Fast solver for band Toeplitz-block-band Toeplitz system

Houssam Khalil

Résolution rapide d'un système de Toeplitz bande par blocs Toeplitz bandes

Houssam Khalil*

Resumé. Soit T une matrice de Toeplitz par blocs de Toeplitz à coefficients dans un corps $\mathbb{K}$ et de taille N. On suppose qu'elle est formée de m blocs de taille $n \times n$; de plus elle est bande par blocs, c'est à dire qu'en dehors des $2k_1 + 1$ diagonales par blocs centrales les blocs sont nuls ; les blocs eux-même sont bande : en dehors des $2k_2 + 1$ diagonales centrales, les éléments des blocs sont nuls. On donne dans ce papier trois méthodes de résolution rapide, en $\mathcal{O}(N^{3/2})$, pour résoudre le système linéaire. Ces méthodes sont plus rapides que les méthodes associées aux matrices creuses. On donne aussi une statistique qui montre que la matrice T devient de plus en plus mal conditionnée quand les largeures des bandes décroissent. Cette remarque n'est pas vrai pour le cas d'une matrice de Toeplitz sclaire.

Mots clés. Toeplitz matrix, rational interpolation, syzygie

1 Introduction

Soit T une matrice de Toeplitz par blocs de Toeplitz à coefficients dans un corps $\mathbb{K}$. On suppose qu'elle est formée de m blocs de taille $n \times n$; de plus elle est bande par blocs, c'est à dire qu'en dehors des $2k_1 + 1$ diagonales par blocs centrales les blocs sont nuls ; les blocs eux-même sont bande : en dehors des $2k_2 + 1$ diagonales centrales, les éléments des blocs sont nuls. T est donc de la forme suivante :

$$T = \begin{pmatrix} T_0 & T_{-1} & \dots & T_{-k_1} & 0 & \dots & 0 \\ T_1 & T_0 & \dots & T_{-k_1+1} & T_{-k_1} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \\ T_{k_1} & \ddots & \ddots & \ddots & \ddots & \ddots & T_{-k_1} \\ 0 & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & T_{-1} \\ 0 & \dots & 0 & T_{k_1} & \dots & T_1 & T_0 \end{pmatrix}, \tag{1}$$

*Institut Camille Jordan, 43 boulevard du 11 novembre 1918, 69622 Villeurbanne cedex France (khalil@math.univ-lyon1.fr).

et chaque T_j est de la forme :

$$T_j = \begin{pmatrix} T_{0,j} & T_{-1,j} & \dots & T_{-k_2,j} & 0 & \dots & 0 \\ T_{1,j} & T_{0,j} & \dots & T_{-k_2+1,j} & T_{-k_2,j} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \\ T_{k_1,j} & \ddots & \ddots & \ddots & \ddots & \ddots & T_{-k_2,j} \\ 0 & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & T_{-1,j} \\ 0 & \dots & 0 & T_{k_2,j} & \dots & T_{1,j} & T_{0,j} \end{pmatrix}. \quad (2)$$

Soit $N = nm$, et on va supposer que m et n sont de même ordre, c'est à dire $C^{-1}m \leq n \leq Cm$ pour une certaine constante $C > 0$. Par conséquent $m = \mathcal{O}(\sqrt{N}),\ n = \mathcal{O}(\sqrt{N})$. Il est classique que la résolution d'un système $n \times n$ de structure bande de largeur k coûte $\mathcal{O}(k^2 n)$ opérations par la méthode de Gauss. Dans notre cas, le coût serait donc en $\mathcal{O}(N^2)$ opérations avec des méthodes directes pour matrices creuses.

Dans cet article, on cherche à exploiter la structure *Toeplitz bande par blocs Toeplitz bande* pour donner une estimation en $\mathcal{O}(N^{3/2})$.

Remarquons tout d'abord que des statistiques expérimentales sur le nombre de conditionnement des matrices aléatoires de Toeplitz et des matrices aléatoires de Toeplitz par blocs de Toeplitz montrent des comportements dont il importe de tenir compte dans l'analyse des résultats.

La section (2) décrit ces statistiques ainsi que les moyens utilisés pour les obtenir. Il faut mentionner ici que les comportements dans le cas Toeplitz et dans le cas Toeplitz par blocs de Toeplitz sont qualitativement différents.

2 Statistiques pour des matrices bandes Toeplitz et pour des matrices bandes Toeplitz par blocs bande Toeplitz

2.1 Etat des connaissances

Un certain nombre de résultats théoriques sont connus pour des matrices aléatoires à structure bande, ou des matrices aléatoires à structure de Toeplitz. Notons en particulier un théorème de limite centrale pour un modéle de matrice bande symétrique par Anderson et Zeitouni [1], un résultat sur la norme spectrale des matrices bandes hermitiennes par Khorunzhy[4] (? ?), un résultat sur la distribution limite des valeurs propres pour les grandes matrices bandes symétriques ou hermitiennes par Molchanov et al. [6], un résultat de grandes déviations pour un modèle de matrices hermitiennes par Guionnet [3], des résultats sur les développements asymptotiques et les échelles d'universatitéspectrale par Khorunzhy et kirsch [5]. D'autre part il y a des résulats pour les matrices de Toeplitz aléatoires (à citer).

aucun de ces résultats ne porte sur le nombre de conditionnement de l'une ou l'autre fammille de modèles de matrice, alors que c'est l'objet essentiel pour l'analyse numérique des méthodes de résolution de systèmes linéaires.

On a donc décidé d'obtenir des informations expérimentales pour des matrices aléatoires bande Toeplitz et bande Toeplitz par blocs bande Toeplitz et on a procédé comme suit.

2.2 Algorithmes

Notons k la largeur impaire de bande dans le cas de Toeplitz scalaire et k_1, k_2 les deux largeurs impaires de bande dans le cas par blocs. Remarquons que les diagonales non nulles occupent une zone symétrique. Plus précisément, dans le cas scalaire, les diagonales sont nulles en dehors de $\{-\lfloor k/2 \rfloor \dots \lfloor k/2 \rfloor\}$, et dans le cas par blocs, les diagonales de blocs sont nulles en dehors de $\{-\lfloor k_1/2 \rfloor \dots \lfloor k_1/2 \rfloor\}$ et les diagonales dans les blocs sont nulles en dehors de $\{-\lfloor k_2/2 \rfloor \dots \lfloor k_2/2 \rfloor\}$.

On a généré des coefficients pseudo-aléatoires gaussiens en nombre approprié, c'est à dire k dans le cas scalaire et $k_1 k_2$ dans le cas par blocs, ce qui permet de décrire une matrice t en structure creuse.

On a effectué un décomposition LU creuse, par l'algorithme"superLU", de T et sa transposée, ce qui a permis de calculer la plus petite valeur singulière par la méthode de la puissance inverse. D'autre part on a obtenu la plus grande valeur singulière par la méthode de la puissance.

On a constaté la convergence rapide de la méthode de la puissance inverse, et la convergence très lente de la méthode de la puissance. Ce qui n'est pas entièrement surprenant.

Le rapport entre la plus grande et la plus petite des valeurs singulières fournit le nombre de conditionnement $\kappa(T) = \|T\|_2 \|T^{-1}\|_2$, la norme $\|.\|_2$ étant la norme spectrale, c'est à dire la racine carrée de la plus grande valeur propre de T^*T.

2.3 Les résultats dans le cas scalaire

On se reportera aux figures pour voir des histogrammes du logarithme en base 10 du nombre de conditionnement.

On remarque que ces histogrammes dépendent peu de la largeur de bande.

2.4 Les résultats dans le cas par blocs

Dans ce cas on constate qu'avec l'accroissement d'au moins une des largeurs de bande, l'histogramme devient de plus en plus étroit, et la moyenne du logarithme du nombre de conditionnement décroît.

Définition 1. *Dans la suite : Z_n (ou Z s'il y a pas de confusion) dénote la matrice de taille n telle que $Z_{ij} = 1$ si $i = j + 1$ et 0 sinon, e_i est le ième élément de la base canonique de $\mathbb{R}$. Notons $\mathbb{K}^{p\times q}$ l'espace vectoriel des matrices à p lignes et q colonnes, I_p la matrice identité $p \times p$, M^T la transposée d'une matrice M.*

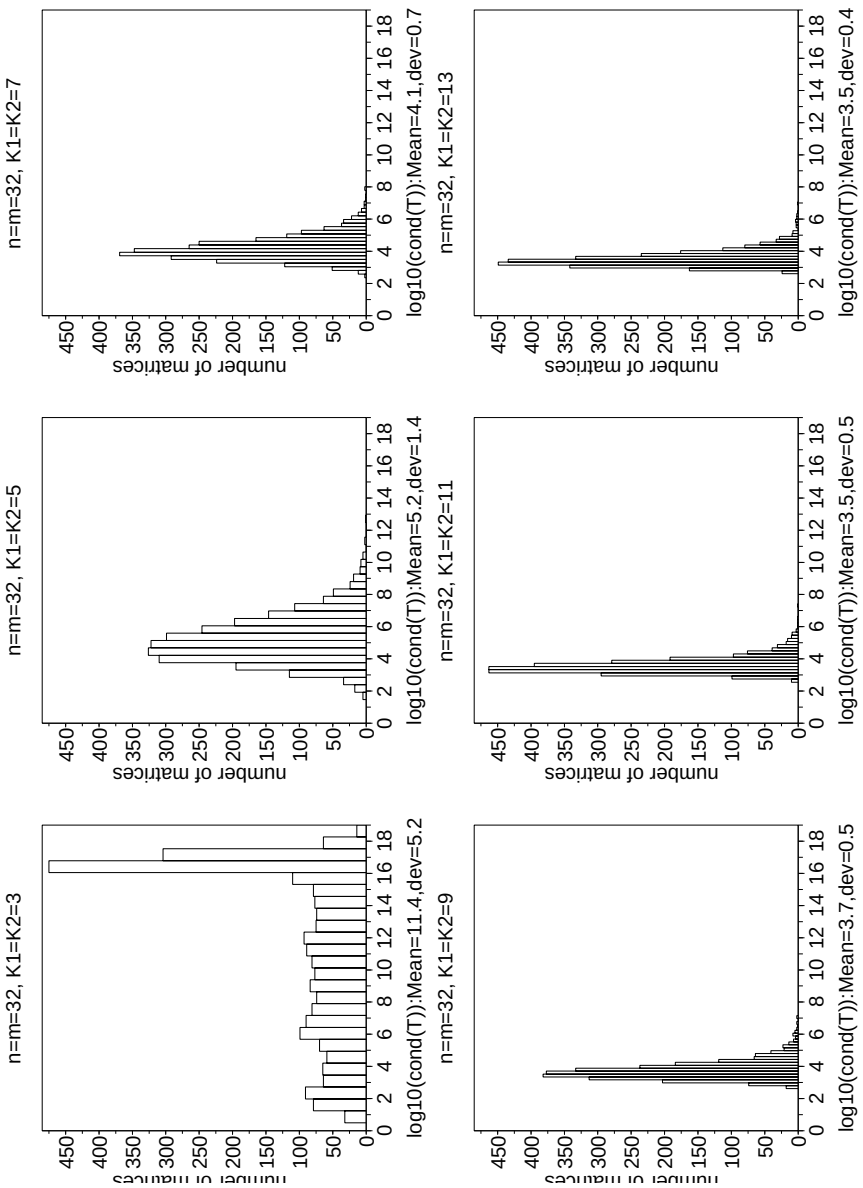


FIG. 1 – On considère des matrices TBT de taille $32^2 \times 32^2$ et de largeur de bande $2k_1 - 1 = 2k_1 - 1$ allant de 3 à 13. On a fait 2500 essaies pour chaque cas.

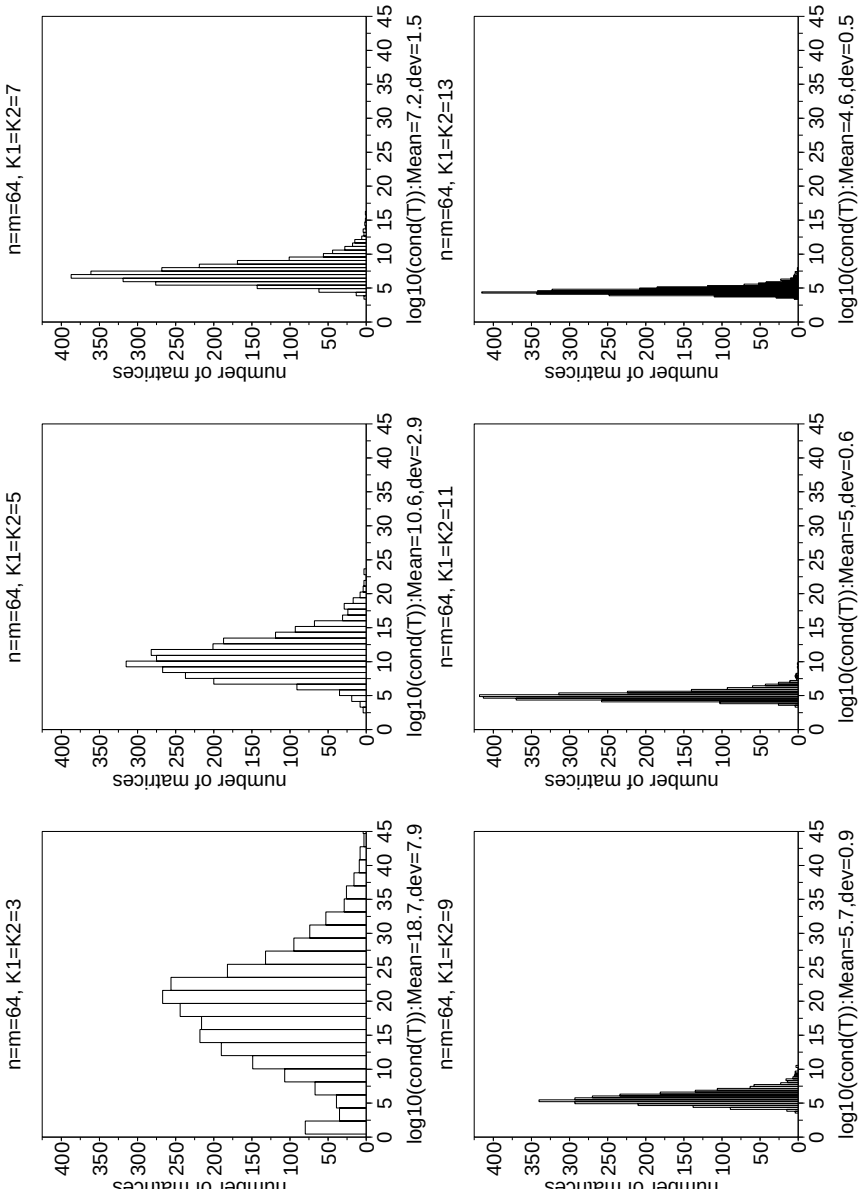


FIG. 2 – Les matrices ici sont de taille $64^2 \times 64^2$. c'est la même chose pour la largeur des bandes et le nombre des essaies

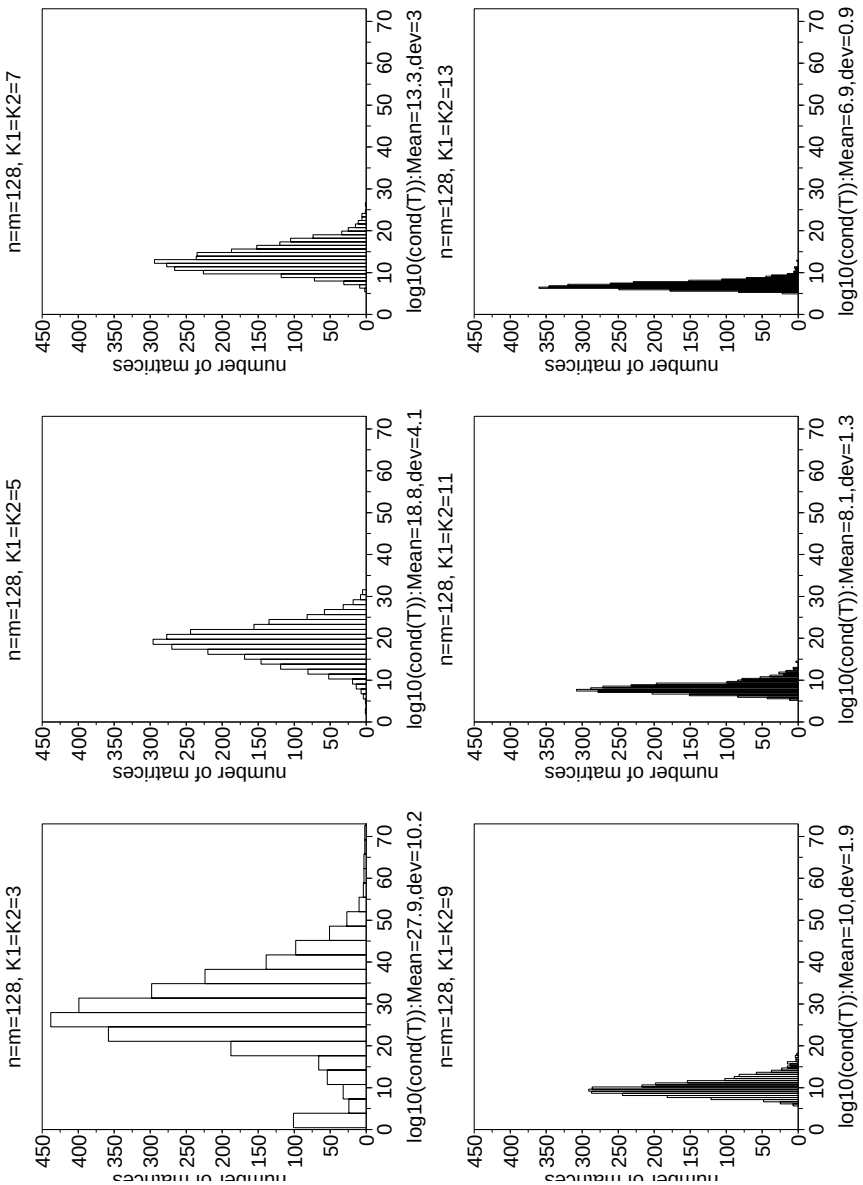


FIG. 3 – Matrices de Taille $128^2 \times 128^2$.

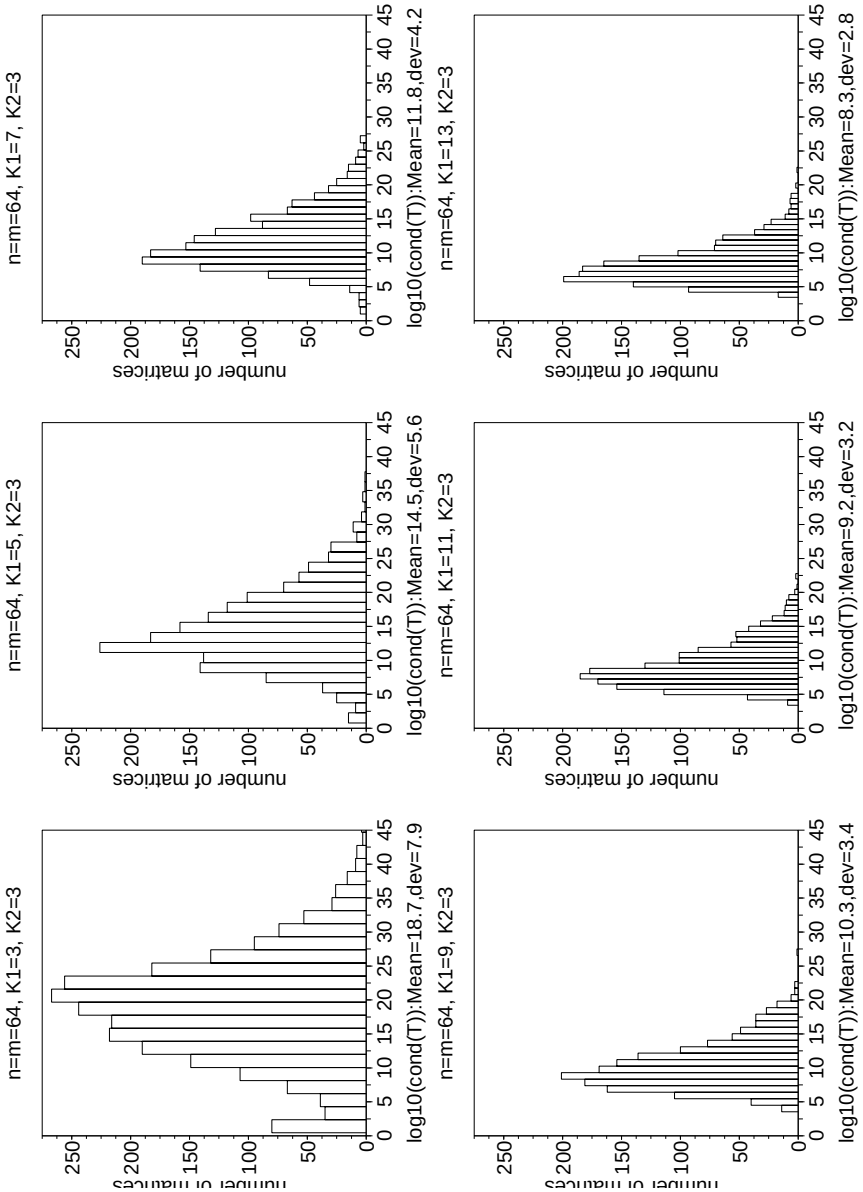


FIG. 4 – Les matrices sont de traille $64^2 \times 64^2$. Ici $k_1 = 1$ est fixe, $2k_2 - 1$ varie entre 3 et 13. Cet histogramme met en évidence dépendance entre largeur des bandes et nombre de conditionnement quand la largeur d'une seule bande est variable.

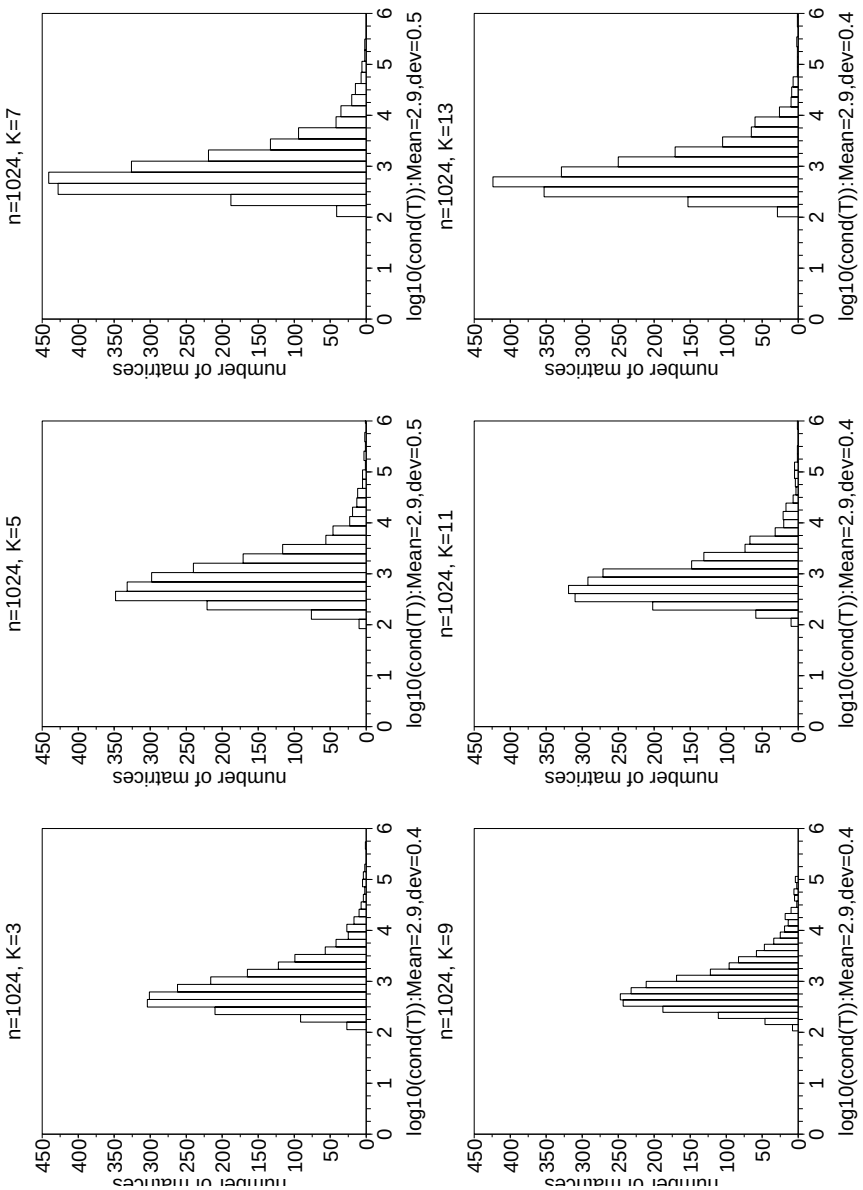


FIG. 5 – Les matrices ici sont de Toeplitz bande de taille 1024×1024, la largeur de bande varie entre 3 et 13. On remarque que le nombre de conditionnement ne varie pas avec la largeur de la bande.

3 Cas scalaire

Les idées de ces algorithmes proviennent de Bini et Pan [2], et les auteurs mentionnent un problème d'instabilité pour le premier algorithme.

3.1 Transformation en matrice circulante plus matrice de petit rang

Rappelons la Formule de Sherman-Morrison-Woodbury :

Théoreme 1. *Soient $A \in \mathbb{K}^{n\times n}$, G, $H \in \mathbb{K}^{n\times k}$. Si $I_k + H^T A^{-1} G$ n'est pas singulière, alors*

$$(A + GH^T)^{-1} = A^{-1} - A^{-1}G(I_k + H^T A^{-1} G)^{-1} H^T A^{-1}. \tag{3}$$

Démonstration. Une matrice par blocs peut se factoriser comme suit :

$$L = \begin{pmatrix} M & N \\ P & Q \end{pmatrix} = \begin{pmatrix} I & 0 \\ PM^{-1} & I \end{pmatrix} \begin{pmatrix} M & 0 \\ 0 & S \end{pmatrix} \begin{pmatrix} I & M^{-1}N \\ 0 & I \end{pmatrix},$$

avec S le complément de Schur donné par $S = Q - PM^{-1}N$. Bien entendu cette factorisation n'a de sens que si M est inversible. Par symétrie, on a également

$$L = \begin{pmatrix} M & N \\ P & Q \end{pmatrix} = \begin{pmatrix} I & NQ^{-1} \\ 0 & I \end{pmatrix} \begin{pmatrix} \tilde{S} & 0 \\ 0 & Q \end{pmatrix} \begin{pmatrix} I & 0 \\ Q^{-1}P & I \end{pmatrix},$$

avec $\tilde{S} = M - NQ^{-1}P$. En inversant les deux termes de l'égalité et en identifiant les deux expressions du bloc de la première ligne et de la première colonne on tire l'identité :

$$M^{-1} + M^{-1}NS^{-1}PM^{-1} = \tilde{S}^{-1}$$

qu'on applique à

$$L = \begin{pmatrix} A & G \\ H^T & -I_k \end{pmatrix}.$$

□

Proposition 1. *Soit $A \in \mathbb{K}^{n\times n}$ une matrice de Toeplitz bande, de largeur de bande $2k+1$. Alors on peut décomposer A sous la forme $C + R$ avec C une matrice circulante et R une matrice de rang au plus $2k$.*

Démonstration. La décomposition suivante est immédiate :

$$A = \begin{pmatrix} a_0 & a_{-1} & \dots & a_{-k} & 0 & \dots & 0 \\ a_1 & a_0 & \dots & a_{-k+1} & a_{-k} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \\ a_k & \ddots & \ddots & \ddots & \ddots & \ddots & a_{-k} \\ 0 & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & a_{-1} \\ 0 & \dots & 0 & a_k & \dots & a_1 & a_0 \end{pmatrix}$$

$$= \begin{pmatrix} a_0 & \dots & a_{-k} & 0 & a_k & \dots & a_1 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ a_k & \ddots & \ddots & \ddots & \ddots & \ddots & a_k \\ 0 & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \\ a_{-k} & \ddots & \ddots & \ddots & \ddots & \ddots & a_{-k} \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ a_{-1} & \dots & a_{-k} & 0 & a_k & \dots & a_0 \end{pmatrix} - \begin{pmatrix} 0 & \dots & 0 & a_k & \dots & a_1 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & \ddots & a_k \\ a_{-k} & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ a_{-1} & \dots & a_{-k} & 0 & \dots & 0 \end{pmatrix}$$

$$= C((a_0 \dots a_k\, 0 \dots 0\, a_{-k} \dots a_{-1})^T) + R = C + R.$$

□

Soit f une fonction d'entier à valeurs positives à l'infini. Rappelons que $\mathcal{O}(f(n))$ est une quantité bornée par $f(n)$ multiplié par une constante.

Corollaire 1. *La résolution du système $Ax = b$ coûte $\mathcal{O}(n \log n) + \mathcal{O}(nk \log k) + \mathcal{O}(k^3)$ opérations.*

Démonstration. En écrivant $R = GH^T$, avec $H, G \in \mathbb{K}^{n \times r}$ ($r = 2k$), et en utilisant la formule de Sherman-Morrison-Woodbury sur le système $(C+R)x = b$ on obtient

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

Remarquons tout d'abord que G est creuse avec $k(k+1)$ éléments non nuls :

$$G = \begin{pmatrix} 0_{k\times k} & g \\ 0_{(n-2k)\times k} & 0_{(n-2k)\times k} \\ f & 0_k \end{pmatrix} \text{ et } H = \begin{pmatrix} I_k & 0_{k\times k} \\ 0_{(n-2k)\times k} & 0_{(n-2k)\times k} \\ 0_{k\times k} & I_k \end{pmatrix},$$

avec $f = L(a_{-k} \dots a_{-1})^T$ et $g = U(a_k \dots a_1)^T$, et $L(v)$ (resp. $U(v)$ la matrice de Toeplitz triangulaire inférieure (resp. triangulaire supérieure) de première colonne (resp. première ligne) égale à v. Ainsi la multiplication d'une matrice de taille $n \times n$ par G coûte $\mathcal{O}(nk^2)$ opérations (la multiplication d'une matrice creuse qui contient p éléments non nuls par un vecteur coûte $2p$ opérations). On

peut même faire mieux, en multipliant G par une matrice $n \times n$ en $\mathcal{O}(nk \log k)$ comme f et g sont de Toeplitz.

Par suite x est obtenu en faisant :

- $v_1 = C^{-1}b$: $\mathcal{O}(n \log n)$ opérations,
- $v_2 = H^T v_1$: 0 opérations, parcequ'il n'y a pas que des 0 et des 1 dans H,
- $C^{-1}v_1$, avec $v_1 = G(I_r + H^T C^{-1} G)^{-1} H^T C^{-1} b$.
- $H^T C^{-1} G$: $\mathcal{O}(n \log n) + \mathcal{O}(nk \log k)$ opérations (calculer C^{-1} en $\mathcal{O}(n \log n)$ opérations puis la multiplier par G en $\mathcal{O}(nk \log k)$ opérations),
- $v_3 = (I_r + H^T C^{-1} G)^{-1} v_2$: $\mathcal{O}(k^3)$ opérations,
- $v_4 = G v_3$: $\mathcal{O}(k \log k)$ opérations,
- $v_5 = C^{-1} v_4$: $\mathcal{O}(n \log n)$ opérations.

□

Remarque 1. *On a $\mathcal{O}(nk \log k)$ est plus 'grand' que $\mathcal{O}(n \log n)$ si $k = \mathcal{O}(n^\alpha)$ pour un $\alpha \in [0, 1[$.*

3.1.1 Instabilité de l'algorithme

Dans leur livre, *Polynomial and matrix computation*, Victor Pan et Dario Bini, énoncent que cet algorithme rencontre des problèmes de stabilité. Les problèmes peuvent être dûs à l'instabilité propre du système linéaire. En effet, on pourra constater sur les figures que la comparaison entre nombre de conditionnement et erreur est normale. Pour se faire, nous avons généré des matrices de Toeplitz (et de Toeplitz bande par blocs Toeplitz bandes. Nous avons testé l'algorithme sur les deux types de matrices) bande pseudo aléatoires uniformes. Nous avons également généré des vecteurs x alatoires et nous avons calculé l'erreur entre x et $\tilde{x}$ qui est la *solution numérique*, via notre algorithme, de $T\tilde{x} = b$, b étant égal à Tx. Cette comparaison semble montrer que l'instabilité ne vient pas de l'algorithme.

3.2 Plongement dans une matrice engendrée par $Z + Z^T$

Dans cette section on travaille avec des matrices carrées de taille n. Soit τ_n (τ s'il y a pas de confusion) l'algèbre engendrée par $W = Z + Z^T$. Les matrices de Toepltiz dans cette section sont symétriques.

Proposition 2. *Soit $A \in \tau$, alors ses coefficients vérifient*

$$\begin{aligned} &a_{i-1,j} + a_{i+1,j} = a_{i,j+1} + a_{i,j-1} \\ &a_{0,j} = a_{n+1,j} = a_{i,0} = a_{i,n+1} = 0 \end{aligned} \tag{4}$$

Démonstration. Si A est dans τ, elle est de la forme $A = \sum_{i=0}^{n-1} \alpha_i W^i$. Or $W^{k+1} = W^k.W = W.W^k$ et si M une matrice alors $(MW)_{ij} = M_{i,j+1} + M_{i,j-1}$ et $(WM)_{i,j} = M_{i-1,j} + M_{i,j+1}$. Donc

$$(W^{k+1})_{ij} = (W^k)_{i,j-1} + (W^k)_{i,j+1} = (W^k)_{i-1,j} + (W^k)_{i+1,j}. \tag{5}$$

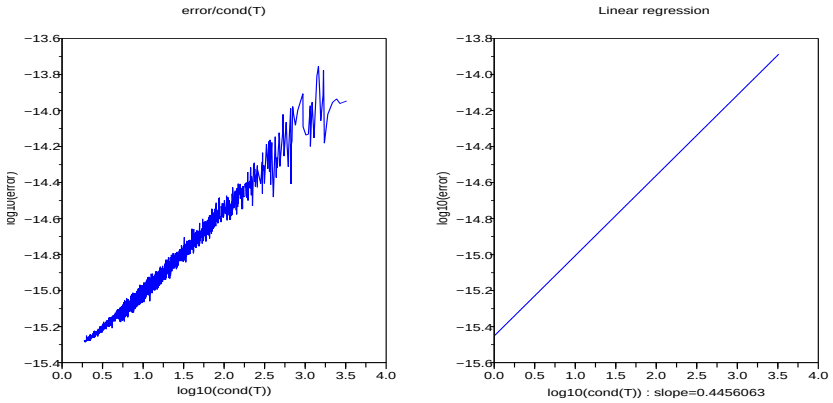


FIG. 6 – Les matrices ici sont de Toeplitz bande de taille 400×400 et la largeur de la bande égale 5. Pour 3000 essaies, on trace le logarithme à base 10 de l'erreur dû à notre algorithme par rapport au logarithme à base 10 du nombre de conditionnement. On trouve que la pente de la droite de regression est plus petit que 0.5 !

Montrons par recurrence sur k que chaque W^k vérifie la propriété (4) : W la vérifie. Supposons que W^k est telle que

$$(W^k)_{i,j-1} + (W^k)_{i,j+1} = (W^k)_{i-1,j} + (W^k)_{i+1,j},$$

D'après (5) :

$$(W^k)_{i,j-1}+(W^k)_{i,j+1} = (W^k)_{i-1,j-1}+(W^k)_{i+1,j-1}+(W^k)_{i-1,j+1}+(W^k)_{i+1,j+1},$$

et

$$(W^k)_{i-1,j}+(W^k)_{i+1,j} = (W^k)_{i-1,j-1}+(W^k)_{i-1,j+1}+(W^k)_{i+1,j-1}+(W^k)_{i+1,j+1},$$

parsuite $(W^k)_{i,j-1} + (W^k)_{i,j+1} = (W^k)_{i-1,j} + (W^k)_{i+1,j}$. □

Corollaire 2. *De là nous déduisons que l'algèbre τ peut être identifiée à*

$$B = \{B_k \in \tau; B_k e_1 = e_k\}.$$

Pour $A \in \tau$,

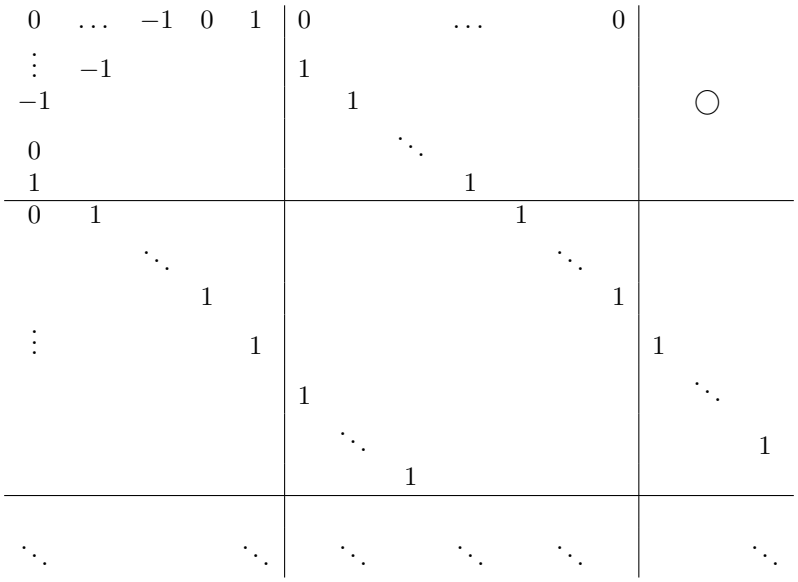
$$A = \sum_{i=1}^{n} a_i B_i,$$

avec $(a_1, \ldots, a_n)^T$ la première ligne de A.

Démonstration. La démonstration est immédiate : nous pouvons donner explicitement une base de l'espace vectoriel des matrices bandes Toeplitz construite à partir de $\tau_{n+2\lfloor k/2 \rfloor}$. □

Proposition 3. *Soit Δ l'espace vectoriel des matrices de Toeplitz $n \times n$, de largeur de bande $2k+1$. Soit dans $\tau_{n+2\lfloor k/2 \rfloor}$, le sous espace vectoriel engendré par les matrices B_i telles que $B_i e_1 = e_i - e_{i-2}$, avec $i = 3 \dots \lfloor k \rfloor + 2$, ainsi que $B_1 = I, B_2 = W$. Soit $E = \{\lfloor k/2 \rfloor + 1, \dots, n - \lfloor k/2 \rfloor - 1\}$. Alors les matrices $\tilde{B}_i = ((B_i)_{lp})_{l,p \in E}$ forment une base de Δ.*

Démonstration. En utilisant la technique donnée dans la proposition précédente pour calculer un élément dans τ, on remarque que B_i est de la forme :



La matrice B_i comporte un centre, c'est à dire l'intersection des lignes et des colonnes indexées par E=$\{\lfloor k/2 \rfloor$+1,...,$n-\lfloor k/2 \rfloor-1\}$, et une périphérique qui est son complémentaire

et $\tilde{B}_i$ est la matrice qui contient des 1 sur les diagonales qui commencent par les éléments $(1,i)$ et $(i,1)$ respectivement. Par suite les $\tilde{B}_i$ génèrent Δ $\square$

Soit $T \in \Delta$ telle que $T = \sum_{i=1}^{k+1} a_i \tilde{B}_i$. On peut donc la plonger dans une matrice $M \in \tau$ avec $M = \sum_{i=1}^{k+1} a_i B_i$ de taille $n + 2\lfloor k/2 \rfloor$. M a donc la forme suivante :

$$M = \begin{pmatrix} M_{11} & M_{12} & M_{13} \\ M_{12}^T & T & M_{23} \\ M_{13}^T & M_{23}^T & M_{33} \end{pmatrix},$$

avec M_{11}, M_{13} et M_{33} de taille $\lfloor k/2 \rfloor \times \lfloor k/2 \rfloor$, M_{12} et M_{23}^T de taille $\lfloor k/2 \rfloor \times n$. On cherche à résoudre le système $Tx = b$. Etudions le système suivant :

$$\begin{pmatrix} M_{11} & M_{12} & M_{13} \\ M_{12}^T & T & M_{23} \\ M_{13}^T & M_{23}^T & M_{33} \end{pmatrix} \begin{pmatrix} 0 \\ x \\ 0 \end{pmatrix} = \begin{pmatrix} b_1 \\ b \\ b_3 \end{pmatrix}.$$

Dans ce système, les inconnues sont x, b_1 et b_3 et on a $Tx = b$. Pour trouver b_1 et b_3 on procède comme suit : soit

$$M^{-1} = \begin{pmatrix} \mu_{11} & \mu_{12} & \mu_{13} \\ \mu_{12}^T & \mu_{22} & \mu_{23} \\ \mu_{13}^T & \mu_{23}^T & \mu_{33} \end{pmatrix}.$$

Les vecteurs b_1 et b_3 vérifient :

$$\begin{pmatrix} \mu_{11} & \mu_{12} & \mu_{13} \\ \mu_{12}^T & \mu_{22} & \mu_{23} \\ \mu_{13}^T & \mu_{23}^T & \mu_{11} \end{pmatrix} \begin{pmatrix} b_1 \\ b \\ b_3 \end{pmatrix} = \begin{pmatrix} 0 \\ x \\ 0 \end{pmatrix},$$

et par suite résolvent le système

$$\begin{cases} \mu_{11} b_1 + \mu_{13} b_3 = -\mu_{12} b, \\ \mu_{13}^T b_1 + \mu_{33} b_3 = -\mu_{23}^T b. \end{cases}$$

Ce système $2\lfloor k/2 \rfloor \times 2\lfloor k/2 \rfloor$ donne b_1 et b_3 en $\mathcal{O}(k^3)$ opérations. Pour construire ce système on a besoin aussi de calculer μ_{11}, μ_{12}, μ_{13} et μ_{23}, puis de calculer $\mu_{12} b$ et $\mu_{23}^T b$. comme μ_{12} et μ_{23}^T sont de taille $\lfloor k/2 \rfloor \times n$, ce calcul va coûter $\mathcal{O}(nk)$ opérations. Le calcul de μ_{11}, μ_{12}, μ_{13}, μ_{23} coûte $\mathcal{O}(n \log^2(n) + \mathcal{O}(k^2 n)$ opérations (voir corollaire ci-dessus).

Proposition 4. *Soit $M \in \tau_n$. On peut resoudre le système $Mx = b$ en $\mathcal{O}(n \log^2 n)$ opérations.*

Démonstration. $M \in \tau$ Comme M est dans τ_n, elle est de la forme

$$M = \sum_{i=1}^{n-1} m_i W^i.$$

Or les valeurs propres et vecteurs propres de W sont donnés par :

$$\lambda_k = 2 \cos \frac{k\pi}{n+1}, \quad v_k = \left(\sin \frac{jk\pi}{n+1} \right)_{1 \le j \le n}, \quad k = 1 \dots n,$$

comme on peut le vérifier simplement. Soit donc s la matrice de la transformation de Fourier en sinus :

$$s = \begin{pmatrix} \sin \frac{\pi}{n+1} & \dots & \sin \frac{n\pi}{n+1} \\ \vdots & \ddots & \vdots \\ \sin \frac{n\pi}{n+1} & \dots & \sin \frac{n^2\pi}{n+1} \end{pmatrix} = \left(\sin \frac{ij\pi}{n+1} \right)_{1 \le i,j \le n}.$$

Classiquement, $s^{-1} = 2s/(n+1)$. Si on pose $S = s\sqrt{2/(n+1)}$, alors $S^{-1} = S$ et $W = SDS$ avec D la matrice diagonale des valeurs propres. Nous en déduisons la diagonalisation de M :

$$M = S \left(\sum_{i=0}^{n-1} m_i D^i \right) S.$$

Le calcul de la somme des $m_i D^i$ se fait en $\mathcal{O}(n \log^2 n)$ (évaluation d'un polynôme (avec coefficients m_i) aux n valeurs propres λ_k). □

Corollaire 3. *Soit $M \in \tau_{n+2\lfloor k/2 \rfloor}$. Pour calculer la périphérie de taille $\lfloor k/2 \rfloor$ de M^{-1} on a besoin de $\mathcal{O}((n+k)\log^2(n+k)) + \mathcal{O}(nk^2)$ opérations.*

Démonstration. La matrice M est symétrique, elle est aussi symétrique par rapport à l'anti-diagonale. Pour calculer donc les $\lfloor k/2 \rfloor$ premières lignes, les $\lfloor k/2 \rfloor$ premières colonnes, les $\lfloor k/2 \rfloor$ dernières lignes et les $\lfloor k/2 \rfloor$ dernières colonnes de M^{-1} (la périphérie de M^{-1}) il suffira de calculer les $\lfloor k/2 \rfloor$ premières colonnes. On a donc besoin de resoudre $\lfloor k/2 \rfloor$ systèmes linéaires avec la matrice M. D'après la proposition précédente ce calcul coûtera $\mathcal{O}(k(n+k)\log^2(n+k))$ opérations. Ou, comme $M^{-1} \in \tau$, on peut calculer la première colonne de M^{-1}, en $\mathcal{O}((n+k)\log(n+k))$ opérations, et construire les autres colonnes demandées en utilisant la technique de la proposition (2), ce qui demande $\mathcal{O}(nk^2)$ opérations. □

Corollaire 4. *La résolution de $Tx = b$ coûte $\mathcal{O}(n \log^2 n) + \mathcal{O}(k^3)$ opérations.*

4 Cas par blocs

Le cas par blocs est une generalisation directe du cas scalaire mais où y rencontre quelques complications, et les comptes d'opérations sont différents.

4.1 Transformation en matrice circulante plus matrice de petit rang

Soit T une matrice comme dans (1) et (2).

Proposition 5. *On peut décomposer T en $T = C + R$, avec C une matrice circulante par blocs circulants et R de rang au plus $2(k_1 n + k_2 m)$, et au plus $\mathcal{O}(k_1^2 k_2 n + k_2^2 k_1 m)$ éléments non nuls.*

Démonstration. Ecrivons $T_i = C_i + R_i$, avec $-k_1 \leq i \leq k_1$, où C_i est une matrice circulante et R_i une matrice de rang au plus $2k_2$. Nous pouvons alors

décomposer T comme suit :

$$T = \begin{pmatrix} C_0 & \dots & C_{-k_1} & 0 & C_{k_1} & \dots & C_1 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ C_{k_1} & \ddots & \ddots & \ddots & \ddots & \ddots & C_{k_1} \\ 0 & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \\ C_{-k_1} & \ddots & \ddots & \ddots & \ddots & \ddots & C_{-k_1} \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ C_{-1} & \dots & C_{-k_1} & 0 & C_{k_1} & \dots & C_0 \end{pmatrix} -$$

$$\begin{pmatrix} 0 & \dots & 0 & C_{k_1} & \dots & C_1 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & \ddots & C_{k_1} \\ C_{-k_1} & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ C_{-1} & \dots & C_{-k_1} & 0 & \dots & 0 \end{pmatrix} - \begin{pmatrix} R_0 & \dots & R_{-k_1} & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ R_{k_1} & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & R_{-k_1} \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & R_{k_1} & \dots & R_0 \end{pmatrix}$$

$$= C + R_1 + R_2 = C + R.$$

Nous remarquons que R_1 est au plus de rang $2k_1n$ et R_2 est de rang $2k_2m$.

Dans R_1 il y a $2 \times k_1(k_1+1)/2$ blocs non nuls, et dans chaque bloc C_i il y a $n + 2(n-1) + 2(n-2) + \cdots + 2(n-k_2) + (k_2^2 + k_2) = 2k_2n + n$ éléments non nuls. Ainsi R_1 contient $(k_1^2 + k_1)(2k_2n + n) = \mathcal{O}(k_1^2k_2n)$ élément non nuls ; la situation est analogue pour R_2. □

En écrivant $R = GH^T$, avec $G, H \in \mathbb{K}^{N \times r}$ $(r = 2(k_1n + k_2m))$, et en appliquant la formule de Sherman-Morrison-Woodbury sur le système $(C + R)x = b$ on obtient

$$x = C^{-1}b - C^{-1}G(I_r + H^TC^{-1}G)^{-1}H^TC^{-1}b.$$

Proposition 6. *G est creuse avec $\mathcal{O}(k_1^2k_2n) + \mathcal{O}(k_2^2k_1m)$ éléments non nuls, et H ne comprend que des 0 et des 1.*

Démonstration. Les matrices G et H se décomposent en

$$G = (G_1\, G_2) \text{ et } H^T = \begin{pmatrix} H_1^T \\ H_2^T \end{pmatrix}$$

avec

$$G_1 = \begin{pmatrix} 0_{nk_1} & E \\ \vdots & 0 \\ 0 & \vdots \\ F & 0_{nk_1} \end{pmatrix},$$

pour décrire la matrice G_2, notons

$$E_l = \{(l-1)n+1, \ldots, (l-1)n+k_2\} \cup \{ln-k_2+1, \ldots, ln\}$$

et

$$E = \cup_{l=1}^{m} E_l$$

alors

$$G_2 = ((R_2)_{ij})_{\substack{1 \le i \le mn \\ j \in E}},$$

$$H_1^T = \begin{pmatrix} I_{nk_1} & 0 & \ldots & 0_{nk_1} \\ 0_{nk_1} & \ldots & 0 & I_{nk_1} \end{pmatrix},$$

et

$$H_2^T = \left(\begin{array}{cc|c|cc} I_{k_2} & 0_{k_2 \times (n-k_2)} & \ldots & I_{k_2} & 0_{k_2 \times (n-k_2)} \\ 0_{k_2 \times (n-k_2)} & I_{k_2} & & 0_{k_2 \times (n-k_2)} & I_{k_2} \end{array} \right)$$

chaque bloc dans H_2 est de taille $2k_2 \times m$.

Donc, en utilisant le compte fait dans la propostion précédente on voit que G contient $(k_1^2+k_1)(2k_2n+n)+(k_2^2+k_2)(2k_1m+m) = \mathcal{O}(k_1^2k_2n)+\mathcal{O}(k_2^2k_1m)$. □

Corollaire 5. *La multiplication de G par un vecteur coûte $\mathcal{O}(k_1^2k_2n)+\mathcal{O}(k_2^2k_1m)$ opérations.*

Corollaire 6. *En supposant que k_1 et k_2 sont petit devant m et n respectivement alors la résolution du système $Tx = b$ donné en (1) et (2) coûte $\mathcal{O}(N^{3/2})$ opérations.*

Démonstration. Soit $N = nm$, $K = k_1^2k_2n + k_2^2k_1m$, et $r = 2(k_1n + k_2m)$.

On a $x = C^{-1}b - C^{-1}G(I_r + H^TC^{-1}G)^{-1}H^TC^{-1}b$, donc x est obtenu en faisant :

- $v_1 = C^{-1}b : \mathcal{O}(N \log N)$.
- $v_2 = H^Tv_1$: 0 opération, car il n'y a que des 0 et des 1 dans H.
- $H^TC^{-1}G : \mathcal{O}(N \log N) + \mathcal{O}(NK) = \mathcal{O}(N \log N) + \mathcal{O}(N^{3/2})$, le calcul de C^{-1} demande $\mathcal{O}(N \log N)$ opérations et la multiplication par G demande $\mathcal{O}(NK)$ opérations car G est creuse.
- $v_3 = (I_r + H^TC^{-1}G)^{-1}v_2 : \mathcal{O}(r^3) = \mathcal{O}(N^{3/2})$, résolution classique.
- $v_4 = Gv_3 : \mathcal{O}(K) = \mathcal{O}(N^{1/2})$, multiplication de G par un vecteur, donc proportionel au nombre d'éléments non nuls dans G.
- $v_5 = C^{-1}v_4 : \mathcal{O}(N \log N)$.

□

4.1.1 Stabilité

Comme dans le cas scalaire, on a appliqué cet algorithme sur des matrices de Toeplitz bande par blocs Toeplitz bandes pseudo aléatoires pour obtenir les erreurs entre la solution de $Tx = b$ calculée par notre algorithme et la solution exacte. En traçant ces erreurs par rapport aux nombres de conditionnement, on

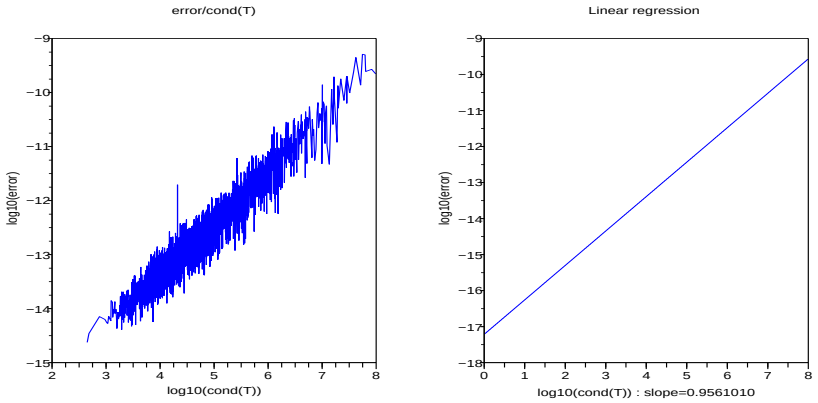


FIG. 7 – Les matrices sont de Toeplitz bande par blocs Toeplitz bande de taille $32^2 \times 32^2$ de largeur de bande $K_1 = K_2 = 5$. Pour 3000 essaies, on trace le logarithme à base 10 de l'erreur dû à notre algorithme par rapport au logarithme à base 10 du nombre de conditionnement. On trouve que la pente de la droite de regression est à peu près égale 1.

a obtenu une ligne de regression de pente $\simeq 1$. Ce qui prouve qu'il y a pas un problème de stabilité pour cet algorithme. Mais on remanque que la pente est sinificativement augmenté de $\simeq 0.5$ pour le cas scalaire à $\simeq 1$ pour le cas par blocs!!!

4.2 Plongement dans une matrice engendrée par $Z + Z^T$

Dans cette section T est une matrice de Toeplitz, par blocs de Toeplitz symétrique par blocs, et les blocs sont symétriques, bande par blocs, et les blocs sont aussi bandes. Les entiers m, n, k_1 et k_2 sont les dimensions utilisées dans les sections précédents. Pour résoudre le système $Tx = b$, on va essayer de plonger la matrice T dans une matrice de l'algèbre $\tau_{m,n}$ (ou τ s'il y a pas de confusion) engendrée par $W = (Z_m + Z_m^T) \otimes (Z_n + Z_n^T)$. Une matrice $M \in \tau$ est donc de la forme

$$M = \sum_{\substack{0 \le i \le m-1 \\ 0 \le j \le n-1}} \alpha_{ij} (Z_m + Z_m^T)^i \otimes (Z_n + Z_n^T)^j = \sum \alpha_{i,j} W^{(i,j)}.$$

Si M est de plus bande par blocs bandes, les bornes supérieures respectives de i et j dans la sommation sont k_1 et k_2. On va essayer de procéder comme dans le cas scalaire pour avoir un algorithme rapide.

Proposition 7. *Soit M dans τ. On peut resoudre le système $Mx = b$ en $\mathcal{O}(nm \log^2 mn)$ opérations.*

Démonstration. Notons s_k la matrice de la transformation en sinus de dimension k, $S_k = s_k\sqrt{2/(k+1)}$ et D_k la matrice diagonale $\mathrm{diag}(\lambda_1, \ldots, \lambda_k)$. Comme $M \in \tau$ alors $M = \sum \alpha_{ij} W^{(i,j)}$ avec $W := W_m \otimes W_n$ et $W^{(i,j)} := (Z_m + Z_m^T)^i \otimes (Z_n + Z_n^T)^j$. Or $W = W_m \otimes W_n = (S_m D_m S_m) \otimes (S_n D_n S_n) = (S_m \otimes S_n)(D_m \otimes D_n)(S_m \otimes S_n) =: SDS$. Donc M admet la diagonalisation :

$$M = SES \text{ avec } E = \sum_{\substack{0 \le i \le m-1 \\ 0 \le j \le n-1}} \alpha_{ij} D_m^i \otimes D_n^j.$$

E peut être calculé en $\mathcal{O}(mn \log^2 mn)$ opérations : il s'agit d'évaluer un polynôme à deux variable en mn points. Comme M^{-1} admet la decomposition $M^{-1} = SE^{-1}S$, la multiplication de M^{-1} par un vecteur coûte 2 transformées rapide en sinus et une multiplication par une matrice diagonale, le tout en $\mathcal{O}(mn \log mn)$ opérations. □

Soit T une matrice bande Toeplitz symétrique par blocs bande Toeplitz symétrique, comme en (1). De même technique que pour le cas scalaire, nous la plongons dans une matrice $M \in \tau$, en la plongeant tout d'abord par blocs, puis par blocs et après en plongeant chaque bloc.

La matrice M obtenue est une matrice de taille $m + 2\lfloor k_1/2 \rfloor$ par blocs, et chaque bloc est de taille $n + 2\lfloor k_2/2 \rfloor$; donc M est de taille $(m + 2\lfloor k_1/2 \rfloor)(n + 2\lfloor k_2/2 \rfloor) \simeq mn + mk_1 + nk_2 + k_1k_2$. Elle a la forme suivante :

$M =$

$$\left(\begin{array}{c|c|c}
M_{11} & M_{12} & M_{13} \\
\hline
M_{21} & \begin{array}{ccc|c|ccc} \times & \times & \times & & \times & \times & \times \\ \times & T_0 & \times & \cdots & \times & T_{n-1} & \times \\ \times & \times & \times & & \times & \times & \times \\ \hline \multicolumn{3}{c|}{\vdots} & \ddots & \multicolumn{3}{c}{\vdots} \\ \hline \times & \times & \times & & \times & \times & \times \\ \times & T_{n-1} & \times & \cdots & \times & T_0 & \times \\ \times & \times & \times & & \times & \times & \times \end{array} & M_{23} \\
\hline
M_{31} & M_{32} & M_{33}
\end{array}\right), \tag{6}$$

avec $M_{11}, M_{13}, M_{31}, M_{33}$ de taille $\lfloor k_1/2 \rfloor (n + 2\lfloor k_2/2 \rfloor) \times \lfloor k_1/2 \rfloor (n + 2\lfloor k_2/2 \rfloor)$ et M_{12}, M_{32} sont de taille $\lfloor k_1/2 \rfloor (n + 2\lfloor k_2/2 \rfloor) \times m(n + 2\lfloor k_2/2 \rfloor)$.

On cherche à resoudre $Tx = b$ avec

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, \quad b = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix}, \quad \text{les } x_i \text{ et les } b_i \text{ étant des vecteurs de taille } n.$$

Comme dans le cas scalaire, on va compléter le vecteur x par de nouvelles composantes nulles, dont les numéros de ligne sont les numéros des nouvelles ligne de M par rapport à T. Il faudra donc ajouter des lignes à b, ce qui créera de nouvelles inconnues :

$$\tilde{x} = \begin{pmatrix} 0_k \\ \bar{x} \\ 0_k \end{pmatrix} \text{ et } \tilde{b} = \begin{pmatrix} \hat{b}_1 \\ \bar{b} \\ \hat{b}_2 \end{pmatrix},$$

avec $k = \lfloor k_1/2 \rfloor (n + \lfloor k_2/2 \rfloor)$ et

$$\bar{x} = \begin{pmatrix} 0_{\bar{k}_2} \\ x_1 \\ \hline 0_{\bar{k}_2} \\ \hline \vdots \\ \hline 0_{\bar{k}_2} \\ x_m \\ 0_{\bar{k}_2} \end{pmatrix} \text{ et } \bar{b} = \begin{pmatrix} \hat{b}_{11} \\ b_1 \\ \hat{b}_{13} \\ \hline \vdots \\ \hline \hat{b}_{m1} \\ b_m \\ \hat{b}_{m3} \end{pmatrix},$$

avec $\bar{k_2} = \lfloor k_2/2 \rfloor$, les $\hat{b}_{i1}$ et les $\hat{b}_{i3}$ de taille $\bar{k_2}$. Soit

$$\hat{b} = \begin{pmatrix} \hat{b}_1 \\ \hat{b}_{11} \\ b_1 \\ \hat{b}_{13} \\ \vdots \\ \hat{b}_{m1} \\ b_m \\ \hat{b}m3 \\ \hat{b}_3 \end{pmatrix}$$

Comme dans le cas scalaire, écrivons $\tilde{x} = M^{-1}\tilde{b}$, et en ne gardant dans ce système que les lignes nulles dans $\tilde{x}$ et que les nouvelles inconnues, nous obtenons un système par rapport à $\hat{b}$, dont la matrice est de taille $\simeq nk_1 + mk_2$. Pour décrire ce système on va écrire M^{-1} d'une façon équivalente à l'écriture de M

donnée en (6), elle sera donc donnée comme suit :

$$M^{-1} = \left(\begin{array}{c|c|c} \mu_{11} & \mu_{12} & \mu_{13} \\ \hline \mu_{21} & \left(\begin{array}{c|c|c} \mu_{11}^{ij} & \mu_{12}^{ij} & \mu_{13}^{ij} \\ \hline \mu_{21}^{ij} & \mu^{ij} & \mu_{23}^{ij} \\ \hline \mu_{31}^{ij} & \mu_{32}^{ij} & \mu_{33}^{ij} \end{array}\right)_{i,j=1\ldots m} & \mu_{23} \\ \hline \mu_{31} & \mu_{32} & M_{33} \end{array}\right),$$

on découpe aussi μ_{12} (pareil pour μ_{32}) de la façon suivante :

$$\mu_{12} = \left(\ \mu_{12,11} \mid \mu_{12,1} \mid \mu_{12,13} \mid \ldots \mid \mu_{12,m1} \mid \mu_{12,m} \mid \mu_{12,m3}\ \right).$$

Le système à resoudre sera donné par :

$$\begin{cases} \mu_{11}\hat{b}_1 + \sum_{i=1}^{m}(\mu_{12,i1}\hat{b}_{i1} + \mu_{12,i3}\hat{b}_{i3}) + \mu_{13}\hat{b}_3 = -\sum_{i=1}^{m}\mu_{12,i}b_i \\ \mu_{12,j1}^T\hat{b}_1 + \sum_{i=1}^{m}(\mu_{11}^{ij}\hat{b}_{i1} + \mu_{13}^{ij}\hat{b}_{i3}) + \mu_{23,j1}\hat{b}_3 = -\sum_{i=1}^{m}\mu_{12}^{ij}b_i \ j = 1\ldots m \\ \mu_{12,j3}^T\hat{b}_1 + \sum_{i=1}^{m}(\mu_{31}^{ij}\hat{b}_{i1} + \mu_{33}^{ij}\hat{b}_{i3}) + \mu_{23,j3}\hat{b}_3 = -\sum_{i=1}^{m}\mu_{32}^{ij}b_i \ j = 1\ldots m \\ \mu_{31}\hat{b}_1 + \sum_{i=1}^{m}(\mu_{32,i1}\hat{b}_{i1} + \mu_{32,i3}\hat{b}_{i3}) + \mu_{33}\hat{b}_3 = -\sum_{i=1}^{m}\mu_{12,i}b_i \end{cases}$$

Pour former le deuxième membre de ce système il faut $\mathcal{O}(\lfloor k_1/2\rfloor(n+\lfloor k_2\rfloor).m(n+2\lfloor k_2/2\rfloor)) + \mathcal{O}(2m^2.\lfloor k_2/2\rfloor.n) \simeq \mathcal{O}(N^{3/2}) + \mathcal{O}(N^{3/2})$ opérations : 2 multiplications d'une matrice de taille $\lfloor k_1/2\rfloor(n+\lfloor k_2\rfloor) \times m(n+2\lfloor k_2/2\rfloor)$ par un vecteur et $2m^2$ multiplications matrice-vecteur avec des matrices de taille $\lfloor k_2/2\rfloor \times n$.

Proposition 8. *En supposant que k_1 et k_2 sont petit devant m et n respectivement alors le calcul de la périphérie (en deux dimension) de M^{-1} coûte $\mathcal{O}((n+k_1)(m+k_2)\log((n+k_1)(m+k_2))) + \mathcal{O}(N^{3/2})$ opérations.*

Démonstration. Même demonstration que pour le cas scalaire □

Finalement, la résolution du système initiale requiert $\mathcal{O}(N^{3/2})$ opérations pour être résolu.

Références

[1] Greg W. Anderson and Ofer Zeitouni. A CLT for a band matrix model. *Probab. Theory Related Fields*, 134(2) :283–338, 2006.

[2] Dario Bini and Victor Y. Pan. *Polynomial and matrix computations. Vol. 1.* Progress in Theoretical Computer Science. Birkhäuser Boston Inc., Boston, MA, 1994. Fundamental algorithms.

[3] Alice Guionnet. Large deviations and stochastic calculus for large random matrices. *Probab. Surv.*, 1 :72–172 (electronic), 2004.

[4] A. Khorunzhy. Sparse random matrices : spectral edge and statistics of rooted trees. *Adv. in Appl. Probab.*, 33(1) :124–140, 2001.

[5] A. Khorunzhy and W. Kirsch. On asymptotic expansions and scales of spectral universality in band random matrix ensembles. *Comm. Math. Phys.*, 231(2) :223–255, 2002.

[6] S. A. Molchanov, L. A. Pastur, and A. M. Khorunzhiĭ. Distribution of the eigenvalues of random band matrices in the limit of their infinite order. *Teoret. Mat. Fiz.*, 90(2) :163–178, 1992.