GRAVEJAT Philippe

Agrégation

24/11/2

Chapitre I : Résolution de systèmes linéaires

Introduction

Deux problèmes fondamentaux :

(i) Etant donnés une matrice inversible A et un vecteur b, déterminer la solution x du système linéaire : $Ax = b$.

(ii) Etant donnée une matrice carrée A, déterminer des éléments propres de A.

Un objectif : proposer des méthodes pratiques de résolution de ces deux problèmes.

I Rappels d'algèbre linéaire

Soit $N \in \mathbb{N}^*$.

1. Normes subordonnées

Soit $\| \ \|$, une norme sur l'espace vectoriel $\mathbb{C}^N$.

Définition : La norme subordonnée à la norme $\| \ \|$ sur l'espace $\mathcal{M}_N(\mathbb{C})$ est l'application $|||\ |||$ définie par :

$$\forall A \in \mathcal{M}_N(\mathbb{C}), \ |||A||| = \sup \{ \|Ax\|, \ x \in \mathbb{C}^N \text{ t.q. } \|x\| \leq 1 \}$$

Exemples : (i) Soit $\forall x \in \mathbb{C}^N, \ \|x\|_2 = \left(\sum_{i=1}^{N} |x_i|^2 \right)^{\frac{1}{2}}$. L'application $\| \ \|_2$ est une norme sur l'espace vectoriel $\mathbb{C}^N$, et sa norme subor-

donnée vaut :

$\forall A \in \mathcal{M}_n(\mathbb{K}),\ |||A|||_1 = \max_{1 \le j \le n} \sum_{i=1}^{n} |A_{i,j}|$.

(ii) Soit $\forall x \in \mathbb{K}^n,\ \|x\|_\infty = \max\{|x_j|,\ 1 \le j \le n\}$. L'application $\|\ \|_\infty$ est une norme sur l'espace vectoriel $\mathbb{K}^n$, et sa norme subordonnée vaut :

$\forall A \in \mathcal{M}_n(\mathbb{K}),\ |||A|||_\infty = \max_{1 \le i \le n} \sum_{j=1}^{n} |A_{i,j}|$.

Propriétés : (i) La norme subordonnée $|||\ |||$ à la norme $\|\ \|$ est bien définie et est une norme sur l'espace $\mathcal{M}_n(\mathbb{K})$.

(ii) La norme subordonnée $|||\ |||$ est définie de manière équivalente par des caractérisations :

$\forall A \in \mathcal{M}_n(\mathbb{K}),\ |||A||| = \max\{\|Ax\|,\ x \in \mathbb{K}^n \text{ t.q. } \|x\| \le 1\}$

$= \max\{\|Ax\|,\ x \in \mathbb{K}^n \text{ t.q. } \|x\| = 1\}$

$= \min\{\alpha \in \mathbb{R}_+ \text{ t.q. } \forall x \in \mathbb{K}^n,\ \|Ax\| \le \alpha \|x\|\}$.

(iii) En particulier, la norme subordonnée $|||\ |||$ satisfait l'inégalité :

$\forall x \in \mathbb{K}^n,\ \|Ax\| \le |||A||| \, \|x\|$.

(iv) La norme subordonnée $|||\ |||$ est une norme d'algèbre sur l'algèbre $\mathcal{M}_n(\mathbb{K})$:

$|||I_n||| = 1$ et $\forall (A,B) \in \mathcal{M}_n(\mathbb{K})^2,\ |||AB||| \le |||A||| \, |||B|||$.

Preuve :

Immédiate.

2. Relation avec le rayon spectral

Rappel : Soit $A \in \mathcal{M}_n(\mathbb{K})$.

(i) Le spectre $\sigma(A)$ de la matrice A est non vide.

(ii) La matrice A est trigonalisable : il existe une matrice inversible $P \in GL_n(\mathbb{K})$ telle que :

$$A = P^{-1} \begin{pmatrix} \lambda_1 & * \\ & \ddots \\ 0 & & \lambda_N \end{pmatrix} P, \quad \text{où} \quad \sigma(A) = \{\lambda_1, \dots, \lambda_N\}.$$

Définition : Soit $A \in \mathcal{M}_N(\mathbb{C})$. Le rayon spectral $\rho(A)$ de la matrice A est défini par :

$$\rho(A) = \max \{ |\lambda| , \lambda \in \sigma(A) \}.$$

Exemples : (i) $\rho(0) = 0$ et $\rho(I_N) = 1$

(ii) Si une matrice $U \in \mathcal{U}_N(\mathbb{C})$ est unitaire, alors : $\rho(U) = 1$.

Propriété : Soit $|||\ |||$ la norme subordonnée à la norme $\|\ \|$ de $\mathbb{C}^N$.

(i) Si $A \in \mathcal{M}_N(\mathbb{C})$, alors :

$$\rho(A) \leq |||A|||$$

(ii) La réciproque est fausse.

Preuve :

(i) Soit $\lambda \in \sigma(A)$ telle que $|\lambda| = \rho(A)$, et $v \in \mathbb{C}^N \setminus \{0\}$ tel que $Av = \lambda v$

alors : $\rho(A) \|v\| = \|Av\| \leq |||A||| \|v\| \Rightarrow |||A||| \geq \rho(A)$.

(ii) $\rho \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} = 0 < \left|\left|\left| \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \right|\right|\right|$.

Théorème : Soit $\varepsilon \in]0, +\infty[$ et $A \in \mathcal{M}_N(\mathbb{C})$. Il existe une norme subordonnée $|||\ |||_A$ à une norme $\|\ \|_A$ de $\mathbb{C}^N$ telle que :

$$|||A|||_A \leq \rho(A) + \varepsilon.$$

Preuve :

Soit $P \in GL_N(\mathbb{C})$ et $(\lambda_1, \dots, \lambda_N) \in \mathbb{C}^N$ tels que :

$$A = P^{-1} \begin{pmatrix} \lambda_1 & t_{12} & \cdots & t_{1N} \\ & \ddots & \ddots & \vdots \\ & & \ddots & t_{N-1,N} \\ 0 & & & \lambda_N \end{pmatrix} P$$

Pour $\delta > 0$, introduisons la matrice $D_\delta = \begin{pmatrix} 1 & & & 0 \\ & \delta & & \\ & & \ddots & \\ 0 & & & \delta^{N-1} \end{pmatrix}$. Il vient :

$$\left(D_\delta^{-1} P\right) A \left(D_\delta^{-1} P\right)^{-1} = \begin{pmatrix} \lambda_1 & t_{12}\delta & \cdots & t_{1N}\delta^{N-1} \\ & \ddots & \ddots & \vdots \\ & & \ddots & t_{N-1,N}\delta \\ 0 & & & \lambda_N \end{pmatrix},$$

De suite que :

$$\left\| \left(D_\delta^{-1} L\right) A \left(D_\delta^{-1} L\right)^{-1} \right\|_\infty = \max_{1 \le i \le N} |\lambda_i| + \sum_{j=i+1}^{N} |t_{i,j}| \, \delta^{j-i}$$

Fixons alors le choix de δ tel que :

$$\forall\, 1 \le i \le N, \quad \sum_{j=i+1}^{N} |t_{i,j}| \, \delta^{j-i} \le \varepsilon$$

Dans ce cas, nous obtenons :

$$\left\| \left(D_\delta^{-1} L\right) A \left(D_\delta^{-1} L\right)^{-1} \right\|_\infty \le \rho(A) + \varepsilon.$$

Enfin, nous savons que :

$$\begin{aligned} \|A\|_A &:= \left\| \left(D_\delta^{-1} L\right) A \left(D_\delta^{-1} L\right)^{-1} \right\|_\infty \\ &= \min \left\{ \alpha \in \mathbb{R}_+ \text{ t.q. } \forall x \in \mathbb{C}^N, \ \left\| \left(D_\delta^{-1} L\right) A \left(D_\delta^{-1} L\right)^{-1}(x) \right\|_\infty \le \alpha \|x\|_\infty \right\} \\ &= \min \left\{ \alpha \in \mathbb{R}_+ \text{ t.q. } \forall y = \left(D_\delta^{-1} L\right)^{-1}(x) \in \mathbb{C}^N, \ \left\| \left(D_\delta^{-1} L\right) A(y) \right\|_\infty \le \alpha \left\| \left(D_\delta^{-1} L\right) y \right\|_\infty \right\}, \end{aligned}$$

De suite que l'application $\|\cdot\|_A$ est la norme subordonnée à la norme de $\mathbb{C}^N$ définie par :

$$\forall x \in \mathbb{C}^N, \quad \|x\|_A = \left\| D_\delta^{-1} L(x) \right\|_\infty,$$

Ce qui conclut la preuve du théorème.

Théorème : Soit $\forall x \in \mathbb{C}^N, \ \|x\|_2 = \left(\sum_{j=1}^{N} |x_j|^2 \right)^{\frac{1}{2}}$.

(i) L'application $\|\ \|_2$ est une norme sur l'espace vectoriel $\mathbb{C}^N$.

(ii) La norme subordonnée $|||\ |||_2$ vaut :

$$\forall A \in \mathcal{M}_N(\mathbb{C}), \quad |||A|||_2 = \rho\left(A^* A\right)^{\frac{1}{2}} = \rho\left(A A^*\right)^{\frac{1}{2}}.$$

En particulier :

(iii) $\forall A \in \mathcal{M}_N(\mathbb{C}), \ |||A^*|||_2 = |||A|||_2$

(iv) $\forall U \in \mathcal{U}_N(\mathbb{C}), \ |||U|||_2 = 1.$

(v) $\forall A \in \mathcal{M}_N(\mathbb{C}), \ \forall U \in \mathcal{U}_N(\mathbb{C}), \ |||U^* A U|||_2 = |||UA|||_2 = |||AU|||_2 = |||A|||_2.$

(vi) En particulier, si la matrice $N \in \mathcal{M}_N(\mathbb{C})$ est normale, alors :

$$|||N|||_2 = \rho(N).$$

Preuve :

(i) Immédiate.

(ii) Par définition, nous savons que :

$$|||A|||_2^2 = \max \left\{ \|A(x)\|_2^2, \ x \in \mathbb{C}^N \text{ t.q. } \|x\|_2 \le 1 \right\}.$$

Comme

$\forall x \in \mathbb{C}^N$, $\|A(x)\|_2^2 = \langle A^*A(x), x\rangle_2$,

il vient :

$\|A\|_2^2 = \max \{ \langle A^*A(x), x\rangle_2, x \in \mathbb{C}^N \text{ t.q. } \|x\|_2 \leq 1 \}$.

La matrice A^*A est hermitienne positive de sorte qu'il existe une matrice unitaire $U \in \mathcal{U}_N(\mathbb{C})$ et des nombres positifs $\lambda_1 \leq \dots \leq \lambda_N$ tels que :

$$A^*A = U^* \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_N \end{pmatrix} U .$$

Comme l'application $y \mapsto U^*y$ est une isométrie de l'espace vectoriel hermitien $(\mathbb{C}^N, \langle , \rangle_2)$, nous arrivons au fait que :

$$\|A\|_2^2 = \max \left\{ \left\langle \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_N \end{pmatrix} (y), y \right\rangle_2, y \in \mathbb{C}^N \text{ t.q. } \|y\|_2 \leq 1 \right\}$$

$$= \lambda_N$$

$$= \rho(A^*A)$$

Deux cas se présentent alors :

- si $\rho(A^*A) = 0$, alors,

$\|A\|_2 = 0 \Rightarrow A = 0 \Rightarrow \rho(AA^*) = 0 = \rho(A^*A) = \|A\|_2$

- sinon, il existe un vecteur $x \in \mathbb{C}^N \setminus \{0\}$ tel que :

$A^*A(x) = \rho(A^*A)\, x \Rightarrow (AA^*)(A(x)) = \rho(A^*A) A(x)$.

Comme

$\|A(x)\|_2^2 = \langle A^*A(x), x\rangle_2 = \rho(A^*A) \|x\|_2^2 > 0$,

la matrice hermitienne positive AA^* a pour valeur propre $\rho(A^*A)$, de sorte que :

$\rho(A^*A) \leq \rho(AA^*)$.

De même, il vient :

$\rho(AA^*) = \rho((A^*)^*A^*) \leq \rho(A^*A)$,

De sorte que l'égalité a lieu.

(iii) Immédiat

(iv) Immédiat.

(v) Cette propriété découle en particulier du fait que l'application $y \mapsto U^*y$ est une isométrie de l'espace hermitien $(\mathbb{C}^N, \langle , \rangle_2)$.

(vi) Si la matrice N est normale, alors, il existe une matrice unitaire $U \in \mathcal{U}_N(\mathbb{C})$ et

Les nombres complexes $\lambda_1, \dots, \lambda_n$ tels que :

$$N = U^* \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix} U ,$$

On en déduit que :

$$|||N|||_2 = \left|\left|\left| \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix} \right|\right|\right|_2 = \max_{1 \le i \le n} |\lambda_i| = \rho(N).$$

3. Convergence de la suite des puissances itérées

Rappel : Soit $(A_m)_{m \in \mathbb{N}}$ une suite de $\mathcal{M}_n(\mathbb{C})$ et $A_\infty \in \mathcal{M}_n(\mathbb{C})$.

La suite $(A_m)_{m \in \mathbb{N}}$ converge vers la matrice A_∞ dans $\mathcal{M}_n(\mathbb{C})$ si il existe une norme $|||\ |||$ sur l'espace vectoriel $\mathcal{M}_n(\mathbb{C})$ telle que :

$$|||A_m - A_\infty||| \xrightarrow[m \to +\infty]{} 0.$$

En particulier, cette propriété ne dépend pas du choix de la norme $|||\ |||$ sur $\mathcal{M}_n(\mathbb{C})$.

Théorème : Soit $A \in \mathcal{M}_n(\mathbb{C})$.

(i) La suite des puissances itérées $(A^m)_{m \in \mathbb{N}}$ converge vers la matrice nulle dans $\mathcal{M}_n(\mathbb{C})$ si le rayon spectral de la matrice A satisfait l'inégalité : $\rho(A) < 1$.

(ii) Dans ce cas, la matrice $I_n - A$ est inversible, et son inverse est donnée par la série convergente :

$$(I_n - A)^{-1} = \sum_{m=0}^{+\infty} A^m.$$

Preuve :

(i) Si $A^m \xrightarrow[m \to +\infty]{} 0$ dans $\mathcal{M}_n(\mathbb{C})$ et si λ désigne une valeur propre de la matrice A telle que :

$$|\lambda| = \rho(A),$$

alors, il existe un vecteur propre $v \in \mathbb{C}^n \setminus \{0\}$ tel que :

$$A(v) = \lambda v \implies \forall m \in \mathbb{N},\ A^m(v) = \lambda^m v.$$

Il s'ensuit que :

$\rho(A)^n \|x\|_2 = |\lambda|^n \|x\|_2 = \|A^n(x)\|_2 \leq |||A^n|||_2 \, \|x\|_2 \xrightarrow[n\to+\infty]{} 0$,

ce qui entraine que: $\rho(A) < 1$.

Réciproquement, si $\rho(A) < 1$, alors, il existe une norme subordonnée $|||\cdot|||$ sur $\mathcal{M}_n(\mathbb{C})$ telle que:

$|||A||| < 1$.

Comme

$\forall n \in \mathbb{N}$, $|||A^n||| \leq |||A|||^n$,

il est clair que:

$A^n \xrightarrow[n\to+\infty]{} 0$ dans $\mathcal{M}_n(\mathbb{C})$.

(ii) Soit $|||\cdot|||$, une norme subordonnée sur $\mathcal{M}_n(\mathbb{C})$ telle que:

$|||A||| < 1$.

Comme

$\forall n \in \mathbb{N}$, $|||A^n||| \leq |||A|||^n$,

la série $\sum A^n$ est absolument convergente, donc convergente, et:

$(I_n - A)\left(\sum_{n=0}^{+\infty} A^n\right) = \left(\sum_{n=0}^{+\infty} A^n\right)(I_n - A) = I_n$.

La matrice $I_n - A$ est donc inversible d'inverse égal à la somme $\sum_{n=0}^{+\infty} A^n$.

Corollaire (Formule du rayon spectral): Soit $A \in \mathcal{M}_n(\mathbb{C})$ et $|||\cdot|||$, une norme d'algèbre sur l'algèbre $\mathcal{M}_n(\mathbb{C})$. Alors:

$$\rho(A) = \lim_{n\to+\infty} |||A^n|||^{\frac{1}{n}}.$$

Preuve:

Soit $\lambda \in \sigma(A)$ telle que $|\lambda| = \rho(A)$, et $x \in \mathbb{C}^n \setminus \{0\}$ tel que $A(x) = \lambda x$. Alors:

$\rho(A)^n |||x \; {}^t x||| = |||\lambda^n x \; {}^t x||| = |||A^n x \; {}^t x||| \leq |||A^n||| \, |||x \; {}^t x|||$.

Comme $|||x \; {}^t x||| \neq 0$, il vient:

$\rho(A) \leq |||A^n|||^{\frac{1}{n}} \Rightarrow \rho(A) \leq \lim_{n\to+\infty} |||A^n|||^{\frac{1}{n}}$.

Soit alors $\varepsilon > 0$. Il vient:

$\rho\left(\frac{A}{\rho(A)+\varepsilon}\right) < 1$,

De sorte que

$\frac{A^n}{(\rho(A)+\varepsilon)^n} \xrightarrow[n\to+\infty]{} 0$.

En particulier, il existe un entier $n \in \mathbb{N}$ tel que:

$\forall m \geq m_\varepsilon$, $|||A^m|||^{\frac{1}{m}} \leq \rho(A) + \varepsilon \Rightarrow \overline{\lim_{m \to +\infty}} |||A^m|||^{\frac{1}{m}} \leq \rho(A) + \varepsilon$.

En conclusion, nous avons :

$$\rho(A) = \lim_{m \to +\infty} |||A^m|||^{\frac{1}{m}}.$$

II Conditionnement de problèmes linéaires

1. Position du problème

Considérons la matrice : $A := \begin{pmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{pmatrix}$, et le vecteur : $b := \begin{pmatrix} 32 \\ 23 \\ 33 \\ 31 \end{pmatrix}$.

La matrice A est symétrique et inversible, et son inverse A^{-1} est égale à :

$$A^{-1} = \begin{pmatrix} 25 & -41 & 10 & -6 \\ -41 & 68 & -17 & 10 \\ 10 & -17 & 5 & -3 \\ -6 & 10 & -3 & 2 \end{pmatrix}.$$

L'unique solution x du système linéaire $Ax = b$ est égale à : $x = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$.

Les calculs numériques sont approchés et doivent prendre en compte les problèmes d'arrondis : de légères modifications de la matrice A ou du second membre b ne doivent pas entraîner des erreurs trop importantes sur le calcul du vecteur solution x. Qu'en est-il ici ?

Lorsque le vecteur b est remplacé par le vecteur $b' := \begin{pmatrix} 32,1 \\ 22,9 \\ 33,1 \\ 30,9 \end{pmatrix}$, l'unique solution x' est désormais égale à : $x' = \begin{pmatrix} 9,2 \\ -24,6 \\ 4,5 \\ -1,1 \end{pmatrix}$.

Une erreur relative de l'ordre de $1/100$ entraîne une erreur relative sur la solution de l'ordre de $10/1$: l'amplification des erreurs est de l'ordre de 1000, ce qui est colossal.

Lorsque la matrice A est remplacée par la matrice $A' := \begin{pmatrix} 10 & 7 & 8{,}1 & 7{,}2 \\ 7{,}08 & 5{,}04 & 6 & 5 \\ 8 & 5{,}98 & 9{,}89 & 9 \\ 6{,}99 & 4{,}99 & 9 & 9{,}98 \end{pmatrix}$

l'unique solution x' devient : $x' := \begin{pmatrix} -81 \\ 137 \\ -34 \\ 22 \end{pmatrix}$. L'amplification des erreurs est à nouveau colossale, et ce bien que les matrices A, A^{-1}, et le second membre b est l'air des plus raisonnables.

La notion de conditionnement permet de quantifier cette difficulté, et parfois d'y remédier à travers l'introduction de méthodes de préconditionnement.

Notons pour terminer qu'un problème identique apparaît dans la résolution de problèmes de valeurs propres. Le spectre de la matrice $B = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$

est réduit au singleton $\{0\}$, tandis que le spectre de la matrice $B' = \begin{pmatrix} 0 & 0 & 0 & 10^{-4} \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$

est donné par : $\sigma(B') = \{\pm 10^{-1}, \pm 10^{-1} i\}$. L'amplification de l'erreur relative est dans ce cas de l'ordre de 10000 !

2. Conditionnement d'un système linéaire

Soit $N \in \mathbb{N}^*$, et $\| \ \|$, une norme sur l'espace vectoriel $\mathbb{C}^N$. Désignons par la notation $|||\ |||$, la norme subordonnée à la norme $\| \ \|$ sur l'espace $\mathcal{M}_N(\mathbb{C})$.

Définition : Soit $A \in \mathcal{GL}_N(\mathbb{C})$. Le conditionnement de la matrice A **relativement à la norme subordonnée** $|||\ |||$ est le nombre strictement positif :

$$\mathrm{cond}(A) := |||A||| \times |||A^{-1}|||.$$

Exemples : (i) $\mathrm{cond}(I_N) = 1$ (quel que soit le choix de la norme $\| \ \|$ sur $\mathbb{C}^N$)

(ii) Si $|||\ |||_2$ est la norme subordonnée à la norme hermitienne

canonique $\|\ \|_2$ sur $\mathbb{C}^n$, alors :

$\forall\, U \in \mathcal{U}_n(\mathbb{C})$, $\mathrm{cond}_2(U) := \|U\|_2 \, \|U^{-1}\|_2 = 1$.

Propriétés : Soit $A \in \mathcal{GL}_n(\mathbb{C})$. Le conditionnement de la matrice A relativement à la norme subordonnée $\|\ \|$ vérifie les propriétés suivantes :

(i) $\mathrm{cond}(A) \geq 1$.

(ii) $\mathrm{cond}(A) = \mathrm{cond}(A^{-1})$.

(iii) $\forall\, \lambda \in \mathbb{C}^*$, $\mathrm{cond}(\lambda A) = \mathrm{cond}(A)$.

Preuve :

(i) Comme $I_n = A\, A^{-1}$, il vient :

$1 = \|I_n\| \leq \|A\| \, \|A^{-1}\| = \mathrm{cond}(A)$

(ii) (iii) Immédiat.

Dans le cas particulier de la norme subordonnée $\|\ \|_2$ à la norme hermitienne canonique $\|\ \|_2$ de $\mathbb{C}^n$, le conditionnement cond_2 vérifie les propriétés suivantes.

Propriétés : Soit $A \in \mathcal{GL}_n(\mathbb{C})$.

(i) Si $U \in \mathcal{U}_n(\mathbb{C})$, alors : $\mathrm{cond}_2(A) = \mathrm{cond}_2(AU) = \mathrm{cond}_2(UA) = \mathrm{cond}_2(U^* A U)$.

(ii) Si $\nu_2(A)$ et $\nu_n(A)$ désignent la plus petite, resp. la plus grande valeur propre de la matrice hermitienne définie positive A^*A, alors : $\mathrm{cond}_2(A) = \left(\dfrac{\nu_n(A)}{\nu_2(A)}\right)^{\frac{1}{2}}$.

(iii) Si la matrice A est normale, et si son spectre est donné par l'ensemble $\sigma(A) = \{\lambda_1(A), \ldots, \lambda_n(A)\}$, alors :

$$\mathrm{cond}_2(A) = \frac{\max_{1 \leq i \leq n} |\lambda_i(A)|}{\min_{1 \leq i \leq n} |\lambda_i(A)|}.$$

Preuve :

(i) Immédiat

(ii) Rappelons que :

$\forall \Pi \in \mathcal{U}_N(\mathbb{C})$, $\|\Pi\|_2 = \rho(\Pi^*\Pi)^{\frac{1}{2}} = \rho(\Pi\Pi^*)^{\frac{1}{2}}$

De sorte que :

$$\mathrm{cond}_2(A) = \rho(A^*A)^{\frac{1}{2}}\,\rho\left((A^*A)^{-1}\right)^{\frac{1}{2}} = \left(\frac{\nu_N(A)}{\nu_1(A)}\right)^{\frac{1}{2}}.$$

(ii) Lorsque la matrice $A \in \mathcal{U}_N(\mathbb{C})$ est normale, son inverse l'est également. En particulier, il vient :

$$\mathrm{cond}_2(A) = \rho(A)\,\rho(A^{-1}),$$

De sorte que :

$$\mathrm{cond}_2(A) = \frac{\max_{1\leq i\leq N} |\lambda_i(A)|}{\min_{1\leq i\leq N} |\lambda_i(A)|}.$$

Exemple : La formule précédente permet de calculer une valeur approchée du conditionnement relativement à la norme $\|\|_2$ de la matrice

$$A = \begin{pmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{pmatrix},$$ laquelle est égale à : $\mathrm{cond}_2(A) \simeq 2984$.

Comme le soulignent les deux théorèmes qui suivent, cette valeur élevée du conditionnement explique les problèmes d'arrondis mentionnés plus tôt.

Le conditionnement d'une matrice permet d'évaluer les erreurs relatives dans la résolution d'un système linéaire de la façon suivante.

Théorème : Soit $A \in \mathcal{GL}_N(\mathbb{C})$.

(i) Soit $b \in \mathbb{C}^N \setminus \{0\}$ et $\delta b \in \mathbb{C}^N$. Si x et $x+\delta x$ désignent les solutions des équations $Ax = b$, resp. $A(x+\delta x) = b+\delta b$, alors :

$$\frac{\|\delta x\|}{\|x\|} \leq \mathrm{cond}(A)\,\frac{\|\delta b\|}{\|b\|}$$

(ii) De plus, il existe des vecteurs $b \in \mathbb{C}^N \setminus \{0\}$ et $\delta b \in \mathbb{C}^N$ tels que cette inégalité soit une égalité.

Preuve :

(i) Par linéarité, il vient :

$$A(\delta x) = \delta b,$$

De sorte que:

$$\begin{cases} \|\delta x\| \leq |||A^{-1}||| \, \|\delta b\|, \\ \|b\| \leq |||A||| \, \|x\|, \end{cases}$$

Soit:

$$\frac{\|\delta x\|}{\|x\|} \leq \text{cond}(A) \frac{\|\delta b\|}{\|b\|}$$

(ii) Cette inégalité devient une égalité lorsque les vecteurs b et δb sont choisis de façon à satisfaire les identités:

$$\begin{cases} \|A^{-1}(\delta b)\| = |||A^{-1}||| \, \|\delta b\|, \\ \|b\| \leq |||A||| \, \|A^{-1}(b)\|, \end{cases}$$

Ce qui est possible.

Théorème: Soit $A \in GL_n(\mathbb{C})$ et $\delta A \in \mathcal{M}_n(\mathbb{C})$ telle que $A + \delta A$ soit inversible.

(i) Soit $b \in \mathbb{C}^n \setminus \{0\}$. Si x et $x + \delta x$ désignent les solutions des équations $Ax = b$, resp. $(A + \delta A)(x + \delta x) = b$, alors:

$$\frac{\|\delta x\|}{\|x + \delta x\|} \leq \text{cond}(A) \frac{|||\delta A|||}{|||A|||}$$

(ii) De plus, il existe un vecteur $b \in \mathbb{C}^n \setminus \{0\}$ et une matrice $\delta A \in \mathcal{M}_n(\mathbb{C}) \setminus \{0\}$ telle que cette inégalité soit une égalité.

(iii) Supposons que: $|||\delta A||| \leq \frac{1}{2|||A^{-1}|||}$. Si $b \in \mathbb{C}^n \setminus \{0\}$ et si $x \neq x + \delta x$ désignent les solutions des équations $Ax = b$, resp. $(A + \delta A)(x + \delta x) = b$, alors:

$$\frac{\|\delta x\|}{\|x\|} \leq 2 \, \text{cond}(A) \frac{|||\delta A|||}{|||A|||}.$$

Preuve:

(i) Par linéarité, il vient:

$$A(\delta x) + \delta A(x + \delta x) = 0,$$

De sorte que:

$$\|\delta x\| \leq |||A^{-1}||| \, |||\delta A||| \, \|x + \delta x\| \Rightarrow \frac{\|\delta x\|}{\|x + \delta x\|} \leq \text{cond}(A) \frac{|||\delta A|||}{|||A|||}.$$

(ii) Soit $\lambda \in \mathbb{C}^*$ tel que $A + \lambda I_n$ soit inversible et $\delta A := \lambda I_n$. Il existe un vecteur $\delta b \in \mathbb{C}^n \setminus \{0\}$ tel que:

$$\|A^{-1}(\delta b)\| = |||A^{-1}||| \, \|\delta b\|.$$

Si $b := \lambda^{-1}(A(\delta b) + \lambda \delta b)$, alors, l'unique solution de l'équation $A(x) = b$ est égale à:

$x = \lambda^{-1}(\delta b) + A^{-1}(\delta b)$,

De sorte que l'unique solution de l'équation $A(\delta x) + \delta A(x+\delta x) = 0$ vaut :

$\delta x = -\lambda (A + \lambda I_N)^{-1}(x) = -A^{-1}(\delta b)$.

Il vient donc :

$x + \delta x = \lambda^{-1}(\delta b)$,

De sorte que :

$\| \delta x \| = |||A^{-1}||| \, \|\delta b\| = \lambda \, |||A^{-1}||| \, \|x+\delta x\| = \text{cond}(A) \dfrac{|||\delta A|||}{|||A|||} \|x+\delta x\|$.

(iii) Lorsque $|||\delta A||| \leq \dfrac{1}{2|||A^{-1}|||}$, la matrice $(\delta A)A^{-1}$ satisfait la condition :

$|||(\delta A)A^{-1}||| \leq \dfrac{1}{2}$.

La matrice $I_N + (\delta A)A^{-1}$ est donc inversible et son inverse vaut :

$\left(I_N + (\delta A)A^{-1}\right)^{-1} = \sum_{n=0}^{+\infty} \left(-(\delta A)A^{-1}\right)^n$.

En particulier, il vient :

$|||\left(I_N + (\delta A)A^{-1}\right)^{-1}||| \leq \sum_{n=0}^{+\infty} |||(\delta A)A^{-1}|||^n \leq 2$.

Par linéarité, nous savons par ailleurs que :

$(A + \delta A)\,\delta x = -(\delta A)x$,

De sorte que :

$\delta x = -A^{-1}\left(I_N + (\delta A)A^{-1}\right)^{-1}(\delta A)x$,

Ce qui implique :

$\|\delta x\| \leq |||A^{-1}||| \; |||\left(I_N + (\delta A)A^{-1}\right)^{-1}||| \; |||\delta A||| \, \|x\|$,

Soit :

$\dfrac{\|\delta x\|}{\|x\|} \leq 2\,\text{cond}(A) \dfrac{|||\delta A|||}{|||A|||}$

Corollaire : Soit $A \in GL_N(\mathbb{C})$ et $\delta A \in \mathcal{M}_N(\mathbb{C})$ telle que : $|||\delta A||| \leq \dfrac{1}{2|||A^{-1}|||}$. Pour $b \in \mathbb{C}^N \setminus \{0\}$ et $\delta b \in \mathbb{C}^N$, désignons par x et $x+\delta x$ les solutions des équations $Ax = b$, resp. $(A+\delta A)(x+\delta x) = b + \delta b$. Alors :

$$\frac{\|\delta x\|}{\|x\|} \leq 2\,\text{cond}(A)\left[\frac{|||\delta A|||}{|||A|||} + \frac{\|\delta b\|}{\|b\|}\right].$$

Preuve :

Sous les hypothèses du corollaire, la matrice $I_N + (\delta A)A^{-1}$ est inversible et son inverse satisfait l'inégalité :

$|||\left(I_N + (\delta A)A^{-1}\right)^{-1}||| \leq 2$.

En particulier, la matrice $A+\delta A$ est également inversible, de sorte que le vecteur δx est bien défini. Il vient alors:

$$(A+\delta A)(\delta x) = \delta b - \delta A(x),$$

De sorte que:

$$\|\delta x\| \leq |||A^{-1}||| \; |||(I_N + (\delta A)A^{-1})^{-1}||| \left[\|\delta b\| + |||\delta A||| \, \|x\| \right]$$

$$\leq 2 \, |||A^{-1}||| \left[\|\delta b\| + |||\delta A||| \, \|x\| \right].$$

Comme $b = A(x)$, nous savons que:

$$\|b\| \leq |||A||| \, \|x\|,$$

Ce qui suffit à montrer que:

$$\frac{\|\delta x\|}{\|x\|} \leq 2 \, \text{cond}(A) \left(\frac{\|\delta b\|}{\|b\|} + \frac{|||\delta A|||}{|||A|||} \right).$$

En conclusion, les erreurs relatives dans la résolution du système linéaire $A(x) = b$ sont contrôlées par le conditionnement de la matrice A. Afin de minimiser les erreurs d'arrondis, il est pertinent de réduire la valeur du conditionnement de la matrice considérée. Le préconditionnement du système linéaire à résoudre repose (en général) sur la multiplication par une matrice $D \in \mathcal{GL}_N(\mathbb{C})$ adhoc (par exemple, diagonale) de façon à se ramener à un système linéaire équivalent: $DA(x) = D(b)$ tel que: $\text{cond}(DA) < \text{cond}(A)$.

La détermination des matrices D optimales pour cette propriété porte le nom de problème d'équilibrage. Il s'agit d'un problème pratique important afin de contrôler les problèmes d'arrondis dans la résolution des systèmes linéaires.

3. Conditionnement d'un problème aux valeurs propres

Soit $N \in \mathbb{N}^*$. Notons $\mathcal{D}_N(\mathbb{C})$ l'ensemble des matrices diagonales de $\mathcal{M}_N(\mathbb{C})$, et considérons une norme $\|\ \|$ sur l'espace vectoriel $\mathbb{C}^N$ telle que la norme subordonnée $|||\ |||$ à cette norme sur $\mathcal{M}_N(\mathbb{C})$ satisfasse l'identité:

$$\forall D \in \mathcal{D}_N(\mathbb{C}), \quad |||D||| = \max_{1 \leq i \leq N} |D_{i,i}|.$$

Exemple : Les normes $|||\ |||_2$, $|||\ |||_1$ et $|||\ |||_\infty$ subordonnées aux normes $\|\ \|_2$, $\|\ \|_1$ et $\|\ \|_\infty$ satisfont cette propriété.

Définition : Soit $A \in \mathcal{M}_N(\mathbb{C})$ une matrice diagonalisable. Le conditionnement de la matrice A relativement au calcul de ses valeurs propres est le nombre strictement positif : $\nu(A) := \inf\{\mathrm{cond}(P), P \in GL_N(\mathbb{C}) \text{ t.q. } PAP^{-1} \in D_N(\mathbb{C})\}$.

Exemples : (i) $\forall D \in D_N(\mathbb{C})$, $\nu(D) = 1$.

(ii) Si $\|\ \|_2$ désigne la norme subordonnée à la norme hermitienne canonique $\|\ \|_2$ sur $\mathbb{C}^N$, alors :

$\forall N \in \mathcal{N}_N(\mathbb{C})$, $\nu_2(N) = 1$

(car : $\forall U \in U_N(\mathbb{C})$, $\mathrm{cond}_2(U) = 1$).

Le conditionnement d'une matrice par rapport au calcul de ses valeurs propres permet d'évaluer des erreurs quant à la détermination des valeurs propres.

Théorème de Bauer - Fike : Soit $A \in \mathcal{M}_N(\mathbb{C})$ une matrice diagonalisable et $\delta A \in \mathcal{M}_N(\mathbb{C})$, alors : $\sigma(A+\delta A) \subset \sigma(A) + \overline{B}_f(0, \nu(A)|||\delta A|||)$.

Preuve :

Soit $\sigma(A) = \{\lambda_1, \dots, \lambda_N\}$ et $\lambda \in \sigma(A+\delta A)$. Quel que soit le nombre ε strictement positif, considérons une matrice inversible $P \in GL_N(\mathbb{C})$ telle que $PAP^{-1} = D \in D_N(\mathbb{C})$ et :

$\mathrm{cond}(P) \leq \nu(A) + \varepsilon$.

Notons que, si $\lambda \in \sigma(A)$, alors, il n'y a rien à démontrer, et supposons donc que $\lambda \notin \sigma(A)$.

Dans ce cas, la matrice $D - \lambda I_N$ est inversible, de sorte que :

$P^{-1}(A+\delta A - \lambda I_N)P = D - \lambda I_N + P^{-1}(\delta A)P = (D-\lambda I_N)\left(I_N + (D-\lambda I_N)^{-1} P^{-1} \delta A P\right)$.

Comme la matrice $P^{-1}(A+\delta A - \lambda I_N)P$ n'est pas inversible, il est clair que :

$|||(D-\lambda I_N)^{-1} P^{-1}(\delta A) P||| \geq 1$,

De suite que :

$$1 \leq \| (Q - \lambda I_n)^{-1} \| \, \mathrm{cond}(P) \, \| \delta A \|$$

$$\leq \| (Q - \lambda I_n)^{-1} \| \, (r(A) + \varepsilon) \, \| \delta A \|.$$

Par hypothèse, il vient :

$$\| (Q - \lambda I_n)^{-1} \| = \frac{1}{\min_{1 \leq j \leq n} |\lambda_j - \lambda|},$$

Des suite que :

$$\lambda \in \sigma(A) + D_F\left(0, (r(A) + \varepsilon) \| \delta A \|\right).$$

Le théorème s'ensuit à la limite $\varepsilon \to 0$.

En conclusion, les ennuis dans la résolution d'un problème aux valeurs propres ne dépendent plus du conditionnement de la matrice diagonalisable considérée, mais de celui des matrices de passage associées à ce problème de diagonalisation. En particulier, les ennuis éventuels liés au problème aux valeurs propres d'une matrice normale sont peu importants : ces matrices sont très bien conditionnées pour le problème aux valeurs propres. Le détermination des valeurs singulières d'une matrice $A \in \mathcal{M}_n(\mathbb{K})$, soit des valeurs propres de la matrice $A^* A$, est donc peu entachée d'ennuis éventuels.

III Méthodes directes pour la résolution de systèmes linéaires

Au contraire des méthodes itératives que nous étudierons dans la suite de ce cours, les méthodes directes de résolution de systèmes linéaires fournissent (au moins au niveau théorique) la (ou les) solution(s) exacte(s) du système considéré. Les éventuels ennuis ne sont (en principe) liés qu'aux problèmes éventuels (et dépendent donc du conditionnement précédemment introduit).

À noter que ces méthodes ne passent pas par le calcul de l'inverse de la matrice considérée, qui réclame la résolution de N systèmes linéaires (pour chacun des vecteurs $(e_j)_{1 \leq j \leq n}$ de la base canonique de $\mathbb{K}^n$) pour une matrice carrée de taille N.

En particulier, certaines de ces méthodes s'étendent à des systèmes linéaires rectangulaires (même si dans ce cas la solution n'est plus unique).

Cependant, dans le cas où une même matrice $A \in \mathcal{M}_N(\mathbb{K})$ interviendrait dans la résolution de plusieurs systèmes linéaires, il peut être utile de déterminer son inverse, ou au moins une factorisation de cette matrice, qui faciliterait la résolution de ces systèmes.

1. Méthode de Gauss

Soit $N \in \mathbb{N}^*$. Pour $A \in \mathcal{GL}_N(\mathbb{K})$ et $b \in \mathbb{K}^N$, considérons la résolution du système linéaire :

$$A(x) = b.$$

La méthode de Gauss pour résoudre ce système repose sur l'observation que la résolution est simple lorsque la matrice A est triangulaire.

Par exemple, si la matrice $A \in \mathcal{T}_N^+(\mathbb{K})$ est triangulaire supérieure, alors, le système $A(x) = b$ se réduit aux N équations :

$$(\mathcal{L}) \begin{cases} a_{11} x_1 + \dots + a_{1N} x_N = b_1 \\ \ddots \quad \vdots \\ a_{NN} x_N = b_N \end{cases}$$

dans lequel les nombres $(a_{ij})_{1 \leq i \leq j \leq N}$ sont non nuls puisque la matrice A est inversible. Il est ainsi facile de calculer le nombre x_N, qui est égale à :

$$x_N = \frac{b_N}{a_{NN}},$$

puis d'en déduire le nombre :

$$x_{N-1} = \frac{1}{a_{N-1,N-1}} \left(b_{N-1} - a_{N-1,N} x_N \right),$$

et d'itérer ce procédé jusqu'à obtenir la valeur de :

$$x_1 = \frac{1}{a_{11}} \left[b_1 - a_{12} x_2 - \dots - a_{1N} x_N \right].$$

C'est la méthode de la remontée qui fournit la valeur de la solution x en $\frac{N(N-1)}{2}$ soustractions, $\frac{N(N-1)}{2}$ multiplications et N divisions, soit

de l'ordre de N^3 opérations élémentaires (ce qui est honnête sur le plan de la complexité algorithmique). Il s'agit donc de se ramener à la résolution d'un système triangulaire (supérieur).

Cette réduction repose sur la méthode du pivot de Gauss. Considérons la première colonne de la matrice A. Cette colonne n'est pas nulle puisque la matrice A est inversible. Il existe donc un entier $1 \leq i_0 \leq N$ tel que : $a_{i_0,1} \neq 0$, et choisissons ce nombre comme pivot. La multiplication par la matrice de transposition :

$$T(1, i_0) = \begin{pmatrix} 0 & & & & 1 & & \\ & 1 & & & & & \\ & & \ddots & & & & \\ & & & 0 & & 1 & \\ & & & & \ddots & & \\ & & & 1 & & 0 & \\ & & & & & & 1 \end{pmatrix} \begin{matrix} \\ \\ \\ \leftarrow i_0 \\ \\ \leftarrow i_1 \\ \\ \\ \end{matrix}$$

$$\begin{matrix} & & & & \uparrow & & \\ & & & & 1 & & \\ & & & & \text{colonne} & & \end{matrix}$$

permet de permuter la première ligne et la i_1-ième ligne, puis la multiplication de la matrice $A' = T(1, i_0)\, A$ par les matrices de transvection :

$$E(i, 1) = I_N - \frac{a'_{i,1}}{a'_{1,1}} E_{i,1},$$

où les matrices $(E_{i,1})_{1 \leq i \leq N}$ sont définies par les identités :

$$\left[E_{i,1}\right]_{k,l} = \begin{cases} 1 & \text{si } k = i \text{ et } l = 1, \\ 0 & \text{sinon,} \end{cases}$$

conduit à une matrice A_1 de la forme :

$$A_1 = \prod_{i=1}^{N} E(i, 1)\, T(1, i_0)\, A = \begin{pmatrix} a_{i_0,1} & * & \cdots & * \\ 0 & & & \\ \vdots & & A'_1 & \\ 0 & & & \end{pmatrix}.$$

Il s'agit ensuite d'itérer ce procédé sur la seconde colonne de A_1, plus précisément sur la première colonne de la matrice A'_1 définie par l'identification ci-dessus, puis sur chacune des N colonnes des matrices résultant de ces opérations. Il est ainsi possible de construire une matrice inversible $\Pi \in GL_N(\mathbb{K})$ telle que la matrice ΠA soit triangulaire supérieure. Ceci achève la phase dite d'élimination de la méthode de Gauss.

Parallèlement, il est possible de calculer le vecteur $\Pi(b)$ et il reste alors à résoudre le système triangulaire supérieur :

$\Pi A(x) = \Pi(b)$,
par la méthode de la remontée précédemment introduite.

Sur le plan théorique, la méthode de Gauss conduit à l'énoncé suivant.

Théorème: Soit $A \in \mathcal{M}_N(\mathbb{C})$. Il existe une matrice inversible $\Pi \in \mathcal{GL}_N(\mathbb{C})$ telle que la matrice $\Pi A \in \mathcal{L}_N^+(\mathbb{C})$ soit triangulaire supérieure.

Preuve:
Par la méthode de Gauss précédemment décrite, il est clair que le théorème est vrai lorsque la matrice A est inversible.
Lorsqu'elle n'est plus inversible, il est possible qu'une des colonnes à éliminer soit identiquement nulle. Dans ce cas, cette colonne est déjà sous forme triangulaire supérieure, et il ne reste plus qu'à traiter les autres colonnes.

Sur le plan pratique, le nombre d'opérations élémentaires requises pour l'application de la méthode de Gauss est de l'ordre de N^3 additions ou soustractions, et de N^3 multiplications ou divisions, ce qui demeure acceptable. Plus précisément, chaque étape de la procédure d'élimination nécessite de l'ordre de $(N-k)^2$ additions ou soustractions, et $(N-k)^2$ multiplications ou divisions, où l'entier k varie entre 1 et $N-1$. Au total, il est donc nécessaire de réaliser de l'ordre de N^3 additions ou soustractions, et N^3 multiplications ou divisions pour cette première étape d'élimination. Les autres étapes, sont le calcul du vecteur $\Pi(b)$, et la méthode de la remontée, ne requièrent que de l'ordre de N^2 additions ou soustractions, et N^2 multiplications ou divisions.

Un autre point pratique important consiste en le choix des pivots. Plus ceux-ci sont grands, moins les erreurs d'arrondis sont importantes. La stratégie du pivot partiel consiste à choisir le pivot à l'étape $k \in \{1, \dots, N-1\}$ comme l'élé-

ment $a_{i_0 k}$ de la $k^{ième}$ colonne tel que :

$$|a_{i_0 k}| = \max_{k \leq i \leq n} |a_{ik}|$$

La stratégie du pivot total consiste quant à elle à choisir le pivot à l'étape $1 \leq k \leq n-1$ comme l'élément $a_{i_0 j_0}$ tel que :

$$|a_{i_0 j_0}| = \max_{k \leq i, j \leq n} |a_{ij}|$$

Il notre que cette seconde stratégie nécessite de permuter, si nécessaire, des colonnes de la matrice A, ce qui revient à multiplier celle-ci par une matrice de transposition à droite. L'étape finale de remontée se trouve ainsi compliquée.

Par ailleurs, ces deux méthodes limitent certes les erreurs d'arrondis, mais entraînent un surcoût en opérations élémentaires.

Pour conclure, rappelons que la méthode de Gauss peut se prolonger en méthode de Gauss-Jordan afin d'obtenir in fine une matrice diagonale et qu'elle s'applique aussi dans le cas de matrices rectangulaires. Sur le plan théorique, elle permet ainsi de démontrer le résultat suivant.

Théorème : Soit $n \in \mathbb{N}^*$ et $A \in \mathcal{M}_{n,n}(\mathbb{K})$. Il existe deux matrices inversibles $L \in \mathcal{GL}_n(\mathbb{K})$ et $Q \in \mathcal{GL}_n(\mathbb{K})$ et un entier $1 \leq r \leq \min\{n, N\}$ tels que :

$$L A Q = \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix}.$$

Notons enfin que la méthode de Gauss fournit aussi l'inverse de la matrice $A \in \mathcal{M}_n(\mathbb{K})$ considérée, ce qui se révèle utile lorsqu'il s'agit de résoudre le système linéaire $A(x) = b$ pour un grand nombre de valeurs du vecteur $b \in \mathbb{K}^n$.

2. Factorisation LU d'une matrice

L'application de la méthode de Gauss conduit au résultat de factorisation

suivant.

Théorème : Soit $A \in \mathcal{GL}_n(\mathbb{C})$. Pour $1 \leq k \leq n$, introduisons les matrices $A_k \in \mathcal{M}_k(\mathbb{C})$ définies par : $\forall\, 1 \leq i, j \leq k$, $(A_k)_{i,j} = a_{i,j}$. Si les matrices $(A_k)_{1 \leq k \leq n}$ sont inversibles, alors, il existe une matrice triangulaire inférieure $L \in \mathcal{L}_n^-(\mathbb{C})$, avec $\forall\, 1 \leq i \leq n$, $l_{ii} = 1$, et une matrice triangulaire supérieure $U \in \mathcal{L}_n^+(\mathbb{C})$ telles que : $A = LU$. De plus, cette factorisation est unique.

Preuve :

L'existence des matrices L et U résulte de l'application de la méthode de Gauss. Au rang $k = 1$, le coefficient a_{11} est non nul, de sorte que l'on peut le choisir comme pivot pour la première colonne. Il existe ainsi une matrice $E_1 = \begin{pmatrix} 1 & & & 0 \\ -\frac{a_{21}}{a_{11}} & \ddots & & \\ 0 & & \ddots & \\ -\frac{a_{n1}}{a_{11}} & & & 1 \end{pmatrix}$ telle que : $E_1 A = \begin{pmatrix} a_{11} & * \\ 0 & A'_1 \end{pmatrix}$. Supposons alors avoir construit des matrices $(E_j)_{1 \leq j \leq k-1}$ (la j-ième colonne) de la forme :

$\forall\, 1 \leq j \leq k-1$, $E_j = \begin{pmatrix} 1 & & & & 0 \\ & \ddots & & & \\ & & 1 & & \\ & & l_{j+1,j} & \ddots & \\ 0 & & l_{n,j} & & 1 \end{pmatrix}$,

telles que : $E_{k-1} \cdots E_1 A = \begin{pmatrix} a_{1,1} & & * \\ & \ddots & \\ 0 & & A'_{k-1} \end{pmatrix}$. Les règles usuelles de multiplication par blocs permettent d'écrire cette identité sous la forme :

$E_{k-1} \cdots E_1 A = \begin{pmatrix} E'_k & 0 \\ * & I_{n-k} \end{pmatrix} \begin{pmatrix} A_k & * \\ * & * \end{pmatrix} = \begin{pmatrix} a_{1,1} & & * & * \\ & \ddots & & \\ 0 & & a_{k,k} & * \end{pmatrix}$, avec $E'_k = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ * & & 1 \end{pmatrix}$.

De sorte que :

$E'_k A_k = \begin{pmatrix} a_{1,1} & & * \\ & \ddots & \\ 0 & & a_{k,k} \end{pmatrix}$.

Comme la matrice A_k est inversible, il est clair que le nombre $a_{k,k}$ est non nul. Il est donc possible d'introduire une matrice $E_k = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & l_{k+1,k} & 1 \end{pmatrix}$ telle que :

$E_k \cdots E_1 A = \begin{pmatrix} a_{1,1} & & & * \\ & \ddots & & \\ & & a_{k,k} & \\ 0 & & & A'_{k+1} \end{pmatrix}$.

En définitive, il est permis de trouver une matrice triangulaire inférieure $E = E_{N-1} \dots E_1 = \begin{pmatrix} 1 & 0 \\ * & 1 \end{pmatrix}$ telle que :

$E A \in \mathcal{T}_N^+(\mathbb{K})$.

Comme l'inverse $L = E^{-1}$ de la matrice L est de même forme, il existe une matrice triangulaire supérieure $U \in \mathcal{T}_N^+(\mathbb{K})$ telle que :

$A = LU$.

De plus, s'il existe deux autres matrices $L' \in \mathcal{T}_N^-(\mathbb{K})$, avec $\forall 1 \leq i \leq N$, $l'_{ii} = 1$, et $U' \in \mathcal{T}_N^+(\mathbb{K})$ telles que :

$A = L'U'$,

alors :

$LU = L'U' \Rightarrow (L'^{-1})L = U'(U^{-1})$.

Comme $(L'^{-1})L \in \mathcal{T}_N^-(\mathbb{K})$ et $U'(U^{-1}) \in \mathcal{T}_N^+(\mathbb{K})$, les matrices $(L'^{-1})L$ et $U'(U^{-1})$ sont diagonales, et dans ce cas :

$(L'^{-1})L = I_N \Rightarrow L = L' \Rightarrow U' = U$,

D'où l'unicité de la factorisation LU.

En pratique, l'application de la méthode de Gauss fournit de manière immédiate la valeur de la matrice triangulaire supérieure $U \in \mathcal{T}_N^+(\mathbb{K})$, mais également celle de la matrice triangulaire inférieure $L \in \mathcal{T}_N^-(\mathbb{K})$.

Si nous revenons à la preuve précédente et notons :

$E_k = \begin{pmatrix} 1 & & & 0 \\ & \ddots & & \\ & & 1 & \\ 0 & & -\ell_{k+1,k} & \ddots \\ & & & & 1 \end{pmatrix}$,

pour chaque entier $k \in \{1, \dots, N-1\}$, alors, l'inverse de la matrice E_k est égal à :

$E_k^{-1} = \begin{pmatrix} 1 & & & 0 \\ & \ddots & & \\ & & 1 & \\ 0 & & \ell_{k+1,k} & \ddots \\ & & & & 1 \end{pmatrix}$,

Et le produit de ces inverses, soit la matrice L, est :

$L = E_1^{-1} \dots E_{N-1}^{-1} = \begin{pmatrix} 1 & & & 0 \\ \ell_{2,1} & \ddots & & \\ \vdots & & 1 & \\ \ell_{N,1} & \dots & \ell_{N,N-1} & 1 \end{pmatrix}$.

L'intérêt de la factorisation LU réside dans la possibilité de résoudre

rapidement de manière répétée des systèmes linéaires de la forme $A(x) = b$ pour différentes valeurs du vecteur $b \in \mathbb{K}^N$. Ces systèmes sont de fait équivalents aux deux systèmes triangulaires suivants :

$$L(y) = b \quad \text{et} \quad U(x) = y,$$

qui peuvent se résoudre par la méthode de la remontée. Le coût en opérations élémentaires est de l'ordre de N^2 au lieu d'un coût de l'ordre de N^3 par la méthode de Gauss, et ce pour chaque résolution du système linéaire $A(x) = b$.

Notons que ce coût est encore réduit lorsque la matrice A est tridiagonale. Ceci résulte du théorème suivant.

Théorème : Soit $A = \begin{pmatrix} b_1 & c_1 & 0 \\ a_1 & \ddots & \ddots \\ & \ddots & c_{N-1} \\ 0 & a_{N-1} & b_N \end{pmatrix} \in \mathcal{M}_N(\mathbb{K})$ une matrice tridiagonale. Pour $1 \leq k \leq N$, introduisons les matrices $A_k \in \mathcal{M}_k(\mathbb{K})$ définies par : $\forall 1 \leq i, j \leq k$, $(A_k)_{i,j} = a_{i,j}$, et notons : $\delta_k = \det(A_k)$.

(i) Les nombres $(\delta_k)_{1 \leq k \leq N}$ sont donnés par les relations de récurrence

$$\delta_1 = b_1 \text{ et } \forall\, 2 \leq k \leq N,\ \delta_k = b_k \delta_{k-1} - a_k c_{k-1} \delta_{k-2},$$

où nous avons noté $\delta_0 = 1$.

(ii) Si les nombres $(\delta_k)_{1 \leq k \leq N}$ sont tous non nuls, la factorisation LU de la matrice A est donnée par les matrices :

$$L = \begin{pmatrix} 1 & & 0 \\ a_2 \frac{\delta_1}{\delta_2} & \ddots & \\ & \ddots & 1 \\ 0 & a_N \frac{\delta_{N-1}}{\delta_{N-2}} & 1 \end{pmatrix} \quad \text{et} \quad U = \begin{pmatrix} \frac{\delta_1}{\delta_0} & c_1 & 0 \\ & \ddots & c_{N-1} \\ 0 & & \frac{\delta_N}{\delta_{N-1}} \end{pmatrix}.$$

Preuve :

(i) La formule de récurrence résulte du développement par rapport à la dernière ligne du déterminant δ_k.

(ii) Si les nombres $(\delta_k)_{1 \leq k \leq N}$ sont tous non nuls, alors, le théorème précédent garantit l'existence et l'unicité de la factorisation LU de la matrice A. Un calcul direct permet alors d'établir que les matrices L et U ci-dessus

satisfont :

$(LU)_{1,1} = \frac{\delta_1}{\delta_0} = b_1$ et $\forall\, 2 \le k \le N$, $(LU)_{k,k} = \frac{\delta_k + a_k c_{k-1} \delta_{k-2}}{\delta_{k-1}} = b_k$,

et, de même,

$\forall\, 1 \le k \le N-1$, $(LU)_{k+1,k} = a_{k+1}$ et $(LU)_{k,k+1} = c_k$,

De sorte que nous avons bien l'égalité :

$LU = A$

Les formules des matrices L et U du théorème permettent alors de calculer les solutions des systèmes triangulaires :

$L(y) = b$ et $U(x) = y$,

en un nombre d'opérations élémentaires de l'ordre de N, ce qui est un gain considérable par rapport à la méthode de Gauss.

3. Factorisation de Cholesky d'une matrice symétrique définie positive

Comme dans le cas tridiagonal, la factorisation LU d'une matrice A est particulièrement bien adaptée au cas où cette matrice $A \in \mathcal{S}_N^{++}(\mathbb{R})$ est symétrique définie positive. La factorisation de Cholesky qui repose sur cette remarque se présente sous la forme suivante.

Théorème : Soit $A \in \mathcal{S}_N^{++}(\mathbb{R})$. Il existe une unique matrice triangulaire inférieure $B \in \mathcal{T}_N^-(\mathbb{R})$ telle que :

(i) $\forall\, 1 \le i \le N$, $B_{ii} > 0$;

(ii) $A = B\, {}^tB$.

Preuve :

Pour $1 \le k \le N$, introduisons les matrices $A_k \in \mathcal{H}_k(\mathbb{R})$ définies par :

$\forall\, 1 \le i, j \le k$, $(A_k)_{i,j} = A_{i,j}$.

Si q désigne la forme quadratique associée à la matrice A, alors, les matrices A_k sont les matrices des formes quadratiques q_k restrictions de q aux

sous-espaces $\mathrm{Vect}(e_1, \dots, e_k)$, où $\mathcal{B} = (e_1, \dots, e_N)$ désigne la base canonique de l'espace vectoriel $\mathbb{R}^N$. En particulier, les matrices A_k sont symétriques définies positives, de sorte qu'elles sont inversibles. Il est donc permis d'introduire la décomposition $(L, U) \in \mathcal{T}_N^-(\mathbb{R}) \times \mathcal{T}_N^+(\mathbb{R})$ de la matrice A, avec $\forall\, 1 \leq i \leq N$, $l_{ii} = 1$.

Notons alors que pour tout entier $1 \leq k \leq N$:

$$A = LU = \begin{pmatrix} 1 & 0 & \\ * & \ddots & 0 \\ * & * & 1 \end{pmatrix} \begin{pmatrix} u_{11} & * & * \\ 0 & \ddots & * \\ 0 & 0 & u_{kk} \end{pmatrix},$$

De sorte que:

$$\det(A_k) = \prod_{i=1}^{k} u_{ii}.$$

Comme tous ces déterminants sont strictement positifs, tous les coefficients $(u_{ii})_{1 \leq i \leq N}$ sont également strictement positifs. Il est donc permis d'introduire la matrice diagonale $D \in \mathcal{D}_N(\mathbb{R})$ définie par:

$$\forall\, 1 \leq i \leq N, \quad D_{ii} = \sqrt{u_{ii}} > 0.$$

Notons alors:

$$B = LD \quad \text{et} \quad C = D^{-1}U.$$

Il vient:

$$A = BC,$$

De sorte que par symétrie de la matrice A:

$${}^tC\,{}^tB = A \Rightarrow C({}^tD)^{-1} = B^{-1}({}^tC).$$

Remarquons alors que ces deux matrices sont de la forme:

$$C({}^tB)^{-1} = \begin{pmatrix} 1 & * \\ & \ddots \\ 0 & 1 \end{pmatrix} \quad \text{et} \quad B^{-1}({}^tC) = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ * & & 1 \end{pmatrix}.$$

Il s'ensuit que:

$$C({}^tB)^{-1} = B^{-1}({}^tC) = I_N \Rightarrow C = {}^tB.$$

Nous avons donc établi l'existence d'une matrice $B \in \mathcal{T}_N^-(\mathbb{R})$, telle que:

$$\forall\, 1 \leq i \leq N, \quad B_{ii} > 0.$$

Pour terminer, notons que si une matrice B satisfait deux conclusions du théorème,

alors, la matrice diagonale D définie par :

$\forall\, 1 \leq i \leq N,\ D_{ii} = B_{ii} > 0,$

permet d'établir la factorisation suivante de la matrice A,

$A = (BD^{-1})(D^t B),$

qui n'est autre que la factorisation LU de A. Nous avons donc :

$L = BD^{-1}$ et $U = D^t B.$

En particulier, il vient :

$\forall\, 1 \leq j \leq N,\ U_{jj} = B_{jj}^2 \Rightarrow B_{jj} = \sqrt{U_{jj}},$

Puis :

$B = LD,$

Ce qui établit l'unicité de la matrice B.

En pratique, la factorisation de Cholesky d'une matrice $A = (a_{ij})_{1 \leq i,j \leq N} \in \mathcal{M}_N(\mathbb{R})$ symétrique définie positive se calcule par une méthode de coefficients indéterminés. L'identité $A = B\,{}^tB$ s'écrit sous la forme :

$\forall\, 1 \leq i, j \leq N,\ A_{ij} = \sum_{k=1}^{N} B_{ik} B_{jk} = \sum_{k=1}^{\min(i,j)} B_{ik} B_{jk}.$

Comme la matrice A est symétrique, il suffit d'étudier ces équations pour $i \leq j$. Le calcul des coefficients B_{ij} se fait alors par récurrence sur l'entier $i \in \{1, \dots, N\}$. Pour $i = 1$, il vient :

$\forall\, 1 \leq j \leq N,\ A_{1j} = B_{11} B_{j1} \Rightarrow \begin{cases} B_{11} = \sqrt{A_{11}}, \\ \forall\, 1 \leq j \leq N,\ B_{j1} = \dfrac{A_{1j}}{\sqrt{B_{11}}}, \end{cases}$

et il est alors possible de résoudre les équations précédentes pour $i = 2$, puis 3, et ainsi de suite.

Une fois la factorisation de Cholesky déterminée, la méthode de Cholesky pour la résolution du système linéaire $A(x) = b$ repose sur la résolution des deux systèmes triangulaires :

$B(y) = b$ et ${}^tB(x) = y,$

par la méthode de la remontée comme dans le cas de la factorisation LU. Notons pour conclure que cette méthode nécessite de l'ordre de la moitié des opérations élémentaires nécessaires à la méthode de Gauss,

suit néanmoins de l'ordre de N^3 opérations élémentaires. Ceci repose sur l'observation que, dans le cas où la matrice A est symétrique, la connaissance (d'un peu plus) de la moitié de ses coefficients suffit à la déterminer entièrement.

Pour terminer ce paragraphe sur les méthodes directes de résolution des systèmes linéaires, rappelons qu'il existe d'autres méthodes (comme la méthode QR) qui permettent aussi de résoudre de manière exacte de tels systèmes. Nous renvoyons à l'ouvrage "Introduction à l'analyse numérique matricielle et à l'optimisation" de PG. Ciarlet pour de plus amples détails.

II Méthodes itératives pour la résolution de systèmes linéaires

Au contraire des méthodes directes, les méthodes itératives ne cherchent pas à déterminer (au moins au niveau théorique) la (ou les) solution(s) exacte(s) du système à résoudre. Il s'agit au contraire de construire une suite de solutions approchées dont la limite est (l'une des) la solution(s) du système analysé.

Le principe de base pour la construction de cette suite est le théorème du point fixe comme nous allons maintenant le voir. Deux questions se posent quant à l'usage de ce théorème : d'une part, la preuve rigoureuse de la convergence de la suite de solutions approchées vers (une) la solution exacte, d'autre part l'analyse de la vitesse de convergence de cette suite vers sa (possible) limite. On note ici aussi la nécessité du test complet en pratique des critères d'arrêt (qui dépendent du conditionnement précédemment introduit).

1. Principe général des méthodes itératives

Soit $N \in \mathbb{N}^*$. Considérons une matrice inversible $A \in \mathcal{M}_N(\mathbb{R})$ et un vecteur $b \in \mathbb{R}^N$ et intéressons-nous à la résolution du système $Ax = b$. Le principe des méthodes itératives consiste à déterminer une matrice $B \in \mathcal{M}_N(\mathbb{R})$ telle que la matrice

$I_n - B$ soit inversible, et un vecteur $c \in \mathbb{K}^n$, tels que le système $A(x) = b$ soit équivalent au système linéaire :

$x = B(x) + c$.

La solution $x \in \mathbb{K}^n$ est alors l'unique point fixe de la fonctionnelle E définie par :

$\forall y \in \mathbb{K}^n$, $E(y) = B(y) + c$.

Il est alors possible de construire une suite de solutions approchées $(x_m)_{m \in \mathbb{N}}$ par la méthode du point fixe de Picard. Etant donné un vecteur $x_0 \in \mathbb{K}^n$, il s'agit de définir la suite $(x_m)_{m \in \mathbb{N}}$ par la formule de récurrence :

$\forall m \in \mathbb{N}$, $x_{m+1} = E(x_m) = B(x_m) + c$.

Si la suite $(x_m)_{m \in \mathbb{N}}$ converge vers un vecteur $x \in \mathbb{K}^n$, alors, ce vecteur x sera l'unique solution du système $x = B(x) + c$, soit du système $A(x) = b$.

Nous pouvons établir la convergence de cette méthode sous les hypothèses suivantes.

Théorème : Soit $B \in \mathcal{M}_n(\mathbb{K})$ telle que la matrice $I_n - B$ soit inversible et $c \in \mathbb{K}^n$ quel que soit le vecteur $x_0 \in \mathbb{K}^n$, considérons la suite $(x_m)_{m \in \mathbb{N}}$ définie par la formule de récurrence :

$\forall m \in \mathbb{N}$, $x_{m+1} = B(x_m) + c$.

(i) Les suites $(x_m)_{m \in \mathbb{N}}$ sont toutes convergentes ssi le rayon spectral $\rho(B)$ de la matrice B satisfait l'inégalité :

$\rho(B) < 1$.

(ii) Dans ce cas, la limite x des suites $(x_m)_{m \in \mathbb{N}}$ est l'unique solution de l'équation :

$x = B(x) + c$.

Preuve :

Comme la matrice $I_n - B$ est inversible, il existe un unique vecteur $x \in \mathbb{K}^n$ tel que :

$x = B(x) + c$.

Il vient alors :

$\forall n \in \mathbb{N}$, $x_{n+1} - x = B(x_n - x)$,

De sorte que, par récurrence sur l'entier $n \in \mathbb{N}$,

$\forall n \in \mathbb{N}$, $x_n - x = B^n(x_0 - x)$.

Si $\rho(B) < 1$, alors,

$B^n \xrightarrow[n \to +\infty]{} 0$ dans $\mathcal{M}_N(\mathbb{C})$,

Et, dans ce cas,

$x_n \xrightarrow[n \to +\infty]{} x$ dans $\mathbb{C}^N$.

Réciproquement, si la suite $(x_n)_{n \in \mathbb{N}}$ est convergente (quel que soit le choix du vecteur x_0), alors, sa limite y satisfait l'équation:

$y = B(y) + c$,

De sorte que $y = x$. Il s'ensuit que:

$\forall x_0 \in \mathbb{C}^N$, $B^n(x_0 - x) \xrightarrow[n \to +\infty]{} 0$ dans $\mathbb{C}^N$.

Ceci implique que:

$\max_{1 \leq i,j \leq N} \left| (B^n)_{i,j} \right| \xrightarrow[n \to +\infty]{} 0$.

Comme cette quantité est une norme sur l'espace vectoriel $\mathcal{M}_N(\mathbb{C})$, il vient:

$B^n \xrightarrow[n \to +\infty]{} 0$ dans $\mathcal{M}_N(\mathbb{C})$,

D'où il découle que $\rho(B) < 1$.

La valeur du rayon spectral $\rho(B)$ de la matrice B détermine donc la convergence de la méthode itérative correspondante. Cette valeur contrôle également la vitesse de convergence comme le souligne l'énoncé suivant.

<u>Théorème:</u> Soit $B \in \mathcal{M}_N(\mathbb{C})$ telle que la matrice $I_N - B$ soit inversible et $c \in \mathbb{C}^N$. Considérons l'unique solution $x \in \mathbb{C}^N$ de l'équation $x = B(x) + c$, et quel que soit le vecteur $x_0 \in \mathbb{C}^N$, la suite $(x_n)_{n \in \mathbb{N}}$ associée par la formule de récurrence:

$\forall n \in \mathbb{N}$, $x_{n+1} = B(x_n) + c$.

Quelle que soit la norme $\| \ \|$ sur $\mathbb{C}^N$, nous avons la convergence:

$\sup \left\{ \| x_n - x \|^{\frac{1}{n}}, x_0 \in \mathbb{C}^N \text{ t.q. } \| x_0 - x \| = 1 \right\} \xrightarrow[n \to +\infty]{} \rho(B)$.

<u>Preuve:</u>

Soit $|||\ |||$ la norme subordonnée à la norme $\|\ \|$ sur $\mathbb{C}^n$ espace vectoriel $\mathcal{M}_n(\mathbb{C})$. Rappelons que la formule du rayon spectral garantit que:

$$|||B^n|||^{\frac{1}{n}} \xrightarrow[n\to+\infty]{} \rho(B).$$

Comme

$$\sup\left\{\|x_n - x\|,\ x_0 \in \mathbb{C}^n \text{ t.q. } \|x_0 - x\| = 1\right\} = \sup\left\{\|B^n(x_0 - x)\|,\ x_0 \in \mathbb{C}^n \text{ t.q. } \|x_0 - x\| = 1\right\}$$

$$= |||B^n|||,$$

il vient:

$$\sup\left\{\|x_n - x\|^{\frac{1}{n}};\ x_0 \in \mathbb{C}^n \text{ t.q. } \|x_0 - x\| = 1\right\} \xrightarrow[n\to+\infty]{} \rho(B).$$

Ce second théorème permet en particulier de comparer deux méthodes itératives pour la résolution d'un même système linéaire $A(x) = b$. Plus le rayon de convergence $\rho(B)$ de la matrice B sera faible, et meilleure sera la méthode. Et notez que dans tous les cas la convergence sera géométrique (lorsque $\rho(B) < 1$), ce qui est déjà très rapide.

2. Méthodes de Jacobi et de Gauss-Seidel

Soit $N \in \mathbb{N}^*$. Considérons une matrice inversible $A \in \mathcal{GL}_N(\mathbb{C})$ et un vecteur $b \in \mathbb{C}^N$, et intéressons-nous à la résolution du système $A(x) = b$.

Définition: Soit $A \in \mathcal{GL}_N(\mathbb{C})$. Une décomposition régulière de la matrice A est la donnée d'un couple de matrices $(M, N) \in \mathcal{GL}_N(\mathbb{C}) \times \mathcal{M}_N(\mathbb{C})$ tel que:

(i) $A = M - N$,

(ii) La matrice $M \in \mathcal{GL}_N(\mathbb{C})$ est facile à inverser.

La méthode itérative basée sur cette décomposition consiste alors à choisir un vecteur $x_0 \in \mathbb{C}^N$ et à considérer la suite $(x_n)_{n\in\mathbb{N}}$ définie par la formule de récurrence:

$$\forall n \in \mathbb{N},\ x_{n+1} = M^{-1}N(x_n) + M^{-1}(b).$$

Exemples : (i) Méthode de Jacobi :

Soit $A \in \mathcal{GL}_N(\mathbb{C})$ et $D = \begin{pmatrix} a_{11} & & 0 \\ & \ddots & \\ 0 & & a_{NN} \end{pmatrix} \in \mathcal{D}_N(\mathbb{C})$. Si la matrice D est inversible (c'est-à-dire si tous les coefficients diagonaux de A sont non nuls), alors, la méthode de Jacobi est la méthode itérative associée à la décomposition régulière $A = M - N$, où $M = D$ et $N = -A + D$. Cette méthode consiste donc à choisir un vecteur $x_0 \in \mathbb{C}^N$ et à considérer la suite $(x_n)_{n \in \mathbb{N}}$ définie par la formule de récurrence :

$$\forall n \in \mathbb{N}, \quad x_{n+1} = \left(I_N - D^{-1}A\right) x_n + D^{-1}(b).$$

(ii) Méthode de Gauss-Seidel :

Soit $A \in \mathcal{GL}_N(\mathbb{C})$ et $T = \begin{pmatrix} a_{11} & & 0 \\ a_{ij} & \ddots & \\ & & a_{NN} \end{pmatrix} \in \mathcal{T}_N^-(\mathbb{C})$. Si la matrice T est inversible (c'est-à-dire si tous les coefficients diagonaux de A sont non nuls), alors, la méthode de Gauss-Seidel est la méthode de itérative associée à la décomposition régulière $A = M - N$, où $M = T$ et $N = -A + T$. Cette méthode consiste donc à choisir un vecteur $x_0 \in \mathbb{C}^N$ et à considérer la suite $(x_n)_{n \in \mathbb{N}}$ définie par la formule de récurrence :

$$\forall n \in \mathbb{N}, \quad x_{n+1} = \left(I_N - T^{-1}A\right) x_n + T^{-1}(b).$$

En pratique, lorsqu'elle est définie, la méthode de Jacobi consiste donc à construire la suite $(x_n)_{n \in \mathbb{N}}$ de la façon suivante :

$$\forall n \in \mathbb{N}, \ \forall 1 \leq i \leq N, \quad (x_{n+1})_i = \frac{1}{a_{ii}} \left[-\sum_{j=1}^{N} a_{ij} (x_n)_j + b_i \right].$$

En particulier, l'inversion de la matrice D est immédiate.

Il en va de même pour la méthode de Gauss-Seidel qui consiste à construire la suite $(x_n)_{n \in \mathbb{N}}$ de façon à ce que :

$$\forall n \in \mathbb{N}, \ \forall 1 \leq i \leq N, \quad (x_{n+1})_i = \frac{1}{a_{ii}} \left[-\sum_{j=1}^{i-1} a_{ij} (x_{n+1})_j - \sum_{j=i+1}^{N} a_{ij} x(x_n)_j + b_i \right]$$

À noter que la complexité de cette seconde méthode est identique à celle de la

méthode de Jacobi, et qu'elle mobilise de plus moins de places en mémoire.

En ce qui concerne la convergence de ces méthodes, nous avons le résultat suivant.

Théorème: Soit $A \in \mathcal{GL}_n(\mathbb{C})$. Considérons une décomposition régulière $(M,N) \in \mathcal{GL}_n(\mathbb{C}) \times \mathcal{M}_n(\mathbb{C})$ de la matrice A et la méthode itérative associée, qui repose sur la formule de récurrence:

$\forall n \in \mathbb{N}$, $x_{n+1} = M^{-1}N(x_n) + M^{-1}(b)$.

La suite $(x_n)_{n\in\mathbb{N}}$ est alors convergente quel que soit le choix du vecteur $x_0 \in \mathbb{C}^n$ ssi: $\rho(M^{-1}N) < 1$.

Dans ce cas, la limite x de la suite $(x_n)_{n\in\mathbb{N}}$ est l'unique solution de l'équation: $A(x) = b$.

Preuve:

Comme

$I_n - M^{-1}N = M^{-1}A \in \mathcal{GL}_n(\mathbb{C})$,

l'équivalence découle des résultats précédents, et si x désigne la limite de la suite $(x_n)_{n\in\mathbb{N}}$, alors:

$x = M^{-1}N(x) + M^{-1}(b) \Rightarrow A(x) = b$.

Cet énoncé permet d'établir que la méthode de Jacobi est bien définie et convergente dans le cas des matrices à diagonales dominantes.

Définition: Soit $A \in \mathcal{M}_n(\mathbb{C})$. La matrice A est à diagonale dominante ssi:

$$\forall 1 \leq i \leq n, \quad |a_{ii}| > \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}|.$$

Exemple: La matrice identité est à diagonale dominante.

Lemme: Soit $A \in \mathcal{M}_n(\mathbb{C})$. Si la matrice A est à diagonale dominante, alors elle est inversible.

Preuve :

Soit $x \in \mathbb{K}^N$ tel que : $A(x) = 0$, et $i_0 \in \{1, \dots, N\}$ tel que :

$$|x_{i_0}| = \max_{1 \leq i \leq N} |x_i|.$$

Comme $A(x) = 0$, il vient :

$$|a_{i_0 i_0}| \, |x_{i_0}| = \left| - \sum_{\substack{j=1 \\ j \neq i_0}}^{N} a_{i_0 j} x_j \right| \leq \sum_{\substack{j=1 \\ j \neq i_0}}^{N} |a_{i_0 j}| \, |x_{i_0}|$$

$$\Rightarrow x_{i_0} = 0$$

De sorte que le vecteur x est nul, et la matrice A, inversible.

Théorème : Soit $A \in \mathcal{GL}_N(\mathbb{K})$, une matrice à diagonale dominante. La méthode de Jacobi est bien définie et elle est convergente. Quel que soit le choix du vecteur $x_0 \in \mathbb{K}^N$, la suite $(x_n)_{n \in \mathbb{N}}$ définie par la formule de récurrence :

$$\forall n \in \mathbb{N}, \quad x_{n+1} = \left(I_N - D^{-1} A\right)(x_n) + b,$$

où $D = \begin{pmatrix} a_{11} & & 0 \\ & \ddots & \\ 0 & & a_{NN} \end{pmatrix}$ converge vers l'unique solution x de l'équation : $A(x) = b$.

Preuve :

Avec les notations précédentes, les matrices Π et N de la décomposition régulière de A sont égales à :

$$\Pi = D \quad \text{et} \quad N = D - A$$

$$\Rightarrow \Pi^{-1} N = I_N - D^{-1} A = \begin{pmatrix} 0 & & -\frac{a_{ij}}{a_{ii}} \\ & \ddots & \\ -\frac{a_{ij}}{a_{ii}} & & 0 \end{pmatrix}$$

Notons en particulier que les coefficients diagonaux $(a_{ii})_{1 \leq i \leq N}$ sont tous non nuls puisque la matrice A est à diagonale dominante. La méthode de Jacobi est donc bien définie. De plus, si $\lambda \in \mathbb{K}$ vérifie $|\lambda| \geq 1$, alors, la matrice $\lambda I_N - \Pi^{-1} N$ est à diagonale dominante, donc inversible. Aussi le nombre λ n'est-il pas une valeur propre de la matrice $\Pi^{-1} N$. En conclusion,

$$\rho(\Pi^{-1} N) < 1,$$

Et, par le théorème précédent, la méthode de Jacobi est convergente.

Il en va de même pour la méthode de Gauss - Seidel.

Théorème : Soit $A \in \mathcal{GL}_N(\mathbb{C})$, une matrice à diagonale dominante. La méthode de Gauss-Seidel est bien définie et elle est convergente. Quel que soit le choix du vecteur $x_0 \in \mathbb{C}^N$, la suite $(x_n)_{n \in \mathbb{N}}$ définie par la formule de récurrence :

$\forall n \in \mathbb{N}$, $x_{n+1} = (I_N - T^{-1}A)(x_n) + T^{-1}(b)$,

où $T = \begin{pmatrix} a_{11} & & 0 \\ & \ddots & \\ a_{ij} & & a_{NN} \end{pmatrix} \in \mathcal{T}_N^-(\mathbb{C})$, converge vers l'unique solution $x \in \mathbb{C}^N$ de l'équation : $A(x) = b$.

Preuve :

La méthode de Gauss - Seidel est bien définie puisque les coefficients diagonaux $(a_{ii})_{1 \leq i \leq N}$ sont tous non nuls. Afin de montrer sa convergence, il s'agit de vérifier que :

$\rho\left(I_N - T^{-1}A\right) < 1$.

Soit donc $\lambda \in \sigma\left(I_N - T^{-1}A\right)$ et $x \in \mathbb{C}^N \setminus \{0\}$ tels que :

$x - T^{-1}A(x) = \lambda x$.

Notons $1 \leq i_0 \leq N$ tel que :

$|x_{i_0}| = \max_{1 \leq i \leq N} |x_i| \neq 0$.

Il vient alors :

$\lambda T(x) = (T - A)(x) \Rightarrow \lambda \sum_{j=1}^{i_0} a_{i_0 j} x_j = -\sum_{j=i_0+1}^{N} a_{i_0 j} x_j$

De sorte que :

$|\lambda| |a_{i_0 i_0}| |x_{i_0}| \leq \left(|\lambda| \sum_{j=1}^{i_0} |a_{i_0 j}| + \sum_{j=i_0+1}^{N} |a_{i_0 j}|\right) |x_{i_0}|$

$\Rightarrow |\lambda| |a_{i_0 i_0}| < \max\{|\lambda|, 1\} |a_{i_0 i_0}|$

De sorte que $|\lambda| < 1$ et $\rho\left(I_N - T^{-1}A\right) < 1$.

La méthode de Gauss - Seidel est également bien définie et convergente lorsque la matrice A est hermitienne définie positive.

Théorème de Ostrowski - Reich : Soit $A \in \mathcal{H}_N^{++}(\mathbb{C})$. La méthode de Gauss - Seidel est bien définie et elle est convergente. Quel que soit le choix du vecteur

$x_0 \in \mathbb{K}^N$, la suite $(x_n)_{n\in\mathbb{N}}$ définie par la formule de récurrence :

$\forall n \in \mathbb{N}$, $x_{n+1} = (I_N - T^{-1}A)(x_n) + T^{-1}(b)$,

où $T = \begin{pmatrix} a_{11} & & 0 \\ & \ddots & \\ 0 & & a_{NN} \end{pmatrix} \in \mathcal{T}_N^-(\mathbb{K})$, converge vers l'unique solution $x \in \mathbb{K}^N$ de l'équation : $A(x) = b$.

Preuve :

Comme la matrice A est hermitienne définie positive, ses coefficients diagonaux $(a_{ii})_{1\le i\le N}$ sont tous non nuls de sorte que la méthode de Gauss - Seidel est bien définie.

Considérons alors la norme $\|\ \|_A$ définie sur $\mathbb{K}^N$ par l'expression :

$\forall x \in \mathbb{K}^N$, $\|x\|_A = \left[\langle A(x), x \rangle_{\mathbb{K}^N} \right]^{\frac{1}{2}}$,

où $\langle\ ,\ \rangle_{\mathbb{K}^N}$ désigne le produit hermitien canonique de $\mathbb{K}^N$. Soit $|||\ |||_A$ la norme subordonnée à la norme $\|\ \|_A$ sur l'espace vectoriel $\mathcal{M}_N(\mathbb{K})$. Si $v \in \mathbb{K}^N$ tel que $\|v\|_A \le 1$, alors,

$$\begin{aligned} \|v - T^{-1}A(v)\|_A^2 &= \langle A(v - T^{-1}A(v)),\ v - T^{-1}A(v) \rangle_{\mathbb{K}^N} \\ &= \|v\|_A^2 - \langle A(T^{-1}A(v)), v \rangle_{\mathbb{K}^N} - \langle A(v), T^{-1}A(v) \rangle_{\mathbb{K}^N} \\ &\quad + \langle A(T^{-1}A(v)),\ T^{-1}A(v) \rangle_{\mathbb{K}^N} \end{aligned}$$

Or,

$A(v) = T(T^{-1}A(v))$,

De sorte que :

$$\begin{aligned} \|v - T^{-1}A(v)\|_A^2 &= \|v\|_A^2 - \langle T^{-1}A(v),\ T(T^{-1}A(v)) \rangle_{\mathbb{K}^N} + \langle (A-T)(T^{-1}A(v)), \\ &\qquad T^{-1}A(v) \rangle_{\mathbb{K}^N} \\ &= \|v\|_A^2 - \langle D(T^{-1}A(v)),\ T^{-1}A(v) \rangle_{\mathbb{K}^N} \\ &< \|v\|_A^2 \le 1 \end{aligned}$$

En particulier,

$|||I_N - T^{-1}A|||_A = \max \{ \|v - T^{-1}A(v)\|_A,\ v \in \mathbb{K}^N \text{ t.q. } \|v\|_A \le 1 \} < 1$.

Il s'ensuit que $\rho(I_N - T^{-1}A) < 1$, puis que la méthode de Gauss - Seidel est dans ce cas aussi convergente.

Il est enfin possible de comparer les deux méthodes dans le cas des matrices tridiagonales.

Théorème: Soit $A = \begin{pmatrix} b_1 & c_1 & 0 \\ a_2 & \ddots & \ddots \\ 0 & \ddots & c_{N-1} \\ & a_N & b_N \end{pmatrix} \in \mathcal{M}_N(\mathbb{C})$ une matrice tridiagonale.

Considérons la matrice diagonale $D = \begin{pmatrix} b_1 & & 0 \\ & \ddots & \\ 0 & & b_N \end{pmatrix} \in \mathcal{D}_N(\mathbb{C})$ et la matrice triangulaire inférieure $T = \begin{pmatrix} b_1 & & 0 \\ a_2 & \ddots & \\ 0 & a_N & b_N \end{pmatrix} \in \mathcal{T}_N^-(\mathbb{C})$, et supposons que la matrice D est inversible.

Les rayons spectraux des matrices $I_N - D^{-1}A$ et $I_N - T^{-1}A$ satisfont l'identité:

$$\rho\left(I_N - T^{-1}A\right) = \rho\left(I_N - D^{-1}A\right)^2.$$

En particulier, les méthodes de Jacobi et de Gauss-Seidel convergent ou divergent simultanément, et lorsqu'elles convergent, la méthode de Gauss-Seidel converge plus rapidement.

Preuve:

Soit

$$\forall \lambda \in \mathbb{C}, \quad \begin{cases} P_J(\lambda) = \det\left(\lambda D + A - D\right), \\ \text{et} \quad P_{GS}(\lambda) = \det\left(\lambda T + A - T\right). \end{cases}$$

Comme les matrices D et T sont inversibles, nous avons les équivalences:

$$\begin{cases} \forall \lambda \in \mathbb{C}, \ \lambda \in \sigma\left(I_N - D^{-1}A\right) \iff P_J(\lambda) = 0, \\ \text{et} \quad \forall \lambda \in \mathbb{C}, \ \lambda \in \sigma\left(I_N - T^{-1}A\right) \iff P_{GS}(\lambda) = 0. \end{cases}$$

De plus,

$$\begin{aligned} \forall \lambda \in \mathbb{C}, \ P_{GS}(\lambda^2) &= \det\left(\lambda^2 T + A - T\right) \\ &= \det \begin{pmatrix} \lambda^2 b_1 & c_1 & & 0 \\ \lambda^2 a_2 & \ddots & \ddots & \\ & \ddots & \ddots & c_{N-1} \\ 0 & & \lambda^2 a_N & \lambda^2 b_N \end{pmatrix}, \end{aligned}$$

Or,

$$\begin{pmatrix} \lambda^2 b_1 & c_1 & & 0 \\ \lambda^2 a_2 & \ddots & \ddots & \\ & \ddots & \ddots & c_{N-1} \\ 0 & & \lambda^2 a_N & \lambda^2 b_N \end{pmatrix} = \begin{pmatrix} \lambda & & 0 \\ & \ddots & \\ 0 & & \lambda^N \end{pmatrix} \begin{pmatrix} \lambda^2 b_1 & \lambda c_1 & & 0 \\ \lambda a_2 & \ddots & \ddots & \\ & \ddots & \ddots & \lambda c_{N-1} \\ 0 & & \lambda a_N & \lambda^2 b_N \end{pmatrix} \begin{pmatrix} \lambda^{-1} & & 0 \\ & \ddots & \\ 0 & & \lambda^{-N} \end{pmatrix}$$

De sorte que:

$$\forall \lambda \in \mathbb{C}, \ P_{GS}(\lambda^2) = \det \begin{pmatrix} \lambda^2 b_1 & \lambda c_1 & & 0 \\ \lambda a_2 & \ddots & \ddots & \\ & \ddots & \ddots & \lambda c_{N-1} \\ 0 & & \lambda a_N & \lambda^2 b_N \end{pmatrix} = \lambda^N \det\left(\lambda D + A - D\right).$$

En conclusion, il vient :

$\lambda^2 \in \sigma\left(I_N - T^{-1}A\right)\setminus\{0\}$ ssi $\lambda \in \sigma\left(I_N - D^{-1}A\right)\setminus\{0\}$,

De suite que :

$$\rho\left(I_N - T^{-1}A\right) = \rho\left(I_N - D^{-1}A\right)^2.$$

Les autres conclusions du théorème résultent alors des énoncés précédents.

Au moins dans le cas des matrices tridiagonales, la méthode de Gauss - Seidel est meilleure que celle de Jacobi. Il reste pour conclure que la méthode de Gauss Seidel fait partie des méthodes dites de relaxation pour lesquelles il est encore permis d'étendre ce résultat de comparaison de convergence. Nous renvoyons à l'ouvrage "Introduction à l'analyse numérique matricielle et à l'optimisation" de Ph. Ciarlet pour de plus amples détails.

V Méthodes de calculs des valeurs propres et des vecteurs propres

Le problème de la recherche des éléments propres d'une matrice $A \in \mathcal{M}_N(\mathbb{C})$ est délicat. Outre les problèmes d'ennuis que nous avons déjà évoqués, il n'existe pas de méthode directe de recherche des valeurs propres au delà de la dimension $N = 5$. Bien, il n'existerait en effet possible de calculer les racines d'un polynôme de degré supérieur ou égal à 5.

Nous allons donc présenter deux méthodes itératives pour déterminer certaines des valeurs propres de la matrice $A \in \mathcal{M}_N(\mathbb{C})$: les méthodes de la puissance itérée, et de la puissance inverse. Il existe bien sûr d'autres méthodes itératives pour déterminer des éléments propres d'une matrice, par exemple, les méthodes de Jacobi, de Givens - Householder, ou la méthode QR. Nous renvoyons à l'ouvrage "Introduction à l'analyse numérique matricielle et à l'optimisation" de Ph. Ciarlet pour de plus amples détails.

1. Méthode de la puissance itérée

Soit $N \in \mathbb{N}^*$. Considérons une matrice $A \in \mathcal{M}_N(\mathbb{C})$ telle que A est diagonalisable, et notons $\sigma(A) = \{\lambda_1, \dots, \lambda_N\}$ son spectre avec multiplicité, et $(e_i)_{1 \le i \le N}$ une base de vecteurs propres de A telle que:

$$\forall\ 1 \le i \le N,\quad A(e_i) = \lambda_i e_i$$

Tout vecteur $x \in \mathbb{C}^N$ se décompose sous la forme:

$$x = \sum_{i=1}^{N} x_i e_i,$$

De sorte que:

$$\forall\ n \in \mathbb{N},\quad A^n(x) = \sum_{i=1}^{N} x_i A^n(e_i) = \sum_{i=1}^{N} x_i \lambda_i^n e_i.$$

En particulier, s'il existe une valeur propre λ_1 de plus grand module c'est-à-dire telle que:

$$\forall\ 2 \le i \le N,\quad |\lambda_i| < |\lambda_1|,$$

et, si x est un vecteur de $\mathbb{C}^N$ tel que $x_1 \neq 0$, alors,

$$\|A^n(x)\| \underset{n \to +\infty}{\sim} |x_1|\,|\lambda_1|^n \|e_1\|,$$

où $\|\ \|$ désigne une norme quelconque de $\mathbb{C}^N$. La norme des puissances itérées $A^n(x)$ est contrôlée par la plus grande valeur propre de A, ce qui permet donc d'obtenir cette valeur propre. Cette remarque conduit à la méthode de la puissance itérée suivante.

Théorème: Soit $A \in \mathcal{M}_N(\mathbb{C})$ une matrice diagonalisable. Notons $\sigma(A) = \{\lambda_1, \dots, \lambda_N\}$ son spectre avec multiplicité, et $(e_i)_{1 \le i \le N}$ une base de vecteurs propres de A telle que:

$$\forall\ 1 \le i \le N,\ \|e_i\|_2 = 1 \text{ et } A(e_i) = \lambda_i e_i,$$

où $\|\ \|_2$ désigne la norme hermitienne canonique de $\mathbb{C}^N$. Choisissons un vecteur $x_0 \in \mathbb{C}^N$ tel que $\|x_0\|_2 = 1$ et définissons la suite $(x_n)_{n \in \mathbb{N}}$ par la formule de récurrence:

$$\forall\ n \in \mathbb{N},\quad x_{n+1} = \frac{A(x_n)}{\|A(x_n)\|_2}.$$

Posons enfin:

$$\forall\ n \in \mathbb{N},\quad \mu_n = \langle A(x_n), x_n \rangle_2.$$

Supposons que:

(i) $\forall 2 \leq i \leq N,\ |\lambda_i| < |\lambda_2|$

(ii) Le vecteur $x_0 = \sum_{i=2}^{N} (x_0)_i e_i$ vérifie la condition : $(x_0)_2 \neq 0$.

Alors, les suites $(x_n)_{n\in\mathbb{N}}$ et $(\nu_n)_{n\in\mathbb{N}}$ sont bien définies et vérifient :

(i) $\nu_n \xrightarrow[n\to+\infty]{} \lambda_2$,

(ii) $\exists e \in E_{\lambda_2}(A)$ t.q. $\frac{|\lambda_2|^n}{\lambda_2^n} x_n \xrightarrow[n\to+\infty]{} e$.

Preuve :

Comme $\lambda_2 > 0$ et $(x_0)_2 \neq 0$, le vecteur $A(x_0)$ n'est pas nul, de sorte que le vecteur x_1 est bien défini et vaut :

$$x_1 = \frac{A(x_0)}{\|A(x_0)\|_2}.$$

Supposons alors que la suite $(x_n)_{n\in\mathbb{N}}$ est bien définie jusqu'au rang n et que :

$$\forall\ 0 \leq k \leq n,\ x_k = \frac{A^k(x_0)}{\|A^k(x_0)\|_2}.$$

En particulier, il existe un nombre strictement positif λ tel que :

$$x_n = \lambda A^n(x_0) = \lambda \sum_{i=2}^{N} (x_0)_i \lambda_i^n e_i$$

$$\Rightarrow x_n \notin \ker(A).$$

Le vecteur x_{n+1} est donc bien défini et :

$$x_{n+1} = \frac{A^{n+1}(x_0)}{\|A^n(x_0)\|_2} \times \frac{\|A^n(x_0)\|_2}{\|A^{n+1}(x_0)\|_2} = \frac{A^{n+1}(x_0)}{\|A^{n+1}(x_0)\|_2}.$$

D'où le fait que les suites $(x_n)_{n\in\mathbb{N}}$ et $(\nu_n)_{n\in\mathbb{N}}$ sont bien définies et que :

$$\forall n \in \mathbb{N},\ x_n = \frac{A^n(x_0)}{\|A^n(x_0)\|_2}.$$

Comme

$$\forall n \in \mathbb{N},\ A^n(x_0) = \sum_{i=2}^{N} (x_0)_i \lambda_i^n e_i,$$

il vient :

$$A^n(x_0) \underset{n\to+\infty}{\sim} (x_0)_2 \lambda_2^n e_2 \quad \text{et} \quad \|A^n(x_0)\|_2 \underset{n\to+\infty}{\sim} |(x_0)_2|\,|\lambda_2|^n,$$

De sorte que :

$$\nu_n \underset{n\to+\infty}{\sim} \frac{|\lambda_2|^{2n} \lambda_2 (x_0)_2^2}{|\lambda_2|^{2n} (x_0)_2^2} \xrightarrow[n\to+\infty]{} \lambda_2,$$

et

$$\frac{|\lambda_2|^n}{\lambda_2^n} x_n \xrightarrow[n\to+\infty]{} e = \frac{(x_0)_2}{|(x_0)_2|} e_2 \in E_{\lambda_2}(A).$$

L'hypothèse sur le vecteur x_0 n'est pas une difficulté en pratique. De fait des essais d'ordinateur, il est probable que l'un des vecteurs $(x_n)_{n\in\mathbb{N}}$ aura une

composante non nulle suivant le vecteur e_2.

De plus, la décomposition en blocs de Jordan d'une matrice quelconque de $\mathcal{M}_n(\mathbb{C})$ permet d'étendre la méthode aux matrices $A \in \mathcal{M}_n(\mathbb{C})$ telles que $\exists \lambda \in \sigma(A)$ t.q. $\forall \rho \in \sigma(A) \setminus \{\lambda\}$, $|\rho| < |\lambda|$.
Par contre, cette méthode ne fonctionne plus nécessairement lorsque la matrice A a deux valeurs propres **différentes** de même module.

Afin de déterminer les autres valeurs et vecteurs propres de la matrice, il est possible d'utiliser une méthode de déflation. Sous les hypothèses du théorème précédent, la transposée ${}^tA$ de la matrice A est elle aussi **diagonalisable** et son spectre est identique à celui de A. Il existe donc une base $(f_i)_{1 \le i \le N}$ de vecteurs propres de ${}^tA$ tels que:
$\forall\, 1 \le i \le N$, ${}^tA(f_i) = \lambda_i f_i$ et $\|f_i\|_2 = 1$.
En appliquant la méthode de la puissance itérée aux matrices A et ${}^tA$, nous pouvons déterminer des valeurs approchées de λ_2 et des vecteurs propres e_2 et f_2. Supposons alors que:
${}^tf_2 e_2 \neq 0$.
Dans ce cas, il est possible de définir la matrice:
$$B = A - \lambda_2 \frac{e_2 {}^tf_2}{{}^tf_2 e_2}.$$
Il vient alors:
$B(e_2) = 0$
Et,
$\forall\, 2 \le j \le N$, ${}^tf_2 A e_j = \lambda_j {}^tf_2 e_j = \lambda_2 {}^tf_2 e_j \Rightarrow {}^tf_2 e_j = 0$,
On note que:
$\forall\, 2 \le j \le N$, $B(e_j) = \lambda_j e_j$.
Le spectre de la matrice B est donc égal à:
$\sigma(B) = \{0, \lambda_2, \dots, \lambda_N\}$
Si cette matrice vérifie les hypothèses du théorème précédent, alors, il est possible de calculer sa plus grande valeur propre par la méthode de la puissance itérée, et d'en déduire une seconde valeur propre de A, et un

vecteur propre associé.

Il notera cependant que cette méthode de déflation se heurte au problème bien connu en raison desquels les identités précédentes ne sont pas exactes. En pratique, il est préférable d'utiliser la méthode de la puissance inverse pour déterminer les autres éléments propres de la matrice A.

2. Méthode de la puissance inverse

Soit $N \in \mathbb{N}^*$. Afin de déterminer les valeurs propres d'une matrice $A \in \mathcal{M}_N(\mathbb{C})$, la méthode de la puissance inverse repose sur un argument de localisation des valeurs propres de A, puis sur la méthode de la puissance itérée afin de déterminer leurs valeurs approchées.

Plus précisément, la première étape consiste à déterminer des valeurs approchées des valeurs propres de la matrice A. Elle repose sur des arguments de localisation tels que le théorème suivant.

Théorème de Gershgorin - Hadamard : Soit $A \in \mathcal{M}_N(\mathbb{C})$. Si $\lambda \in \mathrm{Sp}(A)$, alors,

$$\exists i \leq N \text{ t.q. } |\lambda - a_{ii}| \leq \sum_{\substack{j=1 \\ j \neq i}}^{N} |a_{ij}|$$

Preuve :

Par l'absurde, si cette conclusion n'est pas vérifiée, alors, la matrice $\lambda I_N - A$ est à diagonale dominante, donc inversible, ce qui est contradictoire.

Une fois localisée grossièrement une valeur propre de A à travers une valeur approchée λ_0, deux cas de figure peuvent se produire :

- si $\lambda_0 \in \mathrm{Sp}(A)$, alors, le problème est résolu et il est possible de chercher les autres valeurs propres de A ;
- sinon, la matrice $A - \lambda_0 I_N$ est inversible et le spectre de son inverse est donné par les nombres $\frac{1}{\mu - \lambda_0}$ où $\mu \in \mathrm{Sp}(A)$. En particulier, si λ_0 est

suffisamment proche de λ_i, alors, la valeur propre $\frac{1}{\lambda - \lambda_0}$ sera la valeur propre dominante de la matrice $(A - \lambda_0 I_N)^{-1}$ et il sera possible de la calculer par la méthode de la puissance itérée.
D'un point de vue théorique, cette discussion est résumée par l'énoncé suivant.

Théorème : Soit $A \in \mathcal{M}_N(\mathbb{C})$ une matrice diagonalisable. Notons $\sigma(A) = \{\lambda_1, \dots, \lambda_N\}$ son spectre avec multiplicité, et $(e_i)_{1 \le i \le N}$ une base de vecteurs propres de A telle que :
$\forall 1 \le i \le N$, $\|e_i\|_2 = 1$ et $A(e_i) = \lambda_i e_i$,
où $\|\ \|_2$ désigne la norme hermitienne canonique de $\mathbb{C}^N$.
Supposons alors donné un nombre complexe μ tel que :
$\mu \neq \lambda_i$ et $\forall 1 \le i \le j \le N$, $|\mu - \lambda_j| > |\mu - \lambda_i|$.
Choisissons alors un vecteur $x_0 = \sum_{i=1}^{N} (x_0)_i e_i \in \mathbb{C}^N$ tel que $\|x_0\|_2 = 1$ et $(x_0)_1 \neq 0$ et définissons la suite $(x_n)_{n \in \mathbb{N}}$ par la formule de récurrence :
$$\forall n \in \mathbb{N}, \quad x_{n+1} = \frac{(A - \mu I_N)^{-1}(x_n)}{\|(A - \mu I_N)^{-1}(x_n)\|}$$
La suite $(x_n)_{n \in \mathbb{N}}$ est alors bien définie et elle satisfait :
(i) $\langle (A - \mu I_N)^{-1}(x_n), x_n \rangle_2 \xrightarrow[n \to +\infty]{} \frac{1}{\lambda_1 - \mu}$
(ii) $\exists e \in E_{\lambda_1}(A)$ t.q. $\frac{(\lambda_1 - \mu)^n}{|\lambda_1 - \mu|^n} x_n \xrightarrow[n \to +\infty]{} e$

Preuve :

Immédiate.

En sus des difficultés inhérentes à la méthode de la puissance itérée, la principale difficulté tient ici au choix de la valeur approchée μ de la valeur propre λ_1. La convergence de la méthode de la puissance inverse est d'autant meilleure que μ est proche de λ_1, mais alors, la matrice $(A - \mu I_N)^{-1}$ est bien plus singulière. En pratique, il s'agit de trouver un équilibre entre ces deux propriétés ce qui n'est pas évident.