Compléments de cours en analyse numérique matricielle

M. Granger

1 factorisation QR et méthode de Householder

1.1

Théorème 1.1 *Pout toute matrice $A \in \mathcal{M}_{m,n}(\mathbb{C})$ avec $m \geq n$ il existe $Q \in U(m)$, unitaire et R triangulaire supérieure telle que*

$$A = QR.$$

Si on ajoute la condition $R_{i,i} > 0$, la factorisation est unique.

N.B. Dire que R est une matrice triangulaire supérieure signifie que ses n premières lignes constituent une telle matrice, les $m - n$ suivantes étant nulles.

On considère pour cela les matrices de Householder. La matrice de Householder H_u, associée à un vecteur unité u, c'est à dire tel que $< u, u >= u^\star.u = 1$ est définie par:

$$H = H_u = I - 2uu^\star$$

Lemme 1.2 *La matrice H_u est hermitienne et unitaire.*

Démonstration Comme $H_u^\star = I - 2(u^\star)^\star u^\star = I - 2uu^\star = H_u$, la matrice H_u est hermitienne. Comme $H_u^\star H_u = H_u^2 = (I - 2uu^\star)(I - 2uu^\star) = I - 4uu^\star + 4u(u^\star u)u^\star = I$, en appliquant $u^\star u = 1$, on voit qu'elle est aussi unitaire. □

On peut voir aussi ce dernier résultat géométriquement de la façon suivante:

$$H_u(u) = u - 2uu^\star u = u - 2u = -u, \quad \text{et si } < v, u >= u^\star v = 0,\ H_u(v) = v - 2uu^\star v = v$$

Ces deux calculs montrent que H_u est la symétrie orthogonale par rapport à l'hyperplan $u^\tau = \{v \in \mathbb{C}^n < v, u >= 0\}$. Une autre façon de faire cette description est de remarquer que $uu^\star = 1/2(I - H_u)$ est la projection orthogonale sur la droite $\mathbb{C}.u$.

Noter au passage l'expression utile de H_u en terme de produit scalaire hermitien:

$$H_u(v) = v - 2 < v, u > u$$

Si $u \in \mathbb{R}^n$, le même calcul montre que la matrice $H_u = I - 2u^t u$ est symétrique et orthogonale. Dans ce qui suit on se contentera pour simplifier de démontrer le théorème 1.1 dans le cas où $A \in \mathcal{M}_{n,n}(\mathbb{R})$, et on trouvera alors pour Q une matrice réelle, donc orthogonale, produit de symétries orthogonales associées à des vecteurs u réels.

Commençons par établir dans ce contexte un résultat préliminaire:

Lemme 1.3 *Soient $a, b \in \mathbb{R}^n$ deux vecteurs non colinéaires avec b unitaire. Alors, il existe u unitaire tel que $\exists \alpha \in \mathbb{R},\ H_u(a) = \alpha b$.*

dessin

Démonstration *du lemme .-* Comme $\|H_u(a)\| = \|a\|$, on obtient en passant aux normes dans l'égalité cherchée, la condition $|\alpha| = \|a\|$, d'où deux possibilités, $\alpha = \epsilon\|a\|$ avec $\epsilon = 1,$ ou -1. L'équation à résoudre s'écrit.

$$a - 2 < u, a > u = \epsilon\|a\|b$$

Le réel $\lambda =< u, a >$ est alors déterminé par: $\lambda^2 = 1/2(\|a\|^2 - \epsilon\|a\| < a, b >)$ D'où on tire deux solutions $\lambda_1 = +\sqrt{1/2(\|a\|^2 - \epsilon\|a\| < a, b >)}$ et $\lambda_2 = -\sqrt{1/2(\|a\|^2 - \epsilon\|a\| < a, b >)}$, et les deux vecteurs u unitaires opposés correspondant sont:

$$u_1 \frac{a - \epsilon\|a\|b}{\lambda_1} \text{ et } u_2 = -u_1$$

Noter que ce calcul a bien un sens car puisque b est unitaire et a non lié à b, on a $| < a, b > | < \|a\|\|b\| = \|a\|$, donc $\|a\|^2 - \epsilon\|a\| < a, b >> 0$ et donc aussi $\lambda > 0$. □

Pour le calcul pratique et afin d'éviter les extractions de racines carrées, on écrit: $v = 2 < u, a > u = a - \alpha b$, avec $\alpha = \epsilon\|a\|$. Les deux vecteurs $v = v_\epsilon$ possibles sont représentés sur le dessin ci-dessous. On trouve alors

$$H_u = I - \frac{1}{\beta} v^t v, \text{ avec } \beta = 2\lambda^2 = \|a\|^2 - \epsilon\|a\| < a, b >=< a, v > .$$

Pour que la précision du calcul de H dans la formule:

$$H = I - \frac{1}{\beta} v^t v = I - \frac{1}{< a, v >} v^t v$$

soit la meilleure possible, il convient de prendre à la fois $\|v\|$, et β, les plus grands ce qui consiste à choisir celui des vecteurs v_+ ou v_- qui est porté par la bissectrice de l'angle aigu des vecteurs a et b.

Démonstration *du théorème 1.1 dans le cas réel.-* Existence de la factorisation. Ecrivons $A = [a_1, \cdots, a_n]$, comme la suite de ses n colonnes de longueur m. Si $a_1 \notin \mathbb{R}.e_1$, on peut trouver u unitaire tel que $H_u(a_1) = \|a_1\|e_1$, d'où une équation de la forme.

$$H_u.A = [cste.e_1, a'_2, \cdots, a'_n] = \begin{pmatrix} \star & a'_{1,2} & \cdots & a'_{1,n} \\ 0 & & \cdots & \\ \vdots & & a'_{i,j} & \\ \vdots & & & \\ 0 & a'_{m,2} & \cdots & a'_{m,n} \end{pmatrix}$$

et en itérant le procédé on trouve

$$H_{u_m}.\cdots.H_{u_1}A = \begin{pmatrix} \star & \cdots & \cdots & \star \\ 0 & \star & \cdots & \star \\ \vdots & & \ddots & \\ 0 & & 0 & \star \\ \vdots & 0 & & 0 \\ 0 & \cdots & & 0 \end{pmatrix}$$

ce qui donne la formule demandée en choisissant simplement:

$$Q = (H_{u_m}.\cdots.H_{u_1})^{-1} (= H_{u_1}.\cdots.H_{u_m})$$

Pour l'unicité, remarquons d'abord que chacune des conditions $(\forall i,\ R_{i,i} \neq 0)$ et $(\forall i,\ R'_{i,i} \neq 0)$, équivaut au fait que A est de rang maximum n. On peut en plus, en mulitipliant Q à droite par les H_{e_i} tels que $R_{i,i} < 0$, trouver une décomposition où la condition de positivité sur la diagonale de R est satisfaite. Supposons alors données deux décompositions $A = QR = Q'R'$. On montre alors que $S = Q'^{-1}Q$, est une matrice à la fois orthogonale et triangulaire. Si $m = n$, cela se déduit simplement de l'égalité $S = R'R^{-1}$ entre matrices carrées $n \times n$. En général soit E_i le sous espace de $\mathbb{R}^n$ engendré par les i premiers vecteurs de la base canonique $\{e_1, \cdots, e_i\}$. L'équation $SR = R'$ donne pour tout i une formule du type $S(R_{i,i}e_i + v_i) = R'_{i,i}e_i + v'_i$ avec $v_i, v'_i \in E_{i-1}$, d'où on tire par récurrence $S(E_i) \subset E_i$ c'est à dire la forme triangulaire supérieure de S.

Une telle matrice à la fois orthogonale et triangulaire est du type

$$S = \begin{pmatrix} \epsilon_1 & 0 & \cdots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & \ddots & \\ 0 & & & \epsilon_n \end{pmatrix}$$

avec $\epsilon_j \in \{-1, +1\}$. En effet $S^{-1} = {}^tS$ est à la fois triangulaire supérieure (comme S^{-1}) et inférieure (comme ${}^tS$), donc diagonale, forcément de la forme indiquée comme matrice orthogonale. Il reste à remarquer que, par un calcul immédiat, on trouve $\epsilon_i = R_{i,i}^{-1}R'_{i,i}$ pour avoir le résultat d'unicité souhaité ($S = I$ donc $Q = Q'$.)

□

1.2 Application aux problèmes de moindres carrés

Dans le domaine des sciences expérimentales on est souvent confronté au problème suivant:

On considère $n + 1$ variables $y, x_1, \cdots, x_n$, supposées liées par une relation linéaire:

$$y = b_0 + b_1x_1 + \cdots + b_nx_n$$

où les b_i sont des coefficients supposés inconnus. On cherche à les déterminer, à l'aide de mesures, en nombre m en général très supérieur à $n+1$. Les résultats de ces mesures sont notés:

$$\overline{y_i}, \overline{x_{i,1}}, \cdots, \overline{x_{i,n}} \text{ pour } i = 1, \ldots, m$$

On note aussi $\overline{Y}, resp. \overline{X_1}, \cdots, resp. \overline{X_n} \in \mathbb{R}^m$ les vecteurs colonnes de coordonnées respectives $\overline{y_i}, resp. \overline{x_{i,1}}, \cdots, resp. \overline{x_{i,n}}$

Si les résultats des mesures étaient des valeurs "exactes", le vecteur $\overline{Y} - \sum b_i \overline{X_i}$, serait nul. Dans la pratique, il s'agira de trouver des b_i, qui rendent la norme de ce vecteur minimale, le minimum étant d'autant plus petit que le modèle linéaire est adapté et les mesures précises. Comme $\overline{Y} - \sum b_i \overline{X_i} = \overline{Y} - \overline{X}B$, où

$$\overline{X} = \begin{pmatrix} 1 & \overline{x_{1,1}} & \cdots & \overline{x_{1,n}} \\ \vdots & & & \vdots \\ \vdots & & & \vdots \\ 1 & \overline{x_{1,m}} & \cdots & \overline{x_{n,m}} \end{pmatrix}$$

et B est le vecteur colonne $\begin{pmatrix} 1 \\ b_1 \\ \vdots \\ b_n \end{pmatrix}$

On est ainsi confronté au problème suivant, dans des notations simplifiées:

Problème Etant donnés $A \in \mathcal{M}_{m,n}(\mathbb{R})$ avec $m > n$ et $b \in \mathbb{R}^m$, trouver un vecteur $x \in \mathbb{R}^n$, qui premette de réaliser le minimum de $\|Ax - b\|$. Dans ce qui suit on suppose que $\| \bullet \| = \| \bullet \|_2$, est la norme euclidienne sur $\mathbb{R}^m$, et on va montrer que dans ce cas:
- Le problème se ramène a la résolution d'un système linéaire.
- La solution est unique si A est de rang maximum n.
- La décomposition de Householder de A est adaptée à la solution du problème.

Théorème 1.4 *Une condition nécessaire et suffisante pour que $E(x) = \|Ax - b\|^2$ atteigne son minimum en x est que x soit solution du système linéaire suivant:*

$$^tAAx = {}^tAb.$$

La solution est unique si et seulement si A est de rang n.

Démonstration.- Considérons la projection orthogonale de b sur l'image de A, notée $b_1 = Ax_1$, et $b_2 = b - b_1$. L'orthogonalité de b_2, à l'image de A, est caractérisée par la condition:

$$\forall y \in \mathbb{R}^m < Ay, b_2 >=< y, {}^tAb_2 >$$

donc finalement en raison de la non-dégénerescence du produit scalaire par ${}^tAb_2 = 0$.

On trouve:

$$E(x) =< Ax - b, Ax - b >= \|A(x - x_1)\|^2 + \|b_2\|^2$$

Le minimum de E est donc atteint sur tout vecteur tel que $Ax = b_1$. Comme $Ay = 0 \Leftrightarrow {}^tAAy = 0$, en raison de l'égalité ${}^ty\,{}^tAAy = \|Ay\|^2$, la condition trouvée devient : ${}^tAA(x - x_1) = 0$, soit ${}^tAA(x) = {}^tAA(x_1) = {}^tAb$ puisque ${}^tAAx_1 = {}^tAb_1 = {}^tAb$.

L'unicité équivaut manifestement à l'inversibilité de la matrice carrée ${}^tAA$, qui équivaut comme on vient de le voir au fait que A a un noyau nul, donc est aussi de rang n.

□

Montrons pour conclure comment la décomposition QR, dans le cas du rang maximum, permet de ramener le problème à une résolution de système triangulaire. Soit $\overline{R}$, la matrice triangulaire inversible telle que $R = \begin{bmatrix} \overline{R} \\ \mathbf{0} \end{bmatrix}$. On a puisque Q est orthogonale :

$$\|Ax - b\|^2 = \|QRx - b\|^2 = \|Rx - c\|^2, \text{ avec } c = Q^{-1}b = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}, \ c_1 \in \mathbb{R}^n \times \{0\}.$$

Donc $\|Ax - b\|^2 = \|\overline{R}x - c_1\|^2 + \|c_2\|^2$.

Donc la solution du problème des moindres carrés s'obtient simplement en résolvant le système triangulaire: $\overline{R}x = c_1$, le minimum obtenu étant $\|c_2\|^2$.

On observe que l'image de A est transformé par Q dans le sous-espace $\mathbb{R}^n \times \{0\}$ de $\mathbb{R}^m$. Si A est seulement une matrice de rang $r < n$, la méthode de Householder donne encore une décomposition QR, non unique, avec $R = \begin{bmatrix} \overline{R} \\ \mathbf{0} \end{bmatrix}$, $\overline{R}$ matrice échelonnée de taille $r \times n$, de rang r. La résolution du système $\overline{R}x = c_1$ fournit encore l'ensemble des solutions du problème des moindres carrés.

Remarque: lien avec la méthode de Gram-Schmidt: On suppose que $m = n$, et que A est inversible: Soient $s_{i,j}$ les coefficients de la matrice $S = R^{-1}$. Notons $A = [a_1, \cdots a_n]$, et $Q = [q_1, \cdots q_n]$, les écritures respectives de A et Q sous forme d'une liste de colonnes $a_i, q_j \in \mathbb{R}^n$. Alors l'égalité $AS = Q$, s'écrit:

$$\begin{cases} q_1 = s_{1,1}a_1 \\ \cdots \\ q_i = s_{1,i}a_1 + \cdots + s_{i,i}a_i \\ \cdots \\ q_1 = s_{1,n}a_1 + \cdots + s_{i,n}a_i + \cdots + s_{n,n}a_n \end{cases}$$

On reconnait la méthode d'orthogonalisation de Gram-Smidt pour l' obtention de bases orthonormées.

2 Méthode du gradient

On s'intéresse à la résolution d'un système $Au = b$, lorsque A est une matrice réelle symétrique définie positive, en utilisant cette hypothèse pour se ramener à une recherche de minimum:

Proposition 2.1 *La solution du système $Au = b$ est l'unique $\overline{u} \in \mathbb{R}^n$ qui réalise le minimum de :*

$$J(u) = < Au, u > -2 < b, u >$$

Démonstration La matrice A est inversible, puisque pour tout $x \in \mathbb{R}^n \setminus \{0\}$ on a $< Au, u > > 0$ donc $Au \neq 0$. Notons à priori $\overline{u}$ l'unique solution du système $Au = b$ et $\|u\|_A = \sqrt{< Au, u >}$ la norme associée à A.

On a donc en tenant compte de la symétrie de A, $< Au, v >=< u, Av >$

$$E(u) :=< A(\overline{u} - u), \overline{u} - u >= \|u\|_A^2 - 2 < u, A\overline{u} > + < A\overline{u}, \overline{u} >= J(u) + < b, \overline{u} >$$

La fonction $E(u)$ qui n'est autre que $\|\overline{u} - u\|_A^2$, atteint clairement sa valeur minimum au seul point $\overline{u}$, il en est donc de même de $J(u)$ qui diffère de E par la constante $\|\overline{u}\|_A^2 =< b, \overline{u} >$. □

Remarquons que J est une expression connue, ce qui n'est pas le cas de E puisque l'objet de cette section est précisément demontrer comment calculer une approximation de $\overline{u}$.

Notation 2.2 *On note $e(u) = u - \overline{u}$,("l'erreur"). On a $E(u) = \|e(u)\|_A$.*
$r(u) = b - Au = -A(e(u))$.

On a $r(u) = -A(e(u))$. On remarque aussi que $\overrightarrow{grad}(E)(u) = -2r(u)$, ce qui se voit en rappelant que $h \mapsto < \overrightarrow{grad}(E)(u), h >= dE(u)(h)$, différentielle de E ou de J au point u, et en considérant le développement:

$$J(u + h) = J(u) + 2 < Au, h > -2 < b, h > +\|h\|_A^2$$

2.1 La méthode de descente.

Elle consiste à calculer par récurrence une suite u_k, avec $u_{k+1} = u_k + \alpha_k \vec{p_k}$ qui tend vers la solution. On note $e_k = e(u_k) = u_k - \overline{u}$ et $r_k = r(u_k) = A(\overline{u} - u_k)$

Proposition 2.3 *Pour toute suite donnée de vecteurs $(\vec{p_k})_{k\in\mathbb{N}}$, il existe un choix optimal de α_k, minimisant $E(u_{k+1})$:*

$$\alpha_k = \frac{< r_k, \vec{p_k} >}{< A\vec{p_k}, \vec{p_k} >}$$

De plus pour tout k, $r_{k+1} = r_k - \alpha_k A\vec{p_k}$, et $< \vec{p_k}, r_{k+1} >= 0$

La dernière condition signifie que le vecteur $\vec{p_k}$ est orthogonal au gradient $\overrightarrow{grad}(E)(u_{k+1})$ ou tangent à la courbe de niveau de E au point u_{k+1}.

Démonstration *de la proposition* Etudions la restriction de E à la droite $u_k + \mathbb{R}\vec{p_k}$

$$\begin{aligned} E(u_k + \alpha\vec{p_k}) &=< A(u_k - \overline{u} + \alpha\vec{p_k}), u_k - \overline{u} + \alpha\vec{p_k} > \\ &=< Ae_k + A\alpha\vec{p_k}, e_k + \alpha\vec{p_k} > \\ &= E(u_k) + 2\alpha < A\vec{p_k}, e_k > +\alpha^2 < A\vec{p_k}, \vec{p_k} > \end{aligned}$$

Le minimum du trinome $\alpha \mapsto E(u_k + \alpha\vec{p_k})$ est atteint pour la valeur de α solution de $\frac{d}{d\alpha}(E(u_k + \alpha\vec{p_k})) = 0$, soit:

$$2 < A\vec{p_k}, e_k > +2\alpha < A\vec{p_k}, \vec{p_k} >= 0$$

Comme $< A\vec{p_k}, e_k >=< \vec{p_k}, Ae_k >= - < \vec{p_k}, r_k >$ on en déduit le résultat annoncé:

$$- < \vec{p_k}, r_k > +\alpha_k < A\vec{p_k}, \vec{p_k} >= 0$$

On trouve ensuite $r_{k+1} = A(\overline{u} - u_k - \alpha_k \vec{p_k}) = r_k - \alpha_k A\vec{p_k}$, et enfin comme annoncé:

$$< \vec{p_k}, r_{k+1} >=< \vec{p_k}, r_k > -\frac{< r_k, \vec{p_k} >}{< A\vec{p_k}, \vec{p_k} >} < A\vec{p_k}, \vec{p_k} >= 0$$

□

Interprétation géométrique.- Considérons l'hypersurface de niveau $\mathcal{E}_k$ de la fonction E passant pas u_k, d'équation $E(u) = E(u_k)$. Le résultat obtenu signifie que la droite $u_k + \mathbb{R}\vec{p_k}$ est tangente à $\mathcal{E}_{k+1}$ en u_{k+1}. C'est la traduction géométrique du fait que l'équation:

$$E(u_k + \alpha \vec{p_k}) = E(u_{k+1})$$

a une racine double en $\alpha = \alpha_k$. Le dernier résultat exprime la même chose par l'orthogonalité de cette droite et du gradient r_{k+1}.

On a par construction $E(u_{k+1}) \leq E(u_k)$.

Problème.- Comment choisir les directions de descente $\vec{p_k}$ pour que la suite u_k converge vers la solution $\overline{u}$, et avec la meilleure vitesse de convergence possible.

Les deux résultats qui suivent fournissent un premier élément de réponse:

Lemme 2.4 *1)* $E(u_{k+1}) = E(u_k)(1 - \gamma_k)$, *avec* $\gamma_k = (\frac{<\vec{p_k}, r_k>}{\|\vec{p_k}\|_A \|r_k\|_{A^{-1}}})^2$
2) On a $\gamma_k \geq \frac{1}{cond_2(A)}(\frac{<\vec{p_k}, r_k>}{\|\vec{p_k}\|\|r_k\|})^2$ *où* $\| \bullet \|$ *est la norme euclidienne usuelle.*

Le minimum du trinôme du second degré $T(\alpha) = a\alpha^2 - 2b\alpha + c$ est égal à $T(\frac{b}{a}) = c - \frac{b^2}{a}$, ce qui appliqué à la démonstration précédente se traduit par l'égalité:

$$E(u_{k+1}) = E(u_k) - \frac{< \vec{p_k}, r_k >^2}{< A\vec{p_k}, \vec{p_k} >}$$

Par définition $< A\vec{p_k}, \vec{p_k} >= \|\vec{p_k}\|_A$. Par ailleurs $E(u_k) =< A(e_k), e_k >) =< r_k, A^{-1}r_k >)$, puisque $-A(e_k) = r_k$, et la formule pour γ_k s'en déduit.

2) Notons $\lambda_i, i = 1, \cdots, n$ les valeurs propres de A, avec $0 < \lambda_n \leq \cdots \leq \lambda_1$. On a pour tout $u \in \mathbb{R}^n$, $\|Au\| \leq \lambda_1 \|u\|$ et $\|A^{-1}u\| \leq \frac{1}{\lambda_n}\|u\|$, donc :

$$< A\vec{p_k}, \vec{p_k} >\leq \lambda_1 \|\vec{p_k}\|^2, \text{ et } < r_k, A^{-1}r_k >\leq \frac{1}{\lambda_n}\|r_k\|^2$$

ce qui donne la minoration :

$$\gamma_k \geq \frac{< \vec{p_k}, r_k >^2}{\lambda_1 \|\vec{p_k}\|^2 \frac{1}{\lambda_n}\|r_k\|^2}$$

Pour conclure il suffit de se rappeler que pour une matrice symétrique le conditionnement euclidien est donné par: $cond_2(A) = \frac{\lambda_1}{\lambda_n}$

Théorème 2.5 *Pour toute suite de directions $\vec{p_k}$, telle que il existe un réel $\mu > 0$ satisfaisnt à la condition :*

$$\frac{< \vec{p_k}, r_k >}{\|\vec{p_k}\|\|r_k\|} \geq \mu > 0$$

la suite u_k associé au choix optimal de α_k, converge: $\lim_{k\to\infty} u_k = \overline{u}$

C'est une conséquence directe du lemme puisque la convergence annoncée par le théorème équivaut à $\lim_{k\to\infty} E(u_k) = 0$, et que d'après le lemme et l'hypothèse, $E(u_k)$ est majoré par une suite géométrique de raison : $0 \leq 1 - \frac{\mu^2}{cond_2(A)}$

Géométriquement l'expression $\frac{<\vec{p_k}, r_k>}{\|\vec{p_k}\|\|r_k\|}$ est le cosinus de l'angle que fait le gradient r_k avec la direction choisie $\vec{p_k}$, et la condition du théorème consiste à écarter les directiopn trop proches des directions tangentes en u_k à $\mathcal{E}_k$.

2.2 La méthode du gradient

Elle consiste à prendre le premier choix évident $p_k = r_k$, qui correspond au choix optimal

$$\alpha_k = \frac{\|r_k\|^2}{< Ar_k, r_k >}$$

Théorème 2.6 *La méthode du gradient avec paramètre optimal converge, et la suite $\|e_k\|_A = \sqrt{E(u_k)}$ est majorée par une suite géométrique cstek^n avec le coefficient de rapidité (on note $K(A) = cond_2(A)$:*

$$k = \frac{K(A) - 1}{K(A) + 1}.$$

Démonstration.- Elle repose sur inégalité de Kantorowicz(admise)

$$\gamma_k = \frac{\|r_k\|^4}{< Ar_k, r_k >< A^{-1}r_k, r_k >} \geq \frac{4\lambda_1\lambda_n}{(\lambda_1 + \lambda_n)^2} = \frac{4K(A)}{(K(A) + 1)^2}$$

En reportant dans l'inégalité du lemme 2.4 on en déduit facilement le résultat cherché:

$$0 \leq E(u_{k+1}) \leq E(u_k)(1 - \frac{4K(A)}{(K(A) + 1)^2}) = E(u_k)(\frac{K(A) - 1}{K(A) + 1})^2$$

□

2.3 Méthode du gradient à paramètre constant.

2.4 Méthode du gradient conjugué.

L'objectif de cette méthode est de choisir à chaque étape une direction $\vec{p_k}$ dans le plan engendré par $\vec{p_{k-1}}, r_k$, $\vec{p_k} = r_k + \beta_k \vec{p_{k-1}}$ de façon à rendre le facteur de réduction de l'erreur maximum.

De la relation d'orthogonalité $< \vec{p_{k-1}}, r_k >= 0$, on tire $< r_k, p_k >=< r_k, r_k + \beta_k \vec{p_{k-1}} >= \|r_k\|^2$ donc la relation du lemme 2.4 devient:

$$E(u_{k+1}) = E(u_k)(1 - \frac{< r_k, p_k >}{< Ap_k, p_k >< A^{-1}r_k, r_k >}) = E(u_k)(1 - \frac{\|r_k\|^2}{< Ap_k, p_k >< A^{-1}r_k, r_k >})$$

et la méthode annoncée consiste donc à considérer $\vec{p} = r_k + \beta \vec{p_{k-1}}$ et à <u>minimiser</u>

$$< Ap, p >= \beta^2 < Ap_{k-1}, p_{k-1} > +2\beta < Ap_{k-1}, r_k > + < Ar_k, r_k >,$$

d'où la solution :

$$\boxed{\beta_k = -\frac{<Ap_{k-1}, r_k>}{<Ap_{k-1}, p_{k-1}>}}$$

On remarque que ce résultat implique l'égalité:

$$< Ap_{k-1}, p_k >= 0$$

En effet, $< Ap_{k-1}, p_k >=< Ap_{k-1}, r_k - \frac{<Ap_{k-1}, r_k>}{<Ap_{k-1}, p_{k-1}>} \vec{p_{k-1}} >= 0$.

Rappelons la définition géométrique qui au vu de ce calcul justifie le nom donné à la méthode:

Définition 2.7 *On dit que $u \in \mathbb{C}^n$ et $v \in \mathbb{C}^n$ sont conjugués (par rapport à A) si et seulement si : $< Au, v >= 0$.*

La relation de conjugaison n'est autre que l'orthogonalité pour le produit scalaire défini par A.

Proposition 2.8 *On a les relations:*

$$\forall k \in \mathbb{N},\ < r_{k+1}, r_k >= 0$$

$$\forall k \geq 1,\ \beta_k = \frac{\|r_k\|^2}{\|r_{k-1}\|^2}$$

Démonstration.- Il s'agit d'un simple calcul:

1)

$$\begin{aligned} < r_{k+1}, r_k > &=< r_k - \alpha_k Ap_k, r_k > \\ &= \|r_k\|^2 - \alpha_k < Ap_k, p_k - \beta_k p_{k-1} > \\ &= \|r_k\|^2 - \alpha_k < Ap_k, p_k >= 0 \end{aligned}$$

2)

$$< Ap_{k-1}, r_k > = \frac{1}{\alpha_{k-1}} < r_{k-1} - r_k, r_k > \qquad = \frac{1}{\alpha_{k-1}} \|r_k\|^2$$

$$< Ap_{k-1}, p_{k-1} > = \frac{1}{\alpha_{k-1}} < r_{k-1} - r_k, p_{k-1} > \qquad = \frac{1}{\alpha_{k-1}} \|r_{k-1}\|^2$$

D'où par quotient le résultat annoncé:

$$\beta_k = \frac{< Ap_{k-1}, r_k >}{< Ap_{k-1}, p_{k-1} >} = \frac{\|r_k\|^2}{\|r_{k-1}\|^2}$$

□

3 Demonstration sur le calcul du parametre optimal de relaxation pour les matrices tridiagonales

Démonstration.- Partie 1: On a vu que $\mathcal{L}_\omega = (\frac{D}{\omega} - E)^{-1}(\frac{1-\omega}{\omega}D + F)$. Notons encore $p_{\mathcal{L}_\omega}$ son polynôme caractéristique. On peut mener des calculs similaires à ceux du théorème **??** avec le polynôme:

$$q_{\mathcal{L}_\omega}(\lambda) = det(E - \frac{D}{\omega})p_{\mathcal{L}_\omega}(\lambda) = det[-\frac{1-\omega}{\omega}D - F + \lambda(\frac{D}{\omega} - E)]$$

Par le lemme de la démonstration du théorème **??**, on en déduit:

$$q_{\mathcal{L}_\omega}(\lambda^2) = det(\frac{\lambda^2 - \omega - 1}{\omega}.D - \lambda^2 E - F) = det(\frac{\lambda^2 - \omega - 1}{\omega}.D - \lambda E - \lambda F)$$

soit finalement

$$\boxed{q_{\mathcal{L}_\omega}(\lambda^2) = \lambda^n.q_J(\tfrac{\lambda^2-\omega-1}{\lambda\omega})}$$

Conclusion: On a établi l'équivalence :

$$\beta = \lambda^2 \in Sp(\mathcal{L}_\omega) \setminus \{0\} \Leftrightarrow \{\frac{\lambda^2 + \omega - 1}{\lambda\omega}, -\frac{\lambda^2 + \omega - 1}{\lambda\omega}\} \subset Sp(J)\}$$

Partie 2: On fixe tel que $\{\alpha, -\alpha\} \subset J$, et on cherche à évaluer en fonction de ω la plus grande des valeurs corrrespondante du spectre de $\mathcal{L}_\omega$.

Pour cela résolvons l'équation

$$\frac{\lambda^2 + \omega - 1}{\lambda\omega} = \pm\alpha \text{ ou } \lambda^2 \mp \alpha\omega\lambda - 1 = 0 \tag{1}$$

Le discriminant est $\Delta(\alpha, \omega) = \alpha^2\omega^2 - 4\omega + 4$ et les deux racines de l'équations sont :

$$\lambda = \frac{\mp\alpha\omega \pm \sqrt{\Delta(\alpha, \omega)}}{2}$$

On trouve 4 valeurs de λ deux à deux opposées, donc en élevant au carré les deux valeurs possibles du spectre de $\mathcal{L}\omega$ qu'on notera $\mu_\pm(\alpha, \omega)$:

$$\boxed{\mu_\pm = \tfrac{1}{4}(\alpha^2\omega^2) \pm \tfrac{\alpha\omega}{2}\sqrt{\Delta} + \tfrac{1}{4}(\alpha^2\omega^2 - 4\omega + 4) = \tfrac{\alpha^2\omega^2 - 2\omega + 2 \pm \alpha\omega\sqrt{\Delta(\alpha,\omega)}}{2}}$$

Partie 3: L'objectif est maintenant l'étude, toujours à α fixé de

$$M(\alpha, \omega) := \max(|\mu_+(\alpha, \omega)|, |\mu_-(\alpha, \omega)|)$$

Remarquons au passage que à la limite $M(0, \omega) = |1 - \omega|$.

On va maintenant utiliser l'hypothèse $Sp(J) \subset \mathbb{R}$. Etant donnée la symétrie entre α et $-\alpha$, on peut supposer $\alpha > 0$.

On va examiner selon les cas si l'équation (1) a des racines réelles ou complexes conjuguées, ce qui dépend du signe de $\Delta(\alpha,\omega) = \alpha^2\omega^2 - 4\omega + 4$. On observe que l'équation $\alpha^2\omega^2 - 4\omega + 4 = 0$ a pour discriminant réduit $\delta' = 1 - \alpha^2$.

Cas i) $\alpha \geq 1$. Dans ce cas $\Delta(\alpha,\omega) \geq 0$ et $2\sqrt{M(\alpha,\omega)} = \alpha\omega + \sqrt{\Delta(\alpha,\omega)}$.

Or $x \mapsto M(x,\omega)$ est une fontion croissante de $x \in [1,\infty[$. Comme pour tout $\omega \in [0,2]$, on a

$$M(1,\omega) = (\omega + \sqrt{\Delta(1,\omega)})^2 = (\frac{\omega + \sqrt{(\omega-2)^2}}{2})^2 = 1$$

On obtient: $M(\alpha,\omega) \geq 0$.

Ceci permet de conclure pour la première partie de l'assertion: si la méthode de Jacobi est non convergente, on a aussi $\rho(\mathcal{L}_\omega) \geq 1$, et la méthode de relaxation diverge pour tout paramètre ω.

Cas ii) $0 \leq \alpha < 1$. Dans ce cas l'équation $\Delta(\alpha,\omega) = \alpha^2\omega^2 - 4\omega + 4 = 0$ a deux racines $\omega_0(\alpha)$ et $\omega_1(\alpha)$, avec :

$$1 < \omega_0(\alpha) = \frac{2 - 2\sqrt{1-\alpha^2}}{\alpha^2} = \frac{2}{1+\sqrt{1-\alpha^2}} < 2 < \omega_1(\alpha) = \frac{2 + 2\sqrt{1-\alpha^2}}{\alpha^2} = \frac{2}{1-\sqrt{1-\alpha^2}}$$

On va montrer que $\omega \mapsto M(\alpha,\omega)$ atteint son minimum sur $]0,2[$ en $\omega_0(\alpha)$.

Il y a deux cas à distinguer pour ω :

- Si $\omega \in [\omega_0(\alpha), 2]$, on est dans le domaine de la variable ω où $\Delta(\alpha,\omega) < 0$, donc où l'équation (1) a deux racines conjuguées. On a donc de même pour les carrés de ces racines $\mu_-(\alpha,\omega) = \overline{\mu_+(\alpha,\omega)}$, donc en se référant à nouveau à l'équation (1), pour le produit de ses racines on trouve:

  $$|\mu_+| = \mu_-\mu_+ = \omega - 1$$

  Par conséquent: $M(\alpha,\omega) = \omega - 1$, sur $[\omega_0(\alpha), 2]$.

- Si $0 < \omega \leq \omega_0(\alpha)$, alors $M(\alpha,\omega) = \mu_(\alpha,\omega) = \lambda(\alpha,\omega)^2$, avec

  $$2.\lambda(\alpha,\omega) = \alpha.\omega + \sqrt{\alpha^2\omega^2 - 4\omega + 4}$$

  Pour terminer la démonstration il reste essentiellement à montrer que cette expression est une fonction décroissante de ω sur l'intervalle $[0,\omega_0(\alpha)]$:

  $$\frac{\partial M}{\partial \omega} = 2\lambda\frac{\partial \lambda}{\partial \omega} = 2\lambda(\alpha + \frac{2\omega\alpha^2 - 4}{2\sqrt{\Delta(\alpha,\omega)}})$$

  En remplaçant $\sqrt{\Delta(\alpha,\omega)}$ par $2\lambda(\alpha,\omega) - \alpha\omega$ cela donne $\frac{\partial M}{\partial \omega} = \frac{\lambda[2\alpha(2\lambda(\alpha,\omega)-\alpha\omega)+2\omega\alpha^2-4]}{\sqrt{\Delta(\alpha,\omega)}}$, soit finalement

  $$\boxed{\frac{\partial M}{\partial \omega} = 4\lambda(\alpha,\omega)\frac{\alpha\lambda(\alpha,\omega)-1}{\sqrt{\Delta(\alpha,\omega)}}}$$

On contrôle pour finir le fait que $\alpha\lambda(\alpha,\omega)-1<0$, si $0<\omega<\omega_0(\alpha)$

En effet, puisque $\lambda(\alpha,\omega)$ est solution de l'équation $X^2-\alpha\omega X+\omega-1$, qu'on peut tranformer en $(\alpha X-1)\omega=X^2-1$, on a bien comme souhaité

$$(\alpha\lambda(\alpha,\omega)-1)\omega=\lambda(\alpha,\omega)^2-1<0$$

puisque :

$$\begin{aligned}\lambda(\alpha,\omega)^2&=\frac{\alpha^2\omega^2-2\omega+2+\alpha\omega\sqrt{\alpha^2\omega^2-4\omega+4}}{2}\\&<\frac{1}{2}(\omega^2-2\omega+2+\omega\sqrt{(\omega^2-2)^2})=\frac{1}{2}(\omega^2-2\omega+2+\omega(2-\omega))=1\end{aligned}$$

Bilan: La fonction $\frac{1}{2}(\alpha.\omega+\sqrt{\alpha^2\omega^2-4\omega+4})$ est une fonction croissante de α et est égale à $M(\alpha,\omega)^{\frac{1}{2}}$ pour $0<\omega\leq\omega_0(\alpha)$ et à $\omega-1$ si $\omega_0(\alpha)\leq\omega$. On en déduit (cf figure suivante) que pour $\alpha<\alpha'$ le graphe de $\omega\mapsto M(\alpha,\omega)$ est entierèment situé en dessous de celui de $\omega\mapsto M(\alpha',\omega)$, que $M(\alpha,\omega)$ atteint son minimum en $\omega_0(\alpha)=\frac{2}{1+\sqrt{1-\alpha^2}}$, et que en définitive le paramètre optimal de relaxation est

$$\omega_0=\omega_0(\rho(J))=\frac{2}{1+\sqrt{1-\rho(J)^2}}$$

□

4 Recherche de vecteurs propres et de valeurs propres

4.1 Méthode de la puissance itérée

4.1.1 Recherche de la valeur propre de plus grand module.

Considérons l'exemple suivant: la matrice $\begin{pmatrix} 10 & 0 \\ -9 & 1 \end{pmatrix}$ admet deux valeurs propres $\lambda = 10$ et $\mu = 1$, avec les vecteurs propres unitaires (pour la norme $\| \bullet \|_\infty$):

$$v_{10} = \begin{pmatrix} 1 \\ -1 \end{pmatrix} \quad \text{et } v_1 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

Considérons alors l'opération $v_k = A^k v_0$, de vecteur initial $v_0 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$. On trouve

$$v_k = \begin{pmatrix} 10^k \\ -10^k + 2 \end{pmatrix}$$

On peut alors faire les constatations suivantes:

1. La direction limite pour v_k est un vecteur propre pour $\lambda = 10$. C'est à dire que

$$\frac{v_k}{\|v_k\|_\infty} = \begin{pmatrix} 1 \\ -1 + 2 \times 10^{-k} \end{pmatrix} \underset{k \to \infty}{\longrightarrow} \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

2. Le quotient $\frac{v_{k+1}(i)}{v_k(i)}$ est défini pour k assez grand et $\frac{v_{k+1}(i)}{v_k(i)} \underset{k \to \infty}{\longrightarrow} \lambda = 10$

Cet exemple nous suggère l'algorithme suivant de recherche d'un valeur propre de module maximum. On fixe une norme $\| \bullet \|$ sur $\mathbb{C}^n$, et une matrice $A \in \mathcal{M}_{n,n}(\mathbb{C})$, et on désigne par $v(j)$ la jième coordonnée d'un vecteur $v \in \mathbb{C}^n$:

ALGORITHME

1. Choisir $q_0 \in \mathbb{C}^n$ tel que $\|q_0\| = 1$.
2. $x_k = A.q_{k-1}$ et si $q_{k-1}(j) \neq 0$, $\lambda_k(j) = \frac{x_k(j)}{q_{k-1}(j)}$
3. Si $\gamma_k = \|x_k\| \neq 0$, on définit $q_k = \frac{x_k}{\gamma_k}$

Remarques:
1) Si $\gamma_k = 0$, l'algorithme s'arrête au cran k, ce qui n'arrive pas pour A inversible ni pour q_0 assez général (précisément, pourvu que A ne soit pas nilpotente, dès que q_0 est en dehors du sous espace caractéristique généralisé relatif à zéro)
2) q_k est unitaire donc pour chaque k: $\exists j,\ q_{k-1}(j) \neq 0$ autrement dit $\lambda_k(j)$ est défini. Si q_k a une limite sur la sphère unité, on peut choisir j indépendant de k, pour k assez grand.
3) $q_k = \frac{A^k q_0}{\|A^k q_0\|}$, par une récurrence immédiate sur k.

Théorème 4.1 *On suppose que A est diagonalisable avec une seule valeur propre notée λ_1 de module maximum, de multiplicité $p < n$. On ordonne les valeurs propres $\lambda_1 = \cdots = \lambda_p, \lambda_{p+1}, \cdots, \lambda_n$ par modules décroissants.*
1) Pour q_0 bien choisi, précisément en dehors d'un sous espace vectoriel strict, la suite construite dans l'algorithme fournit, à la fois, un valeur approchée de λ_1 et un vecteur propre de la façon suivante:

a) $\lim_{k\to\infty} (\frac{\overline{\lambda_1}}{|\lambda_1|})^k q_k = q$ existe et on a :

$$\|q\| = 1, \quad \text{et } q \in E_{\lambda_1}, \text{ sous espace propre associé à } \lambda_1$$

b) D'après a), il existe j tel que $\exists k_0,\ \forall k \geq k_0,\ q_k(j) \neq 0$, et alors :

$$\lambda_1 = \lim_{k\to\infty} \frac{x_{k+1}(j)}{q_k(j)}$$

2) L'erreur $\|(\frac{\overline{\lambda_1}}{|\lambda_1|})^k q_k - q\|$, est majorée par une série géométrique $Cste|\frac{\lambda_{p+1}}{\lambda_1}|^k$.

Démonstration On considère une base $(u_1, \cdots, u_n)$ de vecteurs propres, avec u_j associé à λ_j, et on note α_j, la jème coordonnée de q_0:

$$q_0 = \alpha_1 u_1 + \cdots + \alpha_n u_n$$

Dans la somme directe $\mathbb{C}^n = E_{\lambda_1} \bigoplus (\bigoplus_{\lambda_j \neq \lambda_1} E_{\lambda_j})$ la composante de q_0 sur E_{λ_1} est $u = \alpha_1 u_1 + \cdots + \alpha_p u_p$ et on a

$$q_0 = u + \sum_{j=p+1}^{n} \alpha_j u_j.$$

La condition requise sur q_0 est $u \neq 0$, ce qui équivaut encore à $q_0 \notin \bigoplus_{\lambda_j \neq \lambda_1} E_{\lambda_j}$. Le vecteur $q_k = \frac{A^k q_0}{\|A^k q_0\|}$ est donc le vecteur unitaire associé à

$$\begin{aligned} A^k q_0 &= \lambda_1^k u + \sum_{j=p+1}^{n} \alpha_j \lambda_j^k u_j \\ &= \lambda_1^k (u + \sum_{j=p+1}^{n} \alpha_j (\frac{\lambda_j}{\lambda_1})^k u_j) \end{aligned}$$

Posons $e_k = \sum_{j=p+1}^{n} \alpha_j (\frac{\lambda_j}{\lambda_1})^k u_j$, on a $\lim_{k\to\infty} e_k = 0$, et plus précisément du fait que pour tout $j \geq p+1$, $\lambda_j \leq \lambda p + 1$, on déduit l'existence d'un constante C (en fait $C = \sum |\alpha_j| \|u_j\|$), telle que :

$$\|e_k\| \leq C|\frac{\lambda_{p+1}}{\lambda_1}|^k$$

On en tire

$$q_k = \frac{\lambda_1^k}{|\lambda_1|^k} . \frac{u + e_k}{\|u + e_k\|} = \frac{|\lambda_1|^k}{\overline{\lambda_1}^k} . \frac{u + e_k}{\|u + e_k\|}$$

donc

$$q'_k = \frac{\overline{\lambda_1}^k}{|\lambda_1|^k} q_k \underset{k \to \infty}{\longrightarrow} q = \frac{u + e_k}{\|u + e_k\|}$$

Ceci démontre le premier résultat.
Pour le résultat sur la vitesse de convergence il suffit d'écrire: $q'_k - q = \frac{\|u\|.(u+e_k) - \|u+e_k\|u}{\|u+e_k\|.\|u\|}$, majoré en norme par $\frac{2.\|e_k\|}{\|u+e_k\|.}$ $\square$

4.2 Généralisation

Si A est non diagonalisable mais avec une seule valeur propre, notée λ_1, de module maximum, le résultat est encore valable. Ceci peut se voir sur les blocs de Jordan de A. On note :

$$J_q(\lambda) = \begin{pmatrix} \lambda & 1 & 0 & \cdots & 0 \\ 0 & \lambda & \ddots & & \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & & & \lambda & 1 \\ 0 & & \cdots & 0 & \lambda \end{pmatrix}$$

un bloc de Jordan de taille $q \times q$.

Cas 1 λ_1 est une valeur propre simple. La réduction de Jordan de A s'écrit alors:

$$P^{-1}AP = \begin{pmatrix} \lambda_1 & & & & & \\ & \ddots & & & 0 & \\ & & \lambda_1 & & & \\ & & & J_{q_2}(\lambda_2) & & \\ & 0 & & & \ddots & \\ & & & & & J_{q_p}(\lambda_p) \end{pmatrix}$$

où $\lambda_2, \cdots, \lambda_p$, sont les valeurs propres de A de modules tous $< |\lambda_1|$. On a encore une convergence vers zéro de $q'_k - q$, avec une vitesse de convergence similaire, puisque le reste est majorée par une série polynôme×géométrique proportionnelle à $k^n |\frac{\lambda_2}{\lambda_1}|^k$. Les détails sont laissés en exercice. Cela revient à considérer

$$P^{-1}A^kP = \begin{pmatrix} \lambda_1^k & & & & & \\ & \ddots & & & 0 & \\ & & \lambda_1^k & & & \\ & & & J_{q_2}(\lambda_2)^k & & \\ & 0 & & & \ddots & \\ & & & & & J_{q_p}(\lambda_p)^k \end{pmatrix}$$

et à voir que tous les coefficients de $\frac{J_{q_i}(\lambda_j)^k}{\lambda_1^k}$, sont de la forme $\frac{C_k^l \lambda_2^{k-\ell}}{\lambda_1^k}$, avec ℓ borne par n.

Cas général La convergence est beaucoup plus lente dès qu'il existe des blocs de Jordan de taille ≥ 2 pour λ_1. Supposons pour simplifier les notations, tout en traitant essntiellement la question, que A est égale à un bloc de Jordan: $A = J_n(\lambda)$. Alors, la $(j+1)$ième colonne de A^k est le vecteur:

$$x_k = A^k(e_{j+1}) = \begin{pmatrix} C_k^j \lambda_1^{k-j} \\ \vdots \\ C_k^1 \lambda_1^{k-1} \\ \lambda_1^k \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

et on trouve encore que $q_k = \frac{x_k}{\|x_k\|}$, tend vers e_1, vecteur propre.

Toutefois, $\frac{(q_{k+1})_1}{(q_k)_1} = \lambda_1 \frac{C_{k+1}^j}{C_k^j} = \frac{k+1}{k-j+1}$ tend vers λ_1 très lentement, le reste étant cette fois un $O(\frac{1}{k})$.

Remarques. 1) S'il existe plusieurs valeurs propres de module maximum il n'y a plus convergence. Par exemple, une matrice de rotation a deux valeurs propres $e^{i\theta}$ et $e^{-i\theta}$, et quel que soit $z \in \mathbb{C}^2$, dont la décomposition sur les espaces propres est $z = z_+ + z_-$, la direction du vecteur $Az = e^{in\theta} z_+ + e^{-in\theta} z_-$, ne tend vers aucune des deux directions propres.

2) (Sur le choix de q_0) Dans les problèmes concrets, où la valeur de λ_1, n'est pas connu, il en est de même de l'espace des choix à exclure pour q_0. Mais dans la pratique un choix au hasard contient toujours une composante propre sur A_{λ_1}.

3) Dans une matrice prise au hasard, selon la commande $rand(n, n)$ de scilab par exemple, on est dans les hypothèses du théorème, l'espace des choix à exclure (matrice non diagonalisables ou avec des relations $|\lambda_i| = |\lambda_j|$) étant de mesure nulle.

4) Si on choisit $q_0 \in \bigoplus_{j \geq p+1} \mathbb{C}.u_j$, dans les notations de la démonstration du théorème 4.1, on trouverait de même, et en théorie la valeur propre λ_{p+1}. Dans la pratique, sauf cas simple où l'espace propre E_{λ_1} est connu explicitement, les erreurs d'arrondi sur q_0, entrainent l'apparition d'une composante non nulle sur E_{λ_1}, qui finit toujours par l'emporter.

Pour une donnée initiale très proche de $E_{\lambda_{p+1}}$, , on trouve une suite $k \mapsto \lambda_k(j)$, présentant un palier, en λ_2, avant de converger vers λ_1, comme dans la figure ci-dessous.

Voir dans le Lascaux Theodor des exemples de ce type ainsi que des exemples à convergense lente où $\frac{\lambda_2}{\lambda_1} \sim 0.99$.

4.2.1 La méthode de la puissance inverse

On peut d'abord remarquer que si A est inversible avec des valeurs propres rangées par ordre décroissant de façon que :

$$|\lambda_1| \geq \cdots \geq |\lambda_{n-1}| > |\lambda_n| > 0,$$

la méthode précédente appliquée à A^{-1} donne $\frac{1}{\lambda_n}$, comme plus grande valeur propre de A^{-1}, puis λ_n par inversion.

ALGORITHME pour la plus petite valeur propre.
1) Effectuer une fois pour toute une factorisation LU.
2) Résoudre $Ax_k = q_{k-1}$
3) Poursuivre le reste de l'algorithme comme dans le cas de la recherche de la plus grande valeur propre.

De même, la méthode appliquée à $(A - \mu I)^{-1}$, aboutit à $\frac{1}{\lambda - \mu}$, avec λ valeur propre la plus proche de μ.

4.2.2 Méthode de déflation

Lorsque les valeurs propres de A sont rangées par ordre de module croissant, en supposant pour simplifier ces modules tous distincts :

$$|\lambda_1| > \cdots > |\lambda_n|$$

on peut chercher à calculer successivement $\lambda_1, \lambda_2, ...$ A titre d'exemple nous donnons la méthode par soustraction appropriée pour les matrices hermitiennes. Dans ce cas $\lambda_i \in \mathbb{R}$.

Le principe est le suivant:

Soit q un vecteur propre de A pour la valeur propre λ_1 trouvé par exemple par la méthode de la puissance et qu'on supposera unitaire : $q' * q = 1$.

On note $B = A - \lambda_1 q \star q'$. Alors on reparque que $B \star q = 0$, et que si $v \in q^{\perp}$ est orthogonal à q, on a $B \star v = A \star v$ Comme A est diagonalisable dans une base orthonormée comme matrice hermitienne, on constate que B a les même valeurs propres et vecteurs propres, à l'exception de λ_1 remplace par zéro.

Ainsi de proche en proche on pourrait déterminer tout le spectre.

Malgré son apparente simplicité conceptuelle cette méthoe a des limites car elle n'est pas stable à cause des erreurs d'arrondis.

Cette méthode est utilisée dans les problèmes où on ne souhaite calculer qu'un nombre limité de valeurs propres à partir des plus grandes.

5 Méthode de Jacobi

Il s'agit d'un méthode, assez ancienne, de réduction des matrices symétriques, efficace sur les matrices pleines.

Le principe est le suivant: on construit un suite de matrices orthogonales, Ω_k, et une suite de matrices semblables à A:

$$A_0 = A, \quad \text{et } A_{k+1} = {}^t\Omega_k A_k \Omega_k$$

la matrice Ω_k étant choisie à chaque étape pour annuler un terme non diagonal, et "se rapprocher" d'une matrice diagonale.

On utilise la matrice $\Omega = \Omega_{p,q}(\theta) = \omega_{i,j}$ définie par

$$\begin{cases} \omega_{i,j} = \delta_{i,j}, & \text{si } \{i,j\} \neq \{p,q\} \ , \\ \begin{pmatrix} \omega_{p,p} & \omega_{p,q} \\ \omega_{q,p} & \omega_{q,q} \end{pmatrix} = \begin{pmatrix} cos\theta & sin\theta \\ -sin\theta & cos\theta \end{pmatrix}. \end{cases}$$

On appellera matrice orthogonale élémentaire une telle matrice.

Théorème 5.1 *Soit $B = (b_{i,j})$ la matrice semblable à A, et symétrique:*

$$B = \Omega^{-1} A \Omega = {}^t\Omega A \Omega$$

1) $\sum b_{i,j}^2 = \sum a_{i,j}^2$
2) Si $a_{p,q\neq 0}$, il existe un unique $\theta \in]-\frac{\pi}{4}, \frac{\pi}{4}] \setminus \{0\}$ tel que $b_{p,q} = 0$. Précisément θ est déterminé par l'égalité

$$cotan\theta = \frac{a_{q,q} - a_{p,p}}{2a_{p,q}}$$

3) Avec la valeur de θ trouvée en 2), on a en plus:

$$\sum_{i=1}^{n} b_{i,i}^2 = \sum_{i=1}^{n} a_{i,i}^2 + 2a_{p,q}^2$$

Le dernier résultat montre que dans la somme fixe $\sum a_{i,j}^2$, la part de la diagonale augmente à chaque étape.

Démonstration 1) Un calcul élémentaire montre (exercice!) que $\sum a_{i,j}^2 = tr\ {}^tAA$. La propriété 1) se déduit alors du fait que les matrices ${}^tAA$ et ${}^tBB$ sont semblables, donc de même trace, conséquence de l'orthogonalité de Ω et du calcul suivant:

$${}^tBB = {}^t\Omega\ {}^tA\Omega\ {}^t\Omega A\Omega = {}^t\Omega({}^tAA)\Omega$$

2) Le passage de B à A est un changement de base qui laisse fixe les e_j pour $j \notin \{p,q\}$, et qui dans l'espace engendré par $\{e_p, e_q\}$ est un changement de bases orthonormées directes. Les deux conséquences suivantes s'en déduisent:

- Si $\{i,j\} \cap \{p,q\} = \emptyset$, $b_{i,j} = a_{i,j}$.
- $\begin{pmatrix} b_{p,p} & b_{p,q} \\ b_{q,p} & b_{q,q} \end{pmatrix} = \begin{pmatrix} c & -s \\ s & c \end{pmatrix} \begin{pmatrix} a_{p,p} & a_{p,q} \\ a_{q,p} & a_{q,q} \end{pmatrix} \begin{pmatrix} c & s \\ -s & c \end{pmatrix}$

où $c = \cos\theta$, $s = \sin\theta$. La dernière relation donne alors :

$$\begin{aligned} b_{p,q} = b_{q,p} &= c(a_{p,p}s + a_{p,q}c) - s(a_{q,p}s + a_{q,q}c) \\ &= cs(a_{p,p} - a_{q,q}) + a_{p,q}(c^2 - s^2) = \frac{\sin 2\theta}{2}(a_{p,p} - a_{q,q}) + \cos 2\theta a_{p,q} \end{aligned}$$

La condition requise pour annuler $b_{p,q}$ est donc:

$$\boxed{\frac{\cos 2\theta}{\sin 2\theta} = \frac{a_{q,q} - a_{p,p}}{2a_{p,q}}}$$

équation qui possède bien une solution unique avec $2\theta \in]-\frac{\pi}{2}, \frac{\pi}{2}] \setminus \{0\}$.

3) Les sous-matrices 2×2, $\begin{pmatrix} b_{p,p} & b_{p,q} \\ b_{q,p} & b_{q,q} \end{pmatrix}$ et $\begin{pmatrix} a_{p,p} & a_{p,q} \\ a_{q,p} & a_{q,q} \end{pmatrix}$, extraites de A et B par restriction à l'espace engendré par e_p et e_q, s'échangent aussi par une tranformation orthogonale élémentaire. On peut donc apppliquer à cette matrice le résultat 1) ce qui s'écrit:

$$b_{p,p}^2 + b_{q,q}^2 = a_{p,p}^2 + a_{q,q}^2 + 2a_{p,q}^2$$

et donne l'égalité annoncée en 3). □

On peut illustrer ce théorème par le schéma suivant

$$\begin{pmatrix} & & & b_{1,p} & & & b_{1,q} & & & \\ & = & & \times & & = & \times & & = & \\ & & & \times & & & \times & & = & \\ b_{p,1} & \times & \times & b_{p,p} & \times & \times & 0 & \times & \times & b_{p,n} \\ & & & \times & & & \times & & & \\ & = & & \times & & = & \times & & = & \\ & & & \times & & & \times & & & \\ b_{q,1} & \times & \times & 0 & \times & \times & b_{q,q} & \times & \times & b_{q,n} \\ & & & \times & & & \times & & & \\ & = & & \times & & = & \times & & = & \\ & = & & b_{n,p} & & = & b_{n,q} & & = & \end{pmatrix}$$

Dans ce schéma, les termes contenus dans les cases avec un signe = au centre sont inchangés par rapport aux $a_{i,j}$, alors que tous les $a_{i,j}$ dans lequel un au moins des indices est dans $\{p,q\}$ sont altérés par le transformation.

Voir dans Lascaux theodor, tome 2 p. , l'illustration suivante, aisée à programmer: on représente dans chaque position (i,j) la taille du coefficient $a_{i,j}^{(k)}$ de A_k à la $k$ième itération par un carré de coté $|a_{i,j}^{(k)}|$. L'aire totale de ces carrés est constante selon le résultat 1) du théorème

et selon le résultat 3), on observe une accumulation d'aire sur la diagonale, et sous réserve de convergence l'évanouissement progressif des autres carrés.

N.B. Prendre garde au fait que si à la k ième étape on a obtenu $b_{p,q} = 0$, une rotation ultérieure (avec par exemple un indice selectionné)(p, q') peut faire resurgir in coefficient non nul d'indice (p, q).Il n'est donc pas vrai que le processus converge en $\frac{n(n-1)}{2}$ étapes.

Dans ce qui suit on note $t = tan\theta$. On rappelle que:

$$\cos 2\theta = \frac{1-t^2}{1+t^2},\ \sin 2\theta = \frac{2t}{1+t^2},\ \cot 2\theta = \frac{1-t^2}{2t}$$

La valeur de t issu du point 2) du théorème est donc la solution de module ≤ 1 (ou $t = 1$ dans le cas des racines $-1, +1$) de l'équation

$$t^2 + \frac{a_{q,q} - a_{p,p}}{a_{p,q}} t - 1 = 0,\ t \in]0, 1] \tag{2}$$

ALGORITHME

- Sélectionner un indice (p, q) tel que $|a_{p,q}|$ soit maximum.
- Calculer t solution de l'équation 2.
- Faire $b_{p,q} = b_{q,p} = 0$, et pour $\{i, j\} \cap \{p, q\} \neq \emptyset$ avec $(p, q) \neq (i, j) \neq (q, p)$, calculer:

$$b_{i,j} = ({}^t\Omega A \Omega)_{i,j}$$

Commentaire sur l'algorithme 1) Il s'agit d'une description non formalisée de la $k$ième étape. A chaque étape on réaffecte conformément aux formules les variables $A(i, j)$.
2) Dans l'item 3) il est inutile de préciser que les autre $a_{i,j}$ sont inchangé.
3) On peut programmer une boucle du type pour k de 1 à n, ou en fonction d'une précision choisie avec une instruction conditionnelle: $for\ \ i \neq j,\ while\ |a_{i,j}| > 10^{-N}$.
4) Dans le calcul des $b_{i,j}$ à programmer selon le troisième item, il n'est pas utile de déterminer θ, puisque toutes les formules passent par des expressions algébriques de la variable t. En effet:

$$c = \cos t = \frac{1}{\sqrt{1+t^2}},\ \text{et}\ s = \sin t = ct = \frac{t}{\sqrt{1+t^2}}.$$

Variantes dans la stratégie de choix de l'indice (p, q).
1) Au lieu de choisir le plus grand des $a_{p,q}$, on peut effectuer un balayage de tous les indices

$$(p, q) = (2, 1), \cdots, (n, 1), (3, 2), \cdots, \cdots, (n, n-1)$$

Inconvénient: ne maximise pas a chaque étape le gain sur $\sum a_{,i}^2$. Avantage: gain de temps de calcul consistant à éviter à chaque étape la recherche d'un max avec $n(n-1)/2$ comparaisons.
2) Même stratégie avec instruction contidionnelle d'omettre p, q si $|a_{p,q}| < 10^{-N}$. Elimine les étapes non rentables.

Théorème 5.2 *Dans la méthode de Jacobi classique avec choix du coefficient maximum à chaque étape, la convergence a lieu: $limA_k = D$, matrice diagonale semblable à A*

Démonstration On note $A_k = D_k + B_k$, le résultat de la $k$ième itération et D_k sa partie diagonale $D_k = diag(a_{1,1}^{(k)}, \cdots, a_{n,n}^{(k)})$.

i) Montrons d'abord que la suite B_k tend vers zéro. Soit $\| \bullet \|_E$, la norme (qui n'est pas subordonnée ni même sous-multiplicative!) définie par $\|(a_{i,j})\|^2 = \sum a_{i,j}^2$. En vertu du point 1) du théorème 5.1, on a $\|A_k\| = \|A_{k+1}\|$, et si on pose

$$\epsilon_k = \|B_k\|^2 = \sum_{i \neq j} a_{i,j}^{(k)},$$

il s'agit d'établir le fait que $\epsilon_k \underset{k\to\infty}{\longrightarrow} 0$.

Par le point 3) du théorème 5.1, on a :

$$0 \leq \epsilon_{k+1} = \epsilon_k - 2|a_{p_k,q_k}^{(k)}|^2 < \epsilon_k$$

Par le choix du coefficient maximum à l'étape k on a $:\forall(i,j),\ |a_{i,j}^{(k)}| \leq |a_{p_k,q_k}^{(k)}|$, et on obtient $\epsilon_k \leq n(n-1)|a_{p_k,q_k}^{(k)}|^2$. D'où

$$0 \leq \epsilon_{k+1} \leq \epsilon_k(1 - \frac{2}{n(n-1)})$$

Par récurrence cela donne $\epsilon_k \leq \epsilon_0(1 - \frac{2}{n(n-1)})^k$ et montre que $B_k \underset{k\to\infty}{\longrightarrow} 0$.

ii) On remarque ensuite que la suite D_k n'a qu'un nombre fini de valeurs d'adhérence. Soit en effet D une telle valeur d'adhérence et D_{k_p}, un suite extraite tendant vers D. Comme $B_k = A_k - D_k$ tend vers zéro on a aussi $\lim_{p\to\infty} A_{k_p} = D$, donc le polynôme caractéristique de D est (coefficient par coefficient) la limite de celui de A_{k_p}, qui est semblable à A_0. Il en résulte que $P_D(X) := det(XI - D) = det(XI - A_0)$, est indépendant du choix de D. Les termes de la diagonale de D étant les racines de P_D sont les même que les racines comptées avec multipicités de $P_{A_0} = det(XI - A_0)$ et cela donne au plus $n!$ choix possibles pour D (et même moins en cas de valeurs propres multiples).

iii) On remarque enfin que la suite $D_{k+1} - D_k$ tend vers zéro. En effet on peut extraire des calculs du théorème 5.1, l'évaluation suivante de la diagonale de $B - A$

$$\begin{aligned} b_{p,p} - a_{p,p} &= a_{p,p}c2 + a_{q,q}s^2 - 2a_{p,q}cs - a_{p,p} = (a_{q,q} - a_{p,p})\sin^2\theta - a_{p,q}\sin 2\theta \\ &= 2a_{p,q}\frac{\cos 2\theta}{\sin 2\theta}\sin^2\theta - a_{p,q}\sin 2\theta = -tan\theta.a_{p,q} \end{aligned}$$

Du fait que $|tan\theta| \leq 1$, on en tire $|b_{p,p} - a_{p,p}| \leq |a_{p,q}|$ Appliquée à la k ième itération de l'algorithme de Jacobi on déduit de ce calcul :

$$|a_{p_k,p_k}^{(k+1)} - a_{p_k,p_k}^{(k)}| \leq |a_{p_k,q_k}^{(k)}| \leq |||B_k|||_E$$

On a la même majoration pour $a_{q_k,q_k}^{(k+1)} - a_{q_k,q_k}^{(k)}$, qui est le seul autre coefficient non nul de $D_{k+1} - D_k$ et comme $|||B_k|||_E$ tend vers zéro cela donne le résultat voulu.

Pour conclure que A_k a une limite D, ou ce qui revient au même d'après le pont i) que D_k a une limite D, il reste à appliquer le lemme de topologie suivant, laissé en exercice au lecteur, en remarquant aussi que la suite A_k est bornée (puisque la norme $|||A_k|||_E \geq |||D_k|||_E$ est constante égale à $|||A_0|||_E$) :

Lemme 5.3 *Soit u_n une suite dans un espace vectoriel normé de dimension finie telle que*

$$\lim_{k\to\infty}(u_{k+1}-u_k)=0$$

et n'ayant qu'un nombre fini de valeurs d'adhérences. Alors la suite u_k est convergente ou tend en norme vers l'infini.

Indication sur le lemme: Soient $l_1,\cdots,l_q$ les valeurs d'adhérence de la suite u_n. On choisit $\epsilon>0$ tel que quelque soit la paire d'indices (i,j) distincts on a $|l_i-l_j|\geq 3\epsilon$, et $R>0$, tel que la boule fermée $\overline{B}=\overline{B}(0,R)$ contient toutes les boules $\overline{B}(l_i,2\epsilon)$.

Considérons alors le compact $K=\overline{B}(0,R)\setminus B(l_i,\epsilon)$. On montre facilement qu'il existe n_0, tel que $n\geq n_0\Rightarrow u_n\notin K$. On en tire l'alternative :

$$\big[q=1 \text{ et } \lim_{k\to\infty}u_k=l_1\big] \quad \text{ou} \quad \big[q=0 \text{ et } \lim_{k\to\infty}\|u_k\|=+\infty\big]$$

□