Méthode d'estimation à posteriori d'erreurs

Yves Ducrocq

N° d'ordre

THESE

présentée à la Faculté des Sciences
de l'Université de Grenoble

pour obtenir
le grade de Docteur de Troisième Cycle
"MATHEMATIQUES APPLIQUEES"

par

Yves DUCROCQ
Licencié ès Sciences

METHODE D'ESTIMATION A POSTERIORI D'ERREURS

Thèse soutenue le 5 février 1968, devant la Commission d'Examen :

Monsieur	J. KUNTZMANN	Président
Messieurs	N. GASTINEL	Examinateur
	P.J. LAURENT	Examinateur

N° d'ordre

T H E S E

présentée à la Faculté des Sciences
de l'Université de Grenoble

pour obtenir
le grade de Docteur de Troisième Cycle
"MATHEMATIQUES APPLIQUEES"

par

Yves DUCROCQ
Licencié ès Sciences

METHODE D'ESTIMATION A POSTERIORI D'ERREURS

Thèse soutenue le 5 février 1968, devant la Commission d'Examen :

Monsieur	J. KUNTZMANN	Président
Messieurs	N. GASTINEL	Examinateur
	P.J. LAURENT	Examinateur

Je remercie Monsieur KUNTZMANN, Directeur de l'Institut de Mathématiques Appliquées de Grenoble d'avoir bien voulu présider le Jury de cette thèse,

Je remercie Monsieur GASTINEL, Professeur à la Faculté des Sciences de Grenoble qui m'a suggéré les sujets de cette thèse et qui, par sa patience et ses conseils, a aussi permis son achèvement,

Je remercie aussi Monsieur LAURENT, Maître de Conférences qui m'initiait au calcul numérique, d'avoir bien voulu faire partie du Jury de cette thèse,

et enfin, je remercie tous ceux qui m'ont aidé, soutenu, tous ceux qui ont oeuvré si bien à la réalisation matérielle de cette thèse.

FACULTE DES SCIENCES

LISTE DES PROFESSEURS

DOYENS HONORAIRES :

M. MORET

M. WEIL

DOYEN :

M. BONNIER E.

PROFESSEURS TITULAIRES :

MM.	NEEL Louis	Chaire de Physique Expérimentale
	HEILMANN René	Chaire de Chimie
	KRAVTCHENKO Julien	Chaire de Mécanique Rationnelle
	CHABAUTY Claude	Chaire de calcul différentiel et intégral
	BENOIT Jean	Chaire de Radioélectricité
	CHENE Marcel	Chaire de Chimie Papetière
	WEIL Louis	Chaire de Thermodynamique
	FELICI Noël	Chaire d'Electrostatique
	KUNTZMANN Jean	Chaire de Mathématiques Appliquées
	BARBIER Reynold	Chaire de Géologie Appliquée
	SANTON Lucien	Chaire de Mécanique des Fluides
	OZENDA Paul	Chaire de Botanique
	FALLOT Maurice	Chaire de Physique Industrielle
	KOSZUL Jean-Louis	Chaire de Mathématiques M.P.C.
	GALVANI O.	Mathématiques
	MOUSSA André	Chaire de Chimie Nucléaire
	TRAYNARD Philippe	Chaire de Chimie Générale

	SOUTIF Michel	Chaire de Physique Générale
	CRAYA Antoine	Chaire d'Hydrodynamique
	REULOS R.	Théorie des Champs
	BESSON Jean	Chaire de Chimie
	AYANT Yves	Physique Approfondie
	GALLISSOT	Mathématiques
Melle	LUTZ Elisabeth	Mathématiques
MM.	BLAMBERT Maurice	Chaire de Mathématiques
	BOUCHEZ Robert	Physique Nucléaire
	LLIBOUTRY Louis	Géophysique
	MICHEL Robert	Chaire de Minéralogie et Pétrographie
	BONNIER Etienne	Chaire d'Electrochimie et d'Electrométallurgie
	DESSAUX Georges	Chaire de Physiologie Animale
	PILLET E.	Chaire de Physique Industrielle et Electrotechnique
	YOCCOZ Jean	Chaire de Physique Nucléaire Théorique
	DEBELMAS Jacques	Chaire de Géologie Générale
	GERBER R.	Mathématiques
	PAUTHENET R.	Electrotechnique
	VAUQUOIS B.	Chaire de Calcul Electronique
	BARJON R.	Physique Nucléaire
	BARBIER Jean-Claude	Chaire de Physique
	SILBER R.	Mécanique des Fluides
	BUYLE-BODIN Maurice	Chaire d'Electronique
	DREYFUS B.	Thermodynamique
	KLEIN J.	Mathématiques
	VAILLANT F.	Zoologie et Hydrobiologie
	ARNOUD Paul	Chaire de Chimie M.P.C.
	SENGEL P.	Chaire de Zoologie
	BARNOUD F.	Chaire de Biosynthèse de la Cellulose
	BRISSONNEAU P.	Physique
	GAGNAIRE Didier	Chaire de Chimie Physique

Mme	KOFLER L.	Botanique
MM.	DEGRANGE Charles	Zoologie
	PEBAY-PEROULA J.C.	Physique
	RASSAT A.	Chaire de Chimie Systématique

PROFESSEURS SANS CHAIRE :

MM.	GIDON P.	Géologie et Minéralogie
	GIRAUD P.	Géologie
	PERRET R.	Servomécanismes
Mme	BARBIER M.J.	Electrochimie
Mme	SOUTIF J.	Physique
MM.	COHEN J.	Electrotechnique
	DEPASSEL R.	Mécanique des Fluides
	GASTINEL N.	Mathématiques Appliquées
	ANGLES-d'AURIAC P.	Mécanique des Fluides
	DUCROS P.	Minéralogie et Cristallographie
	GLENAT R.	Chimie
	LACAZE A.	Thermodynamique
	BARRA J.	Mathématiques Appliquées
	COUMES A.	Electronique
	PERRIAUX J.	Géologie et Minéralogie
	ROBERT A.	Chimie Papetière
	BIAREZ J.P.	Mécanique Physique
	BONNET G.	Electronique
	CAUQUIS G.	Chimie Générale
	BONNETAIN L.	Chimie Minérale
	DEPOMMIER P.	Etude Nucléaire et Génie Atomique
	HACQUES Gérard	Calcul Numérique
	POLOUJADOFF M.	Electrotechnique

MAITRES DE CONFERENCES :

MM.	DODU J.	Mécanique des Fluides
	LANCIA Roland	Physique Automatique
Mme	KAHANE J.	Physique
MM.	DEPORTES C.	Chimie
Mme	BOUCHE L.	Mathématiques
MM.	SARROT-RAYNAUD J.	Géologie Propédeutique
Mme	BONNIER M.J.	Chimie
MM.	KAHANE A.	Physique Générale
	DOLIQUE J.M.	Electronique
	BRIERE G.	Physique M.P.C.
	DESPRE P.	Chimie S.P.C.N.
	LAJZEROWICZ J.	Physique M.P.C.
	VALENTIN P.	Physique M.P.C.
	BERTRANDIAS J.P.	Mathématiques Appliquées T.M.P.
	LAURENT P.	Mathématiques Appliquées T.M.P.
	CAUBET J.P.	Mathématiques Pures
	PAYAN J.J.	Mathématiques
Mme	BERTRANDIAS F.	Mathématiques Pures M.P.C.
MM.	LONGEQUEUE J.P.	Physique
	NIVAT M.	Mathématiques Appliquées
	SOHM J.C.	Electrochimie
	ZADWORNY F.	Electronique
	DURAND F.	Chimie Physique
	CARLIER G.	Biologie Végétale
	AUBERT G.	Physique M.P.C.
	DELPUECH J.J.	Chimie Organique
	PFISTER J.C.	Physique C.P.E.M.
	CHIBON P.	Biologie Animale
	IDELMAN S.	Physiologie Animale
	BLOCH D.	Electrotechnique
	BRUGEL L.	I.U.T.
	SIBILLE R.	I.U.T.

TABLE DES MATIERES

S O M M A I R E

La partie A de ce travail est consacrée à l'étude à postériori de l'erreur commise dans la résolution d'un système par les moindres carrés. Cette partie est celle qui est susceptible du plus grand nombre d'applications pratiques dans toutes les sciences expérimentales.

La partie B est intéressante au point de vue théorique. Elle parle de l'estimation à postériori de l'erreur sur une ou plusieurs racines numériques d'un polynôme mais me semble beaucoup moins pratique que l'estimation directe.

C traite, elle, du problème des valeurs propres. Nous commençons par estimer à postériori l'erreur quand nous avons une valeur propre et le vecteur propre correspondant puis nous en déduisons l'erreur à postériori sur la valeur propre seule.

Pour ces trois parties, nous donnons une procédure qui réalise effectivement le calcul de cette erreur à postériori. Nous signalons qu'elles sont toutes assez rapides (temps moyen : 20 millièmes d'heures, soit un peu plus d'une minute).

Enfin dans la partie D nous estimons à postériori l'erreur commise dans la résolution d'un système linéaire et ce, dans le cadre d'un espace de Hilbert. Pour cela nous commençons par définir ce que nous appelerons une double norme, puis, à l'aide d'une hypothèse exprimée en utilisant cette double norme, nous étudierons l'erreur à postériori. Nous ferons une application à la méthode de Galerkin des résultats trouvés. Nous particulariserons pour le problème de Dirichlet et l'équation de Fredholm.

La bibliographie de chaque partie se trouve à la fin de celle-ci.

INTRODUCTION

L'objet de cette thèse étant l'estimation à postériori de l'erreur dans différents cas, nous commencerons par expliquer ce qu'est cette estimation. Nous emprunterons à Jean Gâches [1] ces explications.

Jean Gâches a résolu ce problème dans le cas de systèmes linéaires.

Ayant appliqué un algorithme à un ensemble de données, il importe de connaître la signification du résultat numérique obtenu. C'est-à-dire de préciser dans quelle mesure ce résultat est acceptable, compte-tenu du problème proposé.

La solution exacte ne peut être atteinte qu'exceptionnellement. En effet, on utilise généralement des procédés itératifs qui convergent vers la solution exacte mais on ne peut évidemment aller à la limite, on doit se contenter d'une approximation ; d'autre part il y a aussi les erreurs d'arrondi (dûes au nombre fini et constant de bits par mémoire).

Il y a aussi l'incertitude des données dûes soit à leur nature expérimentale (réalisation non idéale de l'expérience, précision des appareils, qualités de l'observateur) soit, aussi du fait, déjà vu, des troncatures de nombres à leur entrée en machine et de leur transformation en binaire.

Pour caractériser la qualité des résultats obtenus on utilise :

L'analyse directe.

Soit $\tilde{x}$ la solution numérique et x^* la solution exacte d'un problème. C'est le calcul d'erreur habituel et il conduit à des inégalités de la forme $||x^*-\tilde{x}|| \leqslant \varepsilon$ mais en général si le nombre d'opérations est grand cela conduit à des majorations trop grossières.

L'analyse rétrograde.

Ici on recherche de quels problèmes $\tilde{x}$ est-il solution.

(1) $f(\tilde{x}) = \varepsilon$ ε est le résidu

(2) $f'(\tilde{x}) = 0$

$||f - f'|| < \varepsilon$

mais ceci conduit encore à des majorations grossières de l'erreur.

L'analyse à postériori.

C'est un raffinement de la précédente en ce sens que nous chercherons, non pas un problème quelconque vérifié par $\tilde{x}$, mais le problème le plus "proche" vérifié par $\tilde{x}$. Nous dirons que f^* est indiscernable de f si $||f^* - f|| \leq \varepsilon^*$ (quantité donnée à l'avance). Nous dirons alors qu'une solution $\tilde{x}$ est compatible avec un problème f si le problème f^*, vérifié par $\tilde{x}$, le plus proche est indiscernable de f. La valeur de l'analyse à postériori vient du fait que ε^* est aisément calculé à partir des incertitudes sur les données du problème.

ETUDE DE L'ERREUR A POSTERIORI

DANS LE CAS

DES MOINDRES CARRES

CHAPITRE I

INTRODUCTION

On cherche souvent, en physique, à déterminer la valeur numérique d'un groupe de paramètres à l'aide d'une suite nombreuse de mesures. Ces mesures forment un tableau de la forme $A_o X = B_o$. A_o et B_o étant des matrices dont les éléments sont connus avec les marges d'incertitude ε_1 et ε_2. X étant le vecteur inconnu à déterminer.

On obtient généralement la valeur des paramètres par la méthode des moindres carrés, si bien qu'on est amené à étudier le système quadratique $A_o^T A_o X = A_o^T B_o$.

L'objet de la partie A de ce travail est justement de déterminer l'erreur à postériori dans la résolution de ce système.

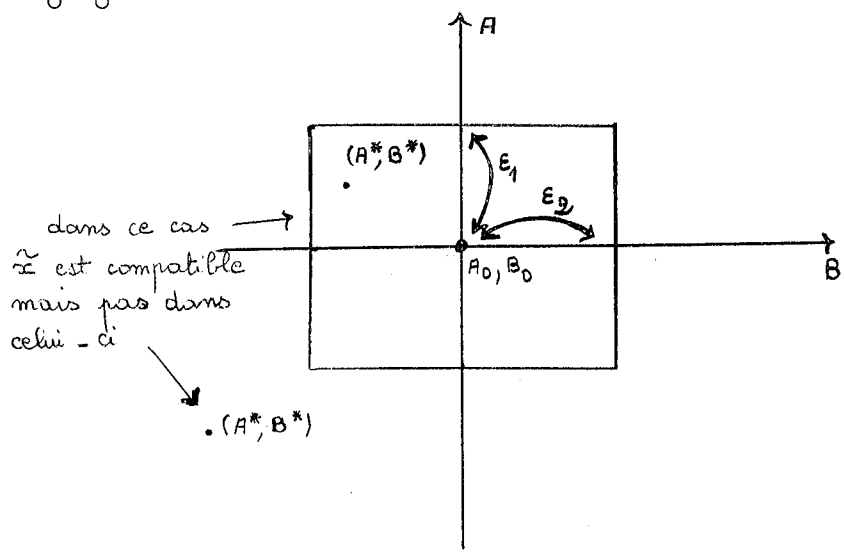
C'est-à-dire, qu'ayant trouvé une solution numérique $\tilde{x}$ de $A_o^T A_o X - A_o^T B_o = o$, nous allons examiner l'ensemble des problèmes dont $\tilde{x}$ est solution exacte et voir s'il en existe un qui ne soit pas trop "différent" du problème véritable. Pour cela, A_o étant de dimension $p \times q$ et B_o de dimension q, on se placera dans l'espace $\mathbb{R}^{q\times(p+1)}$. Dans celui-ci, on cherchera le couple de matrices (A^*, B^*) qui appartient au cône $A^T A\tilde{x} - A^T B = o$ et qui est à la "distance" minimum de (A_o, B_o).

- Dans toute la suite nous utiliserons comme norme de matrices ou de vecteurs la racine carrée de la somme des carrés des éléments,

exemple :

$$||A|| = \sqrt{\sum_{i,j} a_{ij}^2}$$

Si (A^*,B^*) est à l'intérieur du domaine dans lequel nous savons que se trouve (A_o,B_o) nous dirons que la solution $\tilde{x}$ est compatible.



Méthode.

Nous utiliserons la méthode des multiplicateurs de Lagrange que je rappelle ici brièvement :

Soit une fonction de plusieurs variables $f(x_1,x_2,\ldots x_n)$ dont nous cherchons un extrêmum sachant que $x_1, x_2\ldots,x_n$ ne sont pas quelconques mais soumis à "p" contraintes de la forme $g_i(x_1,\ldots,x_n) = o$ $(i=1,2,\ldots p)$. On montre que l'on trouve cet extrêmum en annulant les dérivées partielles par rapport à $x_1,\ldots, x_n$ de la fonction auxiliaire $f(x_1,\ldots,x_n) + \sum_i \lambda_i g_i (x_1,\ldots x_n)$. Les λ_i étant des inconnues auxiliaires, nous aurons donc $n + p$ inconnues à déterminer et, pour cela, les n dérivées partielles et les p contraintes.

Dans le cas présent, nous minimiserons d^2 -carré de la distance euclidienne rapportée à l'espace $\mathbb{R}^{p\times q+q}$ du point (A,B) au point (A_o,B_o) avec la seule contrainte pour (A,B) d'appartenir au cône $A^TA\tilde{X} - A^TB = o$.

Plan Suivi.

Nous commencerons par étudier le cas particulier où A_o est une matrice à 2 lignes et une colonne, x se réduisant alors à un scalaire, puis nous généraliserons le résultat trouvé au cas d'une matrice colonne de longueur p.

Dans la deuxième partie, nous étudierons le cas le plus général à une restriction près : A_o sera une matrice ayant plus de lignes que de colonnes. En effet, en pratique, nous aurons bien plus de mesures que d'inconnues. Nous faisons aussi l'hypothèse que $A_o^T A_o$ sera toujours inversible.

Remarque.

Il est évident que si $\tilde{x} \to x^+$ (solution exacte de $A_o^T A_o x - A_o^T B_o$) alors $(A^*, B^*) \to (A_o, B_o)$.

CHAPITRE II

ETUDE D'UN CAS PARTICULIER

Nous allons étudier le cas particulier suivant :

$$A_o = \begin{pmatrix} a_o \\ a'_o \end{pmatrix} \quad , \quad B_o = \begin{pmatrix} b_o \\ b'_o \end{pmatrix} \quad , \; X = (x).$$

Appliquons la méthode de Lagrange suivant l'exposé du chapitre I. Dans notre cas particulier, cela revient à minimiser dans $\mathbb{R}^4$ la distance euclidienne $d^2 = (a-a_o)^2 + (a'-a'_o)^2 + (b-b_o)^2 + (b'-b'_o)^2$ sachant que $a^{(\prime)}$ et $b^{(\prime)}$ vérifient $(a^2+a'^2)\tilde{x} = ab + a'b'$, autrement dit à annuler les dérivées partielles par rapport à : a, a', b et b' de la fonction :

$$(a-a_o)^2 + (a'-a'_o)^2 + (b-b_o)^2 + (b'-b'_o)^2 + \lambda \left[(a^2+a'^2)\tilde{x} - (ab+a'b')\right].$$

En ajoutant aux quatre dérivées partielles la contrainte, on obtient le système de 5 équations à 5 inconnues a, a', b, b' et λ :

$$\begin{cases} 2(1+\lambda\tilde{x})a - \lambda b = 2a_o \\ 2(1+\lambda\tilde{x})a' - \lambda b' = 2a'_o \\ 2b - \lambda a = 2b_o \\ 2b' - \lambda a' = 2b'_o \\ (a^2+a'^2)\tilde{x} = ab + a'b' \end{cases}$$

En résolvant ce système par rapport à λ, on obtient les quatre valeurs :

$$a = \frac{4a_o + 2\lambda b_o}{4(1+\lambda\tilde{x})-\lambda^2} \qquad a' = \frac{4a'_o+2\lambda b'_o}{4(1+\lambda\tilde{x})-\lambda^2}$$

$$b = \frac{2\lambda a_o+4(1+\lambda\tilde{x})b_o}{4(1+\lambda\tilde{x})-\lambda^2} \qquad b' = \frac{2\lambda a'_o+4(1+\lambda\tilde{x})b'_o}{4(1+\lambda\tilde{x})-\lambda^2}$$

On voit que si $\lambda \to o$ alors $(a,b,a',b') \to (a_o,b_o,a'_o,b'_o)$

Nous allons maintenant résoudre l'équation qui nous donnera λ. Nous l'obtiendrons en reportant les 4 valeurs précédentes dans la contrainte.

L'équation ordonnée en λ s'écrit :

$$\lambda^2\left[a_o b_o + a'_o b'_o + (b_o^2 + b'^2_o)\tilde{x}\right] + 2\lambda\left[a_o^2 + a'^2_o + b_o^2 + b'^2_o\right]$$

$$- 4\left[(a_o^2 + a'^2_o)\tilde{x} - (a_o b_o + a'_o b'_o)\right] = o$$

Nous posons :

$$\begin{cases} a_o^2 + a'^2_o = P \\ b_o^2 + b'^2_o = Q \\ a_o b_o + a'_o b'_o = R \\ \text{et } \frac{R}{P} = x^* \quad \text{(c'est la solution exacte)} \end{cases}$$

l'équation précédente s'écrit alors :

$$\lambda^2 (R+Q\tilde{x}) + 2\lambda (P+Q) - 4P(\tilde{x}-x^*) = o$$

On en déduit les deux solutions.

$$\lambda = \frac{-(P+Q) \pm \sqrt{(P+Q)^2 + 4(R+Q\tilde{x})\ P(x-x^*)}}{R + Q\tilde{x}}$$

nous avons vu que si $\lambda \to o$ alors (a,b,a',b') tend vers (a_o,b_o,a'_o,b'_o) et donc $\tilde{x} \to x^*$ ce qui nous impose ici le choix du signe +. En effet dans ce cas lorsque $\tilde{x} \to x^*$ on a bien $\lambda \to o$.

$$\boxed{\lambda = \frac{-(P+Q) + \sqrt{(P+Q)^2 + 4(R+Q\tilde{x})\ P(\tilde{x}-x^*)}}{R + Q\tilde{x}}}$$

Remarque.

Le discriminant peut s'écrire :

$$\Delta = (P+2R\tilde{x}-Q)^2 + 4(a_o b'_o - a'_o b_o)^2\ (1+\tilde{x}^2)$$

il est donc toujours positif ou nul et donc λ existe toujours.

Calculons maintenant la distance de (A_o,B_o) au cône $A^TA\tilde{x} - A^TB = o$

On a :

$$a^{(\prime)} - a_o^{(\prime)} = \lambda \circ \frac{(\lambda - 4\tilde{x})\ a_o^{(\prime)} + 2\ b_o^{(\prime)}}{4(1+\lambda\tilde{x}) - \lambda^2}$$

$$b^{(\prime)} - b_o^{(\prime)} = \lambda \circ \frac{2\ a_o^{(\prime)} + \lambda\ b_o^{(\prime)}}{4(1+\lambda\tilde{x}) - \lambda^2}$$

On en déduit d^2 qui tous les calculs faits peut s'écrire :

$$\boxed{d^2 = \frac{\lambda^2}{\left[4(1+\lambda\tilde{x}) - \ ^2\right]^2} \circ \left[(\lambda^2+4)(P+Q) + 8(P\tilde{x}-R)(2\tilde{x}-\lambda)\right]}$$

Remarque.

d^2 est une somme de carrés donc $d^2 \geqslant 0$, l'expression de d^2 en fonction de λ (voir page) nous montre que $d^2 \to 0$ lorsque $\lambda \to 0$ et que $d^2 \to \infty$ lorsque $\lambda \to 2(x \pm \sqrt{1+x^2})$ comme d^2 est continue sur $\left]2(x-\sqrt{1+x^2}),\ 2(x+\sqrt{1+x^2})\right[$ on voit que toutes les valeurs entre 0 et ∞ seront prises on peut donc approcher le vrai problème d'aussi près que l'on voudra.

1ère généralisation.

Nous pouvons aisément généraliser les résultats précédents au cas où A_o et B_o sont deux vecteurs colonnes de longueurs quelconques. Il suffit de calculer :

$$P = A_o^T A_o \quad , \quad Q = B_o^T B_o \quad \text{et} \quad R = A_o^T B_o .$$

et de les transporter dans les équations précédentes.

CHAPITRE III

ETUDE DU CAS GENERAL

Nous allons essayer de généraliser les résultats précédents. Nous passerons par l'étape intermédiaire d'une matrice A_o à 3 lignes et 3 colonnes.

$$A_o = \begin{pmatrix} a_o & b_o & c_o \\ a'_o & b'_o & c'_o \\ a''_o & b''_o & c''_o \end{pmatrix} , \quad B_o = \begin{pmatrix} d_o \\ d'_o \\ d''_o \end{pmatrix} , \quad X = \begin{pmatrix} x \\ y \\ z \end{pmatrix} , \quad \lambda = \begin{pmatrix} \lambda \\ \mu \\ \nu \end{pmatrix}$$

écrivons $A^T A\tilde{x} - A^T B$

$$\left[(a^2+a'^2+a''^2)\tilde{x} + (ab+a'b'+a''b'')\tilde{y} + (ac+a'c'+a''c'')\tilde{z} - (ad+a'd'+a''d'')\right]$$
$$\left[(ab+a'b'+a''b'')\tilde{x} + (b^2+b'^2+b''^2)\tilde{y} + (bc+b'c'+b''c'')\tilde{z} - (bd+b'd'+b''d'')\right]$$
$$\left[(ac+a'c'+a''c'')\tilde{x} + (bc+b'c'+b''c'')\tilde{y} + (c^2+c'^2+c''^2)\tilde{z} - (cd+c'd'+c''d'')\right]$$

un simple examen nous montre que les termes contenant une des variables a, a', a'', b, b',..., d'' sont contenues dans une ligne et dans une colonne, cette disposition est générale.

Retour au cas général.

Nous avons à minimiser Trace $\left[(A-A_o)^T(A-A_o)\right] + (B-B_o)^T(B-B_o)$. Nous utiliserons la méthode des multiplicateurs de Lagrange, c'est-à-dire que nous chercherons les extrêmums de

$$\text{TRACE}\left[(A-A_o)^T(A-A_o)\right] + (B-B_o)^T(B-B_o) + \lambda^T(A^T A\tilde{x} - A^T B) \qquad (1)$$

Nous avons comme matrices

$$A_o = \begin{pmatrix} a_{n} & \cdots & a_{1,m} \\ \vdots & & \vdots \\ a_{n,1} & \cdots & a_{n,m} \end{pmatrix} \quad , \quad B_o = \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix} \quad , \quad X = \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix} \quad , \quad \lambda = \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_m \end{pmatrix}$$

Si nous dérivons (1) par rapport à a_{ik}, on obtient :

$$2\, a_{ik} + \lambda_k\, (A_{i\circ} X) + x_k\, (A_{i\circ}\lambda) - \lambda_k\, b_i = 2\, a_{ik}^{o}$$

en dérivant par rapport à b_i, on a :

$2\, b_i = 2\, b_i^{o} + A_{i\circ}\, \lambda$. Nous désignons par $A_{i\circ}$ la $i^{ème}$ ligne de la matrice A.

et, en combinant les deux pour éliminer b_i,

$$\boxed{4\, a_{ik} + 2\, \lambda_k\, (A_{i\circ} X) + 2\, X_k\, (A_{i\circ}\lambda) - \lambda_k\, (A_{i\circ}\lambda) = 4\, a_{ik}^{o} + 2\, \lambda_k\, b_i^{o}}$$

Nous écrivons les unes sous les autres les équations obtenues en faisant varier k. On peut écrire matriciellement :

$$(2)\quad 4\, A_{i\circ}^{T} + 2\lambda\, (A_{i\circ} X) + 2X\, (A_{i\circ}\lambda) - \lambda (A_{i\circ}\lambda) = 4\, A_{i\circ}^{oT} + 2\lambda b_i^{oT}$$

Or $A_{i\circ} X$ et $A_{i\circ}\lambda$ sont des scalaires, on peut donc transposer ces quantités sans changer leur valeur.

$$(2) \text{ s'écrit } 4\, A_{i\circ}^{T} + 2\lambda\, (X^{T} A_{i\circ}^{T}) + 2X\, (\lambda^{T} A_{i\circ}^{T}) - \lambda (\lambda^{T} A_{i\circ}^{T})$$

$$= \left[4\, I + 2(\lambda X^{T} + \lambda\lambda^{T}) - \lambda\lambda^{T}\right] A_{i\circ}^{T} = 4\, A_{i\circ}^{oT} + 2\Lambda b_i^{oT}$$

On pose $D = 4\, I + 2(\lambda X^{T} + X\lambda^{T}) - \lambda\lambda^{T}$

On a donc $A_{i\circ}^T = D^{-1}\ (4A_{i\circ}^{oT}+2\lambda b_i^{oT})$

On en déduit

$$\boxed{\begin{aligned} A^T &= 2\,D^{-1}\left[2A_o^T+\lambda B_o^T\right] \\ A &= 2\left[2A_o+B_o\lambda^T\right]D^{-1} \end{aligned}}$$

Or $b_i = b_i^o + \frac{1}{2} A_{i\circ}\lambda$ et on en tire donc :

$$\boxed{B = B_o + (2A_o+B_o\lambda^T)D^{-1}\lambda}$$

Nous avons obtenu les valeurs de A et B en fonction de λ, il nous faut déterminer λ. On a l'équation encore inemployée : $A^TA\tilde{X} = A^TB = A^T(B_o+\frac{1}{2}A\lambda)$ ce qui s'écrit encore $A^TA(2\tilde{X}-\lambda) = 2A^TB_o$.

Calcul de D^{-1}.

$D = 4\,I + 2(\lambda X^T+X\lambda^T) - \lambda\lambda^T$ est de la forme : $aI + bK$. En étudiant K on s'aperçoit qu'il vérifie l'équation

$$(3)\quad K^3 - \left[2(\lambda^Tx+x^T\lambda)-\lambda^T\lambda\right]K^2 + 4\left[\lambda^Txx^T\lambda-\lambda^T\lambda xx^T\right]K = o$$

or K^2 introduit la matrice $x\,x^T$ qui n'existe pas dans K. On voit donc que K vérifie un polynôme minimal de degré 3. Or (3) ayant le coefficient de K^3 égal à 1 est donc le polynôme minimal

$$m(u) = u^3 - \left[2(\lambda^Tx+x^T\lambda)-||\lambda||^2\right]u^2 + 4\left[(\lambda^Tx)^2-||\lambda||^2||x||^2\right]u$$

K est donc une matrice du 3° degré.

On pourra donc, sauf cas particulier, inverser D par une matrice de la forme $aI + bK + cK^2$. Procédons par identification

$$(4I+K)(aI+bK+cK^2) = 4\,aI + (4b+a)K + (4c+b)K^2 + K^3$$

On remplace K^3 par sa valeur, en fonction de K^2 et K, obtenue à l'aide de (3).

On obtient le système :

$$(4)\qquad \begin{cases} a = \frac{1}{4} \\ 4\,b + c\left\{4\left[(\lambda^T x)^2 - ||\lambda||^2||x||^2\right]\right\} = -\frac{1}{4} \\ b + c\left[4\lambda^T x - ||\lambda||^2\right] = 0 \end{cases}$$

- on en tire a, b et c et on obtient tous les calculs faits :

$$D^{-1} = \frac{\left[4(1+\lambda^T x) - ||\lambda||^2(1+||x||^2) + (\lambda^T x)^2\right] I - (2+\lambda^T x)(\lambda x^T + x\lambda^T) + (1+||x||^2)\lambda\lambda^T + ||\lambda||^2 xx^T}{16(1+\lambda^T x) - 4||\lambda||^2(1+||x||^2) + 4(\lambda^T x)^2}$$

Ceci bien sûr si le système (4) est possible et déterminé, c'est-à-dire le déterminant de la matrice des coefficients est non nul, ou encore si $4\,(1+\lambda^T x) + (\lambda^T x)^2 \neq ||\lambda||^2\,(1+||x||^2)$ la condition sur λ qui en découle est la condition d'inversibilité de D.

Retour au problème.

Nous avons vu que l'équation en λ s'écrivait $A^T A(2x-\lambda) = 2A^T B_o$ donc :

$$4\,D^{-1}(2A_o^T + \lambda B_o^T)(2A_o + B_o\lambda^T)D^{-1}(2x-\lambda) = 2\,D^{-1}(2A_o^T + \lambda B_o^T)B_o$$

ou encore :

$$(5)\qquad (2A_o^T + \lambda B_o^T)(2A_o + B_o\lambda^T)D^{-1}(2x-\lambda) = (2A_o^T + \lambda B_o^T)B_o$$

Calculons $D^{-1}(2x-\lambda)$. On obtient

$$D^{-1}(2x-\lambda) = \frac{x(2+\lambda^T x)-\lambda(1+||x||^2)}{4(1+\lambda^T x)+(\lambda^T x)^2-||\lambda||^2(1+||x||^2)}$$

On pose

$$P = A_o^T A_o$$
$$Q = B_o^T B_o$$
$$R = A_o^T B_o$$

Remarque.

Q est un scalaire

On peut encore écrire (5) sous la forme :

$$(5') \quad (4P+2R\lambda^T+2\lambda R^T+Q\lambda\lambda^T)\cdot\left[(2+\lambda^T\tilde{x})\tilde{x}-(1+||\tilde{x}||^2)\lambda\right]$$
$$= \left[4(1+\lambda^T\tilde{x})-||\lambda||^2(1+||\tilde{x}||^2)+(\lambda^T\tilde{x})^2\right](2R+Q\lambda)$$

Nous poserons que, par définition, x^* est solution exacte du problème posé, c'est-à-dire $Px^* = R$. L'équation (5'), ordonnée en λ s'écrit :

$$\lambda^T\left\{\left[\tilde{x}\tilde{x}^T-(1+||\tilde{x}||^2)I\right]R - Q\tilde{x}\right\}\lambda$$
$$+ 2\left\{P\left[\tilde{x}\tilde{x}^T-(1+||x||^2)I\right]+(R^T\tilde{x}-Q)\ -R\tilde{x}^T\right\}\lambda$$
$$+ 4P(\tilde{x}-x^*) = 0$$

Nous ne pouvons résoudre directement cette équation aussi nous utiliserons la méthode de Newton généralisée $\lambda_{i+1} = \lambda_i - \left[F'(\lambda_i)\right]^{-1} F(\lambda_i)$

Nous voulons la racine la plus proche de zéro aussi nous prendrons comme vecteur de départ $\lambda_o = \begin{pmatrix} o \\ o \\ \vdots \\ o \end{pmatrix}$

ayant cette valeur de λ il suffit de la reporter dans les équations qui donnent $A - A_o$ et $B - B_o$ puis de calculer leurs normes euclidiennes.

CHAPITRE IV

PROCEDURE ALGOL

```
PROCEDURE ERPOSYSQUA(CA,CB,X,U,T,IMPOSSIBLE) ;
Reel tableau CA,CB,X ; entier U,T ; ETIQUETTE IMPOSSIBLE ;
COMMENTAIRE  cette procédure calcule l'erreur à postériori dans la mé-
thode des moindres carrés. CA est la matrice du premier membre, CB celle
du second. CA est de dimension U*T et CB de 1*U  X est la solution numé-
rique du système quadratique ;

Debut reel tableau GA[1:T,1:T], GB[1:T], A[1:T],
C[1:T], B[1:T,1:T], DELTA[1:T], LANDA[1:T],
MA[1:T], MB[1:U] ; Reel D1,D2,DENOM,PSLX,
NORMX,NORML ; Entier COMPT,K ;
Procedure GRESOLSYSLINE(A,B,X,N,IMPOSSIBLE) ;
reel tableau A,B,X ; Entier N ; étiquette impossible ;
   debut triangularisation :
      debut entier i,j,k ; reel R ;
      pour k := 1 pas 1 jusqua N-1 faire
         debut normal :
            debut si ABS(A[K,K]) = o alors allera
            Echange de lignes ;
            pour i := k+1 pas 1 jusqua n faire
              debut R := A[i,k]/A[k,k] ;
              pour j := k+1 pas 1 jusqua N faire
              A[i,j] := A[i,j] - R*A[K,j] ;
              B[i] := B[i] - R*B[k]
              fin
            fin ;
```

```
        Allera Retour ;
        Echange de lignes :
           debut entier L,M ;
           M := k+1
           pour L := M tant que A[L,k] = o          Λ
           L infeg N faire M := M+1 ;
           Si M = N+1 alors allera IMPOSSIBLE ;
           pour j := k pas 1 jusqua N faire
             debut R := A[k,j] ; A[k,j] := A[M,j] ;
             A[M,j] := R ;
             fin ;
           R := B[k] ; B[k] := B[M] ; B[M] := R ;
           Allera NORMAL
           fin ;
        Retour
        fin
     fin triangularisation ;
        RESSYSTRI :
           debut entier i,j ; Reel TX ;
           pour i := N pas -1 jusqua 1 faire
             debut TX := o ;
             pour j := N pas -1 jusqua i+1 faire
             TX := TX-X[j]*A[i,j] ;
             Si A[i,i] = o alors allera impossible ;
             X[i] := (B[i]+TX)/A[i,i]
             fin
           fin Ressystri
           Gresolsysline ;
fin
           Q := o ;
           pour i := 1 pas 1 jusqua U faire
           Q := Q + CB[i]*CB[i] ;
           pour i := 1 pas 1 jusqua T faire
```

```
debut GB[i] := o ;
pour k := 1 pas 1 jusqua U faire
GB[i] := GB[i] + CA[k,i]*CB[k] ;
pour j := 1 pas 1 jusqua T faire
  debut GA[i,j] := o ;
  pour k := 1 pas 1 jusqua U faire
  GA[i,j] := GA[i,j] + CA[k,i]*CA[k,j]
  fin
fin ;
NORMX := o ; H := o ;
pour i := 1 pas 1 jusqua T faire
debut H := H + GB[i]*X[i] ;
C[i] := - 4*GB[i] ;
A[i] := o ;
pour j := 1 pas 1 jusqua T faire
  debut C[i] := C[i] + 4*GA[i,j]*X[j] ;
  GQ[i,j] := X[i]*X[j] ;
  B[i,j] := - GB[i]*X[j]
  fin
fin ;
pour i := 1 pas 1 jusqua T faire
debut GQ[i,i] := GQ[i,i] - (1+NORMX) ;
pour j := 1 pas 1 jusqua T faire
  debut A[i] := A[i] + GQ[i,j]*GB[j] ;
  pour k := 1 pas 1 jusqua T faire
  B[i,j] := B[i,j] + GA[i,k]*GQ[k,j]
  fin ;
A[i] := A[i] - Q+X[i] ;
B[i,i] := B[i,i] + H-Q
fin ;
pour i := 1 pas 1 jusqua T faire
pour j := 1 pas 1 jusqua T faire
B[i,j] := 2*B[i,j] ;
```

```
pour i := 1 pas 1 jusqua T faire
LANDA[i] := o ; COMPT := o ;
ITER : H := P := Q := o ;
pour i := 1 pas 1 jusqua T faire
  debut GB[i] := C[i] ;
  GB[i] := GB[i] + H*LANDA[i]
  fin ;
pour i := 1 pas 1 jusqua T faire
pour j := 1 pas 1 jusqua T faire
  debut GA[i,j] := LANDA[i]*A[j] ;
  GA[i,j] := GA[i,j] + B[i,j] ;
  GB[i] := GB[i] + B[i,j]*LANDA[j]
  fin ;
pour i := 1 pas 1 jusqua T faire
GB[i] := - GB[i] ;
GRESOLSYSLINE(GA,GB,DELTA,T,FICHU) ;
pour i := 1 pas 1 jusqua T faire
  debut P := P + ABS(DELTA[i]) ;
  Q := Q + ABS(LANDA[i])
  fin ;
Si COMPT = o alors allera Suite ;
Si P/Q infeg 10^-3 alors COMPT := COMPT+1 ;
SUITE :
pour i := 1 pas 1 jusqua T faire
LANDA[i] := LANDA[i] + DELTA[i] ;
Si COMPT = o alors COMPT := 1 ;
Si COMPT = 5 alors allera CEFINI sinon
allera ITER ;
CEFINI :
PSLX := o ; NORML := o ;
pour i := 1 pas 1 jusqua T faire
  debut PSLX := PSLX + LANDA[i]*X[i] ;
  NORML := NORML + LANDA[i]*LANDA[i]
  fin ;
```

```
DENOM := 16*(1+PSLX) - 4*NORML(1+NORMX)
         + 4*PSLX*PSLX ;
pour i := 1 pas 1 jusqua T faire
pour j := 1 pas 1 jusqua T faire
  debut GA[i,j] := NORML*X[i]*X[j] + (1+NORMX)
  *LANDA[i]*LANDA[j] - (2+PSLX)*LANDA[i]
  *X[j] + LANDA[j]*X[i]) ;
  GA[i,j] := GA[i,j]/DENOM
  fin ;
pour i := 1 pas 1 jusqua T faire
GA[i,i] := GA[i,i] + (4*(1+PSLX) - NORML*
(1+NORMX) + PSLX*PSLX)/DENOM ;
D1 := D2 := o ;
pour i := 1 pas 1 jusqua U faire
debut pour j := 1 pas 1 jusqua T faire
MA[j] := 4*CA[i,j] + 2*CB[i]*LANDA[j] ;
pour j := 1 pas 1 jusqua T faire
  debut GB[j] := o ;
  pour k := 1 pas 1 jusqua T faire
  GB[j] := GB[j] + MA[k]*GA[k,j]
  fin ;
pour j := 1 pas 1 jusqua T faire
  debut MA[j] := GB[j] ;
  D1 := D1 + (MA[j] - CA[i,j])*(MA[j]-CA[i,j])
  fin ;
MB[i] := o ;
pour j := 1 pas 1 jusqua T faire
MB[i] := MB[i] + (MA[j]*LANDA[j])/2 ;
D2 := D2 + MB[i]*MB[i]
fin ;
SAUTLIGNE ;
ECRIRE("DISTANCE␣DES␣DE","UX␣PRØBLEMES") ;
SAUTLIGNE ;
ECRIRE("NØRME␣DE␣A-AO␣=", RAC2(D1),,
       "NØRME␣DE␣B-BO␣=", RAC2(D2)) ;
```

```
SAUTLIGNE
FICHU : ECRIRE("SYSTEME␣IMPOSSI","BLE") ;
SAUTLIGNE ; Allera IMPOSSIBLE
FIN ERPOSYSQUA ;
```

IV - EXEMPLES NUMERIQUES

Soit le système mal conditionné :

$$A_o = \begin{bmatrix} 3,05 & 4 & 1 & 7 & 6,30 \\ 3,07 & 4,08 & 0,92 & 7,30 & 6,10 \\ 3,01 & 4,01 & 1 & 7,80 & 6,20 \\ 3 & 4,03 & 1,05 & 7,25 & 6,35 \\ 3,05 & 4,08 & 1,02 & 7,50 & 6,24 \\ 3,07 & 4,02 & 1,01 & 7,12 & 6,11 \\ 3,09 & 3,92 & 0,95 & 7,23 & 6,83 \\ 3,12 & 3,84 & 1,05 & 7,14 & 6,37 \\ 3,03 & 4,06 & 0,92 & 7,26 & 6,04 \\ 9,25 & 28,50 & 74,35 & 18,50 & 313,25 \end{bmatrix} \qquad B_o = \begin{bmatrix} 49 \\ 48,98 \\ 49,46 \\ 49,47 \\ 49,16 \\ 48,68 \\ 50,53 \\ 48,90 \\ 48,60 \\ 1165,1 \end{bmatrix}$$

$$x^* = \begin{pmatrix} 2 \\ 4 \\ 1 \\ 1 \\ 3 \end{pmatrix} \quad \text{Gresolsysline} \quad \tilde{x} = \begin{matrix} 2,146 \\ 3,772 \\ 0,814 \\ 1,033 \\ 3,058 \end{matrix}$$

On obtient :

$$||A^* - A_o||^2 = 0,96848 \qquad 10^{-10}$$

$$||B^* - B_o||^2 = 0,15088 \qquad 10^{-11}$$

nous voyons donc qu'une infime variation des coefficients de A_o et B_o (de l'ordre de 10^{-6}) peut donner des variations de l'ordre du dixième pour la solution.

Le même calcul fait avec les coefficients multipliés par 10 puis par 100 donne :

$$\tilde{x}_{10} = \begin{cases} 2,140 \\ 3,776 \\ 0,811 \\ 1,034 \\ 3,058 \end{cases} \qquad \tilde{x}_{100} = \begin{cases} 2,144 \\ 3,772 \\ 0,806 \\ 1,034 \\ 3,060 \end{cases}$$

	X1	X10	X100
$\|\|A^{*} - A_o\|\|^2 =$	$0,96848 . 10^{-10}$	$0,674624 . 10^{-8}$	$0,62370\ 10^{-6}$
$b^{*} - b_o\ ^2 =$	$0,15088\ 10^{-10}$	$0,148445\ 10^{-9}$	$0,97964\ 10^{-8}$

nous voyons que l'erreur a postériori dépend un peu de la façon dont nous rentrons les nombres puisque l'erreur suivant le cas n'est pas multipliée par 100 ou 10000 mais reste inférieure à cette estimation. Cela prouve que, d'après la première remarque, en entrant un nombre en mémoire, nous commettons une erreur d'arrondi qui peut provoquer une erreur sur le résultat assez grave. L'étude du troisième cas où le nombre s'écrit en binaire de façon simple et où il n'y a donc pas d'erreur d'arrondi nous montre bien que c'est la méthode qui est déficiente.

Et voici un exemple mieux conditionné :

$$A_o = \begin{bmatrix} 1 & 12 & 24 & 110 \\ 3 & 4 & 2 & 160 \\ 7 & 1 & 31 & 200 \\ 12 & 7 & 50 & 520 \\ 18 & 9 & 110 & 1040 \\ 34 & 11 & 500 & 2700 \\ 36 & 15 & 23 & 512 \\ 41 & 18 & 4 & 43 \\ 50 & 30 & 1 & 28 \end{bmatrix} \qquad B_o = \begin{bmatrix} 712 \\ 246 \\ 844 \\ 1614 \\ 3366 \\ 12878 \\ 1194 \\ 385 \\ 448 \end{bmatrix}$$

$$x^{*} = \begin{vmatrix} 2 \\ 10 \\ 20 \\ 1 \end{vmatrix} \qquad \tilde{x} = \begin{vmatrix} 2,0000011 \\ 9,9999897 \\ 19,999989 \\ 1,0000019 \end{vmatrix}$$

$$||A^{*} - A_{o}||^{2} = 0,5436398 \quad 10^{-8}$$

$$||B^{*} - B_{o}||^{2} = 0,3213597 \quad 10^{-12}$$

On voit que l'erreur sur x et la distance au problème vérifié le plus voisin sont du même ordre (10^{-6}).

BIBLIOGRAPHIE

PARTIE - A

On trouvera une étude des multiplicateurs de Lagrange dans :

J. BASS - COURS DE MATHEMATIQUES
Masson Editeur 1961

Une introduction plus détaillée à l'analyse à postériori des erreurs peut être trouvée dans :

[1] J. GACHES - COMPATIBILITE D'UNE SOLUTION CALCULEE AVEC LES DONNEES D'UN SYSTEME LINEAIRE A COEFFICIENTS INCERTAINS.
Thèse Besançon 1966

W. DETTLI - W. PRAGER - COMPATIBILITY OF APPROXIMATE SOLUTIONS OF LINEAR EQUATIONS WITH GIVEN ERROR BOUNDS FOR COEFFICIENTS AND RIGHT-HAND SIDES
Numerische Mathematik 6, (405-409) 1964

W. DETTLI - W. PRAGER - J.H. WILKINSON - ADMISSIBLE SOLUTIONS OF LINEAR SYSTEMS WITH NOT SHARPLY DEFINED COEFFICIENTS
J. Siam Numer. Anal. Ser. V. Vol. 2, N° 2 - 1965

On trouvera tous renseignements sur du second degré, polynômes minimaux et la méthode d'inversion de matrice utilisée dans :

N. GASTINEL - MATRICES DU SECOND DEGRE ET NORMES GENERALES EN ANALYSE NUMERIQUE LINEAIRE
Publications scientifique et techniques du Ministère de l'Air - 1962.

ETUDE DE L'ERREUR A POSTERIORI

DANS LE CAS

DU POLYNOME

CHAPITRE I

INTRODUCTION

Considérons une équation de la forme $x^n + a_1^o x^{n-1} + a_2^o x^{n-2} + \ldots + a_{n-1}^o x + a_n^o = o$. Nous avons trouvé, par hypothèse, p solutions numériques de ce polynôme : $\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_p$. Nous allons chercher s'il existe des jeux de coefficients $a_1, a_2, \ldots, a_n$ qui soient tels que $\tilde{x}_1, \tilde{x}_2, \ldots \tilde{x}_p$ en soient des solutions exactes, et si, parmi ces jeux, il s'en trouve un qui soit plus "proche" au sens de la norme euclidienne dans $\mathbb{R}^n$ du jeu $(a_1^o, \ldots, a_n^o)$. Nous utiliserons comme en A les multiplicateurs de Lagrange.

Nous avons ici "p" contraintes pour les a_i.

A savoir que $\sum_{i=1}^{n} a_i \tilde{x}_k^{n-i} + \tilde{x}_k^n = o$ pour $k = 1, \ldots p$.

Suivant un plan similaire de celui suivi dans A nous étudierons d'abord le cas particulier du polynôme du second degré dont on connait une puis deux solutions numériques.

CHAPITRE II

ETUDE D'UN CAS PARTICULIER

1er Cas.

Nous connaissons une racine $\tilde{x}$ d'une équation du second degré.

Nous allons donc chercher le doublet (b^*,c^*) de $\mathbb{R}^2$ sur la droite d'équation $\tilde{x}^2 + b\tilde{x} + \tilde{c} = o$ qui se trouve à la distance minimale de (b_o,c_o).

On appelle ε_i la valeur du polynôme pour $x = \tilde{x}_i$. Appliquons la méthode de Lagrange. On aura donc à dériver $(b-b^o)^2 + (c-c^o)^2 + \lambda(\tilde{x}^2+b\tilde{x}+c)$ ce qui nous donne :

$$\boxed{\begin{aligned} 2(b^*-b^o) + \lambda\tilde{x} &= o \\ 2(c^*-c^o) + \lambda &= o \end{aligned}}$$

d'où l'on tire les valeurs de b^* et c^* en fonction de λ :

$$\boxed{\begin{aligned} b^* &= b_o - \frac{\lambda}{2}\tilde{x} \\ c^* &= c_o - \frac{\lambda}{2} \end{aligned}}$$

Reportons ces valeurs dans la contrainte :

$$\tilde{x}^2 + (b_o - \frac{\lambda}{2}\tilde{x})\,\tilde{x} + c_o - \frac{\lambda}{2} = o$$

$$2(\tilde{x}^2 + b_o\tilde{x} + c_o) = \lambda(1+\tilde{x}^2)$$

$$2\,\varepsilon = \lambda(1+\tilde{x}^2)$$

d'où :

$$\boxed{\lambda = \frac{2\,\varepsilon}{1+\tilde{x}^2}}$$

On en déduit d^2

$$d^2 = (b^*-b_o)^2 + (c^*-c_o)^2$$

$$= \frac{\lambda^2}{4}\tilde{x}^2 + \frac{\lambda^2}{4} = \frac{\lambda^2}{4}(1+\tilde{x}^2)$$

$$= \frac{\varepsilon^2}{(1+\tilde{x}^2)^2}(1+\tilde{x}^2) = \frac{\varepsilon^2}{1+\tilde{x}^2}$$

<u>2ème Cas</u>.

Nous connaissons les deux racines $\tilde{x}$ et $\tilde{y}$.

Nous avons à dériver cette fois :

$$(b-b_o)^2 + (c-c_o)^2 + \lambda(\tilde{x}^2+b\tilde{x}+c) + \mu(\tilde{y}^2+b\tilde{y}+c).$$

Nous obtenons :

$$\boxed{\begin{aligned} 2b^* &= 2b_o - \lambda\tilde{x} - \mu\tilde{y} \\ 2c^* &= 2c_o - \lambda - \mu \end{aligned}}$$

d'autre part : $\begin{cases} \tilde{x}^2 + b^*\tilde{x} + c^* = o \\ \tilde{y}^2 + b^*\tilde{y} + c^* = o \end{cases}$

Si on reporte les valeurs de b^* et c^* dans les deux contraintes on obtient :

$$\lambda(1+\tilde{x}^2) + \mu(1+\tilde{\tilde{x}\tilde{y}}) = 2\,\varepsilon_1$$
$$\lambda(1+\tilde{\tilde{x}\tilde{y}}) + \mu(1+\tilde{y}^2) = 2\,\varepsilon_2$$

d'où l'on tire les valeurs de λ et μ :

$$\boxed{\begin{aligned} \lambda &= 2\circ\frac{(1+\tilde{y}^2)\varepsilon_1 - (1+\tilde{\tilde{x}\tilde{y}})\varepsilon_2}{(1+\tilde{y}^2)(1+\tilde{x}^2) - (1+\tilde{\tilde{x}\tilde{y}})^2} \\ \mu &= 2\circ\frac{(1+\tilde{x}^2)\varepsilon_2 - (1+\tilde{\tilde{x}\tilde{y}})\varepsilon_1}{(1+\tilde{y}^2)(1+\tilde{x}^2) - (1+\tilde{\tilde{x}\tilde{y}})^2} \end{aligned}}$$

CHAPITRE III

ETUDE DU CAS GENERAL

Nous avons un polynôme :

$x^n + a_1^o x^{n-1} + a_2^o x^{n-2} + \ldots + a_{n-1}^o x + a_n^o = o$ et p solutions numériques de ce polynôme : $\tilde{x}_1, \tilde{x}_2 \tilde{x}_3, \ldots, \tilde{x}_p$ $(p \leqslant n)$. En utilisant, comme au chapitre II, la méthode des multiplicateurs de Lagrange, nous aurons à dériver

$$\sum_{i=1}^{n} (a_i - a_i^o)^2 + \sum_{j=1}^{p} \lambda_j \left(\sum_{k=1}^{n} a_k \tilde{x}_j^{n-k} + x_j^n \right).$$

En dérivant par rapport à a_i on obtient :

$$2(a_i^* - a_i^o) + \sum_{j=1}^{p} \lambda_j \tilde{x}_j^{n-i} = o$$

d'où : $$a_i^* = a_i^o - \frac{1}{2} \sum_{j=1}^{p} \lambda_j \tilde{x}_j^{n-i}$$

ou encore avec :

$$A^* = \begin{vmatrix} a_1^* \\ a_2^* \\ \vdots \\ a_n^* \end{vmatrix} \qquad B = \begin{vmatrix} 1 & 1 & \ldots\ldots\ldots & 1 \\ \tilde{x}_1 & \tilde{x}_2 & \ldots\ldots\ldots & \tilde{x}_p \\ \tilde{x}_1^2 & \tilde{x}_2^2 & \ldots\ldots\ldots & \tilde{x}_p^2 \\ \vdots & \vdots & & \\ \tilde{x}_1^n & \tilde{x}_2^n & \ldots\ldots\ldots & \tilde{x}_p^n \end{vmatrix}$$

$$A_o = \begin{vmatrix} a_1^o \\ a_2^o \\ \vdots \\ a_n^o \end{vmatrix} \qquad \lambda = \begin{vmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_p \end{vmatrix}$$

Notre résultat se présente sous la forme matricielle suivante :

$$\boxed{A^* = A_o - \frac{1}{2} B \cdot \lambda}$$

il nous faut maintenant calculer les λ_j.

Nous savons que : $\sum_{i=1}^{n} a_i^* \tilde{x}_k^{n-i} + \tilde{x}_k^n = o$

et que : $\sum_{i=1}^{n} a_i^o \tilde{x}_k^{n-i} + \tilde{x}_k^n = \varepsilon_k$

et par soustraction de ces deux égalités

$$\sum_{i=1}^{n} (a_i^* - a_i^o) \tilde{x}_k^{n-i} = - \varepsilon_k$$

or nous avons vu que :

$$a_i^* - a_i^o = - \frac{1}{2} \cdot \sum_{j=1}^{p} \lambda_j x_j^{n-i}$$

on en déduit :

$$\sum_{i=1}^{n} \left(\sum_{j=1}^{p} \lambda_j x_j^{n-i} \right) x_k^{n-i} = 2 \varepsilon_k$$

Nous avons affaire à des sommes finies, on peut dont inverser les signes sommes et écrire :

$$\boxed{\sum_{j=1}^{p} \left[\sum_{i=1}^{n} (x_j x_k)^{n-i} \right] \lambda_j = 2 \varepsilon_k \quad k = 1, 2, \ldots p}$$

ou encore matriciellement

$$B^T (A^* - A_o) = - \varepsilon$$

Or $$A^* - A_o = - \frac{1}{2} B\lambda$$

donc $$B^T B\lambda = 2\varepsilon$$

et $$\boxed{\lambda = 2(B^T B)^{-1} \varepsilon}$$

Nous avons obtenu un système linéaire de p équations à p inconnues que nous traiterons soit par la méthode de Gauss soit par une autre et on a :

$$\boxed{d^2 = || B(B^T B)^{-1} \varepsilon ||^2}$$

```
PROCEDURE ERPOPOL(AO,X,N,P,IMPOSSIBLE) ;
REEL TABLEAU AO,X ; ENTIER N,P ; ETIQUETTE IMPOSSIBLE ;
COMMENTAIRE  cette procédure calcule l'erreur à postériori commise dans le cas
où nous avons trouvé P solutions numériques réunies dans le tableau X du poly-
nôme du N^ième degré dont les coefficients sont rangés dans le tableau AO de
telle façon que la somme de l'indice et de la puissance soit toujours égale
à N ;

DEBUT ENTIER I,J,K ; REEL TABLEAU GX[1:N,1:P], GA[1:P,1:P],
    EPS[1:P], LANDA[1:P] ;
    REEL D, DELTA ;
    PROCEDURE GRESOLSYSLINE(A,B,X,N,IMPOSSIBLE) ;
    REEL TABLEAU A,B,X ; ENTIER N ; ETIQUETTE IMPOSSIBLE ;
    DEBUT TRIANGULARISATION :
       DEBUT ENTIER I,J,K ; REEL R ;
          POUR K := 1 PAS 1 JUSQUA N-1 FAIRE
          DEBUT NORMAL :
             DEBUT SI ABS(A[K,K]) = 0 ALORS ALLERA ECHANGE
                DE LIGNES ;
                POUR I := K+1 PAS 1 JUSQUA N FAIRE
                DEBUT R := A[I,K]/A[K,K] ;
                   POUR J := K+1 PAS L JUSQUA N FAIRE
                   A[I,J] := A[I,J] - R*A[K,J] ;
                   B[I] := B[I] - R×B[K]
                FIN
             FIN ;
          ALLERA RETOUR ;
          ECHANGE DE LIGNES :
          DEBUT ENTIER L,M ; M := K+1 ;
             POUR L := M TANTQUE A[L,K] = 0 ET L INFEG N
             FAIRE M := M+1 ;
             SI M = N+1 ALORS ALLERA IMPOSSIBLE ;
```

```
            POUR J := K PAS 1 JUSQUA N FAIRE
            DEBUT R := A[K,J] ; A[K,J] := A[M,J] ;
               A[M,J] := R
            FIN ;
            R := B[K] ; B[K] := B[M] ; B[M] := R ;
            ALLERA NORMAL
         FIN ; RETOUR :
      FIN
   FIN TRIANGULARISATION ;
   RESSYSTRI :
   DEBUT ENTIER I,J ; REEL TX ;
      POUR I := N PAS -1 JUSQUA 1 FAIRE
      DEBUT TX := 0 ;
         POUR J := N PAS -1 JUSQUA I+1 FAIRE
         TX := TX-X[J]×A[I,J] ;
         SI A[I,I] = 0 ALORS ALLERA IMPOSSIBLE ;
         X[I] := (B[I]+TX)/A[I,I]
      FIN
   FIN RESSYSTRI ;
FIN GRESOLSYSLINE ;
POUR I := 1 PAS 1 JUSQUA P FAIRE
GX[1,I] := 1.0 ;
POUR I := 2 PAS 1 JUSQUA N FAIRE
POUR J := 1 PAS 1 JUSQUA P FAIRE
GX[I,J] := GX[I-1,J]×X[J] ;
POUR I := 1 PAS 1 JUSQUA P FAIRE
DEBUT EPS[I] := 2·GX[N,I]×X[I] ;
   POUR J := 1 PAS 1 JUSQUA N FAIRE
   EPS[I] := EPS[I] + 2 AO[J]×GX[N+1-J,I]
FIN ;
POUR I := 1 PAS 1 JUSQUA P FAIRE
POUR J := 1 PAS 1 JUSQUA P FAIRE
POUR K := 1 PAS 1 JUSQUA N FAIRE
```

```
        GA[I,J] := GA[I,J] + GX[K,I]×GX[K,J] ;
        GRESOLSYSLINE(GA,EPS,LANDA,P,IMPOSSIBLE) ;
        POUR I := 0 PAS 1 JUSQUA N-1 FAIRE
        DEBUT DELTA := 0 ;
           POUR J := 1 PAS 1 JUSQUA P FAIRE
           DELTA := DELTA + GX[N-I,J]×LANDA[J] ;
           D := D + 0.25 DELTA DELTA
        FIN ;
        SAUTLIGNE ; ECRIRE (D) ;
    FIN ERPOPOL ;
```

Exemples numériques.

Nous avons appliqué les résultats précédents à l'équation :

$$x^4 - 3{,}58\,x^3 + 4{,}8059\,x^2 - 2{,}867222\,x + 0{,}6414408$$

dont les solutions exactes sont : 0,88 ; 0,89 ; 0,90 ; 0,91.

$p = 3$ $\quad \tilde{x}_1 = 0{,}8775$; $\tilde{x}_2 = 0{,}9012$; $\tilde{x}_3 = 0{,}89$

$\quad d^2 = 0{,}515038 \quad 10^{-9}$

$p = 2$ $\quad \tilde{x}_1 = 0{,}8775$; $\tilde{x}_2 = 0{,}9012$

$\quad d^2 = 0{,}460637 \quad 10^{-16}$

$p = 1$ $\quad \tilde{x}_1 = 0{,}88$

$\quad d^2 = 0{,}39201 \quad 10^{-16}$

$p = 1$ $\quad \tilde{x}_1 = 0{,}89989$

$\quad d^2 = 0{,}1158634 \quad 10^{-16}$

nous voyons qu'une variation de l'ordre de 10^{-8} des coefficients entraîne une variation de 10^{-3} sur les solutions. Cette équation est très mal conditionnée car les racines sont très proches. Les derniers chiffres montrent que le seul fait d'introduire un nombre en machine donne, par suite des troncatures dans la transformation de décimal en binaire, une erreur en 10^{-8} qui pourra influer sur les racines de façon sensible.

Equation mieux conditionnée.

$$x^3 - 106\,x^2 + 605\,x - 500$$

les racines sont : 1, 100, 5 ;

$p = 2 \qquad \tilde{x}_1 = 99{,}5 \; ; \; \tilde{x}_2 = 4{,}5 \; ;$

$$d^2 = 0{,}821305 \;\; 10^{+2}$$

$p = 1 \qquad \tilde{x}_1 = 100{,}000002$

$$d^2 = 0{,}5959271 \;\; 10^{-19}$$

nous voyons qu'il faut une variation de l'ordre de 10 pour faire varier les racines de 0,5.

BIBLIOGRAPHIE

PARTIE - B

Aucune nouvelle notion n'est introduite.
On trouvera cependant une étude du même problème mais traitée par des méthodes topologiques dans :

J. GACHES - COMPATIBILITE D'UNE SOLUTION CALCULEE AVEC LES DONNEES D'UN SYSTEME LINEAIRE A COEFFICIENTS INCERTAINS
Thèse Besançon 1966.

NGUYEN HUU VINH - SEMI NORME DUALE GENERALISEE ET APPROXIMATION D'UN VECTEUR
C.R. Acad. Sc. Paris t. 262 - pp. 1456-59 (1966)

- C -

ETUDE DE L'ERREUR A POSTERIORI

DANS LE CAS

DES VALEURS PROPRES

CHAPITRE I

INTRODUCTION

L'analyse à postériori des erreurs dans le cas des valeurs propres n'est pas possible par la méthode que nous avions employée jusqu'à maintenant. En effet la contrainte : la valeur propre numérique $\tilde{\lambda}$ trouvée doit être solution exacte du polynôme caractéristique d'une certaine matrice A, n'est pas du tout aisément manipulable. Nous allons donc utiliser un biais. Au lieu de considérer que nous avons seulement une valeur propre $\tilde{\lambda}$, nous admettrons par hypothèse que nous connaissons aussi un vecteur propre $\tilde{x}$.

Ce qui nous permettra de résoudre, au passage, le problème de l'erreur à postériori dans ce cas. En effet, la contrainte dans ce cas se réduit à $A\tilde{x} = \tilde{\lambda}\tilde{x}$ qui est linéaire. Nous remarquons aussi que $\tilde{x}$ apparaît dans les deux membres, nous pourrons donc considérer dorénavant uniquement les vecteurs $\tilde{x}$ de norme unité.

Ayant résolu ce problème, nous reviendrons au problème initial. Pour cela nous essaierons de minimiser, en faisant varier $\tilde{x}$ sur la boule unité, la norme de l'erreur à postériori dépendante de $\tilde{x}$ que nous avait donnée la solution du problème précédent.

C H A P I T R E II

ERREUR A POSTERIORI QUAND ON CONNAIT UNE VALEUR PROPRE NUMERIQUE $\tilde{\lambda}$ ET LE VECTEUR PROPRE NUMERIQUE $\tilde{x}$ CORRESPONDANT

Ce cas est redevable de la méthode qui nous est maintenant habituelle.

Posons le problème : nous avons une matrice carrée A_o. Nous cherchons une valeur propre et le vecteur propre correspondant par une méthode numérique quelconque (Givens, Jacobi, Rutishauser). Nous trouvons $\tilde{\lambda}$ et $\tilde{x}$. Nous allons chercher s'il existe des matrices A qui admettent $\tilde{\lambda}$ et $\tilde{x}$ comme valeur et vecteur propre exacts, et nous verrons si, parmi ces matrices A, il y en a une qui soit plus "proche", toujours au sens de la norme euclidienne sur R^{n^2}, de A_o. Pour cela nous utiliserons encore la méthode de Lagrange. Nous dériverons donc :

$$\sum_i \sum_j (a_{ij}-a^o_{ij})^2 + \mu^T(A\tilde{x}-\tilde{\lambda}\tilde{x}) \cdot \quad \text{on a } \mu = \begin{matrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{matrix}$$

On obtient :

$$2(a^*_{ij}-a^o_{ij}) + \mu_i \, x^*_j = o$$

ce qui s'écrit matriciellement :

$$\boxed{2(A^*-A_o) + \mu \tilde{X}^T = o}$$

il nous faut maintenant calculer la valeur de μ. Nous disposons pour cela de la contrainte :

$$A^{*}\tilde{x} = \tilde{\tilde{\lambda}}\tilde{x} \quad \text{d'où}$$

$$(A_o - \frac{\mu}{2}\tilde{x}^T)\tilde{x} = \tilde{\tilde{\lambda}}\tilde{x}$$

d'où encore :

$$A_o\tilde{x} - \tilde{\tilde{\lambda}}\tilde{x} = \frac{\mu}{2}$$

nous voyons donc que μ est le double du résidu η on peut donc encore écrire :

$$A^{*} - A_o = \eta\tilde{x}^T$$

et $$||A^{*} - A||^2 = ||(A_o - \lambda I)\,\tilde{\tilde{x}}\tilde{x}^T||^2$$

c'est le résultat cherché.

Remarque.

La matrice, admettant $\tilde{\lambda}$ et $\tilde{x}$ comme valeur et vecteur propre, qui était la plus proche de A_o est donc $A^{*} = A_o - \eta\tilde{x}^T$ or elle se déduit très facilement de l'équation matricielle $A_o\tilde{x} - \tilde{\tilde{\lambda}}\tilde{x} = \eta$ puisque $||\tilde{x}|| = 1$.

C H A P I T R E III

ERREUR A POSTERIORI DANS LE CALCUL DES VALEURS PROPRES $\tilde{\lambda}$

Nous allons utiliser dans ce chapitre les résultats du chapitre II.

Nous avons vu que l'erreur dans le cas où nous connaissions un vecteur propre numérique $\tilde{x}$ en plus de la valeur propre $\tilde{\lambda}$ était égale à $||(A_o-\tilde{\lambda}I)\,\tilde{x}\tilde{x}^T||$ d'autre part $\tilde{x}$ est de norme 1.

Pour résoudre le problème posé dans ce chapitre nous allons essayer d'éliminer $\tilde{x}$ pour cela nous chercherons le minimum de $||(A_o-\tilde{\lambda}I)\,\tilde{x}\tilde{x}^T||$ lorsque $\tilde{x}$ se déplace sur la boule unité. Pour ce faire, nous utiliserons encore une fois les multiplicateurs de Lagrange. Calculons d'abord la valeur des éléments de $(A_o-\lambda I)\,\tilde{x}\tilde{x}^T$ sachant que nous appelons dorénavant B la quantité $A_o - \lambda I$.

On a donc :

$$B\tilde{x} = \left[\sum_i b_{ji}\,\tilde{x}_i\right]_{j = 1\,..\,n}$$

$$B\tilde{x}\tilde{x}^T = \left[\sum_i b_{ji}\,\tilde{x}_i\,\tilde{x}_k\right]_{\substack{j = 1\,...\,n \\ k = 1\,...\,n}}$$

et $$||B\tilde{x}\tilde{x}^T||^2 = \sum_{j=1}^{n}\sum_{k=1}^{n}\left(\sum_{i=1}^{n} b_{ji}\,\tilde{x}_i\,\tilde{x}_k\right)^2 = F$$

Si nous dérivons ce carré de norme par rapport à x_ℓ on obtient :

$$\frac{\partial F}{\partial x_\ell} = 2 \sum_{j=1}^{n} \left[\left(\sum_{i=1}^{n} b_{ji}\, x_i \right) \left\{ b_{j\ell} \left(\sum_{k=1}^{n} x_k^2 \right) + x_\ell \left(\sum_{i=1}^{n} b_{ji}\, x_i \right) \right\} \right]$$

Ce que l'on peut encore écrire :

$$2 \sum_{j} \left[b_{j\ell} \left(\sum_{i} b_{ji}\, x_i \right) \left(\sum_{k} \tilde{x}_k^2 \right) + x_\ell \left(\sum_{i} b_{ji}\, x_i \right)^2 \right]$$

et matriciellement :

$$2 \left[||x||^2 \, B_\ell^T \circ B\tilde{X} + x_\ell X^T B^T B X \right]$$

nous utilisons la méthode des multiplicateurs de Lagrange. Nous avons ici une seule contrainte $||x||^2 = 1$. Nous devons ajouter $\mu \frac{\partial (||x||^2 - 1)}{\partial x_\ell}$ au résultat précédent. Nous obtiendrons, en écrivant toutes les dérivées partielles ($\ell = 1, 2, \ldots n$) les unes sous les autres, un système qui, matriciellement, s'exprimera par :

$$\frac{\partial \varphi}{\partial x} = X^T X B^T B X + X X^T B^T B X = \mu X$$

BXX^TB^T est un scalaire, on a donc

$$\frac{\partial \varphi}{\partial x} = \left[||X||^2 \, B^T B + (||BX||^2 - \mu) I \right] X$$

ou encore :

$$\frac{\partial \varphi}{\partial X} = \left[B^T B + \left(\frac{||BX||^2 - \mu}{||X||^2} \right) I \right] X = o$$

et

$$\left[B^T B - kI \right] X = o$$

pour avoir une solution non nulle de ce problème il faut que

det $\left[B^T B - kI \right] = o$ donc que k soit valeur propre. $B^T B$ est symétrique, donc, d'après le théorème de SCHUR , il y a n valeurs propres réelles. D'autre part, X est un des vecteurs propres. Comme μ ne sert à rien d'autre qu'à assurer la normalisation de X les vecteurs $\tilde{X}$ cherchés seront donc les n vecteurs propres normalisés de $B^T B$.

Or :

$$||B\tilde{X}_i \tilde{X}_i^T||^2 = \text{trace } (\tilde{X}_i \tilde{X}_i^T B^T B \tilde{X}_i \tilde{X}_i^T)$$

Comme

$$B^T B \tilde{X}_i = \lambda_i \tilde{X}_i$$

donc

$$||B\tilde{X}_i \tilde{X}_j^T||^2 = \text{trace } (\lambda_i X_i \underbrace{X_i^T X_i}_{=1} X_i^T)$$

$$= \text{trace } (\lambda_i X_i X_i^T)$$

comme $$\sum_{j=1}^{n} \sum_{i=1}^{n} (x_i x_j)^2 = \sum_{i=1}^{n} x_i^2 \sum_{j=1}^{n} x_j^2 = 1$$

on a $||B\tilde{x}_i \tilde{x}_j^T||^2 = \lambda_i \sum_{j=1}^{n} x_j^2 = \lambda_i$

la solution qui nous intéresse est donc celle qui correspond à λ_i minimum.

CHAPITRE-IV

PROCEDURE ALGOL

```
Procédure ERPOSVALPRO (AoR, LANDAR,N,ERREUR) ;
réel tableau AoR ; Réel LANDAR, ERREUR ; Entier N ;
Commentaire Cette procédure calcule l'erreur a postériori dans le calcul des valeurs
propres d'une matrice N × N appelée Aor. La valeur propre approchée s'appelle LANDAR ;
début entier i, j, k ; réel tableau MUDEU [1 : N],
     VR [1 : N, 1 : N] ,
     réel EPS ;
     procédure JACOBI (N, AR, LAMBDA, VR, EPSILON) ;
     entier N ; réel EPSILON ; réel tableau AR, LAMBDA, VR ;
     début entier i, j, p, q, k ; réel X, Y, WR, E, F, D1, D2, D, MUR, R, CR, G, H,
          UR, ZR, S, T ;
          procédure TRANSFO (R, TR, UR, WR, ZR) ;
          réel R, TR, UR, WR, ZR ;
          début ZR := R * UR + TR * WR ;
               si ABS(ZR) ≤ 10^-15 alors ZR := 0 ;
          fin
          pour i := 1 pas 1 jusqua N faire
          pour j := 1 pas 1 jusqua N faire
          VR[I,j] := si i = J alors 1 sinon 0 ;
          itération : pour p := 1 pas 1 jusqua N-1 faire
          pour Q := P+1 pas 1 jusqua N faire
          si ABS (AR[P,Q]) ≠ 0 alors
          début X := AR[P,P] ; Y := AR[Q,Q] ; WR := AR|P,Q| ;
               E := X-Y ; F = RAC3(E↑2 + 4 * WR↑ 2) ;
               D1 := E+F ; D2 := E-F ;
               D := si ABS (D1) > ABS (D2) alors D1 sinon D2 ;
               MUR := 2 * WR/D ;
               R := 1/RAC2 (1+MUR ↑ 2) ;
               CR := R * MUR ; G := 2 * (CR * AR[P,Q])* R ;
               H := X * R ↑ 2 + Y * CR ↑ 2 ;
               T := Y * R ↑ 2 + X * CR ↑ 2 ;
               AR[P,P] := G + H ; AR|Q,Q| := T-G ;
               AR[P,Q] := 0 ;
               si P ≥ 2 alors pour K := 1 pas 1 jusqua P-1 faire
```

```
        début UR := AR[P,K] ; WR := AR[K,Q] ;
            TRANSFO (R,-CR, UR, WR, ZR) ;
            AR[K,P] := ZR ;
            TRANSFO (R, CR, WR, UR, ZR) ;
            AR[K,Q] := ZR
        fin ;
        si Q > P + 2 alors pour K := P+1 pas 1 jusqua Q-1 faire
        début VR := AR[P,K] ; WR := AR[K,Q] ;
            TRANSFO (R, -CR, WR, UR, ZR) ;
            AR[K,Q] := ZR ;
            TRANSFO (R, CR, UR, WR, ZR) ;
            AR[P,K] := ZR
        fin ;
        si Q < N-1 alors pour K := QM1 pas 1 jusqua N faire
        début UR := AR[P,K] ; WR := AR[Q,K] ;
            TRANSFO (R, CR, UR, WR, ZR) ;
            AR[P,K] := ZR ;
            TRANSFO (R, -CR, WR, UR, ZR) ;
            AR[Q,K] := ZR
        fin ;
        pour i := 1 pas 1 jusqua N faire
        début UR := VR [I,P] ; WR := VR [I,Q] ;
            TRANSFO (R, CR, UR, WR, ZR)
            VR[I,P] := ZR ;
            TRANSFO (R, -CR, WR, UR, ZR) ;
            VR[I,Q] := ZR
        fin ;
fin itération ;
S := T := 0 ;
Pour i := 1 pas 1 jusqua N-1 faire
début pour j := i+1 pas 1 jusqua N faire
      S := S + ABS (AR[I,j]) ; T := T + ABS (AR[I,I]) ;
fin ;
```

```
                si T = 0 alors allera FINI ;
                si S/T > EPSILON alors allera ITERATION ;
                FINI : pour J := 1 pas 1 jusqua N faire
                LAMBDA[J] := AR [J,J]
   fin JACOBI ;
   pour i := 1 pas 1 jusqua N faire
   AoR[I,I] := AoR[I,I] - LANDAR ;
   pour i := 1 pas 1 jusqua N faire
   pour j := 1 pas 1 jusqua N faire
   début VR [i,j] := 0 ;
        pour K := 1 pas 1 jusqua N faire
        VR[i,j] := VR[i,j]+ AoR[I,K] * AoR[J,K] ;
   fin ;
   EPS := 10^-8 ;
   JACOBI (N, VR, MUDEU, AOR, EPS) ;
   ERREUR := MUDEU[1] ;
   pour i := 2 pas 1 jusqua N faire
   si MUDEU[I] < ERREUR alors ERREUR := MUDEU[I] ;
fin ERPOSVALPRO ;
```

Exemples numériques

Soient la matrice

$$\begin{bmatrix} 9 & 1 & -2 & 1 \\ 1 & 8 & -3 & -2 \\ 2 & -3 & 7 & -1 \\ 1 & -2 & -1 & 6 \end{bmatrix}$$

Cette matrice à comme valeur propre exacte 6.
Si on prend comme valeur propre approchée $\tilde{\lambda} = 6{,}0004$ on trouve comme carré de l'erreur $d^2 = 0{,}10425 \quad 10^{-7}$ si $\tilde{\lambda}$ vaut 6,75 alors $d^2 = 0{,}56249$
Considérons maintenant la matrice

$$\begin{bmatrix} 33 & -3 & 0 & -4 & 0 & 8 & 0 & -4 \\ 3 & 33 & 4 & 0 & -8 & 0 & 4 & 0 \\ 0 & 4 & 29 & 1 & -12 & -2 & -8 & -2 \\ 4 & 0 & 1 & 29 & -2 & -12 & -2 & -8 \\ 0 & -8 & -12 & -2 & 25 & 1 & -4 & -2 \\ 8 & 0 & -2 & -12 & 1 & 25 & -2 & -4 \\ 0 & 4 & -8 & -2 & -4 & -2 & 21 & 1 \\ 4 & 0 & -2 & -8 & -2 & -4 & 1 & 21 \end{bmatrix}$$

qui admet 6 et 48 comme valeur propre exacte

$\tilde{\lambda} = 6{,}08$ donne $d^2 = 0{,}63906 \quad 10^{-2}$
et $\tilde{\lambda} = 48.005$ donne $d^2 = 0{,}23329 \quad 10^{-4}$

Le programme est rapide : 22 millièmes d'heures pour 6 calculs d'erreurs appartenantnà des matrices 8 × 8.

BIBLIOGRAPHIE

PARTIE - C

On trouvera des procédés d'analyse directe d'erreur dans le calcul des valeurs propres dans les livres de :

A.M. TURING - ROUNDING-OFF ERRORS IN MATRIX PROCESSES
Quart. J. MECH. APPL. MATH. pp. 287-308

[2] J.H. WILKINSON - ROUNDING ERRORS IN ALGEBRAIC PROCESSES
Her majesty's stationery office 1963

On trouvera d'ailleurs dans ce dernier livre une étude du problème de l'estimation à postériori mais qui n'est résolue que pour une matrice symétrique.
Les méthodes de calcul des vecteurs et valeurs propres sont tirées du livre.

N. GASTINEL - ANALYSE NUMERIQUE LINEAIRE
Hermann Editeur.

- D -

ERREURS A POSTERIORI

DANS LE CAS

DES ESPACES DE HILBERT

CHAPITRE I

INTRODUCTION

Dans cette partie, nous étudierons l'erreur à postériori dans un espace de Hilbert pour les problèmes linéaires. Nous nous placerons dans un espace de Hilbert H séparable sur R. Le produit scalaire en sera noté (,) et la norme correspondante $||\ \ ||$.

Soit $\mathcal{L}(H,H)$ l'ensemble des endomorphismes continus de H dans H. On définira la norme de l'un deux A, par $||A|| = \sup_{||x||=1} ||Ax||$.

Nous définirons ensuite ce que nous appelerons un double produit scalaire et, avec la norme correspondante, on définira une "distance entre problèmes".

Nous pourrons ainsi, à l'aide d'une adaptation de la méthode de Lagrange, estimer le minimum de la distance entre le problème résolu et le problème effectif.

CHAPITRE II

RESULTATS PRELIMINAIRES

Soit donc H un espace de Hilbert séparable sur $\mathbb{R}$ $A \in \mathcal{L}(H,H)$ et soient $\{x_p\}$, $\{y_q\}$ $(p,q=1,2,....)$ deux systèmes orthonormés complets quelconques de H. C'est-à-dire tels que :

$$(x_i,x_j) = (y_i,y_j) = \delta_{ij} \qquad i,j = 1, 2,...$$

Définissons la quantité :

$$N^2(A,\{x_p\},\{y_q\}) = \sum_{p,q} |(Ax_p,y_q)|^2 \qquad (1)$$

Celle-ci va nous servir à caractériser une partie α de $\mathcal{L}(H,H)$ qui sera l'ensemble des endomorphismes A tels que $N^2(A,\{x_p\},\{y_q\})$ soit fini.

Nous allons d'abord montrer que α ne dépend pas de $\{x_p\}$ ou de $\{y_q\}$. Il suffit de vérifier que cela est vrai pour un A quelconque $\in \alpha$ $\sum_{p,q} |(Ax_p,y_q)|$ étant une série à termes positifs nous n'avons pas à nous préoccuper de l'ordre des sommations et nous pouvons donc écrire :

$$\sum_{p,q} |(Ax_p,y_q)|^2 = \sum_{p=1}^{\infty} \left(\sum_{q=1}^{\infty} |(Ax_p,y_q)|^2 \right)$$

mais (Ax_p,y_q) sont les coefficients de Fourier de Ax_p sur la base $\{y_q\}$. On a donc l'égalité de Bessel-Parseval :

$$\sum_{q=1}^{\infty} |(Ax_p,y_q)|^2 = ||Ax_p||^2$$

et (1) s'écrit donc :

$$\sum_{p,q} |(Ax_p, y_q)|^2 = \sum_{p=1}^{\infty} ||Ax_p||^2$$

or $$(Ax_p, y_q) = (x_p, A^* y_q) = (A^* y_q, x_p)$$

donc $$\sum_{p,q} |(Ax_p, y_q)|^2 = \sum_{q=1}^{\infty} ||A^* y_q||^2$$

Comme A^* ne dépend que de A et que les deux systèmes $\{x_p\}$ et $\{y_q\}$ sont indépendants, on voit donc que la quantité N^2 ne dépend que de A (et aussi que $N(A) = N(A^*)$).

Montrons que $\mathcal{a}$ est un sous espace vectoriel de $\mathcal{L}(H,H)$

$\forall \lambda \in \mathbb{R}$ -on a $N(\lambda A) = |\lambda| \cdot N(A)$ donc $\lambda A \in \mathcal{a}$ si A y appartient

Montrons que si A et $B \in \mathcal{a}$ alors $A + B$ aussi

En effet, soit $\{x_p\}$ un système quelconque orthonormé. On a :

$$||Ax_p + Bx_p||^2 = ||Ax_p||^2 + ||Bx_p||^2 + 2(Ax_p, Bx_p)$$

$$||Ax_p - Bx_p||^2 = ||Ax_p||^2 + ||Bx_p||^2 - 2(Ax_p, Bx_p)$$

et en sommant ces deux identités

$$||Ax_p + Bx_p||^2 \leq ||Ax_p + Bx_p||^2 + ||Ax_p - Bx_p||^2 = 2(||Ax_p||^2 + ||Bx_p||^2)$$

ce qui montre bien que $(A+B) \in \mathcal{a}$ si A et B y appartiennent.

Montrons maintenant que N est une norme $N(\lambda A) = |\lambda| \circ N(A)$ est trivial.

Montrons que $N(A+B) \leqslant N(A) + N(B)$

Pour cela considérons l'espace $\ell_2 \quad \ell_2 \subset (R\times R\times R\times \ldots)$. Un élément $x = \{x_1, x_2, \ldots\}$ est dit appartenir à ℓ_2 si $\Sigma\, x_i^2$ existe. C'est d'ailleurs le carré de la norme habituellement attachée à ℓ_2.

$$\Big|\Big| \sum_{p=1}^{\infty} ||Ax_p + Bx_p||_H \Big|\Big|_{\ell_2} \leqslant \Big|\Big| \sum_{p=1}^{\infty} (||Ax_p||_H + ||Bx_p||_H) \Big|\Big|_{\ell_2}$$

$$\leqslant \Big|\Big| \sum_{p=1}^{\infty} ||Ax_p||_H \Big|\Big|_{\ell_2} + \Big|\Big| \sum_{p=1}^{\infty} ||Bx_p||_H \Big|\Big|_{\ell_2}$$

donc $N(A+B) \leqslant N(A) + N(B)$

et $N(A) = o$ entraînant $Ax_p = o \quad \forall x_p \in$ système orthonormé quelconque, on a donc $A = o$

N est bien une norme.

<u>Remarque</u>.

Si $||x|| = 1$ on peut prendre $x = \{x_1, o, o, ..\}$

et donc $N^2(A) \geqslant ||Ax_1||^2 = ||Ax||^2$

donc $||A|| \leqslant N(A)$.

La norme N dérive du double produit scalaire $((A,B)) = \sum_{p=1}^{\infty} (Ax_p, Bx_p)$

(double ne signifie rien en lui-même, ce n'est qu'une notation) on voit que $((A,A)) = N^2(A)$ montrons que si A et B $\in \mathcal{A}$ et si $\{x_p\}$ est un système orthonormé alors $((A,B))$ converge absolument et, de plus, $((A,B))$ ne dépend pas de $\{x_p\}$.

L'inégalité de Schwarz donne immédiatement le résultat :

$$|(Ax_p, Bx_p)| \leq ||Ax_p|| \cdot ||Bx_p||$$

or $$(||Ax_p|| - ||Bx_p||)^2 \geq 0$$

donne $$||Ax_p||^2 + ||Bx_p||^2 \geq 2||Ax_p|| \cdot ||Bx_p||$$

donc $$|(Ax_p, Bx_p)| \leq \frac{1}{2} (||Ax_p||^2 + ||Bx_p||^2)$$

or A et B $\in \mathcal{A}$, on a $\sum_p ||Ax_p||^2$ et $\sum_p ||Bx_p||^2$ qui convergent.

D'autre part, soit $\{y_p\}$ un autre système orthonormé nous fixons p.

L'évaluation du produit scalaire en fonction des coefficients de Fourier dans le système $\{y_q\}$ donne :

$$(Ax_p, Bx_p) = \sum_{q=1}^{\infty} (Ax_p, y_q) \cdot (Bx_p \cdot y_q)$$

On peut donc écrire (les séries étant convergentes on ne se préoccupera pas de l'ordre de sommation

$$((A,B)) = \sum_{p,q}^{\infty} (Ax_p, y_q) \cdot (Bx_p, g_q)$$

$$= \sum_{p,q}^{\infty} (x_p, A^* y_q) \cdot (x_p\ B^* y_q)$$

$$= \sum_{q=1}^{\infty} (A^* y_q, B^* y_q)$$

Cette expression ne dépend pas de $\{x_p\}$. (voir début de II).

$((,))$ est un produit scalaire. (La linéarité est évidente, la commutativité aussi) $\mathcal{A}$ est donc un espace pré-hilbertien.

Montrons que $\mathcal{A}$ est complet.

Soit donc $\{A_n\}$ une suite de Cauchy d'endomorphisme de $\mathcal{A}$.

$\forall\ \varepsilon > 0 \longrightarrow 0$, $\exists$ tel que m et $n \geqslant$ entraîne $N(A_n - A_m) \leqslant \varepsilon$

or $N(A) \geqslant ||A||$ entraîne $||A_n - A_m|| \leqslant \varepsilon$ si m et $n \geqslant$

Soit $\hat{x} \in H$.

On a $||A_n\hat{x} - A_m\hat{x}|| \leqslant ||A_n - A_m|| \cdot ||\hat{x}|| \leqslant \varepsilon \cdot ||\hat{x}||$

donc $\{A_n\hat{x}\}$ est une suite de Cauchy dans H et H étant complet. On montre alors (cf yoshida page 69) que $A_n\hat{x}$ tend vers $T\hat{x}$ et ce quel que soit $\hat{x}$. T est un opérateur linéaire et borné de $\mathcal{L}(H,H)$.

Montrons que $T \in \mathcal{A}$

$N(A_n) \leqslant N(A_{m_o} - A_n) + N(A_{m_o})$ si n et $m_o \geqslant$

alors on a :

$$N(A_n) \leqslant \varepsilon + N(A_{m_o}) = K$$

$N(A_n) \leqslant K$ quel que soit n

Soit donc $\{x_p\}$ une suite orthonormale de H fixons P

$$\sum_{p=1}^{P} ||A_n x_p||^2 < N^2(A_n) \leqslant K^2$$

nous voyons que la sommation de gauche est bornée par la même quantité quel que soit n et P

donc $\sum_{p=1}^{\infty} ||Tx_p||^2 \leqslant K^2$ donc $N(T) < +\infty$ et par conséquent $T \in \mathfrak{a}$.

$\mathfrak{a}$ est donc complet et donc hulbertien.

De plus, $\mathfrak{a}$ est un idéal bilatère pour l'algèbre $\mathfrak{L}(H,H)$.

En effet soit $\{x_p\}$ un système orthonormé $A \in \mathfrak{a}$ et $B \in \mathfrak{L}(H,H)$ alors

$$\sum_{p=1}^{P} ||B \cdot Ax_p||^2 \leqslant ||B||^2 \cdot \sum_{p=1}^{P} ||Ax_p||^2 \leqslant ||B||^2 \cdot N^2(A)$$

Or $\mathfrak{L}(H,H)$ est l'espace des endomorphismes continus donc bornés, par conséquent :

$N(B \cdot A) < +\infty$ donc $B \cdot A \in \mathfrak{a}$ d'autre part si $B \in \mathfrak{L}(H,H)$, B^* aussi si $A^* \in \mathfrak{a}$ alors $B^* \cdot A^* \in \mathfrak{a}$ donc $(B^* \cdot A^*)^*$ aussi et finalement $A \cdot B \in \mathfrak{a}$.

Nous allons étudier un endomorphisme particulier car les résultats obtenus nous servirons pour la suite.

Soient u et v deux élément fixés $\in]-[$

On appelle trans-vection et on note uv^T l'endomorphisme qui à $x \in H$ fait correspondre le vecteur $(v,x)u$

Montrons d'abord que $uv^T \in \mathfrak{a}$

En effet, soit encore $\{x_p\}$ un système orthonormé

$uv^T(x_p) = (v,x_p)\cdot u$ et par conséquent

$$||uv^T(x_p)||^2 = |(v,x_p)|^2 \cdot ||u||^2$$

donc $$N^2(uv^T) = \sum_{p=1}^{\infty} ||uv^T(x_p)||^2 = ||u||^2 \cdot \Sigma\, |(v,x_p)|^2$$

$$= ||u||^2 \cdot ||v||^2$$

d'après l'égalité de Bessel-Parseval.

$$N(uv^T) = ||u|| \cdot ||v||$$

Etudions maintenant le double produit scalaire $((A,uv^T))$

$$((A,uv^T)) = \sum_{p=1}^{\infty} (Ax_p, uv^T x_p)$$

$$= \sum_{p=1}^{\infty} (Ax_p, (v,x_p)u)$$

$$= \sum_{p=1}^{\infty} (v,x_p)(A^*u,x_p)$$

$$= (v,A^*u) = (u,Av)$$

$$((A,uv^T)) = (u,Av)$$

CHAPITRE III

ETUDE THEORIQUE

Soit $A_o \in \mathcal{L}(H,H)$ et $b_o \in H$

Considérons l'équation $A_o x = b_o$

Nous supposons avoir trouvé une solution numérique $\tilde{x}$ de ce problème.

Suivant la démarche habituelle, nous sommes conduit à considérer des couples (A,b)

$A \in \mathcal{L}(H,H)$ et $b \in H$ tels que $A\tilde{x} = b$

Nous appelerons par définition :

distance d'un problème (A_o,b_o) à un autre (A,b) la quantité :

$$\Phi(A,b) = N^2(A-A_o) + ||b - b_o||^2$$

Pour que cette mesure ait un sens nous nous restreindrons par convention à étudier l'erreur à postériori uniquement pour les A tels que $A - A_o \in \alpha$.

Suivant l'habitude, nous conviendrons que $\tilde{x}$ est "compatible" si $\phi(A,b) \leqslant \eta$

η étant un seuil fixé d'avance.

Nous cherchons donc le minimum de $\phi(A,b)$ avec la contrainte suivante : $A\tilde{x} = b$.

Nous utiliserons encore les multiplicateurs de Lagrange mais sous une forme adaptée.

Pour trouver le minimum de $\phi(A,b)$ sachant que $A\tilde{x} = b$ nous étudierons :

$$\phi(A,b) = N^2(A-A_o) + ||b - b_o||^2 + 2(z, A\tilde{x}-b)$$

$z \in H$ étant une espèce de multiplicateur de Lagrange.

Faisons varier A et b de δA et δb

$$\phi(A+\delta A, b+\delta b) = \left((A-A_o+\delta A, A-A_o+\delta A)\right)$$

$$+ (b-b_o+\delta b, b-b_o+\delta b) + 2(z, A\tilde{x}-b+\delta A x-\delta b)$$

$$= N^2(A-A_o) + N^2(\delta A) + 2\left((\delta A, A-A_o)\right) + ||b-b_o||^2$$

$$+ ||\delta b||^2 + 2(\delta b, b-b_o) + 2(z, A\tilde{x}-b) + 2(z, \delta A\tilde{x}-\delta b)$$

$$= \phi(A,b) + N^2(\delta A) + ||\delta b||^2 + 2(z, \delta A\tilde{x}) + 2(\delta b, b-b_o-z)$$

Or nous avons vu que $(z, \delta A\tilde{x}) = \left((\delta A, z\tilde{x}^T)\right)$

donc $$\phi(A+\delta A, b+\delta b) = \phi(A,b) + N^2(\delta A) + ||\delta b||^2$$

$$+ 2\left((\delta A, A-A_o+z\tilde{x}^T)\right)$$

$$+ 2(\delta b, b-b_o-z)$$

nous savons que nous aurons un extrêmum de ϕ lorsque la variation de ϕ en fonction de $\pm\delta A$ et $\pm\delta b$ sera minimum. Dans notre cas cela revient à annuler

$$\left((\delta A, A-A_o+z\tilde{x}^T)\right)$$

et $(\delta b, b-b_o-z)$

Nous ne disposons que de A, b et z

Nous aurons donc les trois équations à trois inconnues

$$\begin{cases} A - A_o + z\tilde{x}^T = o & (1) \\ b - b_o - z = o & (2) \\ A\tilde{x} - b = o & (3) \end{cases}$$

Résolvons par rapport à z

$$A^* = A_o - z\tilde{x}^T$$

$$b^* = b_o + z$$

Reportons dans (3)

$$(A_o - z\tilde{x}^T)x - b_o - z = o$$

posons $A_o\tilde{x} - b = \eta$ alors $z = \dfrac{1}{1+||\tilde{x}||^2}\,\eta$

$$A^* - A_o = \frac{1}{1+||\tilde{x}||^2}\,\eta\tilde{x} \qquad b^* - b_o = \frac{\eta}{1+||\tilde{x}||^2}$$

On en déduit

$$\phi(A^*, b^*) = \frac{||\eta||^2}{1+||\tilde{x}||^2}$$

C H A P I T R E IV

APPLICATIONS AUX METHODES DE GALERKIN

Soit une forme linéaire $a(u,v)$ qui applique $H \times H$ sur $\mathbb{R}$ telle que :

$$\begin{cases} |a(u,v)| \leqslant \gamma \cdot ||u|| \cdot ||v|| & \gamma > 0 \quad ① \\ a(u,u) \geqslant \delta \, ||u||^2 & \delta > 0 \quad ② \end{cases}$$

Soit d'autre part $L(v)$ une forme linéaire qui applique H sur $\mathbb{R}$. L est continue.

Nous voulons résoudre le problème : trouver u tel que $a(u,v) = L(v)$ quel que soit $v \in H$.

Remarquons d'abore que les hypothèses ① et ② associées au lemme de LAX-MILGRAM (Yoshida page 92) nous indiquent qu'il existe un isomorphisme de H sur lui-même. (Endomorphisme bi-univoque et bi-continu) tel que :

$$a(u,v) = (Su,v) \qquad \forall u,v \in H$$

De même, la continuité de L entraîne l'existence de $\varphi \in H$ tel que : $L(v) = (\varphi, v)$ $\forall v$ le problème admet donc la forme variationnelle trouver u tel que $(Su,v) = (\varphi,v)$ ou encore $Su = \varphi$.

Appliquons les résultats trouvés dans III à l'équation $S_o u = \varphi_o$

Nous avons la solution numérique $\tilde{u}$. Elle est solution exacte de $S^* u = \varphi^*$

$$S^* = S_o - \frac{1}{1+||\tilde{u}||^2} (S_o\tilde{u} - \varphi_o) \cdot \tilde{u}$$

$$\varphi^* = \varphi_o + \frac{(S_o\tilde{u} - \varphi_o)}{1+||\tilde{u}||^2}$$

et $$\phi(S^*, \varphi^*) = \frac{||S_o\tilde{u} - \varphi_o||^2}{1+||\tilde{u}||^2}$$

si on appelle η la quantité $S_o\tilde{u} - \varphi_o$

alors

$$\boxed{\phi(S^*, \varphi^*) = \frac{||\eta||^2}{1+||\tilde{u}||^2}}$$

Supposons que pour résoudre le problème nous ayons utilisé la méthode de Galerkine. Pour cela, on dispose d'un système orthonormé $\{\varphi_1, \varphi_2, \ldots\}$ d'éléments de H.

Nous cherchons une solution de la forme

$$u_n = \sum_{i=1}^{n} a_i \varphi_i$$

telle que :

$$a(u_n, \varphi_i) = L(\varphi_i) \quad (i=1,2,\ldots,n)$$

Nous ne connaissons pas η mais nous savons que

$$(\eta, \varphi_j) = a(u_n, \varphi_j) - (\varphi, \varphi_j)$$

$$= a(u_n, \varphi_j) - L(\varphi_j)$$

On voit donc que $\eta = \sum_{i=n+1}^{\infty} [a(u_n, \varphi_i) - L(\varphi_i)] \varphi_i$

et par conséquent :

$$\phi(S^*, \eta^*) = \frac{\sum_{i=n+1}^{\infty} [a(u_n, \varphi_i) - L(\varphi_i)]^2}{1 + ||u_n||^2}$$

On ne peut, bien sûr, avoir seulement qu'une estimation grossière de cette erreur car nous ne pouvons calculer une série infinie

ESTIMATION A POSTERIORI DE L'ERREUR DANS LE CAS DE L'EQUATION DE FREDHOLM

Soit à résoudre l'équation de Fredholm de seconde espèce : trouver $u \in H$ tel que :

$$Au(p) = u(p) - \int_{\Omega} K(p,Q)\, u(Q)\, d\Omega_Q = f(p)$$

sous les hypothèses classiques :

1° K est tel que $\int_{\Omega_p} \int_{\Omega_Q} K^2(p,Q)\, d\Omega_p\, d\Omega_Q < +\alpha$

2° $\int_{\Omega} f^2\, d\Omega$ existe

3° il existe au plus une solution $||u|| < \infty$

la formulation variationnelle de ce problème est classique : il s'agit d'égaler les coefficients de Fourier de Au et de f.
c'est-à-dire qu'il faut que :

$$(Au,v) = (f,v) \text{ quel que soit } v \in H$$

on notera aussi (Au,v) par a(u,v)

$$a(u,v) = (Au,v) = \int_{\Omega_p} \left[u - \int_{\Omega_Q} K(p,Q)\, u(p)\, d\Omega_Q \right] v(p)\, d\Omega_p$$

Montrons que cet opérateur a est borné (ou continu) c'est-à-dire que

$$|a(u,v)| \leqslant \gamma \cdot ||u|| \cdot ||v|| \qquad \gamma > 0$$

$$|a(u,v)| \leqslant \left| \int_{\Omega_p} u \cdot v \cdot d\Omega_p \right| + \left| \int_{\Omega_p} v(p)\, d\Omega_p \cdot \int_{\Omega_Q} K(p,Q)\, u(Q)\, d\Omega_Q \right|$$

or $\left| \int_{\Omega_p} u \cdot v \cdot d\Omega_p \right| \leqslant ||u|| \cdot ||v||$ inégalité de Schwartz.

Etudions l'autre partie :

$$\left| \int_{\Omega_p} v(p)\, d\Omega_p \int_{\Omega_Q} K(p,Q)\, u(Q)\, d\Omega_Q \right| \leqslant \left| \int_{\Omega_p} |v|\, d\Omega_p \left| \int_{\Omega_Q} Ku\, d\Omega_Q \right| \right|$$

$$\leqslant \left| \int_{\Omega_p} ||u|| \cdot v(p) \left[\int_{\Omega_Q} K^2(p,Q)\, d\Omega_Q \right]^{\frac{1}{2}} d\Omega_p \right.$$

à cause de l'inégalité de Schwartz

$$\leqslant ||u|| \cdot \int_{\Omega_p} v(p) \left(\int K^2(p,Q)\, d\Omega_Q \right)^{\frac{1}{2}} d\Omega_p$$

d'après la même inégalité.

$$\leqslant ||u|| \cdot ||v|| \cdot \int_{\Omega_p} \int_{\Omega_Q} K^2(p,Q) \leqslant \alpha \, ||u|| \cdot ||v||$$

donc $\quad |a(u,v)| \leqslant (1+\alpha) \, ||u|| \cdot ||v||$

ce qui montre bien que a est bornée et continue.

Cette démonstration est aussi valable pour l'équation de Fredholin de première espèce :

trouver u tel que $\int_{\Omega} K(p,Q) \, u(Q \, d\Omega_Q = f(p)$

Montrons que l'opérateur A est coercitif c'est-à-dire que (Au,u) ou $a(u,u) > \delta \cdot ||u||^2$ d'après la première démonstration

$$a(u,u) = ||u||^2 - \int_{\Omega_p} u(p) \, d\Omega_p \int_{\Omega_Q} K(p,Q) \, u(Q) \, d\Omega_Q$$

$$\geqslant ||u||^2 - \alpha \, ||u||^2$$

il faut que $a(u,u) > \delta \, ||u||^2 \qquad \delta > 0$

donc que $1 - \alpha > \delta$

$\alpha < 1 - \delta$;

c'est-à-dire que : $\int_{\Omega_p} \int_{\Omega_Q} K^2(p,Q) \, d\Omega_p \, d\Omega_Q < 1 - \delta$

alors le formalisme de IV est applicable.

Soit une base orthonormée $\{\varphi_1, \varphi_2 \ldots, \varphi_n, \ldots\}$

$$a(\varphi_j, \varphi_i) = (A\varphi_j, \varphi_i) = \int_\Omega \left(A\varphi_j(p)\right) \varphi_i(p)\, d\Omega_p$$

$$= \int_\Omega \left[\varphi_j(p) - \int_\Omega K(p,Q)\, \varphi_j(Q)\, d\Omega_Q\right] \varphi_i(p)\, d\Omega_p$$

$$= \int_\Omega \varphi_j(p) \cdot \varphi_i(p)\, d\Omega_p - \int_\Omega \varphi_i(p)\, d\Omega_p \int_\Omega K(p,Q)\, \varphi_j(Q)\, d\Omega_Q$$

$$= \delta_{ij} - \gamma_{ij}$$

Nous cherchons une solution de la forme

$$u_n = \sum_{i=1}^{n} a_i\, \varphi_i$$

avec, comme hypothèse : $(Au_n, \varphi_i) = (f, \varphi_i)$

pour $i = 1, 2, \ldots, n$.

$$(Au_n, \varphi_i) = a_i - \sum_{j=1}^{n} \gamma_{ij}\, a_j = (f, \varphi_i)$$

alors $$r = \sum_{i=n+1}^{\infty} \left[(Au_n, \varphi_i) - (f, \varphi_i)\right] \varphi_i$$

et $$\Phi(S_1, \varphi_1) = \frac{\sum_{i=1}^{n+1} \left[(Au_n, \varphi_i) - (f, \varphi_i)\right]^2}{1 + \Sigma\, a_i^2}$$

CAS DU PROBLEME DE DIRICHLET

Soient $a_{ij}(x)$ et $a_o(x)$ des fonctions mesurables bornées de Ω (ouvert borné de $\mathbb{R}^n$) et $f(x) \in L_2(\Omega)$. En prenant les dérivées faibles (dérivées au sens des distributions) le problème de Dirichlet s'énonce ainsi :

trouver $u \in H_o^1(\Omega)$ tel que :

$$- \sum_{i,j=1}^{n} \frac{\partial}{\partial x_i} \left(a_{ij}(x) \frac{\partial u}{\partial x_j} \right) + a_o(x) \cdot u = f(x)$$

Ce problème admet la forme variationnelle suivante :

$$a(u,v) = \sum_{i,j} \int_\Omega a_{ij}(x) \frac{\partial u}{\partial x_i} \cdot \frac{\partial v}{\partial x_j} \, dx + \int_\Omega a_o(x) \, u \cdot v \, dx = \int_\Omega f \cdot v \, dx$$

il faut trouver u de façon qu'on ait bien l'égalité quel que soit $v \in H_o^1(\Omega)$

le produit scalaire dans H_o^1 sera défini par :

$$(u,v)_{H_o^1} = \int_\Omega \left(u \cdot v + \sum_{i,j=1}^{n} \frac{\partial u}{\partial x_i} \cdot \frac{\partial v}{\partial x_j} \right) dx$$

De plus nous avons les hypothèses suivantes :

ⓐ $$\sum_{i,j}^{n} a_{ij}(x) \, \xi_i \, \xi_j \geqslant \alpha \left(|\xi_1|^2 + \ldots + |\xi_n|^2 \right)$$

$\alpha > 0$

ⓑ $$a_o(x) \geqslant \alpha_o > 0$$

Montrons que nous pouvons appliquer la méthode de Galerkine.

Pour cela, montrons d'abord que $a(u,u)$ est coercitive c'est-à-dire que : $a(u,u) \geqslant \delta \, ||u||^2$

$$a(u,u) = \int \sum_{i,j} a_{ij} \left(\frac{\partial u}{\partial x_i}\right)\left(\frac{\partial u}{\partial x_j}\right) + a_o \, u^2$$

d'après (a) on a :

$$a(u,u) \geqslant \int \alpha \sum \left(\frac{\partial u}{\partial x_i}\right)^2 + \alpha_o \, u^2$$

$$\geqslant \min(\alpha,\alpha_o) \int \sum \left(\frac{\partial u}{\partial x_i}\right)^2 + u^2$$

$$= \min(\alpha,\alpha_o) \, ||u||^2$$

Montrons maintenant que $a(u,v)$ est bornée ou continue, c'est-à-dire qu'il existe γ tel que :

$$|a(u,v)| \leqslant \gamma \cdot ||u|| \cdot ||v|| \qquad \gamma > 0$$

$$|a(u,v)| = \left| \int_\Omega \sum a_{ij} \left(\frac{\partial u}{\partial x_i}\right)\left(\frac{\partial v}{\partial x_j}\right) d\Omega + \int_\Omega a_o \, u \cdot v \, d\Omega \right|$$

$$\leqslant \int \sum |a_{ij}| \cdot \left|\frac{\partial u}{\partial x_i}\right| \cdot \left|\frac{\partial v}{\partial x_j}\right| d\Omega + \int |a_o| \cdot |u| \cdot |v| \, d\Omega$$

prenons $M = \text{Max}\left(a_{ij}(x) \text{ et } a_o(x)\right)$ quels que soient i,j et x.

alors $$|a(u,v)| \leqslant M \int \left(\sum_{i,j} \left|\frac{\partial u}{\partial x_i}\right| \cdot \left|\frac{\partial v}{\partial x_j}\right| + |u| \cdot |v| \right) d\Omega$$

$$\leqslant M \int \left(\left(\sum_i \left|\frac{\partial u}{\partial x_i}\right|\right)\left(\sum_j \left|\frac{\partial v}{\partial x_j}\right|\right) + |u| \cdot |v| \right) d\Omega$$

$$\leqslant M \int \left(\sum_{i=1}^{n} \left|\frac{\partial u}{\partial x_i}\right| + |u| \right)\left(\sum_j \left|\frac{\partial v}{\partial x_j}\right| + |v|\right) d\Omega$$

Or $$\sum_{i=1}^{n} \frac{|x_i|}{n} \leqslant \sqrt{x_1^2 + x_2^2 + \ldots + x_n^2}$$

donc $$\left|a(u,v)\right| \leqslant M \int (n+1)^2 \left(\Sigma(\frac{\partial u}{\partial x_i})^2 + u^2\right)^{\frac{1}{2}} \left(\Sigma(\frac{\partial v}{\partial x_j})^2 + v^2\right)^{\frac{1}{2}} d\Omega$$

et d'après l'inégalité de Schwartz

$$\leqslant M(n+1)^2 \left[\int \Sigma(\frac{\partial u}{\partial x_i})^2 + u^2\right]^{\frac{1}{2}} \left[\int \Sigma(\frac{\partial v}{\partial x_j})^2 + v^2\right]^{\frac{1}{2}}$$

$$\leqslant M(n+1)^2 \, ||u|| \cdot ||v||$$

Posons $M(n^2+1) = \gamma$ alors

$$|a(u,v)| \leqslant \gamma \, ||u|| \cdot ||u||$$

c'est le résultat cherché.

Soit donc une base $\{\varphi_1 \ldots . \varphi_n\}$ de $H_o^1(\Omega)$.

Nous chercherons une solution de la forme

$$u_n = \sum_{i=1}^{n} a_i \, \varphi_i$$

où les a_i sont tels que :

$$a(u_n, \varphi_i) = \int f, \varphi_i \, d\Omega$$

cherchons les composantes de $\eta = Su_n - \varphi$

Ce sont :

$$(Su_n - \varphi, \varphi_j) = a(u_n, \varphi_j) - (\varphi, \varphi_j)$$

$$= a(u_n, \varphi_j) - L(\varphi_j)$$

donc : $\eta = \sum_{i=n+1}^{\infty} a(u_n, \varphi_i) - L(\varphi_i)) \varphi_i$

et

$$\phi = \frac{\sum_{i=n+1}^{\infty} \left[a(u_n, \varphi_i) - L(\varphi_i)\right]^2}{1 + \Sigma a_i^2}$$

CHAPITRE - V

EXEMPLE NUMERIQUE

Soit à résoudre l'équation de Fredholm de seconde espèce

$$u(x) - \int_{-1}^{+1} \frac{3}{16} (x^2+y^2)\, u(y)\, dy = x^2 \qquad (1)$$

$$\int_{-1}^{+1} dx \int_{-1}^{+1} \frac{9}{256} (x^2+y^2)^2\, dy = \frac{1008}{11520}$$

Nous voyons que cette quantité est plus petite que 1 donc on aura bien continuité et coercivité.

Nous pouvons donc utiliser la méthode de Galerkine. Nous utiliserons comme système orthonormé l'ensemble des polynômes de Legendre : p_i $(i=1,2,....)$. Ceux-ci sont bien orthogonaux sur $[-1,+1]$, ils seront orthonormés si nous les multiplions par un facteur convenable. C'est-à-dire pour le polynôme p_n par la quantité $\sqrt{\frac{2}{2n+1}}$.

Nous cherchons une solution numérique de la forme $u_n = \sum_{i=o}^{n} a_i\, p_i$.

Pour cela nous égalerons les n premiers coefficients de Fourier des 2 membres de (1) où nous remplacerons u par u_n.

Cela donne :

$$\int_{-1}^{+1} \left[\sum_{i=o}^{n} q_i\, p_i(x) - \frac{3}{16} \int_{-1}^{+1} \Sigma\, a_i\, p_i(x) \left(x^2+y^2\right) dy \right] p_k(x)\, dx = \int_{-1}^{+1} x^2\, p_k(x)\, dx$$

pour k variant de o à n

nous savons que $p_i(x)\ p_k(x) = o$ si $i \neq k$

donc :

$$\int_{-1}^{+1} x^2\ p_k(x)\ dx = a_k - \frac{3}{16} \int_{-1}^{+1} \Sigma\ a_i\ p_i(x)\ p_k(x)\ dx \int_{-1}^{+1} y^2\ dy$$

$$- \frac{3}{16} \int_{-1}^{+1} \Sigma\ a_i\ p_i(x)\ p_k\ x^2\ dx \cdot \int_{-1}^{+1} dy$$

$$= a_k - \frac{1}{8}\ a_k - \frac{3}{8} \int_{-1}^{+1} \Sigma\ a_i\ p_i(x)\ p_k(x)\ x^2\ dx$$

étudions la dernière intégrale.

Nous savons qu'un polynôme de Legendre est orthogonal à tout polynôme de degré inférieur sur $\left[-1,+1\right]$. On en déduit que les seuls termes qui ne s'annuleront pas sont ceux où la différence entre i et k sera $\leqslant 2$. Calculons toutes les intégrales $\int p_i(x)\ p_k(x)\ x^2\ dx$ qui vérifient cette condition. (Nous nous limiterons à 4).

$$p_o = \sqrt{2}$$

$$p_1 = \sqrt{\frac{2}{3}} \cdot x$$

$$p_2 = \sqrt{\frac{2}{5}} \left(\frac{3}{2} x^2 - \frac{1}{2}\right)$$

$$p_3 = \sqrt{\frac{2}{7}} \left(\frac{5}{2} x^3 - \frac{3}{2} x\right)$$

$$p_4 = \sqrt{\frac{2}{9}} \left(\frac{35}{8} x^4 - \frac{15}{4} x^2 + \frac{3}{8}\right)$$

$\int p_o\, p_o\, x^2 = 1,3333$

$\int p_o\, p_1\, x^2 = o$ nous remarquons que lorsque $i + k$ est impair

$\int p_i\, p_k\, x^2$ est nul sur $\left[-1,+1\right]$

$\int p_o\, p_2\, x^2 = 0,2385$

$\int p_o\, p_3\, x^2 = \int p_o\, p_4\, x^2 = o$

$\int p_1\, p_1\, x^2 = 0,2666$

$\int p_1\, p_2\, x^2 = o$

$\int p_1\, p_3\, x^2 = 0,0498$

$\int p_1\, p_4\, x^2 = o$

$\int p_2\, p_2\, x^2 = 0,0838$

$\int p_2\, p_3\, x^2 = o$

$\int p_2\, p_4\, x^2 = 0,2271$

$\int p_3\, p_4\, x^2 = o$

Nous chercons une solution de la forme $\sum_{i=o}^{2} a_i\, p_i$

On a donc le système :

$$\frac{7}{8}\, a_o + 1,3333\, a_o + 0,2385\, a_2 = \int_{-1}^{+1} x^2\, p_o(x)\, dx$$

$$\frac{7}{8} a_1 + 0{,}2666\ a_1 = \int_{-1}^{+1} x^2\, p_1(x)\, dx$$

$$\frac{7}{8} a_2 + 0{,}2385\ a_o + 0{,}0838\ a_2 = \int_{-1}^{+1} x^2\, p_2(x)\, dx$$

$$\int_{-1}^{+1} x^2\, p_o(x)\, dx = 0{,}9428$$

$$\int_{-1}^{+1} x^2\, p_1(x)\, dx = o$$

$$\int_{-1}^{+1} x^2\, p_2(x)\, dx = 0{,}1696$$

$$2{,}2083\ a_o + 0{,}2385\ a_2 = 0{,}9428$$

$$1{,}1416\ a_1 = o$$

$$0{,}9588\ a_2 + 0{,}2385\ a_o = 0{,}1696$$

On en déduit $a_1 = o$

$a_o = 0{,}4190$

$a_2 = 3{,}8792$

d'où $u_2 = 3{,}8792 \cdot \sqrt{\frac{2}{7}} \left(\frac{3}{2} x^2 - \frac{1}{2}\right) + 0{,}4190 \times \sqrt{2}$

$$\boxed{u_2(x) = 3{,}0879\ x^2 - 0{,}4368}$$

Entamons maintenant le calcul à postériori de l'erreur.

Calculons $a(u_2, P_3) - L(P_3)$, il vaut :

$$\sqrt{\frac{2}{7}} \cdot \int_{-1}^{+1} \left[3{,}0879\ x^2 - 0{,}4368 - \frac{3}{16} \int_{-1}^{+1} (x^2+y^2)\,(3{,}0879\ x^2 - 0{,}4368)\ dy\right].$$

$$\left(\frac{5}{3} x^3 - \frac{3}{2} x\right) dx$$

Le terme dans le crochet est de degré pair, l'intégrale en x portera donc sur des termes de degré impair elle sera donc nulle.

Passons au terme suivant :

$$\sqrt{\frac{2}{9}} \int_{-1}^{+1} \left[3{,}0879\ x^2 - 0{,}4368 - \frac{3}{16} \int_{-1}^{+1} (x^2+y^2)\ (3{,}0879\ x^2 - 0{,}4368)\ dy\right]$$

$$\cdot \frac{35\ x^4 - 30\ x^2 + 3}{8}\ dx =$$

$$\sqrt{\frac{2}{9}} \int_{-1}^{+1} \left[3,0879\ x^2 - 0,4368 - \frac{3}{16} \left(3,0879\ x^2 - 0,4368\right) \left(\frac{5}{3}\right) \right] \times \frac{35\ x^4 - 30\ x^2 + 3}{8}\ dx$$

$$\times \sqrt{\frac{2}{9}} \int_{-1}^{+1} \left(2,1230\ x^2 - 0,3004\right) \frac{35\ x^4 - 30\ x^2 + 3}{8}\ dx$$

$$\frac{\sqrt{2}}{8\sqrt{9}} \int_{-1}^{+1} \left(74,3050\ x^6 - 169,1\ x^4 + 15,408\ x^2 - 0,9012\right) dx$$

$$\frac{\sqrt{2}}{24} \times 37,04 = 2,1855$$

Nous savons que l'erreur vaut :

$$\frac{\sum_{i=3}^{\infty} \left(a(u_2, p_i) - L(p_i)\right)^2}{1 + \sum a_i^2}$$

Nous avons calculé ici :

$$\frac{\sum_{i=3}^{4} \left(a(u_2, p_i) - L(p_i)\right)^2}{1 + (3,08)^2 + (0,43)^2}$$

qui sera une minoration de l'erreur, et on trouve d'ailleurs pour cette erreur tronquée la valeur 0,20.

Ce qui prouve qu'il était inutile de poursuivre les calculs au delà de la deuxième décimale dans le cas où l'on s'arrête à $n = 2$.

BIBLIOGRAPHIE

PARTIE - D

Pour l'étude des espaces de Hilbert, on pourra se référer à :

J. DIEUDONNE - FOUNDATION OF MODERN ANALYSIS
Academic Press 1960

L'utilisation de ces espaces en analyse numérique est traitée dans :

P.J. LAURENT - THEORIE DE L'APPROXIMATION
Cours de l'Université de Grenoble - 1964

On pourra se référer pour des précisions sur les opérateurs à :

K. YOSIDA - FUNCTIONAL ANALYSIS
Springer Verlag - Berlin 1965

N. DUNFORD & J.T. SCWHARTZ - LINEAR OPERATOR
Intersciences Publishers Inc. 1958

On aura de même tous les renseignements sur la méthode de Galerkine, l'équation de Fredholm, le problème de Dirichlet dans les livres suivants :

S.G. MIKHLIN - INTEGRAL EQUATIONS
Pergamon Press. New York 1957

J.L. LIONS - METHODES D'APPROXIMATION NUMERIQUE DES PROBLEMES AUX LIMITES DE LA PHYSIQUE MATHEMATIQUE
Laboratoire de Calcul Numérique de l'Institut Blaise Pascal.

Ce dernier livre donne d'ailleurs beaucoup de renseignements sur les dérivées faibles.

VU

Grenoble, le

Le Président de la Thèse

VU

Grenoble, le

Le Doyen de la Faculté des Sciences

VU, et permis d'imprimer,

Le Recteur de l'Académie de GRENOBLE