Contribution à l'étude de la méthode de Lanczos

Marc Lévy

T H E S E

-:-

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

pour obtenir

Le titre de Docteur de spécialité
Mathématiques Appliquées

par

Marc LEVY

Ingénieur IEG et IMAG

CONTRIBUTION A L'ETUDE DE LA METHODE DE LANCZOS

Thèse soutenue le 26 Juin 1961

Devant la commission d'Examen :

MM. KUNTZMANN : Président

GASTINEL

HACQUES

T H E S E

-:-

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

pour obtenir

Le titre de Docteur de spécialité
Mathématiques Appliquées

par

Marc LEVY

Ingénieur IEG et IMAG

CONTRIBUTION A L'ETUDE DE LA METHODE DE LANCZOS

Thèse soutenue le 26 Juin 1961

Devant la commission d'Examen :

MM. KUNTZMANN : Président

GASTINEL

HACQUES

FACULTE des SCIENCES de L'UNIVERSITE de GRENOBLE

-:-:-:-:-:-:-

Professeurs titulaires classés d'après leur ancienneté comme titulaires ou professeurs à titre personnel :

Doyens honoraires :

MM.	FALLOT P.	1931 Géologie et minéralogie
	GORET L.	1931 Géologie et Minéralogie

Doyen :

M.	WEIL L.	1950 Thermodynamique

MM.	DORIER A.	1941 Zoologie
	HEILMANN R.	1943 Chimie organique
	NEEL L.	1945 Physique expérimentale
	KRAVTCHENKO J.	1946 Mécanique rationnelle
	PARDE M.	1946 Potamologie
	BENOIT J.	1948 Radioélectricité
	CHENE M.	1949 Chimie papetière
	NOBECOURT P.	1949 Micrographie papetière
	FELICI N.	1951 Electrostatique
	KUNTZMANN J.	1952 Mathématiques Appliquées
	BARBIER	1.10.1953 Géologie appliquée
	SANTON L.	1.10.1953 Mécanique des fluides
	CHABAUTY C.	1.10.1954 Calcul différentiel et Intégral
	OZENDA P.	1.10.1954 Botanique
	FALLOT M.	1955 Physique industrielle
	GALVANI O.	1957 Mathématiques
	TRAYNARD Ph.	1.10.1958 Chimie générale
	SOUTIF M.	1.3 .1958 Physique générale

MM.	CRAYA A.	I.II.1953 Hydrodynamique
	REEB G.	I. I.1959 Statistiques Mathématiques
	REULOS R.	I.10.1959 Physique génarale
	WOFLERS F.	I.10.1959 Physique
	BESSON J.	I.10.1959 Chimie
	MOUSSA A.	I.12.1959 Chimie Nucléaire et Radioactivité
	AYANT Y.	I. I.1960 Physique approfondie
Mle	LUTZ E.	I.I. 1960 Mathématiques générales
MM.	GALLISSOT F.	I.I.1960 Mathématiques pures
	BLAMBERT M.	I.10.1960 Mathématiques
	BOUCHEZ R.	I. I.1961 Physique Nucléaire
	LLI BOUTRY	I. I.1961 Physique
	MICHEL	I. I.1961 Géologie et Minéralogie

Professeurs sans chaire

MM.	SILBERT R.	1954 Mécanique des Fluides
	MOUSSIEGT J.	I. I.1959 Electronique
	DESSAUX G.	I.10.1959 Physiologie animale
	PILLET E.	I.10.1960 Electrotechnique

Maîtres de conférence

MM.	BONNIER E.	I.10.1956 Chimie
	BARBIER J.CL.	I.10.1957 Physique
	PAUTHENET R.	2. I.1958 Electrotechnique
	BUYLE-BODIN M.	I.10.1958 Electronique
Mle	NAIM Linda	I.10.1958 Mathématiques
MM.	PERRET R.	I.10.1958 Servomécanismes
	DREYFUS	I.10.1958 Thermodynamique
	VAILLANT F.	I.11.1958 ZOOLOGIE et Hydrobiologie
Mme	SOUTIF J.	I.10.1959 Physique
	DEBELLIAS J.	I.10.1959 Géologie et Minéralogie

Mme KOFLER	I.IO.1958	Botanique
COHEN J.	I.IO.1959	Physique
Mme BARBIER	I.IO.1959	Electrochimie
BRISSONNEAU P.	I.IO.1959	Physique
ARNAUD P.	I.IO.1959	Chimie
VAUQUOIS B.	I.IO.1959	Mathématiques
DEPASSEL R.	I.II.1959	Mécanique des Fluides
ROBERT A.	I.II.1960	Chimie papetière

Délégués

MM. GERBER	I.IO.1960	Mathématiques
GIDON P.	I.IO.1960	Géologie et Minéralogie
DUCROS P.	I.IO.1960	Minéralogie et Cristallographie
HACQUES J.	I.IO.1960	Calcul Numérique
PEBAY - PEROULA	I.IO.1960	Physique
ANGLES D'AURIAC	I.IO.1960	Mécanique des Fluides
COUMES A.	I.IO.1960	Electronique
LAUCIA R.	I.IO.1960	Physique automatique
DODU J.	I.IO.1960	Mécanique des Fluides
BIAREZ	I.IO.1960	Mécanique physique
GASTINEL, LACAZE	Chargés d'Enseignement	Mathématiques Thermodynamique

I N T R O D U C T I O N

L'idée du travail ci-après m'a été suggérée au cours d'un stage effectué à la compagnie des Machines Bull que je tiens à remercier de son obligeance.

C'est ensuite au laboratoire de Calcul de l'Université de Grenoble que l'essentiel de ces études à été accompli sous la direction de Monsieur le Professeur Kuntzmann, dont les conseils et les directives précieuses m'ont guidé continuellement ; qu'il veuille bien trouver ici l'expression de ma profonde reconnaissance.

Je dois également beaucoup à Monsieur Gastinnl qui tant en ce qui concerne les études des erreurs que le calcul matriciel m'a apporté un appui constant et précieux.

Enfin je remercie tout le personnel du laboratoire pour son obligeance et son amabilité, et tout particulièrement Monsieur Bolliet.

J'ai voulu, dans ce travail, étudié les erreurs dans la méthode de Lanczos pour le calcul du polynôme caractéristique d'une matrice. Le sujet est trés vaste et je suis loin de l'avoir épuisé. Mon travail consiste essentiellement en un dégrossissage et un recensement des principales causes d'erreurs.

Cette étude comporte deux aspects absolument différents. Tout d'abord une étude théorique des erreurs et de leurs propagations. J'ai en particulier , au cours de cette partie, été amené à m'interesser, aux erreurs sur un produit scalaire en fonction du nombre de dimensions de l'espace.Le résultat obtenu parait assez interessant et je regrette beaucoup de n'avoir pu faire que des vérifications incomplétes de ce résultat.

Des vérifications complète nécessitant une machine beaucoup plus rapide que celle dontje disposais.

Le second aspect, lui, est essentiellement expérimental ; il s'agissait de vérifier certaines hypothèses et de déterminer certaines règles mais de façon purement expèrimentale.

Ce second aspect fait l'objet de la dernière partie de l'étude intitulée : "expèriences numériques"

Je tiens à remercier Mademoiselle Stankievitch, c'est à ces soins qu'est dûe la qualité de la présentation matérielle de cette thèse.

Que Monsieur Hacques et Monsieur Gastinel veuillent trouver ici l'expression de ma reconnaissance pour avoir eu l'obligeance d'accepter de faire partie de la commission d'examen.

B I B L I O G R A P H I E

a) Etude théorique des valeurs propres

(I) Wedderburn :Lectures on Matrices

(II) O. Schreier et E. Sperner : Introduction to Modern Algebra and Matrix Theory

(III) N. Gastinel : Analyse numérique linéaire (Université de Grenoble)

b) Méthode de Lanczos

(IV) Lanczos C. An iteration method for the solution of the eigenvalue problem of linear differential and integral operators.Journal of Research of the NBS I950,45, p 255.

(V) Rutishauser H. Beiträge zur Kenntnis des Biorthogonalisierungs-Algorithmus von Lanczos ZAMP, I958, 4, p 35

(VI) F. Genuys : La methode de Lanczos : Séminaire de l'AFCAL. I7 mars I958

(VII) Wilkinson : the evaluation of the zeros of ill conditionned polynomials (part II) - Numerische Mathematik I959 p I5C

c) Génération de points sur une hypersphère :

(VIII) Mervin E. Muller. a note on a method for generating points uniformly on N.dimensional spheres. Communications of the A.C.M. April 59 p I9

Première Partie

Remarques Théoriques sur la Méthode de Lanczos

Chapitre I

Présentation Elementaire de la Méthode

I° - **Principe** : Cette méthode consiste à trouver une matrice tridiagonale semblable à la matrice donnée, puis à calculer le polynome caractéristique de cette matrice tridiagonale. Ce dernier calcul est classique et simple. La partie essentielle de la méthode est donc la tridiagonalisation de la matrice.

Je vais d'abord exposer cette méthode dans le cas général ou le polynôme caractéristique est identique au polynôme minimal et où le calcul se déroule correctement. Dans une étude ultérieure, j'étudierai les cas particuliers susceptibles de se produire.

Lanczos a montré [(4)] que sa méthode était valable quelle que soit la matrice considérée : singulière ou régulière, défective ou non.

2° - **Tridiagonalisation** : A partir de maintenant, nous considererons l'espace des vecteurs-colonnes (ou espace des vecteurs à droite de la matrice) désigné par E_{cn}, n étant la dimension de l'e l'espace, ou E_c, si n est égal à la dimension de la matrice, et l'espace des vecteurs lignes (ou espace des vecteurs à gauche de la matrice), désigné par E_{ln} ou E_l.

Un vecteur ligne est orthogonal à un vecteur colonne si la somme des produits des composantes similaires est nul. Soit $b_I \in E_c$ et $C_I \in E_l$, deux vecteurs arbritaires.
Nous désignons par $\{a,b,c\}$ l'espace vectoriel engendré par a,b,c; a,b,c... appartenant à E_c (ou à E_l). Cherchons un vecteur $b_2 \in \{Ab_I ; b_I\}$ et orthogonal à c_I et un vecteur $c_2 \in \{c_IA ; c_I\}$ orthogonal à b_I.

Nous pouvons ecrire :

$$b_2 = Ab_I - \alpha_I b_I$$

d'ou nous tirons $\alpha_I = \dfrac{c_I A\ b_I}{c_I b_I}$

et : $c_2 = c_I A - \alpha'_I c_I$ d'ou nous déduisons $\alpha_I = \alpha'_I$

Nous poserons ci bi $= \sigma_i$

Nous définirons ensuite :

$b_3 \in \{Ab_2 ; b_2 ; b_I\} = \{b_I ; Ab_I ; A^2 b_I\}$; orthogonl à c_I et C_2

Et $c_3 \in \{c_2 A ; c_2 ; c_I\} = \{c_I ; c_I A ; c_I A^2\}$; orthogonl à b_I et b_2

soit : $b_3 = Ab_2 - \alpha_2 b_2 - \beta_I b_I$

d'ou $\alpha_2 = \dfrac{c_2 Ab_2}{\sigma_2}$; $\beta_I = \dfrac{c_I Ab_2}{\sigma_1}$

et $c_3 = c_2 A - \alpha'_2 c_2 - \beta'_I c_I$

d'ou nous tirons $\alpha'_2 = \alpha_2$; $\beta'_1 = \beta$

De façon générale, nous définirons :

$b_{k+I} \in \{Ab_k ; b_k ; b_{k-I} ; \ldots ; b_I\}$ orthogonal à $c_I ; c_2 ; \ldots ; c_k$

$c_{k+I} \in \{c_k A ; c_k ; c_{k-I} ; \ldots ; c_I\}$ orthogonal à $b_I ; b_2 ; \ldots ; b_k$

Il résulte de cès définitions que :

$c_i b_j = 0$ si $i \neq j$ (I)

et que : $c_i Ab_j = 0$ si $|i - j| > I$ (2)

De là il découle que :

$$b_{k+I} = Ab_k - \alpha_k b_k - \beta_{k-I} b_{k-I}$$

et $c_{k+I} = c_k A - \alpha_k c_k - \beta_{k-I} c_{k-I}$

avec $\alpha_k = \dfrac{c_k Ab_k}{\sigma_k}$ et $\beta_{k-I} = \dfrac{c_{k-I} Ab_k}{\sigma_{k-I}}$

Remarquons encore que :

$$c_k A b_{k-I} = c_{k-I} A b_k = c_k b_k = \sigma_k$$

$$\text{d'ou } \beta_{k-I} = \frac{\sigma_k}{\sigma_{k-I}}$$

On constitue ainsi deux suites de vecteurs respectivements orthogonales Il résulte de cette orthogonalité que les vecteurs b_{n+I} et c_{n+I} respectivement orthogonaux à n vecteurs indépendants sont nuls.

On a alors :

$$Ab_I = b_2 + \alpha_I b_I$$

$$Ab_2 = b_3 + \alpha_2 b_2 + \beta_I b_I$$

$$- - - - - - - - - - - - -$$

$$Ab_k = b_{k+I} + \alpha_k b_k + \beta_{k-I} b_{k-I}$$

$$- - - - - - - - - - - - - - - - -$$

$$Ab_n = \alpha_n b_n + \beta_{n-I} b_{n-I}$$

Ceci montre que dans la base b_I ; b_2 ; ... ; b_n, la matrice A est semblable à la matrice.

$$T = \begin{Vmatrix} \alpha_1 & \beta_1 & & & \\ 1 & \alpha_2 & \beta_2 & & \\ & 1 & \alpha_3 & \ddots & \\ & & \ddots & \ddots & \\ & & & \alpha_n & \beta_{n-1} \\ & & & 1 & \alpha_n \end{Vmatrix}$$

Obtenir le polynôme caractéristiqud d'une telle matrice est simple.

Posons :

$$P_I(\lambda) = I$$

$$P_2(\lambda) = \lambda - \alpha_1$$

$$P_3(\lambda) = (\lambda - \alpha_1) P_2(\lambda) - \beta_I P_I(\lambda)$$

$$- -$$

$$P_{k+I}(\lambda) = (\lambda - \alpha_k) P_k(\lambda) - \beta_{k-I} P_{k-1}(\lambda)$$

c'est à dire, encore :

$$\frac{P_{k+I}(\lambda)}{P_k(\lambda)} = \lambda - \alpha_k - \cfrac{\beta_{k-I}}{\lambda - \alpha_{k-I} - \cfrac{\beta_{k-2}}{\lambda - \alpha_{k-2} - \cfrac{\beta_{k-3}}{\ddots \; - \cfrac{\beta_I}{\lambda - \alpha_I}}}}$$

et $P_{n+I}(\lambda) = (\lambda - \alpha_n)\ P_n(\lambda) - \beta_{n-I}\ P_{n-I}(\lambda)$

Alors : $P_{k+I}(\lambda)$ est le polynôme caractéristique de la matrice obtenue en rayant les (n-k) dernières lignes et colonnes de $\|T\|$.

$P_{n+I}(\lambda)$ est donc le polynôme caractéristique de $\|T\|$, c'est à dire celui de A.

3°- Obtention des vecteurs et lignes propres

Supposons que nous ayons calculé les racines $\lambda_I ; \lambda_2 ; \ldots ; \lambda_n$ du polynôme $P_{n+I}(\lambda)$

Soit u_i (i = I ; 2 ; ... n) les vecteurs propres de la matrice A et v_i ses lignes propres.

Nous pouvons écrire :

$$b_I = \sum_{j=I}^{n} a_j u_j$$

d'ou nous déduisons que :

$$b_i = \sum_{j=I}^{n} a_j P_i(\lambda_j) u_j$$

En explicitant ainsi tous les vecteurs b_i nous obtenons un système linéaire.

En résolvant celui-ci nous trouverons :

$$u_i = \sum_{j=I}^{n} a_{ij} b_j$$

Nous pourrons alors écrire :

$$c_k u_i = \sum_{j=i}^{n} a_{ij} c_k b_j = a_{ik} c_k b_k \quad ;$$

d'ou nous tirons $a_{ik} = \frac{c_k u_i}{c_k b_k}$

or : $c_k = \sum_{j=I}^{n} a'_j P_k(\lambda_j) v_j$

d'ou $c_k u_i = P_k(\lambda_i) v_i u_i a'_i$

et $a_{ik} = \frac{P_k(\lambda_i) v_i u_i a'_i}{c_k b_k}$

v_i et u_i étant définis à une homotéthie prés nous pouvons supposer $a_{ik} = \frac{P_k(\lambda_i)}{\sigma_k}$

c'est à dire :

$$u_i = \sum_{k=I}^{n} \frac{P_k(\lambda_i)}{\sigma_k} b_k$$

et de la même façon, nous aurons

$$v_i = \sum_{k=I}^{n} \frac{P_k(\lambda_i)}{\sigma_k} c_k$$

Ce qui permet de déterminer aisement les vecteurs et lignes propres.

Chapitre II

Etude dds Itérés d'un vecteur dans le cas de diviseurs élémentaires nôn linéaires

J'ai utilisé dans ce chapitre les notations employées dans [I] (chapitres II et III) et dans [II] (§ 22 et 23).

En particulier j'appelle sous-espace invariant d'une matrice un sous-espace qui est transformé globalement en lui-même par la transformation linéaire définie par la matrice. A chaque valeur propre de la matrice peut-être attaché un sous-espace invariant. Si à cette valeur propre correspond un diviseur élémentaire non linéaire d'ordre P on définit une suite de vecteurs invariants.

$x^I, x^2, \ldots, x^P$ tels que : $Ax^I = \lambda x^I$

$Ax^2 = \lambda x^2 + x^I$

$- - - - - - - -$

$Ax^i = \lambda x^i + x^{i-I}$

x^I est un vecteur propre et est unique, par contre x^i peut être remplacé par $Z^i = \sum_{j=I}^{j=i} a_i^j x^j$ avec $a_i^i \neq 0$, Une suite de vecteurs invariants : $Z^I, Z^2, \ldots, Z^P$ est dite générée par Z^P et Z^P est appelé générateur de la suite.

Rappelons également que si la nullité de Λ est zéro, l'espace peut-être décomposé suivant les sous-espaces invariants attachés aux différentes valeurs propres de la matrice et que dans chaque espace invariant une suite de vecteurs invariants constitue une base dudit sous-espace.

Ceci étant j'appellerai sous-espace de profondeur i attachée à la valeur propre λ l'ensemble des vecteurs X tels que $(\Lambda - \lambda I)^i X = 0$

Il est clair que cet ensemble de vecteurs constitue bien un sous-espace.

D'autre part ce sous-espace est un sous-espace invariant de la matrice en effet :

$(A - \lambda I)^{i-1} AX = (A - \lambda I)^{i-1} \lambda X$.

donc $(A - \lambda I)^{i} AX = (A - \lambda I)^{i} \lambda X = 0$

Ce sous-espace appartient donc au sou-espace invariant attaché à la valeur propre λ .

Les vecteurs invariants x^j tels que $j < i$ appartiennent au sous-espace de profondeur i.

Profondeur d'un vecteur : Un vecteur appartenant à un sous-espace de profondeur i mais pas au sous-espace de profondeur i+I sera dit de profondeur i. C'est le cas du vecteur invariant x^i (lorsque $i \leqslant p$).

Il est clair qu'un sous-espace de profondeur i contient tous les espaces de profondeurs inférieures.

Profondeur maximale : Si le sous-espace de profondeur K + I est identique au sous-espace de profondeur K, mais que celui-ci soit différent du sous-espace de profondeur K-I, alors K sera dit profondeur maximale des sous-espaces attachée à la valeur propre λ .

Ce sous-espace de profondeur maximale est évidemment confondu avec le sous-espace invariant attaché à la valeur propre λ . Nous en déduisons immédiatement que K=p, ordre du diviseur élémentaire d'ordre le plus élevé attaché à λ .

2° - Etude des vecteurs transformés d'un vecteur par les puissances successives de la matrice.

Nous pouvons choisir comme base de l'espace une base formées de vecteurs invariants de la matrice, il suffit donc d'étudier les transformés d'un vecteur invariant.

Nous poserons $X_j = 0$; $\forall j \leqslant 0$

Et les transformés successifs du vecteur X_j seront :

$$AX_j = \lambda X_j + X_{j-I}$$

$$A^2X_j = A(\lambda X_j + X_{j-I}) = \lambda^2 X_j + 2\lambda X_{j-I} + X_{j-2}$$

- -

$$A^k X_j = \sum_{l=0}^{k} C_k^l \lambda^{k-l} X_{j-l}$$; C_k^l étant le coefficient du binome

Théorème : Considérons les vecteurs : X_i ; AX_i , ... , A^kX_i.

Ces vecteurs sont linéairement indépendants si $k < i$ et linéairement dépendants dans le cas contraire.

En effet, supposons que k étant inférieur à i les vecteurs ne soient pas linéairement indépendants, nous pouvons alors écrire une relation de la forme.

$$a_oX_i + a_IAX_i + ... + a_kA^kX_i = 0$$

or $(A - \lambda I) X_i = X_{i-I}$ et $(A - \lambda_i I)^p X_i = X_{i-p}$

Multiplions la relation précédente par $(A - \lambda I)^{i-I}$

Nous obtenons : $a_oX_i = 0$ soit $a_o = 0$

Multiplions là ensuite par $(A - \lambda I)^{i-2}$ nous obtenons

$a_oX_2 + a_IAX_2 = 0$, comme $a_o = 0$, $a_IAX_2 = 0$ et $a_I = 0$

Multiplions ensuite par $(A - \lambda I)^{i-3}$, etc... jusqu'à $(A - \lambda I)^{i-k-I}$

Nous montrerons ainsi que a_o , a_1 , ... , a_k sont tous nuls ce qui prouve que les vecteurs X_i , AX_i ; .... , A^kX_i sont bien linéairement indépendants.

D'autre part nous avons :

$(A - \lambda I)^i X_i = 0$ or $(A - \lambda I)^i$ est un polynôme de degré i en A, dont les coefficients ne sont pas tous nuls, ce qui nous donne une relation de dépendance linéaire entre les vecteurs :
X_i , AX_i , ... , A^iX_i
Le théorème est donc démontré.

Théorème : Soit une matrice A ayant un polynôme minimal de degré m et un vecteur V ayant pour chaque valeur propre de la matrice une composante non nulle suivant un vecteur de profondeur maximale associé à cette valeur propre ; alors les vecteurs :
V ; AV ; A^2V ; ... ; A^kV seront linéairement indépendants si $k < m$ et linéairement dépendants dans le cas contraire.

En effet soit $V = \sum_{s=1}^{r} V_s$ composante sur le $s^{ième}$ espace propre ; une relation de dépendance linéaire s'écrit $p(A)\ V = 0$; $p(A)$ étant un polynôme en A ; projetons cette relation sur le $s^{ième}$ espace propre elle s'écrit : $p(A)\ V_s = 0$

Or si le $s^{ième}$ espace propre à une profondeur p_s, nous avons $(A - \lambda_s I)^{ps} V_s = 0$, qui est la relation linéaire de plus faible degré sur le $s^{ième}$ espace propre. Donc $p(A)$ contient $(A - \lambda_s I)^{ps}$ en facteur. En faisant ce raisonnement pour chaque espace propre on voit que $p(A)$ est divisible par

$\prod_{i=1}^{r} (A - \lambda_i I)^{pi}$ qui est le polynôme minimal de la matrice ce qui démontre le théroème.

Remarque : Supposons que dans un sous-espace V n'ait pas de composante suivant un vecteur de profondeur maximale, (de profondeur p_h) mais seulement suivant un vecteur de profondeur $p_h - f_h$.
Dans ce sous-espace la relation de dépendance linéaire sera alors seulement de degré : $p_h - f_h$ et non plus de degré p_h ; et la suite des vecteurs V , AV , ... , A^kV sera formée de vecteurs linéairement dépendants à partir de $k = m - f_h$.

Réciproquement : si la suite des vecteurs V ; AV ; ... ; $A^k V$ est formée de vecteurs linéairement dépendants à partir de k = m-f alors dans au moins un sous-espace, V n'a pas de composante suivant un vecteur de profondeur maximale. De plus si f_s représente dans le $s^{ième}$ espace propre la différence entre la profondeur de V et la profondeur maximale de l'espace propre ; on a la relation

$$f = \sum_{s=1}^{r} f_s$$

Chapitre III

Etude de la formation des vecteurs successifs de la méthode de Lanczos

I° - Retour sur la définition des vecteurs b_i et c_i de la méthode de Lanczos

Définition : Soit un sous-espace E_{ck} (resp. E_{1k}) appartenant à E_c (resp E_1)

On appelle sous-espace perpendiculaire à E_{ck} (resp E_{1k}) un sous-espace de E_1 (resp E_c) formé de tous les vecteurs de E_1 (resp E_c) orthogonaux à une base de E_{ck} (resp E_{1k})

Ce sous-espace perpendiculaire à comme dimension n - k.

Nous le désignerons par E^*_{ck} (resp E^*_{1k}) et nous écrirons $E_{ck} \perp E^*_{ck}$

Nous noterons : dimension de E_k = k de la façon suivante : $D(E_k) = k$.

Posons $J_k = \{b_1 ; \ldots ; b_k\} = \{b_1 ; Ab_1 ; \ldots ; A^{k-1}b_1\}$

$I_k = \{c_1 ; \ldots ; c_k\}$

b_{k+1} est alors défini de la façon suivante :

$b_{k+1} \in J_{k+1}$ et $b_{k+1} \in I^*_k$

donc $b_{k+1} \in (J_{k+1} \cap I^*_k)$

Cette intersection est au moins de dimension un, pour qu'elle ne soit pas de dimension supérieure il faut et il suffit que $J_{k+1} \cdot I^*_k = E_c$.

Supposons que $\sigma_j \neq 0$; $j \leq k$; condition qui est nécessaire pour que l'algorithme se déroule normalement

Nous appellerons $\overline{E}$ le complémentaire d'un espace E

Supposons maintenant que $(\overline{J_k \cdot I^*_k}) \neq \emptyset$. Alors il existe certainement un vecteur $v_0 \in E_1$; $v_0 \neq 0$ et perpendiculaire à $J_k \cdot I^*_k$.

$V_o \notin \overline{I_k}$ puisque $V_o \perp I_k^*$ donc $V_o \in I_k$

Nous pouvons donc écrire $V_o = \sum_{i=I}^{k} a_i c_i$

or V_o étant perpendiculaire à J_k les produits scalaires $V_o b_I$; $V_o b_2 \ldots\ldots$; $V_o b_k$ sont tous nuls.

Ce qui peut encore s'écrire :

$$V_o \cdot b_j = \sum_{i=I}^{k} a_i c_i b_j = a_j \sigma_j = 0$$

et ceci pour tout j donc $a_j = 0$; $\forall j$

Et $V_o = 0$ ce qui contredit l'hypothèse. Donc $\overline{J_k \cdot I_k^*} = \emptyset$; l'intersection est donc bien de dimension égale à I et b_{k+I} est déterminé de façon unique.

2°- Expression des quantités σ_k, α_k, β_k, sous forme de déterminants

Considérons la matrice

$$M_{p+I} = \begin{Vmatrix} c_I b_I & c_I A b_I & \ldots\ldots & c_I A^p b_I \\ c_I A b & & & \\ \vdots & & & \vdots \\ c_I A^p b_I & \ldots\ldots & & c_I A^{2p} b_I \end{Vmatrix}$$

Appelons D_{p+I} son déterminant

En effectuant des combinaisons de lignes nous écrirons :

$$D_{p+I} = \begin{vmatrix} c_I b_I & c_I A b_I & \ldots\ldots & c_I A^p b_I \\ c_I b_2 & c_I A b_2 & \ldots\ldots & c_I A^p b_2 \\ - & - & - & - \\ c_I b_{p+I} & c_I A b_{p+I} & \ldots\ldots & c_I A^p b_{p+I} \end{vmatrix}$$

En effectuant alors des combinaisons de colonnes il viendra finalement :

$$D_{p+I} = \begin{vmatrix} c_I b_I & c_2 b_I & \cdots\cdots & c_{p+I} b_I \\ c_I b_2 & c_2 b_2 & \cdots\cdots & c_{p+I} b_2 \\ & & & \\ c_I b_{p+I} & \cdots\cdots & & c_{p+I}\, b_{p+I} \end{vmatrix} = \prod_{i=I}^{p+I} \sigma_i$$

Nous en déduisons immédiatement :

$$\boxed{\sigma_{p+I} = \frac{D_{p+I}}{D_p}}$$

Et comme $\beta_p = \frac{\sigma_{p+I}}{\sigma_p}$

$$\boxed{\beta_{p+I} = \frac{D_{p+I}\, D_{p-I}}{D_p^2}}$$

Considérons maintenant la matrice

$$N_p = \left\| \begin{vmatrix} c_I A b_I & \cdots\cdots & c_I A^p b_I \\ \vdots & & \vdots \\ c_I A^p b_I & \cdots\cdots & c_I A^{2p}\, b_I \end{vmatrix} \right\|$$

Nous pouvons l'écrire :

$$N_p = \left\| \begin{vmatrix} c_I A \\ c_I A^2 \\ \vdots \\ c_I A^p \end{vmatrix} \right\| \quad x \quad \left\| \begin{vmatrix} b_I & Ab_I & \cdots\cdots & A^{p-I} b_I \end{vmatrix} \right\|$$

Nous pouvons effectuer des combinaisons de lignes dans la première matrice du produit pour faire apparaitre : c_2 , c_3 ,....., c_p , et de même des combinaisons de colonnes dans la seconde feront apparaître : b_2 , b_3 , ... , b_p

N_p s'écrira alors :

$$N_p = T \left\| \begin{matrix} c_1 A \\ c_2 A \\ \vdots \\ c_p A \end{matrix} \right\| \times \left\| b_1 \; b_2 \; \ldots\ldots\ldots \; b_p \right\| T^t \; ;$$

T étant une matrice triangulaire unitaire.

On voit alors que :

$$N_p = T \left\| \begin{matrix} \alpha_1 c_1 b_1 & c_2 b_2 & & \\ \beta_1 c_1 b_1 & \alpha_2 c_2 b_2 & c_3 b_3 & \\ & & \ddots & \ddots \\ & & & c_p b_p \\ & & \beta_{p-1} c_{p-1} b_{p-1} & \alpha_p c_p b_p \end{matrix} \right\| T^t$$

Posons Ω_p = Dét (N_p)

Alors :

$$\Omega_p = D_p \begin{vmatrix} \alpha_1 & 1 & & \\ \beta_1 & \alpha_2 & \ddots & \\ & \ddots & \ddots & 1 \\ & & \beta_{p-1} & \alpha_p \end{vmatrix}$$

Ou encore :

$$\frac{\varrho_p}{D_p} = \begin{vmatrix} \alpha_1 & 1 & & & \\ \beta_1 & \ddots & \ddots & & \\ & \ddots & \ddots & 1 & \\ & & \ddots & \ddots & \\ & & & \beta_{p-1} & \alpha_p \end{vmatrix}$$

Nous reconnaissons un déterminant partiel de la matrice tridiagonale et nous aurons alors

$$\frac{\varrho_p}{D_p} = \alpha_p \frac{\varrho_{p-1}}{D_{p-1}} - \beta_{p-1} \frac{\varrho_{p-2}}{D_{p-2}}$$

or $\beta_{p-1} = \dfrac{D_p D_{p-2}}{D_{p-1}^2}$

Donc :

$$\alpha_p \frac{\varrho_{p-1}}{D_{p-1}} = \frac{\varrho_p}{D_p} + \frac{D_p D_{p-2}}{D_{p-1}^2} \cdot \frac{\varrho_{p-2}}{D_{p-2}}$$

et :

$$\alpha_p = \frac{\varrho_p}{D_p} \frac{D_{p-1}}{\varrho_{p-1}} - \frac{D_p \varrho_{p-2}}{D_{p-1}^2} \frac{D_{p-1}}{\varrho_{p-1}}$$

soit encore :

$$\boxed{\alpha_p = \frac{\varrho_p D_{p-1}}{\varrho_{p-1} D_p} - \frac{D_p \varrho_{p-2}}{D_{p-1} \varrho_{p-1}}}$$

Nous allons maintenant étudier les "accidents" qui peuvent se produire dans la méthode de Lanczos.

Normalement l'algorithme doit se poursuivre jusqu'à l'ordre du polynôme minimal. Pour qu'il s'arrête avant, trois cas peuvent se produire :

A un ordre K, inférieur à n (ordre du polynôme minimal)

a) $\sigma_{K+I} = 0$; $b_{K+I} \neq 0$; $c_{K+I} \neq 0$

b) b_{K+I} (ou c_{K+I}) est nul mais c_{K+I} (ou b_{K+I}) ne l'est pas

c) $b_{K+I} = c_{K+I} = 0$

3°- Etude du cas $\sigma_{K+I} = 0$ avec b_{K+I} et $c_{K+I} \neq 0$

b_{K+I} et c_{K+I} étant tous les deux différents de zéros et σ_{K+I} étant nul cela revient à dire que b_{K+I} est perpendiculaire à c_{K+I}

Dans ce qui suit nous supposerons que b_I et c_I ont des composantes suivants tous les vecteurs de profondeurs maximales respectivement à droite et à gauche de la matrice. Lorsque nous parlerons de vecteurs quelconques cela signifiera à cette restriction près.

Sans restreindre la généralité du raisonnement nous pouvons supposer que à chaque valeur propre n'est associé qu'un seul diviseur élémentaire (celui d'ordre le plus élevé, bien entendu). Nous pouvons, en effet, faire cette supposition puisque la formation de la suite des vecteurs s'arrête à l'ordre du polynôme minimal et que seul le diviseur élémentaire d'ordre le plus élevé intervient dans celui-ci.

Par un choix judicieux de la base, nous pouvons toujours supposer que : $b_I = \sum_{h=I}^{r} X_h^{p_h}$; en appelant p_h la profondeur du sous-espace associé à la $h^{ième}$ valeur propre λ_h ; le nombre de valeurs propres distinctes étant r.

Nous savons que :

$$A^t b_I = \sum_{h=I}^{r} A^t X_h^{p_h}$$

$$= \sum_{h=I}^{r} \sum_{l=0}^{t} C_t^l \lambda_h^{t-l} X_h^{p_h-l}$$

Pour définir le vecteur $c_{/}$, nous opérerons de façon différente, nous nous fixerons une base de E_{lm}

$$Y_I^I , \dots , Y_I^{PI} , Y_2^I , \dots , Y_2^{P2} , \dots , Y_r^{Pr} ;$$

Et nous écrirons alors :

$$c_I = \sum_{h=I}^{r} \sum_{i=I}^{i=Ph} a_{hi} \; Y_h^I$$

Effectuons maintenant le produit scalaire :

$$\langle c_{I,} A^t \, b_I \rangle = c_{/} \, A^{\dagger} \, b_{/}$$

$$= \sum_{h=I}^{r} \sum_{l=0}^{t} \theta_t^l \; {}^{t-l}_{h} A_{h,Ph-l}$$

Considérons maintenant le déterminant D_{k+I} (que nous avons défini au paragraphe précédent) il est composé de termes de la forme : $c_I \, A^t \, b_I$

Le développement de ce déterminant sera de la forme

$$\sum_{(\text{permutations})} (-I)^q \prod \left[c_t \, \lambda_h^{t-l} \; a_{h,Ph-l} \right.$$

Le produit comprenant les différents facteurs de la permutation considérée, dont q est la puissance.

Ce qui nous importe ici c'est que si nous considérons les $a_{h,i}$ (composantes de c_I) comme des inconnues, le développement du déterminant est un polynôme homogène de degré k+I.

Si ce polynôme homogène n'est pas identiquement nul ; nous voyons qu'ayant fixé arbitrairement toutes les variables sauf une, nous obtenons un polynôme de degré k+I (au maximum) en la dernière variable et il existe seulement k+I valeurs au maximum qui annulent ce polynôme. Il résulte de cela que si le polynôme homogène n'est pas identiquement nul, il existe une infinité de valeurs de c_I telles que D_{k+I} soit non nul. Or, si D_{k+I} est

différent de zéro, il en est de même de Γ_{k+1}.

Démontrons donc que le polynôme homogène n'est pas identiquement nul. Pour cela nous allons montrer qu'il existe un choix de valeurs telles qu'il n'est pas nul.

Le vecteur b_I étant toujours défini comme ci-dessus nous allons choisir c_I tel que :

$$(I) \quad \left| \begin{array}{l} c_I\, b_I = 0 \\ c_I\, Ab_I = 0 \\ \vdots \\ c_I\, A^{k-1}\, b_I = 0 \\ \text{mais} \\ c_1\, A^{k}\, b_1 = c \neq 0 \end{array} \right.$$

C_1 a n composantes, or nous avons seulement k+1 équations nous allons donc imposer comme condition supplémentaire que seules les k+1 premières composantes de c_I soient non nulles. Soit : c_I^1 ; c_I^2 ; ... ; c_I^{k+1} ces k+1 composantes.

D'autre part nous avons :

$$b_I = \left\| \begin{array}{c} 1 \\ 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \\ 1 \\ \vdots \\ 0 \\ \vdots \end{array} \right\|$$

Les composantes égales à I correspondent aux vecteurs de profondeurs maximales, les composantes nulles aux autres vecteurs.

Nous aurons alors

$$Ab_I = \left\| \begin{matrix} \lambda_I \\ I \\ 0 \\ \vdots \\ \lambda_2 \\ I \\ 0 \\ \vdots \end{matrix} \right\| ; \quad A^2 b_I = \left\| \begin{matrix} \lambda_I^2 \\ 2\lambda_I \\ I \\ 0 \\ \vdots \\ 0_2 \\ \lambda_2^2 \\ 2\lambda_2 \\ I \\ 0 \\ \vdots \end{matrix} \right\| ; \quad \ldots ; \quad A^i b_I = \left\| \begin{matrix} \lambda_I^i \\ c_I^i \lambda_I^{i-I} \\ \vdots \\ I \\ 0_i \\ \lambda_2 \\ \vdots \end{matrix} \right\|$$

Ecrivons maintenant les équations obtenues en effectuant les produits scalaires indiqués :

$$\left| \begin{array}{l} c_I^I + 0 + 0 + \ldots + 0 + c_I^i + 0 + \ldots + 0 + c_I^j + 0 + \ldots\ldots = 0 \\ c_I^I \lambda_I + c_I^2 + 0 + \ldots + 0 + c_I^i \lambda_2 + c_I^{i+I} + 0 + \ldots \\ \qquad\qquad + 0 \ldots + c_I^j \lambda_3 + c_I^{j+I} + 0 + \ldots = 0 \\ c_1^1 \lambda_1^2 + c_1^2 \lambda_1 + c_1^3 + 0 + \ldots\ldots = 0 \\ - - - - - - - - - - - - - - - - \\ - - - - - - - - - - - - - - - - \\ c_1^d \lambda_1^n + - - - - - - - - - - - - = c \neq 0 \end{array} \right.$$

Le déterminant de ce système à la forme suivante :

$$\begin{vmatrix} I & 0 & 0 & \dots & I & 0 & \dots \\ \lambda_I & I & 0 & & \lambda_2 & I & \\ \lambda_I^2 & 2\lambda_I & I & & \lambda_2^2 & 2\lambda_2 & \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \\ \lambda_I^k & K\lambda_I^{k-I} & & & \lambda_2^k & K\lambda_2^{k-I} & \end{vmatrix}$$

La deuxième colonne est la dérivée de la première par rapport à λ_I , la troisième la dérivée de la deuxième , etc...

Pour que le système ait une solution il faut et il suffit que le déterminant ci-dessus soit différent de zéro. Il est évident que les différentes colonnes de ce déterminant sont linéairement indépendantes, donc le déterminant est non nul. Ceci nous montre alors que le vecteur c_I satisfaisant le système (I) existe et est unique.

Le déterminant D_{K+I} correspondant à ce vecteur c_I et au vecteur b_I que nous considérons depuis le début de ce paragraphe est triangulaire inférieur. Ses termes diagonaux sont égaux et valent justement $c_I A^k b_I$; donc ici $D_{k+I} = (c_I A^k b_k)^{k+I}$ qui est différent de zéro, par définition.

Ceci nous montre qu'il existe des vecteurs c_I tels que D_{K+I} soit non nul donc le polynôme homogène obtenu en développant D_{K+I} n'est pas identiquement nul ce que nous voulions démontrer.

Nous pouvons effectuer ce raisonnement pour chaque valeur de K inférieure à n, et nous pourrons ainsi énoncé le théorème :

Thhéorème : Etant donné un vecteur b_i ayant dès composantes suivant un vecteur de profondeur maximale correspondant à chacune des valeurs propres. Il existe une infinité de vecteurs c_I tels que aucun des nombres σ_{K+I} ne soit nul pour $K < n$, ordre du polynôme minimal

Autrement dit : ce ne sera que dans des cas particuliers que un vecteur b_K sera perpendiculaire au vecteur c_K correspondant

5°- $\sigma_{K+I}= 0$ mais b_{K+I} ou c_{K+I} nul

Nous allons montrer que si b_I et c_I ont des composantes selon les vecteurs les plus profonds de leurs espaces respectifs, alors il est impossible que b_{K+I} ou c_{K+I} soit nul, si K < n ordre du polynôme minimal.

Dans le premier paragraphe de ce chapitre nous avons donné une définition géométrique des vecteurs b_i et c_i . Reprenons les notations de ce paragraphe : nous y avons montré que :

$$J_K \cdot I_K^* = E_c$$

Si b_I a des composantes suivant tous les vecteurs les plus profond alors :

$$A^K b_I \not\subset J_K \qquad \text{d'ou} \qquad Ab_K \not\subset J_K$$

mais alors :

$J_{K+I} = J_K \cdot Ab_K$ a une intersection non nulle avec I_K^* or :

$$J_{K+I} \cap I_K^* = b_{K+I} \qquad \text{ce qui montre que } b_{K+I} \neq 0$$

Par conséquent pour que b_{K+I} soit nul il faut que b_I n'ait pas de composantes suivant tous les vecteurs les plus profonds de son espace.

A fortiori pour que b_{K+I} et c_{K+I} soient nuls il faut que pour chacun d'eux, la condition ci-dessus soit remplie.

Finalement nous avons deux sortes d'accidents possibles :

a) b_I ou c_I n'ont pas de composantes suivants tous les vecteurs les plus profonds.

b) b_i et c_i ayant des composantes suivants tous les vecteurs les plus profonds, σ_{K+I} est nul pour $K < m$. Dans ce dernier cas b_{K+I} et c_{K+I} sont non nuls. Mais alors nous avons vu qu'il suffisait de modifier un seul des vecteurs de départ pour faire disparaître l'accident en question.

Deuxième Partie

Etude des erreurs sur un pas de la Méthode de Lanczos

Chapitre I

Erreurs dans un produit scalaire en point décimal flottant

Nous allons maintenant aborder l'étude des erreurs dans la méthode de Lanczos. Cette méthode est théoriquement rigoureuse, mais, du fait du grand nombre d'opérations mises en jeu, on obtient des erreurs de calculs fort importantes.

Ce sont ces erreurs que nous allons étudier.

Les erreurs de calculs dépendant du mode de représentation des nombres, nous supposerons dans ce qui suit que ceux-ci sont écrits en point décimal flottant, c'est-à-dire avec une mantisse de longueur fixe, donnant les chiffres significatifs et un exposant indiquant l'emplacement de la virgule.

Cette représentation est celle qui est le plus souvent utilisée en machine, de plus dans la méthode de Lanczos l'ordre de grandeur des nombres varie dans de larges proportions au cours du calcul et alors le point décimal flottant est à peu près la seule représentation convenable.

Dans ce chapitre, nous commencerons par étudiéer les erreurs dans les opérations en point décimal flottant (en abrégé P.D.F.) en général. C'est seulement au chapitre suivant que nous aborderons les erreurs dans la méthode de Lanczos proprement dite.

Erreur dans le produit de 2 nombres

Soit à effectuer : $a_1 \cdot a_2$

En point décimal flottant si a_1 et a_2 ont chacun n chiffres significatifs le produit obtenu en aura n également, ceci entraine

que dans le produit exact les chiffres significatifs ayant les poids les plus faibles ont été tronqués.

Nous représenterons le résultat réellement obtenu par $a_I \cdot a_2(I+S)$, S étant un coefficient d'erreur

Ce coefficient est variable pour chaque produit puisqu'il représente la partie tronquée du nombre.

Nous supposerons qu'on peut sans commettre d'erreurs appréciables, le considérer comme une constante.

Erreur dans un produit de plus de deux nombres.

Soit a effectuer $\prod_{i=I}^{n} a_i$

Nous poserons $a_I\ a_2\ (I+S) = a_{I2}$

nous effectuos ensuit

$$a_{I2} \cdot a_3 \rightarrow a_{I2}a_3\ (I+S) = a_I a_2 a_3 (I+S)^2$$

en appliquant la convention faite ci-dessus

On voit que finalement on obtient

$$(I+S)^{n-I} \prod_{i=I}^{n} a_i \simeq \left[I+ (n-I)\ S\right] \prod_{i=I}^{n} a_i$$

Erreur dans une somme arithmètique.

Nous allons effectuer

$$\sum_{i=I}^{n} a_i \qquad \text{avec } a_i \geqslant 0$$

La sommation sera effectuée dans l'ordre des indices croissants c'est à dire que nous aurons l'orgonigramme suivant

$$\Big(\Big\{\big[(a_I+a_2)+a_3\big] + a_4\Big\}+\ldots\Big)+a_n$$

Faisant la même hypothèse que dans le cas d'un produit, nous écrirons que nous obtenons: $(a_I+a_2)(I+S)$

Cette hypothèse, assez grossiére, repose sur le fait que, si le nombre de chiffres significatifs dans la représentation utilisée est m; dans le cas d'une somme arithmétique, comme dans

celui d'un produit, l'erreur dûe à la troncature, est, pour un nombre a : $KaI0^{-m}$, K etant compris entre 0,I et I0.

Au résultat obtenu, c'est àdire $(a_I+a_2)(I+S)$ nous ajoutons a_3 et nous obtenons alors: $(a_I+a_2)(I+S)^2+a^3(I+S)$

Lorsque nous aurons effectuer les n-I sommes de notre sommation nous obtiendrons, au lieu du résultat réel,l'expression: $(a_I+a_2)(I+S)^{n-I}+a_3(I+S)^{n-2}+\ldots.+a_n(I+S)$

Soit, en tenat compte de ce que S est infiniment petit:

$$\sum_{i=I}^{n} a_i + (n-I)S(a_I+a_2) + (n-2)Sa_3+..+Sa_n$$

Cette expression, pour n assez grand, peut être, sans erreurs importantes remplacée par l'expression approchée :

$$\sum_{i=I}^{n} a_i + S \sum_{i=I}^{n} (n-i+I)\, a_i$$

Une remarque s'impose, ces expressions dépendent des différents termes, et non du résultat global, comme c'était le cas pour le produit. De plus l'erreur sera plus ou moins importante selon l'ordre adopté pour faire la sommation.

Dans le cas ou nous ne disposerions que de la somme totale et non des différents termes nous pourrions avoir une expression approchée de l'erreur en supposant tous les termes de la somme égaux. Plaçons nous dans le cas ou n est grand ; soit S_n la somme ; l'erreur sera alors donnée par l'expression approchée :

$$\frac{sS_n}{n} \sum_{i=I}^{n} n-i+I = \frac{sS_n}{n}\,\frac{n(n+I)}{2} = \frac{s\,S_n\,(n+I)}{2}$$

Erreur dans une somme algébrique :

Lorsque nous effectuons la différence de deux nombres positifs si ces nombres sont très voisins leur différence sera très faible par contre l'erreur de calcul sera du même ordre que celle effectuée sur la somme des 2 nombres.

En effet, posons

$A = a\ 10^{\alpha}$

$B = b.\ 10^{\beta}$ avec $\alpha \geq \beta$ $A > 0$ et $B > 0$

$A+B = (a + b\ 10^{\beta-\alpha})\ 10^{\alpha}$

$A-B = (a - b\ 10^{\beta-\alpha})\ 10^{\alpha}$

l'erreur dans les deux cas est de l'ordre de $k\varepsilon\ 10^{\alpha}$, k étant compris entre 0,1 et 10.

Il serait donc complétement faux d'écrire l'erreur ε (A-B) comme on serait tenté de le faire par analogie avec la somme arithmétique. Nous aurons une expression donnant une valeur voisine de la valeur exacte de l'erreur si nous prenons comme expression de l'erreur celle correspondant à l'erreur sur la somme des valeurs absolues. Cette expression ne sera qu'approchée mais fournira une évaluation commode de l'erreur effectuée.

Erreur dans un produit scalaire

Dans le méthode de Lanczos l'essentiel des opérations peut-être ramené à une suite de produits scalaires. Il importe donc de connaitre l'erreur faite dans une telle opération.
Soit à effectuer le produit du vecteur
$A(a_1, a_2, \ldots, a_n)$ par le vecteur $B\ (b_1, b_2 \text{ ----}, b_n)$.

Ce produit scalaire sera

$$\sum_{i=1}^{n} a_i b_i$$

Cette opération comrend n multiplications suivies d'une sommation algébrique.
Lorsqu'on effectue un produit $a_i b_i$ on obtient au lieu de $a_i b_i$, $(1+S)a_i b_i$, S étant toujours positif car l'erreur était une erreur de troncature a le signe du produit. Nous supposons en effet que les opérations sont faites sous arrondi, comme c'est le cas sur la machine dont nous disposons.

Donc la somme des erreurs faites sur les produits sera

$S \sum_{i=1}^{n} a_i b_i = s \langle A,B \rangle$, en représentant par $\langle A,B \rangle$ le produit scalaire de A par B.

Il s'agit maintenant d'effectuer la sommation.

Dans la plupart des cas il sera mal commode, voir souvent impossible, de considérer séparément les différents termes de la somme.

Nous utiliserons alors la formule approchée que nous avions obtenue en considérant tous les termes égaux.

Ceci nous donne une erreur sur la somme des termes $a_i b_i$ (indépendamment de l'erreur faite sur les produits)

$$S \frac{n+1}{2} \sum_{i=1}^{n} |a_i \cdot b_i| = S \frac{n+1}{2} \sum_{i=1}^{n} |a| . |b| .$$

Appelons $\overline{A}$ le vecteur dont les composantes sont $|a_1|$, $|a_2|$,..., $|a_n|$; et $\overline{B}$ le vecteur ($|b_1|$, $|b_2|$,... $|b_n|$ nous poserons naturellement $|a_i| = \overline{a_i}$ $|b_i| = \overline{b_i}$

Dès lors:

$$\sum_{i=1}^{n} |a_i| |b_i| = \sum_{i=1}^{n} \overline{a_i} \cdot \overline{b_i} = \langle \overline{A}, \overline{B} \rangle = \| \overline{A} \| . \| \overline{B} \| \cos(\overline{A},\overline{B})$$

$$\text{or} \quad \| \overline{A} \| = \sqrt{\sum_{i=1}^{n} |a_i|^2} = \sqrt{\sum_{i=1}^{n} a_i^2} = \| A \|$$

donc ;

$$\sum_{i=1}^{n} |a_i| |b_i| = \| A \| . \| B \| \cos (\overline{A},\overline{B}).$$

La valeur de cos $(\overline{A},\overline{B})$ est en général inconnue, nous la remplacerons par sa valeur la plus probable. Pour cela nous remarquerons que $\overline{A}$ et $\overline{B}$ étant des vecteurs à composantes toutes positives ou nulles, leurs directions coupent l'hypersphère de rayon I en deux points appartenant au premier quadrant (cette notion généralisant la notion identique, en géométrie à 3 dimensions). Ces points de l'hypersphère ayant une probabilité uniforme sur tout le premier quadrant. Une telle probabilité est proportionnelle à l'élément d'aire sphèrique.

En définitive chercher la valeur la plus probable de l'angle de $\widetilde{A}$ et de $\widetilde{B}$ revient chercher celle de l'angle de deux points X et Y du premier quadrant de l'hypersphère de rayon I. Soit α cet angle

$$I - \cos\alpha = \frac{I\,X\,Y\,I^2}{2}$$

soit $x_I , x_2 , \dots , x_n$ les composantes du point X

$y_I , y_2 , \dots , y_n$ celles du point Y.

Posons

$$x_I = \sin\varphi_I \sin\varphi_2 \dots \sin\varphi_{n-I}$$
$$x_2 = \sin\varphi_I \sin\varphi_2 \dots \sin\varphi_{n-2} \cos\varphi_{n-I}$$
$$x_3 = \sin\varphi_I \sin\varphi_2 \dots \sin\varphi_{n-3} \cos\varphi_{n-2}$$
$$\vdots$$
$$x_{n-I} = \sin\varphi_I \cos\varphi_2$$
$$x_n = \cos\varphi_I$$

Et de façon analogue

$$y_I = \sin\Theta_I \sin\Theta_2 \dots \sin\Theta_{n-I}$$
$$- - - - - - - - - - - - - - - - - - -$$
$$y_n = \cos\Theta_I$$

La distance des deux points s'écrit alors :

$$|XY|^2 = 2-2\left\{\cos(\varphi_{n-I}-\Theta_{n-I})\prod_{i=I}^{i=n-2}\sin\varphi_i\sin\Theta_i + \sum_{p=0}^{p=n-3}\cos\varphi_{p+I}\cos\Theta_{p+I}\prod_{i=I}^{i=p}\sin\varphi_i\sin\Theta_i\right.$$

d'où :

$$\cos\alpha = \cos(\varphi_{n-I}-\Theta_{n-I})\prod_{i=I}^{i=n-2}\sin\varphi_i\sin\Theta_i + \sum_{p=0}^{n-3}\cos\varphi_{p+I}\cos\Theta_{p+I}\prod_{i=I}^{p}\sin\varphi_i\sin\Theta_i$$

L'élément de surface de l'hypersphère, qui représente la probabilité élementaire d'un point sur cette hypersphère, est :

$$\sin^{n-2}\varphi_I \cdot \sin^{n-3}\varphi_2 \ldots \sin^2\varphi_{n-3}\sin\varphi_{n-2}\, d\varphi_I d\varphi_2 \ldots d\varphi_{n-I}$$

La surface du quadrant positif de l'hypersphère est égal à la surface de l'hypersphère : $\frac{2\Pi^{n/2}}{\Gamma(\frac{n}{2})}$, divisée par le nombre de quadrants de cette hypersphère ; soit 2^n ; donc la surface d'un quadrant vaut : $\frac{\Pi^{n/2}}{\Gamma(n/2)\,2^{n-I}}$

La probabilité cherchée sera alors ;

$$E(\cos\alpha) = \frac{\Gamma^2(n/2)2^{2(n-I)}}{\Pi^n} \int_0^{\Pi/2} \cdots \int_0^{\Pi/2} [\cos\alpha] \sin^{n-2}\varphi_I \ldots \sin\varphi_{n-2} \sin^{n-2}\theta_I \ldots \sin\theta_{n-2}\, d\varphi_I \ldots d\theta_{n-I}$$

Le calcul de cette intégrale est tout à fait classique et ne présente pas de difficultés, on trouve finalement :

$$E(\cos\alpha) = \frac{n\,\Gamma^2(n/2)}{I,\Gamma^2(\frac{n+I}{2})}$$

Nous allons chercher une expression approchée de $E(\cos\alpha)$, lorsque n est grand, pour cela nous remplacerons les factorielles par leurs expressions par la formule de Stirling soit :

$$E(\cos\alpha) \# \frac{2\,e\,n}{\Pi(n-I)} \left(I - \frac{I}{n-I}\right)^{n-I}$$

or $\left(I - \frac{I}{n-I}\right)^{n-I}$ tend vers e^{-I} lorsque n tend vers l'infini

Donc $E(\cos\alpha) \quad \frac{2\,e\,n}{\Pi(n-I)} \frac{I}{e} \# \frac{2}{\Pi}$

L'erreur cherchée est alors :

$$s\,\frac{n+I}{2}\,\|A\| \cdot \|B\|\; E(\cos\alpha) \# \frac{s\,(n+I)}{2}\,\|A\| \cdot \|B\|\,\frac{2}{\Pi}$$

$$= \frac{s\,(n+I)}{\Pi}\,\|A\| \cdot \|B\|$$

Cette erreur, rappelons-le est celle faite sur la sommation mais, auparavent, nous avions fait une erreur sur les produits deux à deux. L'erreur globale sur un produit scalaire sera alors :

$$\bar{e}\,(A,B) = S\left[(A,B) + \|A\|\,\|B\|\,\frac{n+1}{2}\,E(\cos\alpha)\right]$$

$$\leqslant S\,\|A\|\,.\,\|B\|\left[1 + \frac{n+1}{2}\,E(\cos\alpha)\right]$$

$$\# S\,\|A\|\,.\,\|B\|\left(1 + \frac{n+1}{\pi}\right)$$

$\bar{e}\,(A,B)$ est l'erreur absolue faite sur le produit scalaire (A,B)
Nous allons maintenant chercher l'erreur relative faite sur ce produit scalaire soit :

$$\frac{\bar{e}\,(A,B)}{(A,B)} = S\left[1 + \frac{\|A\|\,.\,\|B\|\,(n+1)\,E(\cos\alpha)}{2\,(A,B)}\right]$$

or $(A,B) = \|A\|\,.\,\|B\|\cos(A,B) = \|A\|\,.\,\|B\|\cos\beta$

en posant β = angle de A et de B

Donc $\dfrac{\bar{e}\,(A,B)}{(A,B)} = S\left[1 + \dfrac{n+1}{2}\,\dfrac{E(\cos\alpha)}{\cos\beta}\right]$

Contrairement à $\cos\alpha$, $\cos\beta$ n'est pas toujours inconnu. Malheureusement, ceci n'est pas le cas général. S'il est inconnu nous chercherons sa valeur moyenne. Pour faire ce calcul nous supposerons $\cos\beta$ indépendant de $\cos\alpha$ (ce qui en toute rigueur est faux). Nous remarquerons qu'une des deux directions peut toujours être considérée comme l'axe des pôles et qu'en fait une seule direction varie d'ou

$$E(\cos\beta) = \frac{2\,\Gamma(n/2)}{2\,\pi^{n/2}} \int_0^{\pi/2} \sin^{n-2}\varphi_1 \cos\varphi_1 \,d\varphi_1 \prod_{i=1}^{n-2} \int_0^{\pi} \sin^{n-i-1}\varphi_1 \,d\varphi_1 \cdot \int_0^{2\pi} d\varphi_{n-1}$$

$$E(\cos\beta) = \frac{2\,\Gamma(n/2)}{(n-1)\sqrt{\pi}\,\Gamma\left(\frac{n-1}{2}\right)}$$

Appelons $\bar{e}_n$ l'erreur relative

$$\bar{e}_n = S\left[I + \frac{n+I}{2} \frac{E(\cos\alpha)}{E(\cos\beta)} \right] = S\left[I + \frac{(n+I)\, n\Gamma\left(\frac{n}{2}\right)}{2\sqrt{\pi}\,\Gamma\left(\frac{n+I}{2}\right)} \right]$$

Comme précédemment nous évaluerons les factorielles par la formule de Stirling d'ou :

$$\frac{\Gamma(n/2)}{\Gamma\left(\frac{n+I}{2}\right)} \# \left(\frac{I-\frac{I}{2}}{\frac{n-I}{2}} \right)^{\frac{n-I}{2}} \sqrt{\frac{2\,e}{n-I}}$$

or $\left(I - \frac{I}{2} \frac{I}{\frac{n-I}{2}} \right)^{\frac{n-I}{2}} \longrightarrow e^{-I/2}$ quand $n \longrightarrow \infty$

Donc $\frac{\Gamma(n/2)}{\Gamma\left(\frac{n+I}{2}\right)} \# \sqrt{\frac{2}{n-I}}$

Donc : $\bar{e}_n \# S\left[I + \frac{(n+I)\, n}{\sqrt{2\pi(n-I)}} \right]$

Et comme nous avons supposé n assez grand nous confondrons $(n+I)$ et $(n-I)$ avec n, ce qui nous donnera :

$$\bar{e}_n \# S\left(I + \frac{n^{3/2}}{\sqrt{2\pi}} \right)$$

Ce qui nous montre que pour n assez grand l'erreur relative, sur un produit scalaire, varie en moyenne comme $n^{3/2}$

Vérification :

Ce dernier resultat, l'erreur relative sur un produit scalaire proportionnelle en moyenne à $\left(I + \frac{n^{3/2}}{\sqrt{2\pi}} \right)$ est suffisament simple pour que nous puissions espérer arriver à le vérifier.

Principe : On génére des couples de points aléatoires sur l'hypersphère, de rayon un, à n dimensions. Ces points ayant une densité derépartition uniforme.

Ces points aléatoires sont considérés comme des extrémités de vecteurs ayant pour origine le centre de l'hypersphère.

On effectue, pour chaque couple de points, le produit des deux vecteurs. Ce produit est effectué d'une part en simple précision d'autre part en double précision ; la différence entre les deux résultats trouvés représente l'erreur faite en simple précision.

Si on effectue un tel calcul pour un grand nombre de couples de point, la moyenne des erreurs trouvées nous fournira la valeur problable de l'erreur dans l'espace considéré.

Nous savons que l'erreur est de la forme $k\ n^{3/2}$.

Effectuons le calcul ci-dessus pour un espace à n_I dimensions, ceci de calculer effectivement k ; connaissant k nous pouvons prévoir la valeur de l'erreur que nous devons avoir dans un espace à n_2 dimensions ; le calcul précédent effectué alors dans l'espace à n_2 dimension nous permettra de vérifier notre théorie.

On peut d'ailleurs, faire mieux et tracer la courbe erreur en fonction de la dimension de l'espace. Cette courbe doit être semblable à la courbe $y = x^{3/2}$

Mise en oeuvre : Le procédé utilisé pour générer des points sur l'hypersphère est celui donné par Mervin E. Muller dans les communications de l'A.C.M. d'avril 1959. Ce procédé est basé sur la propriété suivante : Si les x_i , (i=I, 2, ..., n) sont des nombres aléatoires suivant une loi normale. Alors le point y, de l'espace à n dimensions, défini par $y_I = \frac{x_i}{\left(\sum_{i=I}^{n} x_i^2\right)^{1/2}}, i=I,2, \ldots,n;$ et $y = (y_I , y_2 , \ldots y_n)$, à une probabilité de répartition uniforme sur la surface de l'hypersphère.

Des nombres aléatoires uniformément répaftis sont générés par le programme, ces nombres sont transformés à l'aide d'une table de la fonction de Gouss en une répartition normale. Les coordonnées du point y sont ensuite calculées. Un deuxième point de l'hypersphère est ensuite calculé. Il ne reste plus qu'a effectué le produit en simple et en double précision et à faire la différence de ces deux produits.

Le rapport de cette différence au produit lui-même nous donne l'erreur relative dans le produit scalaire que nous venons d'effectuer.

Le programme écrit effectue ainsi automatiquement I00 produits scalaires successifs, puis imprime la moyenne des erreurs effectuées sur ces I00 produits scalaires.

Le temps de calcul est rigoureusement proportionnel à la dimension de l'espace. Il est malheureusement assez important : le calcul de I00 produits scalaires réclame I0 minutes pour un espace à 5 dimensions. Or ainsi que nous allons le voir la dispersion des résultats oblige à effectuer plusieurs milliers de produits scalaires pour obtenir une moyenne significative.

Si nous effectuons plusieurs séries de produits scalaires, chaque série comprenant I00 produits scalaires, les erreurs moyennes obtenues dans chacune de ces séries constituent une série d'épreuve d'une variable aléatoire. Variable aléatoire, qui du fait du théorème central limite, suit approximativement une loi de Gouss. Nous avons effectué pour chaque point vingt séries d'essais soit 2000 produits scalaires, malheureusement les résultats sont très dispersés.

Ainsi pour un espace à 5 dimensions la moyenne est de $0,4I7\ I0^{-8}$ et l'écart type(entre les 20 valeurs constituées chacune par le résultat d'une série de I00 essais) est de $0,I82\ I0^{-8}$. Pour I5 dimensions nous avons respectivement :

moyenne : $I,797.I0^{-8}$; écart type $I,967.I0^{-8}$. Ceci nous montre le peu de précision des résultats obtenus et la nécessité de prendre des échantillons d'autant plus importants que la dimension de l'espace est élevé. Or le temps de calcul d'un produit scalaire étant proportionnel à la dimension de l'espace, les temps de calcul deviennent rapidement prohibitifs.

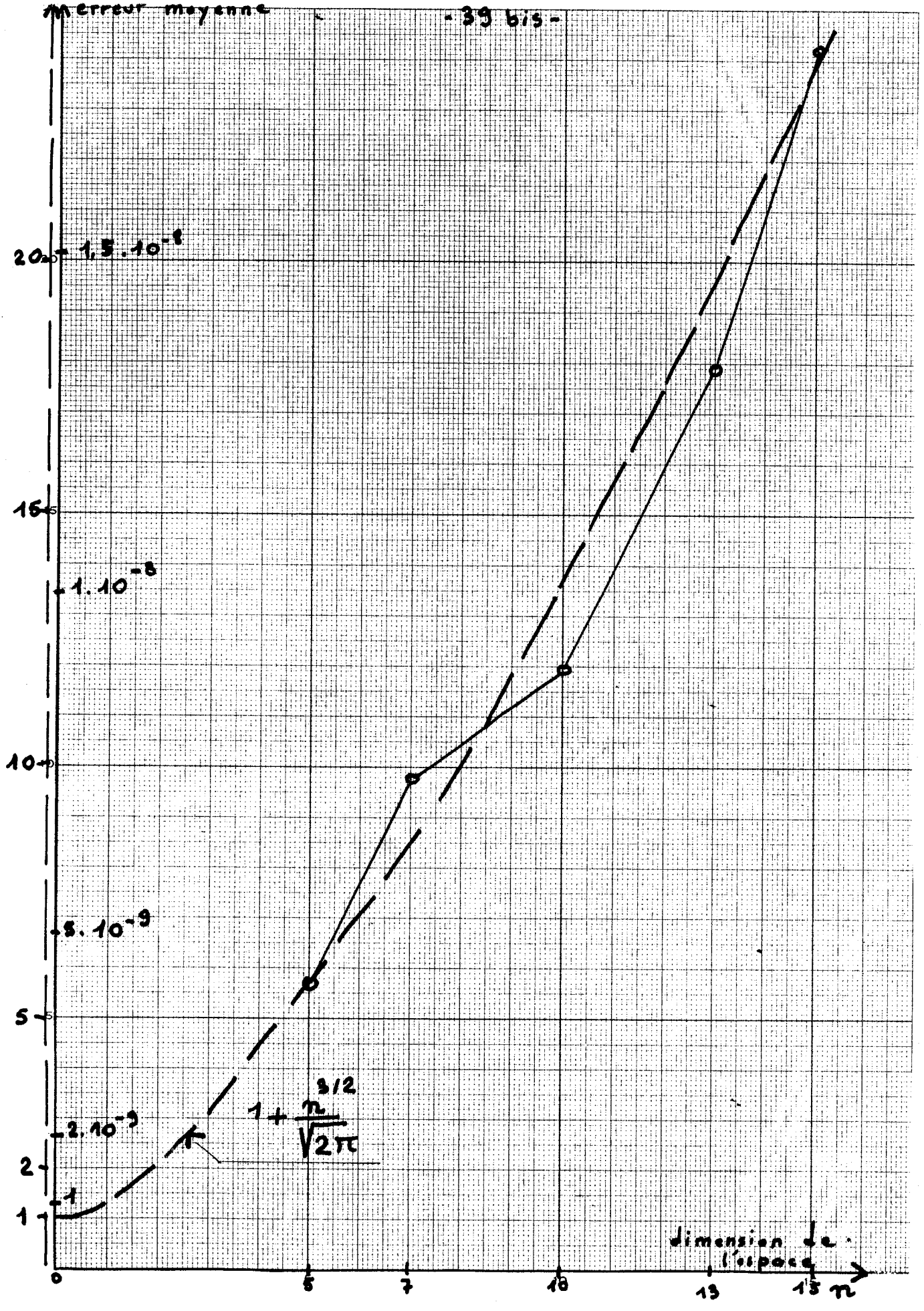
Ces erreurs étant faites, nous avons tracé sur le même graphique la courbe : $y = I + \frac{n^{3/2}}{\sqrt{2\pi}}$; et la courbe des erreurs moyennes, ceci, entre les points 5 et I5.

La courbe d'erreur a été tracée à l'aide des points suivants :

dimension de l'espace	erreur moyenne
5	$0,4I7 \quad I0^{-8}$
7	$0,73I \quad I0^{-8}$
I0	$0,890 \quad I0^{-8}$
I3	$I,328 \quad I0^{-8}$
I5	$I,797 \quad I0^{-8}$

Les échelles des deux courbes ont été choisies de telle sorte que les deux courbes passent, sur le graphique, par le même point pour n = 5.

La coïncidence apparait comme remarquable, ce qui semble bien confirmer, les études précédentes. De plus, nous voyons que la formule obtenue est valable déjà pour des espaces à cinq dimensions, c'est à dire pour un nombre de dimensions assez réduit. Ceci nous permet de voir que le "n assez grand" dont nous avions parlé plus haut est, en fait, assez faible, et de l'ordre de cinq.


erreur moyenne
20
15
10
5
2
1
$1,5.10^{-1}$
1.10^{-8}
5.10^{-9}
2.10^{-9}
1
$1+\frac{n^{3/2}}{\sqrt{2\pi}}$
0
5
7
10
13
15
dimension de l'espace n

Chapitre II

Erreur à un pas de la méthode de Lanczos

1°- Définition d'un pas :

Nous supposons connu b_k , c_k , σ_{k-1}, α_{k-1} , β_{k-2} et les quantités analogues d'indices inférieurs. Nous appellerons alors $k^{ième}$ pas l'ensemble des opérations conduisant à la détermination de σ_k , α_k, β_{k-1} , b_{k+1}, c_{k+1} .

Le premier pas est différent des pas suivants : en effet β_o n'existe pas, nous pouvons le supposer identique en posant

$$\beta_o = 0$$

2°- Invariance de la méthode par un changement de base

Supposons que nous effectuions un changement de base défini par une matrice H.

La matrice A deviendra la matrice $A^* = H^{-1} AH$ le vecteur c_k deviendra $c_k^* = c_k H$ et le vecteur b_k deviendra le vecteur $b_k^* = H^{-1} b_k$

Effectuons alors le produit de $(c_k H)$ par $(H^{-1} b_k)$

$$c_k H H^{-1} b_k = c_k b_k = \sigma_k$$

de même :

$$(c_k H) \cdot (H^{-1} AH) \cdot (H^{-1} b_k) = c_k A b_k$$

donc σ_k, α_k, β_{k-1} vont être invariants.

Dans la nouvelle base nous aurons :

$$\begin{aligned} b_{k+1}^* &= A^* b_k^* - \alpha_k b_k^* - \beta_{k-1} b_{k-1}^* \\ &= (H^{-1} AH)(H^{-1} b_k) - \alpha_k (H^{-1} b_k) - \beta_{k-1} (H^{-1} b_{k-1}) \\ &= H^{-1} b_{k+1} \end{aligned}$$

Ceci nous montre bien l'invariance de la méthode. Cette invariance nous sera utile au chapitre suivant pour étudier la propagation des erreurs.

3°- Hypothèses de calcul

a) Nous supposons que les erreurs sont petites ce qui nous permet de les considérer comme indépendantes.

b) Nous négligerons les erreurs dans les produits simples devant celles dans les produits scalaires.

c) Dans ce chapitre nous étudions les erreurs à un pas. En conséquence nous supposons que tous les calculs effectués antérieurement au $K^{ième}$ pas l'ont été sans erreur.

4°- Recensement des erreurs

Au cours des calculs du $K^{ième}$ pas nous allons d'abord faire des erreurs sur α_K et β_{K-I} ceci aura pour première conséquence d'altérer la matrice tridiagonale.

De plus les erreurs sur α_K et β_{K-I} fausseront b_{K+I} et c_{K+I}.

En plus de ces erreurs sur α_K et β_{K-I}, il y aura les erreurs propres à c_{K+I} et à b_{K+I}, c'est à dire des erreurs faites sur ces vecteurs et ne provenant pas de l'inexactitude de α_K et β_{K-I}

Dans ce qui suit, je supposerai toujours que n est assez grand pour qu'il soit possible d'utiliser pour les produits scalaires les formules simplifiées obtenues à l'aide de la formule de Stirling.

D'autre part soit une quantité a, et a^* la valeur approchée de cette quantité, je désignerai par d (a) l'erreur faite sur cette quantité

Ce qui donnera

$$a^* = a + d(a)$$

5°- Erreurs sur α_K et β_{K-I}

Nous avons deux façons d'envisager l'étude de ces erreurs : ou bien nous utiliserons les erreurs absolues sur les produits scalaires, ou bien les erreurs relatives.

Ces deux procédés ne nous donnerons pas exactement le même résultat. En effet pour passer du stade des erreurs absolues à celui des erreurs relatives nous avons fait une étude statistique supplémentaire ce qui enlève de la précision aux résultats obtenus. Par contre l'utilisation des erreurs absolues pour α_K et β_{K-I} nous conduira à introduire des majorations que nous ne ferions pas si nous utilisions les erreurs relatives. Ajoutons à cela que les formules obtenues à l'aide des erreurs relatives sont plus maniables que celles obtenues en utilisant les erreurs absolues.

Il ne semble pas possible de comparer la précision des résultats obtenus par les deux méthodes. Dans la suite des calculs il sera plus intéressant, suivant le cas, d'utiliser tantôt l'une tantôt l'autre évaluation.

a) Utilisation des erreurs absolues

Etant donné une matrice A, réelle, quelconque, d'éléments a_{ij} ; j'appelle norme de A, la quantité :

$$\left[\sum_{i=I}^{n} \sum_{j=I}^{n} a_{ij}^2 \right]^{I/2} ,$$

Je désignerai cette quantité par N (A).

J'appelle norme de la $i^{ème}$ ligne de A la quantité : $\left[\sum_{j=I}^{n} a_{ij}^2 \right]^{I/2}$ et je la désigne par $\| A_i. \|$

J'ai la relation

$$N(A) = \left[\sum_{i=I}^{n} \| A_{i.} \|^2 \right]^{I/2}$$

Ces définitions étant posées ; considérons les calculs d'un pas de la méthode de Lanczos.

J'ai :

$$\alpha_k = \frac{c_k \, A \, b_k}{\sigma_k} \qquad \text{avec} \quad \sigma_k = c_k \, b_k$$

Pour calculer $c_k \, Ab_k$: je multiplie b_k par A puis j'effectue le produit scalaire du vecteur obtenu par c_k

Donc :

$$d(\alpha_k) = \frac{c_k \, d(Ab_k) + d\left[c_k (Ab_k)\right]}{\sigma_k} - \alpha_k \frac{d(\sigma_k)}{\sigma_k}$$

Considérons Ab_k : Ab_k est un vecteur

La $i^{ème}$ composante de ce vecteur est obtenue par le produit scalaire de la $i^{ème}$ ligne de A par b_k . Appelons $(Ab_k)_i$ cette $i^{ème}$ composante j'ai :

$$d\left[(Ab_k)_i\right] = \|A_{i.}\| \cdot \|b_k\| \, S \, \frac{(\pi + n + I)}{\pi}$$

puisque $\|A_{i.}\|$ est la norme de la $i^{ème}$ ligne de A considéré comme un vecteur.

Soit maintenant le produit scalaire :

$c_k \, d(Ab_k)$. Il est majoré par :

$$\|c_k\| \cdot \|d(Ab_k)\|$$

or
$$\|d(Ab_k)\| = \left[\sum_{i=I}^{n} d\left[(Ab_k)_i\right]^2\right]^{I/2}$$

$$= \left\{\sum_{i=I}^{n} \left[S \, \|A_{i.}\| \, \|b_k\| \, \frac{(\pi + n + I)}{\pi}\right]^2\right\}^{I/2}$$

$$= S \, \|b_k\| \, \frac{(\pi + n + I)}{\pi} \left\{\sum_{i=I}^{n} \|A_{i.}\|^2\right\}^{I/2}$$

$$= S \, \frac{(\pi + n + I)}{\pi} \, \|b_k\| \cdot N(A)$$

Etuidons maintenant d $(c_k\ (Ab_k))$.

Il s'agit d'un produit scalaire et l'erreur sera :

$$S\ \|c_k\| \cdot \|Ab_k\|^{*} \quad \frac{(n+m+1)}{n}$$

Cette erreur est majorée par

$$s\ \frac{(n+m+1)}{n}\ \|c_k\|^{*} \cdot \|b_k\|\ N(A)$$

Enfin

$$d\ (\sigma_k) = S\ \frac{(n+m+1)}{n}\ \|c_k\| \cdot \|b_k\| .$$

En définitive l'erreur sur α_k sera majorée par

$$\boxed{|d\ (\alpha_k)| \leqslant \frac{S\ (n+m+1)}{n\ |\sigma_k|}\ \|c_k\| \cdot \|b_k\| \left[2\ N(A) + |\alpha_k|\right]}$$

Etudions maintenant les erreurs sur β_{k-1} . β_{k-1} a comme expression :

$$\beta_{k-1} = \frac{\sigma_k}{\sigma_{k-1}}$$

Donc

$$d\ (\beta_{k-1}) = \frac{d(\sigma_k)}{\sigma_{k-1}} - \beta_{k-1}\ \frac{d\ (\sigma_{k-1})}{\sigma_{k-1}}$$

Ici se pose une difficulté, théoriquement d (σ_{k-1}) est nul, mais par contre $d\sigma_k$ interviendra dans $d(\beta_k)$ à la même place que $d(\sigma_{k-1})$ dans d (β_{k-1}). C'est pourquoi je raisonnerai pour d (β_{k-1}) comme si d (σ_{k-1}) était différent de zéro. D'ailleurs dans le calcul réel, d (σ_{k-1}) est différent de zéro, et indépendamment de l'erreur transmise par β_{k-2}, fait apparaître une erreur sur β_{k-1} , au $k^{ième}$ pas.

$$d\ (\beta_{k-1}) = \frac{1}{\sigma_{k-1}}\ (d\sigma_k - \beta_{k-1}\ d\sigma_{k-1}).$$

d'ou :

$$|d\beta_{k-1}| \leqslant \left|\frac{1}{\sigma_{k-1}}\right| \frac{3\,(n+n+1)}{n}\left(\|c_k\| \cdot \|b_k\| + |\beta_{k-1}| \cdot \|c_{k-1}\| \cdot \|b_{k-1}\|\right)$$

b) Utilisation des erreurs relatives :

Nous nous plaçons dans les mêmes hypothèses qu'en a). Nous avons alors :

$d\,(Ab_k) = 3\left(1 + \frac{n^{3/2}}{\sqrt{2}\,n}\right) Ab_k$ puisque chaque composante est un produit scalaire.

$c_k \cdot (Ab_k)$ est également un produit scalaire et l'erreur totale sera :

$$\frac{|d\,(c_k\,Ab_k)|}{|c_k\,Ab_k|} = 2\,3\left(1 + \frac{n^{3/2}}{\sqrt{2}\,n}\right)$$

D'autre part

$$\frac{|d\,(\sigma_k)|}{|\sigma_k|} = 3\left(1 + \frac{n^{3/2}}{\sqrt{2}\,n}\right)$$

Donc, comme les erreurs relatives s'ajoutent dans un quotient :

$$\frac{|d\,(\alpha_k)|}{|\alpha_k|} = 3\,3\left(1 + \frac{n^{3/2}}{\sqrt{2}\,n}\right)$$

et

$$\frac{|d\,(\beta_{k-1})|}{|\beta_{k-1}|} = 2\,3\left(1 + \frac{n^{3/2}}{\sqrt{2}\,n}\right)$$

6° - Erreurs sur les vecteurs b_k et c_k

Compte-tenu des hypothèses que nous avons formulées, nous pouvons écrire :

$$d\,(b_{k+1}) = d\,(Ab_k) - d\,(\alpha_k)\,b_k - d\,(\beta_{k-1})\,b_{k-1}$$

Nous voyons qu'en plus des erreurs dûes à α_k et β_{k-1} il y aura le vecteur $d\,(Ab_k)$. Selon l'hypothèse a) ou b) nous aurons pour $d\,(Ab_k)$ les expressions suivantes :

a) $d\,(Ab_k)$ est un vecteur dont les composantes sont :

$$\left[d\,(Ab_k)\right]_i = \|A_{i.}\| \cdot \|b_k\| \; \frac{S\,(n+m+1)}{\eta}$$

b) $$d\,(Ab_k) = S\,\frac{(1+n^{3/2})}{\sqrt{2}\eta}\,Ab_k$$

je vais maintenant chercher quelle est, dans le cas a) la valeur de $\|d\,(b_{k+1})\|$

$$\|d\,(b_{k+1})\| \leq \|d(Ab_k)\| + |d(\alpha_k)|\,\|b_k\| + |d(\beta_{k-1})|\,\|b_{k-1}\|$$

cherchons d'abord

$$\|d\,(Ab_k)\| = N\,(A)\,\|b_k\|\,\frac{S\,(n+m+1)}{\eta}$$

Nous l'avons vu plus haut ,

donc

$$\|d\,(b_{k+1})\| \leq \frac{S(n+m+1)}{\eta}\left\{\frac{1}{|\sigma_k|}\,\|c_k\|\,\|b_k\|^2\left[2N\,(A) + |\alpha_k|\right]\right.$$

$$+ N(A)\,\|b_k\| + \frac{1}{|\sigma_{k-1}|}\left[\|c_k\| \cdot \|b_k\| + |\beta_{k-1}| \cdot \right.$$

$$\left.\left.\|c_{k-1}\| \cdot \|b_{k-1}\|\right]\right\}\|b_{k-1}\|$$

Ou encore en groupant différemment les termes :

$$\|d(b_{k+1})\| < \frac{S(\Pi+n+1)}{\Pi}\left\{\|c_k\|\cdot\|b_k\|\left[\frac{\|b_k\|}{|\sigma_k|}(2N(A)+|\alpha_k|)+\frac{\|b_{k-1}\|}{|\sigma_{k-1}|}\right]\right.$$

$$\left.+\frac{|\beta_{k-1}|}{|\sigma_{k-1}|}\|c_{k-1}\|\cdot\|b_{k-1}\|^2+N(A)\|b_k\|\right\}$$

7°- Importance de σ_k :

Reprenons l'expression $d(\alpha_{k-1}) < \frac{S(\Pi+n+1)}{\Pi}\left[2N(A)+|\alpha_k|\right]$

$$\cdot\frac{\|c_k\|\cdot\|b_k\|}{|\sigma_k|}\,,\ \frac{\|c_k\|\cdot\|b_k\|}{|\sigma_k|}=\frac{1}{\cos(c_k,\ b_k)}$$

cette dernière expression peu varier de I à $+\infty$ alors que $\frac{S(\Pi+n+1)}{\Pi}\left[2N(A)+|\alpha_k|\right]$ varie assez peu , puisqué, seul $|\alpha_k|$ varie.

$\left|\frac{1}{\cos(c_k,\ b_k)}\right|$ intervient également dans $d(\beta_{k-1})$ puisque :

$$|d(\beta_{k-1})| \leqslant \frac{1}{|\sigma_k|}\frac{S(\Pi+n+1)}{\Pi}\|c_{k+1}\|\,\|b_{k+1}\| + \frac{S(\Pi+n+1)}{\Pi}\left|\frac{1}{\cos(c_k,\ k)}\right|$$

Ici son influence est plus faible, mais elle n'en demeure pas moins importante.

Ceci nous montre que dès que l'angle de c_k avec b_k s'approche de $\frac{\Pi}{2}$ les erreurs vont croitre dans des proportions astronomiques il suffit que cos (c_K, b_K) soit divisé par IO pour que l'erreur soit multipliée approximativement par la même quantité.

Calculer à chaque pas cos (c_K , b_K) (et nous verrons ultérieurement que cela est très facile) permettra donc de controler que l'erreur garde une valeur normale. Car il faut remarquer que, si l'angle (c_K, b_K) devient très voisin de $\frac{\Pi}{2}$, ceci est lié au choix des vecteurs de départ. Nous pouvons donc dire que nous avons d'une part une erreur "normale"qu'on ne peut éviter, c'est celle que nous avons étudiée de façon probabiliste dans le chapitre précadent et dans celui-ci, d'autre part une erreur exceptionnelle dûe au mauvais choix de vecteurs de départ ; erreur qui se traduit par un angle (c_K , b_K) voisin de $\frac{\Pi}{2}$. Cette erreur-ci, qui peut devenir beaucoup plus importante que l'erreur normale,peut-être évitée en modifiant le choix des vecteurs de départ.

Remarquons que s'il n'y avait pas d'erreur de calcul $(c_K, b_K) = \frac{\Pi}{2}$ nous donnerait un cas singulier. Avec les erreurs de calcul ce n'est plus seulement la direction perpendiculaire à c_K que b_K ne doit pas avoir, mais c'est dans un certain cône, autour de cette direction, que b_K ne doit pas être.

Chapitre III

Etude de la propagation d'une erreur sur α_k ou β_k

Notations de ce chapitre : Considérons une certaine quantité a nous appelerons a' l'erreur faite sur cette quantité.

I° - Propagation des erreurs sur un pas

Supposons que des erreurs aient été commises au $k^{ième}$ pas et que aux pas suivants les seules erreurs proviennent de la propagation de ces erreurs du $k^{ième}$ pas.

Nous avons :

$$c'_{k+1} = c'_k A - \alpha'_k c_k - \alpha_k c'_k - \beta'_{k-1} c_{k-1} - \beta_{k-1} c'_{k-1} \qquad (1)$$

d'ou nous déduisons que :

$$c'_{k+1} b_{k+P} = c'_k A b_{k+P} - \alpha'_k c_k b_{k+P} - \alpha_k c'_k b_{k+P} - \beta_{k-1} c'_{k-1} b_{k+P} \qquad (2)$$;

ceci pour $P \geq 1$.

remplaçons Ab_{k+P} par : $b_{k+P+1} + \alpha_{k+P} b_{k+P} + \beta_{k+P-1} b_{k+P-1}$

Nous obtenons :

$$c'_{k+1} b_{k+P} = c'_k b_{k+P+1} - \alpha_{k+P} c'_k b_{k+P} - \beta_{k+P-1} c'_k b_{k+P-1} - \alpha'_k c_k b_{k+P} - \beta_{k-1} c'_{k-1} b_{k+P} \qquad (3)$$

L'étude faite dans ce chapitre sera basée sur ces trois formules.

2° - Propagation d'une erreur sur α_s

Supposons qu'aucune erreur n'ait été faite avant le sième pas et que à ce pas la seule erreur faite le soit sur α_s : soit α'_s.

Nous allons commencer par étudier l'influence de cette erreur sur les produits $c_j\ b_i$

Nous avons par hypothèse : $c'_i \quad 0$

donc $c''_s\ b_i = 0$ pour tout i

L'équation nous permet d'écrire

$$c'_{s+1} = -\alpha'_s\ c_s$$

$$c'_{s+1}\ b_i = -\alpha'_s\ c_s\ b_s$$

donc :

$$c'_{s+1}\ b_i = 0 \quad \forall\ i > s$$

L'équation (I) nous donne ensuite :

$$c''_{s+2} = c'_{s+1}\ A - \alpha'_{s+1}\ c_{s+1} - \alpha_{s+1}\ c'_{s+1}$$

$$c''_{s+2}\ b_{s+2+P} = c'_{s+1}\ A\ b_{s+1+P} \cancel{- \alpha'_{s+1}\ c_{s+1}\ b_{s+2+P}} - \alpha_{s+1}\ c'_{s+1}\ b_{s+2+P}$$

Ceci pour tout $P \geqslant 0$

en remplaçant c'_{s+1} par sa valeur soit : $-\alpha'_s\ c_s$

$$c''_{s+2}\ b_{s+2+P} = -\alpha'_s \cancel{c_s\ A b_{s+1+P}} - \alpha'_{s+1}\ \alpha'_s\ \cancel{c_s\ b_{s+2+P}}$$

donc $c'_{s+2}\ b_{s+2+P} = 0 \qquad \forall\ p \geqslant 0$

Supposons maintenant que :

$$c'_j\ b_i = 0 \ ; \ \forall\ j \leqslant k \ ; \ \forall\ i \geqslant j$$

et démontrons que : $c'_{k+1}\ b_{k+p} = 0 \ ; \ \forall\ P \geqslant 1$

(3) nous permet d'écrire :

$$c'_{k+1}\, b_{k+p} = c''_k\, b_{k+p+1} - \alpha_{k+p}\, c'_k\, b_{k+p} + \beta_{k+p-1} c'_k\, b_{k+p-1} - \alpha_k\, c'_k\, b_{k+p} - \beta_{k-1}\, c'_{k-1}\, b_{k+p}$$

en vertu de la supposition faite nous avons :

$$c'_k\, b_{k+p+1} = 0 \; ; \; c'_k\, b_{k+p} = 0 \; ; \; c'_k\, b_{k+p-1} = 0 \; ; \; c'_k\, b_{k+p} = 0 \; ;$$

$$c'_{k-1}\, b_{k+p} = 0$$

puisque $P \geqslant 1$.

donc : $c'_{k+1}\, b_{k+p} = 0$

Ceci nous montre que si le thèorème est vrai jusqu'à l'ordre k, il est vrai jusqu'à l'ordre k+1. Or nous l'avons démontré directement jusqu'à l'ordre 3.

donc : $\boxed{c'_j\, b_i = 0 \quad \forall\, i \geqslant j \; ; \; \forall\, j}$

Par symétrie nous en déduisons que : $c_j\, b'_i = 0 \; ; \forall\, i \; ; \forall\, j \geqslant i$

Il en résulte que $\boxed{\sigma'_j = 0 \quad \forall\, j}$

en effet :

$$\sigma'_j = c'_j\, b_j + c_j\, b''_j = 0$$

de même, comme $\beta_j = \dfrac{\sigma_{j+1}}{\sigma_j}$

$$\boxed{\beta'_j = 0 \; ; \forall\, j}$$

Etudions maintenant les erreurs sur α_{s+p} !

$$\alpha'_{s+1} = \frac{c'_{s+1}\, A\, b_{s+1} + c_{s+1}\, A\, b'_{s+1}}{\sigma_{s+1}} \quad \text{puisque } \sigma'_{s+1} = 0$$

$$\alpha'_{s+1} = -\alpha'_s\, \frac{c_s\, Ab_{s+1} + c_{s+1}\, Ab_s}{\sigma_{s+1}} = -2\alpha'_s\, \frac{\sigma_{s+1}}{\sigma_{s+1}}$$

donc $\boxed{\alpha'_{s+1} = -2\alpha'_s}$

$$\alpha'_{s+2} = \frac{c'_{s+2}\, Ab_{s+2} + c_{s+2}\, Ab'_{s+2}}{\sigma_{s+2}}$$

Etudions le terme $c'_{s+2}\, Ab'_{s+2}$

$$c'_{s+2} Ab_{s+2} = (c'_{s+1} A - \alpha'_{s+1} c_{s+1} - \alpha_{s+1} c'_{s+1})(b_{s+3} + \alpha_{s+2} b_{s+2} + \beta_{s+1} b_{s+1})$$

$$= c'_{s+1} A (b_{s+3} + \alpha_{s+2} b_{s+2} + \beta_{s+1} b_{s+1}) - \alpha'_{s+1} c_{s+1} \beta_{s+1} b_{s+1}$$

$$= c'_{s+1}\Big[b_{s+4} + \alpha_{s+3} b_{s+3} + \beta_{s+2} b_{s+2} + \alpha_{s+2}(b_{s+3} + \alpha_{s+2} b_{s+2} + \beta_{s+1} b_{s+1}) + \beta_{s+1}(b_{s+2} + \alpha_{s+1} b_{s+1} + \beta_s b_s)\Big]$$

$$- \alpha'_{s+1} \beta_{s+1} \sigma_{s+1}$$

$$= c'_{s+1} \beta_{s+1} \beta_s b_s - \alpha'_{s+1} \beta_{s+1} \sigma_{s+1}$$

$$= c'_{s+1} b_s \frac{\sigma_{s+2}}{\sigma_s} - \alpha'_{s+1} \sigma_{s+2}$$

$$= -\alpha'_s c_s b_s \frac{\sigma_{s+2}}{\sigma_s} - \alpha'_{s+1} \sigma_{s+2}$$

$$= -(\alpha'_s + \alpha'_{s+1})\, \sigma_{s+2}$$

par symétrie :

$$c_{s+2}\, Ab_{s+2} = -(\alpha'_s + \alpha'_{s+1})\, \sigma_{s+2}$$

d'ou : $\alpha'_{s+2} = -2(\alpha'_s + \alpha'_{s+1}) = -2(\alpha'_s - 2\alpha'_s) = 2\alpha'_s$

$$\boxed{\alpha'_{s+2} = 2\alpha'_s}$$

Etudions maintenant le produit scalaire :

$$f'_{s+p} = c'_{s+p}\, Ab_{s+p}$$

$$f'_{s+p} = (c'_{s+p-1}\, A - \alpha'_{s+p-1} c_{s+p-1} - \chi_{s+p-1} c'_{s+p-1} - \beta_{s+p-2} c'_{s+p-2}) Ab_{s+p}$$

$$= (c''_{s+p-1}\, A - \alpha_{s+p-1}\, c'_{s+p-1} - \beta_{s+p-2} c'_{s+p-2})(b_{s+p+1} + \gamma_{s+p}\, b_{s+p} + \beta_{s+p-1}\, b_{s+p-1}) - \alpha'_{s+p-1}\, \sigma_{s+p}$$

$$= c'_{s+p-1}\, A(b_{s+p+1} + \alpha_{s+p}\, b_{s+p} + \beta_{s+p-1}\, b_{s+p-1}) - \alpha'_{s+p-1}\, \sigma_{s+p}$$

$$= c'_{s+p-1}\Big[b_{s+p+2} + \alpha_{s+p+1}\, b_{s+p+1} + \beta_{s+p}\, b_{s+p} + \alpha_{s+p}(b_{s+p+1} + \gamma_{s+p}\, b_{s+p} + \beta_{s+p-1}\, b_{s+p-1}) + \beta_{s+p-1}(b_{s+p} + \alpha_{s+p-1}\, b_{s+p-1} + \beta_{s+p-2}\, b_{s+p-2})\Big] - \alpha'_{s+p-1}\, \sigma_{s+p}$$

$$= c'_{s+p-1}\beta_{s+p-1}\beta_{s+p-2}\, b_{s+p-2} - \alpha'_{s+p-1}\, \sigma_{s+p}$$

$$= \frac{\sigma_{s+p}}{\sigma_{s+p-2}}\, c'_{s+p-1}\, b_{s+p-2} - \alpha'_{s+p-1}\, \sigma_{s+p}\,.$$

$$= \sigma_{s+p}\Big[\frac{(c'_{s+p-2}\, A - \alpha'_{s+p-2} \cdot c_{s+p-2} - \chi_{s+p-2} \cdot c'_{s+p-2}}{\sigma_{s+p-2}}$$

$$\frac{- \beta_{s+p-3}\, c'_{s+p-3})\, b_{s+p-2}}{} - \alpha_{s+p-1}\Big]$$

$$= \sigma_{s+p}\Big[\frac{1}{\sigma_{s+p-2}}(c'_{s+p-2}\, Ab_{s+p-2} - \alpha'_{s+p-2}\, \sigma_{s+p-2}) - \alpha'_{s+p-1}\Big]$$

$$= \sigma_{s+p}\Big[\frac{c'_{s+p-2}\, Abs+p-2}{\sigma s+p-2} - \alpha'_{s+p-1} - \alpha'_{s+p-2}\Big]$$

Il en résulte que :

$$\frac{f'_{s+p}}{\sigma s+p} = \frac{f'_{s+p-2}}{\sigma s+p-2} - \alpha'_{s+p-1} - \alpha'_{s+p-2}$$

posons :

$$\frac{f'_{s+p}}{\Gamma_{s+p}} = g'_{s+p}$$

il vient alors :

$$g'_{s+p} = g'_{s+p-2} - \alpha'_{s+p-I} - \alpha'_{s+p-2} \qquad (4)$$

Calculons maintenant :

$$\alpha'_{s+p} = \frac{c'_{s+p}\ Ab_{s+p} + c_{s+p}\ Ab'_{s+p}}{\Gamma_s}$$

$$= \frac{c'_{s+p-2}\ Ab_{s+p-2} + c_{s+p-2}\ Ab'_{s+p-2}}{\Gamma_{s+p-2}} - 2\left[\alpha'_{s+p-I} + \alpha'_{s+p-2}\right]$$

$$= \alpha'_{s+p-2} - 2\left[\alpha'_{s+p-I} + \alpha'_{s+p-2}\right]$$

$$\alpha'_{s+p} = -\left(2\alpha'_{s+p-I} + \alpha'_{s+p-2}\right)$$

Nous avons donc une formule de récurrence qui nous donne α'_{s+p}.

supposons que pour :

$$I \leq j < p \quad . \quad \alpha'_{s+j} = -\alpha'_{s+j-I}$$

alors $\alpha'_{s+p-I} = -\alpha'_{s+p-2}$ d'ou : $\alpha'_{s+p} = \alpha'_{s+p-2} = -\alpha'_{s+p-I}$

Donc le théorème est encore vrai pour p

Or : $\alpha'_{s+I} = -2\,\alpha'_s$

$$\alpha'_{s+2} = 2\alpha'_s = -\alpha'_{s+I}$$

donc en définitive

$$\alpha'_{s+2r} = 2\alpha'_s$$

$$\alpha'_{s+2r-I} = -2\alpha'_s \qquad r > 0$$

3°- Propagation d'une erreur sur β_s

Supposons qu'aucune erreur n'ait été faite avant le $s^{ième}$ pas et qu'à ce pas la seule erreur soit sur β_s.

donc $c'_s = 0 \quad \alpha'_s = 0 \qquad$ et $\beta_s \neq 0$

$$c_{s+I} = c_s A - \alpha_s c_s - \beta_{s-I} c_{s-I}$$

β_s n'intervenant pas nous aurons :

$$c'_{s+I} = 0 \qquad \text{donc } \sigma'_{s+I} = 0 \qquad \text{et} \qquad \alpha'_{s+I} = 0$$

$$c'_{s+2} = - \beta'_s c_s$$

$$\sigma'_{s+2} = c'_{s+2} b_{s+2} + c_{s+2} b'_{s+2} = - \beta'_s (c_s b_{s+2} + c_{s+2} b_s) = 0$$

et puisque $\sigma'_{s+2} = 0$, $\beta'_{s+I} = 0$

$$\alpha'_{s+2} = \frac{c'_{s+2} A b_{s+2}}{\sigma_{s+2}} + \frac{c_{s+2} A b'_{s+2}}{\sigma_{s+2}} = 0$$

donc $\sigma'_{s+2} = \alpha'_{s+2} = 0$;

considérons maintenant c'_{s+3}

$$c'_{s+3} = c'_{s+2} A - \alpha_{s+2} c'_{s+2}$$

$$= - \beta'_s \left[c_s A - \alpha_{s+2} c_s \right]$$

c'_{s+3} est une combinaison linéaire de $c_s A$ et de c_s ; c'_{s+3} s'exprime donc linéairement en fonction de c_s et de c_{s+I}

Supposons maintenant que c'_{s+k} s'exprime linéairement en fonction de c_s à c_{s+k-2} et ceci pour tout k compris entre I et p inclus.

Soit $c'_{s+k} = \sum_{i=I}^{i=k-2} \theta_{s+i} c_{s+i} \qquad \forall\, I \leq k \leq p$

Démontrons que c'_{s+p+I} s'exprime linéairement en fonction de c_s à c_{s+p-I}

$$c'_{s+k}\ b_{s+k} = \left(\sum_{i=I}^{i=k-2} e_{s+i}\ c_{s+i}\right) b_{s+p} = 0$$

$$c'_{s+k}\ A\ b_{s+k} = \left(\sum_{i=I}^{i=k-2} e_{s+i}\ c_{s+i}\right) A b_{s+k} = 0$$

donc $\quad \forall\ I \leq k \leq p \quad \alpha'_{s+k} = 0 \quad$ et $\quad \gamma'_{s+k} = 0$

$$\forall\ I < k \leq p \qquad \beta'_{s+k} = 0$$

On a alors :

$$c'_{s+p+I} = c'_{s+p}\ A - \alpha_{s+p}\ c'_{s+p} - \beta_{s+p-I}\ c'_{s+p-I}$$

$$= \sum_{i=I}^{p-2} e_i\ c_i\ A - \alpha_{s+p} \sum_{i=I}^{p-2} e_i\ c_i - \beta_{p-I} \sum_{i=I}^{p-3} f_i\ c_i$$

$$= \sum_{i=I}^{p-I} h_i\ c_i \qquad \text{c.q.f.d.}$$

Or $c'_{s+I} = 0$

$c'_{s+2} = -\beta'_s\ c_s$, s'exprime linéairement en fonction de c_s .

c'_{s+3} s'exprime linéairement en fonction de c_s et c_{s+I} donc c'_{s+p} s'exprime linéairement en fonction de c_s à c_{s+p-2} pour p quelconque

Mais alors : $\boxed{\alpha'_{s+p} = \beta'_{s+p} = 0 \qquad \forall\ p > 0\ ;\ \alpha'_s = 0}$

Donc une erreur sur β'_s ne donne aucune erreur sur les α et les β ultérieurs.

En ce qui concerne les altérations des coefficients β j'ai obtenu les résultats suivants :

Tout d'abord les coefficients α ne sont absoluments pas modifiés. Quand aux coefficients β, voici ce qu'ils deviennent :

Si je multiplie β_3 par I,00I j'obtiens β_3= 9,787025 ; β_4= 5,34II23 ; β_5= 5,7I6857

Alors que je devais avoir pour β_4 et β_5 les valeurs suivantes : β_4= 5,34III3 ; β_5= 5,7I6838

Donc il y a une erreur parasite sur β_4 et β_5 mais cette erreur est très faible elle vaut, en effet, environ $\beta'_3 \times 10^{-3}$

Si je multiplie β_2 par I,0I j'obtiens :

β_2 = 9,088II9 ; β_3 = 9,778I48 ; β_4 = 5,342839 ; β_5 = 5,722064

I'erreur parasite ici est de l'ordre $\beta'_2 \times 10^{-2}$

Ceci nous montre donc que si les erreurs transmises sur les coefficients α suivent rigoureusement la règle il se produit une erreur parasite sur les coefficients β particulièrement lorsque l'erreur transmise provient d'une erreur initiale sur un coefficient β mais cette erreur parasite est faible, trés faible même pour les erreurs provenant d'une erreur initiale sur un coefficient α .

De plus il faut remarquer que l'erreur de I/I00 faite dans notre dernier essai est d'un ordre très supérieur à celui des erreurs de calculs courantes.

Or ainsi que nous le montre l'essai précédent l'erreur parasite est d'autant plus réduite (en valeur relative par rapport à l'erreur initiale) que l'erreur initiale est faible. Donc on peut dire que la propagation des erreurs sur α et β suit très étroitement la règle trouvée.

4°- Vérifications :

Pour vérifier l'exactitude de ces résultats j'ai utilisé un programme accessoire qui multiplie au choix un des coefficients α ou β par I,0I ou I,00I. N'importe lequel des coefficients α ou β peut ainsi être modifié. j'ai effectué de nombreux essais, tant avec I,0I qu'avec I,00I ; et toujours les erreurs trouvées se sont révelées parfaitementconformes aux formules ci-dessus.

Ainsi pour une ma trice 6 x 6 dont la forme tridiagonale exacte est :

$$\begin{pmatrix} -0,5 & 14,9166 & & & & \\ 1 & -0,5 & 8,9981 37 & & & \\ & 1 & -0,5 & 9,777248 & & \\ & & 1 & -0,5 & 5,341113 & \\ & & & 1 & -0,5 & 5,716333 \\ & & & & 1 & -0,5 \end{pmatrix}$$

J'ai obtenu comme matrice tridiagonale, après multiplication de α_3 par I,0I , la matrice ci-dessous :

$$\begin{pmatrix} -0,5 & 14,9166 66 & & & & \\ 1 & -0,5 & 8,998137 & & & \\ & 1 & -0,505 & 9,777273 & & \\ & & 1 & -0,4900001 & 5,341147 & \\ & & & 1 & -0,5099998 & 5,716889 \\ & & & & 1 & -0,4900005 \end{pmatrix}$$

Si j'appelle e l'erreur prévue par l'étude ci-dessus on voit que sur α l'erreur est environ de $e(1+10^{-5})$

Sur β l'erreur parasite est environ 10 fois plus grande que sur α_3 (soit 2.10^{-4}, e étant l'erreur sur α puisqu'il n'est pas prévu d'erreur sur β)

Donc on peut dire que la formule est parfaitement vérifiée.

Chapitre IV

Propagation d'une erreur sur b_k

Une erreur faite sur b_k va commencer par altérer γ_k et $c_k Ab_k$ donc α_k et β_{k-1}. Ce n'est pas tout : indépendamment des erreurs faites sur α_k et β_{k-1} nous aurons une erreur sur b_{k+1} dûe à ce que b_k est erroné et dans l'expression : $b_{k+1} = (A - \alpha_k I)\, b_k - \beta_{k-1}\, b_{k-1}$, il faudra remplacer b_k par $(b_k + b'_k)$.

En définitive les quantités qui nous intéressent sont les coefficients de la matrice tridiagonale. Or nous connaissons la propagation d'une erreur sur un de ces coefficients.

Nous allons en conséquence procéder de la façon suivante pour faire nôtre étude :

Nous supposons une erreur sur b_k, cette erreur fausse α_k et β_{k-1} de α'_k et β'_{k-1}. Mais une fois calculés α'_k et β'_{k-1} nous ne nous en préoccuperont plus, car nous connaissons leurs propagations b'_k fausse également b_{k+1} de b'_{k+1} puis en chaine tous les b_{k+p}. Ces erreurs sur les b_i donnent des erreurs sur les coefficients de la matrice tridiagonale mais nous ne nous en soucions pas, car nous connaissons la propagation de telles erreurs.

Par conséquent, nous allons commencer par étudier la propagation de b'_k par les vecteurs b_{k+p} sans tenir compte des erreurs commises sur les α_i et β_i (Ceci est possible en égard à l'hypothèse de linéarité des erreurs). Ensuite nous étudierons l'erreur engendrée sur α_k et β_{k-1}

Notations : nous prenons les mêmes qu'au chapitre précédent mais nous faisons en plus l'hypothèse que b'_k peut être décomposé suivant les vecteurs propres u_i

Nous poserons alors $b'_k = \sum_{i=1}^{n} (b'_k)_i \; u_i$

I°- Propagation d'une erreur par les vecteurs b_{k+p}

Nous supposons, ce que nous pouvons toujours faire, que par une transmutation l'espace est rapporté aux vecteurs propres.

$b_{k+1} = (A - \alpha_k I)\, b_k - \beta_{k-1}\, b_{k-1}$

Donc, si b_k est remplacé par $b'_k + b_k$

$$b'_{k+1} = (A - \alpha_k I)\, b'_k = \sum_{i=1}^{n} (A - \alpha_k I)\, (b'_k)_i \, u_i$$

$$= \sum_{i=1}^{n} (\lambda_i - \alpha_k)(b'_k)_i \, u_i = \sum_{i=1}^{n} (b'_{k+1})_i \; u_i$$

en appelant λ_i la valeur propre associée au vecteur propre u_i

$b_{k+2} = (A - \alpha_{k+1} I)\, b_{k+1} - \beta_k\, b_k$. Donc

$b'_{k+2} = (A - \alpha_{k+1} I)\, b'_{k+1} - \beta_k\, b'_k$

$$b'_{k+2} = \sum_{i=1}^{n} \left[(\lambda_i - \alpha_{k+1})(b'_{k+1})_i - \beta_k (b'_k)_i \right] u_i$$

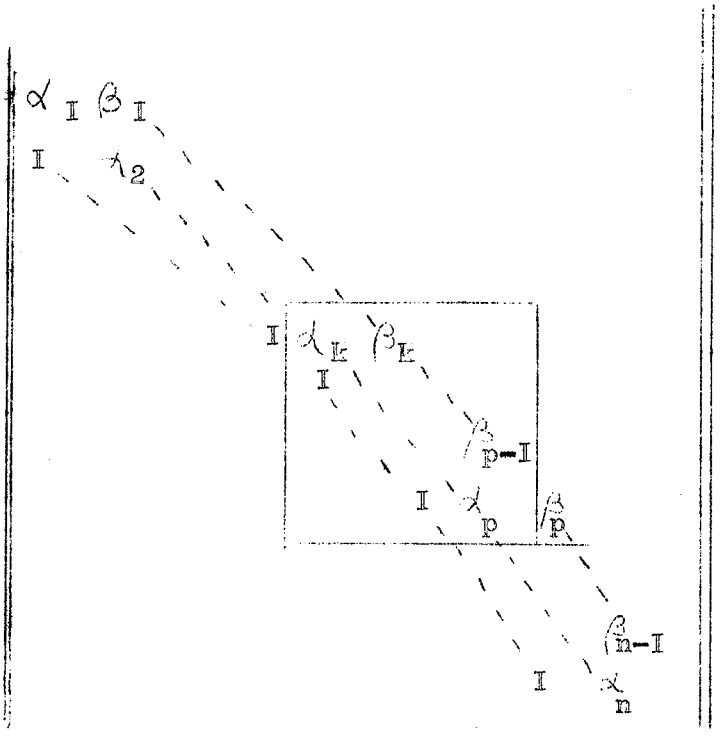
et de façon générale

$$b'_{k+p+1} = (A - \alpha_{k+p} I)\, b'_{k+p} - \beta_{k+p-1} b'_{k+p-1}$$

$$= \sum_{i=1}^{n} \left[(\lambda_i - \alpha_{k+p})(b'_{k+p})_i - \beta_{k+p-1} (b'_{k+p-1})_i \right] u_i$$

Nous voyons immédiatement apparaitre l'analogie avec la formation du polynôme caractéristique de la matrice tridiagonale.

Considérons cette matrice



Nous avons isolé une sous-matrice, également tridiagonale. Nous appellerons une telle sous-matrice B_k^p

C'est une matrice carrée d'ordre (p-k+I) matrice ayant la même diagonale de la matrice tridiagonale entre les termes de rang k et de rang p

Il est clair que l'erreur sur le vecteur b_{p+I} aura pour composante suivant le vecteur propre u_i

$$\text{Det} \left| B_k^p - \lambda_i \, I \right| (b_k'')_i$$

d'où

$$\boxed{b''_{p+I} = \sum_{i=I}^{n} \left[\text{Dét} \left| B_k^p - \lambda_i \, I \right| (b_k'')_i \; u_i \right]}$$

2°- Erreur sur les termes α_k et β_k dûe à une erreur b_k

Au lieu de calculer γ_k nous calculons γ_k^* qui vaut

$$\gamma_k^* = b_k \cdot c_k + b'_k \, c_k = \gamma_k + \gamma'_k$$

De même au lieu de calculer $c_k \, Ab_k$ nous calculons

$$(c_k \, A)(b_k + b''_k) = \alpha_k \, Ab_k + \alpha_k \, Ab'_k$$

Donc au lieu de λ_k nous aurons:

$$\lambda_k^* = \frac{c_k \, Ab_k + c_k \, Ab'_k}{c_k \, b_k + c_k \, b'_k} \neq \frac{c_k \, Ab_k + c_k \, Ab'_k}{c_k \, b_k} \left(I - \frac{c_k \, b'_k}{c_k \, b_k} \right)$$

d'ou $$\alpha'_k = \frac{c_k\, Ab'_k}{c_k\, b_k} - \frac{\alpha_k\, c_k\, b'_k}{\sigma_k}$$

$$\alpha'_k = \frac{1}{\sigma_k}\left[c_k A - \alpha_k\, c_k\right] b'_k$$

$$\alpha'_k = \frac{1}{\sigma_k}\left[c_{k+1} + \beta_{k-1}\, c_{k-1}\right] b'_k$$

Pour $\beta_{k-1} = \frac{\sigma_k}{\sigma_{k-1}}$ nous obtenons de même

$$\beta^*_{k-1} = \frac{\sigma_k + \sigma'_k}{\sigma_{k-1}}$$ ce qui nous donne une erreur

$$\beta'_{k-1} = \frac{\sigma'_k}{\sigma_{k-1}} = \frac{c_k\, b'_k}{\sigma_{k-1}} = \frac{\beta_{k-1}}{\sigma_k}\, c_k\, b'_k$$

En résumé nous obtiendrons comme erreurs :

$$\boxed{\alpha'_k = \frac{c_{k+1}\, b'_k}{\sigma_k} + \frac{c_{k-1}\, b'_k}{\sigma_{k-1}}}$$

$$\boxed{\beta'_{k-1} = \frac{c_k\, b'_k}{\sigma_{k-1}}}$$

Nous pouvons majorer ces erreurs en remplaçant chaque produit scalaire par le produit des normes de vecteurs, nous obtenons alors :

$$|\alpha'_k| \leq \left(\frac{\|c_{k+1}\|}{|\sigma_k|} + \frac{\|c_{k-1}\|}{|\sigma_{k-1}|}\right) \|b'_k\|$$

et $$|\beta'_{k-1}| \leq \frac{\|c_k\| \cdot \|b'_k\|}{|\sigma_{k-1}|}$$

Tous les α et β postérieurs seront aussi entachés d'erreurs, les formules seront analogues il suffit de remplacer b'_k par b'_{k+p} pour avoir l'erreur sur α_{k+p} et β_{k+p-1}

Comme b'_{k+p} est une erreur dûe à la propagation de b'_k . Il serait intéressant d'avoir α'_{k+p} et β'_{k+p} en fonction de b'_k

Or $$b'_{k+p} = \sum_{i=1}^{n}\left[\mathrm{Det}\left|B_k^{k+p-1} - \lambda_i\, I\right| (b'_k)_i\, u_i\right]$$

Nous pouvons facilement trouver un majorant de α'_k et de β'_k

En effet $\left| (b'_{k+p})_i \right| \leq \operatorname{Max}_i \left| \operatorname{Det} \left[B_k^{k+p-I} - \lambda_i I \right] \right| \left| (b')_i \right|$

On en déduit que $\| b'_{k+p} \| \leq \operatorname{Max}_i \left| \operatorname{Det} (B_k^{k+p-I} - \lambda_i I) \right| \| b'_k \|$

D'ou les formules

$$|\alpha'_{k+p}| \leq \operatorname{Max}_i \left| \operatorname{Det} (B_k^{k+p-I} - \lambda_i I) \right| |\alpha'_k|$$

$$|\beta'_{k+p-I}| \leq \operatorname{Max}_i \left| \operatorname{Det} B_k^{k+p-I} - \lambda_i I \right| |\beta'_{k-I}|$$

Remarque : Sur α_{k+p} il y a outre cette erreur les erreurs propagées à partir des erreurs sur les α_{k+j} antérieures ($j < p$), erreurs qui sont dues aux produits scalaires faux. Sur β'_{k+p-I}, au contraire il n'y a pas d'erreurs propagées.

Nous allons maintenant chercher l'erreur totale causée sur un terme α_{k+p}.

Nous avons : $|\alpha'_{k+j}| \leq \operatorname{Max}_i \left| \operatorname{Det} (B_k^{k+j-I} - \lambda_i I) \right| |\alpha'_k|$

Cette erreur se transmet sur les termes suivants de la façon suivante :

$$\alpha'_{k+j+2i} = 2\alpha'_{k+j}$$

$$\alpha'_{k+j+2i-I} = -2\alpha'_{k+j}$$

Donc $|\alpha'_{k+j+i}| = 2 |\alpha'_{k+j}|$

Je vais désigner par $\overline{\alpha'_{k+p}}$ l'erreur totale sur α_{k+p}, on aura :

$$|\overline{\alpha'_{k+p}}| < |\alpha'_k| + 2 \sum_{j=I}^{j=p-I} \left\{ \operatorname{Max}_i \left| \operatorname{Det} (B_k^{k+j-I} - \lambda_i I) \right| \right\} |\alpha'_k| + \operatorname{Max}_i \left| \operatorname{Det} (B_k^{k+p-I} - \lambda_i I) \right| |\alpha'_k|$$

donc

$$|\alpha'_{k+p}| < |\alpha'_k| \left\{ I + \operatorname{Max}_i \left| \operatorname{Det} (B_k^{k+p-I} - \lambda_i I) \right| + 2 \sum_{j=I}^{j=p-I} \operatorname{Max}_i \left| \operatorname{Det} (B_k^{k+j-I} \lambda_i I) \right| \right\}$$

et en remplaçant $|\alpha'_k|$ par sa valeur

$$|\overline{\alpha'_{k+p}}| \leq \|b'_k\| \left(\frac{\|c_{k+I}\|}{|\tau_k|} + \frac{\|c_{k-I}\|}{|\tau_{k-I}|} \right) \left\{ I + \operatorname{Max}_i \left| \operatorname{Det} (B_k^{k+p-I} \lambda_i I) \right| + 2 \sum_{j=I}^{j=p-I} \operatorname{Max}_i \left| \operatorname{Det} (B_k^{k+j-I} - \lambda_i I) \right| \right\}$$

Tandis que

$$|\beta'_{k+p-I}| \leq \operatorname{Max}_i \left| \operatorname{Det} (B_k^{k+p-I} - \lambda_i I) \right| \frac{\|c_k\| \cdot \|b'_k\|}{|\tau_{k-1}|}$$

Or nous avons vu au chapitre II de cette partie une majoration de $\| b'_k \|$

$$\| b'_k \| < \frac{S\,(\Omega+n+1)}{\Omega} \left\{ \| c_{k-1} \| \cdot \| b_{k-1} \| \left[\frac{\| b_{k-1} \|}{|\sigma_{k-1}|} \left(2N(A) + |\alpha_{k-1}| \right) + \frac{\| b_{k-2} \|}{|\sigma_{k-2}|} \right] \right.$$

$$\left. + \frac{|\beta_{k-2}|}{|\sigma_{k-2}|} \| c_{k-2} \| \, \| b_{k-2} \|^2 + N(A) \, b_{k-1} \right.$$

Ceci nous permet donc d'écrire une majoration de $|\overline{\alpha'_{k+p}}|$ et de $|\overline{\beta'_{k+p-1}}|$ fonction uniquement de quantités soit connues, soit calculables :

Nous aurons alors :

$$|\overline{\alpha'_{k+p}}| < \frac{S\,(\Omega+n+1)}{\Omega} \left\{ \| c_{k-1} \| \cdot \| b_{k-1} \| \left[\frac{|b_{k-1}|}{|\sigma_{k-1}|} \left(2N\,(A) + |\alpha_{k-1}| \right) + \frac{\| b_{k-2} \|}{\| \sigma_{k-2} \|} \right] \right.$$

$$\left. + \frac{|\beta_{k-2}|}{|\sigma_{k-2}|} \| c_{k-2} \| \cdot \| b_{k-2} \|^2 + N\,(A) \| b_{k-1} \| \right\} \left\{ I + \right.$$

$$\left. \operatorname*{Max}_i \left| \operatorname{Det} \left(B_k^{k+p-1} - \lambda_i I \right) \right| + 2 \sum_{j=1}^{j=p-1} \operatorname*{Max}_i \left| \operatorname{Det} \left(B_k^{k+j-1} - \lambda_i I \right) \right| \right\}$$

$$\cdot \left(\frac{\| c_{k+1} \|}{|\sigma_k|} + \frac{\| c_{k-1} \|}{|\sigma_k|} \right)$$

et :

$$|\overline{\beta'_{k+p-1}}| < \operatorname*{Max}_i \left| \operatorname{Det} \left(B_k^{k+p-1} - \lambda_i I \right) \right| \frac{S(\Omega+n+1)}{\Omega} \left\{ \| c_{k-1} \| \cdot \| b_{k-1} \| \right.$$

$$\left[\frac{\| b_{k-1} \|}{|\sigma_{k-1}|} \left(2\,N\,(A) + |\alpha_{k-1}| \right) + \frac{\| b_{k-2} \|}{|\sigma_{k-2}|} \right] + \frac{|\beta_{k-2}|}{|\sigma_{k-2}|} \| c_{k-2} \| \cdot \| b_{k-2} \|^2$$

$$\left. + N\,(A) \, \| b_{k-1} \| \right\} \frac{\| c_k \|}{|\sigma_{k-1}|}$$

3°- Vérifications :

Le principe consiste à ajouter au vecteur b_k un vecteur b'_k notablement supérieur au vecteur d'erreur dûe au calcul.

Malheureusement il est à peu près impossible de calculer effectivement $|\overline{\alpha'_{k+p}}|$ Ce qui fait que pratiquement seule l'erreur faite au premier pas pourra être vérifier.

J'ai les formules

$$\alpha'_k = \frac{c_{k+1}\, b'_k}{\Gamma_k} + \frac{c_{k-1}\, b'_k}{\Gamma_{k-1}}$$

et

$$\beta'_{k-1} = \frac{c_k\, b''_k}{\Gamma_{k-1}}$$

Les produits scalaires Γ sont calculés au cours de la méthode de Lanczos. Un programme accessoire m'a permis de les extraire et de faire les produits scalaires : $\langle c_{k+1}, b'_k \rangle$; $\langle c_k, b'_k \rangle$; $\langle c_{k-1}, b'_k \rangle$;

Avec une matrice 8x8 dont j'avais altéré le vecteur b_3 en y ajoutant le vecteur $b''_3 = (0,1 ; 0,1 ; \ldots 0,1)$

Vecteur erreur extrêment important puisque les composantes de b_3 étaient comprises entre 0,5 et 2,5 (en valeurs absolues). J'ai obtenu les résultats suivants :

$\alpha_3 = -0,5$
α_3 modifié $= -0,5112\ 1116\ 47$
d'ou $d'_3 = -0,0112\ 1116\ 47 \neq 0,012$

d'autre part :

$c_4\, b'_3 = -93,5111\ 485$
$c_2\, b''_3 = 0,1149\ 999\ 94$
$\Gamma_2 = 200$
$\Gamma_3 = 3379,699\ 92$

donc on doit avoir

$$\alpha'_3 = \frac{-93,5111\ 485}{3379,699\ 92} + \frac{3,1149\ 999\ 94}{200}$$

$$= -0,0111\ 9113\ 576 \neq 0,012$$

Pour β_2 j'obtiens

$\beta_2 = 16,673\ 266$
β_2 modifié $= 16,7311\ 137$
d'ou $\beta'_2 = 0,057\ 9211 \neq 0,058$
or $c_3\, b''_3 = 11,7$
et $\Gamma_2 = 200$
d'ou $\beta'_2 = \frac{11,7}{200} = 0,0585 \neq 0,058$

Nous voyons donc que l'erreur ne diffère que de très peu de la valeur prévue. Il faut noter d'ailleurs que nous sommes dans un cas qui ne se produira jamais en réalité, en effet on n'obtiendra jamais d'erreur aussi importante que celle introduite ici. Or l'étude faite suppose la linéarité des erreurs, celle-ci sera vérifiée bien plus étroitement dans les cas réels qu'ici ; ce qui permet d'estimer que la formule d'erreur trouvée est satisfaisante.

Troisième Partie

Expériences Numériques

Chapitre I

Détail du programme

Un programme déstiné à mettre en oeuvre la méthode de Lanczos a été écrit sur le calculateur du laboratoire de calcul de l'Université de Grenoble, un Gamma ET Bull. Je n'entrerai pas ici dans le détail de ce programme il me suffira d'indiquer qu'il utilise des nombres écrits en point décimal flottant (P.D.F.) en simple précision, c'est à dire, avec une mantisse de 9 chiffres et un exposant pouvant varier de -50 à + 49. Dans les opéeations les troncatures sont effectuées sans arrondis.

Le programme écrit permet de normaliser les vecteurs obtenus et de réorthogonaliser les vecteurs b_K (resp c_K) aux vecteurs c_j (resp b_j) ; $j<K$. Ces deux compléments ont été proposés par Rutishauser et repris depuis par les différents auteurs.

La réorthogonalisation a pour but de maintenir la propriété d'orthogonalité, propriété essentielle, puisque c'est sur elle que repose toète la méthode.

La normalisation, elle, ne vise pas à diminuer les erreurs, mais comme il a été constaté, qu'en général, les modules des vecteurs avaient tendance à diminuer il s'agit d'éviter les pertes de précision dûes aux exposants trop faible.

Voici plus précisement de quoi il s'agit :

<u>Réorthogonalisation</u> :

A chaque pas on remplace le vecteur b_K calculé par b_K^* définit par

$$b_K^* = b_K - \sum_{j=1}^{K-1} \gamma_{K,j}\, b_j \qquad \text{avec} \quad \gamma_{Kj} = \frac{c_j\, b_K}{\sigma_j}$$

Normalisation :

Les δ_k étant des coefficients de normalisation on modifie les équations de la façon suivante :

$$\delta_2 b_2 = Ab_1 - \alpha_1 b_1 \qquad \delta_2 c_2 = c_1 A - \alpha_1 c_1$$

$$\delta_{k+1} b_{k+1} = Ab_k - \alpha_k b_k - \gamma_{k-1} b_{k-1} \qquad \delta_{k+1} c_{k+1} = c_k A - \alpha_k c_k - \gamma_{k-1} c_{k-1}$$

avec $\alpha_k = \dfrac{c_k Ab_k}{\sigma_k}$ $\qquad \gamma_{k-1} = \delta_k \dfrac{\sigma_k}{\sigma_{k-1}}$

Rutishauser a montré que α_k est indépendant du choix des δ_k ainsi que $\beta_{k-1} = \gamma_{k-1}\,\sigma_k$

La matrice tridiagonale à alors la forme suivante :

$$\begin{pmatrix} \alpha_1 & \gamma_1 & & & \\ \delta_2 & \alpha_2 & \gamma_2 & & \\ & \ddots & \ddots & \ddots & \\ & & & \gamma_{n-1} & \\ & & \delta_n & \alpha_n & \end{pmatrix}$$

Et le polynôme caractéristique est défini par la formule

$$P_{k+1}(\lambda) = \frac{(\lambda - \alpha_k) P_k(\lambda) - \gamma_{k-1} P_{k-1}(\lambda)}{\delta_{k+1}}$$

Les coefficients δ_k ont été choisis égaux à la plus petite des deux normes, des vecteurs b_k et c_k ; ce choix reposant sur le fait que ces normes ont tendance à décroitre.

Ceci nous oblige à calculer $\|b_k\|$ et $\|c_k\|$, mais comme nous disposons de σ_k, nous n'aurons besoin d'aucun calcul pour effectuer le contrôle prévu dans la deuxième partie sur cos (c_k, b_k).

Dans le programme, la réorthogonalisation, comme la normalisation peuvent être supprimées à volonté.

Chapitre II

Essais

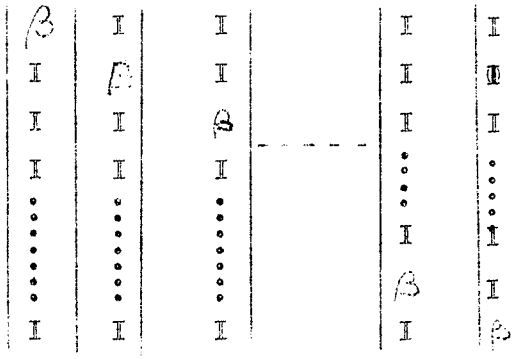
Nous avons utilisé le programme précédent pour étudier expérimentalement trois problèmes différents :

- L'utilité de la réorthogonalisation
- Le choix des vecteurs de départ
- L'influence des valeurs caractéristiques de la matrice étudiée.

I°- Matériel expérimental.

Pour faire ces expériences j'ai utilisé un programme de génération de matrices ayant des vecteurs et des valeurs propres connus, que je dois à l'obligeance de Monsieur Gastinel.

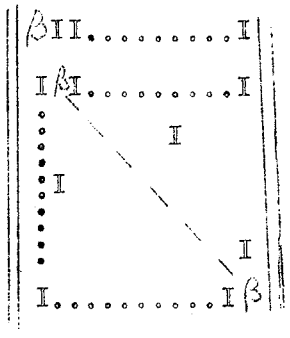
Ce programme fabrique des matrices ayant les vecteurs propres suivants :



Et des valeurs propres quelconques.

Le coefficient β est à la disposition de l'utilisateur.

si on pose



B est la matrice des vecteurs propres et alors on a :

$$B^{-I} = \begin{Vmatrix} \alpha'+\beta' & \beta' & \beta' & \cdots & \beta' \\ \beta' & \alpha'+\beta' & \beta' & & \\ \vdots & \beta' & \alpha'+\beta' & \ddots & \\ \beta' & & & \beta' & \alpha'+\beta' \end{Vmatrix}$$

avec $\alpha' = \frac{I}{\beta - I}$ et $\beta' = -\frac{I}{(\beta-I)(\beta-I+n)}$

n étant, bien entendu l'ordre de la matrice.

B^{-I} est la matrice des lignes propres.

Les vecteurs propres ainsi formés sont régulièrement répartis, ils forment un parapluie dont β caractérise l'ouverture.

si $\beta = I$, ils sont tous confondus ; si $\beta = \infty$ ils sont tous orthogonaux.

Donc, dans notre matrice nous aurons à notre disposition les valeurs propres et le parapluie des vecteurs propres.

J'ai effectué mes essais sur des matrices I5xI5

Interprétation des résultats.

On pourrait résoudre le polynôme caractéristique obtenue, malheureusement ceci est long et la résolution introduit des erreurs importantes.

J'ai préféré la méthode qui consiste à mesurer la précision par le nombre de chiffres exacts sur les coefficients du polynôme caractéristique. Cette méthode est critiquable : en effet, les études sur la résolution des polynômes montrent qu'à de faibles erreurs sur les coefficients ne correspondent pas obligatoirement de faibles erreurs sur les racines. De plus tous les coefficients ne sont pas également sensible aux erreurs.

A cela il faut ajouter ici les faits suivants : le nombre de chiffres exacts n'est pas absolument constant d'un coefficient à l'autre. D'autre part le problème se pose de connaitre les coefficients

exacts. Or si les coefficients extrêmes peuvent être assez facilement calculés exactement à partir des racines il n'en est pas de même pour les coefficients centraux.

En pratique j'ai après quelques tatonnements mis au point le procédé suivants : je calcule directement les coefficients extrêmes. De quelques résolutions faites dans les cas les plus favorables et de ces calculs exacts je tire une table des coefficients du polynôme avec, par exemple, 6 chiffres exacts. (Ce sont les chiffres communs aux différents calculs) Comme tous les autres calculs seront fait dans des cas moins favorables, cette table est en pratique suffisante et suffisamment sûre.

2°- Influence de la réorthogonalisation :

Pour réorthoganaliser le vecteur b_K on est amené à calculer K-I produits scalaires, à faire K-I divisions et à additionner K-I vecteurs.

Ceci va demander beaucoup de temps. D'autre part le nombre d'opérations effectuées est grand et on commet des erreurs sur ces opérations. On peut alors se demander s'il y aura réellement amélioration.

On peut d'autant plus se le demander que la situation est beaucoup plus grave qu'il ne parait à première vue.

En effet b_K devait être orthogonal à c_j et s'il n'y est pas, il diffère peu d'un vecteur orthogonal donc $(c_j , b_K) \# \frac{\Pi}{2}$.

Nous sommes là justement dans le cas ou les erreurs de calculs sont très importantes.

Cependant, il faut remarquer que plus le calcul direct manquera de précision, meilleure sera la correction. En effet, plus b_K s'écartera de la normale à c_j plus la correction sera efficace. On Peut donc prévoir que la correction sera en quelque sorte "auto-compensatrice" et qu'elle sera d'autant plus efficace que le vecteur à corriger sera moins bon.

Le but des calculs effectués dans ce cadre devait nous permettre de vérifier ces hypothèses :

Tout d'abord j'ai constaté que la réorthogonalisation double approximativement le temps de calcul, son coût est donc loin d'être négligeable.

En ce qui concerne son efficacité, je n'ai pas constaté de cas où elle n'apporta pas une amélioration. Toutefois, dans les cas que je qualifierai, de favorables (il s'agissait dans ces cas de matrices ayant leurs valeurs propres bien distinctes, et d'une étude menée avec des vecteurs de départ bien choisis) l'amélioration était très faible.

Voici par exemple, le résultat obtenu avec une matrice ayant comme valeurs propres : I ; -2 ; 3 ; -4 ; ... ; -I4 ; I5, et un $\beta = 2$. Les vecteurs de départ étant choisis pour donner de bons résultats. Le polynôme caractéristique trouvé en utilisant la réorthogonalisation est :

	x^{I5} - 8,000 000 24 x^{I4}
- 587,999 997	x^{I3} + 4255,999 99 x^{I2}
+ I32 90I, 995	x^{II} - 858 I08, 985 x^{I0}
- I4 66I I28, 2	x^{9} + 8I 228 I25, 0 x^{8}
+ 828 780 788	x^{7} - 3 8II 85I 680 x^{6}
- 28 038 682 600	x^{5} + 82 808 254 800 x^{4}
+ 278 348 808 000	x^{3} - 689 324 758 000 x^{2}
- 948 550 I82 000	x + I 807 674 280 000

Les erreurs relatives sur les coefficients sont, au maximum de $5 . I0^{-8}$, ainsi, le deuxième coefficient est exactement : -8 ; le troisième : - 588 ; le dernier I 807 674 868 000

Le même essai, mais sans réorthogonalisation, nous a donné les résultats suivants :

	x^{I5} - 8,000 000 57 x^{I4}
- 0,587 999 978 . $I0^{3}$	x^{I3} + 0,425 600 0I3 . $I0^{4}$ x^{I2}
+ 0,I32 90I 994 . $I0^{6}$	x^{II} - 0,858 I04 086 . $I0^{6}$ x^{I0}
- 0,I46 6II 282 . $I0^{8}$	x^{9} + 0,8I2 28I 822 . $I0^{8}$ x^{8}
+ 0,828 780 789 . $I0^{9}$	x^{7} - 0,38I I85 2I0 . $I0^{I0}x^{6}$
- 0,280 386 826 . $I0^{II}$	x^{5} + 0,828 082 685 . $I0^{II}x^{4}$
+ 0,278 348 8I0 . $I0^{I2}$	x^{3} - 0,689 324 960 . $I0^{I2}x^{2}$
- 0,948 550 II8 . $I0^{I2}$	x + 0,I80 767 507 . $I0^{I3}$

Comme on peut le voir le nombre de chiffres exacts passe en gros de 7 à 6.

Si on essaye la même matrice, mais avec un vecteur b_I très voisin du vecteur propre associé à la valeur propre de plus grand module et un vecteur c_I voisin de la ligne propre associé à la même valeur propre ; ce qui, comme nous le verrons ci-après, est un très mauvais choix de vecteurs de départ. Alors ; avec la réorthogonalisation, la précision des résultats n'est pas modifiée.

Par contre sans réorthogonalisation, on obtient de très mauvais résultats :

$$
\begin{aligned}
& & x^{15} &- 0{,}795\ 209\ 316 \cdot 10^{1}\, x^{14} \\
&- 0{,}588\ 273\ 272 \cdot 10^{3} & x^{13} &+ 0{,}422\ 775\ 814 \cdot 10^{4}\, x^{12} \\
&+ 0{,}138\ 037\ 808 \cdot 10^{6} & x^{11} &- 0{,}846\ 744\ 653 \cdot 10^{6}\, x^{10} \\
&- 0{,}146\ 858\ 319 \cdot 10^{8} & x^{9} &+ 0{,}805\ 377\ 937 \cdot 10^{8}\, x^{8} \\
&+ 0{,}830\ 765\ 925 \cdot 10^{10} & x^{7} &- 0{,}377\ 440\ 640 \cdot 10^{10} x^{6} \\
&- 0{,}231\ 143\ 558 \cdot 10^{11} & x^{5} &+ 0{,}318\ 619\ 460 \cdot 10^{11} x^{4} \\
&+ 0{,}274\ 430\ 059 \cdot 10^{12} & x^{3} &- 0{,}680\ 371\ 731 \cdot 10^{12} x^{2} \\
&- 0{,}952\ 024\ 409 \cdot 10^{12} & x &+ 0{,}129\ 150\ 227 \cdot 10^{13}
\end{aligned}
$$

Ainsi qu'on peut s'en rendre compte il n'y a plus, en moyenne, que deux chiffres exacts.

De nombreux essais ont tous donné le même résultat, à savoir : influence très faible de la réorthogonalisation dans les cas favorables, mais par contre amélioration substantielle dans les autres cas.

3°- Influence du choix du vecteur de départ (sans réorthogonalisation)

Il s'agissait surtout ici de voir s'il était interessant de partir d'un vecteur voisin du vecteur propre associé à la valeur propre de plus grand module. En effet, un tel vecteur peut être obtenu facilement en multipliant un vecteur arbitraire par les puissances de la matrice.

Si nous supposons un vecteur décomposé suivant les vecteurs propres u_i de la matrice , nous aurons :

$b = \sum_{i=I}^{n} a_i u_i$ a_i étant la composante suivant le vecteur u_i

$A\,b = \sum_{i=I}^{n} a_i \lambda_i u_i$

Appelons u_i le vecteur propre associé à la valeur propre de plus grand module

Si $|a_i| << |a_I| \quad \forall i \neq I$

Comme $|\lambda_I| > |\lambda_i| \ \forall i \neq I \quad \frac{|a_i \lambda_i|}{|a_1 \lambda_1|} < \frac{|a_i|}{|a_1|}$

Alors Ab est intèrieur au cône centré sur u_I et de génératrice b La méthode de Lanczos va avoir à former une base avec des vecteurs tous intérieurs à ce cône, donc tous très voisins on conçoit qu'alors les erreurs vont être importantes.

L'expérience devait verifier cette hypothèse

Dans la matrice considérée dans le 2ème § le vecteur propre associé à la valeur propre de plus grand module était

$$\begin{vmatrix} I \\ I \\ I \\ \vdots \\ 2 \end{vmatrix}$$

et la ligne propre correspondante (I,I,...,-15).

en partant avec un vecteur $b_I = \begin{vmatrix} I \\ I \\ \cdot \\ \cdot \\ 2,2 \end{vmatrix}$

et un vecteur c_I=(I,I,.....,-I4) On obtenait le polynôme indiqué à la fin du 2ème § on le voit les résultats sont très mauvais.

Par contre en partant avec les vecteurs:

$$b_I = \begin{vmatrix} I \\ I \\ \vdots \\ I \\ 3 \end{vmatrix} \qquad C_I = (I,I,....,I,-I2)$$

qui sont relativement éloigné la précision était du même ordre qu'avec les vecteurs $\begin{vmatrix} I \\ I \\ . \\ . \\ I \end{vmatrix}$ et $(I,I,\ldots,I)$

vecteurs qui, égaux à la somme des vecteurs propres se sont révélés donner de trés bons résultats.

J'ai alors éssayé de partir prés d'un autre vecteur propre, en particulier de celui associé à la valeur propre de plus petit module. Je n'ai constaté alors aucune influence defavorable.

Jai également éssayé de partir avec un vecteur différant peu d'un vecteur orthogonal ou vecteur propre associé à la valeur propre de plus grand module (et d'un vecteur ligne équivalent). Un tel vecteur est perpendiculaire à un vecteur A^k b.Les resultats ont été absoluments catastrophiques; j'ai éssayé ceci pour plusieurs matrices,pour toutes j'ai sorti des polynômes caractéristiques ayant des coefficients sans aucun chiffre exact.

Tout au plus l'ordre de grandeur de l'exposant était-il conservé. Ce résultat peut s'expliquer par le fait que la composante sur le vecteur propre associé à la valeur propre de plus grand module est voisine de zero, or nous avons vu que si une composante suivant un vecteur propre est nulle, il s'agit d'un cas particulier et que l'algorithme ne peut s'achever.

Nous somme donc trés prés d'un cas particulier c'est ce qui explique le résultat.

Un fait ressort de ces expériences: c' est l'influence trés grand des composantes correspondantes au vecteur et à la ligne propres associés à la valeur propre de plus grand module.

Le résultat pratique etant qu'il ne faut partir ni suivant leurs direction, ni perpendiculairement.

4me : Influence des valeurs propres voisines.

Soit une matrice ayant deux valeurs propres voisines: λ et $\lambda+d\lambda$. λ est associée au vecteur propre u_I et $\lambda+d\lambda$ au vecteur propre u_2.

Je pose $u=u_I+u_2$ et j'appelle (u,u^I) une base de l'espace $(u_I \times u_2)$ dont un des vecteurs de la base est justement $u=u_I+u_2$

J'appelle E_{n-I} l'espace

$(u \times u_3 \times u_4 \times \ldots \times u_n)$

et

En l'espace $(u_I \times u_2 \times u_3 \times \ldots \times u_n) = E_{n-I} \times u^I$

Je pose $b_I = \sum_{i=I}^{n} u_i$ Ce que j'ai toujours le droit de faire

Alors

$$Ab_I = \sum_{i=3}^{n} \lambda_i u_i + \lambda u + u_2 d\lambda$$

Or $u_2 d\lambda$ est un vecteur infiniment petit.

Et je peux écrire $u_2 d\lambda = k_I u\, d\lambda + k_2 u^I d\lambda$.

Seul $k_2 u^I d\lambda$ n'appartient pas à E_{n-I}

Donc la partie n'appartenant pas à E_{n-I} est un infiniment petit je peux écrire:

$b_2 = b_2^* + db_2$

avec $b_2^* \in E_{n-I}$ et $db_2 \in u^I$ posons $b_2^* = \sum_{i=I}^{n} (b_2^*)_i u_i$

$$Ab_2 = \sum^{n} \lambda_i (b_2^*)_i u_i + \lambda \left[(b_2^*)_I + (b_2^*)_2 \right] u + \text{infinements petits}$$

De même à chaque pas la composante selon u^I sera infiniment petite.

Au pas n-I nous aurons déterminé dans E_{n-I} n-I vecteurs indépendants : la composante de b_n appartenant à E_{n-I} sera nulle, seule celle selon u^I sera non nulle.

Or cette composante est très petite par rapport aux vecteurs des pas précédents. Les erreurs faites aux pas précédents ont été déterminées par le module des vecteurs à ces pas. Ces erreurs sont

donc très grandes par rapport à la composante suivant u^I. Donc sur cette composante une erreur très importante sera commise.

D'autre part si on fait tendre $d\lambda$ vers zéro, la composante suivant u^I va tendre elle aussi vers zéro. Or pour une valeur propre double on a $d\lambda = 0$ et pas de composante suivant u^I.

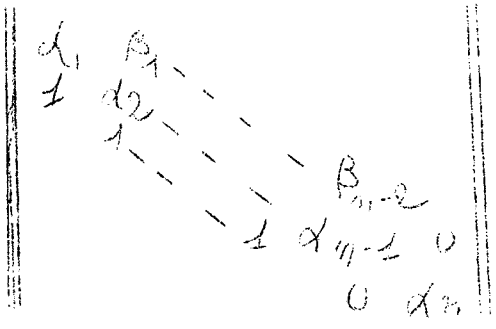
Donc, lorsque $d\lambda$ tend vers zéro on tend de façon continue vers le cas des valeurs propres doubles.

On voit ainsi que la discontinuité théorique entre valeurs propres distinctes et confondues disparait en pratique du fait des erreurs de calcul.

D'autre part $\sigma_n = c_n\, b_n$ sera très petit

donc $\beta_{n-I} = \dfrac{\sigma_n}{\sigma_{n-I}}$ le sera aussi

Ainsi β_{n-1} va tendre vers zéro lorsque deux valeurs propres vont se rapprocher, or dans le cas de deux valeurs propres confondues on a une factorisation du polynôme caractéristique qui se traduit par la forme tridiagonale suivante



Donc nous aurons également une tendance continue vers cette forme du fait des erreurs de calcul.

Ceci apparait plus nettement avec la normalisation : en effet

$$\|b_n\| \leq I$$
$$\|c_n\| \leq I$$

Or $\|\sigma_n\, b_n\| \ll I$ donc $\sigma_n \ll I$

Comme $\alpha_{n-I} = \sigma_n \dfrac{\sigma_n}{\sigma_{n-I}}$; $|\alpha_{n-I}| \ll I$

Ceci apparait parfaitement aux essais :
Les matrices essayées avaient comme valeurs propres
I; -2; 3; -4; 5; 5,0I; -6; 7; -8; 9; -I0; II; -I2; I3; -I4.
Alors que les différents coefficients σ_k et γ_k sont de l'ordre de quelques unités. γ_{n-I} et σ_n sont de l'ordre du centième: donc la factorisation apparait nettement.

Quand au polynôme caractéristique alors qu'il devrait être

$$x^{I5} + 0,I99\ 000 \cdot I0\ x^{I4} - 0,5I3\ 070 \cdot I0^3\ x^{I3}$$
$$- 0,569\ I69 \cdot I0^3\ x^{I2} + 0,I03\ 042 \cdot I0^6\ x^{II} + 0,266\ 852 \cdot I0^5\ x^{I0}$$
$$- 0,998\ 679 \cdot I0^7\ x^9 + 0,487\ 846 \cdot I0^7\ x^8 + 0,494\ 955 \cdot I0^9\ x^7$$
$$- 0,539\ 470 \cdot I0^9\ x^6 - 0,I20\ 333 \cdot I0^{II} x^5 + 0,I77\ 3I7 \cdot I0^{II} x^4$$
$$+ 0,I24\ 454 \cdot I0^{I2} x^3 - 0,I9I\ 990 \cdot I0^{I2} x^2 - 0,374\ 876 \cdot I0^{I2} x$$
$$+ 0,436\ 763\ 239 \cdot I0^{I2}$$

(Je n'ai indiqué les coefficients qu'avec 6 chiffres car je n'ai pas pu obtenir une précision supérieure, sauf pour les coefficients extrêmes).

J'obtenais avec les vecteurs (I, I, ..., I) et $\begin{vmatrix} I \\ I \\ \vdots \\ I \end{vmatrix}$

$$x^{I5} + 0,I98\ 782 \cdot I0\ x^{I4} - 0,5I8\ 086 \cdot I0^3\ x^{I3}$$
$$- 0,568\ II2 \cdot I0^3\ x^{I2} + 0,I03\ 049 \cdot I0^6\ x^{II} - 0,264\ 9I8 \cdot I0^5\ x^{I0}$$
$$- 0,998\ 785 \cdot I0^7\ x^9 + 0,489\ 539 \cdot I0^7\ x^8 + 0,495\ 020 \cdot I0^9\ x^7$$
$$- 0,540\ I98 \cdot I0^9\ x^6 - 0,I20\ 358 \cdot I0^{II} x^5 + 0,I77\ 46I \cdot I0^{II} x^4$$
$$+ 0,I24\ 487 \cdot I0^{I2} x^3 - 0,I92\ I0I \cdot I0^{I2} x^2 - 0,375\ 002 \cdot I0^{I2} x$$
$$+ 0,436\ 957 \cdot I0^{I2}$$

Comme on peut le constater il n'y a plus, en moyenne, que 3 chiffres exacts par coefficient.

Avec la réorthogonalisation la précision monte à 6 chiffres (au lieu de 7, lorsque les valeurs propres sont toutes bien

séparées). Donc on voit, ainsi que je l'ai dit dans le paragraphe 2, que la réorthogonalisation est très utile dans les cas ou la précision des résultats a tendance à diminuer.

S O M M A I R E

-:-:-:-