mimics the matrix A so that $P^{-1}A$ has either condition number close to 1, or has eigenvalues in a narrow interval $[\alpha, \beta]$ containing 1, except for **few outliers**

For Toeplitz matrices P can be chosen in a trigonometric algebra. In this case

- each step of PCG costs $O(n \log n)$
- the spectrum of $P^{-1}A$ is clustered around 1

Example of preconditioners

If A_n is associated with the symbol $a(\theta) = a_0 + 2\sum_{i=1}^{\infty} a_i$ and $a(\theta) \geq 0$, then $\mu(A_n) \to \max a(\theta) / \min a(\theta)$

Choosing $P_n = C_n$, where C_n is the symmetric circulant which minimizes the Frobenius norm $\|A_n - C_n\|_F$, then the eigenvalues of $B_n = P_n^{-1} C_n$ are clustered around 1. That is, for any ϵ there exists n_0 such that the eigenvalues of $P_n^{-1} A$ belong to $[1 - \epsilon, 1 + \epsilon]$ except for a few outliers.

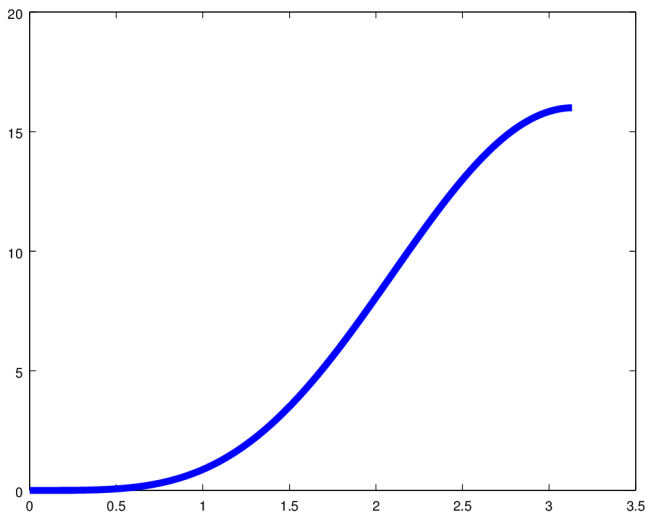
Effective preconditioners can be found in the τ and in the Hartley algebras, as well as in the class of banded Toeplitz matrices

Example of preconditioners

Consider the $n \times n$ matrix A associated with the symbol $a(\theta) = 6 + 2(-4\cos(\theta) + \cos(2\theta))$, that is

$$A = \begin{bmatrix} 6 & -4 & 1 & & \\ -4 & 6 & -4 & 1 & \\ 1 & -4 & \ddots & \ddots & \ddots \\ & \ddots & \ddots & \ddots & \ddots \end{bmatrix}$$

Its eigenvalues are distributed as the symbol $a(\theta)$ and its cond is $O(n^4)$



The eigenvalues of the preconditioned matrix $P^{-1}A$, where P is circulant, are clustered around 1 with very few outliers.

Example of preconditioners

The following figure reports the log of the eigenvalues of A (in red) and of the log of the eigenvalues of $P^{-1}A$ in blue

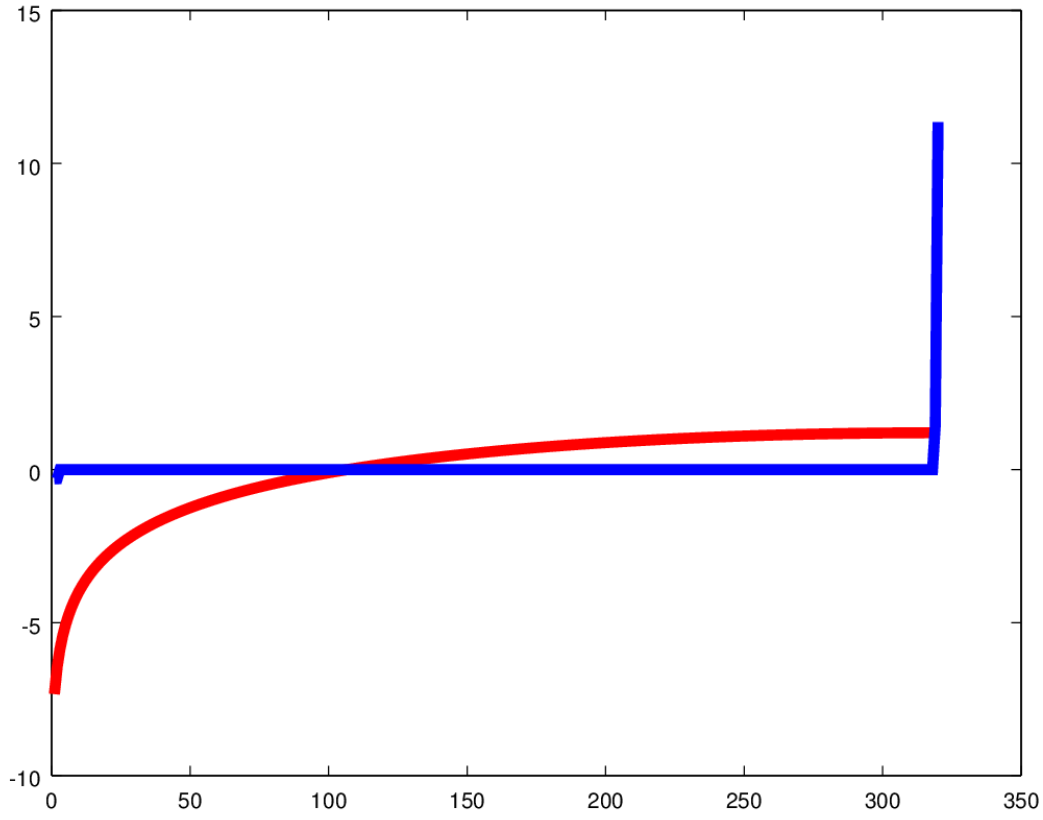


Figure: Log of the eigenvalues of A (in red) and of $P^{-1}A$ in blue

Wiener-Hopf factorization and matrix equations

Consider the equations

$$BX^2 + AX + C = 0, \quad CY^2 + AY + B = 0, \tag{1}$$

where we assume that A, B, C are $n \times n$ matrices and that there exist solutions X, Y with spectral radius $\rho(X) = \eta < 1$, $\rho(Y) = \nu < 1$.

The two equations can be rewritten in terms of infinite block Toeplitz systems. For instance, the first equation takes the form

$$\begin{bmatrix} A & B & & & \\ C & A & B & & \\ & C & A & B & \\ & & \ddots & \ddots & \ddots \end{bmatrix} \begin{bmatrix} X \\ X^2 \\ X^3 \\ \vdots \end{bmatrix} = \begin{bmatrix} -C \\ 0 \\ 0 \\ \vdots \end{bmatrix}.$$

Similarly we can do for the second equation.

This infinite system can be solved by means of the *Cyclic Reduction (CR)* method introduced by Gene Golub for the numerical solution of the discrete Poisson equation over a rectangle and here adjusted to the infinite block Toeplitz case. The CR technique works this way:

- permute block rows and block columns in the above equation by writing the even numbered ones first, followed by the odd numbered ones and get the system

$$\left[\begin{array}{ccc|ccc} A & & & C & B & \\ & A & & & C & \ddots \\ & & \ddots & & & \ddots \\ \hline B & & & A & & \\ C & B & & & A & \\ & \ddots & \ddots & & & \ddots \end{array}\right] \left[\begin{array}{c} X^2 \\ X^4 \\ \vdots \\ \hline X \\ X^3 \\ \vdots \end{array}\right] = \left[\begin{array}{c} 0 \\ 0 \\ \vdots \\ \hline -C \\ 0 \\ \vdots \end{array}\right]$$

- eliminate the unknowns $X^2, X^4, \ldots$ by taking a Schur complement and arrive at the system

$$\begin{bmatrix} \widehat{A}_1 & B_1 & & & \\ C_1 & A_1 & B_1 & & \\ & C_1 & A_1 & B_1 & \\ & & \ddots & \ddots & \ddots \end{bmatrix} \begin{bmatrix} X \\ X^3 \\ X^5 \\ \vdots \end{bmatrix} = \begin{bmatrix} -C \\ 0 \\ 0 \\ \vdots \end{bmatrix}$$

  where

$$\begin{aligned} A_1 &= A_0 - B_0 A_0^{-1} C_0 - C_0 A_0^{-1} B_0 \\ B_1 &= -B_0 A_0^{-1} B_0 \\ C_1 &= -C_0 A_0^{-1} C_0 \\ \widehat{A}_1 &= \widehat{A}_0 - B_0 A_0^{-1} C_0 \end{aligned}$$

  with $A_0 = A, B_0 = B, C_0 = C, \widehat{A}_0 = A$.

where we assume that A is nonsingular.

This latter system has almost the block Toeplitz structure of the original one except for the $(1,1)$ block. Therefore we can repeat the same procedure by generating the sequence of block triangular systems with blocks C_i, A_i, B_i and $\widehat{A}_i$ such that

$$\begin{bmatrix} \widehat{A}_i & B_i & & & \\ C_i & A_i & B_i & & \\ & C_i & A_i & B_i & \\ & & \ddots & \ddots & \ddots \end{bmatrix} \begin{bmatrix} X \\ X^{2^i+1} \\ X^{2*2^i+1} \\ X^{3*2^i+1} \\ \vdots \end{bmatrix} = \begin{bmatrix} -C \\ 0 \\ 0 \\ \vdots \end{bmatrix}$$

where

$$\begin{aligned} A_{i+1} &= A_i - B_i A_i^{-1} C_i - C_i A_i^{-1} B_i \\ B_{i+1} &= -B_i A_i^{-1} B_i \\ C_{i+1} &= -C_i A_i^{-1} C_i \\ \widehat{A}_{i+1} &= \widehat{A}_i - B_i A_i^{-1} C_i \end{aligned}$$

Here, we assume that all the blocks A_i generated this way are nonsingular.

The first equation of this system takes the form

$$\widehat{A}_i X + B_i X^{2^i+1} = -C$$

Moreover, $\|B_i\| = O(\nu^{2^i})$ so that $X_i = -\widehat{A}_i^{-1} C$ provides an approximation to the solution X with error $O((\nu\eta)^{2^i})$

This makes CR one of the fastest algorithms for this kind of problems

Besides this formulation given in terms of Toeplitz matrices, there is a more elegant formulation given in functional form which provides a generalization of the Graeffe iteration. More precisely, define $\varphi_i(z) = z^{-1} C_i + A_i + z B_i$ and find that

$$\varphi_{i+1}(z^2) = \varphi_i(z) A_i^{-1} \varphi_i(-z),$$

that is a generalization to the case of matrix polynomials of the celebrated Graeffe-Lobachewsky-Dandelin iteration (Ostrowski 1940)

Another nice interpretation of CR can be given in terms of the matrix functions $\psi_i(z) = \varphi_i(z)^{-1}$ defined for all the $z \in \mathbb{C}$ where $\varphi_i(z)$ is nonsingular. In fact, one can easily verify that

$$\psi_{i+1}(z^2) = \frac{\psi_i(z) + \psi_i(-z)}{2}$$
$$\psi_0(z) = (z^{-1}C + A + zB)^{-1}$$

This formulation enables one to provide the proof of convergence properties just by using the analyticity of the involved functions.

Moreover, the same formulation allows to define the functions $\psi_i(z)$ in the cases where there is a break-down in the construction of the sequence $\varphi_i(z)$ due to the singularity of some A_i.

The solutions G and R of the matrix equations in (1) provide the Wiener-Hopf factorization of $\varphi(z)$

$$\varphi(z) = (I - zR)W(I - z^{-1}G), \quad W = B + AG$$

which in matrix form takes the following expression

$$\begin{bmatrix} A & B & & \\ C & A & B & \\ & \ddots & \ddots & \ddots \end{bmatrix} = \begin{bmatrix} I & -R & & \\ & I & -R & \\ & & \ddots & \ddots \end{bmatrix} \begin{bmatrix} W & & & \\ & W & & \\ & & \ddots & \end{bmatrix} \begin{bmatrix} I & & & \\ -G & I & & \\ & -G & \ddots & \\ & & \ddots & \end{bmatrix}$$

The same technique can be extended to matrix equations of the kind

$$\sum_{i=-1}^{\infty} A_i X^i = 0$$

and to the computation of the Wiener-Hopf factorization of the function $A(z) = \sum_{i=-1}^{\infty} z^i A_i$, that is, the block UL factorization of the infinite block Toeplitz matrix in block Hessenberg form associated with $A(z)$.

A recent application

In the Erlangian approximation of Markovian fluid queues, one has to compute

$$Y = e^X = \sum_{i=0}^{\infty} \frac{1}{i!} X^i$$

where

$$X = \begin{bmatrix} X_0 & X_1 & \dots & X_\ell \\ & \ddots & \ddots & \vdots \\ & & X_0 & X_1 \\ & & & X_0 \end{bmatrix}, \quad m \times m \text{ blocks } X_0, \dots, X_\ell,$$

X has negative diagonal entries, nonnegative off-diagonal entries, the sum of the entries in each row is nonpositive

Clearly, since block triangular Toeplitz matrices form a matrix algebra then Y is still block triangular Toeplitz

What is the most convenient way to compute Y in terms of CPU time and error?

A recent application

Embed X into an infinite block triangular block Toeplitz matrix X_∞ obtained by completing the sequence $X_0, X_1, \ldots, X_\ell$ with zeros

$$X_\infty = \begin{bmatrix} X_0 & \ldots & X_\ell & 0 & \ldots & \ldots \\ & X_0 & \ddots & X_\ell & 0 & \ddots \\ & & \ddots & \ddots & \ddots & \ddots \\ & & & \ddots & \ddots & \ddots \end{bmatrix}$$

Denote $Y_0, Y_1, \ldots$ the blocks defining $Y_\infty = e^{X_\infty}$

Then Y is the $(\ell+1) \times (\ell+1)$ principal submatrix of Y_∞

We can prove the following decay property

$$\|Y_i\|_\infty \leq e^{\alpha(\sigma^{\ell-1}-1)}\sigma^{-i}, \quad \forall \sigma > 1$$

where $\alpha = \max_j(-(X_0)_{j,j})$.
This property is fundamental to prove error bounds of the following different algorithms

A recent application

Using ϵ-circulant matrices

Approximate X with an ϵ-circulant matrix $X^{(\epsilon)}$ and approximate Y with $Y^{(\epsilon)} = e^{X^{(\epsilon)}}$. We can prove that if, $\beta = \|[X_1, \ldots, X_\ell]\|_\infty$ then

$$\|Y - Y^{(\epsilon)}\|_\infty \leq e^{|\epsilon|\beta} - 1 = |\epsilon|\beta + O(|\epsilon|^2)$$

and, if ϵ is purely imaginary then

$$\|Y - Y^{(\epsilon)}\|_\infty \leq e^{|\epsilon|^2\beta} - 1 = |\epsilon|^2\beta + O(|\epsilon|^4)$$

Using circulant matrices

Embed X into a $K \times K$ block circulant matrix $X^{(K)}$ for $K > \ell$ large, and approximate Y with the $K \times K$ submatrix $Y^{(K)}$ of $e^{X^{(K)}}$.

We can prove the following bound

$$\|[Y_0 - Y_0^{(K)}, \ldots, Y_\ell - Y_\ell^{(K)}]\|_\infty \leq (e^\beta - 1)e^{\alpha(\sigma^{\ell-1}-1)}\frac{\sigma^{-K+\ell}}{1-\sigma^{-1}}, \quad \sigma > 1$$

A recent application

Method based on Taylor expansion

The matrix Y is approximated by truncating the series expansion to r terms

$$Y^{(r)} = \sum_{i=0}^{r} \frac{1}{i!} X^i$$

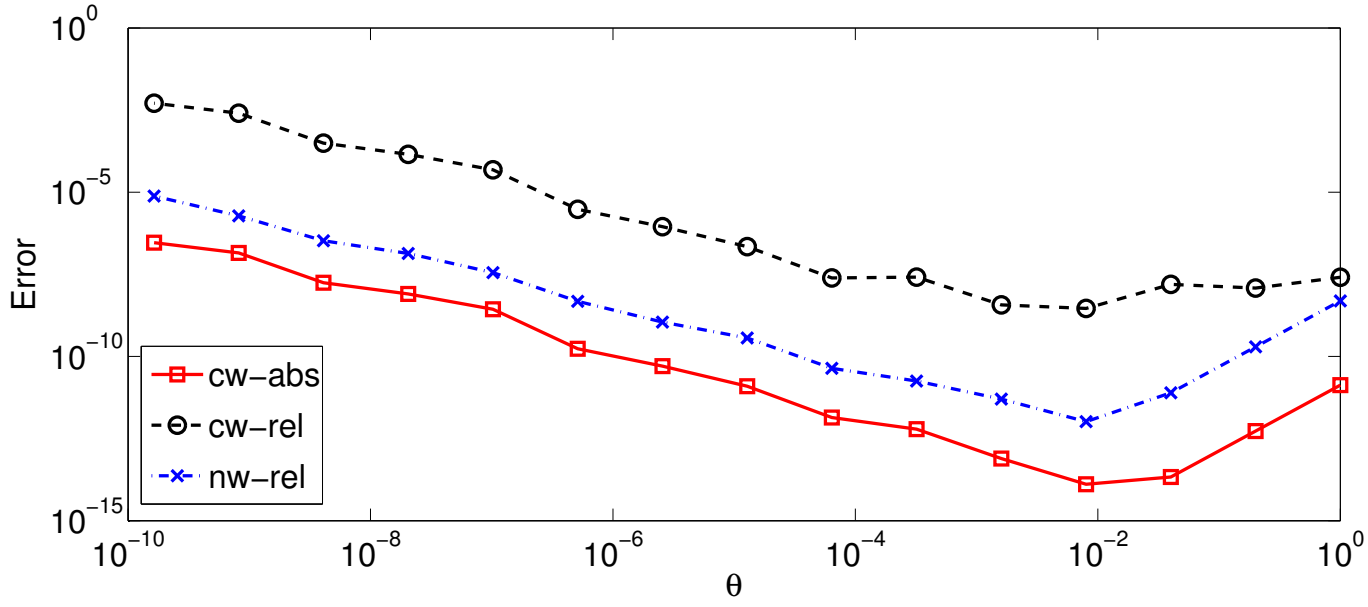


Figure: Norm-wise error, component-wise relative and absolute errors for the solution obtained with the algorithm based on ϵ-circulant matrices with $\epsilon = \underline{i}\theta$.

A recent application

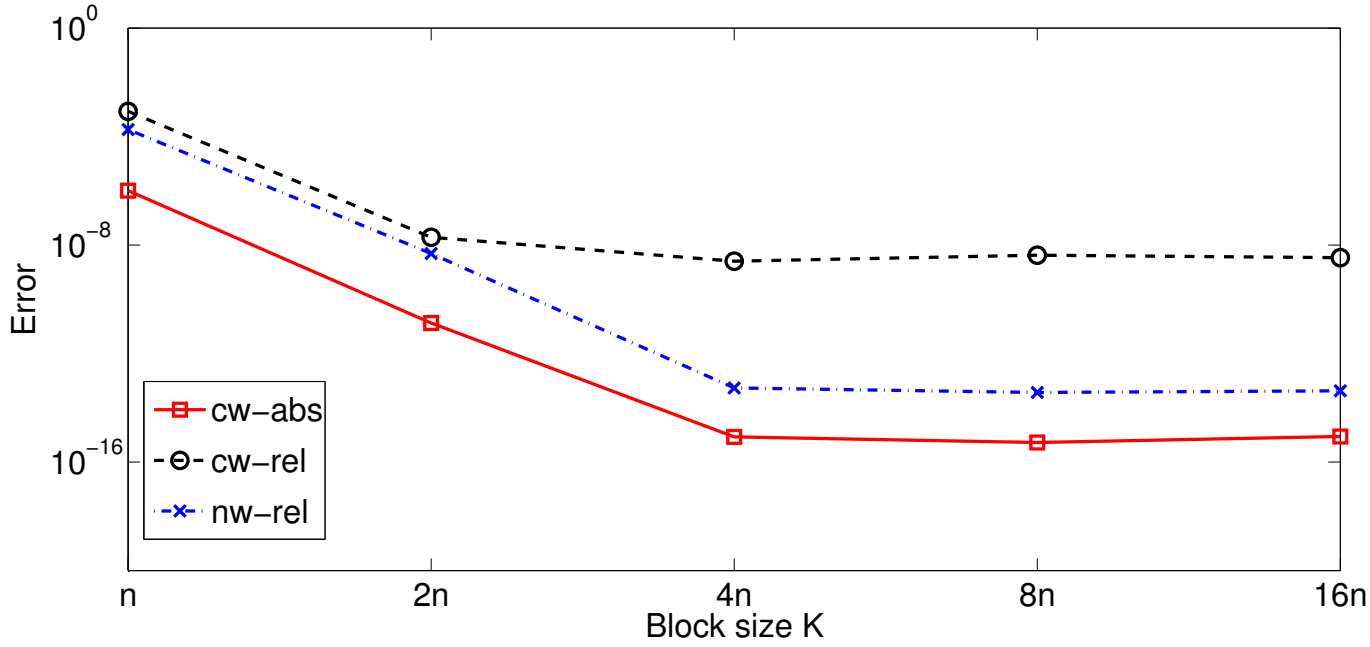


Figure: Norm-wise error, component-wise relative and absolute errors for the solution obtained with the algorithm based on circulant embedding for different values of the embedding size K.

A recent application

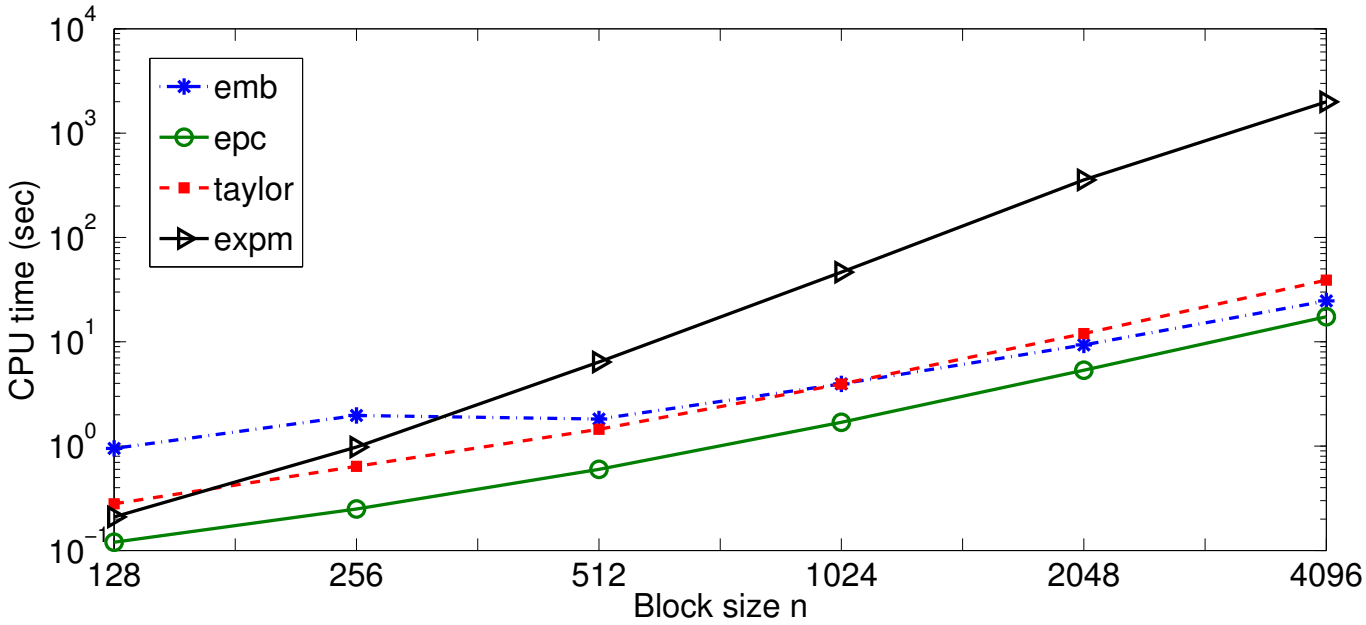
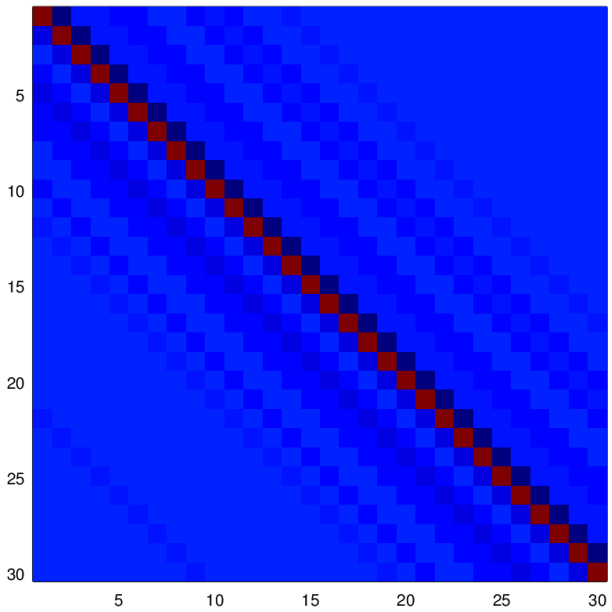


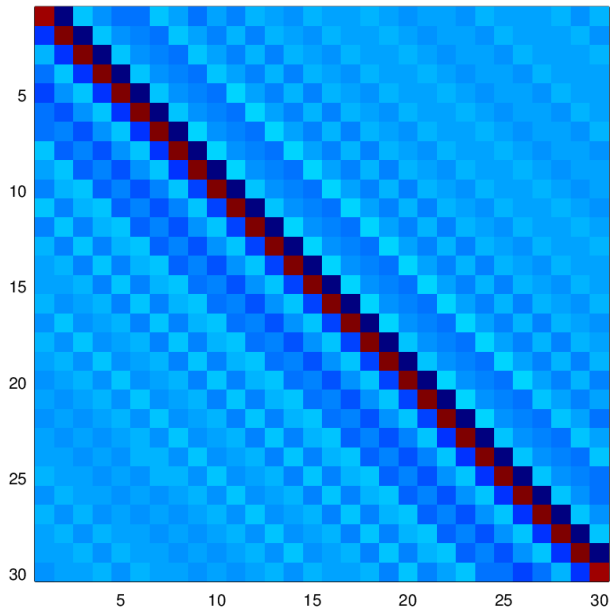
Figure: CPU time of the Matlab function `expm`, and of the algorithms based on ϵ-circulant, circulant embedding, power series expansion.

A recent application

Open issues

Can we prove that the exponential of a general block Toeplitz matrix does not differ much from a block Toeplitz matrix? Numerical experiments confirm this fact but a proof is missing.
Can we design effective ad hoc algorithms for the case of general block Toeplitz matrices?
Can we apply the decay properties of Benzi, Boito 2014 ?

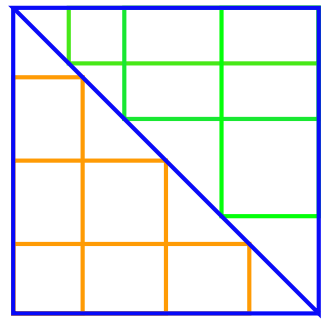




Rank structured matrices

Informally speaking, a rank-structured matrix is a matrix where its submatrices located in some part of its support have low rank

Example of quasi-separable matrices: the submatrices strictly contained in the upper or in the lower triangular part have rank at most 1.



Tridiagonal matrices are quasi-separable

The inverse $B = A^{-1}$ of an irreducible tridiagonal matrix A is quasi-separable

$$b_{i,j} = \begin{cases} u_i v_j & \text{for } i > j \\ w_i z_j & \text{for } i < j \end{cases}$$

that is, $\texttt{triu}(B) = \texttt{triu}(wz^T)$, $\texttt{tril}(B) = \texttt{tril}(uv^T)$
The vectors u, v, w, z are called *generators*

Rank structured matrices

In general, we say that A is (h, k) quasi-separable if the submatrices strictly contained in the lower triangular part have rank at most h, the submatrices strictly contained in the upper triangular part have rank at most k

If $h = k$ we say that A is k quasi-separable

Band matrices are an example of (h, k) quasi-separable matrices and it can be proved that their inverses still share this property

Rank structured matrices are investigated in different fields like integral equations, statistics, vibrational analysis

There is a very wide literature on this subject, and recent books by Van Barel, Vandebril and Mastronardi; Eidelman

Basic properties of rank structured matrices

Let A be k quasi-separable

1. If A is invertible then A^{-1} is k quasi-separable.
2. If $A = LU$ is the LU factorization of A then L and U are quasi-separable of rank $(k, 0)$ and $(0, k)$, respectively
3. If $A = QR$ is a QR factorization of A then Q is quasi-separable of rank k and and U is quasi-separable of rank $(0, 2k)$.
4. The matrices L_i, U_i, A_i defined by the LR iteration $A_i =: L_i U_i$, $A_{i+1} = U_i L_i$ are quasi-separable of rank $(k, 0)$, $(0, k)$, k, respectively

Moreover, there are algorithms for

1. computing A^{-1} in $O(nk^2)$ ops;
2. solving the system $Ax = b$ in $O(nk^2)$ ops;
3. computing the LU and the QR factorization of A in $O(nk^2)$ ops;

Companion matrices

Let $a(x) = \sum_{i=0}^{n} a_i x^i$ be a monic polynomial, i.e., such that $a_n = 1$. A companion matrix associated with $a(x)$ is a matrix A such that $\det(xI - A) = a(x)$

Among the most popular companion matrices we recall the first and second Frobenius forms

$$F_1 = \begin{bmatrix} -a_{n-1} & -a_{n-2} & \dots & -a_0 \\ 1 & 0 & & \\ & \ddots & \ddots & \\ & & 1 & 0 \end{bmatrix}, \quad F_2 = F_1^T,$$

Companion matrices

Both matrices are quasi-separable, F_1 has a generator concerning the upper triangular part while F_2 has a generator concerning the lower triangular part.

Both matrices can be written as an orthogonal (permutation) matrix, that is the unit circulant, plus a correction of rank 1, that is,

$$F_1 = \begin{bmatrix} 0 & \dots & 0 & 1 \\ 1 & \ddots & & 0 \\ & \ddots & \ddots & \vdots \\ & & 1 & 0 \end{bmatrix} - \begin{bmatrix} a_{n-1} & \dots & a_1 & 1+a_0 \\ 0 & \dots & 0 & 0 \\ \vdots & \dots & \vdots & \vdots \\ 0 & \dots & 0 & 0 \end{bmatrix} =: C - uv^T$$

The shifted QR iteration, i.e.,

$$A_i - \alpha_i I =: Q_i R_i, \quad A_{i+1} := R_i Q_i + \alpha_i I$$

generates quasiseparable matrices in the form unitary plus low-rank

There are algorithms of cost $O(n)$ for performing the QR step (Gu, Xia, Zhu, Chandrasekaran; Boito, Eidelman, Gemignani; Aurentz, Mach, Vandebril, Watkins; Frederix, Delvaux, Van Barel, Van Dooren; Del Corso)

Comrade matrix

Define the sequence of orthogonal polynomials $p_i(x)$ satisfying the following three-term recurrence

$$p_0(x) = 1, \quad p_1(x) = x - b_1,$$
$$p_{i+1}(x) = (x - b_{i+1})p_i(x) - c_i p_{i-1}(x), \quad i = 1, 2, \ldots, n-1,$$

where $c_i > 0$. Consider a monic polynomial $p(x)$ represented in this orthogonal basis as $p(x) = \sum_{i=0}^{n} d_i p_i(x), \quad d_n = 1$ Then $p(x) = \det(xI - A)$, where A is the *comrade* matrix (Barnett 1975)

$$A = \begin{bmatrix} b_1 & c_1 & & \\ 1 & b_2 & \ddots & \\ & \ddots & \ddots & c_{n-1} \\ & & 1 & b_n \end{bmatrix} - [0, \ldots, 0, 1] \begin{bmatrix} d_0 \\ \vdots \\ d_{n-3} \\ \widehat{d}_{n-2} \\ \widehat{d}_{n-1} \end{bmatrix}$$

and $\widehat{d}_{n-2} = -d_{n-2} + c_{n-1}$, $\widehat{d}_{n-1} = -d_{n-1} + b_n$.

This matrix is $(1, 2)$ quasi-separable

Colleague matrix

Another companion matrix is the *colleague matrix* (Good 1961, Werner 1983)

$$C = \begin{bmatrix} x_1 & & & & -a_0 \\ 1 & x_2 & & & -a_1 \\ & 1 & \ddots & & \vdots \\ & & \ddots & x_{n-1} & -a_{n-2} \\ & & & 1 & x_n - a_{n-1} \end{bmatrix}$$

This matrix provides the representation of a polynomial $p(x)$ in the Newton basis. More precisely, one can prove that

$$\begin{aligned} \det(xI - C) =& a_0 + a_1(x - x_1) + a_2(x - x_1)(x - x_2) + \cdots \\ & + a_{n-1} \prod_{i=1}^{n-1} (x - x_i) + \prod_{i=1}^{n} (x - x_i). \end{aligned}$$

Arrowhead companion matrix

Similarly, given a monic polynomial $p(x)$ of degree n, choose n pairwise different values $x_0, x_1, \ldots, x_{n-1}$ and consider the *arrowhead companion* pencil of size $n+1$ defined by $xC_1 - C_0$ where

$$C_0 = \begin{bmatrix} x_0 & & & & p_0 \\ & x_1 & & & p_1 \\ & & \ddots & & \vdots \\ & & & x_{n-1} & p_{n-1} \\ -\ell_0 & -\ell_1 & \ldots & -\ell_{n-1} & 0 \end{bmatrix}, \qquad C_1 = \mathrm{diag}(1, 1, \ldots, 1, 0),$$

$$\ell_i = 1/\prod_{j=1, j\neq i}^{n-1}(x_i - x_j)$$

$$p_i = p(x_i)$$

Computing $\det(xC_1 - C_0)$ by means of the Laplace rule along the last column provides the following expression

$$\det(xC_1 - C_0) = \sum_{i=0}^{n} p_i L_i(x), \quad L_i(x) = \prod_{j=1, j\neq i}^{n-1} (x - x_j),$$

that is, the Lagrange representation of the polynomial $p(x)$. Also the pencil $xC_1 - C_0$ is quasiseparable of rank 1.

Smith companion matrix

The *Smith companion* matrix given by Smith in 1970 and considered by Golub in 1973, has the following form

$$S = \mathrm{diag}(b_1, \ldots, b_n) - ew^T, \quad e = (1, \ldots, 1)^T,$$
$$w = (w_i), \; w_i = \frac{p(b_i)}{\prod_{j=1, j\neq i}^{n}(b_i - b_j)}$$

where $p(x)$ is a monic polynomial of degree n, and $b_1, \ldots, b_n$ are pairwise different numbers.

It is easy to show that $\det(xI - S) = p(x)$, that is, S is a companion matrix for $p(x)$. Also in this case, S is a quasiseparable matrix given in terms of a generator. In fact S is expressed as a diagonal plus a rank 1 matrix.

Smith companion

Applications

- locating the zeros of $p(x)$: the set of disks of center x_i and radius $r_i = n \left| p(x_i) / \prod_{j=1, j\neq i}^{n} \right|$ is a set of inclusion disks
- The polynomial root-finding problem reduced to an eigenvalue problem leads to the *secular equation*

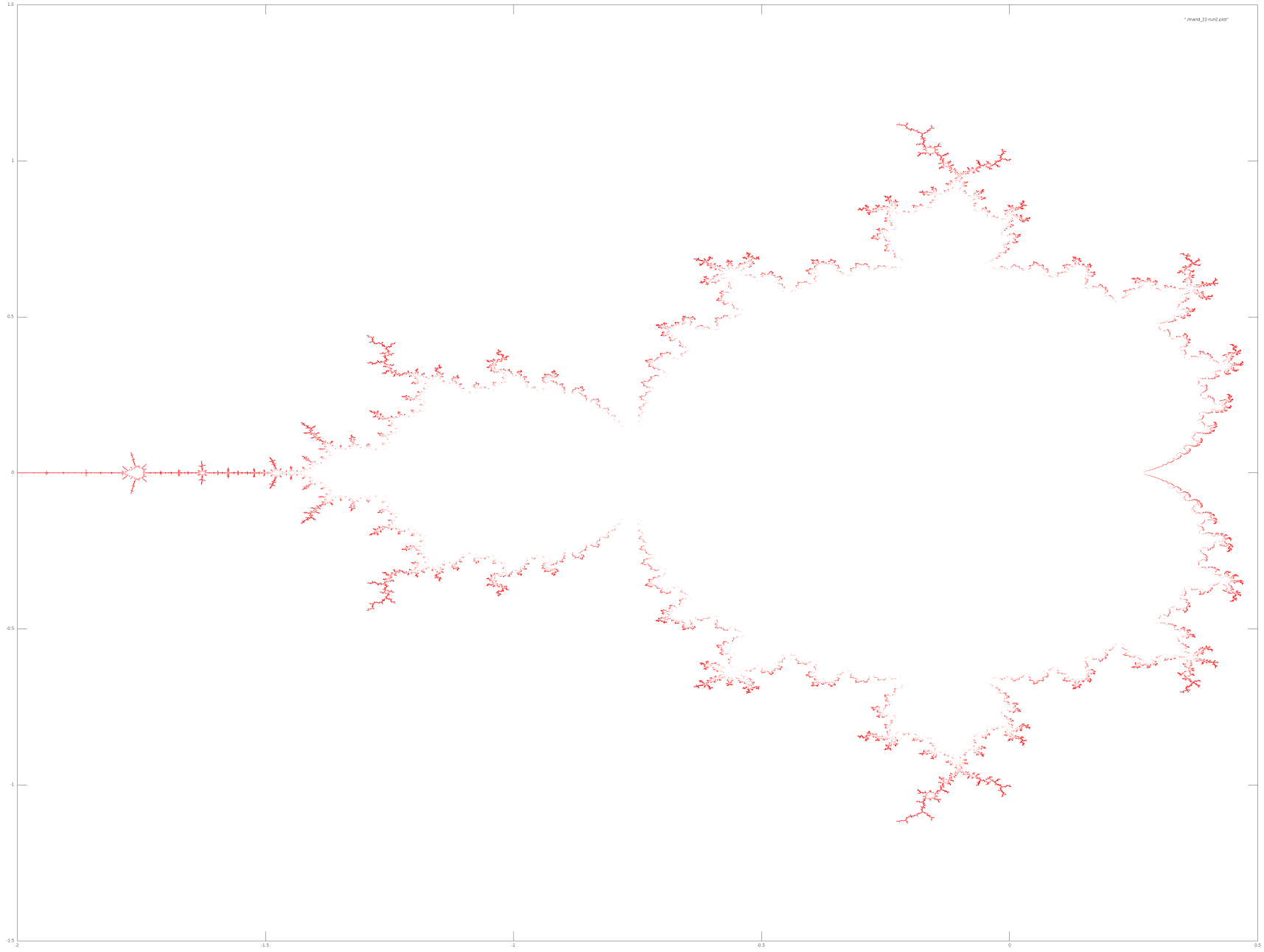
$$\sum_{i=1}^{n} \frac{w_i}{x - b_i} - 1 = 0$$

- the condition number of the zeros of $p(x)$ as function of w_i converges to zero as b_i converge to the polynomial zeros (B., Robol 2014)
- These properties are used in the package MPSolve v.3.1.4 to approximate polynomial zeros with any guaranteed precision.

On computing polynomial zeros

At the moment the fastest software for computing polynomial zeros is MPSolve `http://numpi.dm.unipi.it/mpsolve`

- It relies on Aberth iteration and on the Smith companion matrix
- Its cost is $O(n^2)$ ops per step. In most cases, the number of iterations is independent of n.
- It can exploit multi-core architectures
- On a 20 core computer it can solve the Mandelbrot polynomial of degree 2^{20} in a couple of days of CPU, about a week is needed for degree 2^{21} and about one month for degree 2^{22}
- In principle the Aberth iteration, used for shrinking inclusion disks, could be replaced by the QR iteration based on quasi-separable matrix technology
- At the moment, the Ehrlich-Aberth iteration still performs better than the best available QR algorithms


"./mand_22-run2.plot"
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-1.5
-1
-0.5
0
0.5

Extensions to matrix polynomials

Given $m \times m$ matrices A_i, $i = 0, \ldots, A_n$, with $A_n \neq 0$, we call $A(x) = \sum_{i=0}^{n} x^i A_i$ a matrix polynomial of degree n. The polynomial eigenvalue problem consists in computing the solutions of the polynomial equation $\det A(x) = 0$, given the matrix coefficients of $A(x)$. Throughout, we assume that $A(x)$ is regular, that is $\det A(x)$ is not constant.

the pencil

$$\mathcal{A}(x) = x \begin{bmatrix} I & & & \\ & \ddots & & \\ & & I & \\ & & & A_n \end{bmatrix} - \begin{bmatrix} -A_{n-1} & -A_{n-2} & \cdots & -A_0 \\ I & 0 & & \\ & \ddots & \ddots & \\ & & I & 0 \end{bmatrix}.$$

is such that $\det A(x) = \det \mathcal{A}(x)$

Extensions to matrix polynomials

Similarly, we can extend to matrix polynomials the colleague and the comrade companion. In fact the pencil

$$\mathcal{A}(x) = \begin{bmatrix} (x-x_1)I & & & & A_0 \\ -I & (x-x_2)I & & & A_1 \\ & -I & \ddots & & \vdots \\ & & \ddots & (x-x_{n-1})I & A_{n-2} \\ & & & -I & (x-x_n)A_n + A_{n-1} \end{bmatrix}$$

is such that $\det \mathcal{A}(x) = \det A(x)$, thus provides an extension of the colleague pencil to matrix polynomials.

Extensions to matrix polynomials

Similarly, representing $A(x)$ in the basis formed by the orthogonal monic polynomials $p_i(x)$, $i = 0, \ldots, n$ such that $A(x) = \sum_{i=0}^{n} D_i p_i(x)$, then the extension of the comrade pencil is

$$\mathcal{A}(x) = x\mathrm{diag}(I, \ldots, I, D_n) - \begin{bmatrix} b_1 I & c_1 I & & \\ I & b_2 I & \ddots & \\ & \ddots & \ddots & c_{n-1} I \\ & & I & b_n I \end{bmatrix} + [0, \ldots, 0, I] \begin{bmatrix} D_0 \\ \vdots \\ D_{n-3} \\ \widehat{D}_{n-2} \\ \widehat{D}_{n-1} \end{bmatrix}$$

where $\widehat{D}_{n-1} = D_{n-1} + b_n D_n$ and $\widehat{D}_{n-2} = D_{n-2} + c_{n-1} D_n$

That is, one can prove that $\det A(x) = \det \mathcal{A}(x)$.

Extension to matrix polynomials: a recent result

A first generalization of the Smith companion matrix (B., Robol)

Let $b_i(x)$ be polynomials of degree d_i for $i = 1, \ldots, k$ such that $\sum_{i=1}^{k} d_i = n$ and $\gcd(b_i(x), b_j(x)) = 1$ for $i \neq j$

Define $b(x) = \prod_{i=1}^{k} b_i(x), \quad c_i(x) = \prod_{j=1, j\neq i}^{n} b_j(x)$

Then there exists unique the decomposition

$$p(x) = b(x) + \sum_{i=1}^{k} w_i(x) c_i(x)$$

$$w_i(x) = p(x)/c_i(x) \mod b_i(x)$$

Consequently,

$$p(x) = \det P(x), \quad P(x) = \begin{bmatrix} b_1(x) & & \\ & \ddots & \\ & & b_k(x) \end{bmatrix} + \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} [w_1(x), \ldots, w_k(x)]$$

$P(x)$ is an ℓ-ification of $p(x)$ where $\ell = \max_i d_i$

Extension to matrix polynomials: a recent result

Remarks:

- If $k = n$, then $d_i = 1$, $b_i(x) = x - \beta_i$ and we get the Smith companion form

$$P(x) = xI - \left(\begin{bmatrix} \beta_1 & & \\ & \ddots & \\ & & \beta_n \end{bmatrix} - \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} [w_1, \ldots, w_n] \right)$$

- The left and right eigenvectors of the matrix polynomial $P(x)$ can be explicitly given in terms of the zeros $\xi_1, \ldots, \xi_n$ of $p(x)$
- For $k = n$, if β_i are close to ξ_i then the eigenvalues of $P(x)$ are well conditioned. More precisely $\lim_{\beta_i \to \xi_i} \mathrm{cond}(\xi_j) = 0$ for any j.
  In the case of multiple zeros, the property is true provided that all the b_is converging to a multiple root do not collapse before convergence
  (B., Robol, J.CAM 2014)

Extension to matrix polynomials: a recent result

- Let $A(x) = \sum_{i=0}^{n} A_i x^i$, $A_i \in \mathbb{C}^{m\times m}$ be nondegenerate, $A_n \neq 0$
- Let $b_i(x)$ be pairwise prime monic polynomials of degree d_i, $i = 1, \ldots, k$ such that $\sum_{i=1}^{k} d_i = n$
- Define $B_i(x) = b_i(x)I$, $i = 1, \ldots, k-1$, $B_k(x) = b_k(x)A_n + sI$ where $s \in \mathbb{C}$ is such that $\det B_k(x) \neq 0$ for x zero of $b_i(x)$.

Extension to matrix polynomials: a recent result

- Let $A(x) = \sum_{i=0}^{n} A_i x^i$, $A_i \in \mathbb{C}^{m\times m}$ be nondegenerate, $A_n \neq 0$
- Let $b_i(x)$ be pairwise prime monic polynomials of degree d_i, $i = 1, \ldots, k$ such that $\sum_{i=1}^{k} d_i = n$
- Define $B_i(x) = b_i(x)I$, $i = 1, \ldots, k-1$, $B_k(x) = b_k(x)A_n + sI$ where $s \in \mathbb{C}$ is such that $\det B_k(x) \neq 0$ for x zero of $b_i(x)$.

Then there exists unique the decomposition

$$A(x) = B(x) + \sum_{i=1}^{k} W_i(x)C_i(x)$$

where $B(x) = \prod_{i=1}^{k} B_i(x)$, $\quad C_i(x) = \prod_{j=1,\, j\neq i}^{k} B_j(x)$ and

$$W_i(x) = \frac{A(x)}{\prod_{j=1,\, j\neq i}^{k-1} b_j(x)} B_k(x)^{-1} \mod b_i(x), \quad i = 1, \ldots, k-1$$

$$W_k(x) = \frac{A(x)}{\prod_{j=1}^{k-1} b_j(x)} - sI - s\sum_{j=1}^{k-1} \frac{W_j(x)}{b_j(x)} \mod b_k(x)$$

Moreover

$$\det A(x) = \det \mathcal{A}(x), \quad \mathcal{A}(x) = D(x) + \begin{bmatrix} I \\ \vdots \\ I \end{bmatrix} [W_1(x), \ldots, W_k(x)]$$

where

$$D(x) = \begin{bmatrix} b_1(x)I & & & \\ & \ddots & & \\ & & b_{k-1}(x)I & \\ & & & b_k(x)A_n + sI \end{bmatrix}$$

Remarks

- For $k = n$, $b_i(x) = x - \beta_i$ one has

$$W_i = \frac{A(\beta_i)}{\prod_{j=1, j\neq i}^{n-1}(\beta_i - \beta_j)}((\beta_i - \beta_n)A_n + sI)^{-1}$$

$$W_n = \frac{A(\beta_n)}{\prod_{j=1}^{n-1}(\beta_n - \beta_j)} - sI - s\sum_{j=1}^{n-1}\frac{W_j}{(\beta_n - \beta_j)}$$

- Moreover,

$$\mathcal{A} = x\begin{bmatrix} I & & \\ & \ddots & \\ & & I \\ & & & A_n \end{bmatrix} - \begin{bmatrix} \beta_1 I & & & \\ & \ddots & & \\ & & \beta_{n-1} I & \\ & & & \beta_n A_n + sI \end{bmatrix} + \begin{bmatrix} I \\ I \\ \vdots \\ I \end{bmatrix} [W_1, \ldots, W_n]$$

- For $A_n = I$, one may choose $s = 0$ and get

$$W_i = \frac{A(\beta_i)}{\prod_{j=1, j\neq i}^{n}(\beta_i - \beta_j)}, \quad i = 1, \ldots, n$$

Remarks

- Assume for simplicity $A_n = I$. Set $k = n$, $\beta_i = \omega_n^i$, where ω_n is a primitive nth root of 1. Define $F_n = \frac{1}{\sqrt{n}}(\omega_n^{ij})_{i,j=1,n}$ the Fourier matrix. Then

$$(F^* \otimes I)\mathcal{A}(x)(F \otimes I) = xI - C$$

  where C is the block Frobenius matrix associated with $A(x)$
- If $\beta_i = \alpha\omega_n^i$ then

$$(F^* \otimes I)\mathcal{A}(x)(F \otimes I) = xI - D_\alpha^{-1} C D_\alpha, \quad D_\alpha = \mathrm{diag}(1, \alpha, \ldots, \alpha^{n-1})$$

- The condition number of the eigenvalues of the new pencil is not worse than that of the scaled block companion
- Choosing β_i with different moduli leads to a dramatic reduction of the condition number (experimental verification)

ℓ-ification and strong ℓ-ification

There exist unimodular $mk \times mk$ matrix polynomial $E(x)$, $F(x)$ such that

$$E(x)\mathcal{A}(x)F(x) = I_{mk-k} \oplus A(x)$$

If $d_1 = \cdots = d_k$ then there exist unimodular $mk \times mk$ matrix polynomial $\widehat{E}(x)$, $\widehat{F}(x)$ such that

$$\widehat{E}(x)\mathcal{A}^{\#}(x)\widehat{F}(x) = I_{mk-k} \oplus A^{\#}(x)$$

where $A^{\#}(x) = \sum_{i=0}^{n} A_{n-i}x^i$ denotes the "reversed polynomial"

That is, if the $b_i(x)$ have the same degree then $\mathcal{A}(x)$ is a strong ℓ-ification of $A(x)$ in the sense of De Terán, Dopico, Mackey 2013

Eigenvectors

If $A(\lambda)v = 0$ then

$$\begin{bmatrix} \prod_{j\neq 1} B_j(\lambda)v \\ \vdots \\ \prod_{j\neq k} B_j(\lambda)v \end{bmatrix}$$

is a right eigenvector of $\mathcal{A}$ corresponding to λ

If $u^T A(\lambda) = 0$ then

$$\left[u^T W_1 \prod_{j\neq 1} B_j(\lambda), \ldots, u^T W_k \prod_{j\neq k} B_j(\lambda) \right]$$

is a left eigenvector of $\mathcal{A}$ corresponding to λ

Block companion form

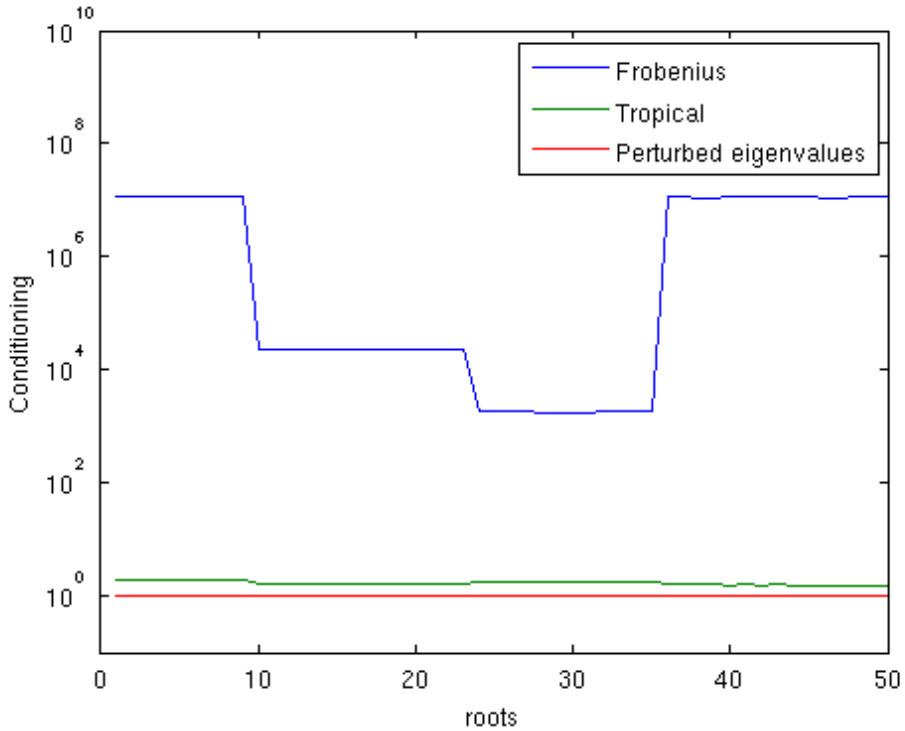
Let $L = \begin{bmatrix} I & & & \\ -I & \ddots & & \\ & \ddots & \ddots & \\ & & -I & I \end{bmatrix}$ then

$$L\mathcal{A}(x) = \begin{bmatrix} B_1(x) + W_1(x) & W_2(x) & \dots & W_{k-1}(x) & W_k(x) \\ -B_1(x) & B_2(x) & & & \\ & -B_2(x) & \ddots & & \\ & & \ddots & B_{k-1}(x) & \\ & & & -B_{k-1}(x) & B_k(x) \end{bmatrix}$$

Numerical properties: scalar polynomials

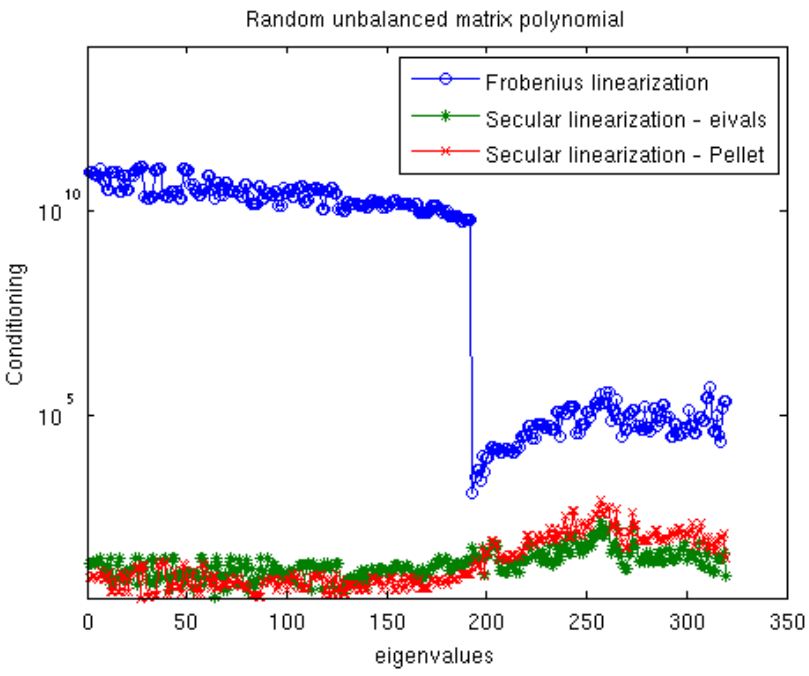
Random scalar polynomial of degree 50 with unbalanced coefficients
Conditioning of the eigenvalues
`a = exp(12*randn(1,n+1))`

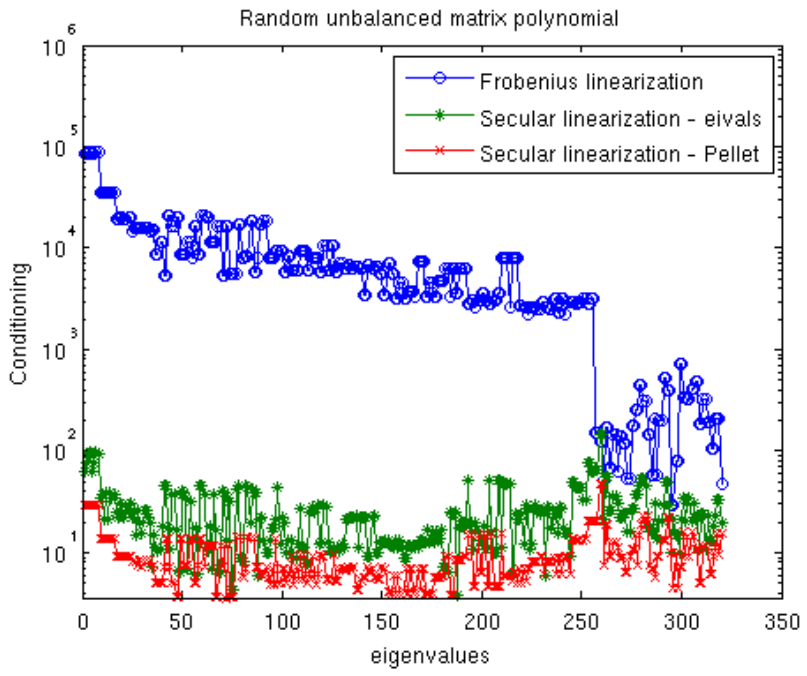
– Frobenius matrix (blue)
– secular linearization $\beta_i = \lambda_i + \epsilon_i$ (red)
– secular linearization β_i equal to the tropical roots multiplied by unit complex numbers (green)

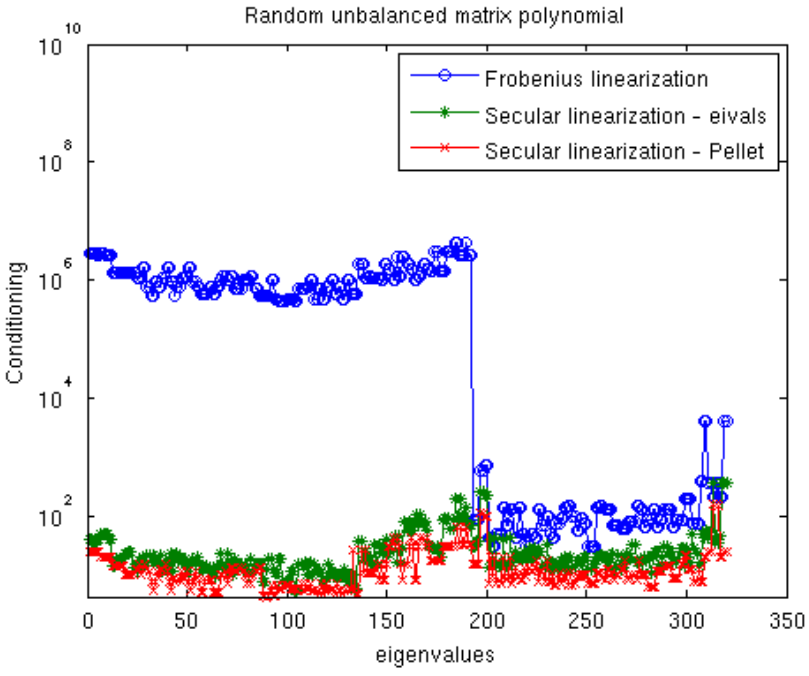


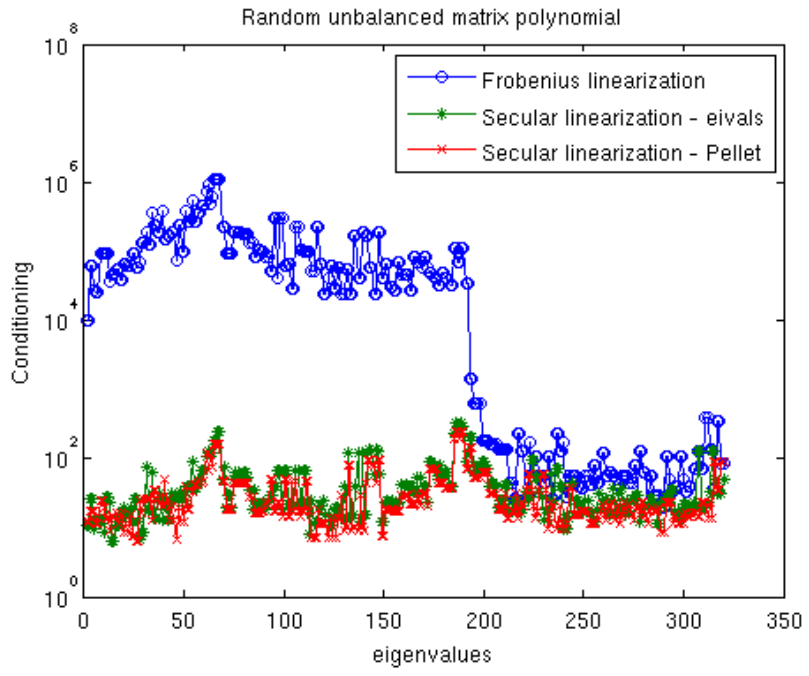
Numerical properties: matrix polynomials

Random matrix polynomial, $m = 64$, $n = 5$,
– Frobenius matrix (blue)
– secular linearization, β_i derived from the eigenvalues (red)
– secular linearization, β_i obtained from the tropical roots (green)






Random unbalanced matrix polynomial
Frobenius linearization
Secular linearization - eivals
Secular linearization - Pellet
Conditioning
eigenvalues
10^10
10^8
10^6
10^4
10^2
0
50
100
150
200
250
300
350


Random unbalanced matrix polynomial
Frobenius linearization
Secular linearization - eivals
Secular linearization - Pellet
Conditioning
eigenvalues
10^8
10^6
10^4
10^2
10^0
0
50
100
150
200
250
300
350

Orr Sommerfeld problem from the NLEVP collection $n = 4$, $m = 64$

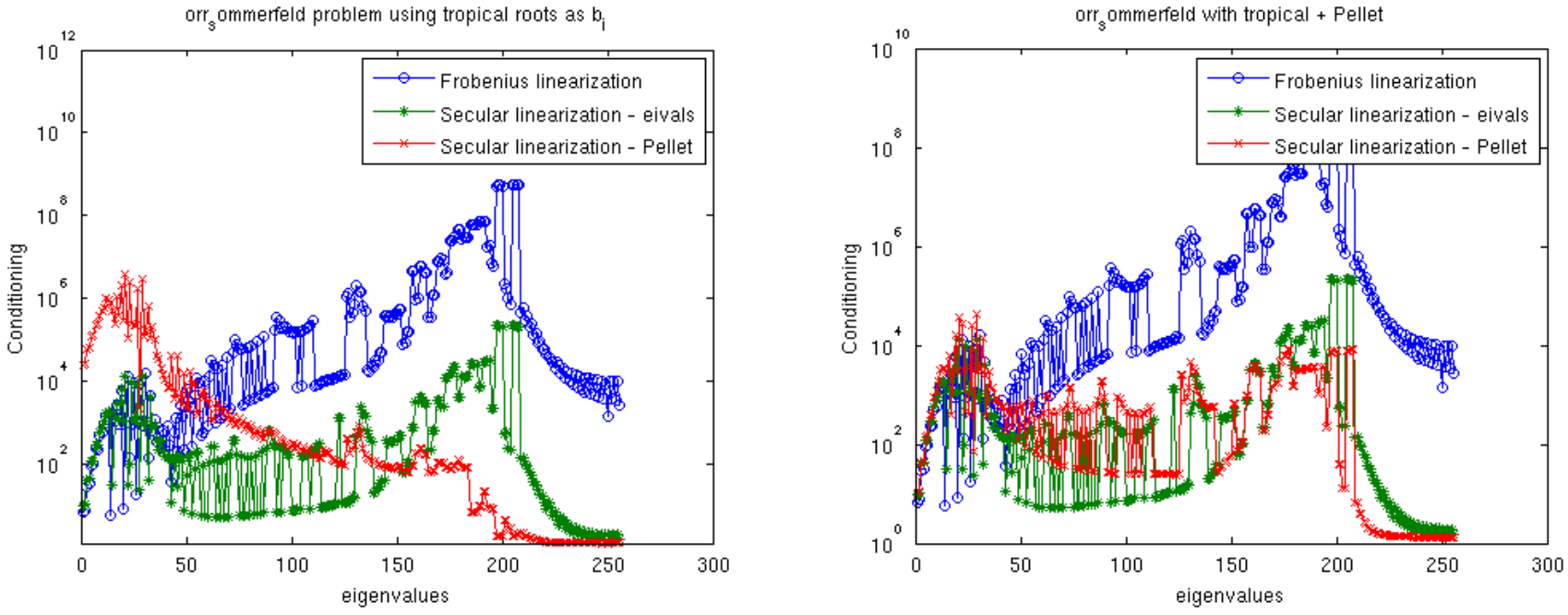
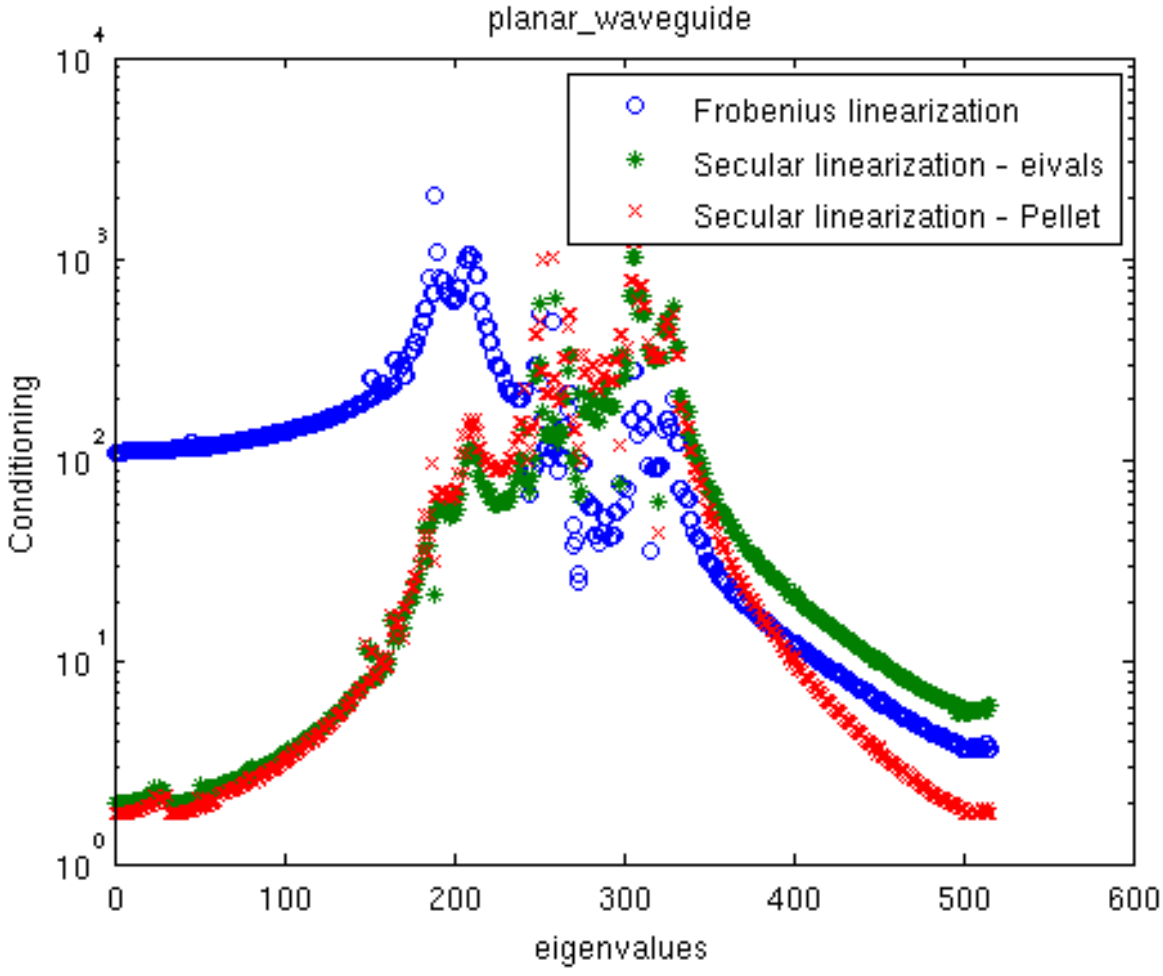


Figure: On the left, the conditioning of the Frobenius and of the secular linearization with the choices of β_i as block mean of eigenvalues and as the estimates given by the tropical roots. On the right, the tropical roots are coupled with estimates given by the Pellet theorem.

Planar waveguide problem from the NLEVP collection $n = 4$, $m = 129$



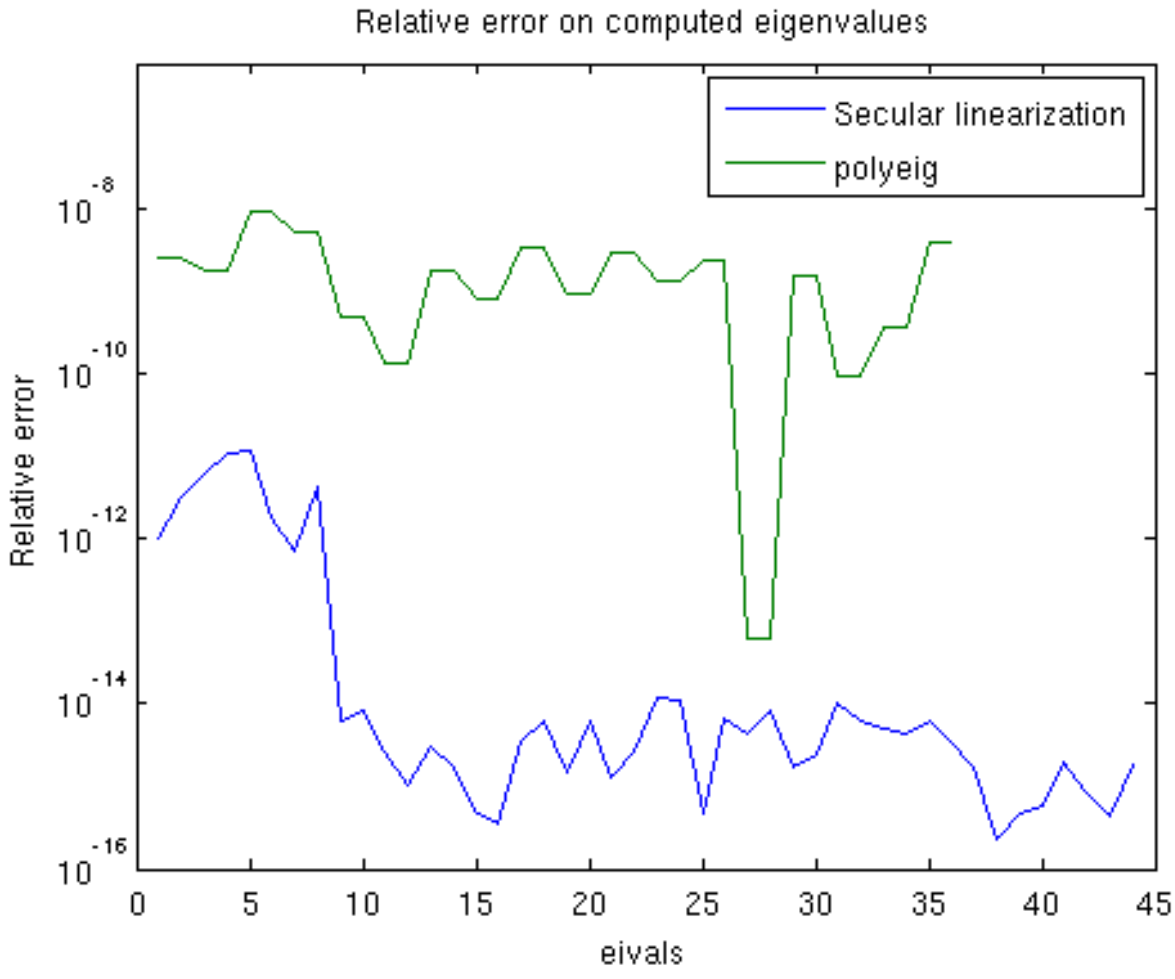


Figure: The accuracy of the computed eigenvalues using polyeig and the secular linerization with the b_i obtained through the computation of the tropical roots.

Recent applications and work in place

Properties of quasi-separable matrices have been exploited to arrive at a matrix polynomial rootfinder with the same feature of MPSolve

Algorithms to compute $p(x) = \det A(x)$ as well as $p(x)/p'(x)$ have been designed with cost $O(nm^2)$ ops, based on the following strategy

- Preprocessing 1. Given $A(x)$ generate a block Smith companion linearization matrix S in $O(n^2m^3)$ ops.
- Preprocessing 2. Reduce S into upper Hessenberg form H in $O(n^2m^3)$ ops. The matrix H is quasi-separable with upper rank $2m-1$.
- Compute $\det A(x) = \det(xI - H)$ at x in $O(nm^2)$ ops by means of the Hyman method, using the quasi-separability of H

Overall cost for applying MPSolve to $\det A(x)$: $O(n^2m^3)$ ops instead of $O(n^2m^3 + nm^4)$ ops

There are still some problems with the numerical stability of the reduction to quasi-separable Hessenberg form which need further investigation

Conclusions

- Toeplitz matrices are encountered in many applications
- They can be associated with Fourier series (symbols)
- Their spectral properties are related to the values of the symbol
- Some matrix algebras, related to fast discrete transforms, can be used for computational purposes: fast product and preconditioning Toeplitz systems
- Displacement operators and their properties can be used to design fast and superfast Toeplitz solvers
- Quadratic matrix equations can be solved by means of Toeplitz computations through Cyclic Reduction, i.e., Graeffe iteration
- The quasi-separable structure is preserved under many matrix transformation
- The considered companion matrices are quasi-separable
- Companion matrices, extended to matrix polynomials, are quasi-separable
- New generalizations of the Smith companion have been given
- Their role in the design of a matrix polynomial rootfinder has been investigated