`Séance du 07/02/2000`

1 Équations numériques

1.1 Introduction

L'objet de cette section est d'étudier les équations de la forme $f(x) = 0$ où f est une fonction numérique et x une variable (ou n-uplet de variables) le plus souvent réelle(s), et parfois complexe(s). Se présentent immédiatement deux possibilités : soit déterminer une formule exacte pour ces racines, soit les calculer de façon approchée.

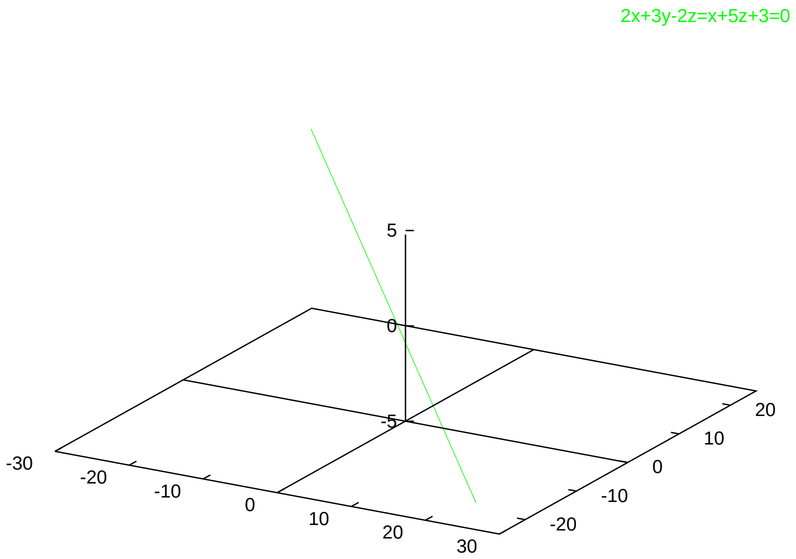
Encore faut-il donner un sens à ces deux questions. La première est affaire de contexte et son sens varie donc selon les besoins. La définition implicite $f(x) = 0$ est parfois la seule chose que l'on désire savoir sur x, surtout si l'on décrit ainsi une famille (une droite, une parabole etc.) mais on aime parfois avoir une représentation paramétrique ou une formule que l'on estime calculable, c'est-à-dire un peu plus explicite que $x = f^{-1}(0)$ (si tant est que cette dernière écriture a un sens). Par exemple dans l'étude des systèmes linéaires, on cherche à décrire les sous-espaces affines d'un certain espace vectoriel :

$$(x, y, z) \in \mathbf{R}^3 \qquad \text{tel que} \qquad \left\{ \begin{array}{rcl} 2x + 3y - 2z & = & 0 \\ x + 5z & = & -3 \end{array} \right.$$

est une équation d'une certaine droite de l'espace. On pourrait se contenter de cette description, et d'ailleurs on le fait dans de nombreux cas. Évidemment quand le sous-espace affine se réduit à un point, c'est-à-dire quand le système linéaire admet une unique solution, on préfère décrire le point par ses coordonnées que par les équations qu'il vérifie. Ce que l'on appelle résoudre un système linéaire au lycée ou en premier cycle, c'est trouver une paramétrisation linéaire de l'ensemble des solutions :

$$(x, y, z) \in \mathbf{R}^3 \qquad \text{tel que} \qquad x = -5z - 3 \quad \text{et} \quad y = 4z + 2$$

dans l'exemple précédent. On a ici le sentiment d'avoir effectivement trouvé quelque chose de plus simple que le système de départ, mais si on étudie un espace défini par trois équations dans $\mathbf{R}^9$, est-il plus simple ou plus compliqué de le décrire par des équations paramétriques en fonction de six variables libres ? Le point essentiel est, en fait, que la représentation paramétrique permet un calcul et donc un tracé des solutions : il suffit de fixer les paramètres et de calculer. C'est donc cela le sens de la question « trouver une formule exacte pour les racines de l'équation » dans ce cas là.



Mais parfois il est plus ardu de se convaincre que l'on a vraiment « calculé » les solutions. Les mathématiques sont truffées de notations pour des objets qui deviennent peu à peu partie intégrante du langage. On convient aisément de la notation pour les nombres entiers ou rationnels. Mais quand viennent les nombres réels (comme e ou π) l'affaire se complique. Même en restant avec $\sqrt{2}$, qu'est-ce d'autre qu'une notation pour l'unique solution positive de $x^2 - 2 = 0$? On a donc envie de se convaincre que l'on peut « toucher » $\sqrt{2}$ par exemple en l'encadrant par des rationnels de plus en plus proches.

Ici survient la seconde problématique. Qu'entend-on par proche et que souhaite-t-on comme approximation ? Ici encore tout est affaire de contexte. Les grecs avaient tendance à préférer les approximations par fractions. Celles-ci s'obtiennent par exemple en faisant tendre une suite de rationnels vers le nombre que l'on veut calculer. Ainsi Archimède (3ème siècle avant Jésus Christ) dans « La mesure du cercle » utilise l'algorithme d'Euclide pour encadrer $\sqrt{2}$. Par la suite Héron d'Alexandrie (1er siècle) donne une méthode pour calculer $\sqrt{a}$ dans « Les métriques ». Pour ce faire il commence par trouver la partie entière b de ce nombre puis utilise la suite récurrente définie par

$$u_0 = b \qquad \text{et} \qquad u_{n+1} = \frac{1}{2}\left(u_n + \frac{a}{u_n}\right)$$

qu'il trouve par une sorte de développement limité (inconnu à l'époque) de $\sqrt{x}$ en 1. François Viète (1540 – 1603) et John Wallis (1616 – 1703) reprennent le procédé pour résoudre certaines équations algébriques et Joseph Louis Lagrange (1736 – 1813) utilise le développement en fractions continues pour ses approximations en théorie des nombres.

Mais l'écriture des nombres dans le système décimal (inventé par Simon Stevin, 1548 – 1620) suggère un autre critère d'approximation, celui fondé sur la distance : le nombre de décimales exactes dans le développement. C'est ainsi qu'au 16ème siècle la méthode de dichotomie est systématiquement employée pour isoler les racines d'un polynôme. Vu sa lenteur de convergence elle est accélérée grâce une méthode d'interpolation dite *regula falsi* (fausse position) développée par François Viète, Johannes Kepler (1571 – 1630) et René Descartes (1596 – 1650).

Mais ce sont les méthodes basées sur des propriétés de point fixe qui sont les plus efficaces, du moins au voisinage de ces points fixes. Leur émergence fut assez longue. Par exemple Al Kashi (environ 1400) améliore les résultats de Claude Ptolémée (100 – 170) concernant le calcul, important en astronomie, de $\sin 1°$ (ce dernier utilisait une interpolation entre $45'$ et $1°30'$ obtenus par dichotomie). Al Kashi part de la valeur en $3°$ obtenue par dichotomie et en cherchant le point fixe de

$$x \mapsto \frac{1}{3}\left(4x^3 + \sin 3°\right) \ .$$

Ce procédé a été repris de nombreuses fois par Johannes Kepler et Isaac Newton (1642 – 1727).

Ce dernier a également mis au point une technique de linéarisation par le calcul différentiel afin de résoudre l'équation de Kepler $x = a + e \sin x$. Cette méthode redonne et généralise les techniques de Héron d'Alexandrie. Elles ont été exposées de façon systématique par J. Raphson en 1690. Cette méthode (dite des tangentes, de Newton ou de Newton-Raphson) est de loin la plus performante quand on sait facilement calculer la dérivée de la fonction considérée.

Se posent maintenant deux questions : peut-on estimer la vitesse de convergence afin de majorer le nombre d'itérations nécessaires pour obtenir une précision donnée et quel est le nombre de calculs nécessaires à chaque étape ? Il est à noter qu'une méthode rapidement convergente peut être catastrophique du point de vue des performances à cause d'un coût élevé en calculs. Par exemple la trichotomie (technique consistant à diviser un intervalle en trois pour localiser une racine) a une vitesse de convergence en 3^{-n} et semble donc plus rapide que la dichotomie qui affiche 2^{-n}, mais l'implémentation de la trichotomie est bien plus lourde et on lui préfère tout naturellement la dichotomie.

C'est Augustin Louis Cauchy (1789 – 1857) qui a précisé de façon rigoureuse la convergence des méthodes de Newton et du point fixe, mais c'est Émile Picard (1856 – 1941) qui a intégré dans un même schéma ces deux méthodes et a systématisé leur utilisation notamment dans le domaine des équations différentielles.

1.2 Dichotomie

Elle est fondée sur le théorème des valeurs intermédiaires. Soit f une fonction numérique continue définie sur un intervalle I de $\mathbf{R}$. Si on se donne deux points u et v tels que $f(u)f(v)$ soit négatif, alors f possède un zéro entre u et v. Malheureusement cela ne garantit pas qu'il y ait unicité de ce zéro ni qu'il ne puisse exister de zéro quand cette hypothèse n'est pas vérifiée. Pour pouvoir être plus précis il faut supposer f de classe C^2

sur I et telle que f et f' n'ont pas de zéro en commun. Dans ce cas on a les résultats suivants :

Lemme 1 *Si* $[a, b]$ *est un intervalle contenu dans* I *et si*

$$\left| f\left(\frac{a+b}{2}\right) \right| > \frac{b-a}{2} \sup_{a \leq t \leq b} |f'(t)|$$

alors f *n'admet pas de zéro sur* $[a, b]$.

En effet cela résulte de l'inégalité des accroissements finis. Si x appartient à $[a, b]$, on a

$$\begin{aligned}
|f(x)| &\geq \left| f\left(\frac{a+b}{2}\right) \right| - \left| f(x) - f\left(\frac{a+b}{2}\right) \right| \\
&\geq \left| f\left(\frac{a+b}{2}\right) \right| - \left| x - \frac{a+b}{2} \right| \sup_{a \leq t \leq b} |f'(t)| \\
&\geq \left| f\left(\frac{a+b}{2}\right) \right| - \frac{b-a}{2} \sup_{a \leq t \leq b} |f'(t)| \\
&> 0
\end{aligned}$$

Lemme 2 *Si* $[a, b]$ *est un intervalle contenu dans* I *et si*

$$\left| f'\left(\frac{a+b}{2}\right) \right| > \frac{b-a}{2} \sup_{a \leq t \leq b} |f''(t)|$$

alors f *admet au plus un unique zéro sur* $[a, b]$.

En effet, d'après le lemme précédent appliqué à f', celle-ci ne s'annule pas sur $[a, b]$ et donc f y est strictement monotone. Le lemme en résulte.

Lemme 3 *Le nombre de zéros de* f *sur* I *est fini.*

Montrons tout d'abord que les zéros de f sont isolés, c'est-à-dire que si x est un zéro de f, il existe un intervalle I_x centré en x tel que f admette x comme unique zéro. En effet si x est un zéro de f, alors $f'(x)$ est non nul par hypothèse sur f et donc f est localement strictement monotone au voisinage de x : il existe un intervalle I_x centré en x tel que f' soit de signe constant sur I_x et donc f n'y admet que x comme zéro.

Supposons maintenant que f ait une infinité de zéros sur I. Par compacité de cet intervalle, on pourrait donc construire une suite $(x_n)_{n \in \mathbf{N}}$ de zéros de f tous distincts et convergente dans I. Notons x sa limite. Par continuité de f sur I, $f(x) = f(\lim x_n) = \lim f(x_n) = 0$ et donc, d'après ce qui précède il existe un intervalle I_x centré en x tel que x soit l'unique zéro de f sur I_x. Ceci est en contradiction avec le fait que x soit la limite de $(x_n)_{n \in \mathbf{N}}$.

Lemme 4 *Soit Z l'ensemble des zéros de f, $m_1 = \min_{x\in Z}|f'(x)|$ et $M_2 = \sup_{t\in I}|f''(t)|$. Pour tout x dans Z et tout t dans I*

$$M_2|t-x| \leq \frac{m_1}{2} \Rightarrow |f'(t)| \geq \frac{m_1}{2}\ .$$

Cela résulte immédiatement de l'inégalité des accroissements finis appliquée à f' entre x et t. En effet si x est un zéro de f et t est dans I, on a

$$|f'(t)| \geq |f'(x)| - |t-x|M_2 \geq m - |t-x|M_2$$

et le lemme en résulte.

Lemme 5 *Soit α et β des majorants de $|f'|$ et $|f''|$ sur un intervalle $[a,b]$ inclus dans I. Il existe un entier n tel que, pour tout point de $[a,b]$ l'une au moins des deux conditions suivantes sont vérifiées :*

1. $|f(x)| > \alpha(b-a)2^{-n}$
2. $|f'(x)| > \beta(b-a)2^{-n}$

Si M_2 est nul (i.e. f est affine) il n'y a rien à démontrer (par exemple en utilisant le lemme précédent). On se place donc dans le cas où M_2 n'est pas nul.

On note J la réunion des intervalles $]x - m_1/2M_2, x + m_1/2M_2[\cap[a,b]$ pour x dans Z et K son complémentaire dans $[a,b]$.

K est compact en tant que fermé dans l'intervalle compact $[a,b]$ et donc la fonction continue $|f|$ y atteint son minimum. Ce dernier n'est pas nul par définition de Z. Notons-le m.

Soit maintenant n tel que $\beta(b-a)2^{-n}$ soit inférieur à $m_1/2$ et $\alpha(b-a)2^{-n}$ soit inférieur à m. D'après le lemme précédent, si t appartient à J, $|f'(x)|$ est supérieur à $m_1/2$ et s'il appartient à K, $|f(x)|$ est supérieur à m. Le lemme en résulte.

Théorème 1 *Soit α, β et n comme dans le lemme précédent. On note $(c_k)_{0\leq k\leq n}$ la subdivision de I de pas constant égal à $(b-a)2^{-n}$; f possède au plus un zéro dans chacun des intervalles $[c_k, c_{k+1}]$ (pour $0 \leq k < n$).*

Ceci résulte immédiatement de la définition de n et des lemmes 1 et 2.

Algorithme 1 *On détermine α et β empiriquement. Si $c = (a+b)/2$ vérifie l'une des relations 1 ou 2, le signe de $f(a)f(b)$ détermine le nombre de zéros de f sur $[a,b]$. Sinon on subdivise $[a,b]$ en deux : $[a,c]$ et $[c,b]$. Sur chacun de ces intervalles on détermine éventuellement de nouvelles valeurs pour α et β et on recommence sur chacun d'eux le test précédent (en $(a+c)/2$ et $(c+b)/2$ respectivement). Le théorème assure qu'en au plus n subdivisions tous les milieux vérifieront le test.*

1.3 Regula falsi

On suppose connu un encadrement $]a,b[$ d'un zéro x de f de telle sorte que f' soit strictement positive sur $[a,b]$. On suppose que f n'est pas un polynôme du premier degré et on l'approxime par un tel polynôme. Autrement dit on résout l'équation

$$f(a) + (f(b) - f(a))\frac{t-a}{b-a} = 0\ .$$

On suppose toujours f de classe C^2 sur $[a,b]$.

Soit $P = P_{a,b}$ le polynôme de degré 1 prenant les mêmes valeurs que f en a et en b ; on a donc

$$P(t) = \frac{(t-a)f(b) - (t-b)f(a)}{b-a}\ .$$

Lemme 6 *Pour tout t dans $[a,b]$, il existe ξ_t dans $[a,b]$ tel que*

$$f(t) - P(t) = \frac{1}{2}f''(\xi_t)(t-a)(t-b)\ .$$

Pour t égal à a ou b, c'est clair. Sinon on peut trouver une constante y telle que $u \mapsto f(u) - P(u) - y(u-a)(u-b)$ s'annule en t. Comme cette fonction s'annule déjà en a et b, sa dérivée admet d'après le théorème de Rolle un zéro entre a et t et un autre entre t et b. Par conséquent sa dérivée seconde s'annule entre a et b, toujours d'après le théorème de Rolle. Mais cette dérivée seconde n'est autre que la fonction $f'' - 2y$. Si ξ_t est son zéro, on a donc $y = f''(\xi_t)/2$.

Lemme 7 *Soit x' la racine de P, il existe deux éléments ξ et ξ' de $[a,b]$ tels que*

$$x' - x = \frac{f''(\xi)}{2f'(\xi')}(x'-a)(x'-b)\ .$$

On a en effet d'après le lemme précédent et l'égalité des accroissements finis appliquée à f entre x et x' :

$$f(x') = P(x') + \frac{1}{2}f''(\xi_{x'})(x'-a)(x'-b) = \frac{1}{2}f''(\xi_{x'})(x'-a)(x'-b)$$

et, pour un certain ξ' compris entre x et x',

$$f(x') = f(x') - f(x) = (x'-x)f'(\xi')\ .$$

D'où le lemme.

Théorème 2 *Soit m_1 un minorant de $|f'|$ sur $[a,b]$ et M_2 un majorant de $|f''|$ également sur $[a,b]$. Soit q donné par*

$$q = \frac{M_2}{8m_1}(b-a)$$

et $(a_n, b_n)_{n\in\mathbf{N}}$ les suites définies par les conditions initiales $a_0 = a$ et $b_0 = b$ ainsi que par la relation de récurrence suivante : soit x' l'unique racine de P_{a_n,b_n} et $\varepsilon = M_2(x' - a_n)(b_n - x')/2m_1$, on pose

$$(a_{n+1}, b_{n+1}) = \begin{cases} (\max(a_n, x' - \varepsilon), x') \\ (x', \min(b_n, x' + \varepsilon)) \end{cases}$$

selon que $f(x')$ est du signe de $f(b_n)$ ou de celui de $f(a_n)$.

Si $q < 1$ alors les suites $(a_n, b_n)_{n\in\mathbf{N}}$ sont adjacentes et convergent vers x. De plus, pour tout entier n, on a

$$|b_n - a_n| \leq \frac{8m_1}{M_2} q^{2^n} .$$

Les deux suites sont bien définies d'après le lemme (puisque x' appartient à $[a, b]$) avec $(a_n)_{n\in\mathbf{N}}$ croissante et $(b_n)_{n\in\mathbf{N}}$ décroissante et on a

$$|b_{n+1} - a_{n+1}| \leq \varepsilon \leq \frac{M_2}{m_1} \frac{(b_n - x')(x' - a_n)}{2} \leq \frac{M_2}{m_1} \frac{(b_n - a_n)^2}{8}$$

et donc

$$|b_{n+1} - a_{n+1}| \leq \frac{8m_1}{M_2} q^{2^n}$$

par une récurrence immédiate. Les deux suites sont donc bien adjacentes. Comme x appartient à tous les intervalles $[a_n, b_n]$, c'est donc leur limite commune.

1.4 La méthode de Newton-Raphson

Cette fois-ci on remplace f par un développement limité à l'ordre 1, autrement dit on résout

$$f(c) + (x - c)f'(c) = 0 .$$

On se place toujours dans le cas où f est de classe C^2 sur un intervalle I avec f' strictement positive et f'' non nulle (i.e. f n'est pas affine).

Théorème 3 *Soit m_1 le minimum de $|f'|$ sur I et M_2 le maximum de $|f''|$ sur ce même intervalle. On suppose que l'on a trouvé un point x_0 tel que $|f(x_0)|$ soit inférieur à $m_1^2/4M_2$. Dans ces conditions il existe un unique zéro x de f dans l'intervalle $[x_0 - c, x_0 + c]$ pour $c = 2|f(x_0)|/m_1$ et la suite définie par la condition initiale x_0 et la relation de récurrence*

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

converge vers x. De plus, pour tout entier n, on a

$$|x - x_n| \leq \frac{2m_1}{M_2} \left(\frac{f(x_0)M_2}{m_1^2} \right)^{2^n} .$$

Montrons par récurrence sur l'entier naturel non nul n que $|x_n - x_{n-1}|$ est inférieur à $c2^{-n}$, $|f(x_{n-1})|$ est inférieur à $cm_1 2^{-n}$ et x_n appartient à $[x_0 - c, x_0 + c]$.

Pouur $n = 1$, on a $|f(x_0)| \leq cm_1/2 = |f(x_0)|$ par construction et

$$|x_1 - x_0| = \left|\frac{f(x_0)}{f'(x_0)}\right| \leq \frac{|f(x_0)|}{m_1} = \frac{c}{2} .$$

Montrons maintenant que la propriété est héréditaire : on a

$$f(x_n) = f(x_n) - f(x_{n-1}) - (x_n - x_{n-1})f'(x_{n-1})$$

et donc

$$|f(x_n)| \leq |x_n - x_{n-1}|^2 \frac{M_2}{2} \leq |x_n - x_{n-1}|cM_2 \leq \frac{|x_n - x_{n-1}|m_1}{2} \leq \frac{cm_1}{2^{n+1}} .$$

On a aussi

$$|x_{n+1} - x_n| = \left|\frac{f(x_n)}{f'(x_n)}\right| \leq \frac{c}{2^{n+1}} .$$

Il en résulte $|x_{n+1} - x_0| \leq c$.

De plus, la série $\sum(x_n - x_{n-1})$ est donc absolument convergente et la convergence de la suite $(x_n)_{n\in\mathbf{N}}$ en découle. Par continuité de f sa limite est nécessairement un zéro de f. Si y était un autre zéro de f, on aurait pour tout entier n

$$|f'(x_n)(y - x_{n+1})| = |f(y) - f(x_n) - f'(x_n)(y - x_n)| \leq \frac{|y - x_n|m_1}{2}$$

et donc $|y - x_{n+1}| \leq |y - x_n|/2$. En faisant tendre n vers l'infini y est donc la limite de $(x_n)_{n\in\mathbf{N}}$.

Enfin en reprenant le calcul précédent, on a

$$|x - x_{n+1}| \leq \frac{|x - x_n|^2 M_2}{2m_1}$$

et donc, pour tout entier naturel n

$$|x - x_n| \leq \frac{2m_1}{M_2}\left(\frac{f(x_0)M_2}{m_1^2}\right)^{2^n} .$$

Séance du 14/02/2000

2 Équations polynomiales

Dans le domaine des équations polynomiales en nombres entiers, rationnels, réels ou complexes, on a longtemps cherché des solutions données par des formules n'autorisant que les quatre opérations élémentaires et la prise de radicaux. Si l'équation du deuxième degré est facilement résolue, il n'en est pas de même pour les équations de degré supérieur. C'est Gerolamo Cardano (1501 – 1576) qui donne une solution pour le degré trois dans son *Artis magnœ sive de regulis algebraicis*. Comme on le comprendra plus tard, il est obligé de sortir du champ réel même (et surtout) pour résoudre les équations à coefficients réels et trois racines réelles. Puis c'est son disciple, Ludivico Ferrari (1522 – 1565) qui obtient la solution du degré quatre.

Mais tout ceci ne résout pas tout. En effet, numériquement, encore faut-il savoir extraire des radicaux et, surtout, ce n'est pas nécessairement la meilleure méthode, surtout pour les équations de degré au moins 5, car on sait que celles-ci ne sont pas, en toute généralité, résoluble par radicaux.

Le premier problème numérique est d'isoler les racines. Si la méthode de dichotomie marche très bien pour les polynômes, il en existe d'autres plus spécifiques inventées par René Descartes (*La géométrie* 1638), Isaac Newton (*L'aritmétique universelle* 1685), Michel Rolle (1652 – 1719), James Stirling (1692 – 1770), Joseph Fourier (1768 – 1830), Augustin-Louis Cauchy, Charles Sturm (1803 – 1855) et Charles Hermite (1822 – 1901).

2.1 Majoration du module des racines d'un polynôme

Lemme 8 *Soit $P = \sum_{k=0}^{n} a_k X^k$ un polynôme à coefficients réels, unitaire, à coefficient constant non nul et dont tous les coefficients (hormis le coefficient dominant) sont négatifs ou nuls. Alors toutes les racines de P sont de module inférieur à*

$$1 + \sup_{0 \leq k < \deg P} |a_k| \qquad \textit{et à} \qquad \max\left(1, \sum_{0 \leq k < \deg P} |a_k|\right) .$$

Montrons tout d'abord que P a une unique racine réelle strictement positive. En effet x dans $\mathbf{R}_+^*$ est racine de P si et seulement s'il est un zéro de $f(x) = x^{-n}P(x) = \sum_{k=0}^{n} a_k x^{k-n}$. Or cette fontion est strictement croissante, admet $-\infty$ comme limite à droite en 0 et 1 comme limite en $+\infty$ et ceci assure l'existence et l'unicité de la racine strictement positive ρ de P.

Remarquons que, grâce à l'inégalité triangulaire et au signe des coefficients de P, on a pour tout complexe z

$$|P(z)| \geq P(|z|)$$

et par conséquent toute racine de P est de module inférieur à ρ. Pour conclure il nous suffit donc de démontrer que ρ est majoré par $1+\sup_{0\leq k<\deg P}|a_k|$ et $\max\Big(1, \sum_{0\leq k<\deg P}|a_k|\Big)$.

Remarquons que

$$\begin{aligned}
\left|\sum_{k=0}^{n-1} a_k(1+\sup_{0\leq j<\deg P}|a_j|)^k\right| &= \sum_{k=0}^{n-1}|a_k|(1+\sup_{0\leq j<\deg P}|a_j|)^k \\
&\leq \sum_{k=0}^{n-1}\sup_{0\leq j<\deg P}|a_j|(1+\sup_{0\leq j<\deg P}|a_j|)^k \\
&\leq \sum_{k=0}^{n-1}\left((1+\sup_{0\leq j<\deg P}|a_j|)^{k+1}-(1+\sup_{0\leq j<\deg P}|a_j|)^k\right) \\
&\leq (1+\sup_{0\leq j<\deg P}|a_j|)^n - 1 \\
&< (1+\sup_{0\leq j<\deg P}|a_j|)^n \ .
\end{aligned}$$

Par conséquent P prend une valeur positive en le réel positif $1+\sup_{0\leq k<\deg P}|a_k|$. Il en résulte que ρ est inférieur à $1+\sup_{0\leq k<\deg P}|a_k|$.

De même

$$\begin{aligned}
\left|\sum_{k=0}^{n-1} a_k \max\left(1, \sum_{0\leq k<\deg P}|a_k|\right)^k\right| &= \sum_{k=0}^{n-1}|a_k|\max\left(1, \sum_{0\leq k<\deg P}|a_k|\right)^k \\
&\leq \sum_{k=0}^{n-1}|a_k|\max\left(1, \sum_{0\leq k<\deg P}|a_k|\right)^{n-1} \\
&\leq \max\left(1, \sum_{0\leq k<\deg P}|a_k|\right)^n
\end{aligned}$$

et donc $P(\max\Big(1, \sum_{0\leq k<\deg P}|a_k|\Big))$ est strictement positif, i.e. ρ est inférieur au maximum de 1 et $\sum_{0\leq k<\deg P}|a_k|$. Ceci achève la démonstration du lemme.

Théorème 4 *Si $P=\sum_{k=0}^n a_kX^k$ un polynôme à coefficients complexes, unitaire, à coefficient constant non nul, alors toutes les racines de P sont de module inférieur à*

$$1+\sup_{0\leq k<\deg P}|a_k| \qquad \textit{et à} \qquad \max\left(1, \sum_{0\leq k<\deg P}|a_k|\right) \ .$$

On considère Q le polynôme donné par

$$Q(X) = X^n - \sum_{k=0}^{n-1}|a_k|X^k \ .$$

Si z est une racine de P, on a

$$|z|^n = \left|\sum_{k=0}^{n-1} a_k z^k\right| \leq \sum_{k=0}^{n-1} |a_k|.|z|^k$$

et donc $Q(|z|)$ est négatif. Par conséquent z est de module inférieur à la plus grande racine de Q. Le théorème résulte donc du lemme précédent.

Algorithme 2 (Méthode de Lagrange) *Si P est un polynôme à coefficients réels et à racines simples $(x_i)_{1\leq i\leq n}$, il existe un polynôme unitaire Q dont les racines sont les nombres $((x_i - x_j)^2)_{1\leq i,j\leq n}$. Les coefficients de Q s'expriment de façon rationnelle en fonction des coefficients de P. Si δ est un majorant des racines de $X^{n(n-1)/2}Q(1/X)$, alors les racines de P sont espacées d'au moins $\delta^{-1/2}$.*

En effet Q est un polynôme symétrique en les racines de P et ses coefficients sont donc des fractions rationnelles en les coefficients de P d'après les formules de Newton. Si δ majore les racines de $X^{n(n-1)}/2Q(1/X)$, alors il majore les $(x_i - x_j)^{-2}$ et le résultat en découle.

2.2 Règle de Descartes

Soit P un polynôme à coefficients réels, donné sous la forme

$$P(X) = \sum_{k=0}^{n} a_k X^{b_k}$$

avec $0 = b_0 < b_1 < \ldots < b_n$ et les a_k tous non nuls. On note $V(P)$ le nombre de variations de signes des coefficients de P. Autrement dit

$$V(P) = \mathrm{card}\,\{0 \leq k < n \mid a_k a_{k+1} < 0\} \ .$$

Théorème 5 *Le nombre de racines réelles (comptées avec multiplicité) strictement positives de P, noté $n_+(P)$ est majoré par $V(P)$. Il lui est égal si P a toutes ses racines réelles.*

Montrons la partie majoration de ce théorème par récurrence sur n, le nombre de monômes non nuls apparaissant dans P.

Si n vaut 1, P est un polynôme constant non nul. On a donc $n_+(P) = 0 = V(P)$.

Supposons n supérieur ou égal à 2 et que P a au moins une racine strictement positive (sinon il n'y a rien à démontrer). Notons $(x_1, \ldots, x_j)$ les zéros de P sur $\mathbf{R}_+^*$ et $(m_1, \ldots, m_j)$ leurs multiplicités. On a

$$P' = X^{b_1-1}\left(b_1 a_1 + b_2 a_2 X^{b_2-b_1} + \ldots + b_n a_n X^{b_n-b_1}\right) \ .$$

Posons $Q = \sum_{k=1}^{n} b_k a_k X^{b_k - b_1}$. Puisque tous les b_k sont positifs, on a $V(Q) = V(P) - 1$ ou $V(Q) = V(P)$ selon que $a_0 a_1$ est négatif ou non.

Si l'on suppose la propriété de récurrence vraie à l'ordre $n-1$, il en résulte que Q et donc P' admet au plus $V(P)-1$ ou $V(P)$ racines strictement positives selon que $a_0 a_1$ est négatif ou non.

Si $a_0 a_1$ est positif, $|P|$ est croissant au voisinage de 0. Par conséquent P' s'annule sur $]0; x_1[$. Comme il s'annule entre deux racines consécutives de P, il a au moins j racines sur $\mathbf{R}_+^*$ en dehors des x_j. Ce qui lui fait au moins $j + \sum_{k=1}^{j}(m_k - 1)$, soit $\sum_{k=1}^{j} m_k$, racines. On a donc $\sum_{k=1}^{j} m_k \leq V(P)$.

Si au contraire $a_0 a_1$ est négatif, le raisonnement précédent garantit au moins $j - 1 + \sum_{k=1}^{j}(m_k - 1)$ racines pour P' et donc $-1 + \sum_{k=1}^{j} m_k \leq V(P) - 1$. D'où le résultat.

En appliquant ce résultat à $R(X) = P(-X)$, on en déduit que le nombre $n_-(P)$ de racines négatives (comptées avec multiplicité) de P est inférieur à $V(R)$. Notons c_k les coefficients de R, on a $c_k = (-1)^{b_k} a_k$ et donc $c_k c_{k-1} = (-1)^{b_k - b_{k-1}} a_k a_{k-1}$. Par conséquent si, pour un certain indice k, $a_k a_{k-1}$ et $c_k c_{k-1}$ sont négatifs, alors b_k est supérieur à $b_{k-1}+2$. Notons $V(P,R)$ le nombre de ces indices, on a

$$b_n = \sum_{k=1}^{n}(b_k - b_{k-1}) \geq (V(P) - V(P,R)) + (V(R) - V(P,R)) + 2V(P,R) = V(P) + V(R)$$

en discutant suivant que l'indice k est tel que $a_k a_{k-1} < 0$, $c_k c_{k-1} < 0$ ou les deux à la fois.

Par conséquent si P a toutes ses racines réelles

$$V(P) + V(R) \geq n_+(P) + n_-(P) = \deg(P) \geq V(P) + V(R)$$

et il y a forcément égalité dans toutes ces inégalités. En particulier $n_+(P)$ est égal à $V(P)$.

2.3 Suites de Sturm

Soit P un polynôme à coefficients réels de degré n (strictement positif) et x un réel. On note $a(x)$ la suite $(P(x), P'(x), \ldots, P^{(n)}(x))$ à laquelle on a retiré les termes nuls ainsi que $v_x(P)$ le nombre de variations de signes dans la suite $a(x)$. Autrement dit

$$\operatorname{card}\left\{0 \leq k < n \mid \exists \ell \in [k; n]\ P^{(k)}(x) P^{(\ell)}(x) < 0 \text{ et } P^{(k+1)}(x) = \ldots = P^{(\ell-1)}(x) = 0\right\} .$$

Théorème 6 (Fourier-Baudan) *Soit $I = [a, b]$ un intervalle de $\mathbf{R}$. Si a et b ne sont pas racines de P, le nombre de racines (comptées avec multiplicité) de P dans I est majoré par $v_a(P) - v_b(P)$. Si de plus P n'a que des racines réelles, il y a en fait égalité.*

Soit Z l'ensemble de tous les zéros de P et de ses polynômes dérivés. C'est un ensemble fini. Notons $z_1 < z_2 < \ldots < z_p$ ses éléments. Pour tout indice j entre 1 et $p-1$, ni P ni aucune de ses dérivées ne s'annulent sur $]z_j, z_{j+1}[$ et ils y gardent donc tous un signe constant. Par conséquent $v_x(P)$ est constante sur cet intervalle.

Pour $x > z_p$, toutes ces fonctions sont du signe de leur coefficient dominant et on a $v_x(P) = 0$. Par contre pour x négatif et inférieur à z_1, $P^{(k)}$ est du signe de $(-1)^{n-k}$ fois le coefficient dominant et donc $v_x(P) = n$.

Étudions maintenant ce qu'il se passe en un élément c de Z. Soit k et l deux indices tels que

$$P^{(k)}(c) = a_k \neq 0\,, \qquad P^{(l)}(c) = a_l \neq 0\,, \qquad \text{et si } k < j < l \qquad P^{(j)}(c) = 0\;.$$

D'après la formule de Taylor, on a au voisinage de c, pour $k < j \leq l$,

$$P^{(j)}(x) = \frac{a_l}{(l-j)!}(x-c)^{l-j}\;.$$

Par conséquent, si $c = z_i$, pour x dans $]c, z_{i+1}[$ (ou $x > c$ quand $i = p$), la suite $P^{(j)}(x)$ change au plus une fois de signe pour $k \leq j \leq l$ et ce uniquement si $a_k a_l$ est négatif.

Par contre pour x dans $]z_{i-1}, c[$ (ou $x < c$ si $i = 1$), elle change de signe $l - k - 1$ fois pour $k < j \leq l$ et une fois de plus entre k et $k+1$ si $a_k a_l$ est du signe de $(-1)^{l-k}$. La différence de variations est donc au moins $l - k - 2$ ce qui est positif sauf peut-être si $l = k+1$. Dans ce dernier cas il ne se passe rien en c pour $k \leq j \leq l$ et on en conclut que la différence de variation entre à gauche de c et à droite de c est positive, pour $k \leq j \leq l$. Par conséquent si c n'est pas un zéro de P, on peut découper $\{0, \ldots, n\}$ entre morceaux du type précédent (car $P(c) \neq 0$ et $P^{(n)}(c) \neq 0$) et

$$\lim_{x\to c, x<c} v_x(P) \geq \lim_{x\to c, x>c} v_x(P)\;.$$

Si maintenant c est un zéro de P, il faut considérer sa multiplicité m. Dans ce cas $P^{(j)}(c) = 0$ pour $j < m$ et $P^{(m)}(c) \neq 0$. La formule de Taylor montre que la suite $P^{(j)}(x)$ pour $j < m$ est de signe constant si x est supérieur à c (et proche), et change m fois de signe si x est inférieur à c (toujours en étant proche). Il en résulte qu'en toute généralité

$$\lim_{x\to c, x<c} v_x(P) - \lim_{x\to c, x>c} v_x(P) \geq m_P(c)$$

où $m_P(c)$ dénote la multiplicité de la racine c dans P (par convention c'est nul si c n'est pas racine de P).

Pour conclure il ne reste qu'à remarquer que si $z_i < \ldots < z_j$ sont les éléments de Z compris entre a et b, on a

$$v_a(P) - v_b(P) = \sum_{k=i}^{j}\left(\lim_{x\to z_k, x<z_k} v_x(P) - \lim_{x\to z_k, x>z_k} v_x(P)\right) \geq \sum_{k=i}^{j} m_P(z_k)\,,$$

ce qui est exactement le résultat annoncé.

En ce qui concerne le cas d'égalité, il suffit d'appliquer le résultat à un intervalle contenant Z pour obtenir l'égalité dans toutes les inégalités précédentes :

$$\lim_{x\to z_k, x<z_k} v_x(P) - \lim_{x\to z_k, x>z_k} v_x(P) = m_P(z_k)$$

et le théorème en résulte.

Soit $(P_k)_{0\leq k\leq n}$ la suite de polynômes définie par $P_0 = P$, $P_1 = -P'$ et, pour $1 \leq k \leq n-1$, P_{k+1} est l'opposé du reste dans la division euclidienne de P_{k-1} par P_k. Soit m le plus grand indice tel que P_m ne soit pas le polynôme nul.

Lemme 9 *P_m est le pgcd de P et P' et il divise tous les termes de la suite $(P_k)_{0\leq k\leq m}$.*

Au signe près la suite fournit l'algorithme d'Euclide de recherche de pgcd de deux polynômes. On a en fait $pgcd(P_k, P_{k+1}) = pgcd(P_{k-1}, P_k)$ pour tout indice k compris entre 1 et m. Notons A ce pgcd commun. Comme $P_{m+1} = 0$, on a $A = pgcd(P_m, P_{m+1}) = P_m$.

On note Q_k le polynôme P_k/P_m pour k compris entre 0 et m et $w_x(P)$ le nombre de variations de la suite $(Q_k(x))_{0\leq k\leq n}$ à laquelle on a retiré les termes nuls.

Théorème 7 (Sturm) *Soit $I = [a, b[$ un intervalle de $\mathbf{R}$. Le nombre de racines (comptées sans multiplicité) de P dans I est égal à $w_b(P) - w_a(P)$.*

Supposons tout d'abord que P n'a que des racines simples (i.e. P_m est un polynôme constant). Dans ces conditions

1. Q_m ne s'annule pas.
2. Si c est une racine réelle de P, $Q_0'(c)Q_1(c) < 0$.
3. Si c est une racine réelle de Q_k, pour $1 \leq k \leq m-1$, $Q_{k-1}(c)Q_{k+1}(c) < 0$.

Le premier point est clair. La second résulte de $Q_0'Q_1 = -(P')^2/P_m^2$ et le dernier de $Q_{k-1}(c) = -Q_{k+1}(c)$ et du fait que Q_k et Q_{k+1} sont premiers entre eux (i.e. c ne peut être racine de Q_{k+1} s'il l'est déjà de Q_k).

Soit Z l'ensemble fini des zéros de tous les Q_k, notés $z_1 < \ldots < z_p$. Comme précédemment, la fonction $w_x(P)$ est constante sur les $]z_i, z_{i+1}[$.

Si c n'est pas un zéro de P et un zéro de Q_k pour un certain k, on a $Q_{k-1}(c)Q_{k+1}(c) < 0$ et le nombre de variations de la suite ne varie pas entre la gauche de c et la droite de c, comme on l'a déjà vu précédemment.

Si par contre c est un zéro de P, au voisinage de c, $Q_0(x)$ est équivalente à $Q_0'(c)(x-c)$ et le nombre de variations augmente donc de 1 quand on passe au-dessus de c :

$$\lim_{x\to c, x>c} w_x(P) - \lim_{x\to c, x<c} w_x(P) = 1 .$$

Plaçons maintenant dans le cas des racines multiples et montrons que les trois propriétés dégagées dans la démonstration précédentes sont encore vraies. La première est claire.

Si $P_{k-1} = A_k P_k - P_{k+1}$, alors $Q_{k-1} = A_k Q_k - Q_{k+1}$ et donc les polynômes Q_k et Q_{k+1} sont premiers entre eux pour tout indice k entre 0 et $m-1$. La troisième propriété est alors immédiate.

Si maintenant c est racine de P (i.e. de Q_0) de multiplicité k, alors on peut écrire

$$Q_1 = -\frac{P'}{P_m} = -\frac{P'Q_0}{P} \qquad \text{et} \qquad Q_0'Q_1 = -Q_0'\frac{Q}{X-c}\frac{(X-c)P'}{P} \sim -kQ_0'(c)^2 \ .$$

La seconde propriété est donc claire aussi.

Séance du 21/02/2000

2.4 La méthode Newton pour les polynômes

Soit P un polynôme à coefficients réels, de degré n strictement positif, ayant toutes ses racines réelles :

$$P(X) = a_n X^n + a_{n-1} X^{n-1} + \ldots + a_1 X + a_0 = a_n \prod_{i=1}^{p} (X - \lambda_i)^{m_i}$$

où $\lambda_1 > \ldots > \lambda_p$ sont les racines (distinctes) de P.

Lemme 10 *Soit Q un polynôme à coefficients réel et λ une racine de Q. On appelle multiplicité de λ (dans Q) la puissance maximale de $X - \lambda$ qui divise Q. La multiplicité de λ est m si et seulement si*

$$Q(\lambda) = Q'(\lambda) = \ldots = Q^{(m-1)}(\lambda) = 0$$

et

$$Q^{(m)}(\lambda) \neq 0 \ .$$

Remarquons tout d'abord que λ est racine d'un polynôme Q si et seulement si $X - \lambda$ divise Q. En effet, on peut effectuer la division euclidienne de Q par $X - \lambda$ dans l'anneau $\mathbf{R}[X]$ et on a $Q = A(X - \lambda) + B$ où B est un polynôme de degré strictement inférieur à celui de $X - \lambda$, i.e. un polynôme constant. En spécialisant cette égalité en λ, on a $Q(\lambda) = B$. Il y a donc équivalence entre $B = 0$, i.e. $X - \lambda$ divise Q, et $Q(\lambda) = 0$.

Appliquons maintenant la formule de Taylor avec reste intégral au polynôme Q. On a, pour tout réel x,

$$\begin{aligned} Q(x) &= Q(\lambda) + (x - \lambda)Q'(\lambda) + \ldots + \frac{(x-\lambda)^{m-1}}{(m-1)!} Q^{(m-1)}(\lambda) \\ &\quad + (x-\lambda)^m \int_0^1 \frac{(1-t)^{m-1}}{(m-1)!} Q^{(m)}\left(tx + (1-t)\lambda\right) dt \ . \end{aligned}$$

Remarquons que l'intégrand est un polynôme en x et le dernier terme est donc un polynôme.

Par conséquent si $Q^{(k)}(\lambda)$ est nul pour k strictement inférieur à m et $Q^{(m)}(\lambda)$ est non nul, alors $(X - \lambda)^m$ divise Q. Comme de plus, pour x égal à λ, on a

$$\int_0^1 \frac{(1-t)^{m-1}}{(m-1)!} Q^{(m)}\left(tx + (1-t)\lambda\right) dt = \int_0^1 \frac{(1-t)^{m-1}}{(m-1)!} Q^{(m)}(\lambda) dt = \frac{Q^{(m)}(\lambda)}{m!}$$

λ est bien racine de multiplicité m de Q.

Réciproquement si λ est racine de multiplicité m de Q alors Q s'écrit $(X-\lambda)^m\tilde{Q}(X)$ pour un certain polynôme $\tilde{Q}$ ne s'annulant pas en λ. On a donc

$$Q'(X) = (X-\lambda)^{m-1}\left(m\tilde{Q}(X) + (X-\lambda)\tilde{Q}'(X)\right)$$

$$Q''(X) = (X-\lambda)^{m-2}\left(m(m-1)\tilde{Q}(X) + 2m(X-\lambda)\tilde{Q}'(X) + (X-\lambda)^2\tilde{Q}''(X)\right)$$

et, plus généralement, pour k inférieur ou égal à m,

$$Q^{(k)}(X) = (X-\lambda)^{(m-k)}\sum_{l=0}^{k} C_k^l \frac{m!}{(m-l)!}(X-\lambda)^{m-l}\tilde{Q}^{(m-l)}$$

et donc $Q^{(k)}(\lambda)$ est nul pour k strictement inférieur à m tandis que $Q^{(m)}(\lambda)$ vaut $m!\tilde{Q}(\lambda)$ et est donc non nul.

Lemme 11 *Le polynôme P' a également toutes ses racines réelles. Celles-ci sont les racines multiples de P (avec un ordre de multiplicité de moins) ainsi qu'une nouvelle racine simple entre chaque paire de racines successives de P.*

Le théorème de Rolle nous assure de l'existence d'une racine de P' dans chaque intervalle $]\lambda_{i+1};\lambda_i[$, pour i entre 1 et $p-1$. De plus, grâce au lemme précédent, on sait aussi que P' admet λ_i avec la multiplicité $m_i - 1$ (si $m_i - 1$ est nul cela signifie que λ_i n'est pas racine de P').

Au total cela fait au moins $p - 1 + \sum_{i=1}^{p}(m_i - 1) = p - 1 + n - p = n - 1$ racines pour P'. Puisque P' est de degré $n-1$, P' n'a pas d'autres racines et ses racines sont exactement de l'ordre que l'on a trouvé, i.e. les λ_i sont de multiplicité $m_i - 1$ et les μ_i sont des racines simples.

Lemme 12 *Pour k inférieur à n, le polynôme $P^{(k)}$ est de signe strictement positif sur* $]\lambda_1;+\infty[$.

On démontre par récurrence sur l'entier naturel k inférieur à n que toutes les racines de $P^{(k)}$ sont réelles et sont supérieures à λ_1.

Pour $k=0$, c'est l'hypothèse initiale sur P. Si c'est vrai pour $P^{(k)}$ avec $n-k = deg(P^{(k)}) \geq 0$, ça l'est pour $P^{(k+1)}$ en appliquant le lemme précédent à $P^{(k)}$.

Il en résulte qu'aucun des $P^{(k)}$ (pour k inférieur à n) ne s'annule sur $]\lambda_1;+\infty[$ et donc y garde un signe constant égal à celui qu'il a quand la variable tend vers $+\infty$. Ce signe est donc le signe du coefficient dominant de $P^{(k)}$, i.e. $n(n-1)\ldots(n-k+1)$ et on en déduit

$$\forall t > \lambda_1 \ , \qquad P^{(k)}(t) > 0 \ .$$

Lemme 13 *Soit μ_1 la plus grande racine de P' strictement inférieure à λ_1 (si elle n'existe pas, on posera $\mu_i = -\infty$). La fonction définie sur $]\mu_1; +\infty[$ par*

$$\begin{cases} g(t) = t - \frac{P(t)}{P'(t)} & si\ t \neq \lambda_1 \\ g(\lambda_1) = \lambda_1 \end{cases}$$

est de classe C^∞. De plus si n est strictement supérieur à 1, pour $t > \lambda_1$, on a

$$\lambda_1 < g(t) < t\ .$$

On écrit encore P sous la forme $(X - \lambda_1)^{m_1} Q_1(X)$ avec Q ayant des racines toutes strictement inférieures à λ_1. Comme λ_1 est racine de P' avec la multiplicité $m_1 - 1$, on peut aussi écrire P' sous la forme

$$P'(X) = (X - \lambda_i)^{m_i - 1} R_1(X)$$

avec R_1 n'ayant que des racines strictement inférieures à λ_1. Il en résulte, pour $t > \mu_1$ et t distinct de λ_1,

$$g(t) = t - (t - \lambda_1)\frac{Q_1(t)}{R_1(t)}$$

où Q_1/R_1 est une fraction rationnelle sans pôle (ni racine) sur $]\mu_1; +\infty[$. Cette formule est en fait aussi valable en λ_1. Sous cette forme on a clairement affaire à une fonction de classe C^∞ sur $]\mu_1; +\infty[$.

Comme P et P' sont strictement positifs pour $t > \lambda_1$, on a $g(t) < t$ sur ce même domaine.

Si, de plus, n est au moins égal à 2, P'' est également strictement positif pour $t > \lambda_1$ et

$$g' = 1 - \frac{P'^2 - P''P}{P'^2} = \frac{P''P}{P'^2}\ .$$

Donc g' est strictement positive pour $x > \lambda_1$ et g est strictement croissante sur ce domaine. Soit donc $t > \lambda_1$. Pour tout couple (x, y) tel que $\lambda_1 < x < y < t$, on a $g(x) < g(y) < g(t)$ et donc, en passant à la limite quand x tend vers λ_1 par valeurs supérieures, on obtient, par continuité de g,

$$\lambda_1 = g(\lambda_1) \leq g(y) < g(t)\ .$$

D'où l'encadrement désiré.

Remarque : si la multiplicité de λ_1 dans P est supérieure à 2, alors elle est la même dans PP'' et dans $(P')^2$, à savoir $2m_1 - 2$ et donc $g'(\lambda_1) > 0$. Dans ce cas g est donc strictement croissante sur $[\lambda_1; +\infty[$.

Théorème 8 *Soit b un réel quelconque supérieur à λ_1. La suite définie par la relation de récurrence $x_{k+1} = g(x_k)$ et de valeur initiale $x_0 = b$ converge en décroissant vers λ_1.*

Si $n = 1$, g est identiquement égale à λ_1 et il n'y a rien à démontrer. On se place donc dans le cas $n \geq 2$.

On montre par récurrence sur l'entier naturel n que $\lambda_1 < x_{k+1} < x_k$.

Pour $n = 0$, on a $x_0 = b > \lambda_1$ par hypothèse et $x_1 = g(x_0)$ vérifie $\lambda_1 < x_1 < x_0$ d'après le lemme précédent.

Si l'hypothèse est vraie au rang k, on a en particulier $\lambda_1 < x_{k+1}$ et donc elle est vraie au rang $k+1$ toujours en appliquant le lemme précédent, cette fois-ci à $t = x_{k+1}$.

Par conséquent $(x_k)_{k\in\mathbf{N}}$ est une suite décroissante à valeurs dans $[\lambda_1; +\infty[$. Étant minorée elle converge donc vers un réel λ. Par passage à la limite λ appartient à $[\lambda_1; +\infty[$ et vérifie $g(\lambda) = \lambda$. D'après ce qui précède il en résulte $\lambda = \lambda_1$.

Proposition 1 *Pour t strictement supérieur à λ_1, on a*

$$0 \leq g'(t) \leq 1 - \frac{1}{n} .$$

De plus si a est un réel quelconque inférieur à λ_1, on a, avec les hypothèses du théorème précédent,

$$0 \leq x_k - \lambda_1 \leq \left(1 - \frac{1}{n}\right)^k (b-a)$$

pour tout entier k.

On a déjà démontré que g est croissante sur $[\lambda_1; +\infty[$. Par ailleurs

$$1 - g' = \left(\frac{P}{P'}\right)' = \left(\left(\frac{P'}{P}\right)^{-1}\right)' = -\frac{(P'/P)'}{(P'/P)^2} .$$

Or

$$\frac{P'}{P} = d\log(P) = \sum_{i=1}^{p} \frac{m_i}{X - \lambda_i}$$

et donc, pour tout $t > \lambda_1$,

$$1 - g'(t) = \frac{\sum_i \frac{m_i}{(t-\lambda_i)^2}}{\left(\sum_i \frac{m_i}{t-\lambda_i}\right)^2} .$$

L'inégalité de Cauchy-Schwarz appliquée aux vecteurs

$$\left(\sqrt{m_1}, \ldots, \sqrt{m_p}\right) \quad \text{et} \quad \left(\frac{\sqrt{m_1}}{t-\lambda_1}, \ldots, \frac{\sqrt{m_p}}{t-\lambda_p}\right)$$

donne le résultat voulu. On a effet

$$\left(\sum_i \frac{m_i}{t-\lambda_i}\right)^2 \leq \sum_i m_i . \sum_i \frac{m_i}{(t-\lambda_i)^2} = n \sum_i \frac{m_i}{(t-\lambda_i)^2} .$$

Soit $t \geq \lambda_1$, d'après l'inégalité des accroissements finis appliquée à g sur $[\lambda_1; t]$, on a

$$|g(t) - \lambda_1| \leq \left(1 - \frac{1}{n}\right) |t - \lambda_1| \ .$$

Montrons par récurrence sur l'entier naturel k que

$$|x_k - \lambda_1| \leq \left(1 - \frac{1}{n}\right)^k |x_0 - \lambda_1| \ .$$

Pour $k = 0$, c'est même une égalité. Si l'inégalité est vraie au rang k, on a, d'après ce qui précède,

$$|x_{k+1} - \lambda_1| \leq \left(1 - \frac{1}{n}\right) |x_k - \lambda_1| \leq \left(1 - \frac{1}{n}\right)^{k+1} |x_0 - \lambda_1| \ .$$

La propriété annoncée en résulte. Si k est un entier naturel, de $a \leq \lambda_1 \leq x_0 = b$, on déduit immédiatement

$$0 \leq x_k - \lambda_1 \leq \left(1 - \frac{1}{n}\right)^k (b - a) \ .$$

Lemme 14 *Soit m inférieur à n et $\rho_1 \geq \rho_2 \geq \ldots \geq \rho_m$ les m premières racines de P écrites sans multiplicité (on peut donc avoir $\rho_1 = \rho_2$) et P_{m-1} le polynôme à coefficients réels tel que*

$$P(X) = P_{m-1}(X) \prod_{i=1}^{m-1} (X - \rho_i) \ .$$

Si b est un réel supérieur à ρ_m, la suite récurrente définie par $x_0^{(m)} = b$ et

$$x_{k+1}^{(m)} = x_k^{(m)} - \frac{P_{m-1}(x_k^{(m)})}{P'_{m-1}(x_k^{(m)})}$$

converge vers ρ_m.

Il suffit d'apliquer le théorème précédent à P_{m-1} dont toutes les racines sont réelles et ρ_m est la plus grande.

L'écriture

$$\begin{aligned}
\frac{P'_{m-1}}{P_{m-1}} &= d\log(P_{m-1}) = \sum_{j \geq m} \frac{1}{X - \rho_j} \\
\frac{P'}{P} &= d\log(P) = \sum_{j} \frac{1}{X - \rho_j} \\
\frac{P'_{m-1}}{P_{m-1}} &= \frac{P'}{P} - \sum_{j<m} \frac{1}{X - \rho_j} = \frac{P' - \sum_{j<m} \frac{P}{X - \rho_j}}{P}
\end{aligned}$$

et donc

$$\frac{P_{m-1}(t)}{P'_{m-1}(t)} = \frac{P(t)}{P'(t) - \sum_{j<m} \frac{P(t)}{t-\rho_j}}$$

permet de calculer la suite $x_k^{(m)}$ en n'utilisant que des termes connus ou approchés : P, P' (qui sont connus grâce à leurs coefficients et non, évidemment, grâce à leurs racines) et les ρ_j pour $j < m$.

Séance du 28/02/2000

3 Recherche de valeurs propres

3.1 Introduction

L'algèbre linéaire sur un corps commutatif telle qu'on la connaît aujourd'hui s'est progressivement dégagée au cours du XIX^e siècle et au début du XX^e de la théorie des équations linéaires et de la géométrie. Les champs d'application en sont nombreux : systèmes d'équations linéaires, équations différentielles ou intégrales linéaire, calcul vectoriel dans les espaces affines, transformation des espaces projectifs, dualité etc.

C'est peut-être Carl Friedrich Gauß (1777 – 1855) le premier qui a permis cet essor en donnant l'interprétation géométrique des nombres complexes. De la sorte il devenait concevable d'additionner et de multiplier par un scalaire des vecteurs et, plus généralement, la nécessité d'un calcul sans référence à un choix de coordonnées devenait claire. Ce sont par la suite Hermann Grassmann (1809 – 1877), August Ferdinand Möbius (1790 – 1868) et William Rowan Hamilton (1805 – 1865) qui dégagent les règles du calcul vectoriel. La notion de rang d'un système est introduite par Georg Ferdinand Frobenius (1849 – 1917) et le calcul matriciel par Arthur Cayley (1821 – 1895)

L'algèbre multilinéaire, et au premier chef la théorie des déterminants, a pris naissance dans la théorie des invariants, notamment en géométrie différentielle. Ce sont Leopold Kronecker (1823 – 1891) et Karl Theodor Wilhelm Weierstraß (1815 – 1897) qui donnent la définition axiomatique des déterminants.

Dès 1888 Giuseppe Peano (1858 – 1932) donne la définition axiomatique des espaces vectoriels et des applications linéaires, mais c'est l'analyse qui fournit les plus importants exemples d'espaces vectoriels de dimension infinie. À propos de recherches sur les équations aux dérivées partielles, David Hilbert (1862 – 1943) introduit le célèbre espace de Hilbert quelques années avant que Stefan Banach (1892 – 1945) étudie systématiquement les opérateurs et la dualité dans les espaces de fonctions.

3.2 Quelques rappels d'algèbre linéaire

Si E et F sont deux espaces vectoriels munis de bases respectives B et B' et u une application linéaire de E dans F, on appelle matrice de u relativement aux bases B et B' le tableau de nombres obtenus en mettant en colonnes les coordonnées dans la base B' des images par u des vecteurs de la base B. Autrement dit si $B = (e_1, \dots, e'_m)$ et $B' = (e'_1, \dots, e'_n)$ alors la matrice de u relativement à B et B' est le tableau

$$M = \mathrm{mat}_{B,B'}(u) = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1m} \\ a_{21} & a_{22} & \dots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nm} \end{pmatrix} = \left(\begin{array}{c|c|c|c} u(e_1) & u(e_2) & \dots & u(e_m) \end{array} \right)$$

avec $u(e_i) = a_{1i}e'_1 + \ldots + a_{ni}e'_n$. Pour B et B' fixées, il y a une bijection entre l'ensemble des matrices de taille (m, n) et l'ensemble des applications linéaires de E dans F. Ceci permet de définir l'addition, la multiplication par un scalaire et la multiplication des matrices comme le pendant de l'addition, la multiplication par un scalaire et la composition des applications linéaires. Quand on étudie une matrice sans autre référence il est en fait sous-entendu que l'on la considère comme la matrice d'un endomorphisme relativement aux bases canoniques de $\mathbf{R}^m$ et $\mathbf{R}^n$ (ou de $\mathbf{C}^m$ et $\mathbf{C}^n$).

Le rang de M est par définition le rang de u, c'est-à-dire la dimension de l'espace engendré par les vecteurs colonnes de M ou, ce qui revient au même, celui de l'espace engendré par les vecteurs lignes de M.

On dit qu'un vecteur x de E est un vecteur propre de u s'il existe un scalaire λ tel que $u(x) = \lambda x$. On dira que x est associé à λ. Matriciellement, en écrivant X le vecteur colonne formé des coordonnées de x dans la base B, cela s'écrit

$$MX = \lambda X \ .$$

On remarquera que 0 est toujours vecteur propre de u.

On dit qu'un scalaire λ est valeur propre de u s'il existe un vecteur propre non nul associé à u.

Soit E un espace vectoriel de dimension n. Rappelons que l'ensemble des applications n-multilinéaires alternées sur E, i.e. les applications

$$\begin{array}{cccc} \varphi & : & E \times \ldots \times E & \to & E \\ & & (x_1, \ldots, x_n) & \mapsto & \varphi(x_1, \ldots, x_n) \end{array}$$

linéaires en chacune des variables et nulles dès que deux des vecteurs x_i sont égaux, est un espace vectoriel de dimension 1. De plus si φ est n-multilinéaire alternée et non identiquement nulle $\varphi(x_1, \ldots, x_n)$ est nul si et seulement si $(x_1, \ldots, x_n)$ est une base de E.

Soit B une base de E. L'unique application n-multilinéaire alternée prenant la valeur 1 sur B est appelée déterminant relativement à la base B et est notée $\det_B$.

Si u est un endomorphisme de E et si $B = (e_1, \ldots, e_n)$, on note $\det_B(u)$ le scalaire $\det_B(u(e_1), \ldots, u(e_n))$. Si M est la matrice de u relativement à la base B (ce qui sous-entend que l'on a choisi B' égal à B) on définit le déterminant de M comme ce scalaire. Avec cette définition et la propriété caractéristique des déterminants on voit qu'une application linéaire est bijective ou, ce qui revient au même, une matrice est inversible si et seulement si son déterminant (relativement à une base quelconque) est non nul.

En particulier le rang d'une matrice est la dimension de sa plus grande matrice carrée extraite de déterminant non nul. On en déduit également que λ est valeur propre de la matrice carrée M si et seulement si $\det(M - \lambda Id)$ est nul, où Id représente la matrice de l'endomorphisme identité. La quantité $\det(M - \lambda Id)$ est en fait un polynôme en λ, nommé polynôme caractéristique de M.

Soit B_1 et B_2 sont deux bases de E et x un vecteur. On note X_1 et X_2 les vecteurs colonnes formés des coordonnées de x respectivement dans les bases B_1 et B_2. On a la formule

$$X_1 = P_{B_1|B_2} X_2$$

où $P_{B_1|B_2}$ est par définition la matrice de passage de la base B_1 à la base B_2, c'est-à-dire la matrice de l'identité relativement aux bases B_2 et B_1 dans cet ordre :

$$P_{B_1|B_2} = \mathrm{mat}_{B_2,B_1}(Id)$$

i.e. $P_{B_1|B_2}$ est la matrice dont les colonnes sont les vecteurs de la base B_2 exprimés dans la base B_1. Dans ces conditions, si B_1' et B_2' sont deux bases de F et u une application linéaire de E dans F on a

$$\mathrm{mat}_{B_1',B_2'}(u) = P^{-1}_{B_2|B_2'} \mathrm{mat}_{B_1,B_2}(u) P_{B_1|B_1'} \qquad \text{et} \qquad P^{-1}_{B_2|B_2'} = P_{B_2'|B_2} \ .$$

On a les formules suivantes :

1. Si M et N sont deux matrices carrées de même taille
$$\det(MN) = \det(M)\det(N) \ .$$
2. Si $M = (a_{ij})$ est une matrice carrée et si M_{ij} est son mineur d'ordre (i,j), c'est-à-dire la matrice carrée obtenue à partir de M en supprimant la $i^{\text{ème}}$ ligne et la $j^{\text{ème}}$ colonne, on a
$$\det(M) = \sum_i (-1)^{i+j} a_{ij} \det(M_{ij}) = \sum_j (-1)^{i+j} a_{ij} \det(M_{ij}) \ .$$
3. (Formules de Cramer.) Si A est une matrice carrée inversible d'ordre n et B un vecteur colonne de taille n, l'équation $AX = B$ admet une unique solution. La $i^{\text{ème}}$ coordonnée de X, notée x_i, est obtenue par la formule
$$x_i = \frac{\det(B_i)}{\det(A)}$$
où B_i est obtenue à partir de A en remplaçant la $i^{\text{ème}}$ colonne par B.
4. (Théorème de Rouché-Fontené.) Si A est une matrice de taille (n,m) de rang r et B un vecteur colonne de taille n, l'équation $AX = B$ soit n'admet aucune solution soit admet comme espace de solutions un espace vetoriel de dimension $n - r$. Si M est une matrice carrée d'ordre r inversible extraite de A
$$M = \begin{pmatrix} a_{i_1 j_1} & a_{i_1 j_2} & \dots & a_{i_1 j_r} \\ a_{i_2 j_1} & a_{i_2 j_2} & \dots & a_{i_2 j_r} \\ \vdots & \vdots & \ddots & \vdots \\ a_{i_r j_1} & a_{i_r j_2} & \dots & a_{i_r j_r} \end{pmatrix}$$

avec $i_1 < i_2 < \ldots i_r$ et $j_1 < j_2 < \ldots < j_r$, la condition pour que le système linéaire $AX = B$ admette des solutions est que les déterminants

$$\begin{pmatrix} a_{i_1j_1} & a_{i_1j_2} & \ldots & a_{i_1j_r} & b_{i_1} \\ a_{i_2j_1} & a_{i_2j_2} & \ldots & a_{i_2j_r} & \\ \vdots & \vdots & \ddots & \vdots & \\ a_{i_rj_1} & a_{i_rj_2} & \ldots & a_{i_rj_r} & b_{i_r} \\ a_{ij_1} & a_{ij_2} & \ldots & a_{ij_r} & b_i \end{pmatrix}$$

soient tous nuls pour tout i n'appartenant pas à $\{i_1, \ldots, i_r\}$. Dans ce cas on obtient les solutions du système en fixant arbitrairement les $n - r$ inconnues x_i pour i n'appartenant pas à $\{i_1, \ldots, i_r\}$ et en résolvant le système de Cramer donné par les r équations en les r inconnues $(x_{i_1}, \ldots, x_{i_r})$, représenté par M. Toutes ces considérations ne dépendent pas de la matrice M choisie (pourvu que son rang soit celui de A).

3.3 La méthode de la puissance

Le but de cette méthode est de trouver, quand elle existe, la valeur propre dominante d'une matrice ainsi qu'un vecteur propre associé. En la prolongeant on peut espérer trouver une base diagonalisation de A, quand elle existe.

Soit A une matrice carrée d'ordre n à coefficients complexes. Une valeur propre λ de A est dite dominante si son module est strictement supérieur au module de toutes les autres valeurs propres de A.

Soit A une matrice diagonalisable admettant $\lambda = \lambda_1$ comme valeur propre dominante, $(\lambda_2, \ldots, \lambda_p)$ les autres valeurs propres de A. Si X est un vecteur colonne, on peut l'écrire

$$X = X_1 + X_2 + \ldots + X_p$$

où $(X_1, X_2, \ldots, X_p)$ sont des vecteurs propres de A associés respectivement aux valeurs propres $(\lambda_1, \lambda_2, \ldots, \lambda_p)$. (On prendra garde que les espaces propres de A ne sont pas nécessairement de dimension 1.)

Pour tout entier m on a donc

$$\lambda_1^{-m} A^m X = X_1 + \left(\frac{\lambda_2}{\lambda_1}\right)^m X_2 + \ldots + \left(\frac{\lambda_p}{\lambda_1}\right)^m X_p$$

et, par conséquent la suite $\lambda_1^{-m} A^m X$ converge vers X_1.

Si X_1 n'est pas nul, notons x_i l'une de ses coordonnées non nulles et $x_i^{(m)}$ la coordonnée correspondante de $A^m X$. On a donc $x_i^{(m)} \sim \lambda_1^m x_1$ et, par conséquent,

$$\lim_{m \to \infty} \frac{x_i^{(m+1)}}{x_i^{(m)}} = \lambda_1 \ .$$

Si λ_1 est réel positif, on a également

$$\lim_{m\to\infty} \frac{A^m X}{||A^m X||} = \frac{X_1}{||X_1||}$$

et ce dernier est un vecteur propre (unitaire) de A et on en déduit λ_1 en calculant son image par A.

Algorithme 3 *On choisit un vecteur X_0 non nul initial quelconque. Supposons X_m construit pour un certain entier naturel m, on forme AX_m et on le normalise en le multipliant par un scalaire adéquat, par exemple de façon à ce que ses coefficients ne soient ni trop petits ni trop grands ou qu'il soit unitaire. On appelle X_{m+1} ce nouveau vecteur. À chaque étape X_m est proportionnel à $A^m X_0$. Si X_m converge vers un vecteur non nul X, sa limite est un vecteur propre de A pour la valeur propre λ obtenue comme quotient de deux coordonnées correspondantes de AX et X.*

Il faut prendre garde qu'il peut y avoir des problèmes. Si par exemple la projection de X_0 sur l'espace propre associé à λ_1 est nulle, on obtiendra un vecteur propre pour une valeur propre différente de λ_1. Néanmoins cela est assez rare pour deux raisons : primo l'espace propre en question est (quand A est diagonalisable et non scalaire) de mesure nulle, secundo une erreur d'arrondi peut très bien permettre à cette projection de ne plus être nulle et donc de permettre à la suite de converger vers un vecteur propre de A pour λ_1.

Mais ce ne sont pas là les seuls problèmes. Il est rare que l'on puisse savoir à l'avance qu'une matrice donnée est diagonalisable, sauf quand elle est par exemple symétrique. Enfin il n'y a aucune raison pour que A admette une valeur propre dominante. Néanmoins si on perturbe un petit peu A en considérant $A + \varepsilon Id$, son spectre est translaté par ε et elle admet en général une valeur propre dominante : si λ et μ sont des valeurs propres distinctes de A telles que $|\lambda| = |\mu|$, alors $\lambda - \varepsilon$ et $\mu - \varepsilon$ sont valeurs propres de $A + \varepsilon Id$ et on a

$$\begin{aligned}
|\lambda - \varepsilon| = |\mu - \varepsilon| &\Leftrightarrow |\lambda - \varepsilon|^2 = |\mu - \varepsilon|^2 \\
&\Leftrightarrow |\lambda|^2 - 2Re(\lambda\bar{\varepsilon}) + |\varepsilon|^2 = |\mu|^2 - 2Re(\mu\bar{\varepsilon}) + |\varepsilon|^2 \\
&\Leftrightarrow Re((\lambda - \mu)\bar{\varepsilon}) = 0 \\
&\Leftrightarrow \varepsilon \in i(\lambda - \mu)\mathbf{R}
\end{aligned}$$

et donc, pour ε général, $|\lambda - \varepsilon|$ est distinct de $|\mu - \varepsilon|$. S'il est, de plus, assez petit, les translatés des valeurs propres de A ne pourront pas devenir de modules égaux. Par exemple si le spectre de A est réel, tout ε assez petit fera l'affaire.

`Séance du 06/03/2000`

Exemple : prenons la matrice symétrique

$$A = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}$$

dont on sait que les valeurs propres sont 2, $2-\sqrt{2}$ et $2+\sqrt{2}$ et choisissons $X_0 = e_1+e_2+e_3$. On a $X_1 = e_1 + e_3$, $X_2 = AX_1/2 = e_1 - e_2 + e_3$ etc. $X_7 = 99e_1 - 140e_2 + 99e_3$, $AX_7 = 338e_1 - 478e_2 + 338e_3$ et on a

$$\frac{338}{99} \simeq 3.41414 \qquad \text{et} \qquad \frac{-478}{-140} \simeq 3.41429$$

et on est donc assez proche de la valeur propre $\lambda_1 = 2 + \sqrt{2} \simeq 3.414214$. De plus le vecteur $X_8 = AX_7/338 \simeq (e_1 - 1.41420e_2 + e_3)$ est déjà très proche du vecteur propre $e_1 - \sqrt{2}e_2 + e_3$.

Pour obtenir les autres valeurs propres il faut créer une matrice qui a les mêmes espaces propres et dont on peut relier les valeurs propres à celles de A. Dans le cas général il n'y a pas d'autres choix que prendre $A - \lambda Id$ pour un certain scalaire λ. Quand A est symétrique, on sait qu'elle est diagonalisable dans une base orthogonale et on peut donc former la matrice $A - \lambda_1 P_1$ où P_1 est la matrice de la projection orthogonale sur l'espace propre que l'on vient de trouver. Cette matrice est encore symétrique et a les mêmes espaces propres que A. Ses valeurs propres sont les mêmes à l'exception de λ_1 qui a disparu (si elle était simple). Concrètement si X est le vecteur limite que l'on a trouvé, on itère le procédé avec

$$A - \frac{AX^tX}{{}^tXX} \; .$$

On prendra garde que AX^tX est une matrice carrée de taille n alors que ${}^tXX$ est un scalaire. Pour justifier cette formule il suffit de remarquer que si Y est orthogonal à X alors ${}^tXY$ est nul et donc aussi AX^tXY et pour X on obtient AX, c'est-à-dire $\lambda_1 X$.

3.4 La méthode de la puissance inverse

On part de la remarque suivante : si A est inversible, les valeurs propres de A^{-1} sont les inverses des valeurs propres de A et leurs espaces propres sont les mêmes. En effet

$$AX = \lambda X \Leftrightarrow X = \lambda A^{-1}X \Leftrightarrow A^{-1}X = \lambda^{-1}X \; .$$

Par conséquent si on applique la méthode de la puissance à la matrice A^{-1} on trouve, si tout va bien, un vecteur propre associé à la valeur propre ayant le module d'inverse le plus grand, i.e. la valeur propre de module le plus petit.

Ceci va servir en particulier à affiner les résultats de la méthode de la puissance. Si (λ, X) est une estimation d'une valeur propre et d'un vecteur propre associé, en appliquant la méthode de la puissance à l'inverse de $A-\lambda Id$ en partant de X, on doit converger vers la valeur propre telle que $\lambda_i - \lambda$ est de plus petit module, ce qui ne saurait être que la valeur propre que l'on vient d'approcher, si on ne s'est pas trompé.

Cette méthode a néanmoins un gros inconvénient : la matrice $A - \lambda Id$ a une valeur propre très petite, donc son déterminant risque fort d'être très petit. Par conséquent les calculs donnant son inverse risquent de mettre en jeu des inverses de nombres très petits. La perte de précision dans un tel calcul est très grande dès qu'on s'approche des limites de la précision de l'ordinateur :

$$|x - 10^{-20}| < 10^{-20} \Leftrightarrow \frac{1}{x} > \frac{1}{2}10^{20} .$$

Pour cette raison on utilise des méthodes de calculs permettant de trouver $A^{-1}Y$, étant donné Y, sans calculer A^{-1}. L'une de ces méthodes consiste à mettre A sous forme LU avec L triangulaire inférieure et U triangulaire supérieure. De la sorte on a

$$X = A^{-1}Y \Leftrightarrow Y = AX \Leftrightarrow (Y = LZ \quad \text{et} \quad Z = UX)$$

et on trouve Z en résolvant un système triangulaire, puis X en résolvant également un système triangulaire. Évidemment en termes de coût de calcul, la recherche d'une décomposition LU prend du temps mais elle évite des erreurs d'arrondis qui peuvent être catastrophiques.

On notera néanmoins que cette décomposition LU n'existe pas toujours. Le résultat général est qu'on peut toujours mettre une matrice inversible sous la forme PLU avec P une matrice de permutation (i.e. associée à un endomorphisme $e_i \mapsto e_{\sigma(i)}$ pour une permutation σ des entiers entre 1 et n), L triangulaire inférieure à diagonale composée de 1 et U triangulaire supérieure.

Exemple : pour une matrice carrée d'ordre 2

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

inversible, i.e. $\det(A) = ad - bc \neq 0$, on étudie l'équation

$$A = \begin{pmatrix} 1 & 0 \\ x & 1 \end{pmatrix} \begin{pmatrix} \alpha & y \\ 0 & \beta \end{pmatrix} = \begin{pmatrix} \alpha & y \\ \alpha x & \beta + xy \end{pmatrix} .$$

Comme le déterminant est multiplicatif et le déterminant de A est non nul, il doit en être de même pour les deux matrices triangulaires intervenant dans cette décomposition. On a donc nécessairement $\alpha\beta \neq 0$. Il en résulte que l'on peut résoudre l'équation générale si et seulement si a est non nul et alors

$$\alpha = a \qquad y = b \qquad x = \frac{c}{a} \qquad \text{et} \qquad \beta = \frac{ad - bc}{a} .$$

Séance du 13/03/2000

3.5 La méthode de Givens-Householder

Le but de cette méthode est mettre une matrice symétrique sous une forme tridiagonale, i.e. avec des coefficients a_{ij} nuls dès que $|i-j|$ est strictement supérieur à 1. Pour cela opère des changements de bases grâce à ses symétries hyperplanes.

Remarquons tout d'abord que si u et v sont deux vecteurs de E, il existe une symétrie hyperplane échangeant u et v si et seulement s'ils sont de même norme. En effet il est nécessaire que u et v soient de même norme puisqu'une symétrie est une isométrie. Réciproquement si u et v sont distincts, la symétrie par rapport à leur hyperplan médiateur répond à la question (puisque cet hyperplan contient 0). Sinon toute symétrie par rapport à un hyperplan contenant $u=v$ convient.

Supposons que E soit de dimension 3. Soit $u=(u_1,u_2,u_3)$ un vecteur de E. On note, pour tout réel θ, u_θ le vecteur $\cos\theta e_2+\sin\theta e_3$. La symétrie par rapport au plan perpendiculaire à u_θ envoie u dans le plan (e_1,e_2) si et seulement si

$$\cos(2\theta)u_3=\sin(2\theta)u_2\ .$$

En particulier un tel u_θ existe.

En effet, la symétrie par rapport au plan perpendiculaire à u_θ, qui est de norme 1, envoie u sur le vecteur

$$v=u-2\langle u,u_\theta\rangle u_\theta\ .$$

La dernière coordonnée de ce vecteur est

$$u_3-2(u_2\cos\theta+u_3\sin\theta)\sin\theta\ ,$$

soit $(1-2\sin^2\theta)u_3-2u_2\sin\theta\cos\theta$. Par conséquent v appartient au plan (e_1,e_2) si et seulement si

$$\cos(2\theta)u_3=\sin(2\theta)u_2\ .$$

Lemme 15 *Soit u et v deux vecteurs de E et P un plan contenant v. Il existe une symétrie hyperplane de E laissant fixe v et envoyant u dans P.*

Si u appartient à P, toute symétrie par rapport à un hyperplan contenant P convient.

Sinon soit F l'espace de dimension 3 contenant P et u. On applique le résultat précédent en prenant comme base de E une base telle que v à soit proportionnel à e_1. Il existe donc w dans F, unitaire, orthogonal à v et tel que

$$u-2\langle u,w\rangle w$$

appartienne à P. Il en résulte que la symétrie par rapport à l'hyperplan orthogonal à w laisse fixe v et envoie u dans P.

Lemme 16 *Soit ψ un endomorphisme symétrique de E. Il existe σ une symétrie hyperplane telle que*

$$\sigma\psi\sigma = \sigma^{-1}\psi\sigma = {}^t\sigma\psi\sigma$$

est un endomorphisme symétrique de E envoyant e_1 dans le plan engendré par e_1 et e_2.

Puisque σ est à la fois orthogonale et involutive, on a $\sigma = \sigma^{-1} = {}^t\sigma$.

Soit maintenant σ une symétrie envoyant $\varphi(e_1)$ dans le plan engendré par (e_1, e_2) et fixant e_1. On a donc

$$\sigma\psi\sigma(e_1) = \sigma\left(\psi(e_1)\right)$$

et ce dernier vecteur appartient au plan engendré par e_1 et e_2.

L'assertion en résulte.

Si l'on demande à σ de fixer e_1 et si ce vecteur n'est pas vecteur propre de ψ, il y a en fait deux choix possibles. Sinon il y a une infinité de choix possibles a priori.

Théorème 9 *Soit A une matrice symétrique. Il existe une matrice orthogonale P telle que $P^{-1}AP = {}^tPAP$ soit tridiagonale.*

On démontre par récurrence sur l'entier k strictement inférieur à $n-1$ la propriété (H) : tout endomorphisme symétrique de E est orthogonalement semblable à un endomorphisme envoyant e_i dans l'espace engendré par (e_{i-1}, e_i, e_{i+1}) pour i compris entre 1 et k. Pour $i = 1$ on convient $e_0 = 0$.

Pour $k = 1$, cela résulte du lemme précédent. Soit maintenant k strictement supérieur à 1, E_k l'espace engendré par $(e_1, \ldots, e_{k-1})$ et F_k son supplémentaire orthogonal. Comme on a supposé k strictement inférieur à $n-1$, F_k est de dimension au moins 3. En supposant la propriété vraie au rang $k-1$, on se donne ρ unitaire tel que ${}^t\rho\psi\rho$ envoie e_i dans l'espace engendré par (e_{i-1}, e_i, e_{i+1}) (qui n'est rien d'autre que $E_{i+2} \cap F_{i-1}$). En particulier, pour $i \leq k-2$, on a

$$\langle {}^t\rho\psi\rho(e_i), e_k\rangle \ .$$

Comme ${}^t\rho\psi\rho$ est un endomorphisme symétrique cela entraîne

$$\langle {}^t\rho\psi\rho(e_k), e_i\rangle$$

et donc ${}^t\rho\psi\rho(e_k)$ appartient à F_{k-1}. Écrivons

$${}^t\rho\psi\rho(e_k) = \alpha e_{k-1} + u$$

avec u dans F_k et α réel. On sait qu'on peut trouver une symétrie hyperplane de F_k fixant e_k et envoyant u dans l'espace engendré par (e_k, e_{k+1}), c'est-à-dire un vecteur unitaire v de F_k orthogonal à e_k tel que

$$u - 2\langle u, v\rangle v \in F_k \cap E_{k+2} \ .$$

Puisque v appartient à F_k et est orthogonal à e_k, c'est qu'il appartient en fait à F_{k+1}. La symétrie σ_k de E par rapport à l'hyperplan orthogonal à v fixe donc E_{k+1} et ainsi envoie ${}^t\rho\psi\rho(e_k)$ sur

$$\sigma_k(\alpha e_{k-1} + u) = \alpha e_{k-1} + \sigma_k(u) ,$$

ce qui est un élément de l'espace engendré par (e_{k-1}, e_k, e_{k+1}). Comme σ_k fixe e_k, on a même

$${}^t\sigma_k{}^t\rho\psi\rho\sigma_k(e_k) = \sigma_k{}^t\rho\psi\rho(e_k) .$$

De plus, comme v appartient à F_{k+1}, σ_k fixe E_{k+1} donc pour i strictement inférieur à k, on a $e_i \in E_k$ et

$${}^t\sigma_k{}^t\rho\psi\rho\sigma_k(e_i) = \sigma_k{}^t\rho\psi\rho(e_i) .$$

Comme ${}^t\rho\psi\rho(e_i)$ appartient à E_{i+2}, il appartient à E_{k+1} et est donc fixé par σ_k. Il en résulte que ${}^t\sigma_k{}^t\rho\psi\rho\sigma_k$ est orthogonalement semblable à ψ et envoie tout e_i, pour i inférieur ou égal à k, dans l'espace engendré par (e_{i-1}, e_i, e_{i+1}).

Par le principe de récurrence la propriété est donc vraie pour $k = n - 2$. Il en résulte que la matrice de l'endomorphisme u dans une base orthogonale bien choisie a tous ses coefficients en-desous de la sous-diagonale nuls. Comme cette matrice est évidemment symétrique, elle est tridiagonale. Ce qui est l'assertion recherchée.

Avec la méthode précédente on a à chaque fois deux choix pour la symétrie σ_k. Comme σ_k fixe e_k, si on écrit

$${}^t\rho\psi\rho(e_k) = \alpha e_{k-1} + \beta e_k + u'$$

avec u' dans F_{k+1}, on a

$$\sigma_k(\beta e_k + u') = \beta e_k + \sigma_k(u')$$

et, comme u' est perpendiculaire à e_k, σ_k préserve l'orthogonalité et envoie $\beta e_k + u'$ dans l'espace engendré par (e_k, e_{k+1}), on a

$$\sigma_k(\beta e_k + u') = \beta e_k \pm ||u'||e_{k+1} .$$

Pour que les erreurs d'arrondis ne soient pas trop importantes on a donc intérêt à ce que, à chaque étape, $||u'||$ ne soit pas trop petit. On fait donc le choix de sorte ${}^t\sigma_k{}^t\rho\psi\rho\sigma_k(e_{k+1})$ ait une projection sur F_{k+2} de norme la plus grande possible.

Comme on a obtenu un endomorphisme semblable à ψ, il a les mêmes valeurs propres que ψ et on est donc ramené à savoir calculer les valeurs propres d'un endomorphisme tridiagonal.

Remarque : Si l'on cherche, avec cette méthode, à diagonaliser ψ, on est conduit à prendre σ envoyant $\psi(e_1)$ dans la droite engendrée par e_1 mais alors $\sigma\psi\sigma$ ne fixe pas cette droite (sauf si e_1 était déjà vecteur propre) ou bien à chercher σ tel que e_1 soit vecteur propre de $\sigma\psi\sigma$. Il revient au même de trouver un vecteur propre de ψ (qui sera $\sigma(e_1)$) or ceci est impossible puisque l'on n'a pas encore déterminé les valeurs propres de ψ. En conclusion la forme tridiagonale est la meilleure que l'on puisse espérer avec cette méthode.

Soit S une matrice symétrique tridiagonale. On note, pour tout entier i compris entre 1 et n, S_i la matrice carrée extraite de S obtenue en ne gardant que les i premières lignes et colonnes. On note P_i le polynôme caractéristique de S_i, $(a_1, \ldots, a_n)$ la diagonale de S et $(b_1, \ldots, b_{n-1})$ la surdiagonale de S (ou la sous-diagonale puisque c'est la même chose) :

$$S = \begin{pmatrix} a_1 & b_1 & 0 & \ldots & 0 \\ b_1 & a_2 & b_2 & \ldots & 0 \\ 0 & b_2 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & b_{n-1} \\ 0 & \ldots & \ldots & b_{n-1} & a_n \end{pmatrix} .$$

On suppose enfin qu'aucun des coefficients b_i, pour i entre 1 et $n-1$, n'est nul.

On pose $P_0 = 1$. On a

$$P_2 = \begin{vmatrix} a_1 - X & b_1 \\ b_1 & a_2 - X \end{vmatrix} = (a_2 - X)(a_1 - X) - b_1^2 = (a_2 - X)P_1 - b_1^2 P_0 .$$

Pour i supérieur ou égal à 2, on a

$$P_{i+1} = \begin{vmatrix} a_1 - X & b_1 & 0 & \ldots & 0 \\ b_1 & a_2 - X & b_2 & \ldots & 0 \\ 0 & b_2 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & b_i \\ 0 & \ldots & \ldots & b_i & a_{i+1} - X \end{vmatrix}$$

et donc, en développant par rapport à la dernière ligne

$$P_{i+1} = (a_{i+1} - X)P_i - b_i \begin{vmatrix} a_1 - X & b_1 & 0 & \ldots & 0 & 0 \\ b_1 & a_2 - X & b_2 & \ldots & 0 & 0 \\ 0 & b_2 & \ddots & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & \ddots & b_{i-2} & 0 \\ 0 & \ldots & \ldots & b_{i-2} & a_{i-1} - X & 0 \\ 0 & \ldots & \ldots & 0 & b_{i-1} & b_i \end{vmatrix}$$

et donc, en développant ce dernier déterminant par rapport à la dernière colonne, on obtient

$$P_{i+1} = (a_{i+1} - X)P_i - b_i^2 P_{i-1} .$$

Si x est une racine de P_i, on a

$$P_{i+1}(x)P_{i-1}(x) = -b_i^2 P_{i-1}(x) \leq 0 .$$

Montrons en fait que P_i et P_{i-1} n'ont pas de racines communes. La formule de récurrence donnant les P_k montre qu'une racine commune à P_i et P_{i-1} est également une racine commune à P_{i-1} et P_{i-2} (si i est plus grand que 2) et donc, finalement, une racine commune à P_1 et P_0. Comme P_0 ne s'annule pas, une telle racine n'existe pas et on a

$$P_i(x) = 0 \Rightarrow P_{i+1}(x)P_{i-1}(x) < 0 .$$

Pour $i = 0$, P_0 est constant égal à 1 et ses limites en $+$ et $-$ l'infini sont égales à 1. Si i est strictement positif, le terme dominant de P_i est $(-X)^i$ et donc

$$\lim_{x\to-\infty} P_i(x) = +\infty \qquad \text{et} \qquad \lim_{x\to+\infty} P_i(x) = \begin{cases} +\infty & \text{si } i \text{ est pair} \\ -\infty & \text{si } i \text{ est impair} \end{cases}$$

Montrons par récurrence sur l'entier i compris entre 1 et n la propriété (H) suivante : P_i est scindé à racines simples et, si i est supérieur à 2, les racines de P_{i-1} sont intercalées entre celles de P_i.

Pour $i = 1$, $P_1 = a_1 - X$ et (H) est vraie.

Soit i supérieur à 1. Supposons (H) vraie pour tout j inférieur à i et montrons qu'elle est vraie pour $i + 1$. Puisque P_{i-1} est à racines simples, il change de signe en chacune de ses racines : P_{i-1} est donc positif avant $\lambda_{i-1,1}$, négatif entre $\lambda_{i-1,1}$ et $\lambda_{i-1,2}$ etc. De par la propriété des racines de P_i, on en déduit que P_{i-1} prend des signes alternés en les racines de P_i. Comme P_{i+1} et P_{i-1} sont de signes opposés en ces mêmes racines, c'est que P_{i+1} prend des signes alternés entre les racines de P_i.

Par le théorème des valeurs intermédiaires on en déduit que P_{i+1} admet une racine entre chacune des racines de P_i, ce qui lui en fait au moins $i - 1$. Or P_{i-1} est positif avant $\lambda_{i-1,1}$, donc en $\lambda_{i,1}$ et P_{i+1} y est négatif. De par son comportement à l'infini, cela impose à P_{i+1} d'avoir une racine avant $\lambda_{i,1}$.

De même P_{i_1} est du signe de $(-1)^{i-1}$ après $\lambda_{i-1,i-1}$ donc en $\lambda_{i,i}$. Aussi P_{i+1} y est-il du signe de $(-1)^i$. Ceci impose à P_{i+1} d'avoir une racine après $\lambda_{i,i}$. La propriété (H) est donc vérifiée pour $i + 1$ et le principe de récurrence entraîne sa validité pour tout entier i compris entre 1 et n.

Lemme 17 *Soit x un réel et $\omega(x)$ le nombre de changements de signes dans la suite $(P_0(x), \ldots, P_n(x))$ dans laquelle on a retiré les zéros, i.e.*

$$\omega(x) = \mathit{Card}\,\{i \in [0; n-1] \mid \exists j \in [i+1; n] \quad P_i(x)P_j(x) < 0 \quad \mathit{et} \quad P_k(x) = 0 \ \mathit{si}\ i < k < j\} .$$

ω est constante sur tout intervalle ne contenant aucune valeur propre de S.

Les valeurs propres de S sont les racines de P_n. Soit V l'ensemble de ces racines et W l'ensemble des racines de tous les polynômes P_i pour i entre 0 et n.

Si x n'appartient pas à W, par continuité des P_i, il existe un intervalle contenant x sur lequel aucun d'eux ne s'annule et donc ω est constante sur cet intervalle. Par conséquent ω est constante sur tout intervalle ne rencontrant pas W.

Si maintenant x appartient à W mais pas à V. Par continuité des P_i et le fait qu'ils sont à racines simples, il existe un intervalle contenant x sur lequel, pour tout i entre

0 et n, soit P_i ne s'annule pas, soit P_i admet un unique 0 (i.e. x) et y change de signe. Puisque les P_i n'ont pas de racine commune, la suite des $P_i(x)$ ne s'annule pas deux fois consécutivement et là où elle s'annule ses valeurs de par et d'autre sont de signes opposés. Comme x n'appartient pas à V et comme P_0 ne s'annule pas, il y a bien des éléments non nuls de la suite $P_i(x)$ de par et d'autre de tout éventuel 0 de cette suite. Autrement dit tout 0 dans la suite $P_i(x)$ correspond à un changement de signe (mais la réciproque n'est pas vraie).

Or si $P_i(x)$ est nul, P_{i+1} et P_{i-1} sont non nuls en x et, par construction, ils ne s'annulent pas sur l'intervalle considéré. Comme ils sont de signes opposé sur cet intervalle quelque soit t dans cet intervalle et quelque soit le signe de $P_i(t)$, la suite $P_{i-1}(t), P_i(t), P_{i+1}(t)$ admet un et un seul changement de signe. Autrement dit le fait que P_i s'annule ne modifie pas le nombre de changements de signes.

Par conséquent ω est constante sur tout intervalle ne rencontrant pas V, i.e. sur tout intervalle ne contenant aucune valeur propre de S.

Théorème 10

$$\lim_{x\to y, x>y} \omega(x) - \lim_{x\to y, x<y} \omega(x)$$

est nul si y n'est pas valeur propre de S et vaut 1 sinon.

Puisque ω est constante sur tout intervalle ne rencontrant pas de valeur propre de S, elle y est continue et donc si x n'est pas une valeur propre de S

$$\lim_{x\to y, x>y} \omega(x) - \lim_{x\to y, x<y} \omega(x) = 0 .$$

Si maintenant x est valeur propre de S, le raisonnement de la question précédente montre encore que le fait que P_i s'annule en x, pour $i < n$, ne modifie pas le nombre de changements de signes dans la suite $P_i(x)$. Si par contre $P_n(x)$ s'annule, comme P_{n-1} est de signe constant au voisinage de x, le fait que P_n change de signe en x montre que le nombre de changements de signes est différent au voisinage de x selon que l'on est avant x ou après. Pour montrer qu'il s'accroît il faut donc montrer que P_n est du signe de P_{n-1} avant x et du signe opposé après. Or P_{n-1} prend des signes alternés en les racines de P_n et est positif en $\lambda_{n,1}$, il est donc du signe de $(-1)^{j-1}$ au voisinage de $\lambda_{n,j}$. Comme P_n change de signe en chacune de ses racines en commençant par être positif en $-\infty$, il est du signe de $(-1)^{j-1}$ à gauche de $\lambda_{n,j}$ et du signe de $(-1)^j$ à droite de cette même racine. L'assertion en résulte.

Algorithme 4 *Pour isoler les valeurs propres de S, on commence par majorer leur valeur absolue par un réel M (soit en majorant les racines du polynôme caractéristique de S grâce à ses coefficients, soit en faisant des majorations directes grâce aux coefficients de S). Ensuite on effectue une dichotomie sur l'intervalle $[-M; M]$ jusqu'à trouver des points où ω augmente d'au plus 1 entre chaque point de la subdivision.*

Remarque : si certains b_i sont nuls, la matrice S est diagonale par blocs. On peut donc se restreindre aux calculs des polynômes caractéristiques des blocs, sur lesquels la méthode précédente s'applique.

Séance du 20/03/2000

4 Les méthodes linéaires en analyse de données

4.1 Introduction

On cherche à trouver une relation entre des données x (a priori connues précisément) et des grandeurs y fruits d'une expérience mettant en jeu les données x. On va donc décrire cette expérience par des couples (x, y) ou, autrement dit, par une fonction $y = \varphi(x)$.

On suppose[1] que y est assertie d'erreurs et/ou de perturbations (erreurs de mesure, phénomènes négligés etc.) modélisées par une variable suivant une loi normale centrée d'écart-type donné (disons σ). En d'autres termes on écrit que la probabilité de mesurer y et de trouver une valeur égale à $\varphi(x)$ à ε près est égale à

$$P\left(|y-\varphi(x)| \leq \varepsilon\right) = \int_{-\varepsilon}^{\varepsilon} e^{-t^2/2\sigma^2} \frac{dt}{\sqrt{2\pi}\sigma} .$$

En fait, plus précisément, la probabilité de mesurer une valeur de y dans un intervalle I est

$$P(y \in I) = \int_I e^{-(t-\varphi(x))^2/2\sigma^2} \frac{dt}{\sqrt{2\pi}\sigma} .$$

Supposons maintenant que l'on mesure successivement des couples (x_i, y_i) pour i allant de 1 à n. On aimerait qu'en terme de probabilité ce soit la valeur exacte $y_i = \varphi(x_i)$ qui ait le plus de chances d'être mesurée. Soit donc y_i une certaine valeur (observée) ; notons Y la variable aléatoire qui modélise notre expérience, c'est-à-dire que Y suit une loi normale centrée en $\varphi(x_i)$ et de variance σ^2. Par continuité de la fonction $t \mapsto e^{-t^2}$, on a

$$P(|Y - y_i| \leq \varepsilon) \simeq \frac{2\varepsilon}{\sqrt{2\pi}\sigma} e^{-(y_i-\varphi(x_i))^2/2\sigma^2} .$$

Notons maintenant $Y = (Y_1, \ldots, Y_n)$ une variable aléatoire reflétant notre expérience, autrement dit, pour un n-uplet $(x_1, \ldots, x_n)$ fixé, un n-uplet de variables aléatoires suivant une loi normale de variance σ^2 et centrées respectivement en $(x_1, \ldots, x_n)$. La probabilité d'observer approximativement des valeurs $(y_1, \ldots, y_n)$ est le produit des quantités précédentes, soit

$$P\left(\sup_{1\leq i\leq n} |Y_i - y_i| \leq \varepsilon\right) \simeq \left(\frac{2\varepsilon}{\sqrt{2\pi}\sigma}\right)^n e^{-\sum_{i=1}^n (y_i-\varphi(x_i))^2/2\sigma^2}$$

et, par conséquent, la probabilité de mesurer effectivement $(\varphi(x_1), \ldots, \varphi(x_n))$ est maximale lorsque la quantité

$$\sum_{i=1}^n (y_i - \varphi(x_i))^2$$

[1] Cette supposition repose sur le théorème de la limite centrale qui exprime grosso modo qu'en moyenne les réalisations successives d'une expérience se distribuent selon une loi normale.

est minimale.

Évidemment si on ne fixe pas de contraintes sur φ, il suffit de trouver une fonction qui vaut exactement y_i en x_i pour minimiser la somme des carrés précédente. Mais en général on attend une certaine forme pour φ, dépendant de certains paramètres, et alors on minimise la somme des carrés relativement à cet ensemble de paramètres. Cette problématique est connue sous le nom de « maximum de vraisemblance ».

4.2 La droite des moindres carrés

On se donne des observations sous la forme de couples (x, y) et on cherche à reconnaître si, aux erreurs d'expérience près, les deux variables sont liées par une relation linéaire. Autrement dit on cherche un couple de réels (a, b) de sorte que la droite d'équation $y = ax + b$ passe le près possible des données observées.

Il faut bien entendu définir cette notion de plus proche. On choisit ici le principe du maximum de vraisemblance et il s'agit donc de minimiser (relativement à a et b) la fonction

$$(a, b) \mapsto \sum_{i=1}^{n} |y_i - (ax_i + b)|^2 \, .$$

L'idée pour résoudre ce problème est de l'interpréter comme un problème d'algèbre linéaire (alors qu'il semblerait qu'il n'ait rien de linéaire). Pour cela introduisons

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \qquad \text{et} \qquad A = \begin{pmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{pmatrix}$$

il s'agit donc de minimiser la quantité

$$||Y - AU||^2 \qquad \text{pour} \quad U = \begin{pmatrix} a \\ b \end{pmatrix}$$

où $||.||$ désigne la norme euclidienne sur $\mathbf{R}^n$.

Il est clair que si Y appartient à l'image de A, U est tout simplement un de ses antécédents. Notons que A est de taille $(n, 2)$ et peut donc être interprétée comme la matrice d'une application linéaire de $\mathbf{R}^2$ dans $\mathbf{R}^n$. Son image est donc, au mieux, un plan de $\mathbf{R}^n$. Il y a donc finalement très peu de chances que Y appartienne à son image, sauf si, bien entendu, on a de bonnes raisons pour croire que ce soit le cas.

Notons que A est de rang 2 ou, ce qui revient au même, est injective si et seulement si tous les x_i sont distincts. On supposera cela dans la suite. Dans ce cas tout élément de $Im(A)$ a un unique antécédent par A et il est obtenu en résolvant un système qui se doit d'être de Cramer.

Remarquons aussi que AU décrit, lorsque U varie, l'image de A et donc on va chercher à minimiser la distance de Y à un point de cette image. Or la distance de Y à un sous-espace vectoriel de $\mathbf{R}^n$ est égale à la distance de Y à sa projection orthogonale sur ce sous-espace, d'après le théorème de Pythagore.

Il nous faut donc calculer la projection orthogonale d'un point Y de $\mathbf{R}^n$ sur $Im(A)$.

On va pour cela étudier une situation générale. Soit A une application linéaire de E dans F, deux espaces vectoriels euclidiens. On note $\langle .,.\rangle_E$ et $\langle .,.\rangle_F$ les produits scalaires sur E et F respectivement.

Lemme 18 *Soit φ une forme linéaire sur E, i.e. une application linéaire de E dans $\mathbf{R}$ ou encore un élément de E^*, il existe un unique vecteur U de E tel que*

$$\forall X \in E \qquad \varphi(X) = \langle X, U\rangle_E \ .$$

L'application $\varphi \mapsto U$ réalise un isomorphisme entre les espaces vectoriels E et E^*. En particulier elle est linéaire.

On étudie l'application f de E dans E^* qui à U associe la forme linéaire $X \mapsto \langle X, U\rangle_E$. C'est une application linéaire puisque le produit scalaire est une application bilinéaire.

Montrons que f est injective. En effet $f(U)$ est nul si et seulement si, pour tout X dans E, $f(U)(X) = 0$, i.e.

$$\forall X \in E \qquad \langle X, U\rangle_E = 0 \ .$$

Puisqu'un produit scalaire est défini, ceci nécessite $U = 0$ (autrement dit $E^\perp = \{0\}$) et donc f est injective. Comme de plus E et E^* sont de même dimension, f est donc bijective, d'où l'assertion.

Lemme 19 *Il existe une unique application linéaire de F dans E, notée A^*, telle que*

$$\forall (X,Y) \in E \times F \qquad \langle AX, Y\rangle_F = \langle X, A^*Y\rangle_E \ .$$

Cette application est nommée adjointe de A et admet pour matrice (dans les bases canoniques) la transposée de celle de A.

Soit Y fixé dans F ; l'application de E dans $\mathbf{R}$ qui à X associe $\langle AX, Y\rangle_F$ est linéaire puisque A l'est et que le produit scalaire est bilinéaire. Il existe donc un unique vecteur U de E tel que

$$\forall X \in E \qquad \langle AX, Y\rangle_F = \langle X, U\rangle_E \ .$$

Notons A^*Y ce vecteur U. Il nous reste à montrer que A^* est une application linéaire.

Soit donc Y et Y' dans F et a, b dans $\mathbf{R}$. On a

$$\forall X \in E \qquad \langle AX, Y\rangle_F = \langle X, A^*Y\rangle_E \qquad \text{et} \qquad \langle AX, Y'\rangle_F = \langle X, A^*Y'\rangle_E$$

et, par bilinéarité du produit scalaire,

$$\begin{aligned}\forall X \in E \qquad \langle AX, (aY + bY')\rangle_F &= a\langle AX, Y\rangle_F + b\langle AX, Y'\rangle_F \\ &= a\langle X, A^*Y\rangle_E + b\langle X, A^*Y'\rangle_E \\ &= \langle X, (aA^*Y + bA^*Y')\rangle_E \ .\end{aligned}$$

Or, par unicité dans le lemme précédent, ceci impose

$$aA^*Y + bA^*Y' = A^*(aY + bY')$$

et donc A^* est linéaire.

Soit maintenant $(e_i)_{i\in I}$ la base canonique de E et $(f_j)_{j\in J}$ celle de F. Le coefficient d'indice (k,l) de A^* est, par définition, la k^e composante de l'image par A^* du l^e vecteur de base de F, autrement dit c'est la composante suivant e_k de A^*f_l. Comme nous avons affaire à des bases orthonormées, c'est donc

$$\langle A^*f_l, e_k\rangle_E$$

soit

$$\langle Ae_k, f_l\rangle_F$$

i.e. le coefficient d'incide (l,k) de A.

En particulier, le rang d'une matrice M étant la dimension maximale d'une matrice carrée inversible extraite de M et cette notion étant invariante par transposition, le rang de A et celui de A^* sont les mêmes.

Proposition 2 *On a*

$$E = Ker(A) \oplus^\perp Im(A^*) \qquad F = Ker(A^*) \oplus^\perp Im(A)\ .$$

Si U appartient au noyau de A et $V = A^*W$ à l'image de A^*, alors

$$\langle U, V\rangle_E = \langle U, A^*W\rangle_E = \langle AU, W\rangle_F = \langle 0, W\rangle_F = 0\ .$$

Autrement dit $Ker(A)$ et $Im(A^*)$ sont orthogonaux.

A fortiori ils sont en somme directe, en effet si U appartient à la fois à $Ker(A)$ et $Im(A^*)$, il est orthogonal à lui-même, i.e. $\langle U, U\rangle_E = ||U||_E^2 = 0$ et donc U est nul, c'est-à-dire $Ker(A) \cap Im(A^*) = \{0\}$.

La dimension de cette somme est donc la somme des dimensions des deux sous-epaces et donc, par le théorème du rang, on a

$$dim(KerA) + dim(ImA^*) = dim(KerA) + rang(A^*) = dim(KerA) + rang(A) = dim(E)$$

et il en résulte

$$Ker(A) \oplus^\perp Im(A^*) = E\ .$$

La seconde assertion résulte de la première appliquée à A^*.

Maintenant que l'on a trouvé l'orthogonal de $Im(A)$, il reste à comprendre comment calculer la projection orthogonale d'un Y de F sur $Im(A)$. D'après ce qui précède c'est l'unique Y_0 tel que

$$Y = Y_0 + Y_1$$

avec Y_0 et Y_1 dans $Im(A)$ et $Ker(A^*)$ respectivement. Notons qu'on a alors $A^*Y = A^*Y_0$ puisque $A^*Y_1 = 0$.

Proposition 3 *On a*

$$Ker(A^*A) = Ker(A) \qquad et \qquad Im(A^*A) = Im(A^*)$$

ainsi que

$$Ker(AA^*) = Ker(A^*) \qquad et \qquad Im(AA^*) = Im(A)\ .$$

Pour tout vecteur colonne U, on a

$$AU = 0 \Rightarrow A^*AU = 0$$

et

$$A^*AU = 0 \Rightarrow \langle A^*AU, U\rangle_E = 0 \Rightarrow ||AU||_F^2 = 0$$

et donc $Ker(A) = Ker(A^*A)$. De plus $Im(A^*A) \subset Im(A^*)$ et

$$rg(A^*) = dim(E) - dim(Ker(A^*)) = dim(E) - dim(Ker(AA^*)) = rg(A^*A)\ .$$

Il en résulte

$$Im(A^*A) = Im(A^*)\ ,$$

ce qui est la première assertion. La seconde résulte de la première appliquée à A^*.

Soit donc de nouveau Y dans F, A^*Y appartient à $Im(A^*)$ et donc, d'après ce qui précède à $Im(A^*A)$. On peut donc trouver U_0 dans E tel que $A^*A(U_0) = A^*Y$. Cet U_0 est défini à un élément de $Ker(A^*A)$ près, c'est-à-dire à un élément de $Ker(A)$ près. Par conséquent AU_0 est uniquement déterminé et on a donc $Y_0 = AU_0$.

En résumé, on a démontré

Théorème 11 *Soit Y dans F, l'ensemble des antécédents par A de la projection orthogonale de Y sur $Im(A)$ est constitué des solutions de l'équation (pour U variant dans E)*

$$A^*A(U) = A^*Y\ .$$

Cet ensemble est un espace affine de dimension $dim(Ker(A))$.

Revenons à notre cas particulier. Ici A est injective et donc l'équation $A^*A(U) = A^*Y$ admet une unique solution en U qui est le vecteur que l'on cherche.

Remarquons maintenant que A^*A est un matrice symétrique puisque

$${}^t(A^*A) = {}^t({}^tAA) = {}^tA{}^t({}^tA) = {}^tAA = A^*A$$

et elle est donc diagonalisable dans une base orthonormée. De plus, dans notre cas, elle de dimension 2 puisque A^* est de taille $(2, n)$ et A de taille $(n, 2)$.

Menons donc le calcul :

$$A = \begin{pmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{pmatrix}$$

et donc

$$A^*A = \begin{pmatrix} x_1^2 + \ldots + x_n^2 & x_1 + \ldots + x_n \\ x_1 + \ldots + x_n & n \end{pmatrix}$$

ou encore

$$A^*A = n \begin{pmatrix} E(X^2) & E(X) \\ E(X) & 1 \end{pmatrix}$$

en notant E l'espérance mathématique (ou moyenne) : si f est une fonction numérique et si $X = (x_1, \ldots, x_n)$ représente un échantillon statistique, on note

$$E(f(X)) = \frac{1}{n} \sum_{i=1}^{n} f(x_i) \ .$$

En particulier $E(X)$ est la moyenne de X et $var(X) = E(X^2) - E(X)$ est la variance de X.

De plus

$$A^*Y = \begin{pmatrix} x_1y_1 + \ldots + x_ny_n \\ y_1 + \ldots + y_n \end{pmatrix}$$

et donc

$$A^*Y = n \begin{pmatrix} E(XY) \\ E(Y) \end{pmatrix} \ .$$

On est donc amené à résoudre le système linéaire $A^*AU = A^*Y$ soit encore

$$\begin{pmatrix} E(X^2) & E(X) \\ E(X) & 1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} E(XY) \\ E(X) \end{pmatrix} \ .$$

Le discriminant de ce système est $E(X^2) - E(X)^2$, c'est-à-dire la variance de X (on retrouve le fait qu'il n'est nul que si tous les x_i sont égaux). Les formules de Cramer donnent donc

$$a = \frac{\begin{vmatrix} E(XY) & E(X) \\ E(Y) & 1 \end{vmatrix}}{var(X)}$$

et

$$b = \frac{\begin{vmatrix} E(X^2) & E(XY) \\ E(X) & E(Y) \end{vmatrix}}{var(X)} \ .$$

Or

$$E(XY) - E(X)E(Y) = E\left((X - E(X)).(Y - E(Y))\right) = cov(X, Y)$$

est la covariance de X et Y (c'est une définition !) et

$$\begin{aligned} E(X^2)E(Y) - E(XY)E(X) &= var(X)E(Y) + E(X)^2E(Y) - E(XY)E(X) \\ &= var(X)E(Y) - E(X)cov(X, Y) \end{aligned}$$

soit

$$a = \frac{cov(X,Y)}{var(X)} \qquad \text{et} \qquad b = E(Y) - \frac{E(X)cov(X,Y)}{var(X)} \ .$$

Par conséquent la droite des moindres carrés admet pour équation

$$y - E(Y) = \frac{cov(X,Y)}{var(X)}\,(x - E(X)) \ .$$

4.3 Quelques généralisations

L'endomorphisme adjoint, que nous avons noté A^* dépend en fait fortement du produit scalaire que l'on a choisi sur E et F. Ici on a fait le choix usuel de prendre des produits scalaires euclidiens, mais il en existe bien d'autres. D'autres choix donneraient d'autres résultats. Par exemple si on choisit le produit scalaire d'indice p (pour p réel strictement supérieur à 1) :

$$||X||_p = \left(\sum_{i=1}^{n} x_i^p\right)^{1/p}$$

l'adjoint A^* est alors défini par $a^*_{ij} = a^{p-1}_{ji}$ et par conséquent la droite minimisant la quantité $\sum_{i=1}^n |y_i - (ax_i - b)|^p$ admet pour équation

$$y - E(Y) = \frac{cov(X^{p-1},Y)}{cov(X^{p-1},X)}\,(x - E(X)) \ .$$

Si l'on désire approcher les valeurs y par des polynômes en les données x, on se ramène encore une fois à un problème linéaire en les coefficients. En effet le problème précédent est tout simplement le cas où l'on approche y par un polynôme de degré 1 en x. Pour le cas général on prend pour A :

$$A = \begin{pmatrix} x_1^d & \cdots & x_1 & 1 \\ \vdots & \ddots & \vdots & \vdots \\ x_n^d & \cdots & x_n & 1 \end{pmatrix} \ .$$

Séance du 27/03/2000

4.4 Une autre droite des moindres carrés

Géométriquement on a vu que $\sum_{i=1}^{n} |y_i - (ax_i + b)|^2$ représente la somme des carrés des écarts verticaux des valeurs observées y_i à la droite d'équation $y = ax + b$. On a vu que cette problématique est issue de la supposition que la grandeur x est connue alors que y est mesurée. Si on essayait au contraire de mesurer x à partir de données connues y, on inverserait le rôle de x et de y et on obtiendrait ainsi la droite des moindres carrés de x relativement à y. Celle-ci a évidemment comme équation

$$x - E(X) = \frac{cov(X,Y)}{var(Y)} (y - E(Y)) \ .$$

Celle-ci n'est donc pas la même que la précédente. Elles passent toutes les deux par le point $(E(X), E(Y))$ et l'angle θ entre ces droites est donné par la formule

$$\tan(\theta) = \frac{\frac{var(Y)}{cov(X,Y)} - \frac{cov(X,Y)}{var(X)}}{1 - \frac{var(Y)}{cov(X,Y)} \frac{cov(X,Y)}{var(X)}} = \frac{var(X)var(Y) - cov^2(X,Y)}{cov(X,Y).(var(X) - var(Y))} \ .$$

Si maintenant les deux quantités x et y sont mesurées avec des erreurs, on peut se poser différentes questions. On peut commencer par comparer les deux droites des moindres carrés et se contenter de cela si elles sont « proches » l'une de l'autre. Mais on peut également tenter de minimiser la somme des carrés des écarts à une droite et non pas les écarts verticaux ($|y - (ax + b)|$) ou horizontaux ($|x - (ay + b)|$).

Rappelons que si une droite Δ admet pour équation $a + bx + cy = 0$ avec (b, c) distinct de $(0, 0)$, la distance d'un point M de coordonnées (x, y) à Δ est donnée par

$$d^2(M, \Delta) = \frac{(a + bx + cy)^2}{b^2 + c^2} \ .$$

Si maintenant λ est un scalaire non nul, remarquons que les droites d'équations $a + bx + cy = 0$ et $\lambda(a + bx + cy) = 0$ sont les mêmes. Aussi en choisissant $\lambda = (b^2 + c^2)^{-1/2}$ on peut ne considérer que les équations de droites pour lesquelles $b^2 + c^2 = 1$. Autrement dit on cherche parmi toutes les droites, modélisées par un triplet de réels (a, b, c) tel que $b^2 + c^2 = 1$, celles pour laquelle la somme des carrés des distances entre les points $M_i = (x_i, y_i)$ observés lors de l'expérience et cette droite est minimale, i.e. on cherche le minimum de la fonction :

$$(a, b, c) \mapsto \sum_{i=1}^{n} \frac{(a + bx_i + cy_i)^2}{b^2 + c^2} = \sum_{i=1}^{n} (a + bx_i + cy_i)^2$$

sur le domaine de $\mathbf{R}^3$ défini par $b^2 + c^2 = 1$.

Écrivons une fois de plus le problème matriciellement. On pose

$$A = \begin{pmatrix} 1 & x_1 & y_1 \\ \vdots & \vdots & \vdots \\ 1 & x_n & y_n \end{pmatrix} \qquad \text{et} \qquad U = \begin{pmatrix} a \\ b \\ c \end{pmatrix} .$$

On a donc

$$AU = \begin{pmatrix} a + bx_1 + cy_1 \\ \vdots \\ a + bx_n + cy_n \end{pmatrix}$$

de sorte que

$$\sum_{i=1}^{n} (a + bx_i + cy_i)^2 = ||AU||^2 .$$

Finalement on est ramené à chercher, étant donné A, un vecteur U tel que $||AU||$ soit minimal sous la condition que sa projection sur le plan engendré par (e_2, e_3) soit unitaire.

On va s'appuyer sur le théorème suivant :

Théorème 12 (Procédé d'orthogonalisation de Gram-Schmidt) *Soit $(u_1, \dots, u_n)$ une base de $\mathbf{R}^n$. On note E_i l'espace engendré par $(u_1, \dots, u_i)$ pour i entier entre 1 et n. Il existe une base orthonormée $(v_1, \dots, v_n)$ de $\mathbf{R}^n$ telle que, en notant F_i l'espace engendré par $(v_1, \dots, v_i)$ pour i entier entre 1 et n, on ait $E_i = F_i$ pour tout indice i.*

Si de plus on impose à $\langle u_i, v_i \rangle$ d'être positif alors cette base est unique. On dit alors que $(v_1, \dots, v_n)$ est obtenue par orthonormalisation de Gram-Schmidt à partir de $(u_1, \dots, u_i)$.

On construit la base $(v_1, \dots, v_n)$ par récurrence sur l'indice i.

Comme $E_1 = \mathbf{R}u_1$, si on veut $F_1 = E_1$ il nous faut prendre v_1 proportionnel à u_1. Pour qu'il soit unitaire et pour que le produit scalaire $\langle u_1, v_1 \rangle$ soit positif, il nous faut nécessairement choisir

$$v_1 = \frac{u_1}{||u_1||} .$$

Supposons maintenant avoir construit une famille orthonormale $(v_1, \dots, v_k)$ pour un certain entier k compris entre 1 et $n-1$ de sorte que $E_i = F_i$ et que $\langle u_i, v_i \rangle$ soit positif pour i inférieur à k.

Comme $E_{k+1} = E_k \oplus \mathbf{R}u_{k+1}$, pour que F_{k+1} soit égal à E_{k+1}, il nous faut prendre v_{k+1} dans E_{k+1} de telle sorte qu'il ne soit pas E_k, i.e. $v_{k+1} = \alpha u_{k+1} + u$ avec u dans E_k et α un scalaire non nul. Le vecteur u est déterminé par le fait que v_{k+1} doit être orthogonal à tous les v_i pour i inférieur à k ou, autrement dit, $-u$ est la projection orthogonale de αu_{k+1} sur E_k. On a donc

$$v_{k+1} = \alpha \left(u_{k+1} - pr_{E_k}(u_{k+1}) \right)$$

et donc, pour que ce vecteur soit unitaire et ait un produit scalaire avec u_{k+1} positif, il nous faut nécessairement choisir

$$v_{k+1} = \frac{u_{k+1} - pr_{E_k}(u_{k+1})}{||u_{k+1} - pr_{E_k}(u_{k+1})||} = \frac{u_{k+1} - \sum_{i=1}^{k} \langle u_{k+1}, v_i \rangle v_i}{||u_{k+1} - \sum_{i=1}^{k} \langle u_{k+1}, v_i \rangle v_i||} .$$

(Ce choix est possible car u_{k+1} n'appartient pas à E_k et donc la norme par laquelle on divise ne peut être nulle.)

Réciproquement on a bien construit une famille orthonormale $(v_1, \ldots, v_{k+1})$ telle que $E_i = F_i$ et que $\langle u_i, v_i \rangle$ soit positif pour i inférieur à $k+1$.

Par le principe de récurrence il existe donc une unique famille vérifiant ces propriétés pour $k = n$, ce qui est l'assertion du théorème.

Soit maintenant A une matrice rectangulaire de taille (n, m) avec $n \geq m$. Soit $(e_1, \ldots, e_m)$ les vecteurs de la base canonique de $\mathbf{R}^m$. Si les vecteurs $(Ae_1, \ldots, Ae_m)$ forment une partie libre de $\mathbf{R}^n$, on peut les compléter en une base $(u_1, \ldots, u_n)$ de $\mathbf{R}^n$ (avec donc $u_i = Ae_i$ pour i entier compris entre 1 et m). Soit $(v_1, \ldots, v_n)$ la base orthonormée de $\mathbf{R}^n$ obtenue par orthonormalisation de Gram-Schmidt à partir de $(u_1, \ldots, u_i)$.

Puisque $(v_1, \ldots, v_n)$ est orthonormale la matrice admettant ces vecteurs pour colonnes est une matrice orthogonale Q. C'est-à-dire qu'elle vérifie

$$^tQQ = Id_n .$$

De plus, de par la propriété $E_i = F_i$ du théorème précédent, la matrice de passage de $(u_1, \ldots, u_n)$ à $(v_1, \ldots, v_n)$ de même que son inverse la matrice de passage de $(v_1, \ldots, v_n)$ à $(u_1, \ldots, u_n)$ sont triangulaires supérieures. Enfin la propriété $\langle u_i, v_i \rangle$ du théorème précédent montre que les diagonales de ces matrices de passage sont strictement positives.

Soit B la matrice dont les colonnes sont données par $(u_1, \ldots, u_n)$, on a donc $B = QT$ où T est une matrice triangulaire supérieure à diagonale strictement positive. En ne prenant que les m premières colonnes des deux termes de cette égalité on obtient :

Théorème 13 (Décomposition QR d'une matrice) *Soit A une matrice rectangulaire de taille (n, m) avec $n \geq m$ et de rang m ; il existe une matrice Q orthogonale de taille n et une matrice rectangulaire de taille (n, m) et « triangulaire supérieure à diagonale positive » R telles que*

$$A = QR .$$

Si $(r_{ij})_{1 \leq i \leq n, 1 \leq j \leq m}$ sont les coefficients de R, la condition « triangulaire supérieure à diagonale positive » siginifie qu'ils sont nuls dès que i est stritement supérieur à j et qu'ils sont strictement positifs lorsque i est égal à j.

Revenons à notre problème de moindres carrés. On cherche U de sorte que $||AU||$ soit minimal parmi les U tels que leur projection sur le plan engendré par (e_2, e_3) soit unitaire. Soit $A = QR$ la décomposition de A obtenue grâce au théorème précédent. On a donc

$$||AU|| = ||QRU|| = ||RU||$$

puisque Q est orthogonale (i.e. est la matrice d'une isométrie : elle préserve la norme). En notant $(r_{ij})_{1\leq i\leq n, 1\leq j\leq 3}$ les coefficients de R, on cherche donc à minimiser

$$|ar_{11} + br_{12} + cr_{13}|^2 + |br_{22} + cr_{23}|^2 + |cr_{33}|^2$$

parmi tous les triplets (a, b, c) tels que $b^2 + c^2 = 1$. Comme r_{11} est strictement positif il nous est permis de choisir a de façon à annuler le premier terme (une fois déterminés b et c). Il nous faut donc minimiser

$$|br_{22} + cr_{23}|^2 + |cr_{33}|^2$$

parmi les couples (b, c) tels que $b^2 + c^2 = 1$.

Notons B la matrice triangulaire supérieure (extraite de T)

$$B = \begin{pmatrix} r_{22} & r_{23} \\ 0 & r_{33} \end{pmatrix} .$$

On cherche donc à minimiser

$$\left|\left| B \begin{pmatrix} b \\ c \end{pmatrix} \right|\right|^2$$

pour des vecteurs unitaires (b, c). Or, si X est un vecteur de $\mathbf{R}^2$,

$$||BX||^2 = \langle BX, BX \rangle = \langle B^*BX, X \rangle$$

et, comme B^*B est une matrice symétrique définie positive, le minimum de cette expression pour X unitaire est obtenu lorsque X est vecteur propre de B^*B associé à la valeur propre minimale de B^*B.

On a ici

$$B^*B = \begin{pmatrix} r_{22}^2 & r_{22}r_{23} \\ r_{22}r_{23} & r_{23}^2 + r_{33}^2 \end{pmatrix}$$

dont le polynôme caractéristique est

$$X^2 - (r_{22}^2 + r_{23}^2 + r_{33}^2)X + r_{22}^2 r_{33}^2$$

et donc sa plus petite valeur propre est

$$\lambda = \frac{1}{2}\left(r_{22}^2 + r_{23}^2 + r_{33}^2 - \sqrt{(r_{22}^2 + r_{23}^2 + r_{33}^2)^2 - 4r_{22}^2 r_{33}^2} \right) .$$

1. Si $r_{23} = 0$ et $r_{22} = r_{33}$, B^*B est scalaire et on prend (b, c) comme on veut (et $a = -(r_{12}b + r_{13}c)/r_{11}$). La valeur minimale de $||AU||^2$ est donc r_{22}^2.

2. Sinon on pose $\alpha = \frac{1}{2}(r_{22}^2 - r_{23}^2 - r_{33}^2)$ et $\beta = r_{22}r_{23}$, de sorte qu'un vecteur propre de B^*B associé à λ est donné par l'équation
$$\left(\alpha + \sqrt{\alpha^2+\beta^2}\right) b + \beta c = 0$$
et on trouve donc
$$b = \pm \frac{\beta}{\sqrt{2\left(\alpha+\sqrt{\alpha^2+\beta^2}\right)\sqrt{\alpha^2+\beta^2}}} \qquad \text{et} \qquad c = \mp\sqrt{\frac{\alpha+\sqrt{\alpha^2+\beta^2}}{2\sqrt{\alpha^2+\beta^2}}}\ .$$
ainsi que $a = -(r_{12}b + r_{13}c)/r_{11}$.

Appliquons cela à notre matrice A. Pour simplifier les calculs, on va chercher l'équation de notre droite sous la forme
$$a + b(x - E(X)) + c(y - E(Y)) = 0\ ,$$
ce qui revient à supposer $E(X) = E(Y) = 0$. On a ici
$$u_1 = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \quad , u_2 = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \qquad \text{et} \qquad u_3 = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}\ .$$

On a donc

1. $v_1 = u_1/||u_1|| = n^{-1/2}u_1$. Par conséquent $r_{11} = n^{-1/2}$.
2. $\langle u_2, v_1\rangle = n^{-1/2}\langle u_2, u_1\rangle = n^{1/2}E(X) = 0$ et donc $r_{12} = 0$.
3. $v_2 = (u_2 - pr_{E_1}u_2)/||u_2 - pr_{E_1}u_2|| = u_2/||u_2|| = (n.var(X))^{-1/2}u_2$ et donc $r_{22} = (n.var(X))^{-1/2}$.
4. $\langle u_3, v_1\rangle = n^{1/2}E(Y) = 0$ et donc $r_{13} = 0$.
5. $p_{E_2}u_3 = \langle u_3, v_2\rangle v_2 = \langle u_3, u_2\rangle u_2/(n.var(X)) = E(XY)u_2/var(X)$ et donc
$$v_3 = \frac{u_3 - cov(X,Y)var(X)^{-1}u_2}{||u_3 - cov(X,Y)var(X)^{-1}u_2||} = \frac{var(X)u_3 - cov(X,Y)u_2}{||var(X)u_3 - cov(X,Y)u_2||}\ .$$
Comme
$$\begin{aligned} ||var(X)u_3 - cov(X,Y)u_2||^2 &= var(X)^2||u_3||^2 - 2var(X)cov(X,Y)\langle u_2, u_3\rangle \\ &\quad + cov^2(X,Y)||u_2||^2 \end{aligned}$$

on a

$$||var(X)u_3 - cov(X,Y)u_2||^2 = n.var(X)\left(var(X)var(Y) - cov^2(X,Y)\right)$$

et il en résulte

$$r_{23} = -\frac{cov(X,Y)}{\sqrt{n.var(X)\left(var(X)var(Y) - cov^2(X,Y)\right)}}$$

et

$$r_{33} = \sqrt{\frac{var(X)}{n\left(var(X)var(Y) - cov^2(X,Y)\right)}}\,.$$

En particulier, comme $r_{12} = r_{13} = 0$, on voit que a est nul : la droite des moindres carrés passe par le point $(E(X), E(Y))$, comme prévu.

Le cas $r_{23} = r_{22} - r_{33} = 0$ est obtenu lorsque $cov(X,Y) = 0$ et $var(X) = var(Y)$. Dans ce cas toute droite passant par $(E(X), E(Y))$ convient ...Sinon on termine les calculs comme dans le cas général.

Séance du 17/04/2000

5 Conditionnement

5.1 Introduction

Quand on veut résoudre le système général donnant la solution des moindres carrés pour une approximation par un polynôme de degré m, i.e. quand on veut minimiser

$$\sum_{i=1}^{n} (y_i - (a_0 + a_1 x_i + \ldots + a_m x_i^m))^2$$

par rapport à $(a_0, \ldots, a_m)$, on est amené à considérer

$$A = \begin{pmatrix} 1 & x_1 & \ldots & x_1^m \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & \ldots & x_n^m \end{pmatrix} \qquad \text{et} \qquad U = \begin{pmatrix} a_0 \\ \vdots \\ a_m \end{pmatrix} .$$

On prend n strictement supérieur à m pour que le problème soit intéressant et on s'intéresse au système $A^*AU = A^*Y$, i.e.

$$\begin{pmatrix} E(1) & E(X) & \ldots & E(X^m) \\ \vdots & \vdots & \ddots & \vdots \\ E(X^m) & E(X^{m+1}) & \ldots & E(X^{2m}) \end{pmatrix} U = \begin{pmatrix} E(Y) \\ \vdots \\ E(X^m Y) \end{pmatrix} .$$

Supposons que les $(x_i)_{1 \leq i \leq n}$ soient uniformément distribués sur $[0,1]$ (comme on est tenté de le faire si on veut décrire précisément un phénomène continu). On aura alors

$$E(X^k) \simeq \int_0^1 x^k dx = \frac{1}{k+1}$$

et on est donc amené à considérer la matrice (dite de Hilbert)

$$H = H_{m+1} = \begin{pmatrix} 1 & \frac{1}{2} & \ldots & \frac{1}{m+1} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{m+1} & \frac{1}{m+2} & \ldots & \frac{1}{2m+1} \end{pmatrix} .$$

Cette matrice à l'inconvénient d'être très instable. C'est-à-dire que si U est solution de $HU = V$ alors la solution de $H'X = V'$ pour H' et V' respectivement proches de H et V n'est pas nécessairement proche de U. Voyons cela pour H_3 et $v = e_1$. On a

$$\begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{3} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \end{pmatrix} U = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \Leftrightarrow U = \begin{pmatrix} 9 \\ -36 \\ 30 \end{pmatrix} .$$

Effectuons un calcul à 10^{-2} près. On a donc

$$H' = \begin{pmatrix} 1 & 0.5 & 0.3 \\ 0.5 & 0.3 & 0.25 \\ 0.3 & 0.25 & 0.2 \end{pmatrix}$$

et V est inchangé. On trouve alors comme solution

$$U' = \begin{pmatrix} -7.59 \\ 42.12 \\ -40.50 \end{pmatrix},$$

ce qui est bien loin du U précédent.

Prenons maintenant 10^{-2} comme précision, i.e. effectuons les calculs en ne gardant que deux chiffres après la virgules dans tous les calculs. La réduction par pivot de Gauß donne

$$\left\{ \begin{array}{rcrcrcl} 1.0a & + & 0.5b & + & 0.33c & = & 1.0 \\ & & 0.08b & + & 0.08c & = & -0.5 \\ & & & & 0.01c & = & 0.17 \end{array} \right.$$

et donc

$$U' = \begin{pmatrix} 7 \\ -23 \\ 17 \end{pmatrix},$$

ce qui est de pire en pire.

5.2 Origine des erreurs

La représentation des coefficients d'un système linéaire dans un ordinateur, ne possédant qu'un nombre limité de chiffres significatifs, entraîne automatiquement des erreurs. C'est le cas pour $1/3$ dans l'exemple précédent. Avec n chiffres significatifs, on commet une erreur relative de 10^{-n} puisque $(1/3 - 0.33...3)/(1/3) = 1 - 0.99...9$.

Vient ensuite l'utilisation d'un algorithme et l'utilisation d'opérations élémentaires : addition, soustraction, multiplication, division. Chaque résultat est obtenu par arrondi. En conséquence au lieu de résoudre $AX_0 = Y_0$, on résout souvent $(A+\Delta A)X = Y_0 + \Delta Y$. Évidemment les perturbations sont inconnues mais on sait toujours les majorer via leur origine. Puisqu'on s'intéresse à la solution du système linéaire, on aimerait estimer la différence $X - X_0$ (par exemple en norme) en fonction de majorations similaires sur ΔA et ΔY.

Donnons un nouvel exemple :

$$A = \begin{pmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{pmatrix} \qquad Y_0 = \begin{pmatrix} 32 \\ 23 \\ 33 \\ 31 \end{pmatrix}$$

et

$$\Delta A = \begin{pmatrix} 0 & 0 & 0.1 & 0.2 \\ 0.08 & 0.04 & 0 & 0 \\ 0 & -0.02 & -0.11 & 0 \\ -0.01 & -0.01 & 0 & -0.02 \end{pmatrix} \qquad \Delta Y = \begin{pmatrix} 0.01 \\ -0.01 \\ 0.01 \\ -0.01 \end{pmatrix} .$$

On trouve $X_0 = e_1 + e_2 + e_3$. La solution de $AX = Y_0 + \Delta Y$ est

$$\begin{pmatrix} 1.82 \\ -0.36 \\ 1.35 \\ 0.79 \end{pmatrix}$$

et donc une perturbation relative de Y de l'ordre de 3.10^{-4} engendre une erreur relative de 0.8 sur X, soit une multiplication des erreurs par 2500.

Pire encore la solution de $(A + \Delta A)X = Y_0$ est

$$\begin{pmatrix} -81 \\ 137 \\ -34 \\ 22 \end{pmatrix}$$

et on ne peut même plus parler de perturbation dans ce cas !

5.3 Conditionnement

Définition 1 *Soit $||.||$ une norme matricielle ; le conditionnement d'une matrice inversible A, associé à cette norme, est le nombre $cond(A)$ défini par $cond(A) = ||A||.||A^{-1}||$.*

On notera $cond_p(A)$ le conditionnement associé à la norme $||.||_p$ sur $\mathbf{R}^n$, pour p dans $[1; +\infty]$. Rappelons que si X est un vecteur de coordonnées $(x_i)_{1 \leq i \leq n}$, on définit

$$||X||_p = (x_1^p + \ldots + x_n^p)^{1/p} \qquad \text{et} \qquad ||X||_\infty = \sup_{1 \leq i \leq n} |x_i|$$

et, pour une matrice A

$$||A||_p = \sup_{||X||_p = 1} ||AX||_p = \sup_{X \neq 0} \frac{||AX||_p}{||X||_p} .$$

Théorème 14 *Soit A une matrice inversible, α un scalaire, p dans $[1; +\infty]$ et $\lambda_n \leq \ldots \leq \lambda_1$ les valeurs singulières de A (i.e. les racines carrées des valeurs propres de A^*A). On a*

1. $cond_p(\alpha A) = cond_p(A)$

2. $cond_p(A) \geq 1$

3. $cond_2(A) = \lambda_1/\lambda_n$

4. $cond_2(A) = 1$ *si et seulement si* A *est une similitude, i.e. de la forme* λQ *avec* Q *orthogonale et* λ *un scalaire non nul.*

En effet, on a

1. $||\alpha A||_p = |\alpha|.||A||_p$ et $(\alpha A)^{-1} = \alpha^{-1}A^{-1}$ et donc
$$cond_p(\alpha A) = |\alpha|.||A||_p.|\alpha|^{-1}.||A^{-1}||_p = cond_p(A) \; .$$

2. Si A et B sont deux matrices et X un vecteur
$$||ABX||_p \leq ||A||_p.||BX||_p \leq ||A||_p||B||_p||X||_p$$
et donc $||AB||_p \leq ||A||_p||B||_p$. Comme $||Id||_p = 1$, on en déduit
$$1 = ||AA^{-1}||_p \leq ||A||_p||A^{-1}||_p = cond_p(A) \; .$$

3. Si X est un vecteur et A une matrice
$$||AX||_2^2 = \langle AX, AX\rangle = \langle X, A^*AX\rangle$$
et A^*A est une matrice symétrique. Soit $\lambda_1^2 \geq \ldots \geq \lambda_n^2$ ses valeurs propres (elles sont strictement positives en raison de la relation précédente et du fait que A est inversible) et $(u_1, \ldots, u_n)$ une base orthonormée de vecteurs propres associée à ces valeurs propres. Si X est un vecteur, on peut l'écrire
$$X = \alpha_1 u_1 + \ldots + \alpha_n u_n$$
et donc
$$A^*AX = \lambda_1^2\alpha_1 u_1 + \ldots + \lambda_n^2\alpha_n u_n \qquad \text{et} \qquad \langle X, A^*AX\rangle = \lambda_1^2\alpha_1^2 + \ldots + \lambda_n^2\alpha_n^2 \; .$$
On en tire
$$||AX||_2^2 \leq \lambda_1^2||X||_2^2$$
avec égalité quand X est proportionnel à u_1 et donc $||A||_2 = \lambda_1$. Comme, de plus,
$$\left(A^{-1}\right)^* A^{-1} = (AA^*)^{-1} = \left(A(A^*A)A^{-1}\right)^{-1}$$
les valeurs propres de $(A^{-1})^* A^{-1}$ sont les inverses de celles de AA^* et ces dernières sont les mêmes que celles de A^*A. Par conséquent la plus grande des valeurs propres de $(A^{-1})^* A^{-1}$ est l'inverse de la plus petite des valeurs propres de A^*A, i.e.
$$||A^{-1}||_2 = \frac{1}{\lambda_n} \; .$$
La troisième assertion en résulte.

4. Pour que $cond_2(A)$ vaille 1, il faut et il suffit que la plus grande et la plus petite des valeurs propres de A^*A soient égales. Comme cette matrice est diagonalisable (car symétrique), cela signifie qu'elle est scalaire. On a donc $A^*A = \lambda_1^2 Id$ et donc $\lambda_1^{-1}A$ est une matrice orthogonale (i.e. son produit avec sa transposée est égal à l'identité). Il revient au même de dire que A est la matrice d'une similitude de rapport λ_1.

Il faut faire attention que, même s'il est relié aux valeurs propres, le conditionnement d'une matrice ne dépend que très peu du déterminant de la matrice. Par exemple si A est la matrice de diagonale 1 et de surdiagonale 2 (tous les autres coefficients étant nuls), on a $Ae_i = e_i + 2e_{i-1}$ pour i entre 2 et n et $Ae_1 = e_1$ et donc l'inverse de A est donné par

$$A^{-1}e_i = \sum_{j=1}^{i}(-2)^{i-j}e_j$$

et, par conséquent,

$$||A||_\infty = ||A(e_1 + e_2)||_\infty = 3$$

et

$$||A^{-1}||_\infty = ||A^{-1}(e_1 - e_2 + \ldots + (-1)^n e_n)||_\infty = 1 + 2 + \ldots + 2^{n-1} = 2^n - 1$$

de sorte que

$$cond_\infty(A) = 3(2^n - 1) >> det(A) = 1 \ .$$

Théorème 15 *Si A est inversible et si X et $X+\Delta X$ vérifient $AX = Y$ et $A(X+\Delta X) = Y + \Delta Y$ pour un certain Y non nul et un certain $Y + \Delta Y$, alors*

$$\frac{||\Delta X||_p}{||X||_p} \leq cond_p(A)\frac{||\Delta Y||_p}{||Y||_p} \ .$$

En effet, on a $A(\Delta X) = \Delta Y$ et donc $\Delta X = A^{-1}(\Delta Y)$. Par conséquent

$$||\Delta X||_p \leq ||A^{-1}||_p.||\Delta Y||_p$$

et (puisque Y est non nul, X ne l'est pas non plus)

$$||Y||_p = ||AX||_p \leq ||A||_p.||X||_p \qquad \text{soit} \qquad \frac{1}{||X||_p} \leq \frac{||A||_p}{||Y||_p} \ ,$$

d'où le résultat.

Remarquons au passage que $cond_p(A)$ est la meilleure valeur possible puisque l'on peut trouver un couple $(Y, \Delta Y)$ pour lequel les inégalités sont des égalités.

Théorème 16 *Si A et $A+\Delta A$ sont des matrices inversibles et si X et $X+\Delta X$ vérifient $AX = Y$ et $(A+\Delta A)(X+\Delta X) = Y$ pour un certain Y non nul, on a*

$$\frac{||\Delta X||_p}{||X + \Delta X||_p} \leq cond_p(A)\frac{||\Delta A||_p}{||A||_p} \ .$$

En soustrayant les deux équations, on obtient

$$A(\Delta X) + \Delta A(X + \Delta X) = 0$$

et donc

$$\begin{aligned} ||\Delta X||_p &= || - A^{-1}.\Delta A.(X + \Delta X)||_p \leq ||A^{-1}||_p.||\Delta A||_p.||X + \Delta X||_p \\ &= cond_p(A)\frac{||\Delta A||_p}{||A||_p}.||X + \Delta X||_p \, . \end{aligned}$$

Le théorème en résulte puisque Y et donc $X + \Delta X$ ne sont pas nuls.

Effectuons un retour sur l'exemple avec

$$A = \begin{pmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{pmatrix} \qquad \text{et} \qquad Y_0 = \begin{pmatrix} 32 \\ 23 \\ 33 \\ 31 \end{pmatrix} \, .$$

On a $||Y_0||_2 \simeq 60$ et $||\Delta Y_0||_2 = 0.02$. Comme A est symétrique, $||A||_2$ est la plus grande des valeurs propres de A, soit environ 30 tandis que $||A^{-1}||_2$ est l'inverse de la plus petite des valeurs propres de A, soit environ 100. D'où un conditionnement $cond_2(A) \simeq 3000$.

On a trouvé X_0 tel que $||X_0||_2 = 2$, $||\Delta X_0||_2 \simeq 1.64$ et donc

$$\frac{||\Delta X_0||_2}{||X_0||_2} \simeq 0.82 \sim 3000.\frac{0.02}{60} = 1 \, .$$

Remarquons également que A admet 1 pour déterminant.

5.4 Application aux moindres carrés

On étudie la projection orthogonale d'un certain Y sur l'image de A. Soit X tel que AX soit cette projection (on supposera X unique, i.e. A de taille (m, n) et de rang n, avec $n \leq m$). Soit $R = Y - AX$ le « résidu » du problème des moindres carrés et θ l'angle entre AX et Y. On a donc

$$\sin\theta = \frac{||R||_2}{||Y||_2} \, .$$

Soit maintenant $X + \Delta X$ la projection orthogonale du même Y sur l'image d'une matrice perturbée $A + \Delta A$. On suppose que $||\Delta A||_2/||A||_2$ est négligeable devant les autres quantités. Dans ce cas on a

$$\begin{cases} \frac{||\Delta X||_2}{||X||_2} & \leq \quad \varepsilon.cond_2(A)\,(1 + cond_2(A).\tan(\theta)) + O(\varepsilon^) \\ \frac{||\Delta R||_2}{||Y||_2} & \leq \quad 2\varepsilon.cond_2(A) + O(\varepsilon^2) \, . \end{cases}$$

On a en effet, par soustraction,

$$\Delta R = R + \Delta R - R = (Y - AX) - (Y - (A + \Delta A)(X + \Delta X)) = -A.\Delta X - \Delta A(X + \Delta X) \; .$$

Comme

$$A^*R = A^*Y - A^*AX = 0 \qquad \text{et} \qquad (A + \Delta)^*(R + \Delta R) = 0$$

on a

$$\begin{array}{rcl} (\Delta A)^*R + (A + \Delta A)^*\Delta R & = & 0 \\ (\Delta A)^*R - (A + \Delta A)^*.A.\Delta X - (A + \Delta A)^*.\Delta A.(X + \Delta X) & = & 0 \\ (\Delta A)^*R - (A + \Delta A)^*.\Delta A.(X + \Delta X) & = & (A + \Delta A)^*.A.\Delta X \\ (\Delta A)^*R - A^*.\Delta A.X + O(\varepsilon^2) & \simeq & A^*A.\Delta X + O(\varepsilon^2) \end{array}$$

et donc, puisque A^*A est de rang égal à celui de A, i.e. n, et est de taille (n, n), elle est inversible et on a

$$\Delta X = (A^*A)^{-1}.(\Delta A)^*.R - (A^*A)^{-1}.A^*.\Delta A.X + O(\varepsilon^2) \; .$$

On a également

$$\begin{array}{rcl} \Delta R & = & -A.\Delta X - \Delta A(X + \Delta X) \\ & = & -A.\Delta X - \Delta A.X + O(\varepsilon^2) \\ & = & -A(A^*A)^{-1}(\Delta A)^*R - \Delta A.X - A(A^*A)^{-1}A^*\Delta A.X + O(\varepsilon^2) \\ & = & - - A(A^*A)^{-1}(\Delta A)^*R - (Id - A(A^*A)^{-1}A^*)\Delta A.X + O(\varepsilon^2) \end{array}$$

et donc

$$\left\{ \begin{array}{rcl} ||\Delta X||_2 & \leq & ||(A^*A)^{-1}||_2.||\Delta A||_2.||R||_2 + ||(A^*A)^{-1}A^*||_2.||\Delta A||_2.||X||_2 + O(\varepsilon^2) \\ ||\Delta R||_2 & \leq & ||A(A^A)^{-1}||_2.||(\Delta A)^*||_2.||R||_2 + ||Id - A(A^*A)^{-1}A^*||_2||\Delta A||_2.||X||_2 + O(\varepsilon^2) \; . \end{array} \right.$$

La matrice A^*A est symétrique de taille (n, n). Soit donc Q orthogonale de taille (n, n) telle que $Q^*(A^*A)Q$ soit diagonale avec les λ_i^2 sur la diagonale. On note u_i le vecteur $\lambda_i^{-1}AQe_i$. Cette famille de vecteurs de $\mathbf{R}^m$ est orthonormale et on peut donc la compléter en une base orthonormée de $\mathbf{R}^m$. Notons U la matrice orthogonale dont les colonnes sont les $(u_i)_{1 \leq i \leq m}$. On a donc

$$U^*AQe_i = U^*(\lambda_i u_i) = \lambda_i e_i$$

soit

$$U^*AQ = \left(\begin{array}{ccc} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \\ \hline & 0 & \end{array} \right)$$

et donc

$$A = U \left(\begin{array}{ccc} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \\ \hline & 0 & \end{array} \right) Q^* \ .$$

On en déduit immédiatement

$$\begin{array}{rcl}
A^*A & = & Q \begin{pmatrix} \lambda_1^2 & & \\ & \ddots & \\ & & \lambda_n^2 \end{pmatrix} Q^* \\
(A^*A)^{-1}A^* & = & Q \left(\begin{array}{ccc|c} \lambda_1^{-1} & & & \\ & \ddots & & 0 \\ & & \lambda_n^{-1} & \end{array} \right) U^* \\
A(A^*A)^{-1} & = & U \left(\begin{array}{ccc} \lambda_1^{-1} & & \\ & \ddots & \\ & & \lambda_n^{-1} \\ \hline & 0 & \end{array} \right) Q^* \\
A(A^*A)^{-1}A^* & = & \left(\begin{array}{c|c} Id & 0 \\ \hline 0 & 0 \end{array} \right) \ .
\end{array}$$

Par conséquent

$$\left\{ \begin{array}{rcl} ||\Delta X||_2 & \leq & \frac{1}{\lambda_n^2}\varepsilon\lambda_1||R||_2 + \frac{1}{\lambda_n}\varepsilon\lambda_1||X||_2 + O(\varepsilon^2) \\ ||\Delta R||_2 & \leq & \frac{1}{\lambda_n}\varepsilon\lambda_1||R||_2 + 1\varepsilon\lambda_1||X||_2 + O(\varepsilon^2) \ . \end{array} \right.$$

On a de plus

$$\frac{||R||_2}{||X||_2} = \frac{||AX||_2}{||X||_2}\frac{||R||_2}{||AX||_2} \leq \tan(\theta)\lambda_1 \ ,$$

d'où

$$||\Delta X||_2 \leq \varepsilon.cond_2(A)\left(1 + \tan(\theta)cond_2(A)\right) + O(\varepsilon^2)$$

et

$$\left\{ \begin{array}{rcl} ||X||_2 & = & ||(A^*A)^{-1}AY||_2 \leq ||(A^*A)^{-1}A||_2||Y||_2 \leq \frac{1}{\lambda_n}||Y||_2 \\ ||R||_2 & = & ||Y - AX||_2 \leq ||Y||_2 \end{array} \right.$$

et, finalement,

$$\frac{||\Delta R||_2}{||Y||_2} \leq 2\varepsilon.cond_2(A) + O(\varepsilon^2) \ .$$