Matrices structurées et matrices de Toeplitz par blocs de Toeplitz en calcul numérique et formel

Houssam Khalil

Institut Camille Jordan

GALAAD, INRIA Sophia-Antipolis

Directeurs : Michelle Schatzman & Bernard Mourrain

25 juillet 2008

Plan

Systèmes linéaires

Applications

Tous les domaines scientifiques :

- Mathématiques Appliquées
- Ingénierie
- Physique
- Biologie
- $\vdots$

$\Longrightarrow$

$$Ax = b$$

Résolution par des méthodes directes

- 1750 : Cramer $\longrightarrow \mathcal{O}(n(n+1)!)$ opérations
- 1810 : Gauss $\longrightarrow \mathcal{O}(n^3)$ flops, $(2n^3/3$ flops$)$
- Actuellement autour de $\mathcal{O}(n^3)$ flops

Exemple

Exemple

Différences finies pour le laplacien en dimension trois, avec un maillage uniforme de pas $\frac{1}{100}$ sur un cube

$\Longrightarrow$

Matrice obtenue

A de taille $10^6 \times 10^6$.

$\Downarrow$

$2 \times 10^{18}/3$ opérations ;
21 ans à 1 Gflops

Solution

Chercher à utiliser la structure pour réduire le temps de calcul :

- Toeplitz, Hankel, Vandermonde, Cauchy, matrices structurées
- Matrices structurées multiniveaux

Plan

Toeplitz $T = (t_{i-j})_{i,j=0}^{n-1}$

$$\begin{pmatrix} t_0 & t_{-1} & \dots & t_{-n+1} \\ t_1 & t_0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & t_{-1} \\ t_{n-1} & \dots & t_1 & t_0 \end{pmatrix}$$

Hankel $H = (h_{i+j})_{i,j=0}^{n-1}$

$$\begin{pmatrix} h_0 & \dots & h_{n-2} & h_{n-1} \\ \vdots & \iddots & \iddots & h_n \\ h_{n-2} & h_{n-1} & \iddots & \vdots \\ h_{n-1} & h_n & \dots & h_{2n-2} \end{pmatrix}$$

Vandermonde $V = (x_i^{j-1})_{i,j=1}^{n}$

$$\begin{pmatrix} 1 & x_1 & \dots & x_1^{n-1} \\ 1 & x_2 & \dots & x_2^{n-1} \\ \vdots & \vdots & & \vdots \\ 1 & x_n & \dots & x_n^{n-1} \end{pmatrix}$$

Cauchy $C = (\frac{1}{s_i - t_j})_{i,j=0}^{n-1}$

$$\begin{pmatrix} \frac{1}{s_1-t_1} & \dots & \frac{1}{s_1-t_n} \\ \vdots & & \vdots \\ \frac{1}{s_n-t_1} & \dots & \frac{1}{s_n-t_n} \end{pmatrix}$$

Algorithmes rapides

Multiplication Matrice $\times$ Vecteur rapide

- $\mathcal{O}(n \log n)$ pour Toeplitz et Hankel
- $\mathcal{O}(n \log^2 n)$ pour Vandermonde et Cauchy

Résolution rapide et ultra-rapide du système linéaire

- algorithmes rapides $\longrightarrow \mathcal{O}(n^2)$
- algorithmes ultra-rapides $\longrightarrow \mathcal{O}(n \log^2 n)$

Structure de Déplacement & Résolution rapide

Plus généralement :

$$D_{M,N}(A) = MA - AN$$

Déplacement

- Type Toeplitz : $\mathrm{rang}(D_{Z,Z}(A)) = r \ll n$
- Type Hankel : $\mathrm{rang}(D_{Z,Z^T}(A)) \ll n$
- Type Vandermonde : $\mathrm{rang}(D_{\mathrm{diag}^{-1}(x),Z^T}(A)) \ll n$
- Type Cauchy : $\mathrm{rang}(D_{\mathrm{diag}(s),\mathrm{diag}(t)}(A)) \ll n$

Algorithmes rapides

- Multiplication rapide : $\mathcal{O}(rn \log^i n)$, $i = 0, 1$
- Résolution
  - rapide : $\mathcal{O}(rn^2)$
  - ultra-rapide :$\mathcal{O}(r^2 n \log^2 n)$

Plan

Matrice TBT

Définition

$$T = \begin{pmatrix} T_0 & T_{-1} & \cdots & T_{-m+1} \\ T_1 & T_0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & T_{-1} \\ T_{m-1} & \cdots & T_1 & T_0 \end{pmatrix} \quad \text{avec}$$

$$T_i = \begin{pmatrix} t_{i,0} & t_{i,-1} & \cdots & t_{i,-n+1} \\ t_{i,1} & t_{i,0} & \ddots & \vdots \\ \vdots & \ddots & \ddots & t_{i,-1} \\ t_{i,n-1} & \cdots & t_{i,1} & t_{i,0} \end{pmatrix}$$

$$T = (t_{\alpha-\beta})_{\alpha,\beta\in\{(i,j);\, 1\leq i\leq m,\, 1\leq j\leq n\}}$$

T est de taille $N = mn$

- Multiplication Matrice $\times$ Vecteur rapide ? Oui en $\mathcal{O}(N \log N)$
- Résolution rapide et ultra-rapide ?
  - rapide oui
    - en $\mathcal{O}(m^3 n \log^2 nm)$ (matrice de type Toeplitz)
    - en $\mathcal{O}(N^2 \log N)$ (Weidemann)
  - ultra-rapide en $\mathcal{O}(N \log^2 N)$???
- Opérateurs de déplacement effectifs ? ???
  - $D_1 = D_{Z_m \otimes I_n, Z_m \otimes I_n}$ et $D_2 = D_{I_m \otimes Z_n, I_m \otimes Z_n}$ exploitent une seule structure
  - $D_1 \circ D_2 = D_2 \circ D_1$ pas un bon opérateur

Explications

Multiplication rapide

Multiplier v par T est équivalente à :

- multiplier deux polynômes de deux variables de degré $(2m-1, 2n-1)$ et $(m-1, n-1)$ resp.
- multiplier un vecteur par une matrice circulante par blocs circulants de taille $4mn \times 4mn$

TBT et type Toeplitz

- $\mathrm{rang}(D_{Z,Z}(T)) = r = 2m$ et $\mathrm{rang}(D_{Z,Z}(PTP^T)) = r = 2n$

  $\Downarrow$

  résolution en $\mathcal{O}(r^2 mn \log^2 mn) = \mathcal{O}(m^3 n \log^2 mn)$ si $m < n$ et $\mathcal{O}(mn^3 \log^2 mn)$ si $n < m$
- T Toeplitz par bloc : généralisation des algorithmes de Toeplitz scalaires $\rightarrow$ résolution en $\mathcal{O}(n^3 m \log^2 m)$.

Algorithmes ultra-rapides

La généralisation des algorithmes scalaires n'exploite qu'une direction de structure :

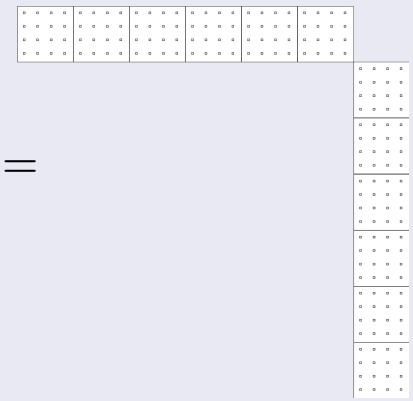
Les blocs perdent leur structure de Toeplitz très facilement

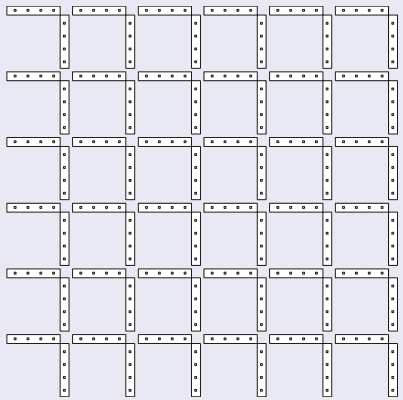
- Les blocs d'un complément de Schur ne sont pas structurés
- Le rang de déplacement de T et ses complément de Schur est au moins $2min(m, n)$

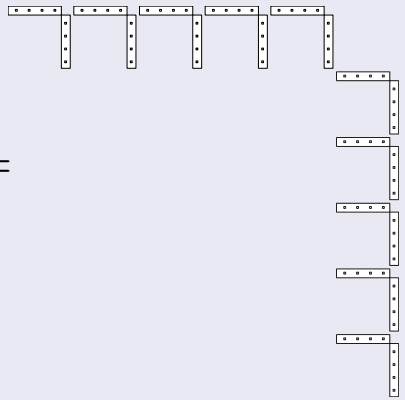
Généralisation des algorithmes scalaires

Les algorithmes de type Levinson, de type Schur, autres algorithmes ultra-rapides utilisent la fait que T et ses compléments de Schur sont de petit rang de déplacement

$D_1 = D_{Z_m \otimes I_n, Z_m \otimes I_n}$, $D_2 = D_{I_m \otimes Z_n, I_m \otimes Z_n}$ et $D = D_1 \circ D_2$

$D_1(T) =$ 

$D_2(T) =$ 

$D(T) =$ 

- $\mathrm{rg}(D_1(T)) = 2m$, $\mathrm{rg}(D_2(T)) = 2n$
- $\mathrm{rg}(D(T)) = 2\min(m, n)$
- $D(T)$ plus creuse, générateurs structurés, valeurs singulières décroissent plus vite... MAIS
  - rang de déplacement GRAND
  - $rg(D(T^{-1})) = 4min(m, n)$ pas $2min(m, n)$
  - les compléments de Schur successifs perdent leurs structures

Plan

$$T = \begin{pmatrix} T_0 & \dots & T_{-k_1} & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ T_{k_1} & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & T_{-k_1} \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & T_{k_1} & \dots & T_0 \end{pmatrix},$$

$$T_j = \begin{pmatrix} T_{0,j} & \dots & T_{-k_2,j} & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ T_{k_2,j} & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & T_{-k_2,j} \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & T_{k_2,j} & \dots & T_{0,j} \end{pmatrix}$$

$$m = \mathcal{O}(\sqrt{N}),\ n = \mathcal{O}(\sqrt{N}), k_1 \ll m,\ k_2 \ll n$$

Elimination de Gauss pour matrices creuses

Largeur de bande $= k_1 n + k_2$

- Elimination gaussienne : $\mathcal{O}((k_1 n + k_2)^2 N) \backsim \mathcal{O}(N^2)$
- Nos algorithmes : $\mathcal{O}(N^{3/2})$

Statistiques

Nombre de conditionnement

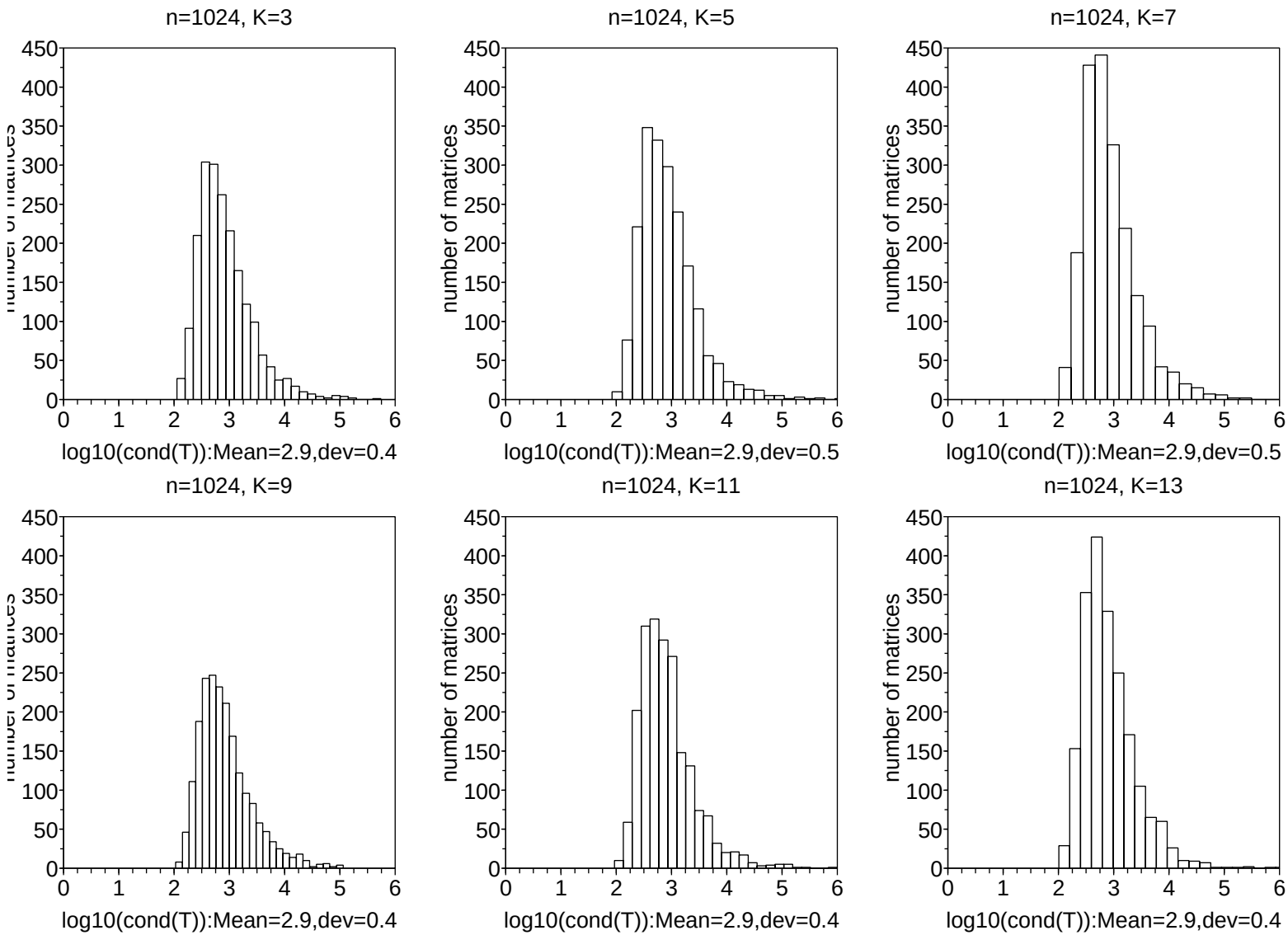
nombre de conditionnement : $\kappa(A) = \|A\|\|A^{-1}\|$. Si δA, δb perturbation de A et b alors l'estimation de l'erreur relative sur x :

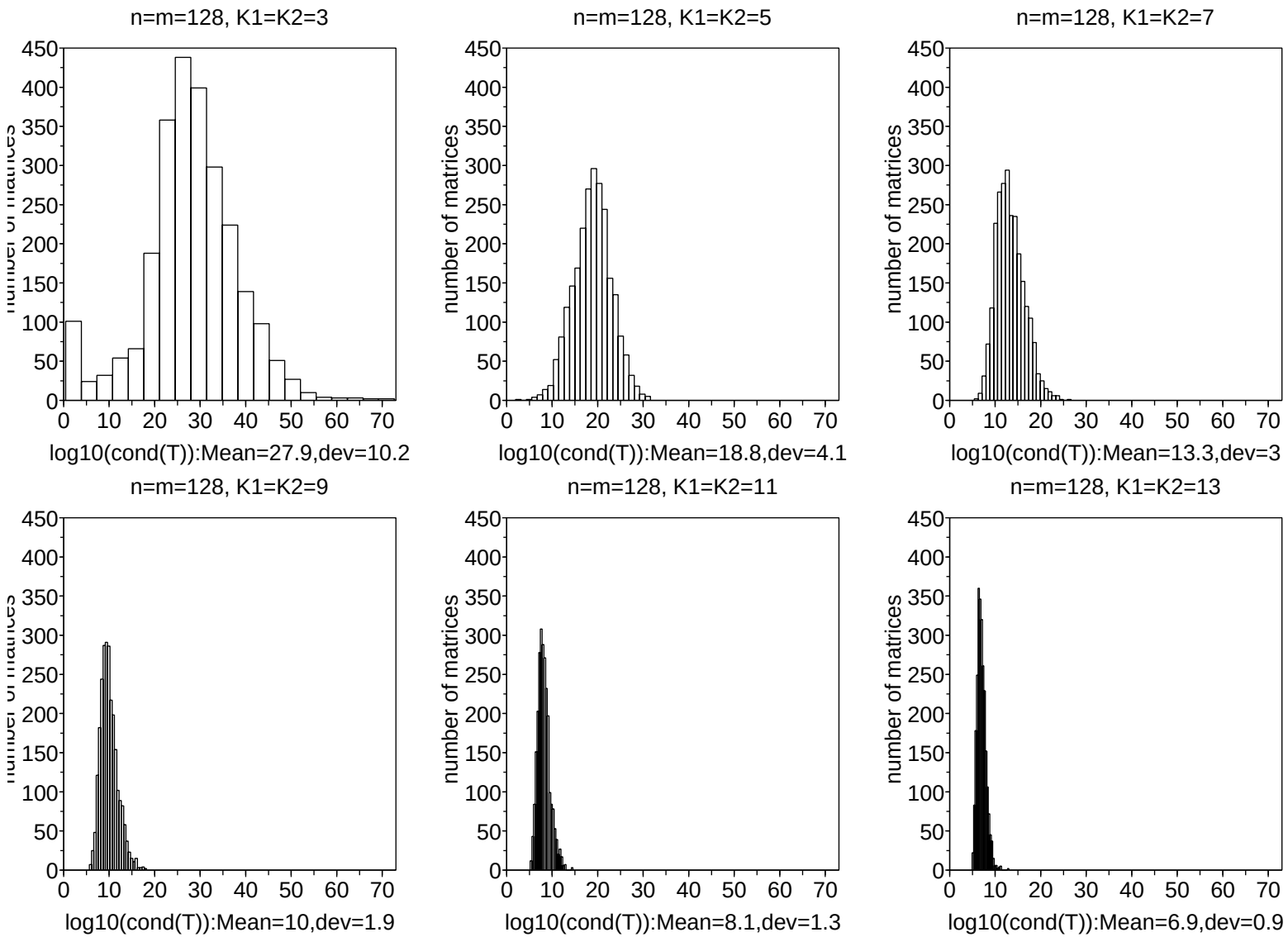
$$\frac{|\delta x|}{|x|} \leq \kappa(A) \left(\frac{|\delta b|}{b} + \frac{\|\delta A\|}{\|A\|} \right) + \text{termes d'ordre supérieur}$$

Statistique

Statistiques expérimentales

- On n'a pas de statistiques théoriques sur le nombre de conditionnement des matrices de Toeplitz bandes
- Des statistiques expérimentales montrent :
  - Dans le cas scalaire, la statistique sur les nombres de conditionnements dépend peu de la largeur de bande
  - Dans le cas par blocs, elle dépend des largeurs de bande : largeurs de bande petites $\longrightarrow$ matrices mal conditionnées
    La distribution est, PEUT ETRE, de Tracy-Widom


n=1024, K=3
n=1024, K=5
n=1024, K=7
n=1024, K=9
n=1024, K=11
n=1024, K=13
number of matrices
log10(cond(T)):Mean=2.9,dev=0.4
log10(cond(T)):Mean=2.9,dev=0.5
log10(cond(T)):Mean=2.9,dev=0.5
log10(cond(T)):Mean=2.9,dev=0.4
log10(cond(T)):Mean=2.9,dev=0.4
log10(cond(T)):Mean=2.9,dev=0.4


n=m=128, K1=K2=3
number of matrices
log10(cond(T)):Mean=27.9,dev=10.2
n=m=128, K1=K2=5
number of matrices
log10(cond(T)):Mean=18.8,dev=4.1
n=m=128, K1=K2=7
number of matrices
log10(cond(T)):Mean=13.3,dev=3
n=m=128, K1=K2=9
number of matrices
log10(cond(T)):Mean=10,dev=1.9
n=m=128, K1=K2=11
number of matrices
log10(cond(T)):Mean=8.1,dev=1.3
n=m=128, K1=K2=13
number of matrices
log10(cond(T)):Mean=6.9,dev=0.9

T vue comme matrice CBC + matrice de petit rang

$$Tx = b$$

Cas scalaire

T une matrice de Toeplitz bande de largeur de bande $2k+1$:

$$T = \begin{pmatrix} t_0 & \dots & t_{-k} & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ t_k & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & t_{-k} \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & t_k & \dots & t_0 \end{pmatrix} =$$

$$\begin{pmatrix} t_0 & \cdots & t_{-k} & 0 & t_k & \cdots & t_1 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ t_k & \ddots & \ddots & \ddots & \ddots & \ddots & t_k \\ 0 & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \\ t_{-k} & \ddots & \ddots & \ddots & \ddots & \ddots & t_{-k} \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ t_{-1} & \cdots & t_{-k} & 0 & t_k & \cdots & t_0 \end{pmatrix} - \begin{pmatrix} 0 & \cdots & \cdots & 0 & t_k & \cdots & t_1 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & t_k \\ 0 & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \\ t_{-k} & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ t_{-1} & \cdots & t_{-k} & 0 & \cdots & \cdots & 0 \end{pmatrix}$$

$$= C + R$$

$$R = GH^T = \left(\begin{array}{ccc|ccc} & & & t_k & \cdots & t_1 \\ & & & & \ddots & \vdots \\ & & & & & t_k \\ t_{-k} & & & & & \\ \vdots & \ddots & & & & \\ t_{-1} & \cdots & t_{-k} & & & \end{array}\right) \left(\begin{array}{c|c} -I_k & \\ \hline & -I_k \end{array}\right)$$

Formule de Sherman-Morrison-Woodbury

Soient $A \in \mathbb{K}^{n\times n}$, G et $H \in \mathbb{K}^{n\times k}$. Si $I_k + H^T A^{-1} G$ est inversible alors

$$(A + GH^T)^{-1} = A^{-1} - A^{-1}G(I_k + H^T A^{-1} G)^{-1} H^T A^{-1}$$

Résolution du système linéaire

$$x = T^{-1}b = C^{-1}b - C^{-1}G(I_k + H^T C^{-1} G)^{-1} H^T C^{-1} b$$

- $C^{-1}v$ en $\mathcal{O}(n \log n)$
- Gv en $\mathcal{O}(k \log k)$ $\quad$ ($H^T v$ en 0 opération)
- $H^T C^{-1} G$ en $\mathcal{O}(n \log n + nk \log k)$
- $(I_k + H^T C^{-1} G)^{-1}$ en $\mathcal{O}(k^3)$

Coût total : $\mathcal{O}(n \log n)$ en supposant que $k \ll n$

Cas par blocs

Toeplitz bande par blocs Toeplitz bandes

- $T_i = C_i + R_i$, R_i de rang $2k_2$. Donc $T = \tilde{C} + R_1$ une matrice de rang $2k_2 m$
- $\tilde{C} \longrightarrow C + R_2$: matrice CBC + matrice de rang $2k_1 n$

Résolution du système linéaire

- $T = C + R_1 + R_2 = C + R, \quad R$ de rang $k = 2(k_1 n + k_2 m)$
- $R = GH^T$, $G, H \in \mathbb{K}^{N,k}$ G contient $\mathcal{O}(k_1^2 k_2 n + k_2^2 k_1 m) = \mathcal{O}(K)$ éléments, H correspond à l'identité

Sherman-Morrison-Woodbury sur $(C+R)x = b$:

$$x = C^{-1}b - C^{-1}G(I_k + H^T C^{-1} G)^{-1} H^T C^{-1} b$$

$$N = nm,\ m = \mathcal{O}(\sqrt{N}),\ n = \mathcal{O}(\sqrt{N}),\ k_1 \ll m,\ k_2 \ll n$$

- $C^{-1}v$ en $\mathcal{O}(N \log N)$
- Gv en $\mathcal{O}(K)$
- $H^T C^{-1} G$ en $\mathcal{O}(N \log N + KN) = \mathcal{O}(N^{3/2})$
- $(I_k + H^T C^{-} G)^{-1}$ en $\mathcal{O}(k^3) = \mathcal{O}(N^{3/2})$

Plan

Cas scalaire

Problème

$$Tu = g$$

avec

$$T = \begin{pmatrix} t_0 & t_{-1} & \dots & t_{-n+1} \\ t_1 & t_0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & t_{-1} \\ t_{n-1} & \dots & t_1 & t_0 \end{pmatrix}$$

et

$$u = \begin{pmatrix} u_0 \\ \vdots \\ u_{n-1} \end{pmatrix}, g = \begin{pmatrix} g_0 \\ \vdots \\ g_{n-1} \end{pmatrix}$$

Approche polynomial et relation avec les syzygies

Notations

- $a = (a_0, \ldots, a_m)^T \longrightarrow a(x) = a_0 + \cdots + a_m x^m$
- $T(x) = \sum_{i=-n+1}^{n-1} t_i x^i = T_-(x) + T_+(x)$
- $\tilde{T}(x) = T_+(x) + x^{2n} T_-(x)$
- $E = \{1, \ldots, x^{n-1}\}$ et Π_E est la projection sur $\mathrm{Vect}(E)$

Théorème

$Tu = g \Leftrightarrow \Pi_E(T(x)u(x)) = g(x)$

Théorème

u solution de $Tu = g \Leftrightarrow \exists\, v(x), w(x) \in \mathbb{K}[x]_{n-1}$;

$$\tilde{T}(x)u(x) + x^n v(x) + (x^{2n} - 1)w(x) = g(x)$$

Définition

Pour $(a, b, c) \in \mathbb{K}[x]^3$ et $d \in \mathbb{K}[x]$ on définie

- $\mathcal{L}(a, b, c) = \{(p, q, r) \in \mathbb{K}[x]^3;\ ap + bq + cr = 0\}$ (ensemble des syzygies de (a, b, c))
- $\mathcal{L}(a, b, c; d) = \{(p, q, r) \in \mathbb{K}[x]^3;\ ap + bq + cr = d\}$

Propositions

- le $\mathbb{K}[x]$-module $\mathcal{L}(\tilde{T}(x), x^n, x^{2n} - 1)$ est libre de rang 2
- Il admet une base $\{(u_1, v_1, w_1), (u_2, v_2, w_2)\}$ telle que $\deg u_1 = \deg v_2 = n$ et les autres sont de degré $< n$ (n-base)

Théorème

Soit $(a(x), b(x), c(x)) \in \mathcal{L}(\tilde{T}(x), x^n, x^{2n} - 1; g(x))$.
$\exists! p_1(x), p_2(x) \in \mathbb{K}[x]$ tels que
$(a, b, c) = p_1(u_1, v_1, w_1) + p_2(u_2, v_2, w_2) + (u, v, w)$
(u, v, w) est ! élément de $\mathcal{L}(\tilde{T}(x), x^n, x^{2n} - 1; g(x)) \cap \mathbb{K}[x]_{n-1}$

Autre forme de la formule de Gohberg-Semencul

Remarque

- (u_1, v_1, w_1) et (u_2, v_2, w_2) tels que

$$\begin{cases} \tilde{T}(x)u_1(x) + x^n v_1(x) + (x^{2n} - 1)w_1 = 1, \\ \tilde{T}(x)u_2(x) + x^n v_2(x) + (x^{2n} - 1)w_2 = \tilde{T}(x)x^n. \end{cases}$$

  forment une $n-$base de $\mathcal{L}(\tilde{T}(x), x^n, x^{2n} - 1)$

- u_1 et u_2 sont les sontion de :
  1. $Tu_1 = e_1$
  2. $Tu_2 = ZTe_n$

Calcul de $n-$ base et de la solution

Algorithmes rapides en $\mathcal{O}(nlog^2 n)$

- En appliquant l'algorithme d'Euclide pour le calcul de PGCD au $p(x) = x^{n-1}T(x)$ et $q(x) = x^{2n-1}$ tronqué en degré $n-1$ on obtient (u_1, v_1, w_1) et (u_2, v_2, w_2)
- On peut faire la division en utilisant l'algorithme de Newton

Matrices Toeplitz par blocs Toeplitz et syzygies

Problème

$$Tu = g$$

$T = (t_{\alpha-\beta})_{\alpha,\beta\in E} \in \mathbb{K}^{N\times N}$, $u = (u_\alpha)_{\alpha\in E}$ et $g = (g_\alpha)_{\alpha\in E}$
$E = \{(i,j); 1 \le i \le m, 1 \le j \le n\}$.

Définition

- $u(x,y) = \sum_{(i,j)\in E} u_{i,j}x^i y^j, \quad g(x,y) = \sum_{(i,j)\in E} g_{i,j}x^i y^j$
- $T(x,y) = \sum_{(i,j)\in E-E} t_{i,j}x^i y^j$
- $\tilde{T}(x,y) = T_{++} + x^{2m}T_{-+} + y^{2n}T_{+-} + x^{2m}y^{2n}T_{--}$

Théorème

$Tu = g \Leftrightarrow \Pi_E(T(x,y)u(x,y)) = g(x,y)$

Théorème

u solution de $Tu = g \Leftrightarrow \exists h_1, \ldots, h_8 \in \mathbb{K}[x,y]_{\substack{m-1 \\ n-1}}$;
$(u(x,y), h_1(x,y), \ldots, h_8(x,y)) \in \mathcal{L}(\mathbf{T}; g(x,y))$ avec
$\mathbf{T} = (\tilde{T}(x,y), x^m, x^{2m}-1, y^n, x^m y^n, (x^{2m}-1)y^n, y^{2n}-1,$
$x^m(y^{2n}-1), (x^{2m}-1)(y^{2n}-1))$

Théorème

Le $\mathbb{K}[x,y]-$module $\mathcal{L}(\mathbf{T})$ est libre de rang 8.

Proposition

$u_1(x,y), u_2(x,y), u_3(x,y) \in \mathbb{K}[x,y]^9_{\substack{m-1\\n-1}}$;

$\mathbf{T}.u_1 = \tilde{T}(x,y)x^m,\ \mathbf{T}.u_2 = \tilde{T}(x,y)y^n,\ \mathbf{T}.u_3 = 1$

les relations suivantes forment une base de $\mathcal{L}(\mathbf{T})$:

$$\begin{cases} \rho_1 = x^m\sigma_1 - u_1 & \rho_5 = y^n\sigma_2 - \sigma_5 \\ \rho_2 = y^n\sigma_1 - u_2 & \rho_6 = x^m\sigma_4 - \sigma_5 \\ \rho_3 = x^m\sigma_2 - \sigma_3 - u_3 & \rho_7 = x^m\sigma_5 - \sigma_6 + \sigma_4 \\ \rho_4 = y^n\sigma_4 - \sigma_7 - u_3 & \rho_8 = y^n\sigma_5 - \sigma_8 + \sigma_2 \end{cases} \quad \text{avec } \sigma_1, \ldots, \sigma_9$$

la base canonique de $\mathbb{K}[x,y]^9$

Théorème

Pour $(a_1, \ldots, a_9) \in \mathcal{L}(\mathbf{T}(x,y); g(x,y))$, $\exists! p_i(x,y) \in \mathbb{K}[x,y]$, $i = 1, \ldots, 8$ tels que $(u, h_1, \ldots, h_8) = (a_1, \ldots, a_9) - \sum_{i=1}^{8} p_i\, \rho_i$

Conclusion et perspectives

$Tx = b$, T est TBT de taille $N \times N$

Conclusion

- On ne sait pas résoudre ce problème en $O(N \log^2 N)$
- On peut le résoudre en $\mathcal{O}(N^2 \log N)$
- Si T est bande, on peut le résoudre en $\mathcal{O}(N^{3/2})$
- On peut le transformer à un problème polynomial
  - dans le cas scalaire on peut le résoudre ultra-rapidement
  - dans le cas par blocs, on ne peut pas

Perspectives

- Essayer de généraliser la notion de la structure de déplacement, en généralisant la notion de “petit rang”
- Essayer de trouver des algorithmes rapides pour le nouveau problème polynomial

Merci