UNIVERSITÉ PARIS XI
U.E.R. MATHÉMATIQUE
91-ORSAY (FRANCE)

N° 20

Temam Roger

Algèbre Linéaire

C3 Analyse Numérique

(1969-1970)

(Publication mathématique d'Orsay)

Table des matières.

CHAPITRE I

RESOLUTION DES SYSTEMES LINEAIRES.

Soit l'équation : $\mathbf{A}.x = b$, où $\mathbf{A}$ est une matrice réelle (n,n) régulière, b un vecteur donné de $\mathbb{R}^n$. L'inconnue est x , vecteur de $\mathbb{R}^n$. Le but de ce chapitre est l'étude de méthodes pratiques de résolution d'un tel système (pour n grand!). On distingue essentiellement deux types de méthodes :

- les méthodes dites directes : x est obtenu "exactement" au bout d'un nombre fini d'opérations

- les méthodes dites itératives : x est obtenu comme la limite d'une suite infinie convergente :

$$x = \lim_{\nu \to +\infty} x^{(\nu)}$$

§I. Méthodes directes.

- Systèmes dont la matrice est triangulaire :

la résolution est immédiate.

. Si $\mathbf{A}$ est triangulaire inférieure, c'est-à-dire :

$$\mathbf{A} = \begin{pmatrix} a_{11} & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & 0 \\ a_{n1} & \cdots & & a_{nn} \end{pmatrix} : a_{ij} = 0 \quad \text{si} \quad j > i \ ;$$

Alors $\det \mathbf{A} = a_{11} \times a_{22} \times \ldots \times a_{nn}$.

Par hypothèse $\mathbf{A}$ est régulière : $\det \mathbf{A}$ est non nul, donc : $a_{ii} \neq 0 \ (i=1,\ldots,n)$

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \quad , \quad b = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} \ ;$$

On calcule aisément les composantes successives $x_1,\dots,x_n$ de x :

$$a_{11}\,x_1 = b_1 \qquad \Longrightarrow x_1 = b_1/a_{11}$$

$$a_{21}\,x_1 + a_{22}\,x_2 = b_2 \qquad \Longrightarrow x_2 = b_2/a_{22} - \frac{a_{21}}{a_{22}}\,x_1$$

$$a_{n1}\,x_1 + a_{n2}\,x_2 + \dots + a_{nn}\,x_n = b_n \Longrightarrow x_n = \frac{1}{a_{nn}}\left(b_n - \sum_{j=1}^{n-1} a_{nj}\,x_j\right)$$

. Si A est triangulaire <u>supérieure</u> :

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ & \ddots & \vdots \\ 0 & & a_{nn} \end{pmatrix} \quad (a_{ij} = 0 \quad \text{si} \quad j < i)$$

La situation est la même, mais cette fois on commence par la $n^{\text{ième}}$ équation, qui donne x_n :

$$a_{nn}\,x_n = b_n \qquad \Longrightarrow x_n = b_n/a_{nn}$$

$$a_{kk}\,x_k + a_{k,k+1}\,x_{k+1} + \dots + a_{k,n}\,x_n = b_k \Longrightarrow x_k = \frac{1}{a_{kk}}\left(b_k - \sum_{j=k+1}^{n} a_{kj}\,x_j\right)$$

$$a_{11}\,x_1 + a_{12}\,x_2 + \dots + a_{1n}\,x_n = b_n \qquad \Longrightarrow x_1 = \frac{1}{a_{11}}\left(b_1 - \sum_{j=2}^{n} a_{kj}\,x_j\right)$$

∎

Nombre d'opérations nécessaires :

- nombre d'additions : $\sum_{k=1}^{n} (n-k) = n^2 - \frac{n(n+1)}{2} = \frac{n(n-1)}{2}$
- nombre de multiplications : $\frac{n(n-1)}{2}$
- nombre de divisions : n .

Cas général . Les différentes méthodes directes ramènent la résolution de $A.x = b$ à la résolution d'un système dont la matrice est triangulaire. Pour cela (sauf dans la méthode de Choleski où le point de vue est légèrement différent), on cherche une matrice régulière B telle que $B.A$ soit triangulaire ; alors :

$$A.x = b \Longleftrightarrow BA.x = B.b \ .$$

En pratique, on détermine B comme produit de matrices élémentaires :

$B = B^{(q)} \ldots B^{(1)}$, les matrices $B^{(k)}$ dépendant de la méthode.

1) Méthode de Gauss.

Considérons une matrice $J = (\sigma_{ij})$ avec :

$$\begin{cases} \sigma_{ii} = 1 \quad , \quad 1 \leqslant i \leqslant n \\ \sigma_{i1} = \alpha_i \quad \text{pour} \quad 2 \leqslant i \leqslant n \\ \sigma_{ij} = 0 \quad \text{dans les autres cas,} \end{cases}$$

soit : $J = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ \alpha_2 & 1 & \ddots & \vdots \\ \vdots & 0 & \ddots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ \alpha_n & 0 & \cdots & 1 \end{pmatrix}$

Cette matrice est triangulaire inférieure, de déterminant 1.

Si C est une matrice d'ordre (n,n) (ou même d'ordre (n,q), $q \neq n$) et si $C' = J.C$, on a le lemme suivant :

Lemme 1 : Soient L'_i $(1 \leqslant i \leqslant n)$ les lignes de C' et L_i $(1 \leqslant i \leqslant n)$ les lignes de C. Alors : $L'_1 = L_1$ et : $L'_i = L_i + \alpha_i L_1$ $(2 \leqslant i \leqslant q)$.

Démonstration : $C = \begin{pmatrix} C_{11} & \cdots & C_{1q} \\ \vdots & \ddots & \vdots \\ C_{n1} & \cdots & C_{nq} \end{pmatrix}$ $\quad C' = \begin{pmatrix} C'_{11} & \cdots & C'_{1q} \\ \vdots & \ddots & \vdots \\ C'_{n1} & \cdots & C'_{nq} \end{pmatrix}$

$\forall i$ et j $(1 \leqslant i \leqslant n \ , \ 1 \leqslant j \leqslant q)$: $C'_{ij} = \sum_{k=1}^{n} \sigma_{ik} C_{kj}$

- Si $i = 1$: $\sigma_{ik} = \delta_1^k$ d'où $C'_{1j} = C_{1j}$

(donc : $L'_1 = L_1$)

- Si $i > 1$: $\sigma_{ik} = \alpha_i \delta_1^k + \delta_i^k$ d'où : $C'_{ij} = \alpha_i C_{1j} + C_{ij}$

(donc : $L'_i = L_i + \alpha_i L_1$).

Décrivons à présent la méthode de Gauss. Nous supposons que l'élément a_{11} de la matrice A est non nul (nous reviendrons ultérieurement sur le cas où cet élément est nul). La première étape de la méthode de Gauss consiste à multiplier $A = A^{(1)}$ par une matrice $J^{(1)}$, du type J ci-dessus, avec :

$$\alpha_2 = -\frac{a_{21}}{a_{11}}, \ldots, \alpha_n = -\frac{a_{n1}}{a_{11}}.$$

On obtient alors une matrice $A^{(2)}$:

$$A^{(2)} = J^{(1)} A^{(1)}$$

qui possède des zéros sous la diagonale, dans la première colonne :

$$A^{(2)} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ 0 & a_{22}^{(2)} & \cdots & a_{2n}^{(2)} \\ \vdots & \vdots & & \vdots \\ 0 & a_{n2}^{(2)} & \cdots & a_{nn}^{(2)} \end{pmatrix}$$

ou encore, en notation matricielle par blocs :

$$A^{(2)} = \begin{pmatrix} A_{11}^{(2)} & A_{12}^{(2)} \\ 0 & A_{22}^{(2)} \end{pmatrix} \begin{matrix} \uparrow 1 \\ \downarrow \\ \uparrow n-1 \\ \downarrow \end{matrix}$$

(1 en haut à gauche, $n-1$ en haut à droite)

Supposons que $a_{22}^{(2)}$ soit différent de zéro : la deuxième étape consiste alors à multiplier $A^{(2)}$ par une matrice $J^{(2)}$:

$$J^{(2)} = \begin{pmatrix} 1 & 0 \\ 0 & J_{22}^{(2)} \end{pmatrix}$$

où $J_{22}^{(2)}$ est du type J à l'ordre $n-1$, de manière à obtenir une matrice $A^{(3)} = J^{(2)} A^{(2)}$ qui possède des 0 sous la diagonale dans les deux premières colonnes ;

On a : $J^{(2)} A^{(2)} = \left(\begin{array}{c|c} A_{11}^{(2)} & A_{12}^{(2)} \\ \hline 0 & J_{22}^{(2)} A_{22}^{(2)} \end{array}\right)$,

la première colonne de $J_{22}^{(2)}$ est : $\begin{pmatrix} 1 \\ -\frac{a_{32}^{(2)}}{a_{22}^{(2)}} \\ \vdots \\ -\frac{a_{n2}^{(2)}}{a_{22}^{(2)}} \end{pmatrix}$

Le processus se poursuivant ainsi, on obtient à la fin de la $(p-1)^{\text{ième}}$ étape, une matrice

$$A^{(p)} = J^{(p-1)} \ldots J^{(1)} . A^{(1)} ,$$

avec : $A^{(p)} = \left(\begin{array}{c|c} A_{11}^{(p)} & A_{12}^{(p)} \\ \hline 0 & A_{22}^{(p)} \end{array}\right)$ ($A_{11}^{(p)}$: $p-1$; $A_{22}^{(p)}$: $n-p+1$) ,

où $A_{11}^{(p)}$ est triangulaire supérieure, et :

$$A_{22}^{(p)} = \begin{pmatrix} a_{pp}^{(p)} & \cdots & a_{pn}^{(p)} \\ \vdots & \ddots & \vdots \\ a_{np}^{(p)} & \cdots & a_{pn}^{(p)} \end{pmatrix}$$

Si $a_{pp}^{(p)} \neq 0$, la $p^{\text{ième}}$ étape consiste à multiplier $A^{(p)}$ par une matrice $J^{(p)}$:

$$\begin{cases} J^{(p)} = \left(\begin{array}{c|c} I_{p-1} & 0 \\ \hline 0 & J_{22}^{(p)} \end{array}\right) \\ I_{p-1} = \text{matrice identité d'ordre } (p-1 , p-1) \\ J_{22}^{(p)} \text{ du type } J , \text{ d'ordre } n-p+1 , \end{cases}$$

de manière à obtenir une matrice : $A^{(p+1)} = J^{(p)} A^{(p)}$, $A^{(p+1)}$ possédant des zéros sous la diagonale dans les p <u>premières colonnes</u>.

On a : $J^{(p)} A^{(p)} = \left(\begin{array}{c|c} A_{11}^{(p)} & A_{12}^{(p)} \\ \hline 0 & J_{22}^{(p)} A_{22}^{(p)} \end{array}\right)$,

et le choix de $J_{22}^{(p)}$ est alors évident : sa première colonne s'écrit :

$$\begin{pmatrix} 1 \\ -a_{p+1,p}^{(p)}/a_{p,p}^{(p)} \\ \vdots \\ -a_{n,p}^{(p)}/a_{p,p}^{(p)} \end{pmatrix}$$

Finalement, on obtient <u>après la</u> $(n-1)^{\text{ième}}$ <u>étape du calcul</u>, une matrice $A^{(n)}$ triangulaire supérieure :

$$A^{(n)} = \begin{pmatrix} a_{11}^{(1)} & \cdots & a_{1n}^{(1)} \\ 0 & \ddots & \vdots \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \; a_{nn}^{(n)} \end{pmatrix}$$

et $A^{(n)} = B.A$, avec $B = J^{(n-1)} \times \ldots \times J^{(1)}$; les matrices $J^{(1)}, \ldots, J^{(n-1)}$ sont régulières, de déterminant $+1$, donc B est régulière et :

$$Ax = b \Longleftrightarrow BA.x = B.b \; .$$ ∎

<u>Cas des pivots nuls</u> : On appelle <u>pivot</u> d'ordre p le terme $a_{pp}^{(p)}$, intersection de la ligne p et de la colonne p de $J^{(p-1)} A^{(p-1)}$. Si après la $(p-1)^{\text{ième}}$ étape $\underline{a_{pp}^{(p)} = 0}$: $A^{(p)}$ est régulière, puisque A est régulière ; puisque $A^{(p)}$ possède des 0 sous la diagonale dans les $(p-1)$ premières colonnes, il existe j $(j > p)$ tel que $a_{jp}^{(p)} \neq 0$.

On échange alors les lignes j et p .

Autrement dit, on pose :

$$\mathbf{A}^{(p+1)} = J^{(p)}\ \Sigma^{(j,p)}\ \mathbf{A}^{(p)} ,$$

où $\Sigma^{(j,p)}$ est la matrice de permutation :

$$\Sigma^{(j,p)} = \begin{pmatrix} 1 & 0 & \cdots & \cdots & 0 \\ 0 & \ddots & & & \\ \vdots & & 0 & \cdots & 1 \\ & & \vdots & \ddots & \vdots \\ & & 1 & \cdots & 0 \\ 0 & & & & \ddots \\ & & & & & 1 \end{pmatrix}$$

(p, j)

La matrice $\Sigma^{(j,p)}$ est régulière, de déterminant -1.

Stratégie des pivots : Il est préférable même lorsque le pivot n'est pas nul de choisir un pivot aussi grand que possible en valeur absolue (afin d'atténuer l'importance des erreurs d'arrondi). Donc, systématiquement, à la $p^{\text{ième}}$ étape du calcul, on amène en position pivot (c'est-à-dire à l'intersection de la $p^{\text{ième}}$ ligne et de la $p^{\text{ième}}$ colonne) le terme $a_{j,p}^{(p)}$ de module maximum, tel que :

$$|a_{k,p}^{(p)}| \leqslant |a_{j,p}^{(k)}| \quad \text{pour} \quad k \geqslant p$$

S'il existe plusieurs indices j tels que $|a_{j,p}^{(p)}|$ soit le plus grand des $|a_{k,p}^{(p)}|$ $(k \geqslant p)$, on prend le plus petit de ces indices j.

On échange alors les lignes j et p.

Traitement du second membre : Les éléments de b doivent subir les mêmes combinaisons que les lignes de la matrice : il est donc commode de considérer b comme une $(n+1)^{\text{ième}}$ colonne de $\mathbf{A}$:

$$b_i = a_{i,n+1} \quad (1 \leqslant i \leqslant n).$$

Conclusion : Une fois l'échange des lignes effectué, on obtient $\mathbf{A}^{(p+1)}$ par les formules :

$$a_{ij}^{(p+1)} = a_{ij}^{(p)} - \frac{a_{ip}^{(p)}}{a_{pp}^{(p)}} a_{pj}^{(p)} \quad \text{si} \quad p+1 \leqslant i \leqslant n \quad \text{et} \quad p+1 \leqslant j \leqslant n+1$$

$$a_{ij}^{(p+1)} = a_{ij}^{(p)} \quad \text{si} \quad 1 \leqslant i \leqslant p \,, \; 1 \leqslant j \leqslant n+1$$

$$a_{ij}^{(p+1)} = 0 \quad \text{si} \quad p+1 \leqslant i \leqslant n \text{ .et } 1 \leqslant j \leqslant p \,.$$

Nombre d'opérations.

A la $p^{\text{ième}}$ étape, nous devons calculer $(n-p+1)(n-p)$ termes (avec les seconds membres), et le calcul de chaque terme nécessite 3 opérations (1 soustraction, 1 multiplication, 1 division). Le nombre total d'opérations est donc :

$$3 \sum_{p=1}^{n-1} (n-p+1)(n-p) = 3\left[\sum_{q=1}^{n-1} q^2 + \sum_{q=1}^{n-1} q\right] = 3\left[\frac{n(n-1)(2n-1)}{6} + \frac{3n(n-1)}{6}\right] = n(n-1)(n+\tfrac{1}{2}).$$

Il faut compter en plus la résolution du système triangulaire, ce qui demande n^2 opérations.

Le nombre total d'opérations est de l'ordre de $\boxed{n^3 \text{ , pour n grand}}$.

2) Méthode des rotations (Givens).

. Considérons une matrice (n,n) de la forme :

$$V(\theta) = \left(\begin{array}{cc|c} \multicolumn{2}{c|}{\overset{2}{\longleftrightarrow}} & \overset{n-2}{\longleftrightarrow} \\ \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & \\ \hline & & \\ 0 & & I_{n-2} \end{array}\right) \begin{array}{l} \updownarrow 2 \\ \updownarrow n-2 \end{array} \qquad (\theta \in \mathbb{R})$$

($V(\theta)$ est la matrice de la rotation d'angle θ dans le plan : $x_3 = \ldots = x_n = 0$).

La matrice $V(\theta)$ est orthogonale, et $\det(V(\theta)) = +1$.

. Si C est une matrice (n,q) (q entier quelconque), et si $C' = V_\theta C$, soient L_i $(1 \leqslant i \leqslant n)$ les lignes de C , L'_i $(1 \leqslant i \leqslant n)$ les lignes de C' :

$$L'_1 = L_1 \cos\theta - L_2 \sin\theta$$

$$L'_2 = L_1 \sin\theta + L_2 \cos\theta$$

Si $i \neq 1$ et 2 : $L_i' = L_i$.

. Considérons la matrice :

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix}$$

Je cherche à amener des 0 sous la diagonale de A , et d'abord dans la première colonne, en multipliant à gauche la matrice A par des matrices convenables.

- si $a_{11} = a_{21} = 0$: je passe à la ligne suivante (la troisième)

- si a_{11} et a_{21} ne sont pas simultanément nuls, posons :

$$Q^{(1,1)} = \left(\begin{array}{cc|c} \cos\theta_1 & \sin\theta_1 & 0 \\ -\sin\theta_1 & \cos\theta_1 & \\ \hline 0 & & I_{n-2} \end{array}\right)$$

avec : $\cos\theta_1 = \dfrac{a_{11}}{\sqrt{a_{11}^2 + a_{21}^2}}$, $\sin\theta_1 = \dfrac{a_{21}}{\sqrt{a_{11}^2 + a_{21}^2}}$.

Alors : $A^{(2,1)} = Q^{(1,1)} A$ possède un zéro à la place de a_{21} :

$$A^{(2,1)} = \begin{pmatrix} a_{11}^{(2,1)} & \cdots & a_{1n}^{(2,1)} \\ 0 & a_{22}^{(2,1)} \cdots & a_{2n}^{(2,1)} \\ a_{31} & a_{32} \; a_{33} \cdots & \vdots \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix}$$

- pour faire apparaître un 0 à la place de a_{31} (s'il n'y en a pas déjà un), posons :

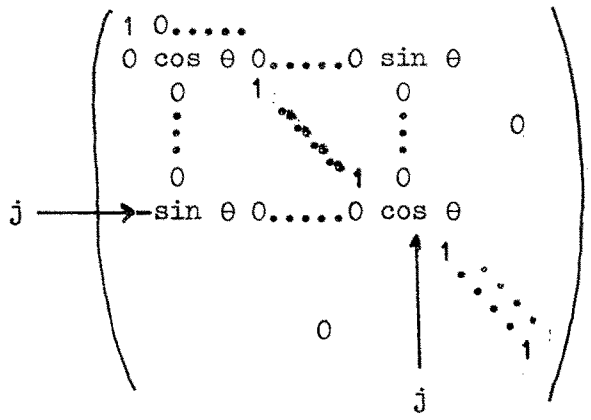
$$Q^{(2,1)} = \left(\begin{array}{ccc|c} \cos\theta_2 & 0 & \sin\theta_2 & \\ 0 & 1 & 0 & 0 \\ -\sin\theta_2 & 0 & \cos\theta_2 & \\ \hline & 0 & & I_{n-3} \end{array}\right)$$

avec $\cos\theta_2 = \dfrac{a_{11}}{\sqrt{a_{11}^2 + a_{31}^2}}$, $\sin\theta_2 = \dfrac{a_{31}}{\sqrt{a_{11}^2 + a_{31}^2}}$, et : $A^{(3,1)} = Q^{(2,1)}\, A^{(2,1)}$.

- Le processus se poursuit jusqu'à la dernière ligne ; nous obtenons :

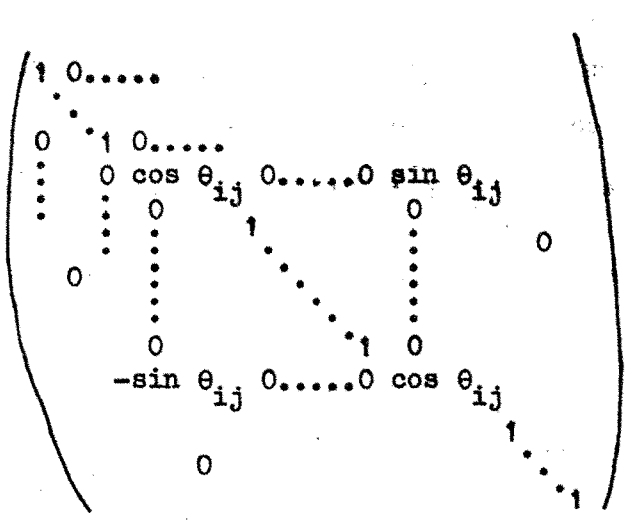
$A^{(2)} = Q^{(n-1,1)} \times \ldots \times Q^{(1,1)} A$, où $A^{(2)}$ possède des zéros sous sa diagonale en première colonne, et où les matrices $Q^{(k,1)}$ sont orthogonales de déterminant $+1$.

- On passe ensuite à la deuxième colonne, où l'on fait apparaître des zéros sous la diagonale en multipliant $A^{(2)}$ à gauche par des matrices de rotation :



Le processus se poursuit, colonne par colonne, jusqu'à l'obtention d'une matrice $A^{(n)}$ triangulaire supérieure : $A^{(n)} = Q.A$, où Q est un produit de matrices de rotation : $Q = \prod_{j=1}^{n-1} \left(\prod_{i=j}^{n} V_{ij}(\theta_{ij}) \right)$

avec $V_{ij}(\theta_{ij}) =$



La matrice Q est orthogonale, de déterminant $+1$.

Nombre d'opérations.

- Pour transformer la première colonne, il faut d'abord calculer les coefficients $\cos \theta_{11}$ et $\sin \theta_{11}$, ce qui demande 2 multiplications, 1 addition, 1 extraction de racine carrée, et 2 divisions (sauf si $a_{11} = a_{21} = 0$).

Il faut ensuite transformer les lignes 1 et 2, ce qui nécessite : $4(n+1)$ multiplications (en tenant compte de la transformation du second membre b) et $2(n+1)$ additions : donc, pour amener un zéro en position a_{21} , il faut $(4n+6)$ multiplications, $(2n+3)$ additions, 2 divisions, 1 extraction de racine carrée.

Pour amener des zéros sous la diagonale en première colonne, il faudra donc : $(n-1)(4n+6)$ multiplications, $(n-1)(2n+3)$ additions, $2(n-1)$ divisions, $(n-1)$ extractions de racine carrée (ou peut-être moins d'opérations, si certains termes se trouvent déjà nuls).

Le nombre total d'opérations nécessaires est donc au maximum :

- pour les multiplications : $\sum_{k=1}^{n-1} (k-1)(4k+6) = \frac{(n-1)}{3} (4n^2 - 17n + 18)$

de l'ordre de $\boxed{\frac{4n^3}{3}}$ quand n est grand

- pour les <u>additions</u> : $\sum_{k=1}^{n-1} (k-1)(2k+3) = \frac{n-1}{6}(4n^2 - 17n + 18)$

- pour les divisions : $2\sum_{k=1}^{n-1} (k-1) = (n-1)(n-2)$

- pour les extractions de racine carrée : $\sum_{k=1}^{n-1} (k-1) = \frac{(n-1)(n-2)}{2}$

quand $n \to +\infty$, le nombre total des opérations varie comme $\boxed{2n^3}$.

3) <u>Méthode de Householder</u>.

Il s'agit encore de faire apparaître des zéros sous la diagonale, d'abord en première colonne, puis en seconde colonne, etc..., jusqu'à la $(n-1)^{\text{ième}}$ colonne.

Dans une première étape on cherche une matrice $B^{(1)}$ telle que $B^{(1)} A^{(1)}_{.1}$ soit parallèle à $e_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$.

($A^{(1)}_{.1}$ désigne la première colonne de $A^{(1)} = A$).

<u>Lemme</u> 1. <u>Soit</u> $u \in \mathbb{R}^n$, <u>de norme</u> $\|u\| = \sqrt{\sum_{i=1}^{n} u_i^2} = 1$. <u>La matrice</u> $H = I - 2u.u^T$ <u>est orthogonale, et elle représente la symétrie par rapport à l'hyperplan</u> P, <u>orthogonal à</u> u.

<u>Démonstration</u> : Tout d'abord, H est symétrique :

$$H^T = I - 2(u.u^T)^T = I - 2u.u^T = H .$$

Montrons que H est orthogonale, c'est-à-dire que $H^T.H = I$:

$$H^T.H = H^2 = (I - 2u.u^T)(I - 2u.u^T) = I - 4u.u^T + 4(u.u^T)(u.u^T).$$

Mais : $(uu^T)(uu^T) = u(u^T.u)u^T$

En fait, u^Tu est un scalaire : $u^Tu = \|u\|^2 = 1$.

Donc $H^TH = I$; H est donc bien orthogonale.

Etudions la transformation sur $\mathbb{R}^n$ associée à H :

$$\forall x \in \mathbb{R}^n \ , \ Hx = x - 2u.(u^T x)$$

et $u^T x$ est en fait le produit scalaire $(u|x)$:

la composante de x suivant u est changée de signe, (multipliée par -1).

La composante suivant P est inchangée : la transformation associée à H est donc la symétrie annoncée.

Pour simplifier, notons a la première colonne de A ; on veut ramener a sur e_1 (mais si a est déjà parallèle à e_1, on ne fait rien).

Lemme 2. Soit $a \in \mathbb{R}^n$, non parallèle à e_1. Il existe $u \in \mathbb{R}^n$, il existe $\alpha \in \mathbb{R}$ tels que : $Ha = \alpha e_1$, avec : $H = I - 2u.u^T$.

N.B. : pour Σ on lira ε.

Démonstration : Nous aurons nécessairement :

$$\|Ha\| = \|\alpha e_1\| = |\alpha|.$$

Mais H est orthogonale d'après le lemme 1 : $\|Ha\| = \|a\|$.

On doit donc prendre : $\boxed{\alpha = \Sigma \|a\|}$ $(\Sigma = +1 \text{ ou } -1)$.

Posons : $\mu = u^T.a$;

$$a - 2u.u^T.a = \alpha e_1 \Longrightarrow a^T.a - 2a^T u.u^T a = \alpha\, a^T e_1$$

$$\Longrightarrow \|a\|^2 - 2\mu^2 = \Sigma \|a\| a_1 \quad (a_1 = a^T e_1).$$

Nous devrons donc avoir : $\mu^2 = \dfrac{\|a\| (\|a\| - \Sigma a_1)}{2}$.

Il faut vérifier que la quantité qui est au second membre est bien positive ; mais cela est vrai, car a n'est pas parallèle à e_1 : il existe $i \neq 1$ $(2 \leqslant i \leqslant n)$ tel que $a^T e_i \neq 0$; $a_i = a^T e_i$:

$$\Sigma a_1 \leqslant |a_1| < \sqrt{a_1^2 + \sum_{i=2}^{n} a_i^2} = \|a\| .$$

Pour chaque valeur de α on a donc deux valeurs possibles, réelles, de μ :

$$\mu = \Sigma' \sqrt{\frac{\|a\| \, (\|a\| - \Sigma \, a_1)}{2}} \qquad (\Sigma' = +1 \quad \text{ou} \quad -1)$$

et $\mu \neq 0$.

Mais u est alors déterminé, car :

$$Ha = \alpha e_1 \Longleftrightarrow a - 2u\mu = \alpha e_1 \; ;$$

$$u = \frac{-\alpha e_1 + a}{2\mu} \; .$$

Le lemme 2 est donc démontré : il y a deux valeurs possibles pour α , et quatre valeurs possibles pour u , telles que $(I - 2u.u^T)a = \alpha e_1$:

$$\alpha = \Sigma \, \|a\| \qquad \Sigma = +1 \quad \text{ou} \quad -1 \; ,$$

$$u = \Sigma' . \frac{a - \Sigma \, \|a\| \, e_1}{\sqrt{2 \, \|a\| \, (\|a\| - \Sigma \, a_1)}} \qquad \Sigma' = +1 \quad \text{ou} \quad -1 \; .$$

La première étape du calcul consiste à former :

$$A^{(2)} = H^{(1)} \, A^{(1)} \qquad (A^{(1)} = A) \; ,$$

où $H^{(1)}$ est la matrice de symétrie fournie par le lemme 2 lorsque $a = A^{(1)}_{.1}$; la première colonne de $A^{(2)}$ est parallèle à e_1 : $A^{(2)}$ contient des 0 sous la diagonale, en première colonne.

Quel est le nombre d'opérations nécessaires pour former $A^{(2)}$?

Il faut d'abord calculer $\|a\| = \|A^{(1)}_{.1}\|$, ce qui demande : n multiplications, (n-1) additions, 1 extraction de racine carrée. Nous devons ensuite former le vecteur $a - \alpha e_1$, ce qui demande n additions, puis calculer le nombre $\sqrt{2 \, \|a\| \, (\|a\| - \alpha)}$, ce qui requiert 1 addition, 2 multiplications, 1 extraction de racine carrée ; il faut ensuite former u , ce qui ne demande que n divisions.

Enfin, nous n'avons plus qu'à former $A^{(2)}$:

$$A^{(2)} = H^{(1)} A^{(1)} = \begin{pmatrix} \Sigma\|A^{(1)}_{.1}\| & A^{(2)}_{12} \\ 0 & \\ \vdots & A^{(2)}_{22} \\ 0 & \end{pmatrix} = (I - 2uu^T)A^{(1)} .$$

Nous avons à former $(n+1)$ vecteurs-colonne $A^{(2)}_{.j}$ (en comptant le second membre).

Or $\forall j \quad (1 \leqslant j \leqslant n+1)$: $A^{(1)}_{.j} - 2u.(u^T A^{(1)}_{.j}) = A^{(2)}_{.j}$.

La formation du produit scalaire : $u^T A^{(1)}_{.j}$ demande $2n$ multiplications et $(n-1)$ additions, le calcul du vecteur : $u(2u^T A^{(1)}_{.j})$ n multiplications et la formation de $A^{(1)}_{.j} - u(2u^T A^{(1)}_{.j})$: n additions.

Donc, pour former $A^{(2)}$ à partir de $A^{(1)}$ il faut, en tout :

- en multiplications : $(n+1) + (n+1)(2n+n) = 3n^2 + 4n + 2$
- en additions : $2n + (n+1)(2n-1) = 2n^2 + 3n - 1$
- en divisions : n
- en extractions de racines carrées : 2. ∎

La $p^{\text{ième}}$ étape du calcul consiste à calculer la matrice : $A^{(p+1)} = H^{(p)} A^{(p)}$, avec :

$$H^{(p)} = \begin{pmatrix} I_{p-1} & 0 \\ 0 & \tilde{H}^{(p)} \end{pmatrix} \begin{matrix} \\ \updownarrow p-1 \\ \updownarrow n-p+1 \end{matrix}$$

(I_{p-1} : $p-1$; 0 : $n-p+1$)

$\tilde{H}^{(p)}$ est une matrice de symétrie dans $\mathbb{R}^{n-p+1}$ fournie par le lemme 2, de telle façon que $A^{(p+1)}$ ait des zéros sous sa diagonale dans les p premières colonnes. (On remarque que la matrice $H^{(p)}$ est orthogonale et symétrique).

Finalement : à la $(n-1)^{\text{ième}}$ étape du calcul, on obtient : $A^{(n)} = B.A$, où $A^{(n)}$ est triangulaire supérieure ; la matrice B est orthogonale (donc régu-

lière) et symétrique.

Nombre total d'opérations dans la méthode de Householder.

- nombre de multiplications :

$$\sum_{p=1}^{n-1} (3(n-p+1)^2 + 4(n-p+1) + 2) = 3\sum_{q=1}^{n-1} q^2 + 4\sum_{q=1}^{n-1} q + 2(n-1)$$

$$= \frac{n-1}{2}(2n^2 + 3n + 4)$$

quand $n \to +\infty$, le nombre de multiplications varie comme $\boxed{n^3}$

- nombre d'additions : $\sum_{q=1}^{n-1} (2q^2 + 3q - 1) = \frac{n-1}{6}(4n^2 + 7n - 6)$

- nombre de divisions : $\sum_{q=1}^{n-1} q = \frac{n(n-1)}{2}$

- nombre d'extractions de racines carrées : $2(n-1) \sim \boxed{2n}$ quand $n \to +\infty$

quand $n \to +\infty$, le nombre total d'opérations varie comme $\boxed{4\frac{n^3}{3}}$

4) Méthode de Choleski.

Cette méthode ne s'adresse qu'à des matrices symétriques, définies positives (toujours régulières).

Lemme. Si A est symétrique, définie positive, il existe une matrice $S = (s_{ij})$ réelle triangulaire supérieure, telle que : $A = S^T.S$, et $s_{ii} > 0$ $(i = 1,\ldots,n)$.

Démonstration : Raisonnons par récurrence sur l'ordre n de la matrice A.

- pour $n = 1$, le lemme est évident.

- supposons la proposition vraie jusqu'à l'ordre $(n-1)$, et écrivons la matrice A (qui est supposée symétrique) sous la forme :

$$A = \left(\begin{array}{c|c} A_{n-1} & C \\ \hline C^T & a_{n,n} \end{array}\right) \begin{array}{l} \updownarrow n-1 \\ \updownarrow 1 \end{array}$$

(avec A_{n-1} : $n-1$; C : 1)

Puisque $\mathbf{A}$ est symétrique, définie positive dans $\mathbf{R}^n$, $\mathbf{A}_{n-1}$ est symétrique définie positive dans $\mathbf{R}^{n-1}$. Par hypothèse de récurrence il existe une matrice S_{n-1} triangulaire supérieure, d'ordre $(n-1)$, telle que :

$$\mathbf{A}_{n-1} = S_{n-1}^T - S_{n-1}$$

(On a : $(\text{dét } S_{n-1})^2 = \text{dét } \mathbf{A}_{n-1} \neq 0$; S_{n-1} est régulière) et S_{n-1} est réelle.

Posons : $$S_n = \begin{pmatrix} \overset{n-1}{\longleftrightarrow} & \overset{1}{\longleftrightarrow} & \\ S_{n-1} & z & \updownarrow\ n-1 \\ 0 & s_{nn} & \updownarrow\ 1 \end{pmatrix}$$

où z et s_{nn} sont encore inconnus.

Nous avons : $$S_n^T S_n = \begin{pmatrix} S_{n-1}^T S_{n-1} & S_{n-1}^T z \\ z^T S_{n-1} & z^T z + s_{nn}^2 \end{pmatrix}$$

Cherchons à identifier $S_n^T S_n$ à la matrice $\mathbf{A}$:

$$c = S_{n-1}^T z \ , \quad \text{et} \quad a_{nn} = z^T z + s_{nn}^2$$

Nous devrons donc prendre :

$$z = (S_{n-1}^T)^{-1} c$$

(ce qui définit z de manière unique).

Montrons qu'alors $a_{nn} - z^T z > 0$

Soit s_{nn} l'une des racines carrées de $(a_{nn} - z^T z)$ (qui, à priori, peuvent être imaginaires).

Par hypothèse : $$S_n = \begin{pmatrix} S_{n-1} & z \\ 0 & s_{nn} \end{pmatrix}$$

et nous voulons avoir :

$$\mathbf{A} = S_n^T S_n$$

Alors : $\det A = (\det S_n)^2 = (\det S_{n-1})^2 . s_{nn}^2$.

Puisque A **est définie positive** $\Rightarrow \det A > 0$; donc $a_{nn} - z^T z = s_{nn}^2 > 0$, et s_{nn} est réel.

Nous choisissons alors pour s_{nn} la racine positive :

$$s_{nn} = +\sqrt{a_{nn} - z^T z}$$

Le lemme est démontré, avec $S_n = S$.

Déterminons effectivement la matrice $S = \begin{pmatrix} s_{11} & \cdots & s_{1n} \\ & \ddots & \vdots \\ 0 & & s_{nn} \end{pmatrix}$

Identifions A et $S^T S$:

$\forall i$ et j , $a_{ij} = \sum_{k=1}^{n} s_{ki}\, s_{kj}$: il suffit d'étudier les termes a_{ij} tels que $i \leqslant j$

- Si $i = j = 1$: $a_{11} = s_{11}^2$ $\qquad s_{11} = \sqrt{a_{11}}$
- Si $i = 1, j > 1$: $a_{1j} = s_{11}\, s_{1j}$ $\qquad s_{1j} = \dfrac{a_{1j}}{s_{11}}$
- Si $i > 1, j = i$: $a_{ii} = \sum_{k=1}^{i} s_{ki}^2$ $\qquad s_{ii} = \sqrt{a_{ii} - \sum_{k=1}^{i-1} s_{ki}^2}$
- Si $i > 1, i < n$: et $n \geqslant j > i$ $\qquad s_{ij} = \dfrac{1}{s_{ii}} \left[a_{ij} - \sum_{k=1}^{i-1} s_{ki}\, s_{kj} \right]$

<u>Nombre d'opérations dans la méthode de Choleski.</u>

- extractions de racines carrées : n
- multiplications : $\sum_{i=2}^{n} (i-1)(n-i+1) = \dfrac{n(n^2-1)}{6} \sim \boxed{\dfrac{n^3}{6}}$ (quand $n \to +\infty$)
- divisions : $\sum_{i=1}^{n} (n-i) = \dfrac{n(n-1)}{2}$
- additions : $\sum_{i=1}^{n} (n-i+1) = \dfrac{n(n-1)}{2}$

Remarques : 1) Le lemme nous assure que les s_{ii} , calculés directement par identification, sont réels positifs.

2) Le lemme est étroitement lié à certains résultats concernant la factorisation des matrices en produit de matrices triangulaires.

(cf. le paragraphe sur la factorisation des matrices, ci-après).

Application : Lorsque l'on a obtenu S telle que

$$A = S^T S ,$$

la résolution de $Ax = b$ se ramène à celle de deux systèmes triangulaires :

$$Ax = b \Longleftrightarrow \begin{cases} S^T.y = b \\ \\ S.x = y \end{cases}$$

On résout d'abord le premier, puis le second système.

5) Factorisation des matrices.

a) Factorisation de la forme A = L.U.

Lorsqu'on résoud l'équation $Ax = b$ par la méthode de Gauss sans permutation de lignes, c'est-à-dire lorsque tous les pivots sont non nuls on écrit :

$$J^{(n-1)} \times \ldots \times J^{(1)} \times Ax = J^{(n-1)} \ldots J^{(1)} . b$$

$$A^{(n)} = J^{(n-1)} \ldots J^{(1)} A$$

$A^{(n)}$ est triangulaire supérieure

$$J = J^{(n-1)} \ldots J^{(1)}$$

$$A = J^{-1} A^{(n)}$$

J étant un produit de matrices triangulaires inférieures, est aussi triangulaire inférieure. J^{-1} l'est donc aussi.

$A^{(n)}$ est triangulaire supérieure. Dans ce cas, on a donc écrit $A = J^{-1} A^{(n)}$ où J^{-1} est triangulaire inférieure et $A^{(n)}$ triangulaire supérieure.

Définition : Soit A une matrice carrée d'ordre n. On dit que A admet une décomposition du type $A = L.U.$ s'il existe une matrice L triangulaire inférieure et une matrice U triangulaire supérieure telles que $A = L.U.$

Définition : Soit A une matrice d'ordre n, soit p inférieur ou égal à n, on appelle mineur fondamental d'ordre p, de A, le déterminant de A_p où $A_p = (a_{ij})_{1 \leqslant i,j \leqslant p}$.

Proposition : Soit A une matrice d'ordre n, non singulière. Alors A admet une décomposition du type $A = L.U.$ si et seulement si les mineurs fondamentaux de A sont non nuls. Dans ce cas la décomposition est unique si on impose la valeur de L ou U sur la diagonale.

Démonstration. 1°) Condition nécessaire

$$\begin{cases} \det \mathbf{A} \neq 0 \\ \mathbf{A} = \mathbf{L}.\mathbf{U}. \end{cases}$$

$$\det \mathbf{A} \neq 0 \Longrightarrow \det \mathbf{L}.\ \det \mathbf{U} \neq 0 \Longrightarrow \begin{cases} \det \mathbf{L} \neq 0 \\ \det \mathbf{U} \neq 0 \end{cases}$$

$$a_{ij} = \sum_{k=1}^{n} L_{ik}\, U_{kj}$$

$a_{ij} = \sum_{1 \leqslant k \leqslant i} L_{ik}\, U_{kj}$ puisque L est triangulaire inférieure. Soit p : $1 \leqslant p \leqslant n$. Alors si i et j sont inférieurs ou égaux à p on a :

$$a_{ij} = \sum_{1 \leqslant k \leqslant p} L_{ik}\, U_{kj}\ , \quad \text{et donc : } \mathbf{A}_p = \mathbf{L}_p.\mathbf{U}_p$$

$$\det \mathbf{A}_p = \det \mathbf{L}_p.\det \mathbf{U}_p = \Big(\prod_{i=1}^{p} L_{ii}\Big)\Big(\prod_{i=1}^{p} U_{ii}\Big)$$

$$\left.\begin{array}{l} \det \mathbf{L} \neq 0 \Longrightarrow \det \mathbf{L}_p \neq 0 \\ \det \mathbf{U} \neq 0 \Longrightarrow \det \mathbf{U}_p \neq 0 \end{array}\right\} \Longrightarrow \det \mathbf{A}_p \neq 0$$

2°) Condition suffisante

On la démontre par récurrence sur n. On suppose que c'est vérifié jusqu'à l'ordre $n-1$. Soit $\mathbf{A}$ une matrice d'ordre n, non singulière, dont les mineurs fondamentaux sont non nuls

$$\mathbf{A} = \left(\begin{array}{c|c} \mathbf{A}_{n-1} & d \\ \hline c & a_{nn} \end{array}\right)$$

Les mineurs fondamentaux de $\mathbf{A}$ sont non nuls par hypothèse. Par conséquent les mineurs fondamentaux de $\mathbf{A}_{n-1}$ sont non nuls. D'après l'hypothèse de récurrence $\mathbf{A}_{n-1}$ admet une décomposition du type $\mathbf{A}_{n-1} = \mathbf{L}_{n-1}.\mathbf{U}_{n-1}$. On cherche alors une décomposition $\mathbf{A} = \mathbf{L}.\mathbf{U}.$ avec $\mathbf{L} = \left(\begin{array}{c|c} \mathbf{L}_{n-1} & 0 \\ \hline x & \ell_{nn} \end{array}\right)$ $\mathbf{U} = \left(\begin{array}{c|c} \mathbf{U}_{n-1} & y \\ \hline 0 & u_{nn} \end{array}\right)$

On procède par identification des blocs ce qui donne les conditions suivantes :

$$\begin{cases} A_{n-1} = L_{n-1} \cdot U_{n-1} \\ c = x \cdot U_{n-1} \\ d = L_{n-1}\, y \\ a_{nn} = xy + \ell_{nn}\, u_{nn} \end{cases}$$

A_{n-1} étant régulière, L_{n-1} et U_{n-1} le sont aussi. L'équation $c = x \cdot U_{n-1}$ permet de déterminer x de manière unique. On obtient y de la même façon à partir de la troisième équation et ensuite : $\ell_{nn}\, u_{nn} = a_{nn} - xy$ donne ℓ_{nn} et u_{nn} (non unicité) L_{n-1} et U_{n-1} sont déterminées de façon unique si on s'est fixé, par exemple, les valeurs de L_{n-1} sur la diagonale.

Si on fixe la valeur de ℓ_{nn}, non nulle, on obtient u_{nn} de manière unique à partir de la quatrième équation.

Pour $n = 1$ l'hypothèse de récurrence est trivialement vérifiée. Ce qui achève la démonstration.

Remarque : Lorsque A s'écrit $A = L.U$ la méthode de Gauss peut se faire sans permutation de lignes et donne effectivement la factorisation car :

$$\left.\begin{array}{l} \det A_1 \neq 0 \\ \vdots \\ \det A_p \neq 0 \end{array}\right\} \Longleftrightarrow \begin{cases} a_{11} \neq 0 \\ \vdots \\ a_{pp}^{(p)} \neq 0 \end{cases}$$

L'interprétation matricielle de la méthode de Gauss avec permutation de lignes est précisée par la proposition suivante :

Proposition : Soit A une matrice non singulière. Alors il existe une matrice de permutation P telle que PA soit factorisable sous la forme $PA = L.U$

(L triangulaire inférieure et U triangulaire supérieure). Toute méthode de Gauss avec permutation de lignes équivaut à un certain choix de P.

Pour les matrices de permutation : voir les rappels sur les matrices de permutation un paragraphe suivant.

Démonstration de la proposition : On la fait par récurrence sur n.

Pour $n = 1$ c'est évident.

On suppose la proposition vraie jusqu'à $n-1$.

Soit $A = (a_{ij})_{\substack{1\leqslant i\leqslant n \\ 1\leqslant j\leqslant n}}$ non singulière.

Si $a_{11} \neq 0$ on pose $\begin{cases} Q = I \\ A' = QA = A \end{cases}$

Si $a_{11} = 0$ A étant non singulière, il existe un indice i tel que $a_{i1} \neq 0$.

On échange la première et la i^e ligne. Il existe donc une matrice de permutation

Q telle que $\begin{cases} A' = QA \\ a'_{11} \neq 0 \end{cases}$

On applique le début de la méthode de Gauss à A' : On fait apparaître des 0 dans la première colonne de A'.

$$JA' = \left(\begin{array}{c|c} a_{11} & a'_{1n} \\ \hline 0 & \\ \vdots & \tilde{A} \\ 0 & \end{array}\right)$$

Ensuite on travaille sur $\tilde{A}$

$$\left.\begin{array}{l} \det A = a'_{11} \cdot \det \tilde{A} \neq 0 \\ a'_{11} \neq 0 \end{array}\right| \Longrightarrow \det \tilde{A} \neq 0$$

On applique l'hypothèse de récurrence à $\tilde{A}$. Il existe $\tilde{P}$ $\tilde{L}$ et $\tilde{U}$ tels que

$\tilde{P}\tilde{A} = \tilde{L}\tilde{U}$

$\tilde{L}$ = matrice triangulaire inférieure

$\tilde{U}$ = matrice triangulaire supérieure

$\Longrightarrow \tilde{A} = \tilde{P}^t \tilde{L} \tilde{U}$.

On pose $P_1 = \left(\begin{array}{c|ccc} 1 & 0 & \cdots & 0 \\ \hline 0 & & & \\ \vdots & & \tilde{P} & \\ 0 & & & \end{array}\right)$

P_1 est une matrice de permutation d'ordre n

$L_1 = \left(\begin{array}{c|ccc} 1 & 0 & \cdots & 0 \\ \hline 0 & & & \\ \vdots & & \tilde{L} & \\ 0 & & & \end{array}\right)$ est triangulaire inférieure

$U_1 = \left(\begin{array}{c|ccc} a'_{11} & \cdots & & a'_{1n} \\ \hline 0 & & & \\ \vdots & & \tilde{U} & \\ 0 & & & \end{array}\right)$ est triangulaire supérieure

$P_1 JA' = \left(\begin{array}{c|ccc} a'_{11} & & & a'_{1n} \\ \hline 0 & & & \\ \vdots & & \tilde{P}\,\tilde{A} & \\ 0 & & & \end{array}\right) = L_1 U_1$

$P_1 J Q A = L_1 U_1$

$P_1 J P_1^t P_1 Q A = L_1 U_1$

$P = P_1 Q$ est une matrice de permutation

$J' = P_1 J P_1^t$

$$J' = \left(\begin{array}{c|c} 1 & 0.....0 \\ \hline \begin{matrix}0\\ \vdots\\ \vdots\\ 0\end{matrix} & \tilde{P} \end{array}\right) \left(\begin{array}{c|c} 1 & 0 \\ \hline \alpha & I_{n-1} \end{array}\right) \left(\begin{array}{c|c} 1 & 0.....0 \\ \hline \begin{matrix}0\\ \vdots\\ \vdots\\ 0\end{matrix} & \tilde{P}^t \end{array}\right) = \left(\begin{array}{c|c} 1 & 0.....0 \\ \hline \tilde{\tilde{P}}_\alpha & I_{n-1} \end{array}\right)$$

qui est triangulaire inférieure

J' est triangulaire inférieure ; donc $J'^{-1} L_1$ l'est aussi.

$PA = J'^{-1} L_1 . U_1$ ce qui achève la démonstration.

b) Factorisation de la forme $A = Q.U$ ← triang. sup.
↑ orthogonale

Dans la méthode de Householder

$A^{(n)} = H^{(n-1)} \ldots H^{(1)} A$ est triangulaire supérieure

$H = H^{(n-1)} \ldots H^{(1)}$ est orthogonale

$A = H^{-1} A^{(n)}$.

Définition : Soit A une matrice d'ordre n ; on dit que A est factorisable sous la forme $A = Q.U$ si il existe une matrice Q orthogonale et une matrice U triangulaire supérieure telles que $A = Q.U$.

Proposition : Soit A non singulière. Alors A est factorisable sous la forme $A = Q.U$ avec Q orthogonale et U triangulaire supérieure. La décomposition est essentiellement unique : Si $Q_1 U_1 = Q_2 U_2 = A$ sont deux telles factorisations alors il existe une matrice diagonale D telle que

$$\begin{cases} dii = \pm 1 \\ U_2 = D U_1 \\ Q_1 = D Q_2 \end{cases}$$

(En complexe on aurait $|dii| = 1$).

Démonstration : L'existence est assurée par la méthode de Householder (ou la méthode des rotations).

Si on a deux factorisations $A = Q_1 \quad U_1 = Q_2 \, U_2$.

On pose $D = Q_2^T \, Q_1 = U_2 \, U_1^{-1}$.

D est orthogonale et triangulaire supérieure.

$D^T = D^{-1}$.

D^T est triangulaire inférieure et D^{-1} est triangulaire supérieure ; donc D est diagonale et comme D est orthogonale alors $dii = \pm 1$ ce qui achève la démonstration.

Remarques. 1°) Lorsque A s'écrit $A = L.U$. La résolution de $Ax = b$ équivaut aux deux résolutions suivantes $\begin{cases} Ly = b \ (1) \\ Ux = y \ (2) \end{cases}$

On résoud d'abord (1) ce qui est facile puisque L est triangulaire. Connaissant y on résoud alors (2) ; d'où l'intérêt de factoriser A sous cette forme. On sait que si A est non singulière il existe une matrice P telle que $PA = L.U$.

Le système $Ax = b$ équivaut à $PAx = Pb$.

2°) Soit A non singulière. On sait que A est factorisable sous la forme $A = QU$. Le système $Ax = b$ s'écrit $QUx = b$ ce qui revient à résoudre deux équations simples :

$$\begin{cases} Qy = b \\ Ux = y \end{cases} \iff \begin{cases} y = Q^{-1}b \\ Ux = y \end{cases}$$

La deuxième équation est facile à résoudre car u est triangulaire supérieure. D'où l'intérêt de cette factorisation.

3°) On verra d'autres applications de la factorisation pour les problèmes de valeurs propres.

§II. Compléments d'algèbre linéaire.

1) Matrices irréductibles.

Rappels sur les matrices de permutation.

- Soit P une matrice de permutation d'ordre n . $P = (p_{ij})$. Il existe une permutation σ de l'ensemble $[1,2,\ldots,n]$ telle que $P_{ij} = \delta_{\sigma(i),j}$.

- Propriétés des matrices de permutation

. P est orthogonale. En effet :

$$(P^T P)_{i,j} = \sum_{k=1}^{n} p_{k,i}\, p_{k,j} = \sum_{k=1}^{n} \delta_{\sigma(k),i}\ \delta_{\sigma(k),j} = \delta_{i,j}$$

. Le produit de deux matrices de permutation est une matrice de permutation

$$(PQ)_{i,j} = \sum_{k=1}^{n} p_{ik}\, q_{kj} = \sum_{k=1}^{n} \delta_{\sigma(i),k}\ \delta_{\tau(k),j} = \delta_{\tau(\sigma(i)),j}$$

. Si $A = \begin{bmatrix} L_1 \\ \vdots \\ L_n \end{bmatrix}$ $\qquad L_i$ = i-ème ligne de A

$A = [c_1, c_2, \ldots, c_n]$ $\quad c_i$ = i-ème colonne de A

$P = (\delta_{\sigma(i),j})$

$$(PA)_{ij} = \sum_{k=1}^{n} \delta_{\sigma(i),k}\ a_{kj} = a_{\sigma(i),j}$$

$$\Longrightarrow PA = \begin{bmatrix} L_{\sigma(1)} \\ \vdots \\ L_{\sigma(n)} \end{bmatrix}$$

$$(AP^t)_{ij} = \sum_{k=1}^{n} \delta_{k\,\sigma(j)}\ a_{ik} = a_{i,\sigma(j)}$$

$$\Longrightarrow AP^t = [c_{\sigma(1)}, \ldots, c_{\sigma(n)}]$$

$$(PA\ P^t)_{ij} = a_{\sigma(i),\sigma(j)}$$

Définition : Une matrice A d'ordre n est dite réductible si il existe une matrice de permutation P telle que $A' = PA\ P^t$ soit du type suivant :

$$\begin{pmatrix} A'_{11} & A'_{12} \\ 0 & A'_{22} \end{pmatrix} \begin{matrix} \updownarrow p \\ \updownarrow n-p \end{matrix}$$

$$\begin{matrix} \leftrightarrow & \leftrightarrow \\ p & n-p \end{matrix}$$

Justification de la définition : quand on a à résoudre $Ax = b$ où A est réductible

$$x = \begin{pmatrix} z \\ - \\ t \end{pmatrix} \begin{matrix} \updownarrow p \\ \updownarrow n-p \end{matrix} \qquad b = \begin{pmatrix} d \\ - \\ e \end{pmatrix} \begin{matrix} \updownarrow p \\ \updownarrow n-p \end{matrix}$$

Alors $Ax = b$ équivaut à :

$$\begin{cases} A'_{22}\, t = e \\ A'_{11}\, z = d - A'_{12}\, t \end{cases}$$

On est donc amené à résoudre un système d'ordre p et ensuite un système d'ordre $n-p$

Lemme : A est réductible si et seulement si il existe une partition de l'ensemble $[1,2,\ldots n]$ en $I \cup J$ telle que $a_{ij} = 0$ pour tout (i,j) de $I \times J$.

Démonstration : 1°) Condition suffisante

$\sigma : [1,2,\ldots n] \to (J,I)$ est une permutation de matrice P

$A' = PA\ P^t$, $a'_{ij} = a_{\sigma(i),\sigma(j)} = 0$ si $\sigma(i) \in I$, $\sigma(j) \in J$

c'est-à-dire $p+1 \leqslant i \leqslant n \quad 1 \leqslant j \leqslant p$.

2°) Condition nécessaire

$A' = PA\,P^t$

$$a'_{ij} = \sum_{1 \leqslant k,\ell \leqslant n} p_{ik}\, a_{k\ell}\, p_{j\ell} = \sum_{1 \leqslant k,\ell \leqslant n} \delta_{\sigma(i),k}\, a_{k,\ell}\, \delta_{\sigma(j)\ell}$$

$$a'_{ij} = a_{\sigma(i),\sigma(j)} = 0 \quad \text{si} \quad \begin{cases} 1 \leqslant j \leqslant p \\ p+1 \leqslant i \leqslant n \end{cases}$$

$I = \sigma([\,p+1,n])$

$J = \sigma([\,1,p])$

I et J forment une partition de $[1,2,\ldots n]$ telle que $a_{ij} = 0$ dès que (i,j) appartient à $I \times J$ ce qui achève la démonstration.

Définition : Une matrice A est dite irréductible si elle n'est pas réductible.

Définition : Soit A une matrice d'ordre n . On dira qu'il existe une chaîne joignant i et j si il existe une suite d'indices (compris entre 1 et n) $k_1 \ldots k_p$ tels que :

$$a_{i,k_1} \neq 0$$
$$a_{k_1,k_2} \neq 0$$
$$\vdots$$
$$a_{k_p,j} \neq 0$$

Lemme : (Caractérisation des matrices irréductibles). A est irréductible si et seulement si pour tout (i,j) de $[1,2,\ldots,n] \times [1,2,\ldots,n]$ il existe une chaîne joignant i et j .

Démonstration : 1°) Condition nécessaire

Soit α dans $[1,2,\ldots,n]$.

$K = \{\beta \in [1,2,\ldots,n] \mid$ il existe une chaîne joignant β et $\alpha\}$.

K est non vide car il y a au moins un indice j tel que $a_{\alpha,j} \neq 0$ (sinon la α-ème ligne de A serait nulle et A serait réductible).

Si $K \neq [1,2,\ldots,n]$ on pose $\begin{cases} I = K \\ J = \complement K \end{cases}$

I et J forment une partition de $[1,2,\ldots,n]$.

Soit (i,j) dans $I \times J$; si $a_{ij} \neq 0$ on aurait j dans K ce qui est impossible. Donc $a_{ij} = 0$. Donc A serait réductible.

L'hypothèse $K \neq [1,2,\ldots,n]$ est donc absurde.

2°) Condition suffisante

Supposons A réductible ; il existe I et J formant une partition de $[1,2,\ldots,n]$ telle que $a_{ij} = 0$ dès que (i,j) est dans $I \times J$.

Soient i dans I et j dans J ; supposons qu'il existe une chaîne joignant i et j

$a_{i,k_1} \neq 0 \ldots, a_{k_{p-1},k_p} \neq 0 , a_{k_p,j} \neq 0$

$a_{i,k_1} \neq 0 \Longrightarrow k_1 \in I$

$a_{k_1,k_2} \neq 0 \Longrightarrow k_2 \in I$

$\vdots$

$k_p \in I$

$a_{k_p,j} \neq 0 \Longrightarrow k_p \in J$. On aboutit à une contradiction.

2) Matrices irréductibles et valeurs propres.

Théorème de Gerschgörin : Soit A une matrice d'ordre n . Les valeurs propres de A sont situées dans la réunion des disques suivants (dits de Gerschgörin) :

$$\begin{cases} |z - a_{ii}| \leqslant \Lambda_i \\ \Lambda_i = \sum\limits_{\substack{1 \leqslant j \leqslant n \\ j \neq i}} |a_{ij}| \end{cases}$$

Démonstration.

Soit X un vecteur propre associé à la valeur propre λ, tel que

$$\max_i |x_i| = 1 = |x_k| .$$

$$Ax = \lambda x$$

$$\sum_{j=1}^{n} a_{ij}\, x_j = \lambda\, x_i$$

$$(\lambda - a_{kk})x_k = \sum_{\substack{1\leqslant j\leqslant n \\ j\neq k}} a_{kj}\, x_j$$

$$|\lambda - a_{kk}|.|x_k| = |\lambda - a_{kk}| \leqslant \sum_{\substack{1\leqslant j\leqslant n \\ j\neq k}} |a_{kj}| = \Lambda_k .$$

Définition : La réunion des disques $\begin{cases} |z - a_{ii}| \leqslant \Lambda_i \\ 1 \leqslant i \leqslant n \end{cases}$ s'appelle la région de Gerschgörin.

Théorème : Soit A une matrice d'ordre n, irréductible. Si une valeur

(*) propre λ est située sur la frontière de la région de Gerschgörin, alors tous les cercles de Gerschgörin passent par λ.

Soit X un vecteur propre associé à λ, tel que $\max_i |x_i| = 1$; alors s'il existe un indice i tel que $|\lambda - a_{ii}| < \Lambda_i$, λ est intérieur à la région de Gerschgörin ce qui contredit l'hypothèse. Donc $|\lambda - a_{ii}| \geqslant \Lambda_i$, $1 \leqslant i \leqslant n$.

Soit $I = \{i \mid |x_i| = 1\}$.

$$i \in I \Longrightarrow |\lambda - a_{ii}||x_i| = |\lambda - a_{ii}| \leqslant \Lambda_i$$

$$\Longrightarrow |\lambda - a_{ii}| = \Lambda_i$$

On veut montrer que $I = [1,2,\ldots,n]$.

On pose $J = p_I$; si $J \neq \emptyset$, I et J forment une partition de $[1,2,\ldots,n]$

$$i \in I \qquad \begin{cases} |\lambda - a_{ii}| = |\sum_{j\neq i} a_{ij}\, x_j| \\ |\lambda - a_{ii}| = \Lambda_i \end{cases}$$

$$\left.\begin{array}{l}\displaystyle\sum_{\substack{1\leqslant j\leqslant n\\ j\neq i}} |a_{ij}|\,(1-|x_j|) \leqslant 0 \\ \\ \displaystyle\max_i |x_i| = 1\end{array}\right\} \Longrightarrow \sum_{\substack{1\leqslant j\leqslant n\\ j\neq i}} |a_{ij}|\,(1-|x_j|) = 0$$

$$\Longrightarrow \forall j \neq i \quad \text{on a} \quad |a_{ij}|\,(1-|x_j|) = 0$$

$$\Longrightarrow \left(\begin{array}{l} i \in I \\ \\ j \in J \end{array}\right] \Longrightarrow a_{ij} = 0\Bigg)$$

A serait donc réductible ce qui est contraire à l'hypothèse. Donc

$I = [1,2,\ldots,n]$ et pour tout i $|\lambda - a_{ii}| = \Lambda_i$ ce qui achève la démonstration.

Définitions.

Soit **A** une matrice carrée d'ordre n.

On dit que :

A est à diagonale dominante si

$$\forall i \qquad |a_{ii}| \geqslant \sum_{\substack{1\leqslant j\leqslant n\\ j\neq i}} |a_{ij}| = \Lambda_i$$

A est à diagonale strictement dominante si

$$\forall i \qquad |a_{ii}| > \sum_{\substack{1\leqslant j\leqslant n\\ j\neq i}} |a_{ij}| = \Lambda_i$$

A est à diagonale fortement dominante si

A étant à diagonale dominante, il existe au moins un i_o tel que l'on ait, pour cet i_o l'inégalité stricte ; c'est-à-dire :

$$\begin{cases} \forall i \qquad |a_{ii}| \geqslant \displaystyle\sum_{\substack{1\leqslant j\leqslant n\\ j\neq i}} |a_{ij}| = \Lambda_i \\ \\ \exists i_o \quad \text{tel que} \quad |a_{i_o i_o}| > \Lambda_{i_o} \end{cases}$$

Proposition.

a) Si **A** est à diagonale strictement dominante, **A** est inversible.

b) Si **A** est à diagonale fortement dominante et irréductible, **A** est inversible.

De plus :

c) Dans les deux cas, si λ est valeur propre de **A** et si tous les éléments a_{ii} de la diagonale sont strictement positifs, alors la partie réelle de λ est strictement positive.

Démonstration.

a) Démontrons le a) de la proposition.

Supposons que **A** soit à diagonale strictement dominante. Il en résulte que :

$$\forall i \qquad |0 - a_{ii}| = |a_{ii}| > \Lambda_i$$

0 n'appartient donc pas à la région de Gerschgörin. D'après le théorème de Gerschgörin 0 n'est pas valeur propre de **A** . Ce qui signifie que **A** est inversible. $(\lambda = 0 \qquad \det(\lambda I - \mathbf{A}) = \det \mathbf{A} \neq 0)$.

b) Démontrons le b) de la proposition.

Si **A** est à diagonale fortement dominante, c'est-à-dire :

$$\begin{cases} \forall i \qquad |0 - a_{ii}| \geqslant \Lambda_i \\ \exists i_o \quad \text{tel que} \quad |0 - a_{i_o i_o}| > \Lambda_{i_o} \end{cases}$$

et si 0 est valeur propre, 0 appartient à la frontière de la région de Gerschgörin.

Comme **A** est irréductible, par hypothèse, d'après le théorème (p.30) *, tous les cercles de Gerschgörin passent par 0 . Donc $\forall i \quad |0 - a_{ii}| = \Lambda_i$.

Ce qui est impossible, pour $i = i_o$.

Donc 0 n'est pas valeur propre de **A** , **A** est inversible.

c) Démontrons, enfin, le c) de la proposition. Montrons que :

- Si $\forall i \quad a_{ii} > 0$ alors $\forall \lambda$ valeur propre de **A** , $\mathrm{Re}\ \lambda > 0$.

λ étant valeur propre de **A** , est dans la région de Gerschgörin. Les cercles de Gerschgörin centrés en a_{ii} $(a_{ii} > 0)$, de rayon Λ_i $(\Lambda_i \leqslant a_{ii})$ sont situés dans le demi plan $\mathrm{Re}\ z \geqslant 0$; ils ne rencontrent pas l'axe $\mathrm{Re}\ z = 0$ si ce n'est au point $z = 0$.

Comme 0 n'est pas valeur propre de **A** , la partie réelle de λ, λ valeur propre de **A** , est strictement positive.

Corollaire : Soit **A** une matrice carrée d'ordre n , hermitienne, telle que tous les éléments de la diagonale de **A** soient strictement positifs.

Si **A** est à diagonale strictement dominante (resp. fortement dominante et irréductible), **A** est définie positive.

Démontrons ce corollaire. D'après la proposition précédente : $\mathrm{Re}\ \lambda > 0$, λ valeur propre de **A** .

A étant hermitienne, toute valeur propre λ est réelle et donc λ est strictement positive, ceci est une condition nécessaire et suffisante pour que **A** soit définie positive.

3) Matrices à éléments positifs.

Définitions.

Soit x un vecteur de R^n $\quad x = \{x_1, \ldots, x_n\}$ on dit que :

x est positif (noté $x \geqslant 0$) si toutes ses composantes x_i sont positives ou nulles.

x est supérieur à y (noté $x \geqslant y$), si x-y est positif

(Remarque : C'est une relation d'ordre non total).

x est strictement positif (noté $x > 0$), si toutes ses composantes sont strictement positives.

De même on dit que x est strictement supérieur à y . Si x-y est strictement positif.

Soit A une matrice de $\mathbb{R}^{n \times n}$.

On dit que :

A est positive (noté $A \geqslant 0$) si tous ses éléments (a_{ij}) $(i = 1,\ldots n$, $j = 1,\ldots n)$ sont positifs ou nuls.

A est supérieure à B si A-B est positive.

A est strictement positive (noté $A > 0$) si tous ses éléments (a_{ij}) sont strictement positifs.

On dit de même que $(A > B)$ si $(A-B > 0)$.

Notations.

Soit x , vecteur de R^n $x = \{x_1,\ldots x_n\}$.

On note $|X|$ le vecteur de R^n ayant pour composante :

$$|X|_i = |X_i|$$

Soit A , matrice de $R^{n \times n}$.

On note $|A|$ la matrice donc les éléments sont égaux à la valeur absolue de ceux de A $\qquad |A|_{ij} = |a_{ij}|$.

Lemme : Soit A une matrice carrée d'ordre n, sur R.

i) A est positive si, et seulement, si Ax est positif, pour tout x positif

$$A \geqslant 0 \Longleftrightarrow Ax \geqslant 0 \quad \forall x \in R^n , x \geqslant 0 .$$

ii) A est strictement positive si, et seulement, si Ax est strictement positif, quel que soit x vecteur de R^n, positif et non nul

$$A > 0 \Longleftrightarrow Ax > 0 \quad \forall x \geqslant 0 , x \neq 0 .$$

Démontrons le lemme.

i) Soit $(e_1,\ldots,e_n)$ la base canonique de R^n.

Soit x un vecteur de R^n :

on a : $x = \sum_{j=1}^{n} x_j e_j$

$$Ae_j = \begin{bmatrix} a_{1j} \\ \vdots \\ a_{nj} \end{bmatrix} = j^{\text{ième}} \text{colonne de } A .$$

- Montrons que la condition est nécessaire :

Si A est positive, Ae_j est positif, quel que soit j :

donc : $Ax = \sum_{j=1}^{n} x_j Ae_j$ est positif si x est positif.

- Montrons que la condition est suffisante :

Supposons donc que Ax est positif, quel que soit x positif

e_j est un vecteur de R^n positif, donc, on a en particulier Ae_j positif, ceci quel que soit j.

A est positive.

ii) On a : $\mathbf{A}x = \begin{bmatrix} \sum_{j=1}^{n} a_{1j} x_j \\ \vdots \\ \sum_{j=1}^{n} a_{nj} x_j \end{bmatrix}$

Montrons que la condition est nécessaire.

Si $\mathbf{A}$ est strictement positive, et si x est positif non nul, il existe au moins un j tel que x_j ne soit pas nul, donc que toutes les composantes de $\mathbf{A}x$ ne soient pas nulles.

Inversement, si on choisit $x = e_j$ on a $\mathbf{A}e_j$ strictement positive par hypothèse, et ceci pour j variant de 1 à n. Donc $\mathbf{A}$ est strictement positive.

Lemme : Soit A matrice carrée d'ordre n à éléments réels. Si A est positive et irréductible alors $(I+\mathbf{A})^{n-1}$ est strictement positive.

Démonstration. $\mathbf{A}$ étant positive $(I+\mathbf{A})^{n-1}$ est positive.

Montrons que $(I+\mathbf{A})^{n-1}$ est strictement positive.

D'après le lemme précédent il nous suffit de montrer que :

$$\forall x \geqslant 0 \quad x \neq 0 \quad (I+\mathbf{A})^{n-1} x > 0$$

Posons $x_o = x \qquad x \in R^n \quad x \geqslant 0 \quad x \neq 0$.

$$x^1 = (I+\mathbf{A})x_o$$

$$x^{(k)} = (I+\mathbf{A})^{(k)}x_o$$

Montrons que $x^{(k+1)}$ a moins de composantes nulles que $x^{(k)}$.

Supposons que $x^{(k)}$ a r composantes nulles. Grâce à une matrice P de permutation, on peut faire en sorte que les composantes nulles soient regroupées, c'est-à-dire :

$$Px^{(k)} = \begin{bmatrix} y^{(k)} \\ \hline 0 \end{bmatrix} \Big]r$$

on a :

$$Px^{(k+1)} = P(I+A)x^{(k)} = Px^{(k)} + PAx^{(k)}$$

$$= Px^{(k)} + PAP^T\, Px^{(k)}$$

$$= Px^{(k)} + A'\, x^{(k)} \quad \text{où} \quad A' = PAP^T = \left[\begin{array}{c|c} A'_{11} & A'_{12} \\ \hline A'_{21} & A'_{22} \end{array}\right]$$

$$= \begin{bmatrix} y^{(k)} + A'_{11}\, y^{(k)} \\ \hline A'_{21}\, y^{(k)} \end{bmatrix}$$

Par hypothèse on a $y^{(k)} + A'_{11}\, y^{(k)} \geqslant y^{(k)} > 0$.

Supposons que $A'_{21}\, y^{(k)}$ soit nul. Ceci entraîne que A'_{21} est nulle, ce qui contredit le fait que A est irréductible.

$Px^{(k+1)}$ a au plus $r-1$ composantes nulles, donc il en est de même pour $x^{(k+1)}$. Comme $x^{(o)}$ a au plus $(n-1)$ composantes nulles $x^{(k)}$ a au plus $n-1-k$ composantes nulles. Donc $x^{(n-1)} = (I+A)^{n-1}x_o$ n'a pas de composante nulle.

$$\forall x \geqslant 0\,,\ x \neq 0\,,\ (I+A)^{n-1}\, x > 0 \ .$$

4) Théorème de Perron-Frobénius.

Théorème : Soit A une matrice de $\mathbb{R}_{(n,n)}$ positive, irréductible.

i) $\rho(A)$ est une valeur propre simple de A à laquelle correspond un vecteur propre strictement positif.

ii) $\rho(A)$ est une fonction strictement croissante des a_{ij} et plus précisément :

Si B est une matrice complexe d'ordre n , telle que : $|B| \leqslant A$ alors $\rho(B) \leqslant \rho(A)$ et $\rho(B) < \rho(A)$ si $|B| \neq A$.

Rappel.

$\rho(A)$ est par définition le rayon spectral de A si $(\lambda_i)_{i\in I}$ sont valeurs propres de A, on a :

$$\rho(A) = \max_i |\lambda_i| .$$

La démonstration du théorème de Perron-Frobenius découle d'une série de lemmes que nous allons énoncer et démontrer.

Auparavant nous définissons $r(x) = r(x,A)$:

Si x est un vecteur de R^n, positif, non nul, on pose :

$$r(x) = \sup_{\substack{\rho \geqslant 0 \\ Ax \geqslant \rho x}} \rho$$

L'ensemble E des nombres ρ de R, positifs ou nuls tel que Ax soit supérieur à ρx est compact.

$$E = \{\rho \mid \rho \geqslant 0 \quad Ax \geqslant \rho(x)\}$$

E est évidemment fermé. Montrons qu'il est borné. x étant non nul, il a au moins une composante non nulle. Soit x_i cette composante. On a :

$$(Ax)_i \geqslant \rho x_i$$

$$\rho \leqslant \frac{(Ax)_i}{x_i}$$

Le sup appartient donc à E. C'est en fait un maximum. On a donc :

$$Ax \geqslant r(x)x .$$

De plus on a :

(1)
$$r(x) = \min_{x_i \neq 0} \frac{\sum_{j=1}^{n} a_{ij} x_j}{x_i} = \min_{x_i \neq 0} \frac{(Ax)_i}{x_i}$$

Démontrons cette égalité.

Posons $\underset{x_i \neq 0}{\text{Min}} \dfrac{(Ax)_i}{x_i} = k$ et montrons que $r(x) = k$.

On a :

$$\begin{cases} \forall j & (Ax)_j \geqslant kx_j \\ \exists i \text{ tel que} & Ax_i = kx_i \end{cases} \qquad (1)$$

Donc k est un élément de E.

Par définition $r(x)$ est supérieur ou égal à k. Supposons que $r(x)$ est strictement supérieur à k. On a alors :

$$\forall j \qquad (Ax)_j \geqslant r(x)x_j > kx_j$$

Ce qui contredit (1). Donc $r(x)$ égal à k.

Nous définissons maintenant le nombre $r = r(A)$ par :

(2)
$$\boxed{r = \sup_{\substack{x \neq 0 \\ x \geqslant 0}} r(x)}$$

Si α est non nul on a $r(\alpha x) = r(x)$.

En effet

$$r(\alpha x) = \underset{\alpha\, x_i \neq 0}{\text{Min}} \frac{\sum_{j=1}^{n} a_{ij}\, \alpha x_j}{\alpha x_i} = \underset{x_i \neq 0}{\text{Min}} \frac{\sum_{j=1}^{n} a_{ij}\, x_j}{x_i} = r(x)$$

Donc $r(x) = r\left(\dfrac{x}{\|x\|}\right)$ et

$$\boxed{r = \sup_{\substack{y \neq 0 \\ y \geqslant 0}} r(y) = \sup_{\substack{y \geqslant 0 \\ \|y\| = 1}} r(y)}$$

On a plus précisément :

Lemme 1 :

(3) $$\boxed{r = \sup_{\substack{x\geqslant 0\\ x\neq 0}} r((I+A)^{n-1}x) = \sup_{\substack{x\geqslant 0\\ \|x\|=1}} r((I+A)^{n-1}x)}$$

Démontrons ce lemme :

on sait que, pour x positif, non nul on a :

$$Ax \geqslant r(x)x .$$

Donc :

$$(I+A)^{n-1}\,Ax = A(I+A)^{n-1}x \geqslant r(x)(I+A)^{n-1}x .$$

Posons $y = (I+A)^{n-1}x$

Il en résulte

$$Ay \geqslant r(x)y .$$

Par définition $r(y)$ est le plus grand ρ tel que Ay soit supérieur ou égal à ρy . Donc on a :

$$r(y) \geqslant r(x).$$

On a :

$$\sup_{\substack{x\geqslant 0\\ x\neq 0}} r((I+A)^{n-1}x) = \sup_{y=(I+A)^{n-1}x} r(y) \geqslant \sup_{\substack{x\geqslant 0\\ x\neq 0}} r(x) = r .$$

Montrons l'inégalité dans le sens contraire.

On a vu que, x étant un vecteur positif, non nul, $(I+A)^{n-1}x$ est strictement positif.

Donc on a l'inégalité suivante :

$$r \geqslant \sup_{\substack{x\geqslant 0\\ x\neq 0}} r((I+A)^{n-1}x)$$

Ce qui entraîne l'égalité

$$r = \sup_{\substack{x \geqslant 0 \\ x \neq 0}} r((I+A)^{n-1}x)$$

Lemme 2 : Soit A une matrice irréductible positive : r est fini et il existe un vecteur z de R^n, $z > 0$ tel que l'on ait : $Az \geqslant rz$ (ce qui signifie que $r = r(z)$).

Démonstration. D'après le lemme 1 on a : $r = \sup_{\substack{x \geqslant 0 \\ \|x\| = 1}} r((I+A)^{n-1}x)$.

La fonction qui, à x, fait correspondre $(I+A)^{n-1}x$ est continue.

Il en est de même de celle qui, à y, fait correspondre $r(y)$, comme on peut le voir sur la formule (1).

La fonction $x \to r((I+A)^{n-1}x)$ étant la composée de deux fonctions continues est continue. Elle atteint donc son maximum sur le compact $\|x\| = 1 \quad x \geqslant 0$:

$$\exists x_o \geqslant 0 \quad \|x_o\| = 1 \text{ tel que } r = r((I+A)^{n-1}x_o).$$

Si nous posons $z = (I+A)^{n-1}x_o$, z vérifie bien la proposition. En effet on a vu que x_o étant un vecteur positif non nul, $(I+A)^{n-1}x_o$ est strictement positif et on a :

$$r = r(z).$$

Ce qui achève la démonstration du deuxième lemme.

Remarque : Un vecteur z positif, non nul, tel que $Az \geqslant rz$ est dit extrémal.

Cette dénomination résulte de ce qu'il n'existe pas de ξ, positif, non nul, tel que l'on ait :

$$A\xi > r\xi .$$

En effet on aurait alors : $r(\xi) > r$, ce qui est impossible.

Nous allons voir que $Az = rz$ lorsque le vecteur z positif est extrémal.

Lemme 3 : Soit A une matrice irréductible, positive ; r est alors une valeur propre de A et r est strictement positif.

Tout vecteur z positif, non nul, extrémal est un vecteur propre associé à r .

De plus z est alors strictement positif.

Démonstration.

D'après le lemme 2 on sait que :

$$\exists z > 0 \quad \text{tel que} \quad Az \geqslant rz$$

Il en résulte, qu'il existe η , vecteur positif tel que l'on ait :

$$Az = rz + \eta \ .$$

Supposons que η ne soit pas nul.

Comme $(I+A)^{n-1}$ est strictement positif, on a :

$$(I+A)^{n-1}\eta > 0$$

De la relation :

$$(I+A)^{n-1} Az = r(I+A)^{n-1} z + (I+A)^{n-1}\eta$$

on déduit donc la relation suivante :

$$(1) \qquad (I+A)^{n-1} Az > r(I+A)^{n-1}z \ .$$

Posons $z' = (I+A)^{n-1}z$.

De (1) il résulte que :

$$Az' > rz'.$$

Ce qui est impossible, d'après la remarque précédente (p.41) ; η est donc

nul, et on a $Az = rz$.

r est donc bien valeur propre de A , et z , vecteur extrémal est un vecteur propre associé à r .

Montrons, de plus, que tout vecteur z extrémal est strictement positif.

On a :

$$(I+A)^{n-1}Az = r(I+A)^{n-1}z = r(I+A)^{n-2}(I+A)z = r(I+A)^{n-2}(1+r)z$$

En raisonnant par récurrence, on obtient :

$$A(I+A)^{n-1}z = r(1+r)^{n-1}z \ .$$

z étant positif, non nul, $(I+A)^{n-1}z$ est strictement positif ainsi que $A(I+A)^{n-1}z$.

Pour que $r(1+r)^{n-1}z$ soit strictement positif il est nécessaire que z soit strictement positif, ainsi que r .

Ce qui achève la démonstration du lemme 3.

Lemme 4 : Soient A une matrice irréductible, positive de $R_{(n\times n)}$. B une matrice complexe d'ordre n telle que l'on ait :

$$|B| \leqslant A \ .$$

Soit μ une valeur propre de B .

On a alors :

(1) $$|\mu| \leqslant r \ .$$

Si l'égalité a lieu dans (1) on a $|B| = A$.

Démonstration. Soit x un vecteur propre associé à la valeur propre μ .

$$Bx = \mu x \ .$$

Soit :

$$\forall i \qquad \mu x_i = \sum_{j=1}^{n} b_{ij}\, x_j$$

$$(2) \qquad \forall i \qquad |\mu||x_i| \leqslant \sum_{j=1}^{n} |b_{ij}|\,|x_j| \leqslant \sum_{j=1}^{n} a_{ij}|x_j|$$

car par hypothèse $|B|$ est inférieur ou égal à A .

Posons : $y = |x|$ c'est-à-dire, $\forall i$, $y_i = |x_i|$.

On a donc : $y \geqslant 0$ $y \neq 0$ et $|\mu| y \leqslant Ay$.

Par définition de $r(y)$ et de r on a :

$$|\mu| \leqslant r(y) \leqslant r .$$

Supposons que $|\mu| = r$.

On a donc

$$Ay \geqslant ry ,$$

y est extrémal et d'après le lemme 3 on a :

$$Ay = ry$$

On a :

$$ry_i = |\mu||x_i| = \sum_{j=1}^{n} a_{ij}\,|x_j|$$

Soit

$$|\mu||x_i| = \sum_{j=1}^{n} |b_{ij}||x_j| \quad \text{d'après la relation (2) ci-dessus}$$

Ce qui entraîne :

$$\forall i \qquad \sum_{j=1}^{n} |a_{ij} - |b_{ij}\|x_j| = 0 .$$

D'après le lemme 3, y , étant extrémal, est strictement positif ; $|x_j|$ est strictement positif. Donc on a :

$$\forall i , \forall j , \qquad |b_{ij}| = a_{ij} .$$

La matrice A est égale à $|B|$ si $|\mu|$ est égal à r.

Démonstration du théorème de Perron-Frobenius. (cf énoncé p.37).

Ecrivons le lemme 4 dans le cas $B=A$. Si μ est valeur propre de A on a $|\mu| \leqslant r$. Le rayon spectral de A est donc inférieur à r :

$$\rho(A) = \max_i |\mu_i| \leqslant r .$$

Comme r est valeur propre de A (d'après le lemme 3) on a plus précisément :

$$\underline{\rho(A) = r}$$

et à $\rho(A)$ correspond un vecteur propre extrémal z strictement positif.

On a ainsi démontré la partie i) du théorème, mis à part le fait que $\rho(A)$ soit une valeur propre SIMPLE.

Montrons la partie ii).

Si on a : $$|B| \leqslant A$$

d'après le lemme 4 on a :

$$\rho(B) \leqslant r$$

D'après la première partie on a $r = \rho(A)$.

Soit

$$|B| \leqslant A \Longrightarrow \rho(B) \leqslant \rho(A).$$

Si $|B|$ est différent de A, le rayon spectral de B est différent de celui de A.

(Raisonnons par l'absurde).

Si $\rho(B) = \rho(A)$ d'après le lemme 4 on a $|B| = A$ ce qui contredit l'hypothèse).

Il ne nous reste plus, pour achever la démonstration du théorème de Perron-Frobenius, qu'à montrer que $\rho(\mathbf{A})$ est une valeur propre simple. Ce que nous ferons au lemme 6 ; le lemme 5 ci-après prépare le lemme 6.

Lemme 5 : Soit $\mathbf{A}$ une matrice irréductible, positive.

Soit C une sous matrice principale de $\mathbf{A}$ c'est-à-dire une matrice $c = (a_{ij})$ $i,j \in I \subset [1,\ldots n]$.

Alors $\rho(C) \leqslant \rho(\mathbf{A})$, et il n'y a égalité que si $C = \mathbf{A}$.

Démonstration.

Grâce à une matrice de permutation P , on peut mettre $\mathbf{A}$ sous la forme :

$$\mathbf{A}' = P\mathbf{A}P^{T} = \left(\begin{array}{c|c} C & \mathbf{A}_{12} \\ \hline \mathbf{A}_{21} & \mathbf{A}_{22} \end{array}\right)$$

et soit alors $B = \left(\begin{array}{c|c} C & 0 \\ \hline 0 & 0 \end{array}\right)$.

Si $\mathbf{A}$ est différent de C , B est inférieure à $\mathbf{A}'$ et ne lui est point égale (en raison de l'irréductibilité) :

$$B \leqslant \mathbf{A}' \quad , \quad B \neq \mathbf{A}'.$$

Nous appliquons le théorème de Perron-Frobenius, comme $B \leqslant \mathbf{A}'$, $B \neq \mathbf{A}'$, on a

$$\rho(B) = \rho(C) < \rho(\mathbf{A}') = \rho(\mathbf{A}).$$

Lemme 6 : Si $\mathbf{A}$ est une matrice positive, irréductible, $\rho(\mathbf{A})$ est une valeur propre simple.

Démonstration. On considère la fonction $\varphi : \lambda \to \varphi(\lambda) = \det(\lambda I - \mathbf{A})$ (le polynôme caractéristique de $\mathbf{A}$).

$\rho(\mathbf{A})$ étant valeur propre de $\mathbf{A}$ on a :

$$\varphi(\rho(\mathbf{A})) = 0 \ .$$

Pour montrer que $\rho(\mathbf{A})$ est une valeur propre simple il nous suffit de montrer que $\varphi'(\rho(\mathbf{A}))$ n'est pas nul.

Or on a :

$$\varphi'(\lambda) = \sum_{i=1}^{n} \det \mathbf{A}_i'$$

où $\mathbf{A}_i'$ est définie de la façon suivante :

- Si $j \neq i$ les colonnes c_j de $\mathbf{A}_i'$ sont égales à celles de $\lambda\mathbf{I}-\mathbf{A}$.

- Si $j = i$ c_i est égale à la dérivée par rapport à λ de la i^e colonne de $\lambda\mathbf{I}-\mathbf{A}$.

C'est-à-dire :

$$\mathbf{A}_i' = \begin{pmatrix} \lambda-a_{11} & a_{12} & 0 & a_{1n} \\ \ddots & \ddots & \vdots & \\ & & 1 & \\ & & 0 & \\ & & \lambda-a_{nn} & \end{pmatrix} \leftarrow i$$

Le déterminant de $\mathbf{A}_i'$ est égal au déterminant du mineur principal C_i de $\lambda\mathbf{I}-\mathbf{A}$ ($\lambda\mathbf{I}-\mathbf{A}$ privée de sa $i^{ème}$ ligne et $i^{ème}$ colonne).

$\varphi'(\lambda)$ est la somme de tous les déterminants des mineurs principaux C_i de $\mathbf{A}$.

Si λ est strictement supérieur à $\rho(C_i)$, $\det(\lambda\mathbf{I}-C_i)$ est différent de zéro (par définition de $\rho(C_i)$), donc garde un signe constant sur $]\rho(C_i),+\infty[$.

Lorsque λ tend vers l'infini, $\det(\lambda I - C_i)$ étant un polynôme en λ de degré $n-1$ devient infini.

$$\det(\lambda I - C_i) = \lambda^{n-1} + \ldots + (-1)^{n-1} \det C_i$$

Le signe de $\det(\lambda I - C_i)$ sur $]\rho(C_i), +\infty[$ est donc le signe plus.

D'après le lemme 5, $\rho(A)$ est supérieur strictement à $\rho(C_i)$, le déterminant de $(\rho(A)I - C_i)$ est donc aussi strictement positif.

$\varphi'(\rho(A))$ est la somme de n déterminants tous strictement positifs. On a donc :

$$\varphi'(\rho(A)) > 0$$

et $\rho(A)$ est une valeur propre simple de A.

Ceci achève la démonstration du théorème de Perron-Frobenius dans le cas où A est une matrice irréductible, positive.

5) Matrices positives réductibles.

Lemme : Si A est une matrice d'ordre n réductible, il existe une matrice de permutation P telle que la transmuée de A, $A' = PAP^T$ soit bloc triangulaire supérieure.

$$A' = \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1p} \\ & A_{22} & \ddots & \vdots \\ 0 & & \ddots & \\ & & & A_{pp} \end{pmatrix}$$

(n_1 lignes, n_2 lignes, n_p lignes ; n_1 colonnes, n_2 colonnes, n_p colonnes)

où chaque A_{ii} est carrée d'ordre n_i et ou bien irréductible ou bien $n_i = 1$, et $A_{ii} = \{0\}$.

Démonstration.

Si **A** est réductible, il existe une matrice de permutation **Q** tel que :

$$QAQ^T = \left(\begin{array}{c|c} B_{11} & B_{12} \\ \hline 0 & B_{22} \end{array}\right)$$

Si une des matrices B_{11} ou B_{22} est réductible, on recommence l'opération précédente, etc... Le processus est terminé lorsque tous les blocs diagonaux sont irréductibles ou d'ordre (1) et égaux à 0 .

Lemme : Soit **A** une matrice bloc triangulaire supérieure.

$$\mathbf{A} = \begin{pmatrix} \overbrace{\mathbf{A}_{11}}^{n_1} & \mathbf{A}_{12} & \cdots & \mathbf{A}_{1p} \\ & \ddots & & \vdots \\ & & \mathbf{A}_{22} & \vdots \\ & & & \ddots \\ & & & \mathbf{A}_{pp} \end{pmatrix}$$

Alors
$$\det \mathbf{A} = \prod_{i=1}^{p} \det \mathbf{A}_{ii}$$

et

le polynôme caractéristique de **A** est égal au produit des polynômes caractéristiques des $\mathbf{A}_{ii}$.

$$\det(\lambda I - \mathbf{A}) = \prod_{i=1}^{p} \det (\lambda I - \mathbf{A}_{ii}).$$

Démonstration.

Nous raisonnons par récurrence sur n .

La proposition est évidente si $n = 1$.

Supposons là vraie à l'ordre $n-1$. Alors on obtient aisément le résultat en développant $\det \mathbf{A}$ suivant la première colonne de **A** .

Soit B_{ij} (resp. C_{ij}) la sous matrice de $\mathbf{A}_{11}$ (resp. **A**) obtenue en supprimant la 1ère colonne et la $i^{ème}$ ligne. On a :

$$\det A = \sum_{i=1}^{n_1} (-1)^{i+1} a_{i1} \det C_{i1} .$$

Par l'hypothèse de récurrence on a :

$$\det C_{i1} = \det B_{i1} \prod_{j=2}^{p} \det A_{jj} .$$

Donc :

$$\det A = \left(\sum_{i=1}^{n_1} (-1)^{i+1} a_{i1} \det B_{i1}\right) \prod_{j=2}^{p} \det A_{jj} = \prod_{j=1}^{p} \det A_{jj} .$$

La proposition concernant les polynômes caractéristiques résulte de ce que $\lambda I - A$ est aussi bloc triangulaire supérieure.

Théorème de Perron-Frobénius (cas général).

Soit A une matrice carrée positive, d'ordre n . Alors

i) $\rho(A)$ est une valeur propre de A et à $\rho(A)$ correspond un vecteur propre x positif (non nul)

ii) Soit B une matrice d'ordre n . Si $|B|$ est inférieur à $\rho(B)$ est inférieur à $\rho(A)$: $|B| \leqslant A \Longrightarrow \rho(B) \leqslant \rho(A)$.

Remarque :

Ce théorème est la généralisation du 1er théorème de Perron-Frobenius, pour une matrice quelconque. On remarque que dans le cas où A est irréductible $\rho(A)$ est une valeur propre simple de A et $\rho(A)$ étant une fonction strictement croissante ; ces propriétés ne sont plus vraies dans le cas général. De même $|B| \leqslant A$, $|B| \neq A$ entraîne $\rho(B) \leqslant \rho(A)$ mais pas nécessairement $\rho(B) < \rho(A)$.

Démonstration.

i) Soit J la matrice d'ordre n dont toues les éléments sont égaux à 1.

$A+\varepsilon J$ est une matrice strictement positive et irréductible.

D'après le théorème de Perron-Frobenius (cas irréductible) $\lambda_\varepsilon = \rho(A+\varepsilon J)$ est une valeur propre de $A+\varepsilon J$.

A λ_ε correspond un vecteur propre x_ε strictement positif. On peut choisir x_ε de norme 1 :

$$\exists\, x_\varepsilon > 0 \ , \ \|x_\varepsilon\| = 1 \ , \ \text{tel que} \ (A+\varepsilon J)x_\varepsilon = \lambda_\varepsilon x_\varepsilon \ .$$

Soit $\varepsilon' < \varepsilon$; on a :

$$0 \leqslant A < A+\varepsilon' J < A+\varepsilon' J \ ;$$

A étant une matrice positive, inférieure strictement à une matrice $A+\varepsilon' J$ irréductible, $\rho(A)$ est strictement inférieur à $\rho(A+\varepsilon' J)$.

De même $\rho(A+\varepsilon' J) < \rho(A+\varepsilon J)$.

Faisons tendre ε vers 0 ; $\rho(A+\varepsilon J)$ étant une suite décroissante bornée inférieurement par $\rho(A)$, elle est convergente : $\rho(A+\varepsilon J)$ tend vers $\bar{\lambda}$, $\bar{\lambda} \geqslant \rho(A)$.

λ_ε étant valeur propre de $A+\varepsilon J$ on a :

$$\det(\lambda_\varepsilon I - A - \varepsilon J) = 0 \ .$$

La fonction det étant une fonction continue, lorsque ε tend vers zéro, $\det(\lambda_\varepsilon I - A - \varepsilon J)$ tend vers $\det(\bar{\lambda} I - A)$; on a donc :

$$\det(\bar{\lambda} I - A) = 0 \ ,$$

et $\bar{\lambda}$ est valeur propre de A . Donc $\bar{\lambda} \leqslant \rho(A)$, et comme $\bar{\lambda} \geqslant \rho(A)$, on a $\bar{\lambda} = \rho(A)$ et $\rho(A)$ est valeur propre de A .

Montrons qu'il lui correspond un vecteur propre x positif ou nul.

Il existe une suite de ε_i ((ε_i) tendant vers 0) telsque :

$\|x_{\varepsilon_i}\| = 1$ et la suite X_{ε_i} a une limite X .

on a $$(A+\varepsilon_i J)X_{\varepsilon_i} = \lambda_{\varepsilon_i} X_{\varepsilon_i}$$

en passant à la limite on obtient :

$$Ax = \bar{\lambda}x \quad \text{où} \quad \bar{\lambda} = \rho(A)$$

et x est positif, non nul (en effet tous les x_{ε_i} sont de norme 1 et positif, donc x est de norme 1).

ii) Supposons $|B| \leqslant A$.

Soit $A' = PAP^T$ la forme bloc triangulaire supérieure de A .

Soit $B' = PBP^T$.

Comme $|B'|$ est inférieur à A', B' a la même structure que A', c'est-à-dire que :

$$B'_{ij} = 0 \quad \text{pour} \quad i < j$$

et d'après le lemme (p.49) on a :

$$\det B' = \prod_{i=1}^{n} \det B'_{ii}$$

$$\det(\lambda I - B') = \prod_{i=1}^{n} \det(\lambda I - B'_{ii})$$

et de même pour A'.

D'après le théorème de Perron-Frobenius (cas irréductible)

on a $$\rho(B'_{ii}) \leqslant \rho(A'_{ii})$$

Donc $$\rho(B) = \max \rho(B'_{ii}) \leqslant \max \rho(A'_{ii}) = \rho(A).$$

§III. Méthodes itératives de résolution des systèmes linéaires.

On cherche à résoudre le système linéaire :

$$\mathbf{A}.\mathbf{x} = b$$

où $\mathbf{A}$ est une matrice réelle de type (n,n), régulière, et b un vecteur donné de $\mathbb{R}^n$.

Résoudre $\mathbf{A}.x = b$ par une méthode itérative, c'est construire une suite de vecteurs $x^{(m)} \in \mathbb{R}^n$, telle que : $\lim_{m\to+\infty} x^{(m)} = x$, et $\mathbf{A}.x = b$.

Pratiquement, on arrêtera évidemment les calculs dès que l'on aura atteint la précision désirée (ou imposée par la machine).

1°) Convergence d'une méthode itérative.

Toutes les méthodes itératives classiques sont du type :

$$\boxed{x^{(m+1)} = B.x^{(m)} + c} \quad (1)$$

où B est une matrice de type (n,n), et $c \in \mathbb{R}^n$. Cherchons à quelles conditions sur B et c les suites vérifiant la relation de récurrence (1) sont convergentes.

On appelle x la solution exacte : $x = \mathbf{A}^{-1}b$;

$$x = Bx + (I-B)x \text{ , d'où :}$$

$$x = Bx + (I-B)\mathbf{A}^{-1}b$$

$$x^{(m+1)} = Bx^{(m)} + c$$

Retranchons membre à membre ces deux égalités :

$$x^{(m+1)} - x = B(x^{(m)} - x) + [c - (I-B)\mathbf{A}^{-1}b] \text{ .}$$

Définition : On dit que la méthode (1) est consistante si, lorsque la limite de la suite $(x^{(m)})$ existe, cette limite est la solution x.

Supposons que la limite de $x^{(m)}$ existe ; soit $y = \lim_{m \to +\infty} x^{(m)}$.

Alors : $y-x = B(y-x) + [c - (I-B)A^{-1}b]$.

Pour que la méthode (1), associée à B et c, soit consistante, il faut et il suffit que :

$$(I-B)(y-x) = c - (I-B)A^{-1}b \Longrightarrow y-x = 0 ;$$

donc la méthode (1) est consistante si, et seulement si :

$$\begin{cases} c = (I-B)A^{-1}b \\ \text{et} \\ (I-B) \text{ est inversible, ou encore : 1 n'est pas valeur propre de } B . \end{cases}$$

Définition : Une méthode itérative est dite convergente si, quel que soit le choix du vecteur initial $x^{(0)}$, $\lim_{m \to +\infty} x^{(m)} = A^{-1}b = x$.

Lemme : On suppose que la méthode (1), associée à B et c est consistante. Alors cette méthode est convergente si, et seulement si : $\lim_{m \to +\infty} B^m = 0$.

Démonstration : Puisque la méthode est consistante, on a :

$$c - (I-B)A^{-1}b ,$$

et $(\forall m)$, $x^{(m+1)} - x = B(x^{(m)} - x)$.

Par récurrence, on obtient : $x^{(m)} - x = B^m(x^{(0)} - x)$.

La méthode est convergente si, et seulement si :

$$B^m(x^{(0)} - x) \to 0 , \quad \forall x^{(0)} \in \mathbb{R}^n$$

c'est-à-dire si et seulement si : $B^m z \to 0$ quand $m \to +\infty$, $\forall z \in \mathbb{R}^n$.

Ceci est équivalent à : $B^m \to 0$ quand $m \to +\infty$.

En effet, si $B^m \to 0$, et $z \in \mathbb{R}^n$, $\|B^m z\| \leqslant \|B^m\| \, \|z\|$. et $B^m z \to 0$.

Réciproquement; si $\forall z \in \mathbb{R}^n$, $B^m z \to 0$ quand $m \to +\infty$, alors prenons successivement pour z les vecteurs de la base canonique e_j $(j = 1,\ldots,n)$: $B^m . e_j \to 0$; et $B^m . e_j$ est le $j^{ième}$ vecteur-colonne de B^m ; donc $B^m \to 0$.

Proposition : Soit B une matrice (n,n) ; alors :

$$\lim_{m \to +\infty} B^m = 0 \Longleftrightarrow \rho(B) < 1 .$$

On va donner deux démonstrations de cette proposition.

1ère démonstration.

La première démonstration, très courte, utilise le théorème suivant (que l'on admettra) :

Théorème (Householder) : $\forall \eta > 0$, $\forall B \in \mathbb{R}_{(n\times n)}$, il existe sur $\mathbb{R}^n$ une norme $\|.\|$ telle que la norme associée de B : $\|B\| = \sup_{\substack{x \in \mathbb{R}^n \\ x \neq 0}} \frac{\|Bx\|}{\|x\|}$ vérifie : $\rho(B) \leqslant \|B\| \leqslant \rho(B) + \eta$.

a) La condition est suffisante :

si $\rho(B) < 1$: $\exists \eta > 0$ tel que $\rho(B) + \eta < 1$. Alors, d'après le théorème cité, il existe une norme $\|.\|$ telle que : $\|B\| \leqslant \rho(B) + \eta < 1$ d'où $\|B\|^m \to 0$ si $m \to +\infty$. Comme, $0 \leqslant \|B^m\| \leqslant \|B\|^m$, on a : $\|B^m\| \to 0$.

b) La condition est nécessaire : supposons $\rho(B) \geqslant 1$. Alors B possède au moins une valeur propre λ, de module : $|\lambda| \geqslant 1$.

Soit x un vecteur propre de B, correspondant à λ :

$Bx = \lambda x$, d'où $B^m x = \lambda^m x$ $(\forall m)$.

Le vecteur x est non nul et $|\lambda^m|$ tend vers 1 ou ∞ quand $m \to +\infty$; donc $B^m x$ ne tend pas vers 0 et B^m ne tend pas vers 0 .

2ème démonstration.

La matrice B est toujours semblable à une matrice de Jordan ; autrement dit, il existe une matrice S , régulière, telle que :

$$B = S\,J\,S^{-1} ,$$

où J est de la forme :

$$J = \begin{pmatrix} J_1 & 0 & 0 \dots 0 \\ & J_2 & 0 \\ & & \ddots \\ 0 & & J_p \end{pmatrix}$$

Pour chaque bloc J_α :

$$J_\alpha = \begin{pmatrix} \lambda_\alpha & 1 & 0 \dots 0 \\ & \ddots & \vdots \\ & & 0 \\ & & 1 \\ & & \lambda_\alpha \end{pmatrix} \quad \begin{matrix} \updownarrow r_\alpha \\ \\ \\ \\ \\ \end{matrix} \qquad \begin{matrix} \\ \\ \\ \\ \\ \leftrightarrow r_\alpha \end{matrix}$$

On a : $B^m = S\,J^m\,S^{-1}$

Donc : $B^m \to 0 \iff J^m \to 0 \iff \forall \alpha\ ,\ J_\alpha^m \to 0\ ,$ (quand $m \to +\infty$)

et d'après les deux lemmes qui suivent, $J_\alpha^m \to 0$ quand $m \to +\infty$ si, et seulement si :

$$|\lambda_\alpha| < 1 \quad (\alpha = 1, \dots, p) ;$$

la proposition en résulte, moyennant les deux lemmes utilisés :

Lemme (1) : Si J_α est un bloc de Jordan :

$$J_\alpha = \begin{pmatrix} \lambda & 1 & 0 & \cdots & 0 \\ & \ddots & \ddots & & \vdots \\ & & \ddots & \ddots & 0 \\ & & & \ddots & 1 \\ & & & & \lambda \end{pmatrix} \quad r$$

(matrice $r \times r$)

<u>alors</u> : $J_\alpha^m = (d_{ij}^{(m)})$,

<u>avec</u> : $d_{ij}^{(m)} = 0$ <u>si</u> $i > j$

$d_{ij}^{(m)} = 0$ <u>si</u> $i \leqslant j$ <u>et</u> $m+i < j$

$d_{ij}^{(m)} = \binom{m}{j-i}\lambda^{m-j+i}$ <u>si</u> $i \leqslant j \leqslant m+i$

<u>Démonstration</u> :

Raisonnons par récurrence sur m.

. Pour $m = 1$: $J_\alpha^1 = J_\alpha$; la vérification est immédiate

. Supposons que la proposition du lemme soit vraie à l'ordre m : montrons qu'elle l'est alors à l'ordre $m+1$

$$J_\alpha^{m+1} = J_\alpha \cdot J_\alpha^m , \quad \text{soit :}$$

$$d_{ij}^{(m+1)} = \begin{cases} \lambda\, d_{ij}^{(m)} + d_{i+1,j}^{(m)} & \text{si } i < r \\ \lambda\, d_{ij}^{(m)} & \text{si } i = r \end{cases}$$

- <u>Si</u> $i > j$: alors $i+1 > j$, donc $d_{ij}^{(m+1)} = 0$

- <u>Si</u> $i \leqslant j$ <u>et</u> $m+1+i < j$: alors $d_{i+1,j}^{(m)} = 0$; et $m+i < j$, d'où $d_{ij}^{(m)} = 0$; donc $d_{ij}^{(m+1)} = 0$.

- <u>Si</u> $i \leqslant j = m+i+1$: $d_{ij}^{(m)} = 0$; $d_{i+1,j}^{(m)} = \binom{m}{j-i-1}\lambda^{m-j+i+1} = 1$; $d_{i,j}^{(m+1)} = d_{i+1,j}^{(m)} = \binom{m+1}{j-i}\lambda^{(m+1)-j+i}$.

- <u>Si</u> $i \leqslant j < m+i+1$: alors $d_{ij}^{(m)} = \binom{m}{j-i}\lambda^{m-j+i}$;

Trois cas sont alors à envisager :

⊗ ou bien $i < r$ et $j > i$: alors $d^{(m)}_{i+1,j} = \binom{m}{j-(i+1)}\lambda^{m-j+i+1}$

d'où : $d^{(m+1)}_{i,j} = \lambda\, d^{(m)}_{ij} + d^{(m)}_{i+1,j} = \lambda^{m-j+i+1}\left[\binom{m}{j-i}+\binom{m}{j-i-1}\right]$

$$d^{(m+1)}_{i,j} = \binom{m+1}{j-i}\lambda^{m+1+j-i}$$

⊗ ou bien $i = j < r$: alors $d^{(m)}_{i+1,j} = 0$,

$$d^{(m+1)}_{ij} = d^{(m+1)}_{ii} = \binom{m}{0}\lambda^m.\lambda = \binom{m+1}{0}\lambda^{m+1} .$$

⊗ ou bien $i = j = r$: $d^{(m+1)}_{rr} = \lambda\, d^{(m)}_{r,r} = \lambda.\binom{m}{0}\lambda^m = \lambda^{m+1}\binom{m}{0}$.

- Dans tous les cas, l'hypothèse de récurrence est vérifiée à l'ordre $m+1$. ■

Lemme (2) : Soit J_α un bloc de Jordan :

$$J_\alpha = \begin{pmatrix} \lambda_\alpha & 1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ 0 & & & \lambda_\alpha \end{pmatrix} \Big\updownarrow r_\alpha$$

alors : $J^m_\alpha \to 0$ (quand $m \to +\infty$) $\Longleftrightarrow |\lambda_\alpha| < 1$.

Démonstration :

Appliquons le lemme (1) à $J^m_\alpha = (d^{(m)}_{ij})$:

$$d^{(m)}_{ij} = \binom{m}{j-i}\lambda_\alpha^{m-j+i} = \frac{m(m-1)\ldots(m-j+i+1)}{(j-i)!}\,\lambda_\alpha^{m-j+i}$$

($\forall i$ et j , dès que $m > r_\alpha$ $\quad 1 \leqslant i \leqslant r_\alpha$, $\quad 1 \leqslant j \leqslant r_\alpha$).

Pour i et j fixés : $\binom{m}{j-i}\lambda_\alpha^{m-j+i} \sim \frac{\lambda_\alpha^{i-j}}{(j-i)!}\, m^{j-i}\, \lambda_\alpha^m$ quand $m \to +\infty$.

Or : $m^{j-i}\,\lambda_\alpha^m \to 0 \Longleftrightarrow |\lambda_\alpha| < 1$

quand $m \to +\infty$

donc : $J^m_\alpha \to 0 \Longleftrightarrow |\lambda_\alpha| < 1$.

2°) Taux de convergence d'une méthode itérative.

Considérons une méthode itérative, définie par :

$$X^{(m+1)} = B\,X^{(m)} + c\ .$$

Nous supposerons qu'elle est convergente.

Définition (1) : - On appelle taux moyen de convergence pour une méthode itérative, définie par une matrice B, le nombre : $R_m(B) = -\mathrm{Log}\,\|B^m\|^{1/m}$.

- On dit que la méthode définie par une matrice B_1 est plus rapide (pour m itérations) que la méthode définie par une matrice B_2 si $R_m(B_1) \geqslant R_m(B_2)$.

Remarques : 1) Le taux moyen de convergence dépend de la norme choisie, et également le fait qu'une méthode itérative soit plus rapide qu'une autre au bout de m itérations.

2) Le nombre $R_m(B)$ permet de "mesurer" le nombre d'itérations nécessaires pour réduire l'erreur $\Sigma^{(m)} = x^{(m)} - x$ d'un facteur k :

$$\Sigma^{(m)} = B^m\,\Sigma^{(0)}$$

$$\|\Sigma^{(m)}\| \leqslant \|B^m\| \cdot \|\Sigma^{(0)}\|\ .$$

Si on veut avoir :

$$\frac{\|\Sigma^{(m)}\|}{\|\Sigma^{(0)}\|} \leqslant k :$$

Cela est réalisé si $\|B^m\| \leqslant k$, soit :

$$\mathrm{Log}\,\|B^m\| = -m\,R_m(B) \leqslant \mathrm{Log}\,k\ .$$

Puisque la méthode converge : $R_m(B) > 0$ (au moins pour m assez grand), et on prendra :

$$m \geqslant -\frac{\text{Log } k}{R_m(B)} \quad (k < 1).$$

On démontre ci-après que $\lim_{m \to +\infty} R_m(B) = -\text{Log}(\rho(B))$ a posteriori.

Ceci justifiera, a posteriori, la définition suivante :

Définition (2) : - On appelle taux asymptotique de convergence d'une méthode itérative (définie par une matrice B) le nombre :

$$R_\infty(B) = \lim_{m \to +\infty} R_m(B).$$

- On dit que la méthode itérative définie par B_1 est asymptotiquement plus rapide que celle définie par B_2 si : $R_\infty(B_1) > R_\infty(B_2)$.

Proposition : $\boxed{R_\infty(B) = \lim_{m \to +\infty} R_m(B) = -\text{Log}[\rho(B)]}$

Conséquence. La méthode (B_1) est asymptotiquement plus rapide que la méthode (B_2) si $\rho(B_1) < \rho(B_2)$.

La proposition va être démontrée à l'aide du lemme suivant :

Lemme : Pour toute matrice B telle que $\rho(B) < 1$, il existe une suite (c_m), avec :

$$\forall m \ , \ 0 < c \leqslant c_m \leqslant c' \ ,$$

telle que, lorsque $m \to +\infty$:

$$\|B^m\| \sim c_m \binom{m}{s-1} \rho(B)^{m-s+1} \ ;$$

l'entier s est, dans la réduite de Jordan J de la matrice B, l'ordre maximum des blocs J_α tels que : $|\lambda_\alpha| = \rho(J) = \rho(B)$.

Démonstration :

Il existe une matrice S régulière telle que : $B = S\,J\,S^{-1}$; alors : $\forall m$, $B^m = S\,J^m\,S^{-1}$.

Etudions la limite de la matrice : $\dfrac{B^m}{\binom{m}{s-1}\rho(B)^{m-s+1}}$

$$J^m = \begin{pmatrix} \boxed{J_1^m} & & & 0 \\ & \boxed{J_2^m} & & \\ & & \ddots & \\ 0 & & & \boxed{J_p^m} \end{pmatrix}$$

λ_β étant l'une des valeurs propres de B, étudions le terme général :

$\dfrac{d_{i,j}^{(m,\beta)}}{\binom{m}{s-1}\rho(B)^{m-s+1}} = g_{ij}^{(m,\beta)}$, du bloc : $\dfrac{J_\beta^m}{\binom{m}{s-1}\rho(B)^{m-s+1}}$.

Soit r_β l'ordre de J_β.

Supposons m assez grand, par exemple $m > n$.

Alors : - si $i > j$: $g_{i,j}^{(m,\beta)} = 0$

- si $i \leqslant j$:

$$g_{ij}^{(m,\beta)} = \frac{\binom{m}{j-i}\lambda_\beta^{m-j+i}}{\binom{m}{s-1}\rho(B)^{m-s+1}} \sim \left[\frac{(s-1)!\lambda_\beta^{i-j}}{(j-i)!\rho(B)^{1-s}}\right] \times \left(\frac{\lambda_\beta}{\rho(B)}\right)^m \frac{m^{j-i}}{m^{s-1}}$$

$(1 \leqslant i \leqslant r_\beta\ ,\ 1 \leqslant j \leqslant r_\beta)$.

On a : $|\lambda_\beta| \leqslant \rho(B)$; trois cas peuvent alors se présenter :

- si $|\lambda_\beta| < \rho(B)$: $g_{i,j}^{(m,\beta)} \to 0$ quand $m \to +\infty$

- si $|\lambda_\beta| = \rho(B)$: alors $r_\beta \leqslant s$; si $r_\beta < s$: $g_{i,j}^{(m,\beta)} \to 0$

- si $|\lambda_\beta| = \rho(B)$ et $r_\beta = s$: $j-i \leqslant r_\beta - 1 = s-1$, donc $g_{ij}^{(m,\beta)} \to 0$,

sauf si $j-i = s-1$.

Mais cela n'a lieu que si $j=s$ et $i=1$, et pour ce terme :

$g_{1,s}^{(m,\beta)} = \left(\dfrac{\lambda_\beta}{|\lambda_\beta|}\right)^{m-s+1}$, donc $|g_{1,s}^{(m,\beta)}| = 1$.

Donc, pour chaque β tel que $|\lambda_\beta| = \rho(B)$ et $r_\beta = s$, on a :

$$\frac{J_\beta^m}{\binom{m}{s-1}\rho(B)^{m-s+1}} = \begin{pmatrix} \mathcal{O}(1) \ldots\ldots \mathcal{O}(1) & \left(\frac{\lambda_\beta}{|\lambda_\beta|}\right)^{m-s+1} \\ \ddots \quad \ddots & \mathcal{O}(1) \\ \quad \ddots \quad \ddots & \vdots \\ & \mathcal{O}(1) \end{pmatrix}$$

où $\mathcal{O}(1)$ - terme qui tend vers 0 pour $m \to \infty$.

Soit T_m la matrice obtenue à partir de la matrice : $\frac{J^m}{\binom{m}{s-1}\rho(B)^{m-s+1}}$ en remplaçant les termes tendant vers 0 par des zéros $\neq$ T_m contient un ou plusieurs termes de module toujours égal à 1, et ses autres termes sont nuls.

Posons $\frac{J^m}{\binom{m}{s-1}\rho(B)^{m-s+1}} = T_m + E_m$: Tous les éléments de E_m tendent vers 0 quand $m \to +\infty$.

$$B^m = \binom{m}{s-1}\rho(B)^{m-s+1}[S\, T_m\, S^{-1} + S\, E_m\, S^{-1}]$$

quand $m \to +\infty$, $\|S\, E_m\, S^{-1}\| \to 0$.

Soit $c_m = \|S\, T_m\, S^{-1}\|$.

- La suite c_m est majorée, car $c_m \leqslant \|S\| . \|T_m\| . \|S^{-1}\|$ et la suite $(\|T_m\|)$ est bornée (sinon les éléments de T_m tendraient vers l'infini).

Donc il existe $c' > 0$ tel que $\forall m : c_m \leqslant c'$.

- La suite c_m est minorée par un nombre $c > 0$; en effet, raisonnons par l'absurde.

Si l'on avait $\lim\limits_{m \to +\infty} \inf c_m = 0$, il existerait une sous-suite (c_{m_i}) tendant vers 0, c'est-à-dire : $\|S\, T_{m_i}\, S^{-1}\| \to 0$: mais alors T_{m_i} tendrait vers 0, ce qui est impossible puisque, par construction, T_m a toujours certains de ses éléments égaux à 1 en module.

Ainsi, il existe c et c' tels que $(\forall m)$: $0 < c \leqslant c_m \leqslant c'$. Ceci étant :

$$c_m - \|S\,E_m\,S^{-1}\| \leqslant \|S\,T_m\,S^{-1} + S\,E_m\,S^{-1}\| \leqslant c_m + \|S\,E_m\,S^{-1}\|$$

donc : $\|S\,T_m\,S^{-1} + S\,E_m\,S^{-1}\| = c_m + \mathcal{O}(1)$.

Finalement : $\|B^m\| = \binom{m}{s-1}\rho(B)^{m-s+1}(c_m + \mathcal{O}(1))$ et le lemme est enfin démontré.

Démonstration de la proposition :

Par définition : $R_m(B) = -\frac{1}{m}\,\mathrm{Log}\,\|B^m\|$.

Appliquons le lemme :

$$R_m(B) = -\frac{1}{m}\,\mathrm{Log}\left[\binom{m}{s-1}\right] - \frac{1}{m}\,\mathrm{Log}(\rho(B)^{m-s+1}) - \frac{1}{m}\,\mathrm{Log}(c_m + \mathcal{O}(1))$$

$$= -\frac{1}{m}\sum_{k=1}^{s+1}\mathrm{Log}(m-k+1) - \frac{1}{m}\,\mathrm{Log}(c_m + \mathcal{O}(1)) + \frac{1}{m}\,\mathrm{Log}((s-1)!) - \frac{m-s+1}{m}\,\mathrm{Log}[\rho(B)]$$

quand $m \to +\infty$: les trois premiers termes tendent vers 0 ; le quatrième tend vers : $-\mathrm{Log}(\rho(B))$; la proposition en résulte.

3°) Principales méthodes itératives.

Elles se construisent toutes de la manière suivante :

on pose : $A = M - N$

où M est inversible, et, en pratique facile à inverser (en particulier matrices diagonales ou triangulaires).

On détermine la suite $x^{(m)}$ de vecteurs de R^n par la relation de récurrence :

$$M\,x^{(m+1)} = N\,x^{(m)} + b$$

c'est-à-dire :

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

Donc $x^{(m+1)} = B\,x^{(m)} + C$ avec $\begin{cases} B = M^{-1} N \\ C = M^{-1} b \end{cases}$.

On a vu qu'une condition nécessaire et suffisante pour qu'une méthode itérative consistante soit convergente est que le rayon spectral de B soit strictement plus petit que 1.

- Ces méthodes itératives sont consistantes, cela est immédiat.

- Les qualités de ces méthodes itératives dépendent donc de la facilité du calcul de l'inverse de M et du rayon spectral de $M^{-1}N$ (qui doit être aussi petit que possible).

On décompose la matrice $A = (a_{ij})$ de la façon suivante :

$$A = D - E - F$$

où a) $D = (d_{ij})$ est diagonale

$$\begin{cases} d_{ij} = 0 & \text{si } i \neq j \\ d_{ij} = a_{ii} & \text{si } i = j \end{cases}$$

b) $E = (e_{ij})$ est strictement triangulaire inférieure

$$\begin{cases} e_{ij} = 0 & \text{si } i \leqslant j \\ e_{ij} = -a_{ij} & \text{si } i > j \end{cases}$$

c) $F = (f_{ij})$ est strictement triangulaire supérieure

$$\begin{cases} f_{ij} = 0 & \text{si } j \leqslant i \\ f_{ij} = -a_{i} & \text{si } j < i \end{cases}$$

Nous allons voir trois méthodes tout-à-fait classiques, applicables dans le cas où tous les éléments de la diagonale de A sont non nuls.

a) Méthode de Jacobi

On pose, pour cette méthode :

$$M = D \quad , \quad N = E+F$$

on a :

$$D\, x^{(m+1)} = (E+F)x^{(m)} + b$$

D est inversible puisque tous les a_{ii} sont non nuls.

Soit :

$$\boxed{x_i^{(m+1)} = -\sum_{\substack{j=1 \\ j\neq i}}^{n} \frac{a_{ij}}{a_{ii}}\, x_j^{(m)} + \frac{b_i}{a_{ii}} \qquad \forall i \quad 1 \leqslant i \leqslant n}$$

Par la méthode de Jacobi, on obtient les composantes de $x^{(m+1)}$ à partir de celles de $x^{(m)}$.

On appelle $B=J$ la matrice de l'itération :

$$J = M^{-1}N$$

Soit

$$J = D^{-1}(E+F)$$

$$\underline{J = L+U} \quad \text{avec} \quad L = D^{-1}E \quad , \quad U = D^{-1}F$$

b) Méthode de Gauss-Seidel

On pose ici $M = D-E$, $N = F$, et on écrit la formule de récurrence :

$$(D-E)x^{(m+1)} = F\, x^{(m)} + b \, .$$

Soit

$$\sum_{j=1}^{i} a_{ij}\, x_j^{(m+1)} = -\sum_{j=i+1}^{n} a_{ij}\, x_j^{(m)} + b_i \qquad \forall i \quad 1 \leqslant i \leqslant n$$

$$a_{ii}\, x_i^{(m+1)} = -\sum_{j=1}^{i-1} a_{ij}\, x_j^{(m+1)} - \sum_{j=i+1}^{n} a_{ij}\, x_j^{(m)} + b_i$$

$$\boxed{x_i^{(m+1)} = -\sum_{j=1}^{i-1} \frac{a_{ij}}{a_{ii}}\, x_j^{(m+1)} - \sum_{j=i+1}^{n} \frac{a_{ij}}{a_{ii}}\, x_j^{(m)} + \frac{b_i}{a_{ii}} \qquad \forall i \quad 1 \leqslant i \leqslant n}$$

Remarque :

Dans la méthode de Jacobi, la $i^{ème}$ composante de $x^{(m+1)}$ est déterminée en fonction des n composantes de $x^{(m)}$ qui doivent donc être conservées en mémoire jusqu'au calcul de $x_n^{(m+1)}$.

Par contre la méthode de Gauss-Seidel donne la $i^{ème}$ composante de $x^{(m+1)}$ en fonction des $(i-1)$ premières composantes de $x^{(m+1)}$ et des $(n-i-1)$ dernières composantes de $x^{(m)}$. Elle ne nécessite pas la conservation en mémoire des autres composantes de $x^{(m)}$; ce qui est un avantage important sur la méthode de Jacobi.

Pour la méthode de Gauss-Seidel on appelle $B = G$ la matrice de l'itération

$$G = M^{-1}N$$

on a

$$G = (D-E)^{-1}F$$

$$= (DD^{-1}D - DD^{-1}E)^{-1}F$$

$$= (D(I-L))^{-1}F$$

$$= (I-L)^{-1}D^{-1}F$$

$$\underline{G = (I-L)^{-1}U}$$

c) Méthode de Relaxation

Pour cette méthode on pose :

$$M = \left(\frac{1}{\omega}D-E\right) \qquad N = \left(\frac{1}{\omega}-1\right)D+F$$

où ω est un nombre réel non nul.

(On verra, ensuite, d'autres limitations pour ω : il faut en fait que $\omega \in]0,2[$; cf aussi plus loin le problème fondamental du choix du ω optimal).

On a la récurrence :

$$(\frac{1}{\omega} D-E)x^{(m+1)} = ((\frac{1}{\omega}-1)D+F)x^{(m)} + b .$$

Soit

$$(D-\omega E)x^{(m+1)} = ((1-\omega)D + \omega F)x^{(m)} + \omega b .$$

$$x_i^{(m+1)} = (1-\omega)x_i^{(m)} - \omega \sum_{j=i+1}^{n} \frac{a_{ij}}{a_{ii}} x_j^{(m)} - \omega \sum_{j=1}^{i-1} \frac{a_{ij}}{a_{ii}} x_j^{(m+1)} + \omega \frac{b_i}{a_{ii}} \quad \forall i \quad 1 \leqslant i \leqslant n .$$

Remarque :

1 - Si $\omega = 1$ on retrouve la méthode de Gauss-Seidel.

2 - On dit qu'il y a sous ou sur-relaxation suivant que $\omega < 1$ ou $\omega > 1$.

Pour la méthode de relaxation on a la matrice d'itération $B = \mathcal{L}\omega$.

$$\mathcal{L}\omega = M^{-1}N$$

$$= (\frac{1}{\omega} D-E)^{-1}((\frac{1}{\omega}-1)D+F)$$

$$= (D-\omega E)^{-1}((1-\omega)D+\omega F)$$

$$= (D-\omega E)^{-1}DD^{-1}((1-\omega)D+\omega F)$$

$$\underline{\mathcal{L}\omega = (I-\omega L)^{-1}((1-\omega)I+\omega U).}$$

4°) Convergence des méthodes de Jacobi et Gauss-Seidel.

Proposition 1.

Supposons que les matrices $L = D^{-1}E$ et $U = D^{-1}F$ définies précédemment soient positives ou nulles.

Alors :

Si la méthode de Jacobi est convergente pour la matrice A, la méthode de Gauss-Seidel est aussi convergente et, est asymptotiquement au moins aussi rapide.

Nous déduirons cette proposition de trois lemmes que nous allons voir maintenant.

Définition.

Pour tout nombre σ positif ou nul, on définit $m(\sigma)$ comme étant le rayon spectral de la matrice $\sigma L + U$ et $n(\sigma)$ comme étant celui de la matrice $L + \sigma U$.

$$m(\sigma) = \rho(\sigma L + U)$$

$$n(\sigma) = \rho(L + \sigma U).$$

Remarque :

1 - On a $m(0) = n(0) = 0$

2 - $m(1) = n(1) = \rho(L+U) = \rho(J)$

3 - Soit λ valeur propre de $\sigma L + U$.

Comme on peut écrire $\sigma L + U$ sous la forme $\sigma(L + \frac{U}{\sigma})$, si σ est non nul, $\frac{\lambda}{\sigma}$ est valeur propre de $(L + \frac{U}{\sigma})$.

D'où :

$$\rho(\sigma L + U) = \sigma\, \rho(L + \frac{U}{\sigma})$$

$$m(\sigma) = \sigma\, n(\frac{1}{\sigma}).$$

Lemme 1.

L et U étant des matrices positives :

$m(\sigma)$ est une fonction croissante de σ.

$n(\sigma)$ est une fonction croissante de σ.

Démonstration.

Soit : $0 \leqslant \sigma \leqslant \sigma'$.

Les matrices L et U étant positives $\sigma L + U$ est inférieur ou égal à $\sigma' L + U$.

D'après le théorème de Perron-Frobenius (cas réductible) on a :

$m(\sigma) = \rho(\sigma L + U) \leqslant \rho(\sigma' L + U) = m(\sigma')$. $m(\sigma)$ est une fonction croissante.

Pour montrer que $n(\sigma)$ est aussi une fonction croissante on ferait le même raisonnement.

Lemme 2.

Si L est une matrice d'ordre n, strictement triangulaire inférieure, on a :

$$(I-L)^{-1} = I + L + L^2 + \ldots + L^{n-1} .$$

Démonstration.

On a $\det(\lambda I - L) = \lambda^n$.

En effet la matrice $(\lambda I - L)$ est triangulaire inférieure et tous les termes de la diagonale sont égaux à λ.

D'après le théorème de Cayley-Hamilton (1) : $L^n = 0$

on a alors :

$$(I-L)(I + L + \ldots + L^{n-1}) = I - L^n = I .$$

D'où

$$(I-L)^{-1} = I + L + L^2 + \ldots + L^{n-1} .$$

Lemme 3.

Soient L et U deux matrices positives respectivement : strictement triangulaire inférieure et strictement triangulaire supérieure.

---

(1) Rappel du théorème de Cayley Hamilton : Si C est une matrice à coefficients dans un corps K et si $p_c(\lambda) = \det(U - \lambda I)$, alors $p_c(C) = 0$.

Soient $J = L+U$ et $G = (I-L)^{-1}U$.

Soit λ le rayon spectral de G .

Si λ est non nul, λ est inférieur ou égal à $m(\lambda)$ et $n(\frac{1}{\lambda})$ est plus grand que 1.

$$\lambda \leqslant m(\lambda) \quad \text{et} \quad n(\tfrac{1}{\lambda}) \geqslant 1 .$$

Démonstration.

L et U étant des matrices positives, G est positive : d'après le lemme 2 on a :

$$G = (I + L + \ldots + L^{n-1})U .$$

D'après le théorème de Perron-Frobenius, λ est valeur propre de G et il existe un vecteur propre x associé à λ , positif ou nul.

$$Gx = \lambda x .$$

Soit $$(I-L)^{-1}Ux = \lambda x .$$

D'où $$Ux = \lambda(I-L)x .$$

$$(U + \lambda L)x = \lambda x .$$

λ est donc valeur propre de $\lambda L + U$

$$\lambda \leqslant \rho(\lambda L + U) = m(\lambda).$$

Si λ n'est pas nul, d'après la remarque 3, on a :

$$\frac{m(\lambda)}{\lambda} = n(\tfrac{1}{\lambda}) \geqslant 1 = m(1).$$

Il en résulte que : $n(\frac{1}{\lambda})$ est supérieur à 1.

Ce qui achève la démonstration de ce lemme.

Proposition 2.

Soient L et U deux matrices positives, respectivement : strictement triangulaire inférieure et strictement triangulaire supérieure,

Soient $J = L+U$ et $G = (I-L)^{-1}U$.

Si $\rho(J) < 1$ alors $\rho(G) \leqslant \rho(J) < 1$.

Démonstration,

Soit λ le rayon spectral de G .

On a : $\rho(J) = \rho(L+U) = m(1) = n(1)$.

Par hypothèse $n(1)$ est donc strictement inférieur à 1.

D'après le lemme 3 on a : $1 \leqslant n(\frac{1}{\lambda})$.

D'où :

$$n(1) < 1 \leqslant n(\frac{1}{\lambda}).$$

n étant une fonction croissante, $\frac{1}{\lambda}$ est strictement supérieur à 1.

m étant aussi une fonction croissante, on a :

$$m(\lambda) \leqslant m(1).$$

D'après le lemme 3, λ est inférieur à $m(\lambda)$. Il en résulte l'inégalité suivante :

$$\rho(G) = \lambda \leqslant m(1) = \rho(J) < 1$$

$$\rho(G) \leqslant \rho(J) < 1 .$$

La proposition 1 en résulte :

Si la méthode de Jacobi est convergente pour la matrice A, c'est-à-dire si le rayon spectral de J est strictement inférieur à 1, la méthode de Gauss-Seidel est aussi convergente, puisque d'après la proposition 2, le rayon spectral

de G est alors strictement inférieur à 1.

D'autre part le taux asymptotique de convergence de la méthode de Gauss-Seidel, $R_\infty(G)$ est supérieur à celui de la méthode de Jacobi, $R_\infty(J)$.

On a, en effet :

$$R_\infty(G) = -\log \rho(G) \geqslant -\log \rho(J) = R_\infty(J)$$

la méthode de Gauss-Seidel est asymptotiquement au moins aussi rapide que celle de Jacobi.

Remarque :

Si L et U ne sont plus positives, dans le cas général, il n'y a plus de théorème de comparaison : une des deux méthodes peut converger pour A et l'autre diverger, et si les deux méthodes sont convergentes, l'une ou l'autre peut, suivant A, être plus rapide.

Nous allons étudier, maintenant, un cas particulier important, pour lequel L et U ne sont plus nécessairement positives.

Proposition.

Soit A une matrice à diagonale strictement dominante (resp. à diagonale fortement dominante et irréductible).

Alors les méthodes de Gauss-Seidel et de Jacobi sont convergentes.

Si de plus, L et U sont positives, la méthode de Gauss-Seidel converge asymptotiquement plus vite.

Démonstration.

a) Montrons que la méthode de Jacobi est convergente.

la matrice de l'itération J est égale à $D^{-1}(E+F)$; soit

$$J = (j_{\alpha\beta}) \qquad \begin{cases} j_{\alpha\beta} = 0 & \text{si} \quad \alpha = \beta \\ j_{\alpha\beta} = -\dfrac{a_{\alpha\beta}}{a_{\alpha\alpha}} & \text{si} \quad \alpha \neq \beta \end{cases}$$

D'après le théorème de Gerschögin, nous savons que les valeurs propres de J sont situées dans la réunion des disques de Gerschgörin, centrés en $j_{\alpha\alpha} = 0$ et de rayon $\sum_{\substack{\beta=1 \\ \beta\neq\alpha}}^{n} |j_{\alpha\beta}|$.

$\mathbf{A}$ étant à diagonale dominante on a :

$$\sum_{\substack{\beta=1 \\ \beta\neq\alpha}}^{n} |j_{\alpha\beta}| = \sum_{\substack{\beta=1 \\ \alpha\neq\beta}}^{n} \frac{|a_{\alpha\beta}|}{|a_{\alpha\alpha}|} \leqslant 1 \qquad \forall\alpha$$

et les valeurs propres de J sont donc dans le cercle du plan complexe de centre 0 et de rayon 1.

1) Si $\mathbf{A}$ est à diagonale strictement dominante tous les cercles de Gerschgörin ont un rayon strictement inférieur à 1, toutes les valeurs propres de J ont donc un module strictement inférieur à 1 et

$$\underline{\rho(J) < 1} \ .$$

2) Si $\mathbf{A}$ est à diagonale fortement dominante et irréductible, il existe α tel que :

$$|a_{\alpha\alpha}| > \sum_{\substack{\beta=1 \\ \alpha\neq\beta}}^{n} |a_{\alpha\beta}| . \qquad (1)$$

Supposons qu'il existe une valeur propre λ , de module 1. Alors λ est sur la frontière de la région de gerschgörin et comme $\mathbf{A}$ est irréductible, tous les cercles de gershgörin passent par λ . Tous ces cercles sont donc centrés en 0 et ont un rayon égal à 1, ce qui contredit la relation (1).

Donc $\underline{\rho(J) < 1}$.

Conclusion.

Si A est à diagonale fortement dominante et irréductible (resp. à diagonale strictement dominante) la méthode de Jacobi converge.

b) Montrons que, dans ce cas, la méthode de Gauss-Seidel est aussi convergente.

On a : $$G = (I-L)^{-1}U = (I + L + \ldots + L^{n-1})U$$

Soit $$|G| \leqslant |I + L + \ldots + L^{n-1}| \, |U|$$

$$\leqslant (I + |L| + \ldots + |L^{n-1}|)|U|$$

$$\leqslant (I - |L|)^{-1}|U|$$

On a de plus :

$$|J| = |L| + |U|$$

(on a égalité car L est une matrice strictement triangulaire inférieure et U une matrice strictement triangulaire supérieure).

De même que nous avons montré que J a un rayon spectral inférieur strictement à 1, on montre que :

$$\rho(|J|) < 1 \, .$$

Il en résulte :

$$\rho(I - |L|)^{-1}|U| \leqslant \rho(|L| + |U|) < 1$$

d'après la proposition 1 (convergence des méthodes de Jacobi et Gauss-Seidel dans le cas où L et U sont positives ou nulles).

D'après le théorème de Perron-Frobenius, puisque $|G| \leqslant (I-|L|)^{-1}|U|$, on a :

$$\rho(G) \leqslant \rho(|G|) \leqslant \rho((I-|L|)^{-1}|U|) < 1 .$$

La méthode de Gauss-Seidel est convergente dans ce cas.

5°) Convergence de la méthode de relaxation.

On rappelle que pour la méthode de relaxation, la matrice d'itération est :

$$\mathcal{L}_\omega = (I-\omega L)^{-1}((1-\omega)I+\omega U).$$

Proposition.

On a : $\rho(\mathcal{L}_\omega) \geqslant |\omega-1|$ et l'égalité n'est possible que si toutes les valeurs propres de $\mathcal{L}_\omega$ ont un module égal à $|\omega-1|$.

Démonstration.

Soit $p(\lambda)$ le polynôme caractéristique de $\mathcal{L}_\omega$. On a :

$$\begin{aligned} p(\lambda) &= \det(\lambda I-\mathcal{L}_\omega) \\ &= \det(\lambda I-(I-\omega L)^{-1}((1-\omega)I+\omega U)) \\ &= \det(I-\omega L)^{-1}\det(\lambda(I-\omega L).((1-\omega)I+\omega U)) \\ &= \det(\lambda(I-\omega L)-((1-\omega)I+\omega U)) \end{aligned}$$

car $I-\omega L$ est une matrice triangulaire inférieure dont les termes diagonaux sont tous égaux à 1.

On a :

$$(*) \qquad (\rho(\mathcal{L}_\omega))^n \geqslant \prod_{i=1}^{n} |\lambda_i| = |p(0)| = |\det((\omega-1)I-\omega U)| = |\omega-1|^n$$

car $((\omega-1)I-\omega U)$ est une matrice triangulaire supérieure dont les termes diagonaux sont tous égaux à $\omega-1$.

D'où :

$$\rho(\mathcal{L}_\omega)^n \geqslant |\omega-1|^n \quad , \quad \text{et} \quad \rho(\mathcal{L}_\omega) \geqslant |\omega-1| \ .$$

S'il y a égalité, il y a aussi égalité partout dans l'expression (*). Il en résulte que l'on a :

$$\rho(\mathcal{L}_\omega)^n = \prod_{i=1}^{n} |\lambda_i|$$

et $$|\lambda_i| = \rho(\mathcal{L}_\omega) \quad \forall i \ , \ 1 \leqslant i \leqslant n \ .$$

Toutes les valeurs propres de ω ont donc alors un module égal à $|\omega-1|$.

Corollaire.

La méthode de relaxation ne peut converger que si $\omega \in]0,2[$.

Nous allons étudier la convergence de la méthode de relaxation dans le cas particulier où **A** est une matrice hermitienne.

Si **A** est hermitienne, F est l'adjointe de E . (Dans le cas réel F est la transposée de E).

On a : $$\mathbf{A} = D-E-E^* \ .$$

De plus D est une matrice réelle.

Proposition.

Soit **A** une matrice carrée d'ordre n , hermitienne, non singulière.

Si la matrice D est définie positive, la méthode de relaxation pour **A** est convergente si et seulement si ω appartient à l'intervalle ouvert $]0,2[$ et si **A** est définie positive.

Démonstration.

a) Calcul préliminaire.

Soit $x \in R^n$, $y = \omega x$ et $z = x-y$.

On a : $$y = (D-\omega E)^{-1}((1-\omega)D+\omega E^*)x$$

$$(D-\omega E)y = ((1-\omega)D+\omega E^*)x \ .$$

Soit $$(D-\omega E)z = \omega Dx - \omega E^* x - \omega Ex = \omega(D-E^*-E)x$$

et $$Dy - \omega Ey + \omega Dy - \omega E^* y = (1-\omega)Dx - \omega Dy + \omega E^* z \ .$$

D'où

$$\begin{cases} (D-\omega E)z = \omega A x \\ (1-\omega)Dz + \omega E^* z = \omega A y \ . \end{cases}$$

En multipliant scalairement la première égalité par x , la seconde par y et en soustrayant membre à membre, on obtient la relation :

(1) $$\omega(x,Ax) - \omega(y,Ay) = ((D-\omega E)z,x) - (1-\omega)(Dz,y) - \omega(E^* z,y) \ .$$

Le second membre est égal à :

(2) $$= (Dz,z) - \omega(Ez,x) + \omega(Dz,y) - \omega(E^* z,y)$$

D étant réelle positive, (2) s'écrit :

(2) $$= (Dz,z) - \omega(z,E^* x) + \omega(Dy,z) - \omega(z,Ey)$$

$$= Dz - \omega(z,(E^* x+Ey)) + \omega(Dy,z) \ .$$

De la relation trouvée précédemment

$$(1-\omega)Dx + \omega E^* x + \omega E_y = Dy \ ,$$

il résulte que l'on a :

$$(2) \qquad = (Dz,z) + \omega(Dy,z) - (Dy,z) + (1-\omega)(Dx,z)$$

$$= (Dz,z) - \omega\, D(z,z) + (Dz,z)$$

$$= (2-\omega)(Dz,z).$$

On a donc le résultat suivant :

$$(2) \qquad \begin{cases} (2-\omega)(Dz,z) - \omega(X,Ax) - \omega(y,Ay) \\ \text{où} \quad x \in R^n , \quad y = \mathcal{L}_\omega\, x \quad \text{et} \quad z = x-y \end{cases}$$

b) Supposons A définie positive et ω appartenant à l'intervalle ouvert $]0,2[$.

Nous allons montrer que la méthode de Relaxation est alors convergente.

Soit x vecteur propre de $\mathcal{L}_\omega$. On a alors :

$y = \mathcal{L}_\omega\, x = \lambda x$ où λ est la valeur propre associée x .

Le résultat (2) trouvé dans le calcul préliminaire devient :

$$(3) \qquad (2-\omega)|1-\lambda|^2(Dx,x) = \omega(1-|\lambda|^2)(x,Ax).$$

Il est exclu que λ soit égal à 1 ; sinon la relation $(D-\omega E)z = \omega Ax$ donnerait $\omega Ax = 0$, soit, ω n'étant pas nul, $Ax = 0$ et A serait singulière.

Le membre de gauche de (3) est donc strictement positif puisque D est définie positive et $\omega < 2$. A étant aussi définie positive, il résulte de (3) que :

$$1-|\lambda|^2 > 0 .$$

Soit

$$|\lambda| < 1 .$$

Le rayon spectral de $\mathcal{L}_\omega$ est donc strictement inférieur à 1. Ce qui entraîne la convergence de la méthode.

c) Réciproquement, supposons que la méthode de relaxation soit convergente et que la matrice A ne soit pas définie positive.

Alors le vecteur initial de l'itération pourra être choisi en sorte que l'erreur initiale $\varepsilon^{(0)}$ vérifie

$$(A\varepsilon^{(0)}, \varepsilon^{(0)}) < 0 .$$

L'égalité (2) avec x remplacé par $\varepsilon^{(0)}$ et y par $\varepsilon^{(1)}$ (on a bien $\varepsilon^{(m+1)} = B\varepsilon^{(m)} = \mathcal{L}_\omega \varepsilon^{(m)}$) entraîne

$$(4) \qquad (2-\omega)(D(\varepsilon^{(0)}-\varepsilon^{(1)}), \varepsilon^{(0)}-\varepsilon^{(1)}) = \omega(\varepsilon^{0}, A\varepsilon^{(0)}) - \omega(\varepsilon^{(1)}, A\varepsilon^{(1)})$$

1 n'étant pas valeur propre de $\mathcal{L}_\omega$, $\varepsilon^{(0)}$ est différent de $\varepsilon^{(1)}$.

Le premier membre de (4) est donc strictement positif. Il en résulte que :

$$0 \geqslant (\varepsilon^{0}, A\varepsilon^{(0)}) > (\varepsilon^{(1)}, A\varepsilon^{(1)})$$

on vérifiera de même que :

$$0 \geqslant (\varepsilon^{(m)}, A\varepsilon^{(m)}) > (\varepsilon^{(m+1)}, A\varepsilon^{(m+1)}) \quad \forall m .$$

Cela est contraire au fait que $\varepsilon^{(m)}$ tend vers 0 lorsque $m \to \infty$ et donc que $(\varepsilon^{m}, A\varepsilon^{(m)})$ tend vers 0 lorsque m tend vers l'infini.

A est donc définie positive.

6°) Méthodes itératives par blocs.

a) Introduction.

Soit A une matrice carrée $n \times n$ qui est supposée donnée sous la forme suivante (matrice bloc) :

$$A = \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1p} \\ A_{21} & & & \vdots \\ \vdots & & & \vdots \\ A_{p1} & \cdots & \cdots & A_{pp} \end{pmatrix}$$

où chaque A_{ij} est d'ordre $n_i \times n_j$ et telle que : $\sum_{i=1}^{p} n_i = n$

Pour tout i, la matrice A_{ii} est donc carrée d'ordre n_i. Pour une telle matrice on pose :

$$\hat{D} = \begin{pmatrix} A_{11} & & & 0 \\ & A_{22} & & \\ & & \ddots & \\ 0 & & & A_{pp} \end{pmatrix} \qquad \hat{E} = \begin{pmatrix} 0 & & & \\ -A_{2,1} & 0 & & 0 \\ & -A_{3,2} & 0 & \\ & & \ddots & \ddots \\ -A_{p,1} & & -A_{p,p-1} & 0 \end{pmatrix}$$

$$\hat{F} = \begin{pmatrix} 0 & -A_{12} & \cdots & -A_{1p} \\ & 0 & -A_{23} & \vdots \\ & & 0 & \\ 0 & & \ddots & -A_{p-1,p} \\ & & & 0 \end{pmatrix}$$

(Comparer à la décomposition ponctuelle $A = D-E-F$).

Utilisant la décomposition, $A = \hat{D}-\hat{E}-\hat{F}$, de A, on va définir des méthodes itératives de résolution d'un système linéaire $Ax = b$.

Ces méthodes seront appelées méthodes itératives par blocs ; et, par opposition, les méthodes itératives précédemment définies sont appelées méthodes itératives ponctuelles.

On supposera toujours que $\hat{D}$ est inversible. Dans la partie du cours concernant les matrices réductibles, on a démontré (lemme 2) que si une matrice A était bloc-triangulaire supérieure alors

$$\det A = \prod_{i=1}^{p} \det A_{ii} .$$

On a donc : $\det \hat{D} = \prod_{i=1}^{p} \det A_{ii}$.

$\hat{D}$ inversible est donc équivalent à : tous les A_{ii} sont inversibles.

b) Méthode de Jacobi par blocs.

On pose $\begin{cases} M = \hat{D} \\ N = \hat{E} + \hat{F} \end{cases}$

L'itération $M X^{(m+1)} = N X^{(m)} + b$ s'écrit alors : $\hat{D} X^{(m+1)} = \hat{E} X^{(m)} + \hat{F} X^{(m)} + b$

alors que dans la méthode de Jacobi ponctuelle on avait :

$D X^{(m+1)} = E X^{(m)} + F X^{(m)} + b$.

On écrit alors les vecteurs X , b de $\mathbb{R}^n$ sous la forme

$$X = \begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix} \qquad b = \begin{pmatrix} b_1 \\ \vdots \\ b_p \end{pmatrix}$$

où X_i et b_i sont des vecteurs de $\mathbb{R}^{n_i}$.

Les formules d'itération deviennent donc :

$$A_{ii} X_i^{(m+1)} = - \sum_{\substack{j=1 \\ j \neq i}}^{n} A_{ij} X_j^{(m)} + b_i$$

Comme les A_{ii} sont inversibles, cela s'écrit aussi :

$$X_i^{(m+1)} = - \sum_{\substack{j=1 \\ j \neq i}}^{n} A_{ii}^{-1} A_{ij} X_j^{(m)} + A_{ii}^{-1} b_i .$$

La matrice de l'itération est $\hat{J} = M^{-1}N$

$$\hat{J} = M^{-1}N = \hat{D}^{-1}(\hat{E}+\hat{F}) = \hat{L}+\hat{U}$$

où $\begin{cases} \hat{L} = \hat{D}^{-1}E \\ \hat{U} = \hat{D}^{-1}F \end{cases}$

c) <u>Méthode de Gauss-Siedel par blocs</u>.

On pose $\begin{cases} M = \hat{D} - \hat{E} \\ N = \hat{F} \end{cases}$

L'itération $M\,X^{(m+1)} = N\,X^{(m)} + b$ s'écrit alors $(\hat{D}-\hat{E})X^{(m+1)} = \hat{F}\,X^{(m)} + b$.

On a donc pour i compris entre 1 et p :

$$\sum_{j=1}^{i} A_{ij}\,X_j^{(m+1)} = -\sum_{j=i+1}^{p} A_{ij}\,X_j^{(m)} + b_i$$

$$X_i^{(m+1)} = -\sum_{j=1}^{i-1} A_{ii}^{-1}\,A_{ij}\,X_j^{(m+1)} - \sum_{j=i+1}^{p} A_{ii}^{-1}\,A_{ij}\,X_j^{(m)} + A_{ii}^{-1}\,b_i \ .$$

On calcule $X_1^{(m+1)},\ldots,X_p^{(m+1)}$ dans cet ordre ; connaissant $X_1^{(m+1)},\ldots,X_{i-1}^{(m+1)}$ on en déduit $X_i^{(m+1)}$ par cette formule.

La matrice de l'itération est $\hat{G} = M^{-1}N$

$$\hat{G} = (\hat{D}-\hat{E})^{-1}\hat{F} = (I - \hat{D}^{-1}\hat{E})^{-1}\,\hat{D}^{-1}\,\hat{F}$$

$$= (I-\hat{L})^{-1}.\hat{U}$$

avec $\begin{cases} \hat{L} = \hat{D}^{-1}\,\hat{E} \\ \hat{U} = \hat{D}^{-1}\,\hat{F} \end{cases}$ comme précédemment.

d) Méthode de relaxation.

Soit ω un nombre réel non nul.

On pose $\begin{cases} M = \frac{1}{\omega}\hat{D} - \hat{E} \\ N = (\frac{1}{\omega} - 1)\hat{D} + \hat{F} \end{cases}$

L'itération s'écrit alors :

$$(\frac{1}{\omega}\hat{D} - \hat{E})X^{(m+1)} = [(\frac{1}{\omega} - 1)\hat{D} + \hat{F}]X^{(m)} + b .$$

La matrice de l'itération est $\hat{\mathcal{L}}_\omega = M^{-1}N$

$$\hat{\mathcal{L}}_\omega = (I - \omega\hat{L})^{-1}[(1-\omega)I + \omega\hat{U}] .$$

7°) Convergence des méthodes de Jacobi et Gauss-Siedel par blocs.

Définition : Soit A une matrice-blocs $A = (A_{ij})$ $1 \leqslant i,j \leqslant p$ où A_{ij} est une matrice à n_i lignes, n_j colonnes et telle que $\sum_{i=1}^{p} n_i = n$. On dit que A est bloc-tridiagonale si pour j strictement plus grand que $i+1$, ou strictement plus petit que $i-1$, A_{ij} est nul ; A est de la forme suivante :

$$A = \begin{pmatrix} A_{11} & A_{12} & & & 0 \\ A_{21} & A_{22} & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & A_{p-1\,p} \\ 0 & & & A_{p,p-1} & A_{pp} \end{pmatrix}$$

Lemme 1 : Soient μ un nombre réel non nul et A une matrice bloc tridiagonale. On pose $A(\mu) = \hat{D} - \mu\hat{E} - \frac{1}{\mu}\hat{F}$. Alors $\det A(\mu) = \det A$.

Démonstration :

Posons $Q_\mu = \begin{pmatrix} \mu I_1 & & & \\ & \mu^2 I_2 & & 0 \\ & & \ddots & \\ & 0 & & \mu^p I_p \end{pmatrix}$ où I_i est la matrice unité de R^{n_i}

D'après le lemme déjà cité dans l'introduction $\det Q_\mu = \mu^{1+2+\ldots+p} \neq 0$ car $\mu \neq 0$.

La multiplication de A par Q_μ à gauche revient à multiplier les blocs-lignes par $\mu, \mu^2, \ldots, \mu^p$. En effet :

$$(Q_\mu A)_{ij} = \sum_{k=1}^{p} (Q_\mu)_{ik} A_{kj} = \mu^i A_{ij} .$$

De même la multiplication d'une matrice B par Q_μ à droite, revient à multiplier les blocs-colonnes de B par $\mu, \mu^2, \ldots, \mu^p$

$$(B Q_\mu)_{i,j} = \sum_{k=1}^{p} B_{i,k} (Q_\mu)_{k,j} = \mu^j B_{i,j} .$$

Dans ces conditions :

$$(Q_\mu A Q_\mu^{-1})_{i,j} = (Q_\mu A Q \tfrac{1}{\mu})_{i,j} = \frac{\mu^i}{\mu^j} A_{ij} .$$

Si $j = i$ $\quad (Q_\mu A Q_\mu^{-1})_{i,i} = A_{i,i}$

Si $j = i-1$ $\quad (Q_\mu A Q_\mu^{-1})_{i,i-1} = \mu . A_{i,i-1}$

Si $j = i+1$ $\quad (Q_\mu A Q_\mu^{-1})_{i,i+1} = \dfrac{A_{i,i+1}}{\mu}$

Si $j > i+1$ ou $j < i-1$ $\quad (Q_\mu A Q_\mu^{-1})_{i,j} = 0$.

Donc $A(\mu) = Q_\mu A Q_\mu^{-1}$.

Et $\det A(\mu) = \det Q_\mu . \det A . \det Q_\mu^{-1} = \det A$ ce qui achève la démonstration.

Proposition : Soit A une matrice bloc-tridiagonale. Les méthodes de Gauss-Siedel et Jacobi divergent ou convergent simultanément ; et lorsqu'elles convergent, la méthode de Gauss-Siedel converge asymptotiquement deux fois plus vite que la méthode de Jacobi.

Démonstration : Le polynôme caractéristique de $\hat{J}$ s'écrit $P_{\hat{J}}(\lambda)$:

$$P_{\hat{J}}(\lambda) = \det(\lambda I - \hat{J}) = \det(\lambda I - \hat{L} - \hat{U}).$$

Mais $\det \hat{D}.P_{\hat{J}}(\lambda) = \det \hat{D} \det(\lambda I - \hat{L} - \hat{U}) = \det(\lambda\hat{D} - \hat{E} - \hat{F})$.

On applique le lemme 1 pour $\mu = -1$.

On trouve : $\det \hat{D} \cdot P_{\hat{J}}(\lambda) = \det(\lambda\hat{D} + \hat{E} + \hat{F}) = (-1)^n \det(-\lambda\hat{D} - \hat{E} - \hat{F})$.

Donc si λ est valeur propre de $\hat{J}$, alors $-\lambda$ est valeur propre de $\hat{J}$.

$P_{\hat{J}}(\lambda)$ peut donc s'écrire sous la forme : $P_{\hat{J}}(\lambda) = p_r(\lambda^2)\, \lambda^{n-2r}$ où p_r est un polynôme de degré r .

A présent le polynôme caractéristique de $\hat{G}$ s'écrit $P_{\hat{G}}(\lambda)$:

$$P_{\hat{G}}(\lambda) = \det[\lambda I - (\hat{D}-\hat{E})^{-1} \hat{F}] .$$

Mais $\det(\hat{D}-\hat{E})P_{\hat{G}}(\lambda) = \det(\lambda(\hat{D}-\hat{E}) - \hat{F}) = \det(\lambda\hat{D} - \lambda\hat{E} - \hat{F})$

$$= \det[\lambda^{\frac{1}{2}}(\lambda^{\frac{1}{2}} \hat{D} - \lambda^{\frac{1}{2}} \hat{E} - \frac{1}{\lambda^{\frac{1}{2}}} \hat{F})] = \lambda^{n/2} \det[\lambda^{\frac{1}{2}} \hat{D} - \lambda^{\frac{1}{2}} \hat{E} - \frac{1}{\lambda^{\frac{1}{2}}} \hat{F}] .$$

Si λ est complexe $\lambda^{\frac{1}{2}}$ désigne une des deux racines carrées complexes de λ .

On applique alors le lemme 1

$$\det(\hat{D}-\hat{E})P_{\hat{G}}(\lambda) = \lambda^{n/2} \det[\lambda^{\frac{1}{2}} \hat{D} - \hat{E} - \hat{F}] = \lambda^{n/2} \det \hat{D} \cdot P_{\hat{J}}(\lambda^{\frac{1}{2}})$$

$$P_{\hat{G}}(\lambda) = \lambda^{\frac{n}{2}} \cdot \lambda^{\frac{n}{2}-r}\, p(\lambda)$$

$$P_{\hat{G}}(\lambda) = \lambda^{n-r}\, p(\lambda^{\frac{1}{2}}).$$

Ainsi, si η est une valeur propre non nulle de $\hat{G}$, les deux racines carrées de η sont valeurs propres de $\hat{J}$ et réciproquement si λ est valeur propre de $\hat{J}$ ($-\lambda$ l'est aussi) et λ^2 est valeur propre de $\hat{G}$. Donc

$$\rho(\hat{G}) = [\rho(\hat{J})]^2 \quad \text{et en particulier : } \rho(\hat{G}) < 1 \Longleftrightarrow \rho(\hat{J}) < 1 .$$

On a $-\text{Log } \rho(\hat{J}) = \frac{1}{2} [-\text{Log } \rho(\hat{G})]$ et pour les taux asymptotiques de convergence :

$$R_\infty(\hat{J}) = \tfrac{1}{2} R_\infty(\hat{G})$$

ce qui achève la démonstration.

8°) Convergence de la méthode de relaxation par blocs.

Proposition : Soit A une matrice carrée d'ordre n décomposée en blocs A_{ij} d'ordre $n_i \times n_j$. Si la méthode de relaxation converge pour ω alors on a : $0 < \omega < 2$.

Démonstration : Elle est absolument identique à celle faite pour la méthode de relaxation ponctuelle.

Proposition : Soit A une matrice bloc, régulière hermitienne, telle que $\hat{D}$ soit définie positive. Alors la méthode de relaxation par blocs converge si et seulement si :

$$\begin{cases} A \text{ est définie positive et} \\ 0 < \omega < 2 . \end{cases}$$

Démonstration : Elle est absolument identique à celle faite pour la méthode de relaxation ponctuelle.

Cas des matrices tridiagonales par blocs.

On considère à nouveau le cas particulier, très important en pratique, des matrices tridiagonales par blocs.

$$A = \begin{pmatrix} A_{11} & A_{12} & & 0 \\ A_{21} & A_{22} & \ddots & \\ & \ddots & \ddots & A_{p-1\,p} \\ 0 & & A_{p\,p-1} & A_{pp} \end{pmatrix}$$

où A_{ij} est d'ordre $n_i \times n_j$ et $\sum_{i=1}^{p} n_i = n$

Proposition : Soit ω différent de 0 et de 1. Si λ est valeur propre de $\hat{J}$ et si η vérifie la relation : $\omega^2 \lambda^2 = \frac{(\eta+\omega-1)^2}{\eta}$ alors η est valeur propre de $\hat{\mathcal{L}}_\omega$ et réciproquement si η est valeur propre de $\hat{\mathcal{L}}_\omega$ et si λ vérifie cette relation alors λ est valeur propre de $\hat{J}$.

Démonstration : Rappelons que $\hat{\mathcal{L}}_\omega = (I-\omega\hat{L})^{-1}[(1-\omega)I+\omega\hat{U}]$.

On écrit le polynôme caractéristique de $\hat{\mathcal{L}}_\omega$: $P_{\hat{\mathcal{L}}_\omega}(\lambda) = P(\lambda)$

$$P(\eta) = \det[\eta I - (\hat{D}-\omega\hat{E})^{-1}((1-\omega)\hat{D}+\omega\hat{F})].$$

Notons déjà que 0 ne peut être valeur propre de $\hat{\mathcal{L}}_\omega$. En effet si $\eta=0$ était valeur propre on aurait : $\det[(1-\omega)\hat{D}+\omega\hat{F}] = 0$.

Or ce déterminant est égal à $(1-\omega)^n \prod_{i=1}^{n} \det A_{ii}$ qui n'est pas nul pour ω différent de 1 (on suppose toujours que les blocs diagonaux A_{ii} sont tous réguliers).

On a alors :

$$P_{\hat{\mathcal{L}}_\omega}(\eta) \times \det(\hat{D}-\omega\hat{E}) = \det[\eta(\hat{D}-\omega\hat{E}) - ((1-\omega)\hat{D}+\omega\hat{F})]$$

$$= \det[(\eta+\omega-1)\hat{D} - \omega\eta\hat{E} - \omega\hat{F}]$$

$$= \det[\eta^{\frac{1}{2}}\omega \{\frac{\eta+\omega-1}{\eta^{\frac{1}{2}}\omega}\hat{D} - \eta^{\frac{1}{2}}\hat{E} - \frac{1}{\eta^{\frac{1}{2}}}\hat{F}\}] .$$

On a donc $P_{\hat{\mathcal{L}}_\omega}(\eta) = C\,\omega^n\,\eta^{\frac{n}{2}}\det\left[\frac{\eta+\omega-1}{\omega\,\eta^{\frac{1}{2}}}\hat{D} - \eta^{\frac{1}{2}}\hat{E} - \frac{1}{\eta^{\frac{1}{2}}}\hat{F}\right]$.

D'après le lemme 1 cela s'écrit :

$$P_{\hat{\mathcal{L}}_\omega}(\eta) = C\,\omega^n\,\eta^{n/2}\det\left[\frac{\eta+\omega-1}{\omega\,\eta^{\frac{1}{2}}}\hat{D} - \hat{E} - \hat{F}\right]$$

$$= C'\omega^n\,\eta^{n/2}\,P_{\hat{J}}\left(\frac{\eta+\omega-1}{\omega\,\eta^{\frac{1}{2}}}\right).$$

(Si η est complexe, $\eta^{\frac{1}{2}}$ désigne l'une quelconque de ses deux racines carrées complexes).

Si η est valeur propre de $\hat{\mathcal{L}}_\omega$ alors η est non nul et $\frac{\eta+\omega-1}{\omega\,\eta^{\frac{1}{2}}}$ est une valeur propre de $\hat{J}$.

Réciproquement si λ est valeur propre de $\hat{J}$ et si η vérifie la relation $\frac{(\eta+\omega-1)^2}{\eta} = \omega^2\lambda^2$ alors η est valeur propre de $\hat{\mathcal{L}}_\omega$.

Ce qui achève la démonstration.

Méthode itérative par blocs : Recherche du paramètre optimal.

Nous allons résoudre à présent, pour le cas d'une méthode de relaxation par blocs pour une matrice tridiagonale par blocs, le problème fondamental du choix du paramètre optimal : c'est-à-dire le ω tel que $R_\infty(\mathcal{L}_\omega)$ soit le plus petit possible, ou donc parmi les ω tels que $\rho(\hat{\mathcal{L}}_\omega) < 1$; le paramètre ω tel que $\rho(\hat{\mathcal{L}}_\omega)$ soit le plus petit possible.

On a le résultat suivant :

Théorème : Si le rayon spectral de $\hat{J}$ est strictement inférieur à 1 et si les valeurs propres de $\hat{J}$ sont réelles, alors la valeur de ω qui rend $\rho(\hat{\mathcal{L}}_\omega)$ minimum est $\omega^* = \frac{2}{1+\sqrt{1-[\rho(\hat{J})]^2}}$, et alors $\rho(\hat{\mathcal{L}}_{\omega_*}) = \frac{1-\sqrt{1-(\rho(\hat{J}))^2}}{1+\sqrt{1-(\rho(\hat{J}))^2}}$.

Démonstration.

Pour $\omega = 1$ $\hat{G} = \hat{\mathcal{L}}_1$.

D'après une proposition énoncée précédemment

$$\rho(\hat{J}) = [\rho(\hat{G})]^2 \Longrightarrow \rho(\hat{J}) = [\rho(\hat{\mathcal{L}}_1)]^2 .$$

Donc si $\rho(\hat{J}) < 1$ on a $\rho(\hat{\mathcal{L}}_1) < 1$ et il existé donc des ω tels que $\rho(\hat{\mathcal{L}}\omega) < 1$. On va chercher le paramètre ω qui rend $\rho(\hat{\mathcal{L}}\omega)$ le plus petit possible.

Les valeurs propres η de $\hat{\mathcal{L}}\omega$ vérifient : $(\eta+\omega-1)^2 = \lambda^2\omega^2\eta$ où λ est valeur propre de $\hat{J}$

$$\eta^2 + (2\omega-2-\omega^2\lambda^2)\eta + (\omega-1)^2 = 0 .$$

Calculons le discriminant Δ de ce trinôme en η

$$\Delta = (2\omega-2-\omega^2\lambda^2)^2 - 4(\omega-1)^2$$

$$= [2\omega-2-\omega^2\lambda^2-2(\omega-1)][2\omega-2-\omega^2\lambda^2+2(\omega-1)]$$

$$= \omega^2\lambda^2[\omega^2\lambda^2 - 4(\omega-1)].$$

Le signe de Δ est le même que celui de $\omega^2\lambda^2 - 4(\omega-1)$ que l'on va étudier

$\omega^2\lambda^2 - 4(\omega-1)$ est un trinôme du second degré en ω de discriminant δ .

$\delta = 16(1-\lambda^2) > 0$ car λ appartient au spectre de $\hat{J}$ et on a supposé $\rho(\hat{J}) < 1$.

$\omega^2\lambda^2 - 4(\omega-1)$ a donc deux racines réelles qui sont ω_1 et ω_2 . Leur somme et leur produit étant positifs, elles sont positives. En outre :

pour $\omega = 1$ $\qquad [\omega^2\lambda^2 - 4(\omega-1)]_{(1)} = \lambda^2 > 0$

pour $\omega = 2$ $\qquad 4\lambda^2 - 4 = 4(\lambda^2-1) < 0$.

On a donc : $0 < 1 < \omega_1 < 2 < \omega_2$.

D'où le signe de Δ pour $\omega \in]0,2[$ c'est-à-dire les seules valeurs de ω pour lesquelles $\rho(\hat{\mathcal{L}}_\omega) < 1$:

ω	0		1		ω_1		2
Δ	0	+		+	0	−	

On va construire la courbe des variations de $\rho(\hat{\mathcal{L}}_\omega)$ en fonction de ω. On considère deux cas :

1er cas $\omega > \omega_1$.

Alors $\Delta < 0$.

$$\eta = \frac{\omega^2\lambda^2 - 2(\omega-1) \pm i\sqrt{-\Delta}}{2}$$

d'où $4\,|\eta|^2 = [\omega^2\lambda^2 - 2(\omega-1)]^2 - \Delta$

$$= [\omega^2\lambda^2 - 2(\omega-1)]^2 - \omega^2\lambda^2[\omega^2\lambda^2 - 4(\omega-1)]$$

$$= [\omega^2\lambda^2 - 2(\omega-1)][\omega^2\lambda^2 - 2(\omega-1) - \omega^2\lambda^2] + 2\omega^2\lambda^2(\omega-1)$$

$$= (\omega-1)[2\omega^2\lambda^2 - 2\omega^2\lambda^2 + 2(\omega-1)]$$

$$= 4(\omega-1)^2$$

$|\eta| = \omega-1$ car $\omega > \omega_1 > 1$.

Donc dans ce cas $\rho(\hat{\mathcal{L}}_\omega) = \omega-1$.

Pour ω compris entre ω_1 et 2, la courbe des variations de $\rho(\hat{\mathcal{L}}_\omega)$ en fonction de ω sera donc la droite $y = \omega-1$.

2ème cas $0 < \omega < \omega_1$.

Alors $\Delta > 0$. Les racines en η sont réelles et vérifient

$(\eta+\omega-1)^2 = \lambda^2\,\omega^2\,\eta$.

Les valeurs propres de $\hat{J}$ sont réelles. Le membre de gauche est réel positif. Le membre de droite l'est donc aussi. Donc les racines en η sont positives.

$\eta = \dfrac{\omega^2\lambda^2 - 2\omega + 2 \pm \sqrt{\Delta}}{2}$, λ est réel.

On peut supposer $\lambda > 0$ puisqu'on sait que λ et $-\lambda$ sont toutes deux valeurs propres et donnent le même η . D'autre part, comme on s'intéresse au rayon spectral de $\hat{\mathcal{L}}_\omega$, on ne s'intéresse qu'à la plus grande des deux valeurs propres associées à λ , c'est-à-dire $\eta_2 = \dfrac{\omega^2\lambda^2 - 2(\omega-1) + \sqrt{\omega^2\lambda^2[\omega^2\lambda^2 - 4(\omega-1)]}}{2}$.

On a : $\omega^2\lambda^2 \geqslant 0$ $\sqrt{\omega^2\lambda^2(\omega^2\lambda^2 - 4(\omega-1)}} \geqslant 0$ donc $\eta_2 \geqslant \dfrac{-2(\omega-1)}{2}$.

Donc $\eta_2 + \omega - 1 \geqslant 0$.

Et alors comme : $(\eta_2 + \omega - 1)^2 = \eta_2\ \omega^2\ \lambda^2$ ceci est équivalent à

$\sqrt{\eta_2}\ \omega\ \lambda = \eta_2 + \omega - 1$ (les deux membres sont $\geqslant 0$).

On pose ici $\varphi = \sqrt{\eta_2}$ on a :

$(*)$ $\varphi^2 - \varphi\omega\ \lambda + \omega - 1 = 0$.

φ est fonction de ω et de λ ; pour $\omega = 0$,

$\varphi_\lambda(0) = 1$.

Etudions les variations de φ en fonction de ω .

On différentie $(*)$: $2\varphi\ d\varphi + d\omega = \lambda\omega\ d\varphi + \lambda\varphi\ d\omega$ d'où :

$$\frac{d\varphi}{d\omega} = \frac{\lambda\varphi - 1}{2\varphi - \lambda\omega} .$$

Pour $\omega = 0$ $\dfrac{d\varphi}{d\omega}(0) = \dfrac{\lambda-1}{2} < 0$ car $\lambda \leqslant \rho(\hat{J}) < 1$. $2\varphi - \lambda\omega = \sqrt{\omega^2\lambda^2 - 4(\omega-1)} \geqslant 0$.

Comme $\dfrac{d\varphi}{d\omega}(0) < 0$ il existe $0 < \varepsilon_1 \leqslant \omega_1$ tel que sur $[0, \varepsilon_1]$, $\varphi(\omega)$ soit décroissant.

Donc sur $[0, \varepsilon_1]$ $\lambda\varphi(\omega) - 1$ reste négatif.

Donc $\dfrac{d\varphi}{d\omega}(\varepsilon_1) < 0$

$\dfrac{d\varphi}{d\omega}$ reste négatif et son module croit.

On peut construire ainsi une suite croissante ε_n $0 < \varepsilon_n \leqslant \omega_1$ telle que pour tout n : $\varphi(\omega)$ soit décroissant sur $[0,\varepsilon_n]$.

La suite ε_n a une limite ε, puisqu'elle est croissante et majorée. Alors $\frac{d\varphi}{d\omega}(\varepsilon) \leqslant 0$.

En effet $\frac{d\varphi}{d\omega}(\varepsilon) = \lim_{n\to\infty} \frac{d\varphi}{d\omega}(\varepsilon_n) \leqslant 0$.

Supposons $\frac{d\varphi}{d\omega}(\varepsilon) = 0$ alors $\lambda\,\varphi(\varepsilon)-1 = 0$ or $\lambda\varphi(\varepsilon)-1 \leqslant \lambda\varphi(0)-1 = \lambda-1 < 0$.

Donc $\frac{d\varphi}{d\omega}(\varepsilon) < 0$.

Alors ε ne peut être différent de ω_1 ; en effet si $\varepsilon < \omega_1$, comme $\frac{d\varphi}{d\omega}(\varepsilon) < 0$ il existerait ε' $\varepsilon < \varepsilon' \leqslant \omega_1$ tel que $\varphi(\omega)$ soit décroissant sur $[0,\varepsilon']$ or ε est la limite croissante de tels ε'.

$$\varphi = \frac{\omega\lambda + \sqrt{\omega^2\lambda^2 - 4(\omega-1)}}{2} \geqslant \frac{\omega\lambda}{2}$$

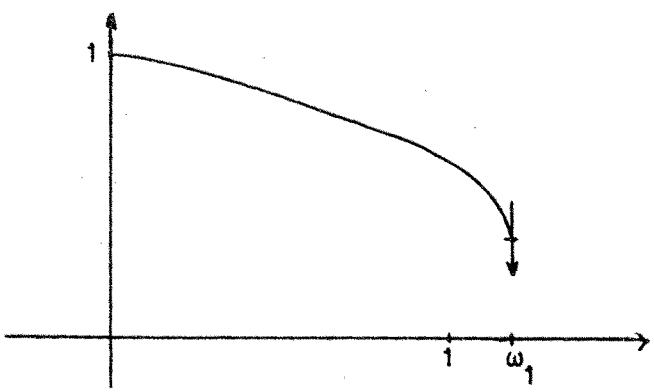
quand $\omega \to \omega_1$ $\quad \frac{d\varphi}{d\omega} \to -\infty$.

Pour $\omega = \omega_1$ $\quad \lambda^2\omega^2 - 4(\omega-1) = 0$

$$\Longrightarrow \lambda^2\omega^2 - 2(\omega-1) = \lambda^2\omega^2 - 4(\omega-1) + 2(\omega-1) = 2(\omega-1)$$

$$\Longrightarrow \eta_2 = \omega_1 - 1 .$$

On peut donc tracer la courbe de variation de η_2 en fonction de ω pour ω compris entre 0 et ω_1



Pour ω fixé entre 0 et ω_1 ; $\omega^2\lambda^2 - 4(\omega-1) > 0$.

Donc $\omega^2\lambda^2 - 4(\omega-1)$ croît avec λ . De même $\omega\lambda$ croît avec λ

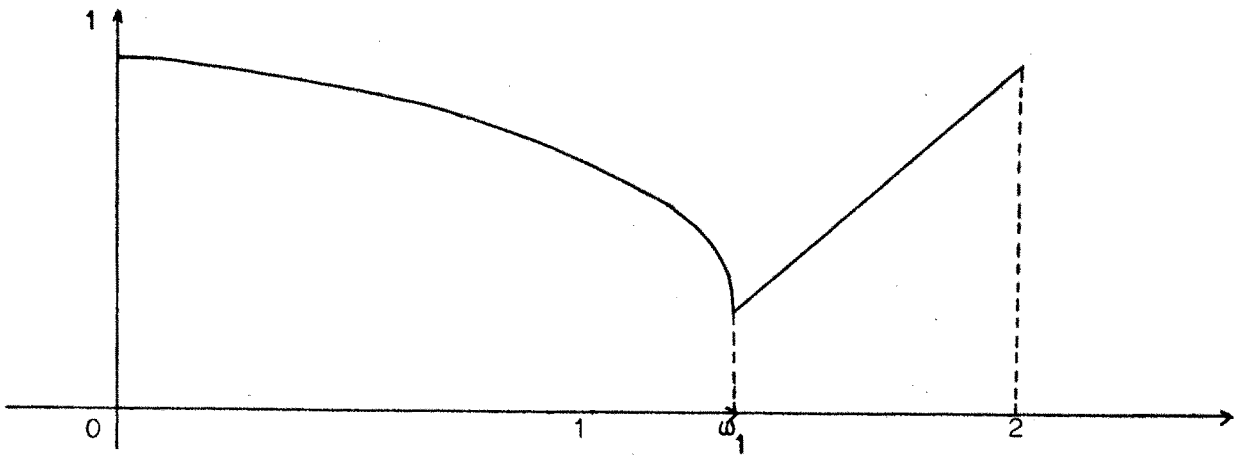
$$\varphi_\lambda(\omega) = \frac{\omega\lambda + \sqrt{\omega^2\lambda^2 - 4(\omega-1)}}{2} .$$

$\varphi_\lambda(\omega)$ croît avec λ . Comme on a choisi λ valeur propre positive de $\hat{J}$, $\varphi_\lambda(\omega)$ est maximum pour $\lambda = \rho(\hat{J})$.

($\rho(\hat{J})$ est valeur propre car les valeurs propres de $\hat{J}$ sont réelles et invariantes par symétrie par rapport à l'origine).

Finalement $\rho(\hat{\mathcal{L}}\omega) = [\varphi_\lambda(\omega)]^2$ lorsque $\lambda = \rho(\hat{J})$.

On peut maintenant tracer la courbe : $\rho(\hat{\mathcal{L}}\omega)$ en fonction de ω ; c'est celle de $|\eta|$ en fonction de ω pour $\lambda = \rho(\hat{J})$



$\rho(\hat{\mathcal{L}}\omega)$ est minimum pour $\omega = \omega_1[\rho(\hat{J})]$.

On appelle cette valeur ω_* .

ω_* est racine de $\sqrt{\omega^2\lambda^2 - 4(\omega-1)}$ pour $\lambda = \rho(\hat{J})$.

Soit $\omega_* = \dfrac{4 - \sqrt{16(1-\rho^2(\hat{J}))}}{2\,\rho^2(\hat{J})}$

$$\omega_* = \frac{2 - 2\sqrt{1-[\rho(\hat{J})]^2}}{[\rho(\hat{J})]^2} = \frac{2}{[\rho(\hat{J})]^2} \times \frac{1-(1-\rho^2(\hat{J}))}{1+\sqrt{1-\rho^2(\hat{J})}} .$$

Finalement $\omega_* = \dfrac{2}{1+\sqrt{1-[\rho(\hat{J})]^2}}$ et alors $\rho(\mathcal{L}_{\omega_*}) = \omega_* - 1 = \dfrac{2}{1+\sqrt{1-\rho^2(\hat{J})}} - 1$

$$\rho(\mathcal{L}_{\omega_*}) = \frac{1-\sqrt{1-[\rho(\hat{J})]^2}}{1+\sqrt{1-[\rho(\hat{J})]^2}} .$$

Conclusion : Relaxation avec paramètre optimal.

Sous les conditions du théorème qu'on vient de démontrer, on commence par chercher $\rho(\hat{J})$ et on calcule ensuite ω_* . Il vaut mieux surestimer ω_* que le sous-estimer car dans le graphe de $\omega \mapsto \rho(\mathcal{L}_\omega)$ la pente de la tangente à gauche de ω_* est infinie, et une petite variation à gauche de ω_* produit une grande variation de $\rho(\mathcal{L}_\omega)$; par contre à droite de ω_* , une variation de ε sur ω_* produira une variation de $O(\varepsilon)$ sur $\rho(\mathcal{L}_{\omega_*})$.

§IV. Notions sur le conditionnement.

Certaines matrices d'apparence simple peuvent donner lieu à des difficultés inattendues pour la résolution de systèmes linéaires : Des petites variations des données produisent de grandes variations sur les solutions. Et ceci est fort gênant puisque tous les calculs sur ordinateurs se font avec des erreurs d'arrondi. Voici un exemple simple très artificiel mais néanmoins très significatif :

on considère la matrice A :

$$A = \begin{pmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{pmatrix}$$

On a $\det A = 1$. Son inverse A^{-1} s'écrit

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

A présent on veut résoudre $Ax = b$

avec $b = \begin{pmatrix} 32 \\ 23 \\ 33 \\ 31 \end{pmatrix}$ on trouve $x = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$

Si on résoud $A(x+\delta x) = b+\delta b$ avec

$$\delta b = \begin{pmatrix} 10^{-1} \\ -10^{-1} \\ 10^{-1} \\ -10^{-1} \end{pmatrix} \quad \text{on trouve} \quad \delta x = \begin{pmatrix} 9,2 \\ -12,6 \\ 4,5 \\ 4,1 \end{pmatrix}$$

Ce qui représente une variation relative énorme de δx en comparaison de la variation δb . Nous allons expliciter la caractéristique d'une matrice A qui fait que de tels phénomènes se produisent. Dans de telles situations, on dit que la matrice est mal conditionnée ; et nous allons définir un nombre qui "mesure" le conditionnement d'une matrice.

1°) Conditionnement des matrices.

On veut résoudre le système $Ax = b$. Si on change les données en ajoutant δb à b on veut savoir comment sera perturbée la solution x . Soit δx la perturbation pour x :

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

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

$$\delta x = A^{-1}\,\delta b$$

$$\Longrightarrow \|\delta x\| \leqslant \|A^{-1}\|\,\|\delta b\|$$

comme $Ax = b$ $\qquad\qquad \|b\| \leqslant \|A\|\,\|x\|$.

Alors $\dfrac{1}{\|x\|} \leqslant \dfrac{\|A\|}{\|b\|}$.

On a donc $\dfrac{\|\delta x\|}{\|x\|} \leqslant \|A\|\,\|A^{-1}\| \cdot \dfrac{\|\delta b\|}{\|b\|}$.

Notons $e_r(b) = \dfrac{\|\delta b\|}{\|b\|}$ l'erreur relative sur b et $e_r(x) = \dfrac{\|\delta x\|}{\|x\|}$ l'erreur relative sur x . On a $e_r(x) \leqslant \|A\| \cdot \|A^{-1}\|\, e_r(b)$.

Définition : On appelle conditionnement de A le nombre $\|A\|\,\|A^{-1}\|$. On écrit $\mathrm{cond}(A) = \|A\|\,\|A^{-1}\|$ on l'appelle aussi le nombre de condition. Une matrice est d'autant mieux conditionnée que son nombre de condition est petit : en effet de petites variations relatives de b entraînent alors une petite erreur

relative sur x .

Propriétés. a) $\text{cond}(A)$ dépend de la norme choisie (sur R^n).

b) Quelle que soit la norme choisie, $\text{cond}(A)$ est supérieur ou égal à 1 ; en effet : $I = A.A^{-1}$ et quelle que soit la norme choisir (sur R^n) $\|I\| = 1$ et alors : $1 \leqslant \|A\| \, \|A^{-1}\| = \text{cond}(A)$.

c) $\text{cond}(A) = \text{cond}(A^{-1})$.

d) $\text{cond}(\lambda A) = \text{cond } A$ pour λ non nul.

Cas particulier où la norme choisie est la norme euclidienne.

$$\|X\| = \left(\sum_{i=1}^{n} X_i^2\right)^{\frac{1}{2}} \quad \text{et} \quad \|A\| = \sup_{\substack{x \in R^n \\ x \neq 0}} \frac{\|Ax\|}{\|x\|} .$$

Proposition : Soit A une matrice carrée d'ordre n , régulière. Alors :

1°) Si μ_1^2 , $\mu_2^2, \ldots, \mu_n^2$ sont les valeurs propres de $A^t A$ avec $0 < \mu_1^2 \leqslant \mu_2^2 \leqslant \ldots \leqslant \mu_n^2$. On a : $\text{cond}(A) = \dfrac{\mu_n}{\mu_1}$.

2°) Si A est symétrique, alors $\text{cond}(A) = \dfrac{|\lambda_M|}{|\lambda_m|}$

λ_M = valeur propre de module maximal de A

λ_m = valeur propre de module mirimal de A .

3°) Si A est orthogonale alors $\text{cond}(A) = 1$.

On démontre d'abord le lemme suivant :

Lemme : Soit A une matrice réelle d'ordre n et soit $B = A^t A$. Alors B est symétrique, semi-définie positive. Ses valeurs propres sont $\mu_1^2, \ldots, \mu_n^2$ telles que $0 \leqslant \mu_1^2 \leqslant \ldots \leqslant \mu_n^2$ et dire que A est inversible équivaut à dire que μ_1^2 est non nul et dans ce cas on a : $\|A^{-1}\| = \dfrac{1}{\mu_1}$.

Démonstration du lemme.

$$\begin{cases} B = B^t \\ (BX,X) = (AX,AX) = \|AX\|^2 \geqslant 0 \qquad \forall X \end{cases}$$

B est donc symétrique et semi-définie positive. Les valeurs propres de B sont donc réelles positives $0 \leqslant \mu_1^2 \leqslant \ldots \leqslant \mu_n^2$ et il existe une base orthogonale $\xi_1, \xi_2, \ldots, \xi_n$ dans laquelle B est diagonale :

$B(\xi_i) = \mu_i^2 \, \xi_i$.

Si $x = \sum_{i=1}^{n} x_i \, \xi_i$ on a $Bx = \sum_{i=1}^{n} x_i \, \mu_i^2 \, \xi_i$ et $(Bx,x) = \sum_{i=1}^{n} x_i^2 \, \mu_i^2$ en sorte que $(Bx,x) \leqslant \mu_n^2 \|x\|^2$.

Pour $x = \xi_n$, $(B\,\xi_n, \xi_n) = \mu_n^2$ et alors $\sup_{\substack{x \in \mathbb{R}^n \\ x \neq 0}} \frac{(Bx,x)}{\|x\|^2} = \mu_n^2$

alors $\|A\|^2 = \sup_{\substack{x \in \mathbb{R}^n \\ x \neq 0}} \frac{\|Ax\|^2}{\|x\|^2} = \sup_{\substack{x \in \mathbb{R}^n \\ x \neq 0}} \frac{(Bx,x)}{\|x\|^2} = \mu_n^2$.

$B = A^t A$ alors $\det B = (\det A)^2$.

Donc si A est inversible, B l'est aussi et réciproquement. Or B inversible équivaut à μ_1^2 non nul.

On a vu que $\|A\|^2 = \mu_n^2$ donc $\|A\| = [\rho(A^t A)]^{\frac{1}{2}}$.

Si A est inversible, on a de même $\|A^{-1}\| = [\rho((A^{-1})^t A^{-1})]^{\frac{1}{2}} = [\rho((A^t)^{-1} A^{-1})]^{\frac{1}{2}}$

$= [\rho((AA^t)^{-1})]^{\frac{1}{2}} = [\rho(B^{-1})]^{\frac{1}{2}}$.

μ_1 est non nul et les valeurs propres de B^{-1} sont : $(\mu_1^2)^{-1}, \ldots, (\mu_n^2)^{-1}$

telles que $0 < \frac{1}{\mu_n^2} \leqslant \frac{1}{\mu_{n-1}^2} \leqslant \ldots \leqslant \frac{1}{\mu_1^2}$ et $\|A^{-1}\| = \frac{1}{\mu_1}$.

Démonstration de la proposition. 1°) est maintenant évident :

$$\text{cond}(A) = \|A\| \cdot \|A^{-1}\| = \mu_n \cdot \frac{1}{\mu_1}$$

2°) si A est symétrique

$$B = A^t A = A^2 .$$

Donc si λ_i est valeur propre de A, λ_i^2 est valeur propre de B et alors

$\mu_n^2 = |\lambda_M|^2$ où λ_M est la valeur propre de plus grand module de A.

$\mu_1^2 = |\lambda_m|^2$ où λ_m est la valeur propre de plus petit module de A.

Et $\text{cond}(A) = \dfrac{|\lambda_M|}{|\lambda_m|}$.

Une matrice symétrique, même définie positive peut-être mal conditionnée (cf. exemple du début du chapitre). Il suffit que $\dfrac{|\lambda_M|}{|\lambda_m|}$ soit grand.

3°) si A est orthogonale

$$A^{-1} = A^t$$

$$B = A^t A = I .$$

Et alors $\text{cond}(A) = \|A^{-1}\| \, \|A\| = [\rho(B)]^{\frac{1}{2}} = 1$.

Les matrices orthogonales sont bien conditionnées (car $A^t = A^{-1}$).

Proposition. Soit A une matrice régulière. Supposons que $Ax = b$ et $A(x + \delta x) = b + \delta b$ alors $\text{cond}(A)$ est le plus petit nombre M tel que :

$$\frac{\|\delta x\|}{\|x\|} \leqslant M \frac{\|\delta b\|}{\|b\|} , \quad \forall b \neq 0 .$$

On va d'abord démontrer le lemme suivant :

Lemme. Soit A une matrice réelle. Alors il existe deux matrices orthogonales U et V telles que $V^t AU = D$ où

$D = \begin{pmatrix} \mu_1 & & 0 \\ & \mu_2 & \\ & & \ddots \\ 0 & & \mu_n \end{pmatrix}$ avec $\mu_1^2, \ldots, \mu_n^2$ les valeurs propres ordonnées de $B = A\,{}^tA$: $0 \leqslant \mu_1^2 \leqslant \mu_2^2 \leqslant \cdots \leqslant \mu_n^2$

Démonstration : Comme B est symétrique il existe une matrice U orthogonale telle que : $U^t BU = D^2$.

Donc $U^t(A\,{}^tA)U = D^2$.

Si on pose $F = AU$ on a donc : $F\,{}^tF = D^2$.

Si $F=(f_{i,j})_{\substack{1\leqslant i\leqslant n \\ 1\leqslant j\leqslant n}}$ cela s'écrit aussi $\sum_{k=1}^{n} f_{ki}\, f_{kj} = \mu_i^2\, \delta_{ij}$ pour tout i et tout j .

Si on désigne par $\vec{f}_i$ la i-ème colonne de F , on peut encore écrire :

$(\vec{f}_i , \vec{f}_j) = \mu_i^2\, \delta_{ij} \quad 1 \leqslant i , j \leqslant n$.

On considère deux cas :

. A est régulière.

Alors μ_i est non nul.

On pose alors $\vec{v}_i = \dfrac{\vec{f}_i}{\mu_i}$

$$(\vec{v}_i , \vec{v}_j) = (\vec{f}_i , \vec{f}_j)\frac{1}{\mu_i\, \mu_j} = \delta_{ij} .$$

On désigne par V la matrice dont les vecteurs colonnes sont les $\vec{v}_i$; alors V est orthogonale et on a : $F = VD$.

Comme $F = AU$ on trouve $AU = VD$ ce qui entraîne $V\,{}^tAU = V\,{}^tVD = D$.

. A est singulière.

Alors B est aussi singulière ; B a $r-1$ valeurs propres nulles qui sont : $\mu_1^2, \ldots, \mu_{r-1}^2$. Les autres valeurs propres étant $\mu_r^2, \ldots, \mu_n^2$:

$$0 = \mu_1^2 = \mu_2^2 = \ldots = \mu_{r-1}^2 < \mu_r^2 \leqslant \ldots \leqslant \mu_n^2 .$$

Pour $r \leqslant i \leqslant n$ on pose $\vec{v}_i = \dfrac{\vec{f}_i}{\mu_i}$.

Alors $\vec{v}_r$, $\vec{v}_{r+1}, \ldots, \vec{v}_n$ forment une famille orthonormale que l'on complète.

Il existe des vecteurs $\vec{v}_1, \vec{v}_2, \ldots, \vec{v}_{r-1}$ tels que $\vec{v}_1, \vec{v}_2, \ldots, \vec{v}_{r-1}$ forment une base orthonormale de R^n . On désigne alors par V la matrice dont les vecteurs colonnes sont les v_i . On a encore $F = VD$ et : $V^t AU = D$. ∎

<u>Démonstration de la proposition</u> :

On sait que $\dfrac{\|\delta x\|}{\|x\|} \leqslant \text{cond}(A) \dfrac{\|\delta b\|}{\|b\|}$.

On va démontrer qu'il n'existe pas de nombre M tel que :

$$\begin{cases} M < \text{cond}(A) \\ \dfrac{\|\delta x\|}{\|x\|} \leqslant M \dfrac{\|\delta b\|}{\|b\|} \quad \forall b \neq 0 . \end{cases}$$

On va montrer qu'il existe b et δb tels que $\dfrac{\|\delta x\|}{\|x\|} = \text{cond}(A) \dfrac{\|\delta b\|}{\|b\|}$ ce qui démontrera la proposition.

D'après le lemme précédent, il existe deux matrices orthogonales U et V telles que $V^t AU = D$.

Posons $b = \mu_n Ve_n$; alors la solution de $Ax = b$ est $x = Ue_n$; en effet : $A U e_n = V D e_n = V(\mu_n e_n) = b$. De même si on pose $\delta b = \mu_1 Ve_1$.

Alors la solution de $A \delta x = \delta b$ est $\delta x = Ue_1$:

$$A Ue_1 = V De_1 = V(\mu_1 e_1) = \delta b .$$

On a alors les relations suivantes : $\|\delta x\| = \|Ue_1\| = \|e_1\|$ car U est orthogonale.

Donc $\|\delta x\| = 1$.

De même $\|x\| = \|Ue_n\| = \|e_n\| = 1$.

$\|\delta b\| = \|\mu_1 \, Ve_1\| = \mu_1 \, \|Ve_1\| = \mu_1 \, \|e_1\|$ car V est orthogonale.

Donc $\|\delta b\| = \mu_1$.

De même $\|b\| = \|\mu_n \, Ve_n\| = \mu_n \, \|e_n\| = \mu_n$. On obtient donc :

$$\frac{\|\delta x\|}{\|x\|} = 1$$

$$\frac{\|\delta b\|}{\|b\|} = \frac{\mu_1}{\mu_n} .$$

Or d'après la proposition $\text{cond}(A) = \dfrac{\mu_n}{\mu_1}$.

On a donc $\dfrac{\|\delta x\|}{\|x\|} = \text{cond}(A) \dfrac{\|\delta b\|}{\|b\|}$.

Stabilité de la solution par rapport aux variations de A .

Proposition : Soit A une matrice $n \times n$ réelle et soit b un vecteur de $\mathbb{R}^n$.

Si $Ax = b$ et $(A+\delta A)(x+\delta x) = b$ alors :

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

et pour toute A il existe δA et il existe b tels qu'on ait l'égalité.

Démonstration :

$$(A+\delta A)(x+\delta x) = Ax + A.\delta x + \delta A.x + \delta A.\delta x = b = Ax$$

d'où $$A.\delta x + \delta A(x+\delta x) = 0$$

et $$\delta x = -A^{-1}.\delta A(x+\delta x).$$

On obtient en passant aux normes :

$$\|\delta x\| \leqslant \|A^{-1}\| . \|\delta A\| . \|x+\delta x\|$$

d'où : $\dfrac{\|\delta x\|}{\|x+\delta x\|} \leqslant \|A^{-1}\| . \|\delta A\| = \|A^{-1}\| . \|A\| . \dfrac{\|\delta A\|}{\|A\|} = \text{cond}(A) . \dfrac{\|\delta A\|}{\|A\|}$.

Cas de l'égalité :

par définition $\|A^{-1}\| = \sup_{Y\neq 0} \frac{\|A^{-1}Y\|}{\|Y\|} = \sup_{\|Y\|\leqslant 1} \frac{\|A^{-1}Y\|}{\|Y\|}$

or la boule unité de $\mathbb{R}^n$ est compacte, donc il existe Y tel que

$\|A^{-1}\| = \frac{\|A^{-1}Y\|}{\|Y\|}$

et $\delta A = \beta I$

alors $A+\delta A = A+\beta I$ donc $(A+\beta I)(x+\delta x) = b$

d'où $x+\delta x - y$.

Or $A.\delta x = -\delta A(x+\delta x)$

$= -\beta I(x+\delta x)$

$= -\beta y$

d'où $\delta x = -\beta A^{-1} y$. En passant aux normes :

$$\|\delta x\| = \beta \|A^{-1}y\| = \|A^{-1}\| . \|y\| = \|\delta A\| . \|A^{-1}\| . \|x+\delta x\|$$

et $\frac{\|\delta x\|}{\|x+\delta x\|} = \text{cond}(A) . \frac{\|\delta A\|}{\|A\|}$.

Cette égalité est vraie même si la norme n'est pas la norme euclidienne.

Lemme : Soit B une matrice carrée réelle. Si $\|B\| < 1$, alors $I-B$ est inversible et : $(I-B)^{-1} = I + B + B^2 + \ldots$ et $\|(I-B)^{-1}\| \leqslant \frac{1}{1-\|B\|}$.

Démonstration : Posons $S_n = I + B + \ldots + B^n$

$$(I-B)S_n = I-B^{n+1} .$$

Or si $\|B\| < 1$, la série $\sum_j B^j$ converge absolument car $\|B^j\| \leqslant \|B\|^j$.

Donc S_n converge vers S et $(I-B)S = I$.

D'autre part :

$$\|(I-B)^{-1}\| \leqslant \sum_j \|B^j\| \leqslant \sum_j \|B\|^j = \frac{1}{1-\|B\|} .$$

Proposition : Soit A une matrice $n \times n$ régulière. Si $\|\delta A\| < \frac{1}{\|A^{-1}\|}$ et si $Ax = b = (A+\delta A)(x+\delta x)$ alors : $\frac{\|\delta x\|}{\|x\|} \leqslant \text{cond}(A) \frac{\|\delta A\|}{\|A\|} \frac{1}{1-\|\delta A\| \cdot \|A^{-1}\|}$

Démonstration : $(A+\delta A)(x+\delta x) = Ax$, d'où : $(A+\delta A)\delta x = -\delta A.x$ en multipliant les deux termes par A^{-1} à gauche : $(I + A^{-1}.\delta A)\delta x = -A^{-1}.\delta A.x$

$$\delta x = -(I + A^{-1}.\delta A)^{-1}.A^{-1}.\delta A.x$$

d'où en passant aux normes :

$$\|\delta x\| \leqslant \|(I + A^{-1}.\delta A)^{-1}\| . \|A^{-1}\| . \|\delta A\| . \|x\|$$

d'après le lemme précédent on a :

$$\|(I + A^{-1}.\delta A)^{-1}\| \leqslant \frac{1}{1-\|A^{-1}\| . \|\delta A\|}$$

car $\|A^{-1}.\delta A\| \leqslant \|A^{-1}\| . \|\delta A\| < 1$.

On obtient donc :

$$\frac{\|\delta x\|}{\|x\|} \leqslant \|A^{-1}\| . \|\delta A\| \frac{1}{1-\|A^{-1}\| . \|\delta A\|} = \text{cond}(A) \frac{\|\delta A\|}{\|A\|} \frac{1}{1-\|A^{-1}\| \|\delta A\|}$$

Si A et δA sont régulières on pose :

$$(A+\delta A)^{-1} = A^{-1} + \delta(A^{-1}).$$

On a alors la proposition :

Proposition : Soit A une matrice $n \times n$ régulière : si $\|\delta A\| < \frac{1}{\|A^{-1}\|}$ alors $A+\delta A$ est inversible et $\frac{\|\delta(A^{-1})\|}{\|A^{-1}\|} \leqslant \text{cond}(A) . \frac{\|\delta A\|}{\|A\|} . \frac{1}{1-\|A^{-1}\| . \|\delta A\|}$

Démonstration : $(A+\delta A)(A^{-1} + \delta(A^{-1})) = I - A.A^{-1}$ d'où

$$\delta A.A^{-1} + A.\delta(A^{-1}) + \delta A.\delta(A^{-1}) = 0$$

$(A+\delta A)\delta(A^{-1}) = -\delta A.A^{-1}$ d'où en multipliant les deux membres par A^{-1} à gauche :

$$(I + A^{-1}.\delta A)\delta(A^{-1}) = -A^{-1}\ \delta A\ A^{-1}$$

donc $\delta(A^{-1}) = -(I + A^{-1}\ \delta A)^{-1}\ A^{-1}\ \delta A\ A^{-1}$.

D'après le lemme $(I + A^{-1}\,\delta A)$ est inversible et de plus :

$$\frac{\|\delta(A^{-1})\|}{\|A^{-1}\|} \leqslant \frac{1}{1-\|A^{-1}\|.\|\delta A\|}.\|\delta A\|.\|A^{-1}\| = \mathrm{cond}(A).\frac{\|\delta A\|}{\|A\|}\frac{1}{1-\|A^{-1}\|.\|\delta A\|}$$

Utilisation des méthodes de résolution des systèmes linéaires.

- GAUSS : utilisée pour les matrices pleines on peut aller jusqu'à des ordres de 50×50 à 300×300 selon les machines.

- GIVENS (rotations) : N'est pas recommandée en pratique : très semblable à la méthode de Householder elle nécessite beaucoup plus d'opérations (2 fois plus de multiplications).

- HOUSEHOLDER : pour des matrices pleines. La méthode est plus stable que Gauss et ne demande pas d'échanges de lignes ; elle sert également pour des algorithmes de calcul de valeurs propres (algorithme Q.R. cf. chap.II).

- CHOLESKI : pour des matrices symétriques plus ou moins pleines. L'ordre peut aller jusqu'à 500×500 et parfois plus.

Les méthodes itératives suivantes sont utilisées pour des matrices creuses provenant de la discrétisation d'équations aux dérivées partielles.

- JACOBI : peu utilisée en pratique ; comparable à la méthode de Gauss-Seidel et moins avantageuse sur plusieurs points.

- GAUSS-SEIDEL : utilisée pour des programmes rapides. Assez bonne précision.

- RELAXATION : ne s'utilise pratiquement qu'avec paramètre optimal (ou voisin de l'optimal) ; elle donne alors des convergences très rapides.

Bibliographie du chapitre I

N. GASTINEL.- Analyse Numérique linéaire. Hermann, Paris (1966).

R.S. VARGA.- Matrix Iterative analysis. Prentice-Hall, New York (1962).

J.H. WILKINSON.- The Algebraïc eigenvalue problem. Oxford University Press, Oxford (1964).

D.K. FADDEEV, V.N. FADDEEV.- Computational Methods of Linear Algebra. W.M. Freeman and Company, San Francisco (1963).

HOUSEHOLDER.- The theory of matrices in Numerical Analysis. Blandell, New-York (1964).

G. FORSYTHE, C.B. MOLER.- Computer Solution of Linear Algebraïc Systems. Prentice-Hall, Inc, Englewood Cliffs, N.J. (1967).

CHAPITRE II

CALCUL DES VALEURS PROPRES ET DES VECTEURS PROPRES D'UNE MATRICE.

Introduction.

La recherche des valeurs propres et vecteurs propres d'une matrice se rencontre dans des problèmes assez divers. En voici deux exemples :

- Lorsque l'on veut résoudre un système linéaire par une méthode de relaxation, il faut connaître la valeur optimale du paramètre ω , et pour cela calculer le rayon spectral de la matrice de Jacobi.

- Lorsque l'on étudie l'équation des ondes $\frac{\partial^2 u}{\partial t^2} - \Delta u = 0$ dans un système physique donné, on est amené à rechercher les vibrations propres de ce système, c'est-à-dire à chercher les nombres ω (pulsations propres du système) et les fonctions $\psi(x)$ associées, tels que les fonctions $u(x,t) = \sin \omega t \, \psi(x)$ soient solutions.

ψ et ω devront vérifier : $\Delta\psi + \omega^2\psi = 0$.

On discrétise alors le problème : on approche la fonction ψ par la suite de ses valeurs sur un ensemble fini de points suffisamment rapprochés et l'on remplace les dérivées partielles par des différences finies.

On considérera que la suite des valeurs qui approchent ψ est un vecteur d'un espace de dimension finie, qui doit-être vecteur propre, associé à la valeur propre $-\omega^2$, d'une certaine matrice remplaçant l'opérateur Δ (les dimensions de cette matrice seront très grandes).

Les complications qui se présentent dans la détermination numérique des éléments propres d'une matrice sont bien plus grandes que pour la résolution des systèmes linéaires. Tout d'abord il n'existe pas ici de méthode "exacte" : toutes les méthodes sont itératives : ceci est compréhensible, puisqu'il s'agit en fait de trouver les zéros d'un certain polynôme, à savoir le polynôme caractéristique de la matrice. De plus on trouvera un certain nombre de cas particuliers où les méthodes générales se trouveront en défaut (valeurs propres multiples, de même module, etc...) (*).

Soit A la matrice réelle, de type (n,n), dont on cherche les valeurs propres et vecteurs propres. On supposera en général A non singulière, puisque, autrement 0 est une valeur propre de A, connue à priori.

1°) Détermination du polynôme caractéristique d'une matrice A.

Pour trouver les éléments propres de la matrice A, la méthode la plus naturelle consiste à déterminer le polynôme caractéristique, à rechercher ses zéros $\lambda_1,\ldots,\lambda_n$, puis à résoudre les systèmes linéaires : $A.x = \lambda_i x$ $(i = 1,\ldots,n)$. Malheureusement, ce schéma n'est pas tres satisfaisant pour les applications numériques.

En effet par les meilleures méthodes, la détermination du polynôme caractéristique nécessite un très grand nombre d'opérations, ce qui signifie une accumulation des erreurs d'arrondi, et une précision faible pour les coefficients du polynôme.

(*) En raison des grandes difficultés inhérentes au problème, ce chapitre ne doit être considéré que comme une introduction au problème les valeurs propres.

En outre, les valeurs propres sont les zéros de ce polynôme ; mais la résolution numérique d'une équation algébrique n'est pas en général très facile, et les résultats sont extrêmement sensibles aux variations des coefficients.

Toutefois, comme la connaissance du polynôme caractéristique peut-être utile en elle-même, indépendamment de la recherche des valeurs propres, nous allons donner une des méthodes numériques connues pour la recherche de ce polynôme :

$$P_A(X) = a_o X^n + a_1 X^{n-1} + \ldots + a_n \qquad (a_o \neq 0).$$

En pratique, cette méthode ne sera utilisable que pour de petites valeurs de n ($n \leqslant 10$).

Remarque : $P_A(X)$ est égal au déterminant de $(A - XI)$ (à un facteur près) ; mais il est exclus de former $P_A(X)$ en développant ligne par ligne, ou colonne par colonne, ce déterminant : le nombre d'opérations serait alors vraiment énorme.

Méthode de Leverrier.

Soient $x_1, x_2, \ldots, x_n$ les valeurs propres, c'est-à-dire les zéros du polynôme P_A. Introduisons les fonctions symétriques élémentaires des racines x_i, c'est-à-dire les n nombres :

$$\sigma_1 = x_1 + x_2 + \ldots + x_n$$

$$\sigma_2 = \sum_{i>j} x_i x_j$$

$$\sigma_3 = \sum_{i>j>k} x_i x_j x_k$$

$$\cdots\cdots\cdots\cdots\cdots\cdots$$

$$\sigma_n = x_1 \times x_2 \times \ldots \times x_n$$

Il est bien connu que les σ_i sont liés aux coefficients a_i du polynôme P_A par les relations :

$$\sigma_i = (-1)^i \frac{a_i}{a_o} .$$

On sait aussi que toute fonction symétrique des racines peut s'exprimer comme une fonction des σ_i , donc comme une fonction des coefficients a_i .

En particulier, considérons les sommes de Newton des x_i :

$$S_1 = x_1 + x_2 + \ldots + x_n$$

$$S_2 = x_1^2 + x_2^2 + \ldots + x_n^2$$

$$S_k = x_1^k + x_2^k + \ldots + x_n^k \quad (k \in \mathbb{N})$$

les sommes S_k sont reliées aux coefficients a_i par les formules dites de Newton :

$$a_1 = -a_o S_1$$

$$a_1 S_1 + 2a_2 = -a_o S_2$$

$$a_1 S_2 + a_2 S_1 + 3a_3 = -a_o S_3$$

$$a_1 S_{k-1} + a_2 S_{k-2} + \ldots + k a_k = -a_o S_k$$

$$a_1 S_{n-1} + a_2 S_{n-2} + a_3 S_{n-3} + \ldots + n a_n = -a_o S_n$$

Connaissant les sommes S_k , il suffit de résoudre un système linéaire triangulaire pour calculer les coefficients $\frac{a_i}{a_o}$. Or les S_k sont facilement accessibles dès que l'on connaît les puissances A^k de la matrice A .

En effet : $S_1 = x_1 + \ldots + x_n$ est égale à la trace de A (somme des éléments diagonaux de A).

S_k n'est autre que la trace de A^k : car si $x_1,\ldots,x_n$ sont les valeurs propres d'une matrice A, $x_1^k,\ldots,x_n^k$ sont les valeurs propres de A^k (pour s'en convaincre il suffit de se rappeler que A est semblable à sa réduite de Jordan: $A = S.J.S^{-1}$; J est triangulaire et a pour éléments diagonaux les valeurs propres de A et $\forall k$: $A^k = S\,J^k\,S^{-1}$).

Nous avons donc une méthode qui fournit "exactement" les coefficients du polynôme caractéristique. Mais le nombre d'opérations nécessaires interdit pratiquement son utilisation pour les grandes valeurs de n.

En effet il faut former les puissances successives de A : calculer A^2 revient à calculer n^2 produits scalaires de vecteurs le R^n, soit à effectuer n^3 multiplications. Pour former $A^k = A.A^{k-1}$ en connaissant A^{k-1}, il faut également n^3 multiplications. Le coût total de la méthode est donc de l'ordre de n^4 multiplications.

Méthode de Leverrier améliorée.

Cette méthode donne directement les coefficients du polynôme caractéristique, sans passer par la résolution d'un système linéaire (mais il faut quand même calculer $(n-1)$ produits de matrices (n,n)).

En effet cette méthode consiste à calculer les $2n$ matrices A_i, B_i $(i=1,\ldots,n)$ et les n nombres p_i définis par :

$$A_1 = A \quad , \; p_1 = \mathrm{trace}(A_1) \; , \; B_1 = A_1 - p_1.I$$

$$A_2 = B_1A \quad , \; p_2 = \frac{1}{2}\,\mathrm{tr}(A_2) \; , \; B_2 = A_2 - p_2.I$$

$$A_n = B_{n-1}A \; , \; p_n = \frac{1}{n}\,\mathrm{tr}(A_n) \; , \; B_n = A_n - p_nI$$

Proposition : Le polynôme caractéristique de A est (à un facteur près) :

$$P_A(\lambda) = \lambda^n - p_1 \lambda^{n-1} \ldots - \lambda p_{n-1} - p_n$$

et nous avons : $B_n = 0$.

Démonstration : Soit $P_A(\lambda)$ le polynôme caractéristique de A :

$$P_A(\lambda) = \det(\lambda I - A) = \lambda^n - b_1 \lambda^{n-1} - \ldots - b_n .$$

Par récurrence sur k , nous allons montrer que b_k est égal à p_k pour tout k $(1 \leqslant k \leqslant n)$.

. pour $k=1$: on sait que k_1 est égal à la somme des valeurs propres, donc à la trace de A :

$$b_1 = tr(A) = tr(A_1) = p_1 .$$

. supposons que $p_i = b_i$ soit vrai pour : $1 \leqslant i \leqslant k-1$.

Montrons qu'alors $p_k = b_k$.

En effet : $A_k = B_{k-1}A = A_{k-1}A - p_{k-1}A = B_{k-2}A^2 - p_{k-1}A = A_{k-2}A^2 - p_{k-2}A^2 - p_{k-1}A$.

En raisonnant par récurrence sur h on vérifie facilement que :

$$A_k = A_{k-h}A^h - p_{k-h}A^h - p_{k-h+1}A^{h-1} - \ldots - p_{k-1}A \qquad (0 \leqslant h \leqslant k-1)$$

d'où, pour $h = k-1$:

$$A_k = A^k - p_1 A^{k-1} - p_2 A^{k-2} - \ldots - p_{k-1}A .$$

Mais alors, par définition de p_k :

$$p_k = \frac{1}{k} \operatorname{trace}(A_k) = \frac{1}{k}[S_k - p_1 S_{k-1} - p_2 S_{k-2} \ldots - p_{k-1}S_1]$$

(où les S_i sont les sommes de Newton des racines de P_A).

Utilisons l'hypothèse de récurrence : $p_i = b_i$ $(1 \leqslant i \leqslant k-1)$; il vient : $p_k = \frac{1}{k}[S_k - b_1 S_{k-1} - b_2 S_{k-2} - \ldots - b_{k-1}S_1]$. Mais le deuxième membre de cette égalité est égal à b_k d'après la $k^{\text{ième}}$ formule de Newton ; donc $p_k = b_k$.

La récurrence montre donc que les p_k sont les coefficients (au signe près) du polynôme caractéristique.

Montrons maintenant que B_n est la matrice nulle :

$$B_n = A_n - p_n I$$

$$B_n = A^n - p_1 A^{n-1} - p_2 A^{n-2} - \ldots - p_n I$$

$$B_n = p_A(A).$$

On a donc $B_n = 0$, grâce au théorème de Cayley-Hamilton.

Remarque : par cette méthode, on obtient indirectement l'inverse de A , lorsque A est inversible ; en effet : $B_n = 0$ signifie $A_n = p_n I$, c'est-à-dire $B_{n-1}A = p_n I$.

Si A est régulière : $p_n = \det(A) \neq 0$, et $A^{-1} = \frac{B_{n-1}}{p_n}$.

2°) Calcul de la valeur propre de A de module maximum.

Soient $\lambda_1, \ldots, \lambda_n$ les valeurs propres de la matrice A : on peut supposer que leurs modules sont rangés dans l'ordre décroissant : $|\lambda_1| \geqslant |\lambda_2| \geqslant \ldots \geqslant |\lambda_n|$.

Il s'agit de calculer λ_1 , et éventuellement les autres valeurs propres de module égal à $|\lambda_1|$, ainsi que les vecteurs propres associés.

Nous supposerons que λ_1 , ainsi que les valeurs propres de même modules, sont simples.

1er cas : $|\lambda_1| > |\lambda_2|$

Supposons qu'il existe dans $\mathbb{R}^n$ une base de vecteurs propres de la matrice A : $X_1, X_2, \ldots, X_n$, associés respectivement à $\lambda_1, \lambda_2, \ldots, \lambda_n$.

Pour tout vecteur $Y^{(o)} \in \mathbb{R}^n$ $(Y^{(o)} \neq 0)$, soient $a_1, \ldots, a_n$ les coordonnées de $Y^{(o)}$ sur $X_1, \ldots, X_n$: $Y^{(o)} = a_1 X_1 + \ldots + a_n X_n$; formons la suite des vec-

teurs $Y^{(k)}$:

$$Y^{(k)} = AY^{(k-1)} = A^k Y^{(o)}.$$

Puisque $X_1, X_2, \ldots, X_n$ sont vecteurs propres :

$$Y^{(k)} = a_1 \lambda_1^k X_1 + a_2 \lambda_2^k X_2 + \ldots + a_n \lambda_n^k X_n .$$

Si a_1 est différent de 0 :

$$Y^{(k)} = a_1 \lambda_1^k \left[X_1 + \frac{a_2}{a_1}\left(\frac{\lambda_2}{\lambda_1}\right)^k X_2 + \ldots + \frac{a_n}{a_1}\left(\frac{\lambda_n}{\lambda_1}\right)^k X_n \right]$$

Or, par hypothèse : $\left|\frac{\lambda_n}{\lambda_1}\right| \leqslant \left|\frac{\lambda_{n-1}}{\lambda_1}\right| \leqslant \ldots \leqslant \left|\frac{\lambda_2}{\lambda_1}\right| < 1$; donc :

$$\left\| \frac{a_2}{a_1}\left(\frac{\lambda_2}{\lambda_1}\right)^k X_2 + \ldots + \frac{a_n}{a_1}\left(\frac{\lambda_n}{\lambda_1}\right)^k X_n \right\| \leqslant \left|\frac{\lambda_2}{\lambda_1}\right|^k \times \left(\left\| \frac{a_2}{a_1} X_2 \right\| + \ldots + \left\| \frac{a_n}{a_1} X_n \right\| \right) .$$

Autrement dit, si $k \to +\infty$:

$$Y^{(k)} = a_1 \lambda_1^k \left[X_1 + 0\left(\left(\frac{\lambda_2}{\lambda_1}\right)^k\right) \right] ,$$

où $0\left(\left(\frac{\lambda_2}{\lambda_1}\right)^k\right)$ désigne un vecteur tendant vers 0, comme $\left(\frac{\lambda_2}{\lambda_1}\right)^k$.

Soit $(e_j)_{j=1,\ldots,n}$ la base canonique dans $\mathbb{R}^n$; il existe un indice j tel que la composante $X_1.e_j$ de X_1 sur e_j soit non nulle (puisque $X_1 \neq 0 \ldots$).

Soit $Y_j^{(k)} = Y^{(k)}.e_j$ la $j^{\text{ième}}$ composante sur la base canonique du vecteur $Y^{(k)}$ ($k \in \mathbb{N}$) :

$$\frac{Y_j^{(k+1)}}{Y_j^{(k)}} = \lambda_1 \times \frac{X_1.e_j + 0(1)}{X_1.e_j + 0(1)} \to \lambda_1 \quad \text{si} \quad k \to +\infty.$$

Donc : si a_1 est différent de 0, alors $\frac{Y_j^{(k+1)}}{Y_j^{(k)}} \to \lambda_1$ quand $k \to +\infty$, pour tout j tel que $X_1.e_j \neq 0$.

La direction du vecteur propre X_1 est également obtenue par passage à la limite :

$$\frac{Y^{(2k)}}{\|Y^{(2k)}\|} \to \varepsilon \frac{X_1}{\|X_1\|} \quad \text{quand} \quad k \to +\infty \quad \left(\varepsilon = \frac{a_1}{|a_1|} , \ a_1 \neq 0\right) \quad \blacksquare$$

Remarques :

1) Afin d'éviter de manipuler des vecteurs $Y^{(k)}$ trop grands, on utilise en pratique les vecteurs normalisés, en posant (pour tout k) :

$$\mu_k = \frac{1}{\|Y^{(k)}\|} \quad , \quad \tilde{Y}(k) = \mu_k Y^{(k)} \quad , \quad X^{(k+1)} = A\tilde{Y}(k) \quad ,$$

c'est-à-dire : $X^{(k+1)} = \mu_k AY^{(k)} = \mu_k Y^{(k+1)}$. Alors (toujours dans les mêmes hypothèses sur les valeurs propres de A et sur $Y^{(0)}$) : $\frac{X_j^{(k+1)}}{\tilde{Y}_j^{(k)}} \to \lambda_1$ pour tout j tel que $X_1 . e_j \neq 0$.

2) Si $a_1 = 0$: le terme prépondérant de $Y^{(k)}$ est alors (en général) $a_2 \lambda_2^k X_2$, et le résultat de convergence est faux. Si a_1 , sans être nul, est petit, il se peut qu'il faille un grand nombre d'itérations avant que le terme : $a_2 \lambda_2^k X_2$ ne devienne négligeable devant $a_1 \lambda_1^k X_1$; dans ce cas, il est préférable de recommencer l'itération avec un autre vecteur initial $Y^{(o)}$ (idem pour $a_1 = 0$ exactement).

3) En fait l'hypothèse que nous avons faite sur l'existence d'une base de vecteurs propres n'était pas nécessaire, bien qu'elle ait simplifié le raisonnement. Soit J la réduite de Jordan de la matrice A : on sait qu'il existe une base $X_1, X_2, \ldots, X_n$ dans laquelle la transmuée de A est égale à J . (quand il existe une base de vecteurs propres, J est diagonale). Pour tout vecteur $Y^{(o)} \in \mathbb{R}^n$, posons $Y^{(k)} = A^k Y^{(o)}$. Soient $a_1, \ldots, a_n$ les composantes de $Y^{(o)}$ dans la base $(X_1, \ldots, X_n)$; dans cette même base, les composantes de $Y^{(k)}$ sont données par le vecteur-colonne :

$$\left[Y^{(k)}\right]_{X_1,\ldots,X_n} = J^k \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} \lambda_1^k & & & \\ 0 & J_2^k & & \\ & 0 & \ddots & \\ & & & J_p^k \end{bmatrix} \times \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}$$

$$= \lambda_1^k \begin{bmatrix} 1 & & & \\ & \theta(1) & & \\ & & \ddots & 0 \\ & 0 & & \theta(1) \end{bmatrix} \times \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix}$$

(λ_1 est une valeur propre simple, donc le bloc de Jordan correspondant dans J est d'ordre 1).

Il est alors possible de tenir les mêmes raisonnements que dans le cas où il existe une base de vecteurs propres de A.

Deuxième cas : $|\lambda_1| = |\lambda_2| > |\lambda_3|$

On supposera encore qu'il existe une base de vecteurs propres :

$X_1, X_2, \ldots, X_n$, correspondant respectivement aux valeurs propres : $\lambda_1, \lambda_2, \ldots, \lambda_n$.

Si $\lambda_1 = -\lambda_2$: λ_1 est réelle.

Soit $Y^{(o)} = a_1 X_1 + a_2 X_2 + a_3 X_3 + \ldots + a_n X_n$ un vecteur quelconque de $\mathbb{R}^n$.

Posons (pour tout entier k) : $Y^{(k)} = AY^{(k-1)} = A^k Y^{(o)}$.

$$Y^{(k)} = a_1 \lambda_1^k X_1 + a_2 \lambda_2^k X_2 + a_3 \lambda_3^k X_3 + \ldots + a_n \lambda_n^k X_n .$$

$$Y^{(2k)} = \lambda_1^{2k}\left[a_1 X_1 + a_2 X_2 + o\left(\left(\frac{\lambda_3}{\lambda_1}\right)^{2k}\right)\right]$$

$$Y^{(2k+1)} = \lambda_1^{2k+1}\left[a_1 X_1 - a_2 X_2 + o\left(\left(\frac{\lambda_3}{\lambda_1}\right)^{2k+1}\right)\right] .$$

Soit $(e_j)_{j=1,\ldots,n}$ la base canonique de R^n ; pour tout indice j tel que : $(a_1 X_1 + a_2 X_2).e_j \neq 0$, on aura :

$$\boxed{\frac{Y_j^{(2k+1)}}{Y_j^{(2k)}} \to \lambda_1^2 \quad \text{quand} \quad k \to +\infty}$$

Déterminons maintenant les directions des vecteurs propres correspondant à λ_1 et $-\lambda_1$ (en supposant par exemple $\lambda_1 > 0$). $(\forall k \in \mathbb{N})$:

$$Y^{(k)} + \lambda_1 Y^{(k-1)} = [a_1 \lambda_1^k X_1 + a_2 \lambda_2^k X_2 + o(\lambda_3^k)] + [a_1 \lambda_1^k X_1 + a_2 \lambda_1 \lambda_2^{k-1} X_2 + o(\lambda_1 \lambda_3^{k-1})]$$

d'où : $Y^{(k)} + \lambda_1 Y^{(k-1)} = \lambda_1^k [2a_1 X_1 + o(1)]$

alors : $\dfrac{Y^{(k)} + \lambda_1 Y^{(k-1)}}{\| Y^{(k)} + \lambda_1 Y^{(k-1)} \|} \to \dfrac{X_1}{\|X_1\|} \cdot \Sigma_1$ quand $k \to +\infty$ $\left(\Sigma_1 = \dfrac{a_1}{|a_1|}\right)$.

De même : $Y^{(2k)} - \lambda_1 Y^{(2k-1)} = \lambda_2^{2k}(2a_2 X_2 + o(1)) = \lambda_1^{2k}(2a_2 X_2 + o(1))$

$$\frac{Y^{(2k)} - \lambda_1 Y^{(2k-1)}}{\| Y^{(2k)} - \lambda_1 Y^{(2k-1)} \|} \to \frac{X_2}{\|X_2\|} \Sigma_2 \quad \text{quand} \quad k \to +\infty \quad \left(\Sigma_2 = \frac{a_2}{|a_2|}\right).$$

Si $\lambda_1 = \overline{\lambda_2}$: λ_1 est complexe ; posons $\boxed{\lambda_1 = re^{i\theta}, \ \lambda_2 = re^{-i\theta}}$.

$(\theta \not\equiv 0 \pmod{2\pi})$.

Soit $Y^{(o)} = a_1 X_1 + \ldots + a_n X_n \in \mathbb{R}^n$ et $Y^{(k)} = A^k Y^{(o)}$

$$Y^{(k)} = r^k \left[a_1 e^{ik\theta} X_1 + a_2 e^{-ik\theta} X_2 + o\left(\left(\frac{\lambda_3}{r}\right)^k\right)\right].$$

Posons : $p = -(\lambda_1 + \lambda_2) = -2r \cos\theta$.

Notons encore $(e_j)_{j=1,\ldots,n}$ la base canonique usuelle sur $\mathbb{R}^n$, et pour tout indice k, soit $Y_j^{(k)}$ la $j^{\text{ième}}$ composante de $Y^{(k)}$ sur cette base.

Les valeurs propres λ_1 et $\lambda_2 = \overline{\lambda_1}$ seront déterminées par la connaissance de r et de $\cos\theta$; donc par celle de r et p. Nous allons obtenir les quantités p et r^2 comme limites, pour k tendant vers l'infini, de suites que nous allons maintenant expliciter.

Pour tout indice j :

$$Y_j^{(k-1)}\,Y_j^{(k+2)} - Y_j^{(k)}\,Y_j^{(k+1)}$$

$$= r^{2k+1}(a_1\, e^{i(k-1)\theta}\, X_1.e_j + a_2\, e^{-i(k-1)\theta}\, X_2.e_j) \times (a_1\, e^{i(k+2)\theta}\, X_1.e_j + a_2\, e^{-i(k+2)\theta} X_2.e_j)$$

$$- r^{2k+1}(a_1\, e^{ik\theta}\, X_1.e_j + a_2\, e^{-ik\theta}\, X_2.e_j) \times (a_1\, e^{i(k+1)\theta}\, X_1.e_j + a_2\, e^{-i(k+1)\theta} X_2.e_j)$$

$$+ 0((\frac{\lambda_3}{r})^{2k+1}).r^{2k+1}$$

$$= a_1\, a_2 (X_1.e_j)(X_2.e_j) r^{2k+1}\; e^{-3i\theta} + e^{3i\theta} - e^{i\theta} - e^{-i\theta} + e((\frac{\lambda^3}{r})^{2n+1}))$$

$$Y_j^{(k-1)}\,Y_j^{(k+1)} - (Y_j^{(k)})^2$$

$$= r^{2k}(a_1\, e^{i(k-1)\theta}\, X_1.e_j + a_2\, e^{-i(k-1)\theta}\, X_2.e_j) \times (a_1\, e^{i(k+1)\theta}\, X_1.e_j + a_2\, e^{-i(k+1)\theta} X_2.e_j)$$

$$- r^{2k}(a_1\, e^{ik\theta}\, X_1.e_j + a_2\, e^{-ik\theta}\, X_2.e_j)^2 + r^{2k}\, 0((\frac{\lambda_3}{r})^{2k})$$

$$= a_1\, a_2 (X_1.e_j)(X_2.e_j) r^{2k}\, (e^{2i\theta} + e^{-2i\theta} - 2 + 0((\frac{\lambda_3}{r})^{2k}) .$$

Or on vérifie facilement que :

$$e^{3i\theta} + e^{-3i\theta} - e^{i\theta} - e^{-i\theta} = (e^{2i\theta} + e^{-2i\theta} - 2)(e^{i\theta} + e^{-i\theta})$$

$$= 2\cos\theta(e^{2i\theta} + e^{-2i\theta} - 2) .$$

Donc, si $a_1\, a_2 \neq 0$, pour tout indice j tel que : $(X_1.e_j) \neq 0$ et $X_2.e_j \neq 0$, on aura :

$$\lim_{k \to +\infty} \frac{Y_j^{(k-1)}\,Y_j^{(k+2)} - Y_j^{(k)}\,Y_j^{(k+1)}}{Y_j^{(k-1)}\,Y_j^{(k+1)} - (Y_j^{(k)})^2} = -p$$

Remarque : $e^{2i\theta} + e^{-2i\theta} - 2$ est non nul car $\theta \neq 0 \pmod{2\pi}$.

De même :

$$Y_j^{(k)} Y_j^{(k+2)} - (Y_j^{(k+1)})^2 = a_1 a_2 (X_1 . e_j)(X_2 . e_j) r^{2k+2} \left(e^{2i\theta} + e^{-2i\theta} - 2 + o\left(\left(\frac{\lambda_3}{r}\right)^{k+2}\right) \right)$$

$$Y_j^{(k-1)} Y_j^{(k+1)} - (Y_j^{(k)})^2 = a_1 a_2 (X_1 . e_j)(X_2 . e_j) r^{2k} \left(e^{2i\theta} + e^{-2i\theta} - 2 + o\left(\left(\frac{\lambda_3}{r}\right)^{2k+2}\right) \right)$$

$$\boxed{\lim_{k \to +\infty} \frac{Y_j^{(k)} Y_j^{(k+2)} - (Y_j^{(k+1)})^2}{Y_j^{(k-1)} Y_j^{(k+1)} - (Y_j^{(k)})^2} = r^2}$$

Remarque : Il s'agit là de l'adaptation au problème des valeurs propres de la méthode de Aitken de détermination de racines imaginaires conjuguées d'un polynôme.

3°) Compléments sur la factorisation des matrices.

On notera dans ce qui suit :

(inf.) (resp. (sup.)) : l'ensemble des matrices triangulaires inférieures (resp. supérieures)

(inf,1) : l'ensemble des matrices triangulaires inférieures ayant des 1 sur la diagonale

(sup,+) : l'ensemble des matrices triangulaires supérieures ayant sur la diagonale des nombres strictement positifs.

Proposition (continuité des factorisations L.U et Q.U).

Soit $(A_k)_{k\geqslant 0}$ une suite de matrices $n\times n$, convergeant vers A lorsque k tend vers l'infini.

i) si A est factorisable sous la forme $L.U$, alors il en est de même de A_k pour k assez grand ; dans ce cas si $A = L.U$ et $A_k = L_k.U_k$ avec $L, L_k \in (\text{Inf},1)$ et $U, U_k \in (\text{sup.})$ on a : $L_k \to L$ et $U_k \to U$ quand $k \to \infty$.

ii) si $A_k = Q_k.U_k$ et $A = Q.U$ avec Q_k et Q orthogonales et $U, U_k \in (\text{sup},+)$ alors : $Q_k \to Q$ et $U_k \to U$ quand $k \to \infty$.

Démonstration :

i) A étant factorisable sous la forme $L.U$ les mineurs fondamentaux de A sont non nuls. Or la suite $(A_k)_{k\geqslant 0}$ convergeant vers A, la suite des mineurs fondamentaux d'ordre p des A_k converge vers le mineur fondamental d'ordre p de A, ceci pour tout p : $1 \leqslant p \leqslant n$.

Donc pour k assez grand tous les mineurs fondamentaux de A_k sont non nuls.

Par conséquent A_k est factorisable sous la forme $L_k.U_k$ avec $L_k \in (Inf,1)$ et $U_k \in (sup)$.

Montrons maintenant que $L_k \to L$ et $U_k \to U$ lorsque $k \to \infty$.

Les A_k étant factorisables sous la forme $L_k.U_k$ pour k assez grand on peut leur appliquer la méthode de Gauss sans permutation de lignes qui donne explicitement L_k et U_k :

$$U_k = J_k^{(n-1)} \ldots J_k^{(1)} . A_k$$

et

$$L_k = (J_k^{(n-1)} \ldots J_k^{(1)})^{-1}$$

$J_k^{(1)}$ est la matrice (cf. chap.I) : $J_k^{(1)} =$

$$\begin{pmatrix} 1 & & & \\ -\frac{a_{21}^{(k)}}{a_{11}^{(k)}} & \ddots & 0 & \\ \vdots & & \ddots & \\ -\frac{a_{n1}^{(k)}}{a_{11}^{(k)}} & 0 & & 1 \end{pmatrix}$$

lorsque $k \to \infty$, $J_k^{(1)} \to J^{(1)} =$

$$\begin{pmatrix} 1 & & & \\ -\frac{a_{21}}{a_{11}} & \ddots & 0 & \\ \vdots & & \ddots & \\ -\frac{a_{n1}}{a_{11}} & 0 & & 1 \end{pmatrix}$$

(car $a_{11} \neq 0$).

Il en résulte que $J_k^{(1)} A_k \to J^{(1)} A$, en sorte que $J_k^{(2)} \to J^{(2)}$ (cf. la déf. de ces matrices $J^{(p)}$ au chap.I). On démontre de même que pour tout r tel que : $n-1 \geqslant r \geqslant 1$ on a : $J_k^{(r)} \to J^{(r)}$.

Par conséquent : $L_k = (J_k^{(n-1)} \ldots J_k^{(1)})^{-1} \to (J^{(n-1)} \ldots J^{(1)})^{-1} = L$.

Ensuite $U_k = J_k^{(n-1)} \ldots J_k^{(1)} A_k \to J^{(n-1)} \ldots J^{(1)} A = U$.

ii) Soit $A_k - Q_k.U_k$ et $A = Q.U$ avec U_k, $U \in (sup,+)$ et Q_k et Q orthogonales. On a alors $\|Q_k\| = \|Q_k^T\| = 1$ pour la norme matricielle associée à la norme euclidienne de R^n. De plus $\|U_k\| = \|Q_k^T.A_k\| \leqslant \|A_k\| < cte$ car A_k converge vers A.

On peut donc extraire de (Q_k) et (U_k) des sous-suites convergentes (Q_{k_i}) et (U_{k_i}).

Soient $\hat{Q} = \lim.Q_{k_i}$ et $\hat{U} = \lim U_{k_i}$.

$\hat{Q}$ est une matrice orthogonale car :

$$\hat{Q}.\hat{Q}^T = \lim.Q_{k_i}.Q_{k_i}^T = I .$$

D'autre part $\hat{U} \in (sup,+)$ comme limite d'une suite d'éléments de $(sup,+)$.

Or $\lim A_k = \lim Q_k U_k = \lim Q_{k_i} U_{k_i} = \hat{Q}\hat{U} = A = Q.U$. On sait d'autre part que la factorisation de A sous la forme $Q.U$ avec Q orthogonale et $U \in (sup,+)$ est unique, par conséquent : $\hat{Q} = Q$ et $\hat{U} = U$. Comme les suites (Q_k) et (U_k) sont bornées et ne possèdent que Q et U pour valeurs d'adhérence, elles convergent respectivement vers ces limites.

Proposition : Soit A une matrice $n \times n$ non singulière. Alors il existe une matrice de permutation P telle que $P.A = LU$ avec $L \in (Inf,1)$ et $U \in (sup.)$ et de plus : $P^TLP \in (Inf,1)$.

Démonstration : La première partie de la démonstration, c'est-à-dire qu'il existe P telle que $P.A = L.U$ figure au paragraphe 5-a du chap.I (p. I.19), et cela correspond à une méthode de Gauss avec échange de lignes. Il s'agit ici de trouver une loi d'échange des lignes (i.e. une matrice P) telle que P^TLP soit en outre dans $(Inf,1)$. Cette stratégie du pivot est la suivante :

On note p : le plus petit entier tel que $a_{p,1} \neq 0$.

A la première étape l'échange de lignes est celui correspondant à la permutation :

$$(1,2,\ldots,n) \rightarrow (p,1,2,\ldots,p-1,p+1,\ldots,n).$$

Soit Q la matrice associée à cette permutation.

On fait ensuite apparaître des 0 dans la 1ère colonne de la matrice QA :

$$JQA = \left(\begin{array}{c|ccc} a_{p,1} & \cdots & \cdots & a_{p,n} \\ \hline & & & \\ 0 & & A_{n-1} & \\ & & & \end{array}\right) \quad \text{avec} \quad J = \begin{pmatrix} 1 & & & \\ 0 & \ddots & & 0 \\ 0 & & \ddots & \\ \alpha_{p+1} & 0 & \ddots & \\ \vdots & & \ddots & \\ \alpha_n & & & 1 \end{pmatrix}$$

$$\alpha_i = -\frac{a_{i1}}{a_{p1}} \qquad p+1 \leqslant i \leqslant n$$

La démonstration se fait par récurrence.

Pour $n=1$, la vérification de la proposition est immédiate.

On la suppose vraie jusqu'à l'ordre $n-1$.

. A_{n-1} est non singulière car $\det(JQA) = a_{p-1} \det A_{n-1} \neq 0$ donc il existe une matrice de permutation P_{n-1} telle que

$$P_{n-1} A_{n-1} = L_{n-1} \cdot U_{n-1} \quad \text{et} \quad P_{n-1}^T L_{n-1} P_{n-1} \in (\text{Inf},1) .$$

On introduit alors les matrices $\tilde{P}_{n-1}$ et $\tilde{L}_{n-1}$ et $\tilde{U}_{n-1}$

$$\tilde{P}_{n-1} = \left(\begin{array}{c|c} 1 & 0 \\ \hline 0 & P_{n-1} \end{array}\right) \qquad \tilde{L}_{n-1} = \left(\begin{array}{c|c} 1 & 0 \\ \hline 0 & L_{n-1} \end{array}\right)$$

$$\tilde{U}_{n-1} = \left(\begin{array}{c|c} a_{p1} & a_{p2} \cdots a_{pn} \\ \hline 0 & U_{n-1} \end{array}\right)$$

. Posons $P = \tilde{P}_{n-1} \cdot Q$ et $L = \tilde{P}_{n-1}\, J^{-1}\, \tilde{P}_{n-1}^T\, \tilde{L}_{n-1}$.

D'après le chap.I, par.5, $L \in (Inf,1)$ et $PA = L.\tilde{U}_{n-1}$.

Il reste à vérifier que $P^T L P \in (Inf,1)$.

$$P^T L P = Q^T\, \tilde{P}_{n-1}^T\, \tilde{P}_{n-1}\, J^{-1}\, \tilde{P}_{n-1}^T\, \tilde{L}_{n-1}\, \tilde{P}_{n-1}\, Q$$

$$= Q^T\, J^{-1}\, \bar{L}_n\, Q \quad \text{avec} \quad \bar{L}_n = \tilde{P}_{n-1}^T\, \tilde{L}_{n-1}\, \tilde{P}_{n-1} .$$

D'après l'hypothèse de récurrence $\bar{L}_n \in (Inf,1)$.

Considérons $J^{-1}\, \bar{L}_n$:

$$J^{-1}\, \bar{L}_n = \begin{pmatrix} 1 & 0 \\ 0 & \\ \vdots & \\ 0 & \\ -\alpha_{p+1} & I_{n-1} \\ \vdots & \\ -\alpha_n & \end{pmatrix} \begin{pmatrix} 1 & 0 \\ & \\ 0 & 0 \end{pmatrix}$$

$$= \begin{pmatrix} 1 & 0 \\ 0 & \\ \vdots & \\ 0 & 0 \\ -\alpha_{p+1} & \\ \vdots & \\ -\alpha_n & \end{pmatrix}$$

. En multipliant $J^{-1}\, \bar{L}_n$ à gauche par Q^T on obtient :

$$Q^T (J^{-1} \bar{L}_n) = \begin{pmatrix} 0 & \\ \vdots & 0 \\ 0 & \\ 1 & 0 \\ -\alpha_{p+1} & \\ \vdots & \\ -\alpha_n & \end{pmatrix}$$

. enfin en multipliant à droite par Q :

$$P^T L P = Q^T (J^{-1} \bar{L}_n) Q = \begin{pmatrix} \triangle & \begin{matrix} 0 \\ \vdots \\ 0 \end{matrix} & 0 \\ 0 \ldots 0 \quad 1 & 0 \ldots 0 & \\ \square & \begin{matrix} -\alpha_{p+1} \\ \vdots \\ -\alpha_n \end{matrix} & \triangle \end{pmatrix}$$

C'est bien une matrice triangulaire inférieure portant des 1 sur la diagonale.

4°) Algorithme L.R .

Description : Partant d'une matrice $n \times n$, A , on cherche à construire une suite (A_k) de matrices semblables à A , donc qui ont les mêmes valeurs propres que A , et qui donne à la limite $(k \to \infty)$ des informations sur les valeurs propres de A .

Les R_i désignent dans la méthode des matrices triangulaires supérieures et les L_i des éléments de (Inf,1). Supposons A factorisable sous la forme L.U , on pose :

$$A_1 = A = L_1 . R_1$$

$$A_2 = R_1 L_1 = L_1^{-1} A L_1$$

si A_2 est factorisable sous la forme LU , on pose :

$$A_2 = L_2 . R_2$$

$$A_3 = R_2 . L_2$$

$$\vdots$$

$A_k = L_k . R_k$ si A_k est factorisable L.U

$$A_{k+1} = R_k . L_k = L_k^{-1} A_k L_k .$$

Formules. On a : $A_{k+1} = L_k^{-1}(L_{k-1}^{-1} A_{k-1} L_{k-1})L_k$.

Par itération, on obtient :

$$A_{k+1} = L_k^{-1}\ldots L_1^{-1} A L_1\ldots L_k .$$

Soit $M_k = L_1\ldots L_k$, $M_k \in (\text{Inf},1)$.

On a alors : $A_{k+1} = M_k^{-1} A M_k$.

Soit encore : $M_k A_{k+1} = A M_k$.

Posons $S_k = R_k\ldots R_1$; $S_k \in (\text{sup.})$.

On a : $M_k S_k = L_1\ldots L_k R_k\ldots R_1$

$= L_1\ldots L_{k-1} A_k R_{k-1}\ldots R_1$

$= A M_{k-1} S_{k-1}$, d'où en réitérant : $M_k S_k = A^k$.

Convergence de l'algorithme L.R .

Soit A une matrice $n\times n$. On suppose que A admet n valeurs propres distinctes en module : $|\lambda_1| > |\lambda_2|\ldots > |\lambda_n|$. On suppose en outre que A et toutes les A_k sont factorisables sous la forme L.U (conditions pour que l'algorithme soit applicable).

Soit D la matrice diagonale des valeurs propres :

$$D = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}$$

Il existe X telle que $A = XDX^{-1} = XDY$ avec $Y = X^{-1}$.

Alors on a : $A^k = XD^kX^{-1} = XD^kY$.

a) On suppose d'abord que X et Y sont factorisables sous la forme L.U ($L \in (\text{Inf},1)$ et $U \in (\text{sup})$).

$$X = L_x . U_x \quad \text{et} \quad Y = L_y . U_y \ .$$

Alors : $A^k = L_x\ U_x\ D^k\ L_y\ U_y$

$$= L_x\ U_x (D^k\ L_y\ D^{-k}) D^k\ U_y \ .$$

Soient $d_{i,j}$ les éléments de $D^k\ L_y\ D^{-k}$ et ℓ_{ij} les éléments de L_y. On a :

$$d_{ij} = \begin{cases} 0 & \text{si } i < j \\ 1 & \text{si } i = j \\ \ell_{ij} \left(\frac{\lambda_i}{\lambda_j}\right)^k & \text{si } i > j \end{cases}$$

Or $\left|\frac{\lambda_i}{\lambda_j}\right| < 1$ pour $i > j$, donc $\left|\frac{\lambda_i}{\lambda_j}\right|^k \to 0$ quand $k \to \infty$.

Par conséquent $D^k\ L_y\ D^{-k} \to I$ quand $k \to \infty$, et on peut écrire

$D^k\ L_y\ D^{-k} = I + E_k$ avec $E_k = o(1)$.

Donc : $A^k = L_x\ U_x (I + E_k) D^k\ U_y$

$$= L_x (U_x + U_x E_k) D^k\ U_y$$

$$= L_x (I + U_x\ E_k\ U_x^{-1}) U_x\ D^k\ U_y \ .$$

Considérons $I + U_x\ E_k\ U_x^{-1}$; cette matrice tend vers I quand $k \to \infty$. D'après le i) de la proposition du paragraphe 3, pour k assez grand :

$$I + U_x\ E_k\ U_x^{-1} = \hat{L}_k . \hat{U}_k \quad \text{avec} \quad \hat{L}_k \in (\text{Inf},1) \quad \hat{U}_k \in (\text{sup.}) \ .$$

De plus $\hat{L}_k \to I$ et $\hat{U}_k \to I$ quand $k \to \infty$.

A^k s'écrit alors : $A^k = L_x\ \hat{L}_k\ \hat{U}_k\ U_x\ D^k\ U_y = M_k\ S_k$.

$L_x\ \hat{L}_k \in (\text{Inf},1)$; $\hat{U}_k\ U_x\ D^k\ U_y \in (\text{sup.})$.

La décomposition de A^k sous forme $L.U$ étant unique lorsque $L \in (\text{Inf},1)$ et $U \in (\text{sup})$, on obtient par identification :

$$L_x\ \hat{L}_k = M_k \quad \text{et} \quad \hat{U}_k\ U_x\ D^k\ U_y = S_k \ .$$

D'où $M_k \to L_x$ lorsque $k \to \infty$ et $A_{k+1} = M_k^{-1}\ A\ M_k \to L_x^{-1}\ A\ L_x$ quand $k \to \infty$.

On a d'autre part :

$$L_x^{-1} A L_x = L_x^{-1} XDY L_x$$
$$= L_x^{-1} L_x U_x D U_x^{-1} L_x^{-1} L_x$$
$$= U_x D U_x^{-1} \in (\text{sup}) .$$

Donc : $A_{k+1} \to U_x D U_x^{-1}$ quand $k \to \infty$.

Or $U_x D U_x^{-1}$ est semblable à D et de la forme : $\begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}$

On a donc la conclusion suivante : (A_k) est convergente et la diagonale de A_k tend vers $(\lambda_1,\ldots,\lambda_n)$.

b) On ne suppose plus maintenant que Y est factorisable sous la forme L.U.

D'après la proposition du paragraphe 3, il existe une matrice de permutation P_y telle que :

$$P_y . Y = L_y . U_y \quad \text{et} \quad P_y^T L_y P_y \in (\text{Inf},1)$$

Donc
$$A^k = X D^k Y$$
$$= X D^k P_y^T L_y U_y$$
$$= X P_y^T (P_y D^k P_y^T) L_y . U_y .$$

. $P_y D^k P_y^T$ est diagonale.

Si on pose $P_y D P_y^T = D_o = \begin{pmatrix} \lambda_{i_1} & & 0 \\ & \ddots & \\ & & \ddots \\ 0 & & \lambda_{i_n} \end{pmatrix}$

On a $P_y D^k P_y^T = D_o^k$.

A^k s'écrit alors : $A^k = X P_y^T (D_o^k L_y D_o^{-k}) D_o^k U_y$

et $D_o^k L_y D_o^{-k} = P_y D^k P_y^T L_y P_y D^{-k} P_y^T = P_y D^k \bar{L}_y D^{-k} P_y^T$

avec $\bar{L}_y = P_y^T L_y P_y \in (Inf,1)$.

Soient $\bar{d}_{ij}$ les éléments de la matrice $D^k \bar{L}_y D^{-k}$ et $\bar{\ell}_{ij}$ les éléments de $\bar{L}_y$. On a alors :

$$\bar{d}_{ij} = \begin{cases} 0 & j > i \\ 1 & j = i \\ \bar{\ell}_{ij}\left(\frac{\lambda_i}{\lambda_j}\right)^k & j < i \end{cases}$$

Donc $D_k \bar{L}_y D^{-k} \to I$ lorsque $k \to \infty$.

On peut donc écrire :

$$P_y D^k \bar{L}_y D^{-k} P_y^T = I + E_k \quad \text{avec} \quad E_k = \mathcal{O}(1).$$

. A^k s'écrit alors :

$$A^k = XP_y^T(I + E_k)D_o^k U_y .$$

. Supposons XP_y^T factorisable sous la forme L.U :

$$XP_y^T = L_x . U_x .$$

Alors
$$A^k = L_x U_x(I + E_k)D_o^k U_y$$
$$= L_x(I + U_x E_k U_x^{-1})U_x D_o^k U_y .$$

. La matrice $I + U_x E_k U_x^{-1}$ tend vers I quand $k \to \infty$.

D'après la proposition i) du par.3 on peut écrire, pour k assez grand, $I + U_x E_k U_x^{-1} = \hat{L}_k . \hat{U}_k$ avec $\hat{L}_k \in (Inf,1)$ et $\hat{U}_k \in (sup.)$. De plus $\hat{L}_k \to I$ et $\hat{U}_k \to I$ quand $k \to \infty$.

Pour k assez grand A^k s'écrit donc :

$$A^k = L_x . \hat{L}_k \hat{U}_k U_x D_o^k U_y = M_k . S_k .$$

$L_x \hat{L}_k \in (\text{Inf},1)$ et $\hat{U}_k . U_x D_o^k U_y \in (\text{sup.})$. on obtient donc par identification :

$$M_k = L_x \hat{L}_k \quad \text{et} \quad S_k = \hat{U}_k U_x D_o^k U_y .$$

Donc M_k tend vers L_x quand $k \to \infty$.

D'où $A_{k+1} = M_k^{-1} A M_k$ tend vers $L_x^{-1} A L_x$ quand $k \to \infty$.

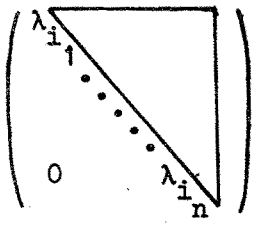
D'autre part $L_x^{-1} A L_x = L_x^{-1} L_x U_x P_y D P_y^T U_x^{-1} L_x^{-1} L_x$

$$= L_x^{-1} X D_o X^{-1} L_x$$

$$= U_x D_o U_x^{-1} .$$

Donc $A_{k+1} \to U_x D_o U_x^{-1}$ quand $k \to \infty$.

Or $U_x D_o U_x^{-1}$ est de la forme



Proposition : Sous les hypothèses suivantes :

i) $|\lambda_1| > |\lambda_2| > \ldots > |\lambda_n|$.

ii) les matrices A_k sont factorisables sous la forme L.U

iii) X et $Y = X^{-1}$ sont factorisables L.U

ou iii bis) XP_y^T est factorisable L.U (P_y matrice de permutation associée à Y).

Alors :

L'algorithme L.R est constructible ; les matrices A_k convergent vers une matrice triangulaire supérieure et la diagonale de A_k tend vers :

- $(\lambda_1, \ldots, \lambda_n)$ dans le cas iii)
- $(\lambda_{i1}, \ldots, \lambda_{in}) = \text{perm.}(\lambda_1, \ldots, \lambda_n)$ dans l'éventualité iii bis).

Remarque : Pour le cas des matrices symétriques, définies > 0 , cf. au par.6 une modification de l'algorithme LR : l'algorithme L.R - Cholesky.

5°) Algorithme Q.R .

Rappel : On appelle (sup,+) l'ensemble des matrices carrées, triangulaires supérieures, dont les éléments diagonaux sont strictement positifs.

Description de l'algorithme QR :

On utilise ici la décomposition des matrices sous la forme Q.U avec Q orthogonale et U dans (sup,+). Dans la méthode les éléments de (sup,+) seront notés R_i . Cette décomposition existe et est unique. En effet soit A une matrice régulière et Q'U' une factorisation QU de A ; alors toutes les autres factorisations seront du type QU avec $\begin{cases} Q = Q'\Delta^{-1} \\ U = \Delta U' \end{cases}$ où Δ est une matrice diagonale unitaire. Il existe Δ unique (et donc une factorisation QU unique) telle que la diagonale de U soit à éléments positifs : on a $U_{ii} = \Delta_{ii}\, U'_{ii}$ on prend donc $\Delta_{ii} = \dfrac{U'_{ii}}{|U'_{ii}|}$ (U'_{ii} est différent de 0 puisque A est régulière et U' triangulaire supérieure).

Soit A une matrice régulière. On construit une suite de matrices A_k de la façon suivante : on pose $A_1 = A = Q_1 . R_1$

$$A_2 = R_1\, Q_1 = Q_1^T\, Q_1\, R_1\, Q_1 = Q_1^T\, A_1\, Q_1 .$$

A_2 est semblable à A_1 et non singulière.

On suppose que l'on a obtenu A_k non singulière et semblable à A . Alors on factorise A_k : $A_k = Q_k\, R_k$ et on pose $A_{k+1} = R_k\, Q_k$

$$A_{k+1} = R_k\, Q_k = Q_k^T\, A_k\, Q_k$$

A_{k+1} est semblable à A_k, donc semblable à A et non singulière.

On construit donc, ainsi, par récurrence, une suite A_k de matrices non singulières, toutes semblables à A, donc ayant les mêmes valeurs propres que A.

Formules : On a $A_{k+1} = Q_k^T Q_{k-1}^T A_{k-1} Q_{k-1} Q_k = Q_k^T \ldots Q_1^T A Q_1 \ldots Q_k$.

Posons $M_k = Q_1 \ldots Q_k$

$S_k = R_k \ldots R_1$

Avec ces notations :

$$A_{k+1} = M_k^T A M_k .$$

Soit $M_k A_{k+1} = A M_k$.

On a d'autre part :

$$\begin{aligned} M_k S_k &= Q_1 \ldots Q_k R_k \ldots R_1 \\ &= Q_1 \ldots Q_{k-1} A_k R_{k-1} \ldots R_1 \\ &= M_{k-1} A_k S_{k-1} \\ &= A M_{k-1} S_{k-1} \end{aligned}$$

D'où $M_k S_k = A^k$.

Mise en oeuvre pratique de la méthode.

En appliquant la méthode de Householder on construit des matrices orthogonales : $H_k^{(1)}, H_k^{(2)}, \ldots, H_k^{(n-1)}$ telles que $U_k = H_k^{(n-1)} \ldots H_k^{(1)} A_k$ soit triangulaire supérieure. Il existe une matrice unique qui est diagonale et unitaire Δ_k telle que : $H_k^{(n-1)} \ldots H_k^{(1)} A_k = \Delta_k R_k$ où R_k est un élément de (sup,+).

Comme $\Delta_k = \Delta_k^{-1}$ on a : $\Delta_k H_k^{(n-1)} \ldots H_k^{(1)} A_k = R_k$.

Δ_k vérifie : $(\Delta_k)_{ii} = \dfrac{(U_k)_{ii}}{|(U_k)_{ii}|}$.

Convergence de l'algorithme QR .

1°) Cas où $|\lambda_1| > |\lambda_2| \ldots > |\lambda_n| > 0$.

Alors **A** est diagonalisable et il existe une matrice **X** inversible telle que $\mathbf{A} = XDX^{-1}$ avec $D = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}$

On pose $Y = X^{-1}$.

a) On suppose que Y admet une factorisation du type LU .

$Y = L_y U_y$ où L_y est un élément de (Inf,1) et U_y est triangulaire supérieure. On sait qu'une telle décomposition est unique.

Alors on écrit $\mathbf{A}^k = X D^k Y = X D^k L_y U_y = X(D^k L_y D^{-k})D^k U_y$

$$(D^k L_y D^{-k})_{ij} = L_{y\,ij} \left(\frac{\lambda_i}{\lambda_j}\right)^k = \begin{cases} 0 & \text{si } j > i \\ 1 & \text{si } i = j \\ (L_y)_{ij} \left(\frac{\lambda_i}{\lambda_j}\right)^k & \text{si } j < i \,. \end{cases}$$

Or si j est strictement inférieur à i on a $\left|\frac{\lambda_i}{\lambda_j}\right| < 1$; donc $\left(\frac{\lambda_i}{\lambda_j}\right)^k$ tend vers 0 quand k tend vers l'infini. $D^k L_y D^{-k}$ converge donc vers la matrice identité.

On pose $E_k = D^k L_y D^{-k} - I$; E_k tend vers 0 quand k tend vers l'infini.

On a $\mathbf{A}^k = X(I + E_k)D^k U_y$.

X admet une décomposition QU où U est élément de (sup, +) : $X = Q_x R_x$.

On en déduit $\mathbf{A}^k = Q_x R_x(I + E_k)D^k U_y = Q_x(I + R_x E_k R_x^{-1})R_x D^k U_y$.

$I + R_x E_k R_x^{-1}$ tend vers I quand k tend vers l'infini, $I + R_x E_k R_x^{-1}$ admet une décomposition unique QU , $U \in$ (sup,+).

$$I + R_x E_k R_x^{-1} = \hat{Q}_k \hat{R}_k .$$

$\hat{Q}_k \hat{R}_k \xrightarrow[k\to\infty]{} I$. Donc d'après les propriétés rappelées au paragraphe précédent sur la continuité des factorisations Q.U et du fait que I s'écrit de façon unique : $I = QR$ avec $Q=I$ et $R=I$, $\hat{Q}_k$ tend vers I et $\hat{R}_k$ tend vers I quand k tend vers l'infini.

$A^k = (Q_x \hat{Q}_k)(\hat{R}_k R_x D^{-k} U_y)$ est une factorisation Q.U de A^k .

D'autre part $A^k = M_k S_k$ en est une autre. La factorisation du type QU étant essentiellement unique, il existe une matrice Δ_k diagonale et unitaire telle que :

$$\begin{cases} M_k = (Q_x \hat{Q}_k)\Delta_k^{-1} \\ S_k = \Delta_k(\hat{R}_k R_x D^{-k} U_y) \end{cases}$$

On a donc : $M_k \Delta_k = Q_x \hat{Q}_k$.

$M_k \Delta_k$ tend donc vers Q_x quand k tend vers l'infini.

On a $A_{k+1} = M_k^T A M_k$.

Donc $\Delta_k^{-1} A_{k+1} \Delta_k = (\Delta_k^{-1} M_k^T) A (M_k \Delta_k)$.

Comme $\Delta_k^{-1} = \Delta_k^T$, $\Delta_k^{-1} A_{k+1} \Delta_k$ tend vers $Q_x^T A Q_x$ quand k tend vers l'infini.

Or $Q_x^T A Q_x = Q_x^T (XDX^{-1}) Q_x$

$= Q_x^T Q_x R_x D R_x^{-1} Q_x^T Q_x$

$= R_x D R_x^{-1}$.

$R_x D R_x^{-1}$ est triangulaire supérieure et semblable à D ; on a :

$$(R_x D R_x^{-1})_{ii} = D_{ii} = \lambda_i .$$

D'autre part Δ_k étant diagonale et unitaire, ses éléments diagonaux sont tous égaux à 1 en module

$$\Delta_k^{-1} A_{k+1} \Delta_k = \Delta_k A_{k+1} \Delta_k .$$

Donc $(\Delta_k^{-1} A_{k+1} \Delta_k)_{ij} = (A_{k+1})_{ij} \Delta_{ii}^{(k)} \Delta_{jj}^{(k)}$.

On en déduit :

$$\begin{cases} |\Delta_k^{-1} A_{k+1} \Delta_k| = |A_{k+1}| \\ (\Delta_k^{-1} A_{k+1} \Delta_k)_{ii} = (A_{k+1})_{ii} \quad \text{pour} \quad 1 \leqslant i \leqslant n \ . \end{cases}$$

Proposition : Sous l'hypothèse que toutes les valeurs propres sont distinctes en module et que Y admet la factorisation LU alors $|A_{k+1}|$ converge vers $|R_x D R_x^{-1}|$ et $(A_{k+1})_{ii}$ converge vers λ_i pour tout i compris entre 1 et n , quand k tend vers l'infini.

b) Cas où Y n'admet pas la factorisation LU .

Dans ce cas, il existe une matrice de permutation P_y telle que $P_y Y = L_y U_y$ où $\bar{L}_y = P_y^T L_y P_y$ est aussi triangulaire inférieure et n'a que des 1 sur la diagonale. La démonstration est sensiblement la même que dans le cas a) :

$$A^k = X D^k Y = X D^k P_y^T L_y U_y = X P_y^T P_y D^k P_y^T L_y U_y = X P_y^T P_y D^k \bar{L}_y P_y^T U_y \ .$$

$$A^k = X P_y^T P_y (D^k \bar{L}_y D^{-k}) P_y^T P_y D^k P_y^T U_y \ .$$

$$(D^k \bar{L}_y D^{-k})_{ij} = (\bar{L}_y)_{ij} (\frac{\lambda_i}{\lambda_j})^k = \begin{cases} 0 & \text{si} \quad j > i \\ 1 & \text{si} \quad j = i \\ (\bar{L}_y)_{ij} (\frac{\lambda_i}{\lambda_j})^k & \text{si} \quad j < i \ . \end{cases}$$

Si $j < i$: $\left|\frac{\lambda_i}{\lambda_j}\right| < 1$. Donc $(\frac{\lambda_i}{\lambda_j})^k$ tend vers 0 quand k tend vers l'infini et $D^k \bar{L}_y D^{-k}$ converge vers la matrice identité.

On pose $E_k = P_y (D^k \bar{L}_y D^{-k}) P_y^T - I$.

E_k tend vers 0 quand k tend vers l'infini.

On a : $A^k = X P_y^T [P_y D^k \bar{L}_y D^{-k} P_y^T] P_y D^k P_y^T U_y = X P_y^T (I + E_k)(P_y D^k P_y^T) U_y$.

P_y est la matrice associée à une permutation. σ des indices 1,2,...,n .

Si on pose : $D_o = P_y\ D\ P_y^T$ alors $(D_o)_{ij} = \lambda_{\sigma(i)}\,\delta_{\sigma(i),\sigma(j)}$.

Donc
$$D_o = \begin{pmatrix} \lambda_{\sigma(1)} & & & \\ & \ddots & & 0 \\ & & \ddots & \\ & 0 & & \lambda_{\sigma(n)} \end{pmatrix}$$

On a donc $A^k = X\ P_y^T(I+E_k)D_o^k\ U_y$. $X\ P_y^T$ admet une factorisation unique QU avec U élément de $(\sup,+)$: $Q_x\ R_x$.

On en déduit $A^k = Q_x\ R_x(I+E_k)D_o^k\ U_y$. Ce que l'on peut écrire encore :

$$A^k = Q_x(I+R_x\ E_k\ R_x^{-1})R_x\ D_o^k\ U_y\ .$$

$I+R_x\ E_k\ R_x^{-1}$ admet une factorisation unique : $I+R_x\ E_k\ R_x^{-1} = \hat{Q}_k\ \hat{R}_k$ et comme $I+R_x\ E_k\ R_x^{-1}$ tend vers I : $\hat{Q}_k$ et $\hat{R}_k$ tendent aussi vers I , quand k tend vers l'infini.

$A^k = (Q_x\ \hat{Q}_k)(\hat{R}_k\ R_x\ D_o^k\ U_y)$ est une factorisation Q.U de A^k .

D'autre part $A^k = M_k\ S_k$ en est une autre. La factorisation du type QU étant essentiellement unique, il existe une matrice Δ_k diagonale et unitaire telle que

$$\begin{cases} M_k = (Q_x\ \hat{Q}_k)\Delta_k^{-1} \\ S_k = \Delta_k(\hat{R}_k\ R_x\ D_o^{+k}\ U_y)\ . \end{cases}$$

$M_k\ \Delta_k$ tend vers Q_x quand k tend vers l'infini.

On a : $A_{k+1} = M_k^T\ A\ M_k$.

Donc $\Delta_k^T\ A_{k+1}\ \Delta_k = (M_k\ \Delta_k)^T\ A\ (M_k\ \Delta_k) = (Q_x\ \hat{Q}_k)^T\ A\ (Q_x\ \hat{Q}_k)$.

Donc $|\Delta_k^T\ A_{k+1}\ \Delta_k| = |A_{k+1}|$ tend vers $Q_x^T\ A\ Q_x$ quand k tend vers l'infini.

On a : $Q_x^T\ A\ Q_x = Q_x^T(Q_x\ R_x\ P_y)D(Q_x\ R_x\ P_y)^{-1}Q_x = R_x\ D_o\ R_x^{-1}$.

$R_x\ D_o\ R_x^{-1}$ est triangulaire supérieure et semblable à D_o ; donc :

$$(R_x\ D_o\ R_x^{-1})_{ii} = (D_o)_{ii} = \lambda_{\sigma(i)} \qquad 1 \leqslant i \leqslant n\ .$$

On a : $\begin{cases} (\Delta_k^T A_{k+1} \Delta_k)_{ii} = (A_{k+1})_{ii} \\ (|\Delta_k^T A_{k+1} \Delta_k|)_{ij} = (|A_{k+1}|)_{ij} \end{cases}$.

Donc $(A_{k+1})_{ii}$ tend vers $\lambda_{\sigma(i)}$ quand k tend vers l'infini.

Proposition : Sous l'hypothèse que les valeurs propres $\lambda_1,\ldots,\lambda_n$ sont toutes distinctes en module $|A_{k+1}|$ converge quand k tend vers l'infini et pour tout i entre 1 et n $(A_{k+1})_{ii}$ tend vers $\lambda_{\sigma(i)}$ où σ est une permutation de $\{1,2,\ldots,n\}$. (Plus précisément la permutation associée à P_y : $P_y Y = L_y U_y$).

2°) Quelques cas où les valeurs propres sont égales.

a) On suppose ceci :

$\begin{cases} |\lambda_1| > |\lambda_2| > \ldots |\lambda_r| = |\lambda_{r+1}| = \ldots = |\lambda_p| > \ldots > |\lambda_n| > 0 \\ \lambda_r = \lambda_{r+1} = \ldots = \lambda_p \\ \text{et l'espace propre associé à } \lambda_r \text{ est de dimension } p-r+1 . \end{cases}$

α) Supposons que Y admette une factorisation LU : $Y = L_y U_y$.

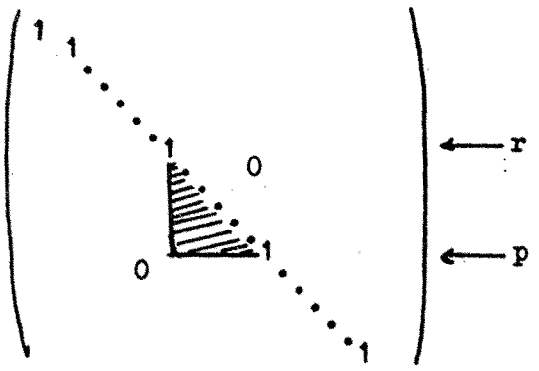
$$A^k = X D^k Y = X D^k L_y U_y = X D^k L_y (D^{-k} D^k) U_y$$

$$(D^k L_y D^{-k})_{ij} = (L_y)_{ij} \left(\frac{\lambda_i}{\lambda_j}\right)^k = \begin{cases} 0 & \text{si } j > i \\ 1 & \text{si } j = i \\ (L_y)_{ij} \left(\frac{\lambda_i}{\lambda_j}\right)^k & \text{si } j < i . \end{cases}$$

. Si $j < i$ et, si j ou i n'appartient pas à $[r,r+1,\ldots,p]$, $(L_y)_{ij} \left(\frac{\lambda_i}{\lambda_j}\right)^k$ tend vers 0 quand k tend vers l'infini.

. Si $j < i$ et si j et i appartiennent à $[r,r+1,\ldots,p]$ $\frac{\lambda_i}{\lambda_j} = 1$.

$$D^k L_y D^{-k} \xrightarrow[k\to\infty]{} \tilde{L} =$$



On pose $E_k = D^k L_y D^{-k} - \tilde{L}$.

E_k tend vers 0 quand k tend vers l'infini.

On a donc $A^k = X(\tilde{L} + E_k)D^{+k} U_y = X\tilde{L}(I + \tilde{L}^{-1} E_k)D^{+k} U_y$.

$X\tilde{L}$ s'écrit de façon unique : $X\tilde{L} = Q_x R_x$.

Donc $A^k = Q_x R_x(I + \tilde{L}^{-1} E_k)D^{+k} U_y = Q_x(I + R_x \tilde{L}^{-1} E_k R_x^{-1})R_x D^{+k} U_y$.

On écrit $I + R_x \tilde{L}^{-1} E_k R_x^{-1} = \hat{Q}_k \hat{R}_k$ de façon unique. On en déduit :

$$A^k = (Q_x \hat{Q}_k)(\hat{R}_k R_x D^{+k} U_y)$$

$\hat{R}_k \in (\sup,+)$

$\hat{Q}_k \xrightarrow[k\to\infty]{} I$

$\hat{R}_k \xrightarrow[k\to\infty]{} I$.

D'autre part $A^k = M_k S_k$, il existe donc une matrice orthogonale telle que

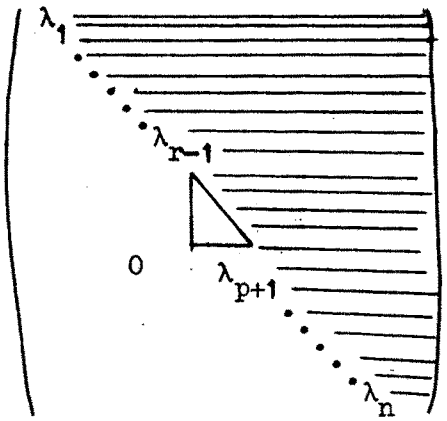
$M_k = Q_x \hat{Q}_k \Delta_k^{-1}$. Donc $M_k \Delta_k = Q_x \hat{Q}_k$ tend vers Q_x

$$\Delta_k^T A_{k+1} \Delta_k = (M_k \Delta_k)^T A (M_k \Delta_k).$$

Donc $\Delta_k^T A_{k+1} \Delta_k$ converge vers $Q_x^T A Q_x$.

Or $Q_x^T A Q_x = Q_x^T(Q_x R_x \tilde{L}^{-1})D(\tilde{L} R_x^{-1} Q_x^T)Q_x = R_x \tilde{L}^{-1} D \tilde{L} R_x^{-1}$ qui est une matrice

"presque" triangulaire supérieure du type :



Conclusion : Dans le cas considéré $|A_{k+1}|$ converge quand k tend vers l'infini et pour tout i n'appartenant pas à $[r,r+1,\ldots,p]$, $(A_{k+1})_{ii}$ converge vers λ_i quand k tend vers l'infini. Cette méthode s'étend au cas de plusieurs groupes de valeurs propres égales en module.

β) Précisons à présent ce qui se passe si Y n'est pas factorisable L.U. ; conformément à un lemme précédemment démontré, il existe P_y matrice de permutation telle que :

$$P_y Y = L_y U_y \quad \text{et} \quad P_y^T L_y P_y = \bar{L}_y \in (\text{inf},1).$$

Reprenons nos calculs avec cette modification.

$$A^k = X D^k Y = X D^k P_y^T L_y U_y = X D^k P_y^T L_y P_y P_y^T U_y$$

$$A^k = X D^k \bar{L}_y D^{-k} D^k P_y^T U_y$$

$$(D^k \bar{L}_y D^{-k})_{ij} = \bar{\ell}_{ij} \left(\frac{\lambda_i}{\lambda_j}\right)^k$$

et de même que précédemment :

$$\bar{\ell}_{ij} \left(\frac{\lambda_i}{\lambda_j}\right)^k = 0 \quad \text{si} \quad i < j$$

$$\bar{\ell}_{ij} \left(\frac{\lambda_i}{\lambda_j}\right)^k = 1 \quad \text{si} \quad i = j$$

$$\bar{\ell}_{ij} \left(\frac{\lambda_i}{\lambda_j}\right)^k \to 0 \quad \text{si} \quad i > j \quad \text{sauf si} \quad r \leqslant j < i \leqslant t .$$

$$D^k \bar{L}_y D^{-k} = \tilde{L} + E_k \quad \text{avec} \quad E_k = o(1) \text{ lorsque } k \to \infty ,$$

$\tilde{L}$ du même type que précédemment mais différente. On a :

$$A^k = X(\tilde{L} + E_k)D^k P_y^T U_y = X\tilde{L}(I + \tilde{L}^{-1} E_k)P_y^T P_y D^k P_y^T U_y$$

$$A^k = X \tilde{L} P_y^T(I + P_y \tilde{L}^{-1} E_k P_y^T)(P_y D^k P_y^T)U_y .$$

Posons :

$$P_y \tilde{L}^{-1} E_k P_y^T = F_k$$

alors $F_k \to 0$ lorsque $k \to \infty$.

D'autre part $X \tilde{L} P_y^T$ est non singulière donc factorisable sous la forme QU :

$$X \tilde{L} P_y^T = QU$$

Donc :

$$A^k = QU(I + F_k)(P_y D^k P_y^T)U_y$$

$$A^k = Q(I + U F_k U^{-1})U P_y D^k P_y^T U_y .$$

Mais $I + U F_k U^{-1} \to I$ lorsque $k \to \infty$, donc d'après une proposition précédente $I + U F_k U^{-1}$ est factorisable sous la forme $\hat{Q}_k \hat{U}_k$, $\hat{U}_k \in (\sup,+)$, et de plus

$$\hat{Q}_k \to I \ , \quad \hat{U}_k \to I \quad \text{lorsque} \quad k \to \infty .$$

On écrit donc :

$$A^k = Q \hat{Q}_k \hat{U}_k U (P_y D^k P_y^T)U_y$$

$Q \hat{Q}_k$ est une matrice orthogonale, $\hat{U}_k U P_y D^k P_y^T U_y$ est une matrice triangulaire supérieure.

Puisque $A^k = M_k S_k$ et que la factorisation Q.U. de A^k est essentiellement unique, il existe une matrice Δ_k, diagonale et orthogonale telle que :

$$M_k = Q \hat{Q}_k \Delta_k^{-1} .$$

Puisque $A_{k+1} = M_k^T A M_k$ et $M_k \Delta_k \to Q$ pour $k \to \infty$,

$$\Delta_k^T A_{k+1} \Delta_k = \Delta_k^T M_k^T A M_k \Delta_k \to Q^T A Q \quad \text{pour} \quad k \to \infty .$$

Mais :

$$A = X D X^{-1} \quad \text{et} \quad X = Q U P_y \tilde{L}^{-1} .$$

Donc :

$$Q^T AQ = Q^T Q U P_y \tilde{L}^{-1} D \tilde{L} P_y^T U^{-1} Q^T Q = U P_y \tilde{L}^{-1} D \tilde{L} P_y^T U^{-1} .$$

Dans ce cas la matrice A^k converge pour $k \to \infty$.

Un cas important est celui des matrices symétriques définies positives (elles ont des valeurs propres réelles et donc différentes en module ou égales). Donc dans les deux cas l'algorithme QR converge mais alors l'espace propre associé à la valeur propre λ multiple a pour dimension l'ordre de multiplicité de λ.

b) Cas des valeurs propres égales en module.

Nous envisageons le cas où :

$$|\lambda_1| > |\lambda_2| > \ldots > |\lambda_r| = |\lambda_{r+1}| = \ldots = |\lambda_t| > |\lambda_{t+1}| > \ldots > |\lambda_n| ,$$

A étant supposée diagonalisable :

$$A = X D X^{-1} = X D Y$$

et Y étant factorisable sous la forme $L_y U_y$ (i.e. les mineurs fondamentaux de Y sont non nuls).

Pour $r \leqslant j \leqslant t$: $\lambda_j = |\lambda_j| \exp i \theta_j$ $(i = \sqrt{-1})$.

Reprenons les calculs précédents :

$$(D^k L_y D^{-k})_{i,j} = \ell_{ij}\left(\frac{\lambda_i}{\lambda_j}\right)^k$$

$$\ell_{ij}\left(\frac{\lambda_i}{\lambda_j}\right)^k = 0 \quad \text{si} \quad j > i$$

$$\ell_{ij}\left(\frac{\lambda_i}{\lambda_j}\right)^k = 1 \quad \text{si} \quad j = i$$

$$\ell_{ij}\left(\frac{\lambda_i}{\lambda_j}\right)^k = \exp\left[ik(\theta_i - \theta_j)\right] \quad \text{pour} \quad j < i \quad \text{et} \quad r \leqslant j < i \leqslant t$$

$$\ell_{ij}\left(\frac{\lambda_i}{\lambda_j}\right)^k \to 0 \quad (k \to \infty) \qquad \text{pour} \quad j < i \quad \text{dans les autres cas.}$$

En résumé, on peut écrire :

$$D^k L_y D^{-k} = \Lambda_k + E_k \quad \text{avec} \quad E_k = o(1) \quad \text{lorsque} \quad k \to \infty \quad \text{et}$$

$$\Lambda_k = \begin{matrix} \\ r \\ t \\ \\ \end{matrix} \begin{pmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & 0 \\ & & & \ddots & \\ & & 0 & & 1 \end{pmatrix} \begin{matrix} \\ \\ \\ r \\ t \end{matrix}$$

Les éléments du triangle étant de la forme $\ell_{ij} \exp\left[ik(\theta_i - \theta_j)\right]$.

Donc :

$$A^k = X(\Lambda_k + E_k)D^k U_y \,.$$

X matrice non singulière, est factorisable sous la forme $Q_x R_x$

$$A^k = Q_x R_x(\Lambda_k + E_k)D^k U_y$$

$$A^k = Q_x(R_x \Lambda_k + R_x E_k)D^k U_y$$

$R_x \Lambda_k$ est une matrice non singulière donc factorisable sous la forme Q.U.

Posons $\beta_k = R_x \Lambda_k$, c'est une matrice de la forme suivante :

$$\beta_k = \begin{pmatrix} \beta_{k_{11}} & \beta_{k_{12}} & \beta_{k_{13}} \\ 0 & \beta_{k_{22}} & \beta_{k_{23}} \\ 0 & 0 & \beta_{k_{33}} \end{pmatrix} \begin{matrix} r \\ t \end{matrix}$$

$$\begin{matrix} r & t \end{matrix}$$

Il est clair qu'elle peut se factoriser sous la forme suivante : $\beta_k = \rho_k \cdot \varepsilon_k$,

ρ_k matrice orthogonale, ε_k triangulaire supérieure +)

$$\rho_k = \begin{pmatrix} I_{r-1} & 0 & 0 \\ 0 & Q_{\rho_k} & 0 \\ 0 & 0 & I_{r-t} \end{pmatrix} \begin{matrix} r \\ t \end{matrix} \qquad \varepsilon_k = \begin{pmatrix} \varepsilon_{k_{11}} & \varepsilon_{k_{12}} & \varepsilon_{k_{13}} \\ 0 & \varepsilon_{k_{22}} & \varepsilon_{k_{23}} \\ 0 & 0 & \varepsilon_{k_{33}} \end{pmatrix}$$

$$\begin{matrix} r & t \end{matrix}$$

On voit (par identification) que :

$$\varepsilon_{k_{11}} = \beta_{k_{11}} \ , \ \varepsilon_{k_{13}} = \beta_{k_{13}} \ , \ \varepsilon_{k_{33}} = \beta_{k_{33}} \ ,$$

$Q_{\rho_k}\ \varepsilon_{k_{22}}$ doit-être une factorisation de $\beta_{k_{22}}$ et de la même façon

$\varepsilon_{k_{23}} = Q^T_{\rho_k}\ \beta_{k_{23}}$.

D'autre part $\rho_k\ \varepsilon_k + R_x\ E_k$ est factorisable Q.U. car pour tout k , $|R_x\ \Lambda_k|$ est bornée non singulière et lorsque k tend vers l'infini, E_k tend vers 0 .

Soit :

$$\rho_k\ \varepsilon_k + R_x\ E_k = \hat{Q}_k\ \hat{R}_k \ ,$$

ce qui donne :

$$I + \rho_k^T\ R_x\ E_k\ \varepsilon_k^{-1} = \rho_k^T\ \hat{Q}_k\ \hat{R}_k\ \varepsilon_k^{-1}$$

et entraîne avec la proposition de continuité des factorisations Q.U. :

$$\rho_k^T\ \hat{Q}_k \to I \quad \text{lorsque} \quad k \to \infty \ .$$

Soit :

$$\rho_k^T \hat{Q}_k = I + \mathcal{O}(1)$$

$$\hat{Q}_k = \rho_k + \mathcal{O}(1) ,$$

puisque ρ_k est bornée.

Finalement :

$$A^k = Q_x \hat{Q}_k \hat{R}_k D^k U_y .$$

On procède alors comme dans les démonstrations précédentes. Nous avons deux factorisations de la forme QU , donc il existe une matrice diagonale et orthogonale Δ_k telle que :

$$M_k = Q_x \hat{Q}_k \Delta_k^{-1} .$$

D'autre part :

$$A_{k+1} = M_k^{-1} A M_k$$

$$\Delta_k^{-1} A_{k+1} \Delta_k = (M_k \Delta_k)^{-1} A (M_k \Delta_k)$$

et donc :

$$\Delta_k^{-1} A_{k+1} \Delta_k = (Q_x \hat{Q}_k)^{-1} A (Q_x \hat{Q}_k) = (Q_x \rho_k)^{-1} A (Q_x \rho_k) + \mathcal{O}(1)$$

mais

$$A = X D Y = Q_x R_x D R_x^{-1} Q_x^T$$

ce qui entraîne

$$\Delta_k^{-1} A_{k+1} \Delta_k = \rho_k^T R_x D R_x^{-1} \rho_k + \mathcal{O}(1)$$

$R_x D R_x^{-1}$ est une matrice triangulaire supérieure ayant les valeurs propres de A sur la diagonale et ne dépendant pas de k . Il est facile de voir que la multiplication à gauche par ρ_k^T et à droite par ρ_k n'altère que les colonnes r à t laissant les autres invariantes.

Ainsi pour $k \to \infty$, certains blocs de $\Delta_k^{-1} A_{k+1} \Delta_k$ sont convergents et d'autres pas. Toutefois,

$$(\Delta_k^{-1} A_{k+1} \Delta_k)_{ii} = (A_{k+1})_{ii} \to \lambda_i$$

pour $k \to \infty$ et $\forall i \notin [r,t]$.

Conclusion : dans le cas considéré,

$$(A_{k+1})_{ii} \to \lambda_i \quad , \quad k \to \infty$$
$$\forall i \notin [r,t]$$

Remarque :

Il est évident que l'on ne saurait toujours avoir,

$$\text{diag } A_k \to [\lambda_1, \ldots, \lambda_n] \quad , \quad k \to \infty .$$

Voici deux contre exemples simples :

a) A est réelle, avec des valeurs propres imaginaires conjuguées. Comme toutes les matrices A_k sont réelles, il est exclu que $\text{diag } A_k$ converge vers $[\lambda_1, \ldots \lambda_n]$.

b) $A = Q$ = une matrice orthogonale. Dans ce cas l'algorithme QR "s'évanouit" puisque toutes les matrices A_k sont égales à A (nous avons $R_k = I \ \forall k$).

6°) **Algorithme LR - Cholesky.**

Description de l'algorithme.

Soit A une matrice symétrique définie positive. On a vu dans le début du cours que, dans ces conditions, il existe une matrice L triangulaire inférieure, d'éléments diagonaux $\ell_{ii} > 0$, telle que

$$A = LL^T .$$

Cette matrice L étant unique (factorisation de Choleski) (L n'appartient pas nécessairement à $\mathrm{Inf}(1)$).

Posons :

$$A_1 = A = L_1 L_1^T$$

$$A_2 = L_1^T L_1 = L_1^{-1} A L_1 = L_1^T A (L_1^T)^{-1}$$

A_2 est une matrice symétrique car : $A_2^T = (L_1^T L_1)^T = L_1^T (L_1^T)^T = A_2$.

A_2 est définie positive : en effet elle est non singulière : A_2 semi-définie positive :

$$(A_2 X, X) = (L_1^T L_1 X, X) = |L_1 X|^2 \geqslant 0$$

$\Longrightarrow (A_2 X, X) = 0 \Longleftrightarrow L_1 X = 0 \Longleftrightarrow X = 0$ car L_1 est non singulière donc on a $(A_2 X, X) > 0 \quad \forall X \neq 0$.

On peut donc lui appliquer la factorisation de Choleski

$$A_2 = L_2 L_2^T \text{ et l'on pose : } A_3 = L_2^T L_2 .$$

Plus généralement si on a déterminé A_k symétrique définie positive, on factorise A_k en :

$$A_k = L_k L_k^T$$

et on pose : $A_{k+1} = L_k^T L_k$ qui est symétrique définie positive (même démonstration que pour A_2).

D'où les formules suivantes :

$$A_{k+1} = L_k^{-1} L_k L_k^T L_k = L_k^{-1} A_k L_k \qquad (1)$$

de même :

$$A_{k+1} = L_k^T A_k (L_k^T)^{-1}$$

Posons $M_k = L_1 \ldots L_k$ alors,

$$A_{k+1} = M_k^{-1} A M_k \quad \text{(par itération de (1))}$$

de la même façon (ou par transposition)

$$A_{k+1} = A_{k+1}^T = M_k^T A (M_k^T)^{-1} .$$

On a :

$$M_k M_k^T = L_1 \ldots L_k . L_k^T \ldots L_1^T = M_{k-1} A_k M_{k-1}^T = A M_{k-1} M_{k-1}^T = A_k$$

(évidemment A^k est symétrique définie positive).

Au lieu comme précédemment de nous intéresser directement à la convergence de l'algorithme, nous allons le comparer à l'algorithme Q.R.

<u>Comparaison avec l'algorithme</u> Q.R.

Pour différentier facilement les deux algorithmes nous noterons A_k les matrices d'itération de L-R Choleski, $\tilde{A}_k$ celles de QR.

D'après les paragraphes précédents :

$$A^k = Q_1 \ldots Q_k R_k \ldots R_1 .$$

Calculons A^{2k} ; puisque A^k est symétrique c'est :

$$A^{2k} = (A^k)^T A^k = (R_k \ldots R_1)^T (Q_k^T \ldots Q_1^T)(Q_1 \ldots Q_k)(R_k \ldots R_1).$$

Puisque les matrices Q_i sont orthogonales, $Q_i^T Q_i = I$ et :

$$A^{2k} = (R_k \ldots . R_1)^T (R_k \ldots . R_1)$$

$(R_k \ldots . R_1) \in (sup,+)$ par construction et ainsi : $(R_k \ldots . R_1)^T \in (inf,+)$.

Mais, d'après LR-Cholesky :

$$A^{2k} = (L_1 \ldots . L_{2k})(L_1 \ldots . L_{2k})^T$$

avec $(L_1 \ldots . L_{2k}) \in (inf,+)$, $(L_1 \ldots . L_{2k})^T \in (sup,+)$.

Ainsi, puisque dans ces conditions la factorisation de Choleski de A^{2k} est unique :

$$L_1 \ldots . L_{2k} = R_1^T \ldots . R_k^T .$$

D'autre part, pour l'algorithme LR-Choleski :

$$A_{2k+1} = L_{2k}^T L_{2k} = (L_1 \ldots . L_{2k})^{-1} A (L_1 \ldots . L_{2k})$$

$$A_{2k+1} = (L_1 \ldots . L_{2k})^T A ((L_1 \ldots . L_{2k})^T)^{-1}$$

mais d'après ce qui précède, ceci est égal à :

$$A_{2k+1} = (R_k \ldots . R_1) A (R_k \ldots . R_1)^{-1} = \tilde{A}_{k+1} .$$

La $(2k+1)$-ième étape de l'algorithme LR-Choleski est donc égale à la $(k+1)$-ième étape de l'algorithme Q.R.

Ainsi l'algorithme LR-Choleski converge toujours, ce qui est une amélioration sensible sur l'algorithme L.R. pour des matrices non symétriques définies positives.

7°) Techniques diverses.

a) Amélioration de la convergence par translation des valeurs propres.

Nous avons vu précédemment que dans l'algorithme Q.R. la vitesse de la convergence de la matrice E_k vers 0 dépend de $(\frac{\lambda_i}{\lambda_j})^k$, $\lambda_i < \lambda_j$, mais si les valeurs propres sont peu séparées, la vitesse de convergence risque d'être très lente.

Nous allons essayer d'améliorer ceci à l'aide de la remarque suivante :

- Si une matrice A a pour valeurs propres $\lambda_1,\ldots,\lambda_n$, la matrice $A-\alpha I$ a pour valeurs propres :

$$\lambda_1-\alpha\ ,\ \lambda_2-\alpha\ ,\ldots,\ \lambda_n-\alpha\ .$$

L'algorithme Q.R. est alors modifié ainsi : à la k-ième étape on factorise

$A_k-\alpha_k I$ $(\alpha_k \neq \lambda_i\ \forall i$ et convenable)

$$A_k-\alpha_k I = Q_k R_k \quad \text{et l'on pose} \quad A_{k+1} = R_k Q_k + \alpha_k I\ .$$

Etablissons quelques formules :

- $A_{k+1} = Q_k^T Q_k(R_k Q_k + \alpha_k I) = Q_k^T(A_k-\alpha_k I)Q_k + \alpha_k I = Q_k^T A_k Q_k$
- $A_{k+1} = M_k^T A M_k$ d'où $M_k A_{k+1} = A M_k$
- $M_k S_k = M_{k-1}(A_k-\alpha_k I)S_{k-1} = (M_{k-1} A_k - \alpha_k M_{k-1})S_{k-1} = (A-\alpha_k I)M_{k-1} S_{k-1}$

donc :

$$M_k S_k = (A-\alpha_k I)\ldots.(A-\alpha_1 I).$$

Supposons A diagonalisable : $A = XDY$ $(Y = X^{-1})$.

Alors : $A-\alpha I = X(D-\alpha I)Y$.

Ainsi : $M_k S_k = X(D-\alpha_k I)\ldots.(D-\alpha_1 I)\ Y = X\ D_k\ Y$,

avec $D_k = (D-\alpha_k I)\ldots.(D-\alpha_1 I)$.

Supposons de plus que Y soit factorisable LU : $Y = L_y\ U_y$

$$M_k\ S_k = X\ D_k\ L_y\ U_y = X(D_k\ L_y\ D_k^{-1})D_k\ U_y$$

$$(D_k\ L_y\ D_k^{-1})_{ij} = \begin{cases} 0 & \text{si } i < j \\ 1 & \text{si } i = j \\ \ell_{ij}\ \dfrac{(\lambda_i-\alpha_k)\ldots.(\lambda_i-\alpha_1)}{(\lambda_j-\alpha_k)\ldots.(\lambda_j-\alpha_1)} \to 0 & \text{si } i > j \end{cases}$$

Nous avons donc intérêt à prendre α_k voisin de la valeur propre la plus petite en module.

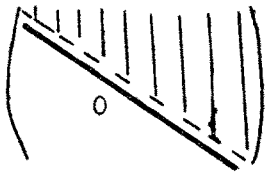
On améliorera ainsi considérablement la vitesse de la convergence.

Il est facile de voir en appliquant les mêmes procédés que dans les paragraphes précédents que la suite de matrices $|A_k|$ converge vers une matrice $|R_x\ D\ R_x^{-1}|$ qui a sur la diagonale les valeurs propres de A rangées dans le bon ordre.

Les mêmes raisonnements peuvent s'adapter aux différents cas envisagés précédemment.

b) Réduction à la forme de Hessenberg.

On cherche une matrice semblable à A de la forme :



matrice triangulaire supérieure plus une sous-diagonale

dite forme de Hessenberg.

Cette forme est intéressante car elle comporte beaucoup de zéros et d'autre part elle est invariante pour l'algorithme Q.R. ce qui diminue considérablement le

nombre d'opérations à effectuer.

La méthode suivante de réduction à la forme de Hessenberg est une adaptation de la méthode de Householder (cf. Chap.I). Définissons une suite de matrices orthogonales (du type Householder) H_k et une suite A_k

$$A_1 = A \ , \ A_2 = H_1 A_1 H_1 \ , \ldots \text{ etc.}$$

On voudrait que A_n soit sous forme de Hessenberg et semblable à A .

Raisonnons par récurrence : à la p-ième étape nous avons :

$$A_p = \left(\begin{array}{c|c} A_{11} & A_{12} \\ \hline \begin{array}{c} X \\ \vdots \\ A_{21} \\ X \end{array} & A_{22} \end{array}\right) \quad \begin{matrix} p \\ n-p \end{matrix}$$

et l'on voudrait $A_{p+1} = \left(\begin{array}{c|c} A'_{11} & A'_{12} \\ \hline X & \\ A'_{21} & A'_{22} \end{array}\right)$ $\begin{matrix} p+1 \\ n-p-1 \end{matrix}$

A_{11} est sous la forme de Hessenberg ainsi que A'_{11}

A_{21} est une matrice dont toutes les colonnes sont nulles sauf la dernière.

A'_{21} est une matrice dont tous les termes sont nuls sauf peut-être celui en haut et à droite.

Soit α la p-ième colonne de A_{21} ; d'après un lemme (page chap.I) il existe $u \in \mathbb{R}^{n-p}$ et $k \in \mathbb{R}$ tels que

$$\tilde{H}_p \alpha = k \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \Big\} n-p \quad \text{avec} \quad \tilde{H}_p = I_{n-p} - 2uu^T$$

$\tilde{H}_p$ est symétrique et orthogonale.

Posons :

$$H_p = \left(\begin{array}{c|c} I_p & 0 \\ \hline 0 & \tilde{H}_p \end{array}\right)$$

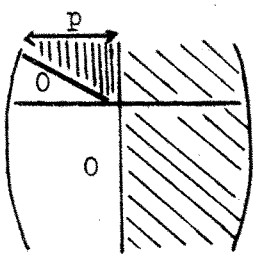
et effectuons le produit $H_p A_p H_p$

$$A_p H_p = \left(\begin{array}{c|c} A_{11} & A_{12}\tilde{H}_p \\ \hline A_{12} & A_{22}\tilde{H}_p \end{array}\right) \quad , \quad H_p A_p H_p = \left(\begin{array}{c|c} A_{11} & A_{12}\tilde{H}_p \\ \hline \tilde{H}_p A_{12} & \tilde{H}_p A_{22}\tilde{H}_p \end{array}\right)$$

On pose donc :

$$A_{p+1} = H_p A_p H_p \quad \text{matrice semblable à} \quad A_p$$

$A_{p+1} =$



Ainsi A_n sera de la forme annoncée.

8°) Conditionnement du problème des valeurs propres.

Supposons A diagonalisable (i.e. $A = X D X^{-1}$) et considérons une variation ΔA de A.

Soit λ une valeur propre de $A + \Delta A$

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

$$X^{-1}(A + \Delta A - \lambda I)X = (D-\lambda I) + X^{-1} \Delta A X .$$

Supposons que λ ne soit pas valeur propre de A ($\lambda \neq \lambda_i \ \forall i , 1 \leqslant i \leqslant n$).

Alors : $\det(D-\lambda I) \neq 0$ et $D-\lambda I$ est inversible et l'on peut écrire

$$X^{-1}(A + \Delta A - \lambda I)X = (D-\lambda I)[I + (D-\lambda I)^{-1} X^{-1} \Delta A X] .$$

Puisque λ est valeur propre de $A + \Delta A$, $A + \Delta A - \lambda I$ est une matrice singulière ; comme d'autre part $(D-\lambda I)$ est non singulière, $I + (D-\lambda I)^{-1} X^{-1} \Delta A X$

doit-être singulière. Or nous connaissons un lemme qui nous dit que $I+B$ est inversible lorsque $\|B\| < 1$.

Nous avons donc nécessairement :

$$\|(D-\lambda I)^{-1}\ X^{-1}\ \Delta A\ X\| \geqslant 1 \ .$$

Soit :

$$1 \leqslant \|X\|\ \|X^{-1}\|\ \|\Delta A\|\ \|(D-\lambda I)^{-1}\| .$$

Supposons que la norme matricielle $\|.\|$ soit telle que :

$$\|D-\lambda I\| = \max_i\ |\lambda_i-\lambda| \ .$$

Alors :

$$1 \leqslant \|X\|\ \|X^{-1}\|\ \|\Delta A\|\ \max_i\ |\lambda_i-\lambda|^{-1} \ .$$

Soit

$$\min_i\ |\lambda_i-\lambda| \leqslant \|X\|\ \|X^{-1}\|\ \|\Delta A\| .$$

L'ensemble des valeurs propres étant fini il existe i tel que

$$|\lambda_i-\lambda| \leqslant \|X\|\ \|X^{-1}\|\ \|\Delta A\| .$$

$\|X\|\ \|X^{-1}\|$ est ce que nous avons appelé conditionnement de X (cond (X)), nous le noterons aussi $\chi_X(A)$.

On voit que cette majoration de l'erreur dépend de la matrice X dont on se sert pour diagonaliser A .

Définissons aussi :

$$\chi(A) = \inf_{X \text{ diag } A}\ \text{cond}(X)$$

la borne inférieure est prise sur l'ensemble des matrices X qui diagonalisent A.

On appelle $\chi(A)$ le conditionnement de A pour le problème des valeurs propres.

On a donc le résultat suivant :

Si λ est valeur propre de $A + \Delta A$, il existe λ_i valeur propre de A telle que

$$|\lambda_i - \lambda| \leqslant \chi(A) \, \|\Delta A\| .$$

Propriétés.

* $\chi(A) \geqslant 1$.

* L'égalité peut-être atteinte : en effet, si A est hermitienne, symétrique réelle, ou orthogonale $\chi(A) = 1$; en effet le théorème de Schur nous dit qu'il existe une matrice X orthogonale telle que $A' = X^T A X$ soit triangulaire supérieure mais :

$$A'^T (X^T A X)^T = X^T (X^T A)^T = X^T A^T X = X^T A X = A'$$

donc A' est symétrique et donc diagonale.

* Invariance de $\chi(A)$.

Si $B = Q A Q^T$ avec Q orthogonale et A diagonalisable $A = X D X^{-1}$

$$\chi(B) = \chi(A).$$

On peut écrire $B = Q X D X^{-1} Q^T = (QX) D (QX)^{-1}$.

Donc

$$X \text{ diag } A \Longleftrightarrow QX \text{ diag } B$$

$$\chi(A) = \inf_{X \text{ diag } A} \text{cond}(X) = \inf_{Y \text{ diag } B} \text{cond}(Q^T Y) \leqslant \chi(B)$$

puisque $\|Q^T\| = 1$

$$\chi(B) = \inf_{Y \text{ diag } B} \text{cond}(Y) = \inf_{X \text{ diag } A} \text{cond}(QX) \leqslant \chi(A)$$

Donc : $\chi(B) = \chi(A)$.

En particulier : la réduction à la forme de Hessenberg ne change pas le conditionnement pour les valeurs propres.

Stabilité des valeurs propres.

Soit A une matrice $n \times n$ ayant n valeurs propres distinctes en module :

(0) $\quad |\lambda_1| > |\lambda_2| > \ldots > |\lambda_n| .$

On étudie la stabilité des valeurs propres et des vecteurs propres lorsque A varie en fonction d'un paramètre $\varepsilon \in [0, \varepsilon_0[$. Si X_i est un vecteur propre associé à $\lambda_i = \lambda_i(\varepsilon)$, on a :

(1) $\quad A(\varepsilon).X_i(\varepsilon) = \lambda_i(\varepsilon).X_i(\varepsilon) .$

On suppose que A , X_i et λ_i sont des fonctions différentiables de ε (*)

Différentiant alors la formule (1), on obtient :

(2) $\quad dA.X_i + A.dX_i = X_i\, d\lambda_i + \lambda_i.dX_i .$

Considérons la matrice A^T . Ses valeurs propres sont les valeurs propres de A . Soit Y_i un vecteur propre de A^T associé à λ_i : $A^T Y_i = \lambda_i Y_i$.

On sait que Y_i est orthogonal à X_j si $\lambda_i \neq \lambda_j$ c'est-à-dire ici si $i \neq j$.

En effet on a : $(AX_j, Y_i) = (\lambda_j X_j, Y_i) = \lambda_j(X_j, Y_i) = (X_j, A^T Y_i) = \lambda_i(X_j, Y_i).$

D'où $(\lambda_j - \lambda_i)(X_j, Y_i) = 0$.

Donc $(X_j, Y_i) = 0$ si $i \neq j$, c'est-à-dire $X_j \perp Y_i$ si $i \neq j$.

Multiplions scalairement les deux membres de l'égalité (2) par Y_i :

$$(dA.X_i, Y_i) + (A.dX_i, Y_i) = d\lambda_i(X_i, Y_i) + \lambda_i(dX_i, Y_i).$$

(*) On peut montrer, grâce à (0) que les $\lambda_i(\varepsilon)$ sont tous distincts en module, pour ε assez petit ; et en outre il est possible de numéroter les $\lambda_i(\varepsilon)$ en sorte que : $\lambda_i(\varepsilon) \to \lambda_i(0)$, $\varepsilon \to 0$, $\forall i$, $1 \leqslant i \leqslant n$.

Or $(\mathbf{A}\, dX_i, Y_i) = (dX_i, \mathbf{A}^T Y_i) = (dX_i, \lambda_i Y_i) = \lambda_i (dX_i, Y_i)$.

D'où : $(d\mathbf{A}.X_i, Y_i) = d\lambda_i (X_i, Y_i)$.

On peut donc écrire :

$$d\lambda_i = \frac{(d\mathbf{A}.X_i, Y_i)}{(X_i, Y_i)} .$$

Ce qui entraîne : $\left|\frac{d\lambda_i}{d\varepsilon}\right| \leqslant \left\|\frac{d\mathbf{A}}{d\varepsilon}\right\| . \frac{|X_i| . |Y_i|}{|(X_i, Y_i)|}$.

Ou encore : $\boxed{\left|\frac{d\lambda_i}{d\varepsilon}\right| \leqslant s_i \left\|\frac{d\mathbf{A}}{d\varepsilon}\right\|}$

avec : $\frac{1}{s_i} = \frac{|(X_i, Y_i)|}{|X_i| . |Y_i|} = |\cos\beta_i|$;

le nombre s_i mesure en quelque sorte la distorsion entre X_i et Y_i .

On déduit de cette inégalité que le calcul de la $i^{ème}$ valeur propre est stable si la distorsion entre X_i et Y_n n'est pas trop grande et en particulier si X_i est parallèle à Y_i . Cela se produit à coup sûr pour tout i si :

- $\mathbf{A}$ est symétrique, $\mathbf{A} = \mathbf{A}^T$.
- $\mathbf{A}$ est orthogonale, $\mathbf{A}^T\mathbf{A} = I$.

9°) Conditionnement des valeurs propres.

Remarque préliminaire : Les valeurs propres étant distinctes, les $X_i(\varepsilon)$ forment, pour tout ε , une base de R^n . Il est donc possible d'écrire :

$$\frac{dX_i}{d\varepsilon} = \sum_{j=1}^{n} \alpha_{ij}(\varepsilon)\, X_j(\varepsilon).$$

Nous allons voir que l'on peut normer les $X_i(\varepsilon)$ en sorte que $\alpha_{ii}(\varepsilon) \equiv 0$, $\forall i$, et c'est la stabilité des $X_i(\varepsilon)$ ainsi normés que l'on va étudier (*).

Le fait que l'on puisse normer les X_i en sorte que $\alpha_{ii} \equiv 0$ se vérifie aisément : soient $\tilde{X}_1(\varepsilon), \ldots, \tilde{X}_n(\varepsilon)$ les vecteurs propres normés arbitrairement et

(*) Il est évident qu'une normalisation fantaisiste des $X_i(\varepsilon)$ entrainerait des instabilités artificielles pour ces vecteurs...

soit $X_i(\varepsilon) = \sigma_i(\varepsilon)\,\tilde{X}_i(\varepsilon)$, la fonction scalaire $\sigma_i(\varepsilon)$ étant à préciser.

On a, les $\tilde{X}_j$ formant une base de $\mathbb{R}^n$:

$$\frac{d\tilde{X}_i(\varepsilon)}{d\varepsilon} = \sum_{j=1}^{n} \tilde{\alpha}_{ij}(\varepsilon)\,\tilde{X}_j(\varepsilon) .$$

D'où : $\dfrac{dX_i}{d\varepsilon} = \dfrac{d\sigma_i(\varepsilon)}{d\varepsilon}\,.\tilde{X}_i(\varepsilon) + \sigma_i(\varepsilon) \sum_{j=1}^{n} \tilde{\alpha}_{ij}(\varepsilon)\,\tilde{X}_j(\varepsilon).$

Cela s'écrit : $\dfrac{dX_i}{d\varepsilon} = \sum_{j=1}^{n} \alpha_{ij}(\varepsilon)\,X_j(\varepsilon)$ avec en particulier :

$$\alpha_{ii} = \frac{1}{\sigma_i}\left(\frac{d\sigma_i}{d\varepsilon} + \sigma_i\,\tilde{\alpha}_{ii}\right) .$$

La fonction $\tilde{\alpha}_{ii}(\varepsilon)$ étant connue, les choix de σ_i rendant $\alpha_{ii}(\varepsilon)$ identiquement nulle sont évidents :

$$\sigma_i(\varepsilon) = c\,\exp\Big(-\int_0^{\varepsilon} \tilde{\alpha}_{ii}(\bar{\varepsilon})\,d\bar{\varepsilon}\Big).$$

Retournons à présent à la formule (2) :

$$dA.X_i + A.dX_i = X_i\,d\lambda_i + \lambda_i.dX_i .$$

Multiplions scalairement par Y_j :

$$(dA.X_i, Y_j) + (A\,dX_i, Y_j) = (X_i\,d\lambda_i, Y_j) + (\lambda_i\,dX_i, Y_j).$$

Or $(X_i.d\lambda_i, Y_j) = 0$ si $i \neq j$, d'où dans ce cas :

$$(dA.X_i, Y_j) + (A.dX_i, Y_j) = (\lambda_i\,dX_i, Y_j).$$

Soit encore : $(dA.X_i, Y_j) = (\lambda_i - \lambda_j)(dX_i, Y_j).$

Comme on a : $dX_i = \sum_{k=1}^{n} \alpha_{ik}\,X_k\,d\varepsilon$, $\alpha_{ii} = 0$, alors

$$(dX_i, Y_j) = \sum_{k=1}^{n} \alpha_{ik}(X_i, Y_j)d\varepsilon = \alpha_{ij}(X_j, Y_j)d\varepsilon$$

car $(X_k, X_j) = 0$ si $k \neq j$.

Par conséquent : $(dA.X_i, Y_j) = (\lambda_i - \lambda_j)\alpha_{ij}(X_j, Y_j)d\varepsilon$.

On obtient donc si $i \neq j$:

$$\alpha_{ij} = \frac{(dA/d\varepsilon \cdot X_i, Y_j)}{(\lambda_i - \lambda_j)(X_j, Y_j)} .$$

D'où : $$\frac{dX_i}{d\varepsilon} = \sum_{\substack{j=1 \\ j\neq i}}^{n} \alpha_{ij} X_j = \sum_{\substack{j=1 \\ j\neq i}}^{n} \frac{(\frac{dA}{d\varepsilon} X_i, Y_j)}{(\lambda_i - \lambda_j)(X_j, Y_j)} \cdot X_j$$

et en passant aux normes :

$$\left|\frac{dX_i}{d\varepsilon}\right| \leqslant \sum_{\substack{j=1 \\ j\neq i}}^{n} \frac{\left\|\frac{dA}{d\varepsilon}\right\| \cdot |X_i| \cdot |Y_j| \cdot |X_j|}{|\lambda_i - \lambda_j| |(X_i, Y_j)|}$$

soit encore :

$$\boxed{\left|\frac{dX_i}{d\varepsilon}\right| \leqslant |X_i| \cdot \left\|\frac{dA}{d\varepsilon}\right\| \left(\sum_{\substack{j=1 \\ j\neq i}}^{n} \frac{s_j}{|\lambda_i - \lambda_j|} \right)}$$

On remarque que cette inégalité montre qu'un vecteur propre peut-être mal conditionné si un seul des s_j est grand, contrairement au conditionnement des valeurs propres où il suffit que le s_i correspondant à λ_i ne soit pas grand

Relation entre les s_i et χ (nbre de conditionnement pour le pb. des v.p.)

. Soit A une matrice $n \times n$ ayant n valeurs propres distinctes. Soit S un matrice qui diagonalise A :

$$A = S D S^{-1} .$$

Soit $(e_1, \ldots, e_n)$ la base canonique de $\mathbb{R}^n$.

. Se_i est un vecteur propre de A pour la valeur propre λ_i car :

$$A S e_i = S D e_i = \lambda_i S e_i .$$

Posons donc $X_i = S e_i$.

On vérifie de même que $(S^{-1})^T e_i$ est un vecteur propre de A^T associé à la valeur propre λ_i et on pose :

$$Y_i = (S^{-1})^T e_i .$$

On a alors :

$$\frac{1}{s_i} = \frac{|(X_i, Y_i)|}{|X_i| \cdot |Y_i|} = \frac{|(Se_i, (S^{-1})^T e_i)|}{|(Se_i| \cdot |(S^{-1})^T e_i|} = \frac{1}{|Se_i| \cdot |(S^{-1})^T e_i|} .$$

D'autre part :

$$|Se_i| \leqslant \|S\| \cdot |e_i| = \|S\|$$

$$|(S^{-1})^T e_i| \leqslant \|(S^{-1})^T\| \cdot |e_i| = \|(S^{-1})^T\| = \|S\| .$$

D'où :

$$s_i = |Se_i| \cdot |(S^{-1})^T e_i| \leqslant \|S\| \cdot \|S^{-1}\| = \chi_S(A).$$

Donc : $s_i \leqslant \chi_S(A)$ pour toute S qui diagonalise A, et par conséquent :

$$s_i \leqslant \chi(A) .$$

Montrons par ailleurs que $\chi(A) \leqslant \sum_{j=1}^{n} s_j$.

Soient $X_1, X_2, \ldots, X_n$ les vecteurs propres normés de A et $Y_1, \ldots, Y_n$ les vecteurs propres normés de A^T. S étant une matrice de passage à une base qui diagonalise A, les colonnes de S sont des vecteurs propres de A.

On peut prendre $S = [X_1\sqrt{s_1}, \ldots, X_n\sqrt{s_n}]$. On vérifie alors que :

$$S^{-1} = \begin{bmatrix} Y_1^T\sqrt{s_1} \\ \vdots \\ Y_n^T\sqrt{s_n} \end{bmatrix} .$$

En effet en effectuant $S^{-1}S$:

$$(S^{-1}S)_{ij} = Y_i^T\sqrt{s_i} \cdot X_j\sqrt{s_j} = \begin{cases} 0 & \text{si } i \neq j \\ 1 & \text{si } i = j \end{cases}$$

Avec ce choix de S, on a par définition :

$$\chi_S(A) = \|S\| \, \|S^{-1}\| \quad \text{où} \quad \|S\| = \sup_{x \neq 0} \frac{|Sx|}{|x|} .$$

Or $(Sx)_i = \sum_{j=1}^{n} s_{ij} x_j$.

Notant $S_{i.}$ le ie vecteur ligne de S , l'inégalité de Schwarz donne :

$$(Sx)_i^2 \leqslant |S_{i.}|^2 \, |x|^2$$

et
$$|Sx|^2 \leqslant \sum_{i=1}^{n} |S_{i.}|^2 \, |x|^2 \, .$$

D'où :

$$\|S\|^2 \leqslant \sum_{i=1}^{n} |S_{i.}|^2$$

et :

$$\|S\|^2 \leqslant \sum_{i=1}^{n} S_i \, |X_i|^2 = \sum_{i=1}^{n} S_i$$

de même
$$\|S\|^2 \leqslant \sum_{i=1}^{n} S_i \, |Y_i|^2 = \sum_{i=1}^{n} S_i \, .$$

D'où :

$$\chi(A) \leqslant \chi_S(A) \leqslant \sum_{i=1}^{n} S_i \, .$$

On a donc pour tout j , $1 \leqslant j \leqslant n$:

$$\boxed{S_j \leqslant \chi(A) \leqslant \sum_{i=1}^{n} S_i}$$