Calcul de valeurs propres de grandes matrices hermitiennes par des techniques de partitionnement

Yousef Saad

THESE

présentée à

UNIVERSITE SCIENTIFIQUE ET MEDICALE DE GRENOBLE

INSTITUT NATIONAL POLYTECHNIQUE DE GRENOBLE

par

Youcef SAAD

pour obtenir le grade de

Docteur de troisième cycle

MATHEMATIQUES APPLIQUEES

Calcul de valeurs propres de grandes matrices hermitiennes par des techniques de partitionnement

Thèse soutenue le 21 mars 1974 devant la commission d'examen

Président Monsieur N. GASTINEL

Examinateurs Madame F. CHATELIN

Monsieur F. ROBERT

UNIVERSITE SCIENTIFIQUE
ET MEDICALE DE GRENOBLE

LISTE DES PROFESSEURS

Président : Monsieur Michel SOUTIF
Vice-Président : Monsieur Gabriel CAU

PROFESSEURS TITULAIRES

MM.	ANGLES D'AURIAC Paul	Mécanique des fluides
	ARNAUD Georges	Clinique des maladies infectieuses
	ARNAUD Paul	Chimie
	AUBERT Guy	Physique
	AYANT Yves	Physique approfondie
Mme	BARBIER Marie-Jeanne	Electrochimie
MM.	BARBIER Jean-Claude	Physique expérimentale
	BARBIER Reynold	Géologie appliquée
	BARJON Robert	Physique nucléaire
	BARNOUD Fernand	Biosynthèse de la cellulose
	BARRA Jean-René	Statistiques
	BARRIE Joseph	Clinique chirurgicale
	BENOIT Jean	Radioélectricité
	BERNARD Alain	Mathématiques Pures
	BESSON Jean	Electrochimie
	BEZES Henri	Chirurgie générale
	BLAMBERT Maurice	Mathématiques Pures
	BOLLIET Louis	Informatique (IUT B)
	BONNET Georges	Electrotechnique
	BONNET Jean-Louis	Clinique ophtalmologique
	BONNET-EYMARD Joseph	Pathologie médicale
	BONNIER Etienne	Electrochimie Electrométallurgie
	BOUCHERLE André	Chimie et Toxicologie
	BOUCHEZ Robert	Physique nucléaire
	BOUSSARD Jean-Claude	Mathématiques Appliquées
	BRAVARD Yves	Géographie
	BRISSONNEAU Pierre	Physique du solide
	BUYLE-BODIN Maurice	Electronique
	CABANAC Jean	Pathologie chirurgicale
	CABANEL Jean	Clinique rhumatologique et hydrologie
	CALAS François	Anatomie
	CARRAZ Gilbert	Biologie animale et pharmacodynamie
	CAU Gabriel	Médecine légale et Toxicologie
	CAUQUIS Georges	Chimie organique
	CHABAUTY Claude	Mathématiques Pures
	CHARACHON Robert	Oto-Rhino-Laryngologie
	CHATEAU Robert	Thérapeutique
	CHENE Marcel	Chimie papetière
	COEUR André	Pharmacie chimique
	CONTAMIN Robert	Clinique gynécologique
	COUDERC Pierre	Anatomie Pathologique
	CRAYA Antoine	Mécanique

Mme	DEBELMAS Anne-Marie	Matière médicale
MM.	DEBELMAS Jacques	Géologie générale
	DEGRANGE Charles	Zoologie
	DESRE Pierre	Métallurgie
	DESSAUX Georges	Physiologie animale
	DODU Jacques	Mécanique appliquée
	DOLIQUE Jean-Michel	Physique des plasmas
	DREYFUS Bernard	Thermodynamique
	DUCROS Pierre	Cristallographie
	DUGOIS Pierre	Clinique de Dermatologie et Syphiligraphie
	FAU René	Clinique neuro-psychiatrique
	FELICI Noël	Electrostatique
	GAGNAIRE Didier	Chimie physique
	GALLISSOT François	Mathématiques Pures
	GALVANI Octave	Mathématiques Pures
	GASTINEL Noël	Analyse numérique
	GEINDRE Michel	Electroradiologie
	GERBER Robert	Mathématiques Pures
	GIRAUD Pierre	Géologie
	KLEIN Joseph	Mathématiques Pures
Mme	KOFLER Lucie	Botanique et Physilogie végétale
MM.	KOSZUL Jean-Louis	Mathématiques Pures
	KRAVTCHENKO Julien	Mécanique
	KUNTZMANN Jean	Mathématiques appliquées
	LACAZE Albert	Thermodynamique
	LACHARME Jean	Biologie végétale
	LAJZEROWICZ Joseph	Physique
	LATREILLE René	Chirurgie générale
	LATURAZE Jean	Biochimie pharmaceutique
	LAURENT Pierre-Jean	Mathématiques appliquées
	LEDRU Jean	Clinique médicale B
	LLIBOUTRY Louis	Géophysique
	LOUP Jean	Géographie
Mle	LUTZ Elisabeth	Mathématiques Pures
MM.	MALGRANGE Bernard	Mathématiques Pures
	MALINAS Yves	Clinique obstétricale
	MARTIN-NOEL Pierre	Seméiologie médicale
	MASSEPORT Jean	Géographie
	MAZARE Yves	Clinique médicale A
	MICHEL Robert	Minéralogie et Pétrographie
	MOURIQUAND Claude	Histologie
	MOUSSA André	Chimie nucléaire
	NEEL Louis	Physique du solide
	OZENDA Paul	Botanique
	PAUTHENET René	Electrotechnique
	PAYAN Jean-Jacques	Mathématiques Pures
	PEBAY-PEYROULA Jean-Claude	Physique
	PERRET René	Servomécanismes
	PILLET Emile	Physique industrielle
	RASSAT André	Chimie systématique
	RENARD Michel	Thermodynamique
	REULOS René	Physique industrielle
	RINALDI Renaud	Physique
	ROGET Jean	Clinique de pédiatrie et de puériculture
	SANTON Lucien	Mécanique
	SEIGNEURIN Raymond	Microbiologie et Hygiène
	SENGEL Philippe	Zoologie
	SILBERT Robert	Mécanique des fluides
	SOUTIF Michel	Physique générale

MM.	TANCHE Maurice	Physiologie
	TRAYNARD Philippe	Chimie générale
	VAILLAND François	Zoologie
	VALENTIN Jacques	Physique nucléaire
	VAUQUOIS Bernard	Calcul électronique
Mme	VERAIN Alice	Pharmacie galénique
M.	VERAIN André	Physique
Mme	VEYRET Germaine	Géographie
MM.	VEYRET Paul	Géographie
	VIGNAIS Pierre	Biochimie médicale
	YOCCOZ Jean	Physique nucléaire théorique

PROFESSEURS ASSOCIES

MM.	BULLEMER Bernhard	Physique
	HANO JUN-ICHI	Mathématiques Pures
	STEPHENS Michaël	Mathématiques appliquées

PROFESSEURS SANS CHAIRE

MM.	BEAUDOING André	Pédiatrie
Mme	BERTRANDIAS Françoise	Mathématiques Pures
MM.	BERTRANDIAS Jean-Paul	Mathématiques appliquées
	BIAREZ Jean-Pierre	Mécanique
	BONNETAIN Lucien	Chimie minérale
Mme	BONNIER Jane	Chimie générale
MM.	CARLIER Georges	Biologie végétale
	COHEN Joseph	Electrotechnique
	COUMES André	Radioélectricité
	DEPASSEL Roger	Mécanique des fluides
	DEPORTES Charles	Chimie minérale
	GAUTHIER Yves	Sciences biologiques
	GAVEND Michel	Pharmacologie
	GERMAIN Jean-Pierre	Mécanique
	GIDON Paul	Géologie et Minéralogie
	GLENAT René	Chimie organique
	HACQUES Gérard	Calcul numérique
	JANIN Bernard	Géographie
Mme	KAHANE Josette	Physique
MM.	MULLER Jean-Michel	Thérapeutique
	PERRIAUX Jean-Jacques	Géologie et Minéralogie
	POULOUJADOFF Michel	Electrotechnique
	REBECQ Jacques	Biologie (CUS)
	REVOL Michel	Urologie
	REYMOND Jean-Charles	Chirurgie générale
	ROBERT André	Chimie papetière
	DE ROUGEMONT Jacques	Neurochirurgie
	SARRAZIN Roger	Anatomie et chirurgie
	SARROT-REYNAULD Jean	Géologie
	SIBILLE Robert	Construction mécanique
	SIROT Louis	Chirurgie générale
Mme	SOUTIF Jeanne	Physique générale

MAITRES DE CONFERENCES ET MAITRES DE CONFERENCES AGREGES

Mle	AGNIUS-DELORD Claudine	Physique pharmaceutique
	ALARY Josette	Chimie analytique
MM.	AMBLARD Pierre	Dermatologie
	AMBROISE-THOMAS Pierre	Parasitologie
	ARMAND Yves	Chimie
	BEGUIN Claude	Chimie organique
	BELORIZKY Elie	Physique
	BENZAKEN Claude	Mathématiques appliquées
	BILLET Jean	Géographie
	BLIMAN Samuel	Electronique (EIE)
	BLOCH Daniel	Electrotechnique
Mme	BOUCHE Liane	Mathématiques (CUS)
MM.	BOUCHET Yves	Anatomie
	BOUVARD Maurice	Mécanique des fluides
	BRODEAU François	Mathématiques (IUT B)
	BRUGEL Lucien	Energétique
	BUISSON Roger	Physique
	BUTEL Jean	Orthopédie
	CHAMBAZ Edmond	Biochimie médicale
	CHAMPETIER Jean	Anatomie et organogénèse
	CHIAVERINA Jean	Biologie appliquée (EFP)
	CHIBON Pierre	Biologie animale
	COHEN-ADDAD Jean-Pierre	Spectrométrie physique
	COLOMB Maurice	Biochimie médicale
	CONTE René	Physique
	COULOMB Max	Radiologie
	CROUZET Guy	Radiologie
	DURAND Francis	Métallurgie
	DUSSAUD René	Mathématiques (CUS)
Mme	ETERRADOSSI Jacqueline	Physiologie
MM.	FAURE Jacques	Médecine légale
	GENSAC Pierre	Botanique
	GIDON Maurice	Géologie
	GRIFFITHS Michaël	Mathématiques appliquées
	GROULADE Joseph	Biochimie médicale
	HOLLARD Daniel	Hématologie
	HUGONOT Robert	Hygiène et Médecine préventive
	IDELMAN Simon	Physiologie animale
	IVANES Marcel	Electricité
	JALBERT Pierre	Histologie
	JOLY Jean-René	Mathématiques Pures
	JOUBERT Jean-Claude	Physique du solide
	JULLIEN Pierre	Mathématiques Pures
	KAHANE André	Physique générale
	KUHN Gérard	Physique
	LACOUME Jean-Louis	Physique
Mme	LAJZEROWICZ Jeannine	Physique
MM.	LANCIA Roland	Physique atomique
	LE JUNTER Noël	Electronique
	LEROY Philippe	Mathématiques
	LOISEAUX Jean-Marie	Physique nucléaire
	LONGEQUEUE Jean-Pierre	Physique nucléaire
	LUU DUC Cuong	Chimie organique
	MACHE Régis	Physiologie végétale
	MAGNIN Robert	Hygiène et Médecine préventive
	MARECHAL Jean	Mécanique
	MARTIN-BOUYER Michel	Chimie (CUS)

MM.	MAYNARD Roger	Physique du solide
	MICHOULIER Jean	Physique (IUT A)
	MICOUD Max	Maladies infectieuses
	MOREAU René	Hydraulique (INP)
	NEGRE Robert	Mécanique
	PARAMELLE Bernard	Pneumologie
	PECCOUD François	Analyse (IUT B)
	PEFFEN René	Métallurgie
	PELMONT Jean	Physiologie animale
	PERRET Jean	Neurologie
	PERRIN Louis	Pathologie expérimentale
	PFISTER Jean-Claude	Physique du solide
	PHELIP Xavier	Rhumatologie
Mlle	RIERY Yvette	Biologie animale
MM.	RACHAIL Michel	Médecine interne
	RACINET Claude	Gynécologie et obstétrique
	RENAUD Maurice	Chimie
	RICHARD Lucien	Botanique
Mme	RINAUDO Marquerite	Chimie macromoléculaire
MM.	ROMIER Guy	Mathématiques (IUT B)
	SHOM Jean-Claude	Chimie générale
	STIEGLITZ Paul	Anesthésiologie
	STOEBNER Pierre	Anatomie pathologique
	VAN CUTSEM Bernard	Mathématiques appliquées
	VEILLON Gérard	Mathématiques appliquées (INP)
	VIALON Pierre	Géologie
	VOOG Robert	Médecine interne
	VROUSSOS Constantin	Radiologie
	ZADWORNY François	Electronique

MAITRES DE CONFERENCES ASSOCIES

MM.	BOUDOURIS Georges	Radioélectricité
	CHEEKE John	Thermodynamique
	GOLDSCHMIDT Hubert	Mathématiques
	SIDNEY STUARD	Mathématiques Pures
	YACOUD Mahmoud	Médecine légale

CHARGES DE FONCTIONS DE MAITRES DE CONFERENCES

Mme	BERIEL Hélène	Physilogie
Mme	RENAUDET Jacqueline	Microbiologie

Fait le 30 mai 1972.

Cette thèse a été réalisée sous la direction de Madame F. CHATELIN, Maître de Conférences à l'Université des Sciences Sociales de Grenoble. Je tiens à lui exprimer ma profonde gratitude pour son aide précieuse et ses encouragements.

Que Monsieur N. GASTINEL, Professeur à l'Université Scientifique et Médicale de Grenoble, reçoive ma plus vive reconnaissance pour avoir bien voulu présider ce jury.

A Monsieur F. ROBERT, Maître de Conférences à l'Université Scientifique et Médicale de Grenoble, j'adresse mes remerciements les plus sincères pour sa présence parmi ce même jury ; ainsi que pour l'intérêt qu'il a porté à mon travail.

Je voudrais également remercier Mademoiselle Cl. PAYERNE qui a réalisé la frappe du manuscrit dans un délai très court ainsi que tous ceux qui ont participé à la réalisation matérielle de cet ouvrage.

TABLE DES MATIERES

INTRODUCTION GENERALE

PREMIERE PARTIE

Partitionnement d'opérateurs

Introduction..... I.1

1. Définitions et notations..... I.2

2. Etude de la résolvante de T dans le s.e.v.E..... I.4

3. Propriétés de S(z)..... I.16

Bibliographie de la première partie..... I.25

DEUXIEME PARTIE

Applications du partitionnement

Introduction..... II.1

CHAPITRE I

Applications aux matrices hermitiennes..... II.2

Introduction..... II.2

1. Matrices hermitiennes : l'algorithme d'Abramov..... II.4

2. Etude de la convergence..... II.5

3. Accélération de la convergence..... II.14

4. Expériences numériques..... II.17

CHAPITRE II

Applications aux matrices tridiagonales..... II.24

1. Partitionnement successifs..... II.25

2. Comportement des algorithmes LR et QR avec translation II.30

3. Application à l'accélération de l'algorithme QR..... II.44

4. Application à la méthode des bissections..... II.67

5. Procédures Algol..... II.73

CHAPITRE III

Application au calcul de valeurs propres d'opérateurs compacts autoadjoints..... II.80

Introduction..... II.80
1. Notations et résultats immédiats..... II.82
2. Corrections des valeurs propres obtenues par la méthode de Galerkin..... II.85
3. Expériences numériques..... II.100
Bibliographie de la deuxième partie..... II.113

TROISIEME PARTIE

Sur le calcul des éléments propres de très grandes matrices

Introduction..... III.1
1. Approximation d'espaces invariants..... III.3
2. La méthode de Galerkin appliquée au calcul d'éléments propres d'opérateurs. Applications aux matrices..... III.9
3. La méthode de Rayleigh généralisée..... III.23
Bibliographie de la troisième partie..... III.63

BIBLIOGRAPHIE GENERALE

INTRODUCTION GENERALE

Nous nous intéressons au calcul des éléments propres d'opérateurs hermitiens par des méthodes utilisant le partitionnement.

La motivation principale de cette étude vient de la constatation pratique suivante : lorsqu'on calcule les élements propres d'un opérateur, ou d'une très grande matrice T , nous sommes souvent amenés pour des raisons liées à la capacité de la mémoire centrale de l'ordinateur, à restreindre T à un sous espace E de dimension inférieure n et a approcher les éléments propres de T par ceux de l'opérateur T_E correspondant.

Il est alors immédiat que deux points de vue se présentent :

1) Soit on considère que les éléments propres de T_E approchent bien ceux de T et on les accepte comme approximations des éléments propres de T.

2) Soit on cherche à améliorer par des corrections successives les éléments propres de T_E.

Dans un premier temps, nous nous sommes plus particulièrement intéressés au deuxième point de vue, le premier point étant la méthode des projections, largement étudiée dans la littérature.

Indépendamment du problème de la précision nous avons cherché à utiliser le partitionnement pour essayer de diminuer les temps d'exécution de certains algorithmes très classiques : alors que les méthodes de partitionnement ne donnent pas de très bons résultats du point de vue du temps d'exécution, on constate que lorsqu'on les combine avec d'autres algorithmes ils donnent lieu à des améliorations très nettes de ces algorithmes (algorithme QR et algorithme de bissection).

Par la suite, il nous a paru que pour les matrices d'un très grand ordre N, si on se place du premier point de vue, il n'est pas suffisant de partitionner une matrice A lorsque celle-ci est écrite dans la base canonique $e_1, e_2, \ldots, e_n, e_{n+1}, \ldots, e_N$, mais qu'il est nécessaire de bien choisir les n premiers éléments $V_1, V_2, \ldots, V_n$, de la base orthonormales dans laquelle A doit être écrite, puisque les valeurs propres approchées par la méthode de Galerkin sont celles de

$$T_E = [V_1, V_2, \ldots, V_n]^T A [V_1, V_2, \ldots, V_n].$$

Nous nous sommes intéressés au cas assez courant où l'on choisit pour (V_i) le système obtenu par orthogonalisation des vecteurs $X_o, AX_o, \ldots, A^{n-1}X_o$ où X_o est un vecteur initial. La question qui s'est posée à nous est la suivante :

Que peut-on dire de l'évolution des erreurs de méthode sur les valeurs propres en fonction de n ?

Pour la présentation de notre travail, nous avons proposé trois parties distinctes :

La première partie, d'ordre théorique, présente les fondements essentiels de la notion de partitionnement par l'intermédiaire de la théorie des perturbations analytiques : nous y verrons en particulier comment une valeur propre λ de T peut être considérée comme une racine de l'équation $\lambda_i(z) = 0$ où $\lambda_i(z)$ est une fonction méromorphe liée au partitionnement de T.

Dans la deuxième partie, nous verrons l'utilisation des résultats précédents d'une part pour le calcul des valeurs propres de matrices hermitiennes (en cherchant les racines de $\lambda_i(z) = 0$) et d'autre part pour l'amélioration des approximations de Galerkin d'opérateurs autoadjoints compacts (en approchant les équations $\lambda_i(z) = 0$).

Enfin dans la troisième partie, nous parlerons de calcul de valeurs propres de très grandes matrices par l'approximation de Galerkin. Nous y développerons le premier point de vue proposé ci-dessus.

PREMIERE PARTIE

P A R T I T I O N N E M E N T D ' O P E R A T E U R S

INTRODUCTION

Cette partie a pour but de définir la notion de partitionnement d'un opérateur dans un espace de Hilbert H, liée à la décomposition $H = E \oplus F$, et de grouper les notions théoriques essentielles utilisées par la suite. Etant donné la variété des applications il est indispensable d'introduire les définitions les plus générales possibles : ainsi nous parlerons d'opérateurs linéaires sur des espaces de Hilbert et nous relierons la notion de partitionnement d'opérateurs à celle de perturbations d'opérateurs.

Le résultat principal auquel nous aboutissons est le suivant :

On peut ramener le calcul des valeurs propres de l'opérateur T au calcul des zéros de certaines fonctions méromorphes dont on connaît la dérivée, ce qui permet d'appliquer l'algorithme de résolution de Newton.

Nous pouvons regrouper les méthodes d'approximation de valeurs propres d'opérateurs en deux méthodes types :

* Le premier type consiste, pour calculer les éléments propres de l'opérateur T dans H, à les approcher par ceux de la partie de T dans E, sous-espace vectoriel de dimension finie de H.

  Ce sont les méthodes de projections telles que celle de Galerkin, Rayleigh-Ritz,...

* Le deuxième type de méthode concerne non plus la résolution du problème des éléments propres de l'opérateur T mais le problème du calcul des éléments propres de la partie T_F de T dans F : on approche alors T_F par la partie de T dans un sur-espace F_n de F.

  C'est en particulier la méthode des problèmes intermédiaires proposée par Weinstein.

Pour le premier type de méthode auquel nous nous intéressons plus particulièrement, une valeur propre exacte λ de T, peut être aussi considérée comme une valeur propre exacte de $T_E+H(\lambda)$, où $H(\lambda)$ qui est défini lorsque λ n'est pas dans le spectre de T_F, est une "perturbation" de T_E dépendant de la valeur propre λ .

Si l'on sait estimer la perturbation $H(\lambda)$, cette remarque permet de corriger la valeur propre de T_E considérée comme valeur propre approchée de T. Nous en verrons des applications dans la deuxième partie.

1. DEFINITIONS ET NOTATIONS

1.1.

+ Dans toute cette partie on considère un espace de Hilbert H séparable, sur $\mathbb{C}$.

Soit T un opérateur linéaire de H dans H défini sur le sous-espace vectoriel $\mathcal{D}(T)$ appelé domaine de T

+ Soit la décomposition en somme directe $H = E \oplus F$, où E et F sont deux sous-espaces vectoriels orthogonaux de H, et où E est de dimension finie.

On appelle . π la projection orthogonale de H, avec $\mathrm{Im}(\pi) = E$.

. π_E l'application de H dans E qui à $x \in H$, associe πx considéré comme élément de E.

+ On dénote par $T|_E$ la restriction de T au sous-espace E :

$T|_E$ est donc l'application de E dans H qui à tout $x \in \mathcal{D}(T) \cap E$ associe $Tx \in H$.

On peut définir π_F et T_F de la même manière, et ceci permet de considérer les opérateurs suivants :

$T_E = \pi_E T|_E$ de $\mathcal{D}(T_E) = \mathcal{D}(T) \cap E$ dans E

$T_F = \pi_F T|_F$ de $\mathcal{D}(T_F) = \mathcal{D}(T) \cap F$ dans F

$U = \pi_E T|_F$ de $\mathcal{D}(U) = \mathcal{D}(T) \cap F$ dans E

$V = \pi_F T|_E$ de $\mathcal{D}(V) = \mathcal{D}(T) \cap E$ dans F

DEFINITION :

On dit que l'on a partitionné T, selon la décomposition $H = E \oplus F$, ce qu'on note schématiquement ainsi :

$$T = \left(\begin{array}{c|c} T_E & U \\ \hline V & T_F \end{array}\right)$$

T_E (respectivement : T_F) est appelé la partie de T dans E (respectivement : dans F) .

Pour la définition de T_E et T_F nous avons suivi ici la terminologie employée par Weinstein, Steinger [2] , Gould [4] et Kato [1].

1.2. RAPPELS SUR LES OPERATEURS [1,3,8,9]

. Nous dénoterons par 1_H, 1_F, 1_E ect... les applications identité sur H, F, E respectivement.

. Soit $z \in \mathbb{C}$, si $T - z1_H$ est injectif alors l'inverse $(T - z1_H)^{-1}$ défini sur $\mathrm{Im}(T-z1_H) = (T-z1_H)\mathcal{D}(T)$ est appelé la résolvante de T en z .

. $\rho(T) = \{z \in \mathbb{C} \ ; \ (T-z1_H)$ est injective et d'image dense dans $H\}$

$= \{z \in \mathbb{C} \ ; \ R(z)$ existe et $\overline{\mathcal{D}(R(z))} = H\}$

est l'ensemble résolvant de T. Son complémentaire dans $\mathbb{C}$, noté $\sigma(T)$, est le spectre de T.

. T est fermé si pour toute suite x_n de $\mathcal{D}(T)$ vérifiant :

i) x_n converge vers x

ii) Tx_n converge vers y

alors $x \in \mathcal{D}(T)$ et $y = Tx$.

. $\mathcal{L}(H)$ dénote l'ensemble des opérateurs bornés de H dans H et tels que $\mathcal{D}(T) = H$

. Si T est fermé alors :

$\rho(T) = \{z \in \mathbb{C} \ ; \ R(z) \in \mathcal{L}(H)\}$.

. Lorsque $\mathcal{D}(T_E) = E$ (i.e. lorsque $\mathcal{D}(T) \supset E$), on appelle, déterminant de T_E le déterminant de la matrice de terme général $(T_{p_j,p_i})_{i,j=1,\ldots,n}$ où $(p_i)_{i=1,\ldots,n}$ forme une base orthonormale quelconque de E.

2. ETUDE DE LA PARTIE DE LA RESOLVANTE DE T DANS LE SOUS-ESPACE VECTORIEL E.

On va présenter dans ce paragraphe des résultats concernant la partie de R(z) dans E. Nous verrons qu'il est possible de caractériser, d'une part, les valeurs propres de T_F par l'équation $\det(R(z)_E) = 0$ et d'autre part, les valeurs propres de T, par l'équation $\det(R(z)_E)^{-1}) = 0$.

Cette dernière équation nous conduira à chercher une expression théorique de $(R(z)_E)^{-1}$ et nous aboutirons alors au résultat suivant :

Les valeurs propres de T n'appartenant pas au spectre de T_F sont caractérisées par l'équation :

$$\det(T_E - z1_E - U(T_F - z1_F)^{-1}V) = 0$$

qui peut être considérée comme une perturbation de l'équation :

$$\det(T_E - z1_E) = 0$$

cette dernière donnant les valeurs propres de T_E .

DEFINITION :

1) Lorsque T_F est fermé, on note, pour $z \notin \sigma(T_F)$,

$$S(z) = T_E - z1_E - U(T_F - z1_F)^{-1}V$$

2) On appelle fonction de Weinstein la fonction

$W(z) = \det(R(z)_E)$ définie pour $z \in \rho(T)$.

PROPOSITION 1 :

On suppose que T et T_F sont fermés et que $\mathcal{D}(T) \supset E$.
Soit $z \in \rho(T) \cap \rho(T_F)$. Alors :

1) $W(z) \neq 0$, autrement dit $R(z)_E$ est inversible.

2) $(R(z)_E)^{-1} = S(z)$.

DEMONSTRATION :

Nous pouvons nous restreindre à faire la démonstration pour $z = 0$ sans rien changer à la généralité du résultat. Comme T et T_F sont supposés fermés,

$$0 \in \rho(T) \cap \rho(T_F)$$

implique que T et T_F sont inversibles au sens suivant :

Il existe un opérateur T^{-1} de domaine H tel que

$$T^{-1}T = 1_{\mathcal{D}(T)} \quad \text{et} \quad TT^{-1} = 1_H$$

Il existe un opérateur $(T_F)^{-1}$ de domaine F tel que

$$(T_F)^{-1}T_F = 1_{\mathcal{D}(T_F)} \quad \text{et} \quad T_F(T_F)^{-1} = 1_F \ .$$

Partitionnons T^{-1} de la même manière que T :

$$T^{-1} = \left(\begin{array}{c|c} T'_{11} & T'_{12} \\ \hline T'_{21} & T'_{22} \end{array}\right)$$

avec $\quad T'_{11} = (T^{-1})E \qquad T'_{21} = \pi_F T^{-1}|_E$

$$T'_{12} = \pi_E T^{-1}|_F \qquad T'_{22} = \pi_F T^{-1}|_F$$

T^{-1} étant de domaine $\mathcal{D}(T^{-1}) = H$, les domaines de T'_{11}, T'_{12}, T'_{21}, T'_{22} sont E, F, E et F respectivement.

On a alors :

$$\begin{cases} T'_{11}\, T_E + T'_{12}\, V = 1_E & (1) \\ T'_{11}\, U + T'_{12}\, T_F = 0_{\mathcal{D}(T_{22}),E} & (2) \end{cases}$$

où 1_E est l'identité de E dans E et où $0_{\mathcal{D}(T_{22}),E}$ est l'application nulle de $\mathcal{D}(T_{22})$ dans E.

En effet, démontrons (1) par exemple :

$$\pi_E T^{-1}|_E \pi_E T|_E + \pi_E T^{-1}|_F \pi_F T|_E = \pi_E T^{-1} \pi T|_E + \pi_E T^{-1}(1_H - \pi)T|_E$$

$$= \pi_E T^{-1}(\pi T_E + (1_H - \pi)T|_E) = \pi_E T^{-1}(\pi + (1_H - \pi))T|_E$$

$$= \pi_E 1_H|_E = 1E$$

(2) se démontre de manière identique.

Comme T_F est supposé inversible, de (2) on peut tirer :

$$T'_{12} = -T'_{11} U (T_F)^{-1} , \qquad (3)$$

opérateur de F dans E.

La composition de ces trois opérateurs a un sens :

$$F \xrightarrow{(T_F)^{-1}} \mathcal{D}(T)\cap F \xrightarrow{U} E \xrightarrow{T'_{11}} E$$

$$\mathcal{D}(U) \qquad \mathcal{D}(T'_{11})$$

(Ceci est dû en fait que $(T_F)^{-1}H = \mathcal{D}(T_F)$) .

L'expression (3) portée dans (1) donne :

$$T'_{11}(T_E - U(T_F)^{-1}V) = 1_E$$

ce qui montre que T'_{11} est inversible et est d'inverse $T_E - U(T_F)^{-1}V$.

C.Q.F.D.

Cette proposition va nous permettre de caractériser, d'une part l'ensemble $\sigma(T_F) \setminus \sigma(T)$ en utilisant la fonction de Weinstein W(z), et d'autre part l'ensemble $\sigma(T) \setminus \sigma(T_F)$ en utilisant S(z).

Ces deux caractérisations sont utiles car en général il est rare qu'une valeur propre de T soit également une valeur propre de T_F. —

. CARACTERISATION DE $\sigma(T_F) \setminus \sigma(T)$:

Grâce à la proposition 1 on voit que si $z \in \rho(T)$ alors si W(z) est nul, z ne peut pas appartenir à $\rho(T_F)$ (puisque sinon $z \in \rho(T_F) \cap \rho(T)$ et donc d'après la proposition $W(z) \neq 0$).

On vient donc de montrer que $z \in \rho(T)$ et $W(z) = 0 \Rightarrow z \in \sigma(T_F)$

De plus nous pouvons montrer la réciproque, c'est-à-dire que finalement nous avons le :

COROLLAIRE 2 :

T et T_F sont supposés fermés et tels que $\mathcal{D}(T) \supset E$.

Soit $z \in \rho(T)$. Alors :

z appartient au spectre de T_F si et seulement si $W(z) = 0$.

DEMONSTRATION :

D'après les remarques précédentes, il reste à montrer seulement: si $z \in \sigma(T_F)$ alors $W(z) = 0$.

Raisonnons par l'absurde : supposons que $W(z) \neq 0$
$z \in \rho(T) \Rightarrow R(z)$ est borné de domaine H. (Car T est fermé).
Partitionnons R(z) selon la décomposition $H = E \oplus F$:

$$R(z) = \left(\begin{array}{c|c} T'_{11} & T'_{12} \\ \hline T'_{21} & T'_{22} \end{array} \right)$$

et de même :

$$T - z1_H = \left(\begin{array}{c|c} T_{11} & T_{12} \\ \hline T_{21} & T_{22} \end{array} \right)$$

De la même manière que pour la démonstration de la proposition 1 il est facile de démontrer que :

$$T_{21}T'_{12} + T_{22}T'_{22} = 1_F \qquad (1)$$

$$T_{21}T'_{11} + T_{22}T'_{21} = 0_{E,F} \qquad (2)$$

et comme on suppose que $W(z) = \det(T'_{11}) \neq 0$, T'_{11} est inversible, et de (2) on tire que :

$$T_{21} = -T_{22}T'_{21}(T'_{11})^{-1} \qquad (3)$$

La justification de cette composition d'opérateurs se fait comme pour la proposition 1. Portons (3) dans (1) :

$$T_{22}(T'_{22}-T'_{21}(T'_{11})^{-1}T'_{12}) = 1_F \qquad (4)$$

Pour conclure notons qu'il est possible de démontrer de la même manière que

$$(T'_{22}-T'_{21}(T'_{11})^{-1}T'_{12})T_{22} = 1_F \qquad (5)$$

en partant des égalités :

$$T'_{21}T_{12} + T'_{22}T_{22} = 1 \quad \mathcal{D}(T_{22})$$

$$T'_{11}T_{12} + T'_{12}T_{22} = 0 \quad \mathcal{D}(T_{22}),E$$

(4) et (5) montrent alors que T_{22} est inversible et son inverse

$$(T_{22})^{-1} = T'_{22}-T'_{21}(T'_{11})^{-1}T'_{12}$$

est de domaine F.

T_{22} étant fermé, $(T_{22})^{-1}$ l'est aussi [8].

$(T_{22})^{-1}$ étant fermé et de domaine $\mathcal{D}(T_{22})^{-1} = F$ est donc borné (théorème du graphe fermé).

Donc : $z \in \rho(T_F)$.

Ce qui contredit l'hypothèse que $z \in \sigma(T_F)$.

C.Q.F.D.

. CARACTERISATION DE $\sigma(T) \setminus \sigma(T_F)$

COROLLAIRE 3 :

On suppose que T et T_F sont fermés et que $\mathcal{D}(T) \supset E$

Soit $z \notin \sigma(T_F)$. Alors :

z appartient au spectre de T si et seulement si $\det(S(z)) = 0$.

DEMONSTRATION :

← Raisonnons par l'absurde : Supposons que $\det(S(z)) = 0$ et que $z \in \rho(T)$. D'après la proposition 1 $z \in \rho(T) \cap \rho(T_F)$ entraîne que $R(z)_E$ est inversible et d'inverse $S(z)$,
$\det(Z(z))$ et donc non nul.

→ On raisonne encore par l'absurde : Supposons que $\det(S(z)) \neq 0$
Alors $(T-z1_H)$ est inversible et son inverse est l'opérateur suivant :

$$R(z) = \left[\begin{array}{c|c} (S(z))^{-1} & (S(z))^{-1} \cup (T_F-z1_F)^{-1} \\ \hline -(T_F-z1_F)^{-1}V(S(z))^{-1} & (T_F-z1_F)^{-1}(1_F+V(S(z))^{-1}U(T_F-z1_F)^{-1}) \end{array}\right]$$

On peut en effet vérifier que cet opérateur est bien $R(z)$ en composant à droite et à gauche avec $T-z1_H$ et en constatant que $R(z)(T-z1_H) = 1_{\mathcal{D}(T)}$ et $(T-z1_H)\,R(z) = 1_H$.

D'autre part le domaine de $R(z)$ est H (vérification immédiate).

- $R(z)$ est fermé (inverse d'un opérateur fermé)
- $R(z)$ est de domaine H.

Donc d'après le théorème du graphe fermé, $R(z)$ est borné.
i.e. $R(z) \in \mathcal{L}(H)$.
Finalement $z \in \rho(T)$ ce qui contredit l'hypothèse de départ $z \in \sigma(T)$.

C.Q.F.D.

Avant de discuter de l'utilisation de ces deux corollaires, nous allons donner une autre forme plus précise du corollaire 3. Nous venons de caractériser $\sigma(T) \setminus \sigma(T_F)$.
Puisque E est de dimension finie et puisque T s'écrit :

$$T = (1-\pi)T(1-\pi)+(T\pi+\pi T(1-\pi)) ,$$

T peut être considéré comme le perturbé de $(1-\pi)T(1-\pi)$ par une perturbation de rang fini et on peut se demander si $\sigma(T) \setminus \sigma(T_F)$ n'est pas constitué uniquement de valeurs propres de T.

La réponse est donnée par la proposition et le corollaire qui suivent :

PROPOSITION 4 :

Supposons que T_F est fermé et que $\mathcal{D}(T) \supset E$.
Soit $\lambda \in \rho(T_F)$. Alors :

λ est une valeur propre de T si et seulement si $\det(S(\lambda)) = 0$
De plus la multiplicité géométrique de λ est égale au corang de $S(\lambda)$.

En comparant la proposition 4 au corollaire 3 il vient :

COROLLAIRE 5 :

Supposons que T et T_F sont fermés et que $\mathcal{D}(T) \supset E$.
Alors le spectre de T est constitué uniquement du spectre de T_F plus un nombre fini de valeurs propres de multiplicités géométriques finies.

DEMONSTRATION DE LA PROPOSITION 4 :

1) Supposons que λ est valeur propre de T. Alors $\exists X \in H \quad X \neq 0$ tel que :

$$(T-\lambda 1_H)X = 0$$

Si on pose

$$\pi_E X = x \quad \text{et} \quad \pi_F X = y$$

alors

$$(T-\lambda 1_H)X = 0$$

s'écrit :

$$\begin{cases} (T_E-\lambda 1_E)x + Uy = 0 & (1) \\ Vx + (T_F-\lambda 1_F)y = 0 & (2) \end{cases}$$

Comme $\lambda \in \rho(T_F)$ de (2) on tire $y = -(T_F-\lambda 1_F)^{-1}Vx$ et (1) donne alors :

$$(T_E-\lambda 1_E-U(T_F-\lambda 1_F)^{-1}V)x = 0$$

i.e. $S(\lambda)x = 0$.

De plus, $x \neq 0$ (car sinon $y = 0$ et donc $X = 0$) donc $\det(S(\lambda)) = 0$.

2) Supposons que $\det(S(\lambda)) = 0$, alors il existe un $x \neq 0$ tel que $S(\lambda)x = 0$. Si on pose encore $y = -(T_F-\lambda 1_F)^{-1}Vx$, on constate que le vecteur X tel que $\pi_E X = x$ et $\pi_F X = y$ vérifie les équations (1) et (2) ci-dessus donc λ est valeur propre de T et X est un vecteur propre associé.

C.Q.F.D.

APPLICATIONS

Le corollaires 2 et 3 permettent de résoudre le problème de l'approximation des valeurs propres d'un opérateur de deux façons différentes

1°) On peut approcher les valeurs propres de T parcelles de T_E. D'après le corollaire 3, il suffit de calculer les zéros de la fonction $\det(S(z))$, dans une région Δ, ne contenant que des valeurs propres isolées de T et aucun point du spectre de T_F, pour avoir les valeurs propres de T situées dans Δ. Cela sera étudié en détail dans la deuxième partie.

2°) On peut approcher les valeurs propres de T_F par celles de T : D'après le corollaire 2, un procédé de calcul de toutes les valeurs propres de T_F situées dans une région Δ de $\mathbb{C}$ qui ne contient que des valeurs propres isolées de T_F et telle que $\Delta \cap \sigma(T) = \emptyset$, consiste à trouver les zéros de la fonction W(z), situés dans Δ. C'est ce procédé qui est utilisé dans la méthode de Weinstein appelée méthode des problèmes intermédiaires de première espèce.

Nous allons décrire très brièvement cette méthode.
Pour une description complète voir [2,4].

METHODE DE PROBLEMES INTERMEDIAIRES DE PREMIERE ESPECE

On se donne un opérateur compact autoadjoint T sur H. On suppose pour simplifier que T est défini positif et on ordonne ses valeurs propres en décroissant. On se donne un sous-espace vectoriel F de H.

On veut résoudre le problème des éléments propres de T_F dans son domaine :
i.e. on cherche $\lambda \in \mathbb{C}$ et $u \in \mathcal{D}(T_F)$ tels que :

$$(1) \qquad T_F u = \lambda u$$

La méthode de Weinstein (développée en 1935) a pour but de donner des bornes supérieures de valeurs propres de T_F et de complèter ainsi les bornes inférieures données par la méthode de Rayleigh-Ritz.

L'intérêt qu'ont porté les physiciens à cette méthode est dû principalement au fait que l'on peut donc encadrer une valeur propre λ_i de T_F, ainsi :

$$\lambda_i \text{ (Rayleigh-Ritz)} \leq \lambda_i \leq \lambda_i \text{ (Weinstein)}$$

En général on applique la méthode de Weinstein à des opérateurs d'inverse compact et c'est de cet inverse que l'on parle ici. Cependant cet inverse n'est pas utilisé explicitement dans les calculs; Cf [2,4].

T est appelé opérateur de base et le problème

$$(2) \qquad Tu = \lambda u \qquad\qquad u \in \mathcal{D}(T)$$

est le problème de base.

Toute la difficulté de la méthode réside dans le choix de ce problème de base : car ce qui est posé, c'est le problème à résoudre (1) et il peut y avoir plusieurs choix possibles pour le problème (2), la difficulté de la résolution du problème dépendant de ce choix.

On fait l'hypothèse fondamentale suivante :
Le problème de base (2) est entièrement résolu i.e. on connait toutes les valeurs propres et tous les vecteurs propres de (2).

Soit E le sous-espace vectoriel supplémentaire orthogonal de F. Ici E n'est plus supposé de dimension finie : soient alors $p_1, p_2, \ldots p_n, \ldots$ une famille infinie d'éléments libres de E. On dénote par E_n le sous espace vectoriel engendré par $(p_1, p_2, \ldots, p_n)$ et par π_n la projection orthogonale sur E_n et Q_n la projection orthogonal sur $F_n = E_n^{\perp}$

$$\pi_E, \pi_{E_n}, Q_F, Q_{F_n}, T_E, \text{ etc...}$$

sont définis comme en 1.

La méthode consiste à approcher le problème (1) par la suite de problèmes suivants :

$$(3) \qquad T_{F_n} u = \lambda u \qquad u \in \mathcal{D}(T_{F_n})$$

Le problème (3) est appelé le $n^{\text{ième}}$ problème intermédiaire.

CONVERGENCE :

Les projections Q_n vérifiant $Q_n x \to Qx \quad \forall x \in H$,
On peut montrer, comme en [3], la convergence uniforme de $Q_n T Q_n$ vers QTQ et cela permet de conclure, comme en [3] , qu'il y a convergence des valeurs propres $\lambda_i^{(n)}$ de (3) vers les valeurs propres λ_i de (1).
De plus le principe du mini-max montre que :

$$\lambda_i \leq \lambda_i^{(n+p)} \leq \lambda_i^{(n)} \qquad \forall n, \forall p$$

On a donc pour cette méthode une convergence des $\lambda_i^{(n)}$ par valeurs supérieur

RESOLUTION DES PROBLEMES INTERMEDIAIRES

Si λ est simultanément une valeur propre du problème de base (2) et du $n^{ième}$ problème intermédiaire (3), on dit que λ est une valeur propre persistante de (3) : c'est évidemment un cas particulier .

On a le théorème suivant [2 , p. 272].

THEOREME :

> λ **est** une valeur propre non persistante de (3) si et seulement si $W(\lambda) = 0$. De plus la multiplicité de λ est égale au corang de $R(\lambda)_E$.

On reconnait ici le corollaire 2 avec des hypothèses plus fortes (T compact autoadjoint défini positif).

Le théorème ci-dessus ignore les valeurs propres persistantes : il existe un procédé qui permet de savoir si une valeur propre du problème de base (2) est persistante. (Ce procédé est basé sur la définition de $W(z) = \det(R(z)_E)$ même lorsque z est valeur propre de T).

La méthode employée par Weinstein consiste, d'abord à chercher parmi les valeurs propres du problème de base celles qui sont persistantes, puis à calculer les zéros de la fonction W(z) qui ne sont pas des valeurs propres de (2) : Les solutions obtenues sont les valeurs propres non persistantes du problème (3) d'après le théorème.

———— X ————

3. PROPRIETES DE S(z)

Nous allons dans ce paragraphe revenir aux applications du corollaire 3 : dans toute la suite nous nous intéresserons plus particulièrement aux méthodes de projections, et nous aurons besoin d'une étude détaillée de l'application $z \to S(z)$.

Nous commençons par montrer que l'application qui à z, appartenant à l'ensemble résolvant de T_F, associe S(z) est une application holomorphe dans $\rho(T_F)$.

Nous en déduirons que, sous certaines conditions [1 pp.63 à 126] il existe localement des fonctions holomorphes $z \to \lambda_i(z)$, où $\{\lambda_i(z)\}$ sont les valeurs propres de S(z).

Alors, puisque $\det(S(z)) = \prod_{i=1}^{n} \lambda_i(z)$ il revient au même de chercher les zéros de det(S(z)) et les zéros des fonctions $\lambda_i(z)$.

3.1. ETUDE DE $z \to S(z)$

PROPOSITION 6 :

Supposons que T_F est fermé et que $\mathcal{D}(T) \supset E$.

Alors $z \to S(z)$ est holomorphe dans $\rho(T_F)$.

La démonstration découle immédiatement des deux lemmes qui suivent :

LEMME 7 :

Sous les hypothèses de la proposition ci-dessus et pour tout x et tout $y \in E$, l'application $z \to (S(z)x,y)$ est holomorphe dans tout domaine Δ ne contenant aucun élément de $\rho(T_F)$.

DEMONSTRATION DU LEMME 2.2. :

Puisque T_F est fermé,

Pour $|z-z_o| < \dfrac{1}{\|(T_F-z_o)^{-1}\|}$ on peut écrire, grâce au développement de la résolvante de T_F autour de z_o :

$$(S(z)x,y) = ((T_E-z_o1_E)x,y)-(z-z_o)(x,y)-(U(T_F-z_o1_F)^{-1}Vx,y)$$
$$-(U(\sum_{n=1}^{\infty}(z-z_o)^n(T_F-z_o)^{-n-1}Vx),y) \qquad (1)$$

Mais U est borné car il est de rang fini, donc

$$U(\sum_{n=1}^{N}(z-z_o)^n(T_F-z_o1_F)^{-n-1}Vx = \sum_{n=1}^{N}(z-z_o)^nU(T_F-z_o1_F)^{-n-1}Vx,$$

a pour limite lorsque N tend vers l'infini :

$$\sum_{n=1}^{\infty}(z-z_o)^n\, U\,(T_F-z_o1_F)^{-n-1}Vx$$

et comme le produit scalaire est continu, le dernier terme de la somme (1) est égal à :

$$\sum_{n=1}^{\infty}(z-z_o)^n(U(T_F-z_o)^{-n-1}Vx,y)$$

finalement nous avons le développement analytique suivant au voisinage de tout point $z_o \in \rho(T_F)$:

$$(S(z)x,y) = (S(z_o)x,y)-(z-z_o)(x,y)$$
$$-\sum_{n=1}^{\infty}(z-z_o)^n(U(T_F-z_o1_F)^{-n-1}Vx,y)$$

C.Q.F.D.

LEMME 8 [I,p. 152] :

Soient X et Y deux espaces de Banach et soit A(z) une famille d'opérateurs bornés de X dans Y, définie pour z appartenant à un domaine Δ du plan complexe. Supposons que $< A(z)x,y >$ soit holomorphe en z dans Δ pour tout $x \in X$, $y \in Y^*$ espace adjoint de Y.
Alors A(z) est holomorphe en z dans Δ au sens de la norme.
(i.e. A(z) est fortement différentiable sur Δ).

La démonstration se trouve en [1, p. 152].

On peut alors montrer facilement la proposition 6 :

DEMONSTRATION DE LA PROPOSITION 6 :

La famille S(z) est bornée pour tout $z \in \Delta = \rho(T_F)$ parce que S(z) est un opérateur de E dans E et E est de dimension finie.
D'autre part grâce au lemme 7 nous sommes exactement sous les hypothèses du lemme 8 et le résultat en découle.

C.Q.F.D.

Grâce à la proposition 6 on peut considérer qu'au voisinage d'un point z_o de $\rho(T_F)$, S(z) est une perturbation holomorphe de $S(z_o)$.

Comme nous l'avons signalé en introduction, il revient au même de chercher les zéros de la fonction det(S(z)) ou les zéros des n fonctions $\lambda_i(z)$ où $\{\lambda_i(z)\}$ sont les valeurs propres de S(z). Plus exactement puisque

$$\det(S(z)) = \prod_{i=1}^{n} (\lambda_i(z)) ,$$

si pour un indice i quelconque $\lambda_i(z) = 0$ alors $\det(S(z)) = 0$ et réciproquement si $\det(S(z)) = 0$, alors une au moins des valeurs propres $\lambda_i(z)$ est nulle. On peut donc écrire d'après le corollaire 3 que :

$\mu \in \rho(T_F)$ est une valeur propre de T si et seulement si il existe une valeur propre $\lambda_i(\mu)$ de $S(\mu)$ qui est nulle.

Ceci nous conduit naturellement à étudier les fonctions $\lambda_i(z)$.

3.2. ETUDE DES VALEURS PROPRES $\lambda_i(z)$ DE $S(z)$

Soit $\lambda(z_o)$ une valeur propre de $S(z_o)$ de multiplicité algébrique m. Si on considère un disque D dont le contour Γ entroure $\lambda(z_o)$ et l'isole du reste du spectre, alors il y a toujours m valeurs propres de S(z) dans D, comptées avec leurs ordres de multiplicité, lorsque z appartient à un voisinage $\mathcal{V}$ de z_o .

Ceci est dû au fait que $z \to P(z) = \frac{-1}{2i\pi} \int_\Gamma (S(z)-\eta 1_E)^{-1} d\eta$

est holomorphe dans $\mathcal{V}$ et donc dim P(z)E = dim $P(z_o)$E = m Cf [1].

En fait on sait [KATO : 1] qu'il existe s fonctions distinctes notées $\lambda_1(z)$, $\lambda_2(z)$, .., $\lambda_s(z)$ $(s \leq m)$, où chaque $\lambda_i(z)$ est une valeur propre de S(z), et qui sont en général des branches de fonctions analytiques lorsque $m \neq 1$.
Si m = 1 il existe évidemment une seule fonction $\lambda(z)$ et cette fonction est analytique.

Cela veut dire que les fonctios $\lambda_i(z)$ sont en général holomorphes sauf peut-être en des points où S(z) admet des valeurs propres multiples (plus généralement en des points appelés points exceptionnels en [1])

Cependant, dans le cas particulier où S(x), $x \in \mathbb{R} \setminus \sigma(T_F)$ forme une famille d'opérateurs autoadjoints, (par exemple lorsque T est autoadjoint) alors on démontre que les fonctions $\lambda_i(x)$ sont toujours analytiques.

Nous nous intéresserons en fait à ce seul cas par la suite. Cependant pour simplifier l'exposition des résultats et également pour montrer qu'en réalité la généralisation est possible lorsque T n'est pas autoadjoint, nous commençons par étudier le cas où $\lambda(z_o)$ est une valeur propre simple de $S(z_o)$ (et où T est quelconque).

1°) CAS OU $\lambda(z_o)$ EST UNE VALEUR PROPRE SIMPLE DE $S(z_o)$

Dans ce cas, nous avons vu que pour tout z dans $\mathcal{V}$, il existe une (seule) valeur propre $\lambda(z)$ de $S(z)$ dans D.
De plus la fonction $z \to \lambda(z)$ est analytique dans $\mathcal{V}$.

Soit φ_o un vecteur propre de $S(z_o)$ associé à la valeur propre simple $\lambda(z_o)$ et soit Ψ_o un vecteur propre de $S(z_o)^*$ associé à la valeur propre $\bar{\lambda}(z_o)$.

φ_o est appelé vecteur propre à droite de $S(z_o)$ et

Ψ_o vecteur propre à gauche.

On sait (Wilkinson : [5]) que $(\varphi_o, \Psi_o) \neq 0$.

Nous voulons d'une part montrer l'existence d'une famille holomorphe de vecteurs propres à droite et à gauche de $S(z)$, associés à $\lambda(z)$, et d'autre part utiliser cette famille pour trouver l'expression de la dérivée de $\lambda(z)$ en un point quelconque de $\mathcal{V}$.

Nous utiliserons le lemme suivant :

LEMME 9 :

Il existe $\varepsilon > 0$ et un voisinage $\mathcal{V}_o$ de z_o tel que :

$\forall z \in \mathcal{V}_o$, $|(P(z)\varphi_o, \Psi_o)| \geq \varepsilon > 0$

DEMONSTRATION :

φ_o est un vecteur propre de $S(z_o)$ associé à $\lambda(z_o)$, donc

$(P(z_o)\varphi_o, \Psi_o) = (\varphi_o, \Psi_o) \neq 0$

et en raison de l'holomorphie de $P(z)$ en z_o l'application

$z \to |(P(z)\varphi_o, \Psi_o)|$

et continue en z_o et il existe donc un voisinage $\mathcal{V}_o$ de z_o tel que :

$\forall z \in \mathcal{V}_o \qquad |(P(z)\varphi_o, \Psi_o)| \geq \frac{1}{2} |(\varphi_o, \Psi_o)| = \varepsilon$.

C.Q.F.D.

PROPOSITION 10 :

T et T_F sont supposés fermés, $\mathcal{D}(T) \supset E$.

Soit $z_o \notin \sigma(T_F)$ et supposons que $\lambda(z_o)$ est une valeur propre simple de $S(z_o)$. Alors il existe un voisinage $\mathcal{V}$ de z_o contenu dans $\rho(T_F)$ et dans lequel :

i) Il existe une fonction holomorphe $z \to \lambda(z)$ où $\lambda(z)$ est une valeur propre de $S(z)$.

ii) Il existe deux fonctions vectorielles $z \to \varphi(z)$ et $z \to \Psi(z)$ holomorphes dans un voisinage $\mathcal{V}_o \subset \mathcal{V}$ de z_o où $\varphi(z)$ et $\Psi(z)$ sont des vecteurs propres à droite et à gauche respectivement de $S(z)$.

iii) La dérivée de la fonction $\lambda(z)$ au point $z \in \mathcal{V}$ est :

$$\lambda'(z) = -1 - \frac{(U(T_F - z1_F)^{-2}V\varphi(z), \Psi(z))}{(\varphi(z), \Psi(z))}$$

iv) Si pour $\mu \in \mathcal{V}$, $\lambda(\mu) = 0$ alors μ est une valeur propre de T .

DEMONSTRATION :

i) et iv) découlent des remarques précédentes. Il reste à montrer ii) et iii).

ii) Il suffit de poser . $\varphi(z) = \dfrac{P(z)\varphi_o}{(P(z)\varphi_o, \Psi_o)}$

. $\Psi(z) = P^*(z)\Psi$

. On constate que $\varphi(z_o) = \dfrac{\varphi_o}{(\varphi_o,\Psi_o)}$ et que $\Psi(z_o) = \Psi_o$

$\Psi(z)$ est holomorphe car $P^*(z)$ l'est et de même

$\varphi(z)$ est holomorphe dans le voisinage $\mathcal{V}_o$ défini par le lemme 9 car $|(P(z)\varphi_o,\Psi_o)| \geq \varepsilon > 0$.

$\varphi(z)$ et $\Psi(z)$ sont évidemment des vecteurs propres à droite et à gauche respectivement de $S(z)$ puisque la multiplicité de $\lambda(z)$ est supposée égale à 1.

De plus , on a : $\forall z \in \mathcal{V}_o$.

$$. \quad (\varphi(z),\Psi(z)) = \frac{(P(z)\varphi_o,P^*(z)\Psi_o)}{(P(z)\varphi_o,\Psi_o)} = \frac{(P^2(z)\varphi_o,\Psi_o)}{(P(z)\varphi_o,\Psi_o)} = 1$$

$$. \quad (\varphi(z),\Psi_o) = 1$$

iii) Il suffit de faire la démonstration au point z_o puisque si $z \in \mathcal{V}$, $z \neq z_o$, la proposition est valable en remplaçant z par z_o .

La démonstration se fait en développant de deux manières différentes l'expression :

$$(S(z)\varphi(z),\Psi_o)-(S(z_o)\varphi(z),\Psi_o)$$

. D'une part c'est égal à :

$$(S(z)\varphi(z),\Psi_o)-(\varphi(z),S^*(z_o)\Psi_o) = (\lambda(z)-\lambda(z_o))(\varphi(z),\Psi_o)$$

or d'après la remarque ci-dessus $(\varphi(z),\Psi_o) = 1$.
Donc c'est égal à $\lambda(z)-\lambda(z_o)$.

. D'autre part c'est aussi $((S(z)-S(z_o))\varphi(z),\Psi_o)$

Or on a :

$$S(z)-S(z_o) = -(z-z_o)1_E-(z-z_o)\, U\,(T_F-z1_F)^{-1}(T_F-z_o1_F)^{-1}V$$

D'où :

$$\frac{\lambda(z)-\lambda(z_o)}{z-z_o} = -\,1\,-\,(U(T_F-z1_F)^{-1}(T_F-z_o1_F)^{-1}V\varphi(z),\Psi_o)$$

Comme $z \in \rho(T_F)$, $(T_F-z1_F)^{-1}$ est holomorphe dans $\mathcal{V}$ et on peut passez à la limite :

$$\lambda'(z_o) = -\,1\,-\,(U(T_F-z_o)^{-2}V\,\varphi(z_o),\Psi_o)$$

On obtient le résultat en remarque que d'après ii)

$$\varphi(z_o) = \frac{\varphi_o}{(\varphi_o,\Psi_o)}$$

D'où :

$$\lambda'(z_o) = -\,1\,-\,\frac{(U(T_F-z_o1_F)^{-2}V\,\varphi_o,\Psi_o)}{(\varphi_o,\Psi_o)}$$

C.Q.F.D.

2°) CAS OU T EST AUTOADJOINT

Lorsque T est autoadjoint, la famille $S(x) = T_E-x1_E-U(T_F-x1_F)^{-1}V$ pour x réel, $x \in \rho(T_F)$, est une famille d'opérateurs autoadjoints.

En effet :

$$S^*(x) = T_E^*-x1_E-V^*(T_F^*-x1_F)^{-1}u^*$$

Comme T est autoadjoint, on a

$$T_E^* = T_E\ ,\ T_F^* = T_F \quad \text{et} \quad V = U^*$$

et le résultat est immédiat. De plus un opérateur autoadjoint est fermé et,

Nous pouvons démontrer le résultat suivant :

PROPOSITION 11 :

T est supposé autoadjoint , $\mathcal{D}(T) \supset E$.
Soit $x_o \in \rho(T_F)$ et
soit $\lambda(x_o)$ une valeur propre de $S(x_o)$ de multiplicité m.
Alors il existe un voisinage $\mathcal{V}$ de x_o dans lequel :

i) Il existe m fonctions analytiques $x \to \lambda_i(x)$, i=1,...,m , où $\lambda_i(x)$ est une valeur propre de S(x) .

ii) Il existe m fonctions vectorielles $x \to \varphi_i(x)$, i=1,...,m , analytiques, où $\varphi_i(x)$ est un vecteur propre de S(x) associé à $\lambda_i(x)$ et où pour tout x la famille $\{\varphi_i(x)\}$ est orthonormale.

iii) La dérivée de $\lambda_i(x)$ en un point x de $\mathcal{V}$ est donnée par

$$\lambda_i'(x) = - 1 - \| (T_F - x1_F)^{-1} V \varphi_i(x) \|^2$$

iv) Si $\mu \in \mathcal{V}$, et si pour une des fonctions $\lambda_i(x)$ on a $\lambda_i(\mu) = 0$ alors μ est une valeur propre de T.

DEMONSTRATION :

Le i) et iii) découlent de la théorie générale de la perturbation des opérateurs linéaires. On sait en effet que pour une famille analytique d'opérateurs autoadjoints S(x) , (x réel) les $\lambda_i(x)$ sont analytiques et il existe des vecteurs propres $\varphi_i(x)$ associés aux $\lambda_i(x)$ tels que les $\varphi_i(x)$ sont analytiques en x [1,pp. 120-121] .

La démonstration du iii) se fait de manière analogue à celle du iii de la proposition 10. Ici (φ_o, Ψ_o) est remplacé par $\|\varphi_i(x)\|$ qui est égal à 1 .

C.Q.F.D.

BIBLIOGRAPHIE POUR LA PREMIERE PARTIE

[1] KATO, T.
"Perturbation theory for linear operators".
Springer Verlag (1966).

[2] WEINSTEIN, A. ; STEINGER, W.
"Methods of intermediate problems for eingenvalues".
Academic Press (1972).

[3] CHATELIN-LABORDE, F.
"Méthodes numériques de calcul des valeurs propres et des vecteurs propres d'un opérateur linéaire".
Thèse, Université de Grenoble (1971).

[4] GOULD, S.H.
"Variational methods for eingenvalue problems".
University of Toronto Press (1957).

[5] WILKINSON, J.H.
"The algebraic eingenvalue problem".
Clarendon Press (1965).

DEUXIEME PARTIE

APPLICATIONS DU PARTITIONNEMENT

Le domaine le plus simple pour les applications des résultats précédents est, sans doute, celui du calcul des valeurs propres de matrices.

Nous allons commencer par étudier un algorithme proposé par Abramov pour calculer des valeurs propres de matrices hermitiennes ([1],1958 ; [6], 1961).

Cet algorithme, comme on le verra, consiste simplement à calculer les valeurs propres comme points fixes des fonctions $\lambda_i(z)+z$, où les $\lambda_i(z)$ sont les fonctions étudiées dans la première partie : en effet la $i^{ème}$ valeur propre cherchée est la solution de $\lambda_i(z) = 0$. Les hypothèses suffisantes pour la convergence seront précisées, grâce à la théorie précédente, et d'autre part des procédés d'accélération de la convergence seront proposés. Ceci fera l'objet du chapitre I. Au chapitre II nous verrons l'utilisation pratique des fonctions $\lambda_i(z)$ pour accélérer l'algorithme QR et l'algorithme de la bissection. Ces deux derniers algorithmes étant employés après la tridiagonalisation des matrices de départ, il nous a paru logique de n'étudier que le cas des matrices tridiagonales.

Pour QR, en particulier, les fonctions $\lambda_i(z)$ donnent lieu à des translations d'origine de nature différente de celles utilisées usuellement. La convergence asymptotique est alors d'ordre 7 au lieu de 3.

Au chapitre III, nous verrons l'utilisation des résultats précédents pour le calcul des valeurs propres d'opérateurs compacts hermitiens.

CHAPITRE I

APPLICATIONS AUX MATRICES HERMITIENNES

INTRODUCTION

Soit A une matrice carrée d'ordre n sur K (K = $\mathbb{R}$ ou $\mathbb{C}$) partitionné ainsi :

$$A = \left(\begin{array}{c|c} a & b \\ \hline c & d \end{array}\right) \quad \begin{matrix} \updownarrow m \\ \updownarrow p \end{matrix}$$

où $a \in \mathcal{M}_{m,m}(K)$ $\quad b \in \mathcal{M}_{m,p}(K)$

$c \in \mathcal{M}_{p,m}(K)$ $\quad d \in \mathcal{M}_{p,p}(K)$

avec $m+p = n$.

On note 1_i la matrice identité sur K^i.

Il y a deux manières de choisir l'espace E introduit dans la partie I précédente.

1) Soit $E = \{(e_1, e_2, \ldots, e_m)\}$, espace engendré par les m vecteurs $e_1, e_2, \ldots, e_m$. Dans ce cas l'opérateur S(z) défini dans la première partie est représenté par :

$$a(z) = a - z1_m - b(d - z1_p)^{-1}c$$

2) Soit $E = \{(e_{m+1}, e_{m+2}, \ldots, e_n)\}$ et alors $S(z)$ est représenté par :

$$d(z) = d - z1_p - c(a - z1_m)^{-1}b$$

On note $\sigma(a)$ l'ensemble des valeurs propres de a.
$a(z)$ et $d(z)$ sont définis en dehors de $\sigma(a)$ et $\sigma(d)$ respectivement.
Comme conséquence immédiate du corollaire 3 on obtient :

LEMME 1 :

Soit μ un nombre complexe qui n'est pas une valeur propre de a (resp. de d). Alors μ est une valeur propre de A si et seulement si le déterminant de $d(\mu)$ (resp. de $a(\mu)$) est nul.

Nous disposons ici d'une démonstration plus simple que celle du corollaire 3, puisque la décomposition

$$A - \mu 1_n = \left(\begin{array}{c|c} a-\mu 1_m & 0 \\ \hline c & d(\mu) \end{array}\right)\left(\begin{array}{c|c} 1_m & (a-\mu 1_m)^{-1}b \\ \hline 0 & 1_p \end{array}\right) \qquad (1.1.)$$

permet de conclure que $\det(d(\mu)).\det(a-\mu 1_m) = \det(A-\mu 1_n)$ et le résultat est immédiat.

La proposition 6 et la proposition 1 de la première partie deviennent respectivement :

LEMME 2 :

L'application qui à z associe $d(z)$ (resp. $a(z)$) est holomorphe dans $\mathbb{C} \setminus \sigma(a)$ (resp. $\mathbb{C} \setminus \sigma(d)$).

PROPRIETE 3 :

Si on note π la matrice $\pi = (e_{m+1}, e_{m+2}, \ldots, e_n)$ et π' la matrice $\pi' = (e_1, e_2, \ldots, e_m)$, alors :

. pour $z \notin \sigma(a) \cup \sigma(A)$ on a $d(z) = (\pi^T (A - z1_n)^{-1} \pi)^{-1}$

. pour $z \notin \sigma(d) \cup \sigma(A)$ on a $a(z) = (\pi'^T (A - z1_n)^{-1} \pi')^{-1}$

Il est également possible de démontrer directement les deux résultats ci-dessus.

1. MATRICES HERMITIENNES : L'ALGORITHME D'ABRAMOV.

Lorsque A est hermitienne, a et d le sont aussi et de plus $c = b^H$.

D'après le lemme 1, pour calculer les valeurs propres de A, qui ne sont pas des valeurs propres de d (resp. de a), il faut résoudre l'équation det (a(z)) = 0 (resp. det (d(z)) = 0).

Puisque A est hermitienne, on peut se restreindre à chercher les racines de l'une des équations ci-dessus sur l'axe réel.

On s'occupera de l'équation det (a(x)) = 0 , mais il est clair que le même travail peut être fait pour l'équation det (d(x)) = 0.

. Soit A une matrice hermitienne : $A = \left(\begin{array}{c|c} a & b \\ \hline b^H & d \end{array}\right)$

où $a \in \mathcal{M}_{m,m}(K)$, $b \in \mathcal{M}_{m,p}(K)$, $d \in \mathcal{M}_{p,p}(K)$.

. Les valeurs propres μ_i de A et $\mu_i^{(a)}$ de a sont ordonnées en croissant (resp. en décroissant).

Pour résoudre l'équation $\det (a-x1_m-b(d-x1_p)^{-1}b^H) = 0$ Abramov a proposé l'algorithme suivant [6] :

En partant de $\mu_i^{(0)} = \mu_i^{(a)}$

faire :

$$\mu_i^{(s+1)} = \text{la } i^{\text{ème}} \text{ valeur propre de}$$

$$a_i^{(s)} = a-b(d-\mu_i^{(s)}1p)^{-1}b^H$$

La matrice $a_i^{(s)}$ ci-dessus n'est autre que $a(\mu_i^{(s)}) + \mu_i^{(s)}1_m$ Pour tout x, les nombres $\lambda_i(x)$ peuvent être ordonnées en croissant (resp. en décroissant) (cf. KATO [10]).

On peut donc dire que $\lambda_i(x)$ est la $i^{\text{ème}}$ valeur propre de a(x). On voit alors aisément que l'algorithme d'Abramov n'est pas autre chose que l'algorithme des approximations successives appliquée à la résolution de l'équation $\lambda_i(x)+x = x$ équivalente à $\lambda_i(x) = 0$.

2. ETUDE DE LA CONVERGENCE.

) Il nous a été possible de démontrer directement (démonstration longue !) le résultat suivant concernant la convergence de l'algorithme d'Abramov :

PROPOSITION :

4.A) Si on ordonne toutes les valeurs propres en croissant et si on suppose que :

i) $d-\mu_i^{(a)}1p$ est définie positive

ii) $\|b(d-\mu_i^{(a)}1p)^{-2}b^H\| < 1$

Alors la suite $\mu_j^{(s)}$ converge vers μ_j, pour j=1,2,...,i-1,i, lorsque s tend vers l'infini.

4.B) Si on ordonne toutes les valeurs propres en décroissant et si on suppose que :

i) $d-\mu_i^{(a)} 1_p$ est définie négative

ii) $\|b(d-\mu_i^{(a)} 1p)^{-2} b^H\| < 1$

Alors la suite $\mu_j^{(s)}$ converge vers μ_j, pour j=1,2,...,i-1,i, lorsque s tend vers l'infini.

REMARQUES SUR LES HYPOTHESES :

1) L'hypothèse ii) est importante : si elle n'est pas vérifiée, il peut y avoir divergence des $\mu_j^{(s)}$. Voici un exemple :

Soit $A = \left(\begin{array}{cc|c} 0 & 0 & \beta \\ 0 & -1 & 0 \\ \hline \beta & 0 & 0 \end{array}\right)$, i.e. $a = \begin{pmatrix} 0 & 0 \\ 0 & -1 \end{pmatrix}$, $b = \begin{pmatrix} \beta \\ 0 \end{pmatrix}$, $d = 0$.
β réel

Si on ordonne les valeurs propres en croissant, on a :

$\mu_1^{(a)} = -1$ et donc $d-\mu_1^{(a)}$ est définie positive. C'est-à-dire que l'hypothèse i) est vérifiée.

* Si $|\beta| > 1$ $\mu_1^{(s)}$ diverge . En effet :

$$a_1^{(s)} = \begin{pmatrix} \beta^2/\mu_1^{(s)} & 0 \\ 0 & -1 \end{pmatrix} \qquad \text{et } \mu_1^{(s+1)} = \min \left\{\frac{\beta^2}{\mu_1^{(s)}}, -1\right\}$$

On a alors :

$\mu_1^{(0)} = -1$, $\mu_1^{(1)} = -\beta^2$, $\mu_1^{(2)} = -1$, ...

et la suite $\mu_1^{(s)}$ alterne entre les nombres -1 et +1 .

* Si $|\beta| < 1$, il y a convergence car $\mu_1^{(s)} = -1$ pour tout s, et la suite $\mu_1^{(s)}$ est donc stationnaire en -1.

Ceci s'explique par le fait que lorsque $|\beta| > 1$, l'hypothèse ii) n'est pas vérifiée alors que si $|\beta| < 1$ elle l'est.

Cette hypothèse est ignorée de l'auteur V. CHICHOV [6] , qui a en fait démontré l'existence de deux nombres $\bar{\mu}_i$ et $\underline{\mu}_i$ avec $\underline{\mu}_i \leq \mu_i \leq \bar{\mu}_i$, et tels que :

. $\mu_i^{(2s+1)}$ converge en décroissant vers $\bar{\mu}_i$ lorsque s tend vers l'infini.

. $\mu_i^{(2s)}$ converge en croissant vers $\underline{\mu}_i$ lorsque s tend vers l'infini.

L'égalité $\bar{\mu}_i = \underline{\mu}_i$ n'est pas montrée.

2) Il peut y avoir divergence si l'hypothèse i) n'est pas vérifiée même lorsque ii) est vérifiée.
Voici un exemple :

$$\text{Soit } A = \left(\begin{array}{cc|cc} 0 & 0 & \sqrt{6/5} & 0 \\ 0 & 0 & 0 & 1/\sqrt{5} \\ \hline \sqrt{6/5} & 0 & -2 & 0 \\ 0 & 1/\sqrt{5} & 0 & 1/2 \end{array}\right)$$

$$\text{i.e. } a = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, \quad b = \begin{pmatrix} \sqrt{6/5} & 0 \\ 0 & 1/\sqrt{5} \end{pmatrix}, \quad d = \begin{pmatrix} -2 & 0 \\ 0 & 1/2 \end{pmatrix}$$

Ordonnons les valeurs propres en décroissant :

$$\mu_1^{(a)} = 0 \quad \text{et donc} \quad d - \mu_1^{(a)} 1_p = \begin{pmatrix} -2 & 0 \\ 0 & 1/2 \end{pmatrix}$$

n'est ni définie positive, ni définie négative.

DEMONSTRATION :

En effet on a démontré (proposition 10 de la première partie) que dans le cas où l'opérateur est hermitien :

$$\lambda_j'(x) = -1-\|(d-x1p)^{-1}b^H\varphi_j(x)\|^2 \quad \text{où } \varphi_j(x) \text{ est}$$

le vecteur propre de a(x) associé à la $j^{ème}$ valeur propre $\lambda_j(x)$.

Si on appelle (α_k) les composantes du vecteur $b^H\varphi_j(x)$ selon la base des vecteurs propres u_k de d on a :

$$(d-x1p)^{-1}b^H\varphi_j(x) = \sum_{k=}^{p} \frac{\alpha_k^2}{(\mu_k^{(d)}-x)^2}$$

où $\mu_k^{(d)}$ est la $k^{ème}$ valeur propre de d.

Comme $d-\mu_i^{(a)}1p$ est définie positive par hypothèse, alors :

$$\mu_k^{(d)}-x1p \geq \mu_k^{(d)}-\mu_i^{(a)} > 0 \quad \text{pour tout} \quad x \leq \mu_i^{(a)} .$$

D'où :

$$\|(d-x1p)^{-1}b^H\varphi_j(x)\|^2 = \sum_{k=1}^{p} \frac{\alpha_k^2}{(\mu_k^{(d)}-x)^2} \leq \Sigma \frac{\alpha_k^2}{(\mu_k^{(d)}-\mu_k^{(a)})^2}$$

$$= \|(d-\mu_i^{(a)}1p)^{-1}b^H\varphi_j(x)\|^2$$

et donc :

$$|\lambda_j'(x)+1| \leq \|(d-\mu_i^{(a)}1p)^{-1}b^H\varphi_j(x)\|^2 = \|b(d-\mu_i^{(a)}1p)^{-2}b^H\varphi_j(x)\|$$

$$\leq \|b(d-\mu_i^{(a)})^{-2}b^H\| < 1$$

C.Q.F.D.

On a :

$$a_1^{(s)} = \begin{pmatrix} \frac{6/5}{2+\mu_1^{(s)}} & 0 \\ 0 & \frac{1/5}{\mu_1^{(s)}-1/2} \end{pmatrix}$$

et donc :

$$\mu_1^{(s+1)} = \max \{ \frac{6/5}{2+\mu_1^{(s)}} , \frac{1/5}{\mu_1^{(s)}-1/2} \}$$

On peut montrer par récurrence que :

$$0 \leq \mu_1^{(3s+1)} \leq 3/10$$

et que

$$\mu_1^{(3s+2)} \geq 2$$

ce qui montre que la suite $\mu_1^{(s)}$ est divergente.

Pour démontrer la convergence de l'algorithme sous les hypothèses 4.A), nous allons montrer d'abord que les fonctions $\lambda_j(x)+x$ sont contractante pour $x \leq \mu_i^{(a)}$. Cela entraînera que les $\mu_j^{(s)}$ vont converger. Il faudra alors montrer que $\mu_j^{(s)}$ converge vers μ_j. Nous le démontrons à l'aide de trois lemmes :

LEMME 5 :

Si $d-\mu_i^{(a)} 1p$ est définie positive et si $\| b(d-\mu_i^{(a)} 1p)^{-2} b^H \| < 1$, alors $j=1,2,\ldots,i$ les fonctions $\lambda_j(x)+x$ sont contractantes dans l'intervalle $]-\infty,\mu_i^{(a)}]$.

DEMONSTRATION :

En effet on a démontré (proposition 10 de la première partie) que dans le cas où l'opérateur est hermitien :

$$\lambda_j'(x) = -1-\|(d-x1p)^{-1}b^H\varphi_j(x)\|^2$$ où $\varphi_j(x)$ est le vecteur propre de a(x) associé à la $j^{ème}$ valeur propre $\lambda_j(x)$.

Si on appelle (α_k) les composantes du vecteur $b^H\varphi_j(x)$ selon la base des vecteurs propres u_k de d on a :

$$(d-x1p)^{-1}b^H\varphi_j(x) = \sum_{k=}^{p} \frac{\alpha_k^2}{(\mu_k^{(d)}-x)^2}$$

où $\mu_k^{(d)}$ est la $k^{ème}$ valeur propre de d.

Comme $d-\mu_i^{(a)}1p$ est définie positive par hypothèse, alors :

$$\mu_k^{(d)}-x1p \geq \mu_k^{(d)}-\mu_i^{(a)} > 0 \quad \text{pour tout} \quad x \leq \mu_i^{(a)} .$$

D'où :

$$\|(d-x1p)^{-1}b^H\varphi_j(x)\|^2 = \sum_{k=1}^{p} \frac{\alpha_k^2}{(\mu_k^{(d)}-x)^2} \leq \Sigma \frac{\alpha_k^2}{(\mu_k^{(d)}-\mu_k^{(a)})^2}$$

$$= \|(d-\mu_i^{(a)}1p)^{-1}b^H\varphi_j(x)\|^2$$

et donc :

$$|\lambda_j'(x)+1| \leq \|(d-\mu_i^{(a)}1p)^{-1}b^H\varphi_j(x)\|^2 = \|b(d-\mu_i^{(a)}1p)^{-2}b^H\varphi_j(x)\|$$

$$\leq \|b(d-\mu_i^{(a)})^{-2}b^H\| < 1$$

C.Q.F.D.

LEMME 6 :

Soit μ un nombre complexe n'appartenant pas au spectre de d. Alors les deux assertions suivantes sont équivalentes :

i) μ est une valeur propre de A de multiplicité γ .

ii) 0 est une valeur propre de $a(\mu)$ de multiplicité γ .

DEMONSTRATION :

Les matrices A et $a(\mu)$ étant hermitiennes, la multiplicité d'une valeur propre est égale au nombre de vecteurs propres linéairement indépendants associés à cette valeur propre. Or, le vecteur $X = \begin{pmatrix} x \\ y \end{pmatrix}$ est un vecteur propre de A associé à la valeur propre μ si et seulement si x est un vecteur propre de $a(\mu)$ associé à la valeur propre 0 et $y = -(d-\mu 1_p)^{-1} b^H x$.
Il est alors facile de voir que pour que $x_1, x_2, \ldots, x_\nu$, soient indépendants, il faut et il suffit que $\begin{pmatrix} x_1 \\ y_1 \end{pmatrix}\begin{pmatrix} x_2 \\ y_2 \end{pmatrix}\ldots\begin{pmatrix} x_\nu \\ y_\nu \end{pmatrix}$ le soient aussi.

C.Q.F.D.

LEMME 7 :

Soit μ_i une valeur propre de A telle que $d-\mu_i 1_p$ est définie positive . Alors :

1°) Pour $j=1,2,\ldots,i$ on a $\lambda_j(\mu_j) = 0$

2°) Réciproquement , soit μ une valeur propre plus petite que μ_i et soit j un indice pour lequel $\lambda_j(\mu) = 0$.
Alors μ est la $j^{ème}$ valeur propre de A.

DEMONSTRATION :

1°) Soient $\Lambda_1, \Lambda_2, \ldots, \Lambda_q$, les valeurs propres distinctes de A, ordonnées en croissant :

$$\Lambda_1 < \Lambda_2 \ldots < \Lambda_q$$

$\gamma(i)$ dénote la multiplicité de Λ_i .

Posons :

$$\eta(i) = \sum_{k=1}^{i} \gamma(k) = \max \{k | \lambda_k \leq \Lambda_i\}$$

Montrons d'abord que $\lambda_1(\mu_1) = 0$.

μ_1 n'étant pas une valeur propre de d, il existe d'après le lemme 1, k tel que $\lambda_k(\mu_1) = 0$.
Soit k_1 le plus petit des indices k tels que $\lambda_k(\mu_1) = 0$.

Alors $k_1 = 1$.

En effet sinon on aurait :

. $\lambda_{k_1-1}(\mu_1) \leq \lambda_{k_1}(\mu_1) = 0$

. $\lambda_{k_1-1}(x) \to +\infty$ lorsque $x \to -\infty$.

Donc λ_{k_1-1} aurait un zéro inférieur ou égal à μ_i .

Appelons μ ce zéro. On a :

. Soit $\mu = \mu_1$, i.e. $\lambda_{k_1-1}(\mu_1) = 0$: ceci contredit le fait que k_1 est le plus petit indice k pour lequel $\lambda_{k_1}(\mu_1) = 0$.

. Soit $\mu < \mu_1$. Or est une valeur propre de A puisque $\lambda_{k_1-1}(\mu) = 0$.

Donc ceci contredit le fait que λ_1 est la plus petite valeur propre de A.

Finalement, on a bien $\lambda_1(\mu_1) = 0$.

Comme $\mu_1 = \mu_2 = \ldots = \mu_{\eta(1)} = \Lambda_1$

$$\lambda_1(\mu_j) = 0 \qquad j=1,2,\ldots,\eta(1) \ ;$$

et d'après le lemme 6 :

$$\lambda_1(\Lambda_1) = \lambda_2(\Lambda_1) = \ldots = \lambda_{\eta(1)}(\Lambda_1) = 0$$

On a donc démontré que $\lambda_j(\mu_j) = 0$ pour $j=1,2,\ldots,\eta(1)$.

Supposons que le résultat soit vrai jusqu'à $\eta(j)$ et montrons que c'est alors vrai jusqu'à $\eta(j+1)$. (Tant que $\mu_{\eta(j+1)} \leq \mu_i$) .

Il suffira de montrer que $\lambda_{\eta(j+1)}(\mu_{\eta(j+1)}) = 0$, car alors en faisant le même raisonnement que ci-dessus sur les multiplicités, on montrerait aisément que $\lambda_k(\mu_k) = 0$ pour $k = \eta(j)+1 \ , \ \eta(j)+2,\ldots, \eta(j+1)$.

Puisque $\mu_{\eta(j)+1} \notin \sigma(d)$, il existe un indice k tel que $\lambda_k(\mu_{\eta(j)+1}) = 0$. (2)

Soit ℓ le plus petit des indices k tels que (2) est vérifiée.

<u>Alors $\ell = \eta(j)+1$</u> .

En effet, sinon on aurait :

- soit $\ell \leq \eta(j)$: ceci est impossible car alors :
  $\lambda_\ell(\mu_{\eta(j)+1}) = 0$ et $\lambda_{k_j}(\mu_{k_j}) = 0$
  (hypothèse de récurrence) impliquent, en raison de l'unicité du zéro de la fonction λ_ℓ , que $\mu_{\eta(j)+1} = \mu_\ell$ ceci est impossible puisque :

$$\mu_\ell \leq \mu_{\eta(j)} < \mu_{\eta(j)+1}$$

. Soit $\ell > \eta(j)+1$.

Alors on aurait :

. $\lambda_{\ell-1}(\mu_{\eta(j)+1}) \leq \lambda_{\ell}(\mu_{\eta(j)+1}) = 0$

. $\lambda_{\ell-1}(\mu_{\eta(j)}) > \lambda_{\ell-1}(\mu_{\eta(j)+1}) = 0$

($\lambda_{\ell-1}$ est strictement décroissante).

D'où $\lambda_{\ell-1}$ admet un zéro μ tel que $\mu_{\eta(j)} < \mu \leq \mu_{\eta(j)+1}$

μ étant une valeur propre de A on ne peut pas avoir $\mu_{\eta(j)} < \mu < \mu_{\eta(j)+1}$ parce qu'il n'y a aucune valeur propre de A strictement comprise entre $\mu_{\eta(j)}$ et $\mu_{\eta(j)+1}$.

Donc il ne reste plus que la possibilité $\mu = \mu_{\eta(j)+1}$ i.e. $\lambda_{\ell-1}(\mu_{\eta(j)+1}) = 0$. Or c'est absurde car ℓ est par définition le plus petit indice k pour lequel (2) est vérifiée. Donc la possibilité $\ell > \eta(j)+1$ est également a écarter. Il reste seulement la possibilité $\ell = \eta(j)+1$.

2°) La deuxième partie du lemme découle de la première partie puisque, les fonctions λ_j étant décroissantes strictement, elles ont au plus un zéro : donc $\lambda_j(\mu) = 0$ et $\lambda_j(\mu_j) = 0$ impliquent que $\mu = \mu_j$.

DEMONSTRATION de la PROPOSITION 4 :

Nous démontrons uniquement le 4.A) : une démonstration identique pourrait être faite pour 4.B).

D'après le lemme 5, et sous les hypothèses de la proposition, les fonctions $x \to \lambda_j(x)+x$ pour $j=1,2,\ldots,i$, sont contractantes dans l'intervalle $]-\infty,\mu_i^{(a)}]$.

Puisque $\mu_j^{(a)} \leq \mu_i^{(a)}$, les procédés des approximations successives $\mu_j^{(s+1)} = \lambda_j(\mu_j^{(s)})+\mu_j^{(s)}$ partant de $\mu_j^{(0)} = \mu_j^{(a)}$ vont converger.

Donc pour $j=1,2,\ldots,i$: $\mu_j^{(s)}$ converge vers un nombre μ tel que

$$\lambda_j(\mu)+\mu = \mu \text{ i.e. } \lambda_j(\mu) = 0 .$$

Donc d'après le deuxièmement du lemme 5, $\mu = \mu_j$.
(Le lemme 5 s'applique parce que $\mu_j \leq \mu_j^{(a)}$ et donc
$d-\mu_j 1_p = d-\mu_j 1_p+(\mu_j^{(a)}-\mu_j)1_p$ est aussi définie positive).

C.Q.F.D.

REMARQUES :

* La démonstration directe, adaptée de [6] , utilise le principe du minimax et des raisonnements analogues à ceux ci-dessus sur les multiplicités des valeurs propres. Cette démonstration très longue et très technique a peu d'intérêt.

* D'après le lemme 5 et la démonstration ci-dessus, la condition ii) de la propositions 4 est une condition suffisante pour que les fonctions $\lambda_j(x)$ $j=1,\ldots,i$ soient contractantes dans l'intervalle $]-\infty,\mu_i^{(a)}]$.

3. ACCELERATION DE LA CONVERGENCE.

Nous venons de présenter l'algorithme d'Abramov comme une méthode d'approximation successives appliquée aux équations $\lambda_i(x)+x = x$.
Cet algorithme permet d'approcher un zéro de la fonction $\lambda_j(x)$ et ce zéro d'après le lemme 1 est une valeur propre de A.

Comme nous verrons que la convergence du procédé proposé par Abramov peut être très lente nous pourrons calculer les zéros de $\lambda_i(x)$ par d'autres procédés classiques de résolutions d'équations.

De plus, les hypothèses i) et ii) de la proposition 4 sont trop restrictives : on sait que pour appliquer la méthode de Newton, il n'est pas nécessaire que les fonctions dont on cherche les zéros soient contractantes.

3.1.

Commençons par étudier la vitesse de convergence de l'algorithme d'Abramov.

. Soit $\mu_i^{(s+1)}$ la $i^{ème}$ valeur propre de $a_i^{(s)}$

. D'après le lemme 1, si μ_i n'est pas une valeur propre de d, alors $\det(a(\mu_i)) = 0$; donc 0 est une valeur propre de $a(\mu_i)$.

Soit alors $x_i^{(s)}$ un vecteur propre de $a_i^{(s)}$ associé à $\mu_i^{(s+1)}$ et x_i un vecteur propre de $a(\mu_i)$ associé à la valeur propre 0 de $a(\mu_i)$.
Supposons que $(x_i^{(s+1)}, x_i) \neq 0$.
Alors il a été démontré par Chichov [6] que :

$$\frac{\mu_i^{(s+1)}-\mu_i}{\mu_i^{(s)}-\mu_i} = - \frac{(b(d-\mu_i 1p)^{-1}(d-\mu_i^{(s)}1p)^{-1}{}_b H x_i^{(s)}, x_i)}{(x_i^{(s)}, x_i)} \qquad (1)$$

Si donc on pose :

$$y_i = -(d-\mu_i 1p)^{-1}{}_b H x_i \quad ,$$

et $$y_i^{(s)} = -(d-\mu_i^{(s)}1p)^{-1}{}_b H x_i^{(s)} \quad .$$

L'égalité ci-dessus devient :

$$\frac{\mu_i^{(s+1)}-\mu_i}{\mu_i^{(s)}-\mu_i} = - \frac{(y_i^{(s)}, y_i)}{(x_i^{(s)}, x_i)}$$

On remarque que le vecteur $X_i = \begin{pmatrix} x_i \\ y_i \end{pmatrix}$ est un vecteur propre de A associé à la valeur propre μ_i (cf. la démonstration de la proposition 4).

D'autre part le vecteur :

$$X_i^{(s)} = \begin{pmatrix} x_i^{(s)} \\ y_i^{(s)} \end{pmatrix}$$

est une approximation de X_i si $\mu_i^{(s+1)}$ est une approximation de μ_i .

La convergence asymptotique est donc réglée par le nombre

$$-\frac{\|y_i\|^2}{\|x_i\|^2}$$

Si $\frac{\|y_i\|}{\|x_i\|} << 1$ la convergence est rapide. C'est le cas lorsque les matrices b et d sont de normes "petites".

Dans le cas contraire, le procédé proposé peut donner une convergence très lente. Nous devons donc étudier des procédés d'accélération de la convergence.

3.2. LE PROCEDE DE NEWTON.

Nous connaissons la dérivée de la fonction $\lambda_i(x)$ en un point x quelconque où λ_i est définie.

Grâce à la proposition 4, on a en effet en appelant $x_i^{(s)}$ le vecteur propre de $a_i^{(s)} = a(\mu_i^{(s)}) + \mu_i^{(s)} 1_m$ associé à $\mu_i^{(s+1)}$:

$$\lambda_i'(\mu_i^{(s)}) = -1 - \|(d-\mu_i^{(s)})^{-1} b^H x_i^{(s)}\|^2$$

$$= -1 - \|y_i^{(s)}\|^2$$

Le procédé de Newton pour résoudre l'équation $\lambda_i(x) = 0$ consiste à construire la suite $\mu_i^{(0)} = \mu_i^{(a)}$,

$$\mu_i^{(s+1)} = \mu_i^{(s)} - \frac{\lambda_i(\mu_i^{(s)})}{\lambda_i'(\mu_i^{(s)})} \quad , \quad s=0,1,2,\ldots, \quad \text{soit}$$

$$\mu_i^{(s+1)} = \mu_i^{(s)} + \frac{\lambda_i(\mu_i^{(s)})}{1+\|y_i^{(s)}\|^2} \quad , \quad s=0,1,2,\ldots$$

Cette itération exige à chaque pas, le calcul de la $i^{ème}$ valeur propre de $a-\mu_i^{(s)}-b(d-\mu_i^{(s)}1p)^{-1}b^H$ et du vecteur propre correspondant, mais permet d'obtenir le vecteur propre $\begin{pmatrix} x_i \\ y_i \end{pmatrix}$ associé à μ_i puisque $\begin{pmatrix} x_i \\ y_i \end{pmatrix}$ est la limite du vecteur $\begin{pmatrix} x_i^{(s)} \\ y_i^{(s)} \end{pmatrix}$.

4. EXPERIENCES NUMERIQUES

Soit $A = \left(\begin{array}{c|c} a & b \\ \hline b^T & d \end{array}\right) \begin{matrix} \updownarrow m \\ \updownarrow p \end{matrix}$ symétrique d'ordre $n = m+p$

Nous faisons des expériences numériques sur la matrice A avec m = p et :

$$a = \begin{pmatrix} 2 & -1 & & & \\ -1 & \ddots & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & -1 \\ & & & -1 & 2 \end{pmatrix} \qquad b = \begin{pmatrix} 0 & -1 & & & \\ & \ddots & \ddots & & \\ & & \ddots & \ddots & \\ & & & \ddots & -1 \\ 1 & & & 1 & 0 \end{pmatrix} \qquad d = \begin{pmatrix} 6 & 1 & & & \\ 1 & \ddots & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & 1 \\ & & & 1 & 6 \end{pmatrix}$$

Les valeurs propres exactes de A sont connues.

4.1.

Nous commençons par calculer les cinq plus petites valeurs propres de A lorsque n = 20 , (m = p = 10) par le procédé d'Abramov-Chichov. L'algorithme d'Abramov-Chichov est arrêté lorsque

$$|\mu_i^{(s+1)}-\mu_i^{(s)}| < 10^{-9}$$

Le tableau 1 indique :

1°) Pour i=1,2,...,5 les valeurs propres exactes μ_i de A.

2°) Pour i=1,2,...,5 les valeurs propres $\mu_i^{(0)}$ de a, valeurs initiales de l'algorithme d'Abramov-Chichov ainsi que les erreurs initiales $\mu_i^{(0)} - \mu_i$.

3°) Pour i=1,2,...,5,... le nombre d'itérations nécessitées, la valeur approchée $\tilde{\mu}_i$ obtenue par l'algorithme et l'erreur $|\tilde{\mu}_i-\mu_i|$.

4.2.

Nous testons ensuite la méthode d'Abramov-Chichov accélérée de la manière suivante :

On prend :

$$\mu_i^{(o)} = \mu_i^{(a)}$$

$$\mu_i^{(s+1)} = \mu_i^{(s)} - \frac{\lambda_i(\mu_i^{(s)})}{\delta^{(s)}} \qquad s=0,1,2,...,...$$

où $\delta^{(0)} = -1$ et $\delta^{(s)} = \dfrac{\lambda_i(\mu_i^{(s)})-\lambda_i(\mu_i^{(s-1)})}{\mu_i^{(s)}-\mu_i^{(s-1)}}$ $\qquad s=1,2,...,...$

A la limite $\delta^{(s)}$ est voisin de la dérivée de λ_i en $\mu_i^{(s)}$.

On prend encore n = 20 et on calcule les cinq plus petites valeurs propres. Le test d'arrêt choisi est celui-ci :

$$|\mu_i^{(s+1)} - \mu_i^{(s)}| < 10^{-8}$$

Les résultats sont ceux du tableau 2.

4.3

Nous faisons exactement les mêmes expériences que précédemment lorsque n = 30 (m = p = 15).
On calcule les huit plus petites valeurs propres de A.
Les tests d'arrêt sont encore

$$|\mu_i^{(s+1)} - \mu_i^{(s)}| < 10^{-9}$$

pour l'algorithme d'Abramov-Chichov et $|\mu_i^{(s+1)} - \mu_i^{(s)}| < 10^{-8}$

pour l'algorithme accéléré.
Les résultats sont ceux des tableaux 3 et 4.

Tableau - 1 -

n = 20 , m = p = 10 .

Calcul des cinq premières valeurs propres par l'algorithme d'Abramov-Chichov.

** VALEURS PROPRES EXACTES **

-01 446766950994858
+00 177708776855437
+00 396124528390323
+00 695044902736020
+01 106779251268069

** VALEURS INITIALES **

* I *	* MU(I,0) *	* ERREURS *
1	-01 810140527710052	-01 363373576715194
2	+00 317492934337638	+00 139784157482201
3	+00 690278532109431	+00 294154003719107
4	+01 116916997399623	+00 474125071260209
5	+01 171537032345343	+00 647577810772737

** METHODE D'ABRAMOV-CHICHOV **

* ITERATIONS *	* MU(I) OBTENUS *	* ERREURS *
5	-01 446766950993626	1.231660'-13
6	+00 177708776865897	1.045966'-11
8	+00 396124528397191	6.867285'-12
10	+00 695044902745335	9.314993'-12
12	+01 106779251270621	2.551581'-11

Tableau - 2 -

n = 20 , m = p = 10 .

Calcul des cinq premières valeurs propres par l'algorithme d'Abramov-Chichov accéléré.

** VALEURS PROPRES EXACTES **

-01 446766950994858
+00 177708776855437
+00 396124528390323
+00 695044902736020
+01 106779251268069

** VALEURS INITIALES **

* I *	* MU(I,0) *	* ERREURS *
1	-01 810140527710052	-01 363373576715194
2	+00 317492934337638	+00 139784157482201
3	+00 690278532109431	+00 294154003719107
4	+01 116916997399623	+00 474125071260209
5	+01 171537032345343	+00 647577810772737

** METHODE D'ABRAMOV-CHICHOV ACCELEREE **

* ITERATIONS *	* MU(I) OBTENUS *	* ERREURS *
3	-01 446766950994850	8.049117'-16
4	+00 177708776855436	9.992007'-16
4	+00 396124528390322	1.484923'-15
5	+00 695044902736020	2.081668'-16
5	+01 106779251268069	0

Tableau - 3 -

n = 30 , m = p = 15 .

Calcul des huit premières valeurs propres par l'algorithme d'Abramov-Chichov.

** VALEURS PROPRES EXACTES **

-01 205227064324194
-01 818802349900220
+00 183442974399805
+00 324168753519077
+00 502613535421671
+00 716946235170895
+00 964967509228836
+01 124413232369725

** VALEURS INITIALES **

* I *	* MU(I,0) *	* ERREURS *
1	-01 384294391935390	-01 179067327611196
2	+00 152240934977427	-01 703606999874047
3	+00 337060775394910	+00 153617800995105
4	+00 585786437626905	+00 261617684107828
5	+00 888859533960797	+00 386245998539125
6	+01 123463313526982	+00 517686900098928
7	+01 160981935596775	+00 644851846738909
8	+01 200000000000000	+00 755867676302749

** METHODE D'ABRAMOV-CHICHOV **

* ITERATIONS *	* MU(I) OBTENUS *	* ERREURS *
4	-01 205227064330184	5.989566'-13
5	-01 818802349840443	5.977704'-12
6	+00 183442974417046	1.724179'-11
8	+00 324168753520616	1.538977'-12
9	+00 502613535416022	5.649536'-12
10	+00 716946235192322	2.142722'-11
11	+00 964967509139426	8.940991'-11
13	+01 124413232362680	7.045209'-11

Tableau - 4 -

n = 30 , m = p = 15 .

Calcul des huit premières valeurs propres par l'algorithme d'Abramov-Chichov accéléré.

** VALEURS PROPRES EXACTES **

-01 205227064324194
-01 8188023499002 20
+00 183442974399805
+00 3241687535190 77
+00 502613535421671
+00 7169462351708 95
+00 9649675092288 36
+01 124413232369725

** VALEURS INITIALES **

* I *	* MU(I,0) *	* ERREURS *
1	-01 384294391935390	-01 179067327611196
2	+00 152240934977427	-01 703606999874047
3	+00 337060775394910	+00 153617800995105
4	+00 585786437626905	+00 261617684107828
5	+00 888859533960797	+00 386245998539125
6	+01 123463313526982	+00 517686900098928
7	+01 160981935596775	+00 644851846736909
8	+01 200000000000000	+00 755867676302749

** METHODE D'ABRAMOV-CHICHOV ACCELEREE **

* ITERATIONS *	* MU(I) OBTENUS *	* ERREURS *
3	-01 205227064324180	1.440688'-15
4	-01 818802349900200	1.332268'-15
4	+00 183442974399803	1.859624'-15
4	+00 324168753519076	1.290634'-15
5	+00 502613535421671	8.326673'-16
5	+00 716946235170894	1.526557'-16
5	+00 964967509228836	4.163336'-17
5	+01 124413232369725	1.776357'-15

CHAPITRE II

APPLICATIONS AUX MATRICES TRIDIAGONALES

On rencontre, parmi les méthodes de calcul d'éléments propres de matrices, la classe très importante des algorithmes qui transforment A sous forme condensée dont les éléments propres sont faciles à calculer.

La forme condensée la plus employée pour les matrices hermitiennes est la forme tridiagonale.

Lorsqu'une matrice a été mise sous forme tridiagonale hermitienne, par l'algorithme de Householder ou de Givens, on calcule ses valeurs propres le plus souvent par l'un des algorithmes suivants :

- Algorithme QR
- Algorithme des bissections [14]

Nous allons utiliser les résultats précédents pour accélérer les deux algorithmes ci-dessus :

* Pour QR il s'agit de donner des translations d'origine se déduisant de l'étude précédente. Nous proposerons des translations d'origine de nature très différentes des translations classiques et qui améliorent la rapidité d'exécution ainsi que la précision obtenue.

* Pour l'algorithme des bissections, nous proposons un moyen de terminer le procédé des bissections par la méthode d'Abramov-Chichov accélérée

En A, nous présentons les notations utilisées pour les matrices tridiagonales.

1. PARTITIONNEMENTS SUCCESSIFS DE MATRICES TRIDIAGONALES

Soit A une matrice quelconque, d'ordre n, partitionnée ainsi :

$$A = \left(\begin{array}{c|c} a_{n-1} & b_n \\ \hline c_n & d_n \end{array}\right)$$

où a_{n-1} est la sous-matrice principale supérieure d'ordre n-1 de A.
Ceci est donc un cas particulier du partitionnement précédent où

$$m = n-1 \quad , \quad \text{et} \quad p = 1 \ .$$

On peut recommencer à partitionner $a_{n-1}, a_{n-2}, \ldots$ de la même manière, c'est-à-dire avec p = 1 à chaque fois.

Nous noterons donc pour $i=2,3,\ldots,n-1,n$:

$$a_i = \left(\begin{array}{c|c} a_{i-1} & b_i \\ \hline c_i & d_i \end{array}\right)$$

où a_j représente la sous-matrice principale supérieure d'ordre j de A pour $j=1,2,..,n-1$.

On prendra : $a_n = A$

c_i^T et b_i sont des vecteurs de K^{i-1} .

Soit maintenant A <u>tridiagonale</u> :

$$A = \begin{pmatrix} d_1 & \gamma_2 & & & \\ \beta_2 & \ddots & \ddots & & \\ & \ddots & \ddots & \gamma_n & \\ & & \ddots & d_n & \\ & & & \beta_n & \end{pmatrix}$$

On a alors, si on pose $\tilde{e}_{i-1} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} \in K^{i-1}$

$$c_i = (0,0,\ldots,0,\beta_i) = \beta_i \tilde{e}_{i-1}^T$$

et de même :

$$b_i = \gamma_i \tilde{e}_{i-1}$$

On supposera que toute la suite que $\beta_i \gamma_i \neq 0$.
La fonction matricielle $z \to d(z)$, définie au chapitre précédent, correspond ici à l'application

$$z \to d_i - z - c_i (a_{i-1} - z 1_{i-1})^{-1} b_i$$

définie pour tout z de $\mathbb{C}$ qui n'est pas une valeur propre de a_{i-1} .

NOTATION :

Pour i=2,...,n on dénote par $\Psi_i(z)$ la fonction

$$\Psi_i(z) = d_i - z - c_i (a_{i-1} - z 1_{i-1})^{-1} b_i$$

définie sur $\mathbb{C} \setminus \sigma(a_{i-1})$

On notera $\Psi(z)$ la fonction $\Psi_n(z)$.

REMARQUE :

D'après l'égalité 1.1. du paragraphe 1 on a :

$$\Psi_i(z) = \frac{\det (a_i - z 1_i)}{\det (a_{i-1} - z 1_{i-1})}$$

On va montrer que l'on peut prendre comme définition équivalente des fonctions $\Psi_i(z)$, les fonctions définies par la récurrence suivante :

$$\Psi_1(z) = d_1 - z$$

$$\Psi_i(z) = \begin{cases} \infty & \text{si} \quad z \in \sigma(a_{i-1}) \\ d_i - z - \dfrac{\beta_i \gamma_i}{\Psi_{i-1}(z)} & \text{si} \quad z \notin \sigma(a_{i-1}) \end{cases} \qquad (2.1)$$

$i=2,3,\ldots,n$.

On a en effet d'après les définitions et les notations :

$$\Psi_i(z) = d_i - z - \beta_i \gamma_i \tilde{e}_{i-1}^T (a_{i-1} - z1_{i-1})^{-1} \tilde{e}_{i-1} \qquad ((2.2)$$

Lorsque $i \geq 3$ on constate d'après la propriété 3 que

- pour $z \notin \sigma(a_{i-1}) \cup \sigma(a_{i-2})$

  on a l'égalité

  $$\tilde{e}_{i-1}^T (a_{i-1} - z1_{i-1})^{-1} \tilde{e}_{i-1} = \frac{1}{\Psi_{i-1}(z)}$$

  ce qui donne la relation (2.1) ci-dessus.

- Mais pour $z_o \in \sigma(a_{i-2}) \cup \sigma(a_{i-1})$ on peut montrer, en appliquant la décomposition 1.1 à $(A - z1_n)^{-1}$, que

  $\tilde{e}_{i-1}^T (a_{i-1} - z_o 1_{i-1})^{-1} \tilde{e}_{i-1} = 0$. i.e. d'après (2.2)

  $\Psi_i(z_o) = d_i - z_o$,

  ce qui revient à donner $\Psi_{i-1}(z_o)$ la valeur infinie dans la formule 2.1.

  La formule 2.1 est valable pour $i = 2$.

CAS DES MATRICES HERMITIENNES.

Si A est hermitienne, le spectre est réel et on peut restreindre le domaine de la fonction Ψ à l'axe réel.

Pour x réel on a :

$$\Psi_i(x) = d_i - x - b_i^H (a_{i-1} - x1_{i-1})^{-1} b_i$$

D'après la proposition 10 de la première partie, la dérivée de Ψ_i est :

$$\Psi_i'(x) = -1 - \| (a_{i-1} - x1_{i-1})^{-1} b_i \|^2$$

Les zéros de la fonction Ψ_i sont les valeurs propres μ_j^i , j=1,2,...,i de a_i.
Les pôles de Ψ_i sont les valeurs propres μ_j^{i-1} j=1,2,...,i-1 de a_{i-1} .

On suppose que toutes les valeurs propres sont ordonnées en croissant.
On démontre que :

$$\mu_j^i < \mu_j^{i-1} < \mu_{j+1}^i \qquad \text{(cf. [14])}$$

On peut alors représenter l'allure de la courbe $\Psi_i(x)$.
Sur la figure 1 on a pris i = 5.

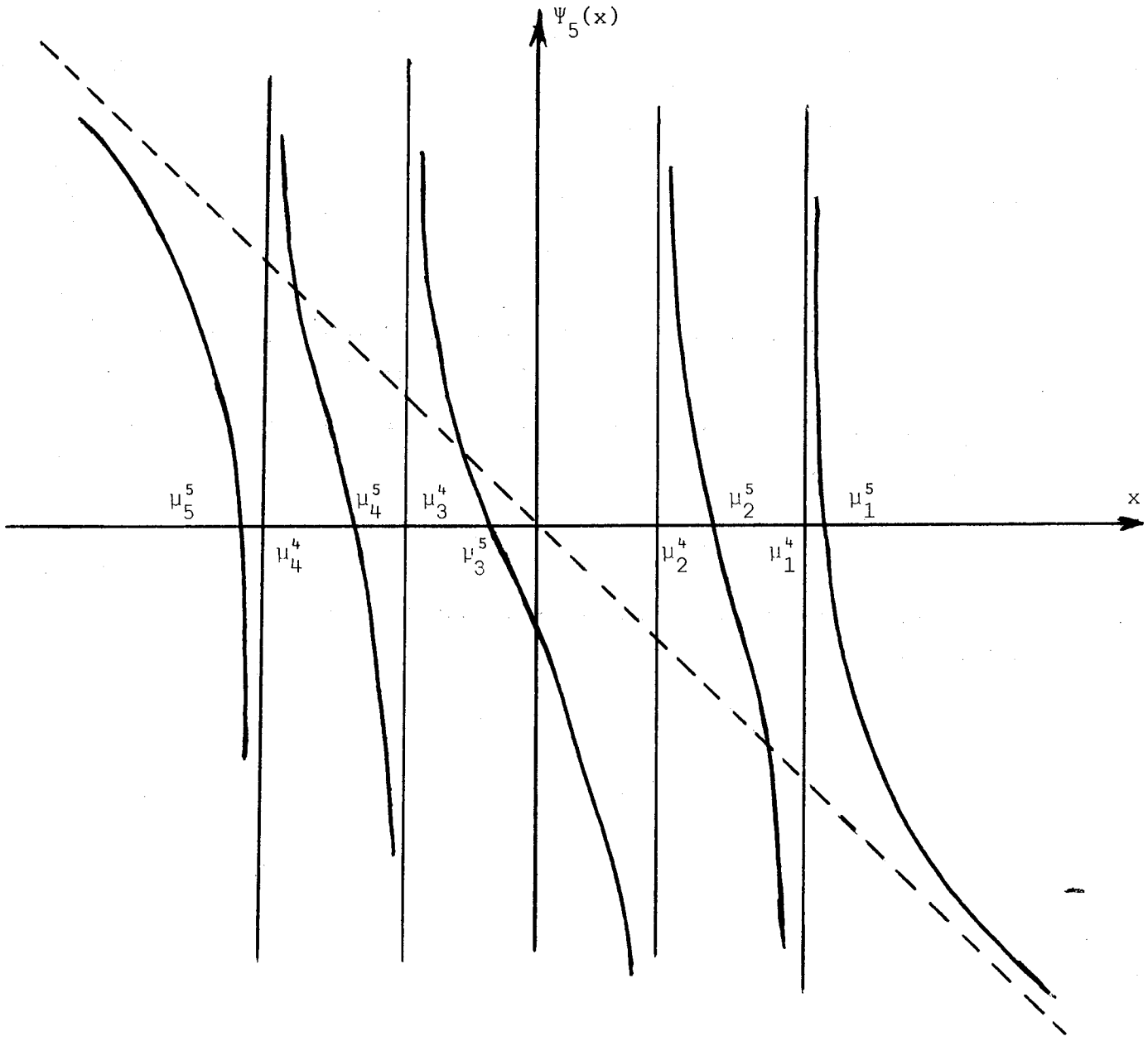


Figure 1

2. COMPORTEMENTS DES ALGORITHMES LR ET QR AVEC TRANSLATION D'ORIGINE.

2.1. LES ALGORITHMES LR ET QR AVEC TRANSLATIONS D'ORIGINE.

Parmi les méthodes les plus couramment employées pour calculer les éléments propres d'une matrice hermitienne donnée, certaines consistent à la transmuer d'abord en une matrice tridiagonale dont on doit ensuite calculer les éléments propres.

Soit donc une matrice tridiagonale :

$$A = \begin{pmatrix} d_1 & \gamma_2 & & & \\ \beta_2 & \ddots & \ddots & & \\ & \ddots & \ddots & \gamma_n & \\ & & \beta_n & d_n & \end{pmatrix}$$

Pour calculer tous les éléments propres de A on utilise le plus souvent :

- Soit QR qui conserve la forme tridiagonale symétrique.
- Soit LR, qui conserve la forme tridiagonale mais pas la symétrie.
- Soit LR, appliqué non pas à A mais à la matrice

$$\begin{pmatrix} d_1 & 1 & & & \\ \alpha_1 & \ddots & \ddots & & \\ & \ddots & \ddots & 1 & \\ & & \alpha_n & d_n & \end{pmatrix} \qquad \text{où } \alpha_i = \beta_i \gamma_i \ .$$

L'algorithme qui en résulte est appelé le Q.D. algorithme.

Ces trois algorithmes peuvent être accélérés par des translations d'origine. Ainsi, pour QR par exemple on construit la suite :

$$A_1 = A \ ,$$

$$A_s - t_s I = Q_s R_s \qquad , \qquad s=1,2,\ldots$$

$$A_{s+1} = R_s Q_s + t_s I \ .$$

On notera les éléments de A_s avec l'indice (s) supérieur.

Si A est hermitienne, ($\gamma_i^{(s)} = \overline{\beta_i^{(s)}}$ pour i=2,...,n) alors les translations d'origine les plus employées sont les suivantes :

a) $t_s = d_n^{(s)}$

b) t_s = la valeur propre de $\begin{pmatrix} d_{n-1}^{(s)} & \overline{\beta_n^{(s)}} \\ \beta_n^{(s)} & d_n^{(s)} \end{pmatrix}$ la plus voisine de $d_n^{(s)}$.

c) $t_s = d_n^{(s)} - \dfrac{|\beta_n^{(s)}|^2}{d_{n-1}^{(s)} - d_n^{(s)}}$

2.2. LE PROBLEME DE LA CONVERGENCE DES ALGORITHMES LR ET QR AVEC TRANSLATIONS .

Le problème de la convergence globale des algorithmes LR et QR avec translations d'origine est un problème difficile. On ne sait pas à notre connaissance donner des conditions simples pour garantir la convergence du procédé pour une matrice quelconque [11].

Pour les matrices tridiagonales les translations b) et c) ci-dessus ont été étudiées par J.H. Wilkinson [16] et la convergence globale de la stratégie b) a été démontrée.

La convergence asymptotique par contre est plus aisée à étudier : en supposant qu'il y a convergence (en un certain sens), on démontre que à la limite, $\beta_n^{(s)}$ tend vers zéro très rapidement (convergence quadratique ou cubique). (cf. [14],[16]).

Lorsque $\beta_n^{(s)}$ devient suffisamment petit on procède à la déflation : on ignore la $n^{ème}$ ligne et la $n^{ème}$ colonne et on poursuit l'algorithme avec la matrice principale supérieure d'ordre n-1 . Ceci nous amène à poser comme définition de la convergence de QR (resp. LR) avec translations d'origine la définition suivante :

> On dira que l'algorithme QR (resp. LR) avec translations d'origine converge si la matrice A_s converge et si $\beta_n^{(s)}$ tend vers zéro.

Il est intéressant d'étudier l'effet d'une itération des algorithmes LR et QR avec translation, sur la $n^{ème}$ ligne de la matrice tridiagonale A.

En raison des considérations du 1°, dans tout ce qui suit LR sera appliqué à une matrice tridiagonale quelconque, et QR à une matrice tridiagonale hermitienne.

2.3. EXPRESSION DE LA DERNIERE LIGNE DE LA TRANSFORMEE $\widetilde{A}$ DE A PAR UNE ITERATION DES ALGORITHMES LR ET QR .

- a - ALGORITHME APPLIQUE A A TRIDIAGONALE QUELCONQUE .

Soit $A = \begin{pmatrix} d_1 & \gamma_2 & & \\ \beta_2 & \ddots & \ddots & \\ & \ddots & \ddots & \gamma_n \\ & & \beta_n & d_n \end{pmatrix}$ partitionnée en $A = \left(\begin{array}{c|c} a_{n-1} & b_n \\ \hline c_n & d_n \end{array}\right)$

Soit $t \notin \bigcup_{i=1}^{n-1} \sigma(a_i)$. On peut vérifier que :

$$A - t1_n = \left(\begin{array}{c|c} \ell & 0 \\ \hline cr^{-1} & 1 \end{array}\right) \left(\begin{array}{c|c} r & \ell^{-1}b \\ \hline 0 & \Psi(t) \end{array}\right) = LR$$

où $\ell r = a - t1_m$ est la décomposition de Gauss de $a_{n-1} - t1_{n-1}$.

Alors :

$$A = RL + t1_n = \left(\begin{array}{c|c} \tilde{a}_{n-1} & \tilde{b}_n \\ \hline \tilde{c}_n & \tilde{d}_n \end{array}\right) = \left(\begin{array}{c|c} \tilde{a}_{n-1} & \tilde{b}_n \\ \hline \Psi(t)cr^{-1} & \Psi(t)+t \end{array}\right)$$

D'où $\tilde{d}_n = \Psi(t)+t$.

$c = (0,0,0,\ldots,\tilde{\beta}_n)$.

Calculons $\tilde{\beta}_n$.

En faisant exactement le même raisonnement que ci-dessus on aurait :

$$r_{n-1,n-1} = \Psi_{n-1}(t)$$

donc

$$cr^{-1} = \beta_n \tilde{e}_{n-1}^T r^{-1} = \beta_n(0,0,\ldots,\frac{1}{r_{n-1,n-1}}) = \frac{\beta_n}{\Psi_{n-1}(t)} \tilde{e}_{n-1}^T .$$

$$\text{i.e. } \tilde{\beta}_n = \frac{\beta_n}{\Psi_{n-1}(t)} \Psi(t) .$$

On a donc :

PROPOSITION 8:

Soit $A = \begin{pmatrix} d_1 & \gamma_2 & & & \\ \beta_2 & \ddots & \ddots & & \\ & \ddots & \ddots & \gamma_n & \\ & & \beta_n & d_n \end{pmatrix}$ et soit $t \notin \bigcup_{i=1}^{n-1} \sigma(a_i)$.

Alors si $\tilde{A}$ est la transformée de A par une itération de l'algorithme LR avec la translation d'origine t, on obtient pour $\tilde{d}_n$ et $\tilde{\beta}_n$, éléments non nuls de la dernière ligne de $\tilde{A}$, les expressions

$$\tilde{d}_n = \Psi(t)+t \quad ; \quad \tilde{\beta}_n = \frac{\Psi(t)}{\Psi_{n-1}(t)} \beta_n$$

- b - ALGORITHME QR APPLIQUE A A TRIDIAGONALE HERMITIENNE.

On cherche des formules analogues aux précédentes, exprimant par l'intermédiaire des fonctions Ψ_i, la dernière ligne de la matrice $\tilde{A}$ résultant d'une itération de l'algorithme QR avec la translation t.
Le résultat est le suivant :

PROPOSITION 9 :

Soit $A = \begin{pmatrix} d_1 & \bar{\beta}_2 & & & \\ \beta_2 & \ddots & \ddots & & \\ & \ddots & \ddots & \ddots & \bar{\beta}_n \\ & & \ddots & \ddots & \\ & & & \beta_n & d_n \end{pmatrix}$ et soit $t \notin \bigcup_{i=1}^{n-1} \sigma(a_i)$

si $\tilde{A}$ est la transformée de A par une itération de l'algorithme QR appliqué avec la translation d'origine t, on obtient pour $\tilde{d}_n$ et $\tilde{\beta}_n$ les expressions :

$$\tilde{d}_n = t - \frac{\Psi(t)}{\Psi'(t)}$$

$$\tilde{\beta}_n = \frac{-\Psi(t)}{C_{n-1}\Psi'(t)\Psi_{n-1}(t)}\,\beta_n = \pm\,\frac{[-\Psi'_{n-1}(t)]^{1/2}\Psi(t)}{\Psi'(t)\Psi_{n-1}(t)}\,\beta_n$$

où C_{n-1} est le cosinus de l'angle θ_{n-1} utilisé dans la $n-1^{ème}$ rotation de Givens au cours de la transformation QR.

REMARQUE :

La démonstration ci-dessous se fait seulement dans le cas où $t \notin \bigcup_{i=1}^{n-1} \sigma(a_i)$. Mais en fait les Ψ_i étant $\mathcal{C}^\infty$ dans $\mathbb{R} - \sigma(a_{i-1})$, on peut prolonger par continuité les formules ci-dessus partout où elles ont un sens c'est-à-dire
- dans $\{t \in \mathbb{R} ;\ t \notin \sigma(a_{i-1})\}$ pour la formule donnant $\tilde{d}_n$
- et dans $\{t \in \mathbb{R} ;\ t \notin \sigma(a_{n-1}) \cup \sigma(a_{n-2})\}$ pour la formule donnant $\tilde{\beta}_n$.

DEMONSTRATION :

On peut simplifier la démonstration sans changer la généralité du résultat, en prenant t = 0 et en supposant que A est (symétrique) réelle.

La méthode de Givens utilise pour transmuer A par l'algorithme QR, les matrices de rotation :

$$P_i = \begin{bmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1\ C_i & +S_i \cdots & \\ & & -S_i & C_i \cdots & \\ & & & & 1 \\ & & & & & \ddots \\ & & & & & & 1 \end{bmatrix} \begin{matrix} \\ \\ (i-1)^{\text{ème}}\ \text{ligne} \\ i^{\text{ème}}\ \text{ligne} \\ \\ \\ \\ \end{matrix}$$

avec $C_i = \cos \theta_i$

$S_i = \sin \theta_i$

Pour i=2,3,...,n , on choisit les angles θ_i de façon à ce que la matrice

$$R_i = P_i P_{i-1} \cdots\cdots P_1 A$$

soit de la forme (1) ci-dessous.

Nous allons utiliser ces matrices pour la démonstration du résultat On convient de prendre pour première rotation $\theta_1 = 0$ i.e. $P_1 = I$.

$$R_i = \begin{bmatrix} r_1 & \beta''_2 & & x_3 & & \\ & r_2 & \ddots & & x_4 & \\ & & \ddots & r_{i-1} & \beta''_i & \ddots \\ & & & & & \\ & & & d'_i & & \beta'_{i+1} \\ & & & \beta_{i+1} & d_{i+1} & & \beta_{i+2} \\ & & & & \ddots & \ddots & \ddots \\ & & & & & d_{n-1} & \beta_n \\ & & & & & & d_n \end{bmatrix} \quad (1)$$

Au pas suivant pour avoir R_{i+1} de la forme ci-dessus avec un zéro à la place de β_{i+1}, il suffit de prendre P_{i+1} avec les paramètres :

$$\begin{cases} C_{i+1} = \dfrac{d'_i}{r_i} & (2) \\ \qquad\qquad \text{où } r_i = \sqrt{d'^2_i + \beta^2_{i+1}} \\ S_{i+1} = \dfrac{\beta_{i+1}}{r_i} & (3) \end{cases}$$

finalement $R = R_n$ sera de la forme :

$$R = \begin{pmatrix} r_1 & \beta''_2 & x_3 & & \\ & r_2 & \beta''_3 & \ddots & x_n \\ & & \ddots & \ddots & \\ & O & & r_{n-1} & \beta''_n \\ & & & & \beta''_n \\ & & & & d'_n \end{pmatrix} \qquad (4)$$

et $\tilde{A} = RP_1^T P_2^T \ldots P_n^T$

on a :

$$\beta''_i = C_i \beta'_i + S_i d_i \qquad (5)$$

$$d'_i = C_i d_i - S_i \beta'_i \qquad (6)$$

$$\beta'_{i+1} = C_i \beta_{i+1} \qquad (\text{pour } i < n) \qquad (7)$$

$$r_i = \sqrt{\beta^2_{i+1} + (d'_i)^2} \qquad (8)$$

La post-multiplication par $P_2^T \ldots P_n^T$ donne :

$$i=1,\ldots,n-1 \qquad \tilde{d}_i = C_i d'_i + S_{i+1} \beta''_{i+1} \qquad (9)$$

$$i=2,\ldots,n-1 \qquad \tilde{\beta}_i = S_i r_i \qquad (10)$$

$$\tilde{d}_n = C_n d'_n \qquad (11)$$

$$\tilde{\beta}_n = S_n d'_n \qquad (12)$$

a) On va montrer par récurrence que $d'_i = C_i \Psi_i(0)$

c'est vrai pour $i = 1$ $(d'_1 = d_1 = C_1 \Psi_1(0) = d_1)$

D'après (4) $d'_i = C_i d_i - S_i \beta'_i$

(1) et (2) $\Rightarrow S_i = C_i \dfrac{\beta_i}{d'_{i-1}}$

(5) $\Rightarrow \beta'_i = C_{i-1} \beta_i$

$$\Rightarrow \quad d'_i = C_i \left(d_i - \frac{\beta_i^2}{d'_{i-1} / C_{i-1}} \right) \qquad (13)$$

Si on suppose maintenant que $d'_{i-1} = C_{i-1} \Psi_{i-1}(0)$ (14)

(13) montre que $d'_i = C_i \left(d_i - \dfrac{\beta_i^2}{\Psi_{i-1}(0)} \right) = C_i \Psi_i(0)$

b) Montrons que :

$$C_i^2 = - \frac{1}{\Psi'_i(0)}$$

Il suffira de montrer que $\Psi'_i(0) = - 1 - tg^2(\theta_i)$

(en effet $\cos^2\theta = \dfrac{1}{1+tg^2\, a}$)

D'après les formules de récurrence (2.1) on a ceci :

$$\Psi'_j(0) = - 1 + \frac{\beta_j^2}{\Psi_{j-1}^2(0)} \Psi'_{j-1}(0) \qquad (15)$$

D'autre part d'après (1) et (2) on a :

$$tg\ \theta_j = \frac{\beta_j}{d'_{j-1}} \qquad \text{comme} \quad d'_{j-1} = C_{j-1} \Psi_{j-1}(0)$$

ceci donne :

$$tg\ \theta_j = \frac{\beta_j}{C_{j-1} \Psi_{j-1}(0)} \qquad (16)$$

Le résultat se démontre alors facilement par récurrence :

* C'est vrai pour i = 1 puisque $\theta_1 = 0$ et que $\Psi_1'(0) = -1$

et donc $\Psi_1'(0) = -1 - tg^2\,\theta_1$

* Supposons le résultat vrai jusqu'à i-1 alors :

$$\Psi_{i-1}'(0) = -\frac{1}{C_{i-1}^2}$$

d'après (15) on a :

$$\Psi_i'(0) = -1 + \frac{\beta_i^2}{\Psi_{i-1}^2(0)} \quad \Psi_{i-1}'(0) = -1 - \frac{\beta_i^2}{\Psi_{i-1}^2(0)\, C_{i-1}^2}$$

et d'après (16)

$$\Psi_i'(0) = -1 - tg^2\,\theta_i \ .$$

c) Le résultat de la proposition 9 découle alors de (13), (14), et de a) et b).

C.Q.F.D.

REMARQUE :

Il est possible de démontrer en fait des formules analogues pour tous les éléments de $\tilde{A}$ (autres que ceux de la dernière ligne).

Pour LR :

(15) pour i=1,2,...,n-1

$$\begin{cases} \tilde{\beta}_{i+1} = \dfrac{\Psi_{i+1}(t)}{\Psi_i(t)}\ \beta_{i+1} \\ \tilde{d}_i = \Psi_i(t) + t + \dfrac{\beta_{i+1}\gamma_{i+1}}{\Psi_i(t)} \\ \tilde{\gamma}_{i+1} = \gamma_{i+1} \end{cases}$$

Pour QR

$$(15) \qquad \text{pour } i=1,2,\ldots,n-1 \quad \begin{cases} \tilde{d}_i = t - \dfrac{\Psi_i(t)}{\Psi_i'(t)} + d_{i+1} + \dfrac{\Psi_{i+1}(t)}{\Psi_{i+1}'(t)} \\ \tilde{\beta}_i = \dfrac{-\Psi_i(t)}{c_{i-1}c_{i+1}\Psi_{i-1}(t)\Psi_i'(t)}\beta_i = \end{cases}$$

$$= \frac{\pm\left[\Psi_{i-1}'(t)\Psi_{i+1}'(t)\right]^{1/2}\Psi_i(t)}{\Psi_{i-1}(t)\Psi_{i+1}(t)}\beta_i$$

Ceci a peu d'intérêt pour ce quenous voulons étudier puisque en pratique on procède par déflation et c'est donc les comportements de $\tilde{d}_n$ et $\tilde{\beta}_n$ qui sont primordiaux.

2.4. APPLICATION A LA COMPARAISON DES COMPORTEMENTS DE LR ET QR.

Il nous est impossible, à l'aide des résultats établis, de comparer globalement les algorithmes LR et QR avec translations. Nous ne pouvons donner que des indications sur les comportements de ces deux algorithmes. Ceci est dû en particulier au fait que LR ne conserve pas la forme symétrique tandis que QR ne conserve pas la forme tridiagonale non symétrique.

- a - COMPARAISON DE $\tilde{d}_n$ POUR LR ET QR

D'après la proposition 9, les meilleures translations t que l'on puisse prendre, sont celles pour lesqelles $\Psi(t) = 0$ (autrement dit les valeurs propres de A !) puisqu'alors $\tilde{\beta}_n = 0$.

Nous allons voir que pour le terme diagonal, une itération de LR où QR correspond à la résolution approchée de $\Psi(t) = 0$.

* Pour LR, $\tilde{d}_n(LR) = \Psi(t)+t$ correspond à un pas de la méthode des approximations successives appliquée à la résolution de $\Psi(x)+x = x$ à partir de $x = t$.

* Pour QR, $\tilde{d}_n(QR) = t - \dfrac{\Psi(t)}{\Psi'(t)}$ correspond à un pas de la méthode de Newton appliquée à $\Psi(x) = 0$ à partir de $x = t$.

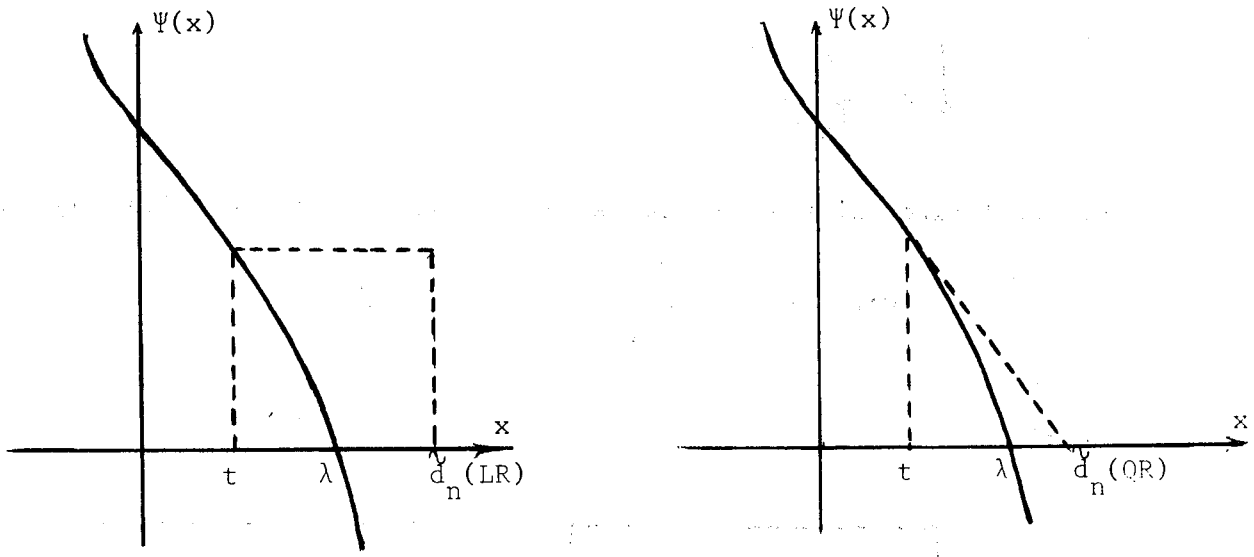


Figure 2

La figure 2 indique que dans le cas général $\tilde{d}_n(QR)$ est plus voisin d'une valeur propre λ que $\tilde{d}_n(LR)$.

- b - COMPARAISON DE $\tilde{\beta}_n$ POUR LR ET POUR QR

$$\text{On a } \tilde{\beta}_n(LR) = \frac{\Psi(t)}{\Psi_{n-1}(t)}\beta_n \quad \text{et} \quad \tilde{\beta}_n(QR) = \frac{-1}{C_{n-1}\Psi'(t)} \cdot \frac{\Psi(t)}{\Psi_{n-1}(t)} \cdot \beta_n$$

On pose pour simplifier les notations :

$$\begin{cases} T = \dfrac{1}{C_{n-1}} \\ u = \dfrac{\beta_n}{\Psi_{n-1}(t)} \end{cases}$$

On peut tirer aisément de la démonstration de la proposition 9 que :

$$-\Psi'(t) = 1 + \frac{\beta_n^2}{\Psi_{n-1}^2(t)}\, T^2 = 1 + T^2u^2$$

d'où :

$$\boxed{\tilde{\beta}_n(QR) = \frac{Tu}{1+T^2u^2}\Psi(t)} \quad \text{et} \quad \boxed{\tilde{\beta}_n(LR) = u\,\Psi(t)}$$

Les deux coefficients de $\Psi(t)$ dans les expressions de $\tilde{\beta}_n(LR)$ et $\tilde{\beta}_n(QR)$ considérés comme des fonctions de u sont respectivement :

$$\ell(u) = \frac{\tilde{\beta}_n(LR)}{\Psi(t)} = u$$

$$q(u) = \frac{\tilde{\beta}(QR)}{\Psi(t)} = \frac{Tu}{1+T^2u^2}$$

Représentons graphiquement ces deux fonctions :

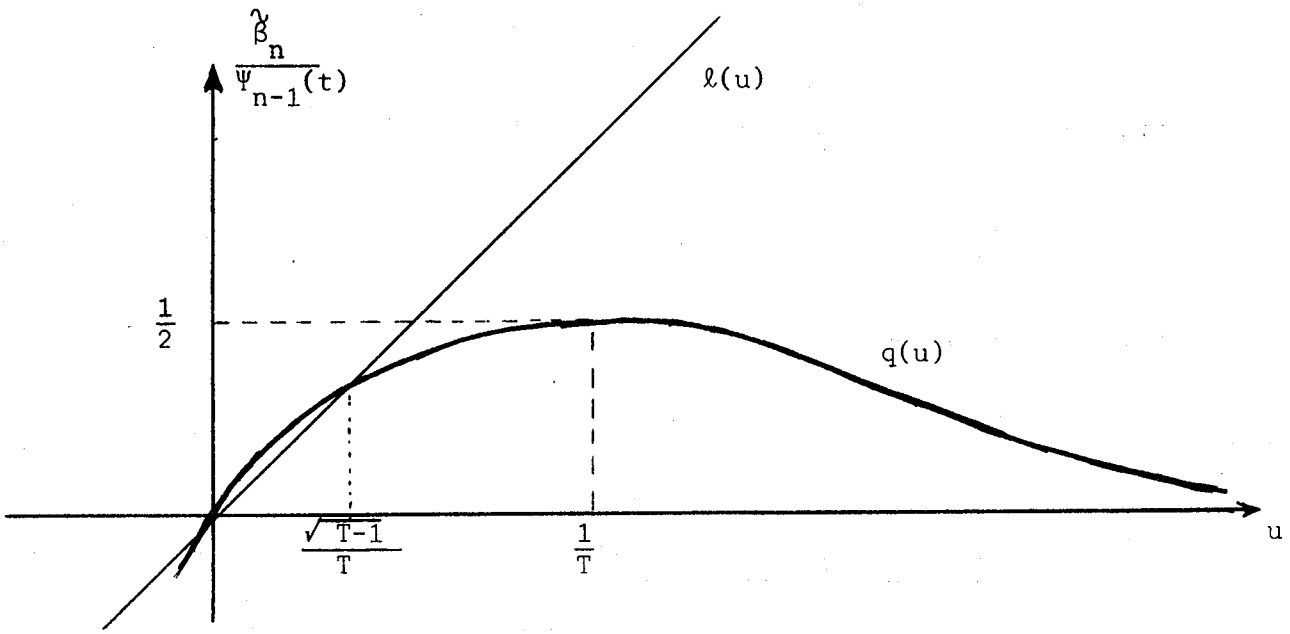


Figure 3

On constate, d'après la figure 3, que si $u = \frac{\beta_n}{\Psi_{n-1}(t)}$ "tend vers l'infini", ce qui correspond à un mauvais choix de t ($\Psi_{n-1}(t) \to 0$) , alors pour LR, $\ell(u)$ tend vers l'infini, alors que pour QR, le coefficient q(u) tend vers zéro. Voyons un exemple simple :

Soit $A = \left(\begin{array}{c|c} 10^{-3} & 10^{2} \\ \hline 10^{2} & 0 \end{array}\right)$

Prenons $t = 0$ (t voisin de 10^{-3}) .

$$u = \frac{\beta_n}{\Psi_{n-1}(t)} = \frac{10^2}{+10^{-3}} = + 10^5$$

$$T_1 = \frac{1}{\cos \theta_1} \qquad \text{d'où} \quad \begin{cases} \ell(u) = 10^5 \\ q(u) \underset{\sim}{>} 10^{-5} \end{cases}$$

Comme $\Psi(t) = d_n - t - \frac{\beta_n^2}{\Psi_{n-1}(t)} = - 10^7$ cela donne :

$$\tilde{\beta}_n(LR) = - 10^{12} \quad , \quad \tilde{\beta}_n(QR) = - 10^2$$

De manière analogue, on trouve :

$$\tilde{d}_n(LR) = -10^7 \quad , \quad \tilde{d}_n(QR) = - 10^{-3}$$

- c - ETUDE DU CAS PARTICULIER $t = d_n$

On trouve alors :

$$\text{pour LR} \begin{cases} \tilde{d}_n = d_n - u\gamma_n \\ \tilde{\beta}_n = - u^2 \gamma_n \end{cases} , \quad \text{pour QR} \begin{cases} \tilde{d}_n - d_n - \frac{u}{1+T\,u}\ \beta_n \\ \tilde{\beta}_n = \frac{Tu^2}{1+t^2u^2}\ \beta_n \end{cases}$$

Ainsi pour QR on montre les résultats suivants :

i) $|\tilde{d}_n - d_n| \leq \frac{1}{2} |c_{n-1}||\beta_n|$

ii) $|\tilde{\beta}_n| \leq |c_{n-1}||\beta_n|$

iii) Lorsqu'il y a convergence, celle-ci est cubique à la limite.

Les résultats ci-dessus sont faciles à montrer.

REMARQUE :

iii) et ii) ont été démontrés par Wilkinson en [14] et [16].

On voit sur l'exemple du -b- comment se comporterait LR avec la même stratégie $t = d_n$.
En effet dans l'exemple on a pris $t = d_n = 0$ et on a obtenu

$$\tilde{\beta}_n(LR) = 10^{12} \quad \text{et} \quad \tilde{d}_n(LR) = -10^{7}) \ .$$

L'exemple correspond à un mauvais choix de t : $t = 0$ est voisin de l'asymptote $x = 10^{-3}$.

Il se produit un "saut" pour LR.
Pour QR avec la stratégie $t = d_n$, i) et ii) montrent qu'il ne peut pas y avoir de "saut".

Il en résulte que QR est plus stable que LR du point de vue numérique .(Ceci est un fait bien connu).

3. APPLICATION A L'ACCELERATION DE L'ALGORITHME QR PAR LES TRANSLATIONS D'ORIGINE.

3.1.

Puisque le meilleur choix de la translation d'origine t consiste à prendre t tel que $\Psi(t) = 0$, on peut proposer en plus des stratégies usuelles décrites en [14], certaines stratégies consistant à résoudre approximativement l'équation $\Psi(x) = 0$: il suffira de faire quelques itérations du procédé de Newton appliqué à cette équation et de prendre la solution approchée comme translation.

Diverses stratégies ont été étudiées en [13], nous reprenons ici les plus intéressantes du point de vue pratique.

. On note $\Psi_{Newt}(x)$ la fonction $\Psi_{Newt}(x) = x - \frac{\Psi(x)}{\Psi'(x)}$

. μ désignera la valeur propre de $\begin{pmatrix} d_{n-1} & \bar{\beta}_n \\ \beta_n & d_n \end{pmatrix}$ qui est la plus voisine de d_n (ou encore le nombre $d_n - \frac{|\beta_n|^2}{d_{n-1}-d_n}$ qui en est une approximation.)

STRATEGIES CLASSIQUES

T1 : $t = d_n$

T2 : $t = \mu$

STRATEGIES PROPOSEES

T3 : $t = \Psi_{Newt}(\mu)$

T4 : $t = \lambda_\varepsilon$,
solution approchée de $\Psi(x) = 0$ calculée par le procédé de Newton : en partant de $t^{(0)} = \mu$ faire $t^{(s+1)} = \Psi_{Newt}(t^{(s)})$ jusqu'à ce que $|t^{(s+1)}-t^{(s)}| \leq \varepsilon$.

Puisque μ est la valeur propre de $\begin{pmatrix} d_{n-1} & \bar{\beta}_n \\ \beta_n & d_n \end{pmatrix}$, la plus voisine de d_n, on constate que si on note $\Psi^{[k]}_{Newt}(x)$, la fonction associée à la sous-matrice principale inférieure d'ordre k de A, de la même manière que Ψ est associée à A, μ n'est autre que $\Psi^{[2]}_{Newt}(\mu)$. Il est donc intéressant de généraliser les stratégies T3, T4, en choisissant comme valeur initiale au lieu de μ , le nombre μ' défini par :

$$\mu' = \Psi^{[k]}_{Newt}(\Psi^{[k']}_{Newt}(\mu))$$

où k et k' sont convenablement choisis.

(On suppose que $1 \leq k' \leq k \leq n-1$).

Si λ est la valeur propre de A la plus voisine de d_n, alors μ' est plus proche de λ que μ dès que β_n est assez petit.

Cette nouvelle valeur μ' nous conduit aux stratégies suivantes plus efficaces du point de vue pratique, que les précédentes :

T5 : $t = \Psi_{Newt}(\mu')$

T6 : Faire T4 en partant de $t^{(0)} = \mu'$

La proposition suivante décrit les comportements asymptotiques de l'algorithme QR avec les stratégies T3, T4, T5, T6 :

PROPOSITION 10 :

i) Soit t une translation d'origine provenant de $t = \Psi_{Newt}(t')$ où t' est tel que $|t-t'| \leq \varepsilon$. Supposons qu'il n'y ait pas de pôle de la fonction Ψ entre t et t'.
Alors il existe θ , $0 < \theta < 1$ tel que si $\xi = t'+\theta(t-t')$ on ait :

$$|\hat{\beta}_n| \leq \frac{\varepsilon^2}{4} |\Psi''(\xi)|$$

ii) Supposons que l'algorithme QR avec la stratégie $t = \Psi_{Newt}(d_n)$ converge, alors la convergence est d'ordre 7 à la limite.

iii) Supposons que l'algorithme QR avec la stratégie T3 converge, alors la convergence est d'ordre 7 à la limite, sauf si $d_n^{(s)} - d_{n-1}^{(s)}$ tend vers zéro, auquel cas elle est d'ordre 5.

DEMONSTRATION :

i) D'après la proposition 9 et en utilisant les notations du paragraphe 2, on a :

$$\hat{\beta}_n = \frac{Tu}{1+T^2u^2}\ \Psi(t)$$

d'où :

$$|\beta_n| \leq \frac{1}{2}\,|\Psi(t)| \qquad (1)$$

S'il n'y a pas d'asymptote entre t et t' on peut écrire d'après la formule de Taylor que :

$$\Psi(t) = \Psi(t')+(t'-t)\Psi'(t') + \frac{1}{2!}\,(t'-t)^2\Psi''(\xi)\ ,$$

où ξ est un nombre compris entre t et t'.

Comme $t = t' - \dfrac{\Psi(t')}{\Psi'(t')}$ le terme $\Psi(t')+(t'-t)\Psi'(t')$ est nul, donc :

$$\Psi(t) = \frac{1}{2!}\,(t'-t)^2\Psi''(\xi) \qquad (2)$$

Le résultat i) découle alors de (1) de (2) et du fait que $|t-t'| \leq \varepsilon$.

ii) Dans tout ce qui suit, l'indice s est supprimé pour simplifier les notations . Ainsi, $\beta_n \to 0$ voudra dire $\beta_n^{(s)} \to 0$ lorsque $s \to \infty$, et $d_n \to \lambda$ voudra dire $\underset{s\to\infty}{d_n^{(s)}} \to \lambda$.

Nous utiliserons souvent le fait que si $A = \left(\begin{array}{c|c} a_{n-1} & b_n^H \\ \hline b_n & d_n \end{array}\right)$

est tridiagonale hermitienne et $|\beta_i| \neq 0$, i=2,...,n alors :

les valeurs propres $\lambda_j^{(n-1)}$ de a_{n-1} et les valeurs propres λ_j de A vérifient lorsqu'elles sont ordonnées en croissant [13] :

$$\lambda_{j-1} < \lambda_{j-1}^{(n-1)} < \lambda_j \qquad (3)$$

Ainsi toutes les matrices A_s générées par l'algorithme n'ont pas de valeur propre multiple et la limite $\bar{A}$ également.

. Pour toutes les stratégies proposées, t est choisi de la forme $t = \Psi_{Newt}(t')$. On peut donc écrire l'égalité (2), puisqu'à la limite, il n'y aura pas d'asymptote entre t et t' :

$$\Psi(t) = \frac{1}{2}(t-t')^2\Psi''(\xi) \qquad (4)$$

D'après l'expression de la dérivée :

$$\Psi'(x) = -1 - |\beta_n|^2 \tilde{e}_{n-1}^T (a_{n-1}-x)^{-2} \tilde{e}_{n-1}$$

on aura

$$\Psi''(\xi) = +2|\beta_n|^2 \tilde{e}_{n-1}^T (a_{n-1}-\xi)^{-3} \tilde{e}_{n-1}$$

D'autre part si $t' = d_n$

$$t = \Psi_{Newt}(d_n) = d_n - \frac{\Psi(d_n)}{\Psi'(d_n)}$$

d'où :

$$t - t' = -\frac{\Psi(d_n)}{\Psi'(d_n)}$$

or on peut écrire :

$$\frac{\Psi(d_n)}{\Psi'(d_n)} = |\beta_n|^2 \frac{1}{\Psi_{n-1}(d_n)\Psi'(d_n)}$$

$$t - t' = -\frac{\Psi(d_n)}{\Psi'(d_n)} = -v_n|\beta_n|^2 \quad \left(\text{où } v_n = \frac{1}{\Psi_{n-1}(d_n)\Psi'(d_n)}\right)$$

L'expression 4 s'écrit donc

$$\Psi(t) = v_n^2 \, \tilde{e}_{n-1}^T (a_{n-1}-\xi)^{-3} \tilde{e}_{n-1} |\beta_n|^6$$

et la proposition 9 (résultat concernant $\tilde{\beta}_n$) donne alors :

$$\tilde{\beta}_n = \frac{-v_n^2 \, \tilde{e}_{n-1}^T (a_{n-1}-\xi 1_{n-1})^{-3} \tilde{e}_{n-1} \, \beta_n |\beta_n|^6}{C_{n-1} \Psi_{n-1}(t) \Psi'(t)} \qquad (5)$$

donc $\quad |\tilde{\beta}_n| = K|\beta_n|^7$

et <u>il faut montrer que s'il y a convergence</u> : K ne tend pas vers l'infini.

Pour cela on montrera que :

a) $C_{n-1} \, \Psi_{n-1}(t)$ ne tend pas vers zéro.

b) $\tilde{e}_{n-1}^T (a_{n-1}-\xi)^{-3} \tilde{e}_{n-1}$ ne tend pas vers l'infini.

c) v_n est borné. (ou $|\frac{v_n}{\beta_n}|$ si $d_{n-1}-d_n \to 0$ pour iii).

(Le terme $|\Psi'(t)|$ est > 1) .

Montrons d'abord b) si $\tilde{e}_{n-1}^T (a_{n-1}-\xi 1_{n-1})^{-3} \tilde{e}_{n-1}$ tendait vers l'infini, cela voudrait dire que, à la limite, ξ serait valeur propre de a_{n-1}. Or comme $t < \xi < t'$ et que t et t' convergent vers une valeur propre $\bar{\lambda}$ de A (puisque A converge et $\beta_n \to 0$), cela voudrait dire que la matrice limite $\bar{A}$ aurait la forme

$$\bar{A} = \left(\begin{array}{c|c} \bar{a}_{n-1} & 0 \\ \hline 0 & \bar{\lambda} \end{array}\right) \qquad (5')$$

où $\bar{\lambda}$ serait une valeur propre de $\bar{a}_{n-1}$.

i.e. $\bar{\lambda}$ est une valeur propre multiple de $\bar{A}$: ceci <u>contredirait les inégalités</u> (3) .

a). Reprenons les formules et les notations de la proposition 9. Il est facile de voir que si A converge et si $\beta_n \to 0$ alors $C_n \to 1$. (et $S_n \to 0$) . [Le produit des rotations P_i est en effet de la forme :

$$Q = \left(\begin{array}{c|c} \multicolumn{2}{c}{\multirow{2}{*}{\quad}} & \\ & & \\ \hline 0 & -S_n & C_n \end{array}\right) \qquad]$$

La relation (14) s'écrit $d'_{n-1} = C_{n-1}\,\Psi_{n-1}(t)$ (5)

la relation (8) s'écrit $r_{n-1} = \sqrt{(d'_{n-1})^2+|\beta_n^2|}$ (6)

la relation (10) s'écrit $\overset{\gamma}{\beta}_{n-1} = s_{n-1}r_{n-1}$ (7)

enfin de (9) on tire :

$$\overset{\gamma}{d}_{n-1} = C_{n-1}d'_{n-1} + s_n\beta''_n + t \qquad (8)$$

Si $C_{n-1}\,\Psi_{n-1}(t) = d'_{n-1}$ tendait vers zéro cela entraînerait que r_{n-1} aussi, par (6), donc de (7) on déduit que $\beta_{n-1} \to 0$. De 8 on déduit que $\overset{\gamma}{d}_{n-1} \to t$) (β''_n est borné sinon la matrice de départ A serait semblable à une suite de matrices dont la norme tend vers l'infini).
$d_n \to t$ donc finalement la matrice limite $\bar{A}$ aurait la forme suivante :

$$\left(\begin{array}{c|cc} \bar{a}_{n-1} & 0 & \\ \hline 0 & t & 0 \\ & 0 & t \end{array}\right)$$

et ceci est impossible pour la même raison qu'en b).

c)

$$|v_n| = \left|\frac{1}{\Psi_{n-1}(d_n)\Psi'(d_n)}\right| < \frac{1}{|\Psi_{n-1}(d_n)|}$$

Or $\Psi_{n-1}(d_n)$ ne peut pas tendre vers zéro sinon on aurait à la limite une matrice de la forme (5') et on a montré que c'est impossible en b).

iii) La démonstration ci-dessus est valable sauf pour la majoration de v_n : Si $t' = \mu$ on a :

$$v_n = \frac{1}{\Psi'(\mu)}\left(\frac{1}{d_{n-1}-\mu} - \frac{1}{\Psi_{n-1}(\mu)}\right)$$

Si $d_{n-1} - \mu \neq 0$ (ce qui est équivalent à $d_{n-1} - d_n \neq 0$), la démonstration du ii,c) ci-dessus pour $\Psi_{n-1}(d_n)$, est valable pour $\Psi_{n-1}(\mu)$ et v_n est borné à la limite dans ce cas.

Si au contraire $d_{n-1} - \mu \to 0$ on peut utiliser le fait que $\left|\frac{\beta_n}{d_{n-1}-\mu}\right| \leq 1$ (en effet μ est la valeur propre de $\begin{pmatrix} d_{n-1} & \bar{\beta}_n \\ \beta_n & d_n \end{pmatrix}$ la plus voisine de d_n et la relation $|\beta_n|^2 = (d_{n-1}-\mu)(d_n - \mu)$ donne l'inégalité).

Donc :

$$\frac{v_n}{\beta_n} = \frac{1}{\Psi'(t)}\left(\frac{\beta_n}{d_{n-1}-\mu} - \frac{\beta_n}{\Psi'_{n-1}(\mu)}\right)$$

est borné. Ce qui, en raison de la relation 5, donne $|\tilde{\beta}_n| = K|\beta_n|^5$ où K est borné. On a donc une convergence d'ordre 5 ici.

C.Q.F.D.

REMARQUES :

1°) On peut montrer que $d_n^{(s)} - d_{n-1}^{(s)}$ tend vers zéro si et seulement si $d_n^{(s)}$ et $d_{n-1}^{(n)}$ tendent vers zéro.

Le cas où $d_n^{(s)} - d_{n-1}^{(s)}$ tend vers zéro est délicat : Parlett, Kahan [12] ont rencontré le même problème pour la translation $t = \mu$. On ne sait pas si un tel cas est possible. Cependant si la matrice initiale A est régulière une telle éventualité est impossible, sinon cela voudrait dire que A est semblable à la matrice $\bar{A}$ dont la dernière ligne est nulle.

2°) La stratégie T4 devient, à la limite, simplement la stratégie T3.

3.2. COMPORTEMENT DE L'ALGORITHME QR AVEC LA STRATEGIE T4.

Cet algorithme peut être considéré comme la combinaison de deux algorithmes :

- l'algorithme QR
- l'algorithme de calcul à la'précision' ε d'un zéro de la fonction Ψ par la méthode de Newton.

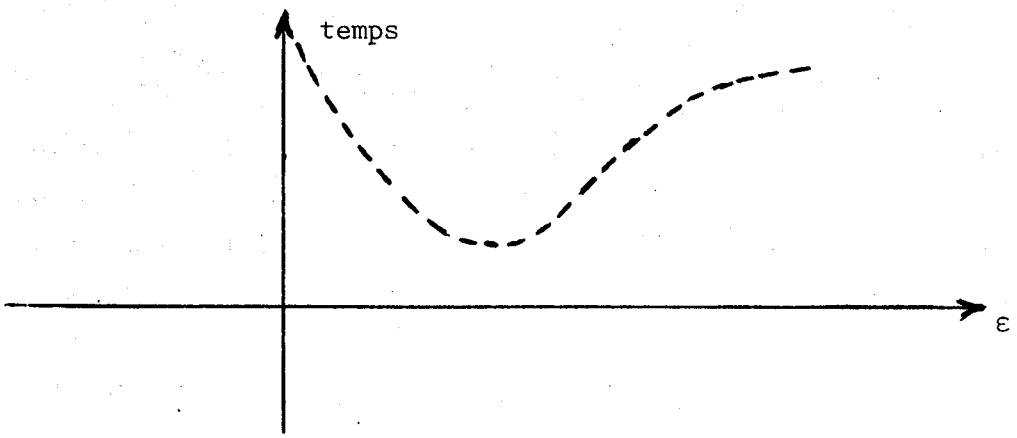
Cette deuxième étape est exactement la méthode d'Abramov accélérée par le processus de Newton(cf. paragraphe 2) avec le partitionnement $m = n-1$, $p = 1$.

$$A = \left(\begin{array}{c|c} a & b \\ \hline c & d \end{array}\right) .$$

La fonction matricielle d(z) est ici réduite à une fonction scalaire

Nous avons essayé d'autres stratégies, utilisant l'algorithme d'Abramov-Chichov classique (par exemple des stratégies telles que $t = \Psi(\mu)+\mu$), mais les résultats étant nettement inférieurs à ceux que nous avons décrit, nous les avons abandonnées.

Nous avons constaté expérimentalement que le temps d'éxécution du programme permettant de calculer toutes les valeurs propres a l'allure suivante :



Ceci est dû au fait que le temps total $T(\varepsilon)$ d'exécution du programme se décompose ainsi :

1 - Le temps $T_1(\varepsilon)$ mis pour calculer toutes les translations t. On fait d'abord toujours $t^{(1)} = \Psi_{Newt}(\mu)$ et puis $t^{(s+1)} = \Psi_{Newt}(t^{(s)})$ jusqu'à ce que $|t^{(s+1)}-t^{(s)}| < \varepsilon$.

Donc : . si $\varepsilon \to \infty$, $T_1(\varepsilon)$ tendra vers une constante (correspondant au calcul de $t^{(1)}$)

. si $\varepsilon \to 0$, $T_1(\varepsilon)$ deviendra infiniment grand.

2 - Le temps $T_2(\varepsilon)$ mis pour exécuter toutes les transmutations. Il diminue avec le nombre d'itérations (donc avec ε).

En faisant la somme de ces deux temps on obtient une courbe du type suivant : $(T(\varepsilon) = T_1(\varepsilon)+T_2(\varepsilon))$:

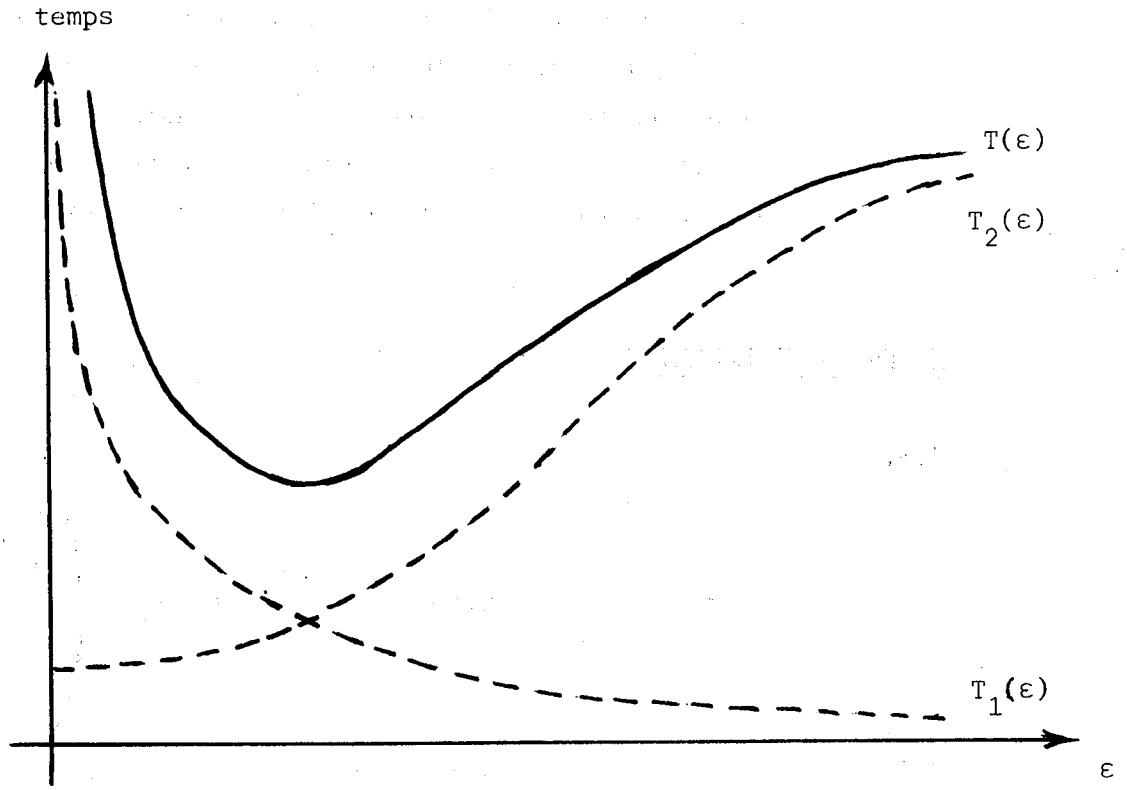


Figure 4

) On constate par ce "raisonnement" qu'il existe un "ε optimal".

b) La convergence étant d'ordre 7 à la limite :

Il est évident que le nombre d'itérations sera considérablement diminué. Il en résulte :

- un gain de temps souvent non négligeable (ceci pour ε bien choisi),
- une réduction des erreurs d'accumulation (car il y a beaucoup moins d'itérations),
- dans le cas très courant où l'on calcule simultanément les vecteurs propres et les valeurs propres, aux transformations QR vont s'ajouter $4KN^2$ opérations (où K = nombre total d'itérations) et

on remarque en faisant un raisonnement analogue à ci-dessus qu'on gagne encore plus de temps dans ce cas et qu'il est plus aisé d'estimer un ε permettant de réduire le temps d'exécution. (On constate expérimentalement que le gain relatif en temps est voisin du gain relatif en itérations, cf. exemples numériques.)

3.3. EXPERIENCES NUMERIQUES.

3.3.1.

On a pris les matrices tridiagonales $D_N = \begin{pmatrix} 2 & -1 & & & \\ -1 & \ddots & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & -1 \\ & & & -1 & 2 \end{pmatrix}$

d'ordre N. i.e. $\beta_i = -1$, $i=2,\ldots,N$

et $d_i = 2$, $i=1,\ldots,N$;

dont les valeurs propres sont $\lambda_i = 2(1-\cos(\frac{i\pi}{N+1})) = 4 \sin^2 (\frac{i\pi}{2(N+1)})$

Pour N assez grand il y aura donc aux extrémités du spectre des valeurs propres très voisines. (C'est donc un cas défavorable pour QR).
Dans tous les exemples qui suivent, on a procédé par déflation : on a donc travaillé successivement sur les matrices déflatées de dimension $\ell = N, N-1, \ldots, N-P, \ldots, 2$. Le test d'arrêt choisi est $|\beta_\ell| < 10^{-9}$.
Pour N = 89 les résultats sont ceux du tableau 1.

On a constaté que le gain en temps augmente avec la dimension de la matrice. En prenant N = 130 on a comparé les stratégies 2 et 5

[on a pris $\mu' = \psi^{[k]}_{Newt}(\psi^{[k']}_{Newt}(\mu))$] .

(Résultats dans le tableau 2).

TABLEAU 1

N = 89 - Calcul des valeurs propres.

stratégie employée	nombre d'itérations	temps d'exécution (secondes)	gain relatif en itérations par rapport à la stratégie T2	gain relatif en temps par rapport à la stratégie T2
T1	179	3,889		
T2	176	3,781		
T3	133	3,558	24 %	6 %
T4 avec $\varepsilon = 10^{-4}$	115	3,473	35 %	8 %
T5 avec $\mu' = \Psi^{[10]}_{Newt}(\mu)$	113	3,184	37 %	16 %

TABLEAU 2

N = 130 - Calcul des valeurs propres.

T2	247	7,502		
T6	173	5,544	30 %	26 %

TABLEAU 3

N = 40 - Calcul des valeurs propres et vecteurs propres.

T2	80	14,63		
T5 avec le test d'arrêt $\lvert\beta_\ell(t^{(s+1)}-t^{(s)})\rvert < 10^{-6}$	44	8,37	45 %	43 %

Le cas le plus intéressant d'après le troisième point de la remarque III.b est celui où l'on calcule simultanément les valeurs propres et les vecteurs propres. On constate sur les résultats du tableau 3 (pour N = 40) que le gain relatif en temps est effectivement voisin du gain relatif en itérations.

3.3.2.

Soit B_N la matrice tridiagonale symétrique d'ordre N où

$$d_i = N \qquad i=1,2,\ldots,N$$

$$\beta_{i+1} = \sqrt{i(N-i)} \qquad i=1,2,\ldots,N-1.$$

Les valeurs propres de B_N sont connues [9] : $\lambda_i = 2i-1$, $i=1,2,\ldots,N$.
Nous avons testé sur ces matrices la stratégies T_5 dans laquelle les paramètres k et k' sont déterminés de la manière suivante :

> Si ℓ est la dimension de la matrice déflatée et si $\ell \geq 10$ alors $k' = [\ell/10]$, (division entière de ℓ par 10) et $k = 3\ell$, sinon si $\ell < 10$ $\quad k' = k = 1$.

On a pris successivement N = 120, 150, 200, 250 et 300.
Ici seules les valeurs propres sont calculées. Nous comparons alors pour les translations T2 et T5, les nombres d'itérations (colonne B) les temps d'exécution (colonne C), les gains relatifs en itérations (colonne D) et en temps (colonne E) par rapport à T2 et enfin les précisions moyennes (colonne F).

(Si μ_i est la $i^{ème}$ valeur propre calculée, la précision moyenne est le nombre $\frac{1}{N} \sum_{i=1}^{N} \frac{|\mu_i-\lambda_i|}{|\lambda_i|}$) .

Les résultats sont ceux du tableau 4.

TABLEAU 4

Calcul des valeurs propres

	A	B	C	D	E	F
N=120	T2	237	7,004			$5,091 \times 10^{-14}$
	T5	142	5,423	40 %	23 %	$1,186 \times 10^{-14}$
N=150	T2	299	10,500			$7,934 \times 10^{-14}$
	T5	170	7,973	43 %	24 %	$2,854 \times 10^{-14}$
N=200	T2	399	18,691			$1,125 \times 10^{-13}$
	T5	220	13,765	42 %	26 %	$2,861 \times 10^{-14}$
N=250	T2	499	29,159			$1,472 \times 10^{-13}$
	T5	271	22,318	45 %	23 %	$4,316 \times 10^{-14}$
N=300	T2	599	43,042			$1,468 \times 10^{-13}$
	T5	324	31,629	46 %	27 %	$4,62 \times 10^{-14}$

Lorsque nous calculons simultanément les valeurs propres et les vecteurs propres par les méthodes T2 et T6, nous constatons dans nos expériences numériques que la précision sur les vecteurs propres est plus sensible au nombre d'itérations que celle sur les valeurs propres. Alors que nous diminuons le nombre d'itérations de moitié environ, la précision sur les valeurs propres qui est déjà excellente par T2 n'est que légèrement améliorée, alors que pour les vecteurs propres la précision est nettement améliorée. Dans l'exemple qui suit on a pris la matrice B_{50} sur laquelle nous avons calculé les valeurs propres et les vecteurs propres simultanément par QRT6 (QR avec les translations d'origine T6).
k et k' ont été choisis ainsi :

$$k' = [\ell/5] \quad , \quad k = 2 \times k'$$

Dans la stratégie T6 on rencontre en plus le problème du choix de ε : lorsqu'on calcule seulement les valeurs propres, ε ne doit pas être trop petit sinon on risque d'obtenir un algorithme trop coûteux (cf. figure 3).

Par contre dans le cas du calcul simultané des valeurs propres et des vecteurs propres nous avons moins de restrictions. Si le test pour la déflation est :

$|\beta_\ell| < \varepsilon$ on prendra :

- ε du même ordre que ε' ($\varepsilon = \varepsilon'$) si on calcule simultanément les valeurs propres et les vecteurs propres
- ε de l'ordre de $\sqrt{\varepsilon'}$ si on calcule seulement les valeurs propres.

Dans cet exemple on a pris $\varepsilon = \varepsilon'$ avec $\varepsilon' = 10^{-9}$.
Nous avons calculé les nombres $\rho_{B_N} = \left(\sum_{i=1}^{N} \|(B_N - \mu_i)\varphi_i\|^2 \right)^{1/2}$ où μ_i et φ_i sont les valeurs propres et les vecteurs propres de B_N, calculés par les deux méthodes ($\|\varphi_i\| = 1$).

Nous obtenons les résultats du tableau 5 où les colonnes A, B, C, D, E, F ont les mêmes contenus que les tableaux précédents et où la colonne supplémentaire G contient les nombres ρ_{B_N} ci-dessus.

Ces nombres ρ_{B_N} nous donnent une indication sur la précision moyenne des vecteurs propres obtenus par les deux méthodes.

TABLEAU 5

N = 50 - Calcul des valeurs propres et des vecteurs propres

A	B	C	D	E	F	G
T2	100	33,211			$4{,}438 \cdot 10^{-15}$	$1{,}949 \cdot 10^{-9}$
T6	57	19,494	43 %	42 %	$3{,}742 \cdot 10^{-15}$	$9{,}058 \cdot 10^{-13}$

Dans les exemples précédents les calculs ont été effectués en double précision. Pour avoir une idée réelle de l'effet des stratégies T5, T6 sur l'accumulation des erreurs d'arrondis il est bon de donner un exemple de calcul en simple précision : Nous avons comparé T2 à T5 sur la matrice B_{60}.

Temps d'exécution : 1,599 s. pour T2

1,312 s pour T5

Précision obtenue sur les 10 premières valeurs propres :

par QRT2 :	par QRT5 :
0.001038551	0.0002117157
0.0003693898	9.250641'-05
0.0002748489	7.362365'-05
0.0002042225	5.013602'-05
0.0001561907	3.994835'-05
0.0001177354	3.554604'-05
9.059906'-05	3.147125'-05
7.998147'-05	1.570383'-05
6.193272'-05	9.873334'-06

On voit le **gain** appréciable en précision apporté par la stratégie T_5 .

3.3.3.

On va tester ici et dans le 3.3.4. les méthodes QR et Jacobi pour les matrices symétriques pleines.

Sur la matrice A d'ordre 30 (cf. Wilkinson [15]) où $a_{ij} = \max(i,j)$ on a comparé les méthodes suivantes :

1°/ QRT2 : algorithme QR avec la stratégie T2 appliquée à la matrice tridiagonale T obtenue par la méthode de Householder appliquée à

2°/ QRT6 : la même méthode avec la stratégie T6.

3°/ JACOBI (cf. Wilkinson [15]).

Dans les trois cas on calcule simultanément les valeurs propres et les vecteurs propres.

- Temps d'exécution :

QRT2	: 11,11 sec.	(47 itérations)
QRT6	: 8,88 sec.	(37 itérations)
JACOBI	: 29,69 sec.	(2539 rotations planes)

QRT6 est donc le plus rapide des trois algorithmes avec 20 % de moins que QRT2 en temps d'exécution.

REMARQUE :

Le temps d'exécution pour QRT2 et QRT6 comprend :

- le temps mis pour tridiagonaliser A (méthode Householder)
- le temps mis pour calculer les éléments propres de T
- la substitution arrière pour calculer les vecteurs propres dans la base initiale.

Une grande partie du temps est prise par la transformation de Householder. C'est pourquoi le gain relatif en temps d'exécutions est ici moins important que celui de l'exemple 1°/ sur la matrice D_{40} .

- Précision :

- Valeurs propres.
  Ne connaissant pas les valeurs propres exactes de cette matrice on ne peut pas comparer la précision obtenue par les trois méthodes sur les valeurs propres. Toutefois, beaucoup de valeurs propres calculées sont pratiquement identiques.

- Vecteurs propres.
  On a évalué les résidus $\|r_A(U_i)\| = \|(A-\mu_i)U_i\|$ où μ_i et U_i sont les valeurs propres et les vecteurs propres de A pour $i=1,\ldots,N$, avec $\|U_i\| = 1$.

On remarque que les résidus obtenus par QRT2 et QRT6 sont pratiquement identiques (et sont moins bons que ceux obtenus par JACOBI). Ceci est dû essentiellement à la transformation de Householder.
On peut le vérifier si on calcule les résidus des vecteurs propres calculés φ_i de la matrice tridiagonale T (obtenue par tridiagonalisation de Householder) :

$$\|r_T(\varphi_i)\| = \|(T-\mu_i)\varphi_i\| \quad , \qquad i=1,\ldots,N \ .$$

Les nombres $\rho_A = \left(\sum_{i=1}^{N} \|r_A(U_i)\|^2\right)^{1/2}$ et $\rho_T = \left(\sum_{i=1}^{N} \|r_T(\varphi_i)\|^2\right)^{1/2}$

ont été calculés. On obtient :

	ρ_A	ρ_T
QRT2	$9{,}526598 \times 10^{-7}$	$5{,}560... \quad 10^{-10}$
QRT6	$8{,}526596 \times 10^{-7}$	$1{,}807... \quad 10^{-12}$
JACOBI	$1{,}652004 \times 10^{-12}$	

Il ressort de ces résultats que dans QRT2 et QRT6 les erreurs sur les vecteurs propres sont dûes essentiellement à la transformation de Householder (sur cet exemple).

Des exemples de résultats sur les valeurs propres et les résidus obtenus par les trois méthodes appliquées à A sont donnés dans le tableau 6.

TABLEAU 6

v.p.	Méthodes	v.p. calculées	résidus	
			$\|r_A(U_i)\|$	$\|r_T(\varphi_i)\|$
λ_1	QRT2	639, 629 434 437 187	$3{,}36 \times 10^{-11}$	$1{,}88 \times 10^{-10}$
	QRT6	639, 629 434 437 187	$2{,}32 \times 10^{-12}$	$1{,}72 \times 10^{-12}$
	JACOBI	639, 629 434 437 187	$1{,}14 \times 10^{-12}$	
λ_2	QRT2	-114, 511 176 460 083	$3{,}53 \times 10^{-11}$	$1{,}94 \times 10^{-10}$
	QRT6	-114, 511 176 460 083	$1{,}09 \times 10^{-12}$	$1{,}04 \times 10^{-12}$
	JACOBI	-114, 511 176 460 084	$1{,}93 \times 10^{-13}$	
λ_{30}	QRT2	- 0, 250 687 020 232 980	$3{,}41 \times 10^{-7}$	$2{,}98 \times 10^{-12}$
	QRT6	- 0, 250 687 020 232 979	$3{,}41 \times 10^{-7}$	$3{,}38 \times 10^{-14}$
	JACOBI	- 0, 250 687 020 232 979	$3{,}41 \times 10^{-14}$	
λ_{23}	QRT2	- 0, 299 589 611 198 079	$9{,}40 \times 10^{-8}$	$1{,}12 \times 10^{-12}$
	QRT6	- 0, 299 589 611 198 079	$9{,}40 \times 10^{-8}$	$8{,}36 \times 10^{-15}$
	JACOBI	- 0, 299 589 611 198 050	$1{,}29 \times 10^{-14}$	

3.3.4.

Les mêmes méthodes QRT2, QRT6 et JACOBI ont été essayées sur la matrice d'ordre 44

$$B = 8D_{44} - 5D_{44}^2 + D_{44}^3 \qquad \text{(cf. WILKINSON [15])}$$

où

$$D_{44} = \begin{pmatrix} 2 & 1 & & & 0 \\ 1 & \ddots & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & 1 \\ 0 & & & 1 & 2 \end{pmatrix} \qquad \text{d'ordre 44.}$$

B est alors à sept diagonales et à la forme suivante :

$$B = \begin{bmatrix} 5 & 2 & 1 & 1 & & & 0 \\ 2 & 6 & 3 & & & & \\ 1 & 3 & 6 & 3 & & & \\ 1 & 1 & 3 & \ddots & \ddots & & \\ & 1 & \ddots & \ddots & \ddots & 3 & \\ & & \ddots & \ddots & 3 & 6 & 2 \\ 0 & & & 1 & 1 & 2 & 5 \end{bmatrix}$$

Les valeurs propres de B sont connues :

de $i=1,\ldots,N$ $\qquad \lambda_i = \mu_i(8-5\mu_i+\mu_i^2)$

avec : $\qquad \mu_i = 4 \sin^2 \left(\frac{i\pi}{2N+2}\right) \qquad (N=44)$

- <u>Temps d'exécution</u> :

QRT2	32,92	sec.	(82 itérations)
QRT6	25,79	sec.	(53 itérations)
JACOBI	116,23	sec.	(6726 rotations planes)

QRT6 prend 22 % moins de temps que QRT2.

- Précision :

Les erreurs sont ici encore dûes essentiellement à la transformation de Householder et la substitution arrière. Les commentaires du 3.3.3. s ont encore valables.

Le tableau 7 contient les résultats relatifs aux valeurs propres se trouvant aux extrémités du spectre.

Dans le tableau 8 sont reproduits les résidus obtenus pour les mêmes valeurs propres par les trois méthodes.

TABLEAU 7

	λ_1	λ_2
v.p. exacte	15. 922 215 640 509 7	15. 691 222 719 611 3
v.p. calculée par QRT2	15. 922 215 640 509 8	15. 691 222 719 611 3
v.p. calculée par QRT6	15. 922 215 640 509 7	15. 691 222 719 611 3
v.p. calculé par JACOBI	15. 922 215 640 509 7	15. 691 222 719 611 2

)

λ_3	λ_{42}
15. 314 010 508 437 9	0. 340 171 322 132 473
15. 314 010 508 438 0	0. 340 171 322 132 481
15. 314 010 508 438 0	0. 340 171 322 132 481
15. 314 010 508 437 9	0. 340 171 322 132 483

λ_{43}	λ_{44}
0. 153 824 064 142 014	0. 038 856 634 456 583 9
0. 153 824 064 142 019	0. 038 856 634 456 587 4
0. 153 824 064 142 019	0. 038 856 634 456 587 4
0. 153 824 064 142 026	0. 038 856 634 456 598 0

TABLEAU 8

v.p.	Résidus par QRT2		Résidus par QRT6		Résidus par JACOBI	
λ_1	7.08	10^{-14}	3.08	10^{-14}	2.15	10^{-14}
λ_2	6.30	10^{-14}	2.88	10^{-14}	2.65	10^{-14}
λ_3	7.66	10^{-14}	2.89	10^{-14}	2.67	10^{-14}
λ_{42}	9.25	10^{-15}	9.09	10^{-15}	1.40	10^{-14}
λ_{43}	6.68	10^{-15}	6.63	10^{-15}	1.49	10^{-14}
λ_{44}	4.99	10^{-15}	5.21	10^{-15}	1.72	10^{-14}

4. APPLICATION A LA METHODE DES BISSECTIONS SUCCESSIVES.

De la propriété des suites de Sturm et du fait que

$$\Psi_i(x) = \frac{\det(a_i - x1_i)}{\det(a_{i-1} - x1_{i-1})}$$

nous déduisons la proposition suivante :

PROPOSITION :

Soit x un nombre réel tel que la suite finie $\Psi_1(x), \Psi_2(x), \ldots, \Psi_n(x)$ existe. Alors :

- Le nombre de $\Psi_i(x)$ appartenant à la suite et qui sont positifs ou nuls est égal au nombre de valeurs propres plus grandes ou égales à x ;
- Le nombre de $\Psi_i(x)$ appartenant à la suite et qui sont négatifs est égal au nombre de valeurs propres strictement inférieurs à x .

Habituellement, dans la méthode des bissections successives on utilise au lieu des $\Psi_i(x)$, les $p_i(x)$ polynomes caractéristiques des a_i :

$$p_i(x) = \det(a_i - x1_i)$$

les p_i vérifient :

$$p_o(x) = 1 \quad , \quad p_1(x) = d_1 - x$$

$$p_i(x) = (d_i - x)P_{i-1}(x) - \beta_i^2 P_{i-2}(x) \qquad i=2,\ldots,n$$

L'idée d'employer les $\Psi_i(x)$ au lieu des $p_i(x)$ a été déjà étudiée (cf. Wilkinson [15] p. 249).

Elle présente de nombreux avantages ; en particulier il y a deux fois moins d'opérations et on évite les dépassements (overflow et underflow).

On notera m(x) le nombre de $\Psi_i(x)$ qui sont < 0 dans la suite $\Psi_1(x),\ldots,\Psi_n(x)$.

Dans l'algorithme classique des bissections successives (voir [1]) on détermine un intervalle $[x_o, x_1]$ dans lequel se trouvent toutes les valeurs propres puis, pour calculer la $i^{ème}$ v.p., on réduit cet intervalle en utilisant la propriété ci-dessus des m(x).

Donc :

ALGORITHME 1 :

Soit $x = \dfrac{x_o + x_1}{2}$

si $m(x) < i$ prendre $x_o = x$ sinon

prendre $x_1 = x$,

jusqu'à ce que $[x_o - x_1]$ soit suffisamment petit.

L'algorithme ci-dessus est très coûteux si on veut calculer un grand nombre de valeurs propres. Dans ce cas une première amélioration possible consiste à remarquer qu'au cours du calcul de la $i^{ème}$ valeur propre, les m(x) successifs peuvent nous permettre d'améliorer la localisation des autres valeurs propres à calculer ([2]) (cette remarque est dûe à Givens).

Malgré cette amélioration l'algorithme est peu conseillé pour calculer un grand nombre de valeurs propres ou toutes les valeurs propres de A et il y a intérêt à utiliser la méthode des bissections ci-dessus seulement pour isoler les valeurs propres dans des intervalles disjoints et à terminer avec une méthode de Newton.

ALGORITHME 2 :

1) Trouver un intervalle $[x_o,x_1]$ dans lequel se trouvent toutes les valeurs propres.

   De $i = m_1$ à m_2 faire :

2) Trouver un intervalle $[x_o^i,x_1^i]$ dans lequel se trouve la seule valeur propre λ_i , par la méthode de bissection avec la remarque ci-dessus.
   Dès que $[x_o^i,x_1^i]$ est trouvé aller à 3.

   Si $[x_o-x_1] < \varepsilon_1$ arrêter le processus pour la $i^{ème}$ v.p. (ε_1 donné).

3) En partant de $\lambda_i^{(0)} = \frac{x_o^{(i)}+x_1^{(i)}}{2}$, faire $\lambda_i^{(s+1)} = \lambda_i^{(s)} - \frac{\psi(\lambda_i^{(s)})}{\psi'(\lambda_i^{(s)})}$

   sans sortir de l'intervalle $[x_o^i,x_1^i]$, jusqu'à ce que $|\lambda_i^{(s+1)}-\lambda_i^{(s)}|$ soit suffisamment petit.

Il est a remarquer que la partie 3 de l'algorithme 2 n'est autre que la méthode d'Abramov-Chichov accélérée.

COMPORTEMENT DE L'ALGORITHME 2.

Il est connu que la méthode des bissections successives est très stable, et converge en un nombre k de pas vers un nombre μ_i tel que $|\mu_i-\lambda_i| < 2^{-k}|x_o-x_1|$ où λ_i est la $i^{ème}$ valeur propre cherchée.

Le procédé de Newton par contre peut présenter des instabilités et donner des valeurs propres assez peu précises pour certaines matrices.

REMARQUES D'ORDRE PRATIQUE :

Analyse de l'erreur de méthode.

Lorsque $t = \Psi_{Newt}(t')$ et que $|t-t'| \leq \varepsilon$ on a vu (cf. prop. 10) que :

$$|\Psi(t)| = \frac{\varepsilon^2}{2}\Psi''(\xi) .$$

. Comme $|\Psi'(x)| > 1$ et que $\Psi(t) = \Psi(\lambda)+(t-\lambda)\Psi'(\xi)$

$$\Psi(\lambda) = 0 \Rightarrow |t-\lambda| = \frac{|\Psi(t)|}{|\Psi'(\xi)|} < \Psi(t) .$$

. On a vu que $\Psi''(\xi) = -2\beta_n^2 \tilde{e}_{n-1}^T(a_{n-1}-\xi 1_{n-1})^{-3}\tilde{e}_{n-1}$

On peut donc majorer $|\Psi''(\xi)|$ par $\frac{2\beta_n^2}{D_\xi^3}$

où D_ξ est la distance de ξ au spectre de a_{n-1}

D'où

$$\boxed{|t-\lambda| \leq \varepsilon^2 \frac{\beta_n^2}{D_\xi^3}}$$

Cette inégalité nous a en particulier servi pour écrire un test d'arrêt convenable pour la phase 3, de l'algorithme 2.

(Une bonne estimation de D_ξ peut être trouvée à chaque itération de Newton grâce à la phase 2 de l'algorithme et au fait qu'à la limite on peut prendre $D_\xi \sim D_{t'}$, distance de t' au spectre de a_{n-1}) .

EXPERIENCES NUMERIQUES

Nous avons calculé les dix premières valeurs propres de D_{40} (cf. 3.3.1.) par l'algorithme 1 puis par l'algorithme 2. La précision demandée était de 10^{-15} pour l'algorithme 2 et de 10^{-12} pour l'algorithme 1. Voici les résultats obtenus :

TABLEAU 9 :

Calcul des dix premières valeurs propres de D_{40} par l'algorithme normal

V.P. NO:	NBRE D'ITS	V.P. CALCULEES
1	42	-02 5863976326392 4
2	42	-01 2343915243955 05
3	42	-01 5260915224471 32
4	42	-01 9320721588838 36
5	42	+00 1449950977962 54
6	42	+00 2076688860774 99
7	42	+00 2303607866256 42
8	42	+00 3641412784668 32
9	42	+00 4570216403558 33
10	42	+00 5581568127984 42

001.16 SECONDS IN EXECUTION

TABLEAU 10

Calcul des dix premières valeurs propres par l'algorithme amélioré.

V.P. NO:	NBRE DE BISECTIONS:	NBRE DE PAS DE NEWTON :
1	10	6
2	3	10
3	3	7
4	3	9
5	3	11
6	2	5
7	4	6
8	5	3
9	2	4
10	3	5

```
 * V.P. CALCULEES*
-02 5868397632519O2
-01 2343915243930 27
-01 5260915224444 13
-01 9320721589013 89
+00 1449509779581 1
+00 2076688860778 39
+00 2808607860255 98
+00 3641412784665 65
+00 4570216403561 14
+00 5589568127984 26
```

000.48 SECONDS IN EXECUTION

TABLEAU 11

Valeurs propres exactes

```
***VP EXACTES  ***
-02 5868397632519O7
-01 2343915243930 29
-01 5260915224444 19
-01 9320721589013 89
+00 1449509779581 1
+00 2076688860778 89
+00 2808607860255 98
+00 3641412784665 65
+00 4570216403561 14
+00 5589568127984 26
```

5. PROCEDURES ALGOL

```
PROCEDURE QRTRIDIAG(INTEGER VALUE M1,N;LONG REAL VALUE EPS ;
   INTEGER RESULT TOTITER ;
   LONG REAL ARRAY E,D(*);
          LONG REAL ARRAY Z(*,*);LOGICAL EINVEC,FAIL) ;
     BEGIN
         COMMENT  CETTE PROCEDURE CALCULE LES  VALEURS PROPRES ET LES
                  VECTEURS PROPRES D'UNE MATRICE TRIDIAGONALE SYMETRI-
                  QUE DONT LA DIAGONALE EST D ET LA SUB-DIAGONALE EST E
                  SI LE PARAMETRE LOGIQUE EINVEC PREND LA VALEUR
                  "VRAI",ALORS LES VECTEURS PROPRES SONT CALCULES AVEC
                  LES VALEURS PROPRES SIMULTANEMENT. LES VECTEURS
                  PROPRES RANGES DANS LE MEME ORDRE QUE LES VALEURS
                  PROPRES , SONT DANS LE TABLEAU Z(1::N,1::N) .
                  LES VALEURS PROPRES SONT RANGEES DANS LA DIAGONALE D
                  QUI EST DONC EFFACEE.LE TABLEAU Z N'EST PAS TOUCHE SI
                  EINVEC=FALSE,ET PEUT DONC ETRE N'IMPORTE QUEL TABLEAU
                  A DEUX DIMENSIONS SI ON CHERCHE SEULEMENT LES VALEURS
                  PROPRES  .       REMARQUE:
                  QRTRIDIAG EST ECRIT POUR POUVOIR ETRE UTILISE
                  INDEPENDAMMENT DES PROCEDURES SUIVANTES  ET
                  CERTAINES INSTRUCTIONS SONT INUTILES SI QRSHIFT
                  EST UTILISE ;

        LONG REAL PROCEDURE MINRAC(LONG REAL VALUE A,B,C) ;
           BEGIN LONG REAL T,S;
                 T:=(A-C)/(2L*B) ;
                 S:=LONGSQRT(1+T*T) ;
                 A:=(IF T<0 THEN T-S ELSE T+S) ;
                      C-B/A
           END MINRAC ;
      LONG REAL PROCEDURE PSINEWTON(INTEGER VALUE M,K;
                   LONG REAL H ;LONG REAL ARRAY E,D(*)) ;
        BEGIN LONG REAL PSI,T,S;
              S:=0L ;PSI:=1L ;
              FOR I:=M UNTIL K DO BEGIN
              T:=(E(I)/PSI)**2 ;
              S:=(1L+S)*T ;
              PSI:=D(I)-H-T*PSI ;
              IF PSI=0L THEN PSI:=1'-20L ;
            END ;
          H+PSI/(1L+S)
        END PSINEWTON ;
        INTEGER ITER,ITR,K1,K2;
        LONG REAL SOM,TRANS,R,S,C,H,EP,ES,DP,MU,MUP,EPS2;
        IF EPS=0L THEN EPS:=1'-16L ;
        SOM:=0L ;    TOTITER:=0;
        EPS:=0.5L*EPS/(N-M1+1) ;
        EPS2:=EPS ;
        IF ¬EINVEC THEN EPS2:=LONGSQRT(EPS2) ;
             IF EINVEC THEN FOR I:=M1 UNTIL N DO
             FOR J:=M1 UNTIL I DO
             Z(I,J):=Z(J,I):=(IF I=J THEN 1L ELSE 0L) ;
```

```
     FOR L:=N STEP-1 UNTIL M1 DO
    BEGIN ITER:=0 ;     MUP:=0L ;
          IF L>10 THEN BEGIN K1:=L DIV 10 ;
          K2:=3*K1   END
               ELSE K1:=K2:=L-1;
          COMMENT     CALCUL DE LA TRANSLATION D'ORIGINE
               KPRIM =L DIV 3  ,ET K=3*KPRIM ;
TRANSLAT:      IF L=M1 THEN GOTO FIN1;
           IF E(L)*E(L)<ABS((SOM+D(L))*(D(L-1)-D(L)))*EPS
             THEN GOTO FIN1 ;
            MU:=MINRAC(D(L-1),E(L),D(L)) ;
            IF (ABS(MUP-MU)>=0.15L*ABS(SOM+MU))AND(TOTITER<10)
                      THEN GOTO TRANSMUT ;
            C:=PSINEWTON(L-K1,L,MU,E,D) ;
            TRANS:=PSINEWTON(L-K2,L,C,E,D) ;
            ITR:=0 ;
             WHILE (ABS(TRANS-C)>EPS2)AND(ITR<3) DO
           BEGIN
               C:=TRANS ;
               ITR:=ITR+1 ;
               TRANS:=PSINEWTON(1,L,C,E,D) ;
           END ;
           FOR I:=M1 UNTIL L DO D(I):=D(I)-TRANS ;
           SOM:=SOM+TRANS ;
         COMMENT   TRANSMUTATION QR ;
TRANSMUT:           DP:=D(1) ;          C:=1L ;
                    MUP:=MU ;           S:=0L ;
          FOR I:=M1+1 UNTIL L DO
           BEGIN
          EP:=C*E(I) ;   H:=DP*C ;
            IF ABS(DP)>ABS(E(I)) THEN
          BEGIN
            C:=E(I)/DP ;       R:=LONGSQRT(1L+C*C) ;
            E(I-1):=S*DP*R ;         S:=C/R ;       C:=1L/R ;
              END
              ELSE
             BEGIN
               C:=DP/E(I) ;
               R:=LONGSQRT(1L+C*C) ;
             E(I-1):=S*E(I)*R ;       S:=1L/R ;      C:=C/R ;
               END;
          ES:=C*EP+S*D(I) ;
          DP:=C*D(I)-S*EP ;
          D(I-1):=H+S*ES ;
          IF EINVEC THEN FOR K:=M1 UNTIL N DO
          BEGIN     H:=Z(K,I-1) ;
                    Z(K,I-1):=C*H+S*Z(K,I) ;
                    Z(K,I):=C*Z(K,I)-S*H ;
          END ;
        END I;
          D(L):=C*DP ;      E(L):=S*DP ;
          ITER:=ITER+1 ;   TOTITER:=TOTITER+1 ;
          IF ITER=25 THEN GOTO DIVERG ELSE GOTO TRANSLAT ;
FIN1:       D(L):=D(L)+SOM ;
           END BOUCLEL;
                    GOTO FIN2 ;
     DIVERG:             FAIL:=TRUE ;
       FIN2:     END QPTRIDIAG;
```

```
PROCEDURE HOUSTRIDIAG(INTEGER VALUE N;LONG REAL ARRAY A(*,*);
        LONG REAL ARRAY D,E(*)) ;

         BEGIN
     COMMENT  :        CETTE PROCEDURE TRIDIAGONALISE PAR LA
          METHODE DE HOUSEHOLDER LA MATRICE A(1::N,1::N) . LA DIACONALE
          EST RANGEE EN D(1::N) ET LES ELEMENTS SUB-DIACONAUX SONT RANGES
          DANS LES N-1 DERNIERES MEMOIRES DE E(1::N) (DE PLUS E(1)=0) .
          LE TRIANGLE STRICTEMENT SUPERIEUR DE A EST CONSERVE ET LE
          TRIANGLE INFERIEUR SERT AU RETOUR A LA BASE INITIALE PAR LA
          PROCEDURE VECTSUBT .LA PRESENTE PROCEDURE EST INSPIREE
          DE WILKINSON(1) P. 219 ;

          LONG REAL F,G,H;  INTEGER L ;
    FOR I:=N STEP-1 UNTIL 3 DO
    BEGIN L:=I-1 ;       H:=0L ;
          FOR K:=1 UNTIL L DO H:=H+A(I,K)*A(I,K) ;
           IF H<1'-15L THEN
              BEGIN      E(I):=0L ;
                         D(I):=A(I,I) ;
                         GOTO SAUT ;
              END ;
                          F:=A(I,L) ;
          E(I):=G:=IF F>0 THEN -LONGSQRT(H) ELSE LONGSQRT(H) ;
            D(I):=A(I,I) ;   A(I,I):=H:=H-F*G ;
            A(I,L):=F-G ;        F:=0L ;
    FOR J:=1 UNTIL L DO
    BEGIN      G:=0L ;
          FOR K:=1 UNTIL J DO G:=G+A(J,K)*A(I,K) ;
          FOR K:=J+1 UNTIL L DO G:=G+A(K,J)*A(I,K) ;
          G:=E(J):=G/H ; F:=F+G*A(I,J) ;
    END J;
          H:=F/(H+H) ;
  FOR J:=1 UNTIL L DO
     BEGIN     F:=A(I,J) ;     G:=E(J):=E(J)-H*F ;
               FOR K:=1 UNTIL J DO
               A(J,K):=A(J,K)-F*E(K)-G*A(I,K)
          END J;
SAUT :
   END I;
                E(1):=0L ;      D(1):=A(1,1) ;
                D(2):=A(2,2) ;        E(2):=A(2,1) ;
END HOUSTRIDIAG;
```

---

(1) Référence [15]

```
PROCEDURE VECTSUBST(INTEGER N; LONG REAL ARRAY A,Z(*,*)) ;
  BEGIN

     COMMENT        PROCEDURE POUR LE RETOUR ARRIERE .
        UTILISE LES INFORMATIONS QUI SONT CONSERVEES
        DANS LE TRIANGLE INFERIEUR DE A AU COURS DE
        LA TRANSFOMATION DE HOUSEHOLDER POUR ECRIRE
        LES VECTEURS PROPRES DANS LA BASE INITIALE  ;

         LONG REAL P ;
     FOR I:=1 UNTIL N DO
     BEGIN
         FOR K:=1 UNTIL N DO
          BEGIN
               P:=0L ;
               FOR J:=1 UNTIL I-1 DO
               P:=P+A(I,J)*Z(J,K) ;
               P:=P/A(I,I) ;
               FOR J:=1 UNTIL I-1 DO Z(J,K):=Z(J,K)-A(I,J)*P ;
          END
      END
  END VECTSUBST ;
```

```
PROCEDURE QRSHIFT(INTEGER VALUE N;INTEGER RESULT TOTITER ;
        LOGICAL VALUE    EINVEC ;    LOGICAL RESULT IMPOSSIBLE ;
        LONG REAL VALUE EPS ;  LONG REAL ARRAY A,VECT(*,*) ;
        LONG REAL ARRAY VP(*)) ;
   BEGIN
 COMMENT :  LES PROCEDURES QRTRIDIAG ,HOUSTRIDIAG ET VECTSUBST ETANT
        PREDECLAREES,LA PROCEDURE SUIVANTE PERMET DE CALCULER TOUTES
        LES VALEURS PROPRES ET TOUS LES VECTEURS PROPRES D'UNE MATRICE
        SYMETRIQUE PLEINE A(1::N,1::N) PAR L'ALGORITHME QR.LES VALEURS
        PROPRES SONT RANGEES DANS LE TABLEAU VP(1::N).LES VECTEURS
        PROPRES NE SONT CALCULES QUE SI EINVEC=TRUE ET SONT ALORS
        RANGES DANS VECT(1::N,1::N).DANS LE CAS CONTRAIRE VECT N'EST
        PAS UTILISE ET PEUT DONC ETRE N'IMPORTE QUEL TABLEAU A DEUX
        DIMENSIONS (PAR EXEMPLE A).LES  VP(I)  NE SONT PAS TOUJOURS
        RANGEES DANS L'ORDRE CROISSANT. SI LE NOMBRE DE TRANSMUTATIONS
        NECESSITEES POUR LE CALCUL D'UNE VALEUR PROPRE VP(L) ,
        DEPASSE 25, IMPOSSIBLE PREND LA VALEUR "VRAI".TOTITER EST LE
        NOMBRE TOTAL D'ITERATIONS NECESSITEES ;

        INTEGER N1,Q ;
        LONG  REAL ARRAY E(1::N) ;
        IF EINVEC THEN
             FOR I:=1 UNTIL N DO
             FOR J:=1 UNTIL I DO
                  Z(I,J):=Z(J,I):=(IF I=J THEN 1L ELSE 0L) ;
             HOUSTRIDIAG(N,A,VP,E) ;
             IMPOSSIBLE:=FALSE ;
             N1:=N+1 ;      Q:=N ;
   DEL:   IF N1>1 THEN GOTO SORTIE ;
        FOR J:=N1-1 STEP-1 UNTIL 1 DO
        IF E(J)=0 THEN
        BEGIN       Q:=N1-1 ;
                    N1:=J ;   GOTO QR  ;
        END ;
    QR :     QRTRIDIAG(N1,Q,EPS,TOTITER,E,VP,VECT,EINVEC,IMPOSSIBLE) ;
             IF IMPOSSIBLE THEN GOTO FIN
                  ELSE GOTO DEL ;
 SORTIE:    IF EINVEC THEN VECTSUBST(N,A,VECT) ;
   FIN :
        END QRSHIFT;
```

```
PROCEDURE BISECT(INTEGER VALUE N,M1,M2;LONG REAL VALUE EPS ;
     LOGICAL IMPOSS ;
     LONG REAL ARRAY E,B,D,VP(*)) ;
BEGIN
     COMMENT      CETTE PROCEDURE CALCULE LES VALEURS PROPRES
     D'UNE MATRICE TRIDIAGONALE SYMETRIQUE PAR LA METHODE DES
     BISSECTIONS COMPLETEE PAR UNE METHODE DE NEWTON (METHODE
     D'ABRAMOV ACCELEREE) .LES VALEURS PROPRES ETANT ORDONNEES
     EN CROISSANT ,L'ALGORITHME CALCULE LES VALEURS PROPRES DE
     VP(M1) A VP(M2) , OU M1 ET M2 SONT DEUX ENTIERS ET M1<M2 .
     LORSQUE  LA PHASE DES BISSECTIONS DEPASSE 80 ITERATIONS ,
     IMPOSS PREND LA VALEUR 'TRUE' ET LEPROCEDE EST ARRETE
     DE MEME SI LA PHASE DE NEWTON DEPASSE 40 ITERATIONS    ;

LONG REAL PROCEDURE PSINEWTON(INTEGER VALUE K ;
          LONG REAL VALUE H ; LONG REAL ARRAY B,D(*)) ;
     BEGIN     LONG REAL PSI,T,S ;
          S:=0L ;   PSI:=D(K)-H ;
     FOR I:=N-1 STEP -1 UNTIL 1 DO
       BEGIN
          T:=B(I+1)/(IF PSI=0 THEN 1'-20L ELSE PSI) ;
          S:=(1L+S)*T/PSI ;
          PSI:=D(I)-H-T ;
          IF PSI=0L THEN PSI:=1'-30L ;
      END ;          H+PSI/(1L+S)
END ;
     LONG REAL X,T,PSI,X1,X0,H,H1,MIN ;
     INTEGER SAVE0,SAVE1,M,ITER1,ITER2,SAVE ;
     IMPOSS:=FALSE ;
     X0:=D(N)-ABS(E(N)) ;
     X1:=D(N)+ABS(E(N)) ;
     FOR I:=1 UNTIL N-1 DO
     BEGIN
     H:=ABS(E(I))+ABS(E(I+1)) ;
     IF D(I)-H<X0 THEN X0:=D(I)-H ;
     IF D(I)+H>X1 THEN X1:=D(I)+H ;
     END ;
     MIN:=X0 ;
     EPS:=EPS/B(N) ;
     FOR I:=M1 UNTIL M2 DO VP(I):=X1 ;
     FOR I:=M1 UNTIL M2 DO
     BEGIN
     ITER1:=ITER2:=0 ;
     X1:=VP(I) ;      X0:=MIN ;
     SAVE0:=0 ;     SAVE1:=N ;
     IF (I>M1) AND (MIN<VP(I-1)) THEN
                         X0:=VP(I-1) ;
INTERVALLE :
          X:=(X0+X1)*0.5L  ;
          ITER1:=ITER1+1 ;
          IF ITER1>80 THEN
                  BEGIN          IMPOSS:=TRUE ;
                                 GOTO FINBOUCLE ;
                  END ;
      M:=0 ;          T:=1L ;
     FOR I:=1 UNTIL N DO
     BEGIN T:=D(I)-X-B(I)/T ;
          IF T=0 THEN T:=1'-30L ;
          IF T<0 THEN  M:=M+1 ;
          IF I=N-1 THEN SAVE:=M ;

      END ;
```

```
            IF M<I THEN
                    BEGIN    X0:=X ;
                             SAVE0:=SAVE ;
                             IF X>MIN THEN MIN:=X ;
                    END
                ELSE
                    BEGIN     X1:=X ;
                              SAVE1:=SAVE ;
                              IF M=I THEN
                               IF MIN<X THEN MIN:=X ;
                             FOR J:=I+1 UNTIL M DO
                             IF J<=M2 THEN
                                 IF VP(J)>X THEN VP(J):=X ;
                    END ;
          IF SAVE0<> SAVE1
            THEN GOTO NEWTON ;
         IF ABS(X0-X1)<EPS THEN
            BEGIN VP(I):=X ;
                   GOTO FINBOUCLE
              END
            ELSE GOTO INTERVALLE ;
NEWTON:    H1:=X0 ;
           H:=(X0+X1)*0.5L ;
           T:=1'-4L ;
       WHILE ((H1-H)/T)**2>T*EPS DO
        BEGIN
         H1:=H ;
           IF ITER2>40 THEN
             BEGIN    IMPOSS:=TRUE ;
                       GOTO FINBOUCLE ;
           END ;
          H:=PSINEWTON(N,H,B,D) ;
          ITER2:=ITER2+1 ;
          IF (H<X0)OR(H>X1) THEN GOTO INTERVALLE ;
          T:=(IF H-X0<X1-H THEN H-X0 ELSE X1-H) ;
          IF T < 1'-4L THEN T:=1'-4L ;
        END ;
          VP(I):=H ;
  FINBOUCLE:          END ;

       FIN1:            END BISECT ;
```

CHAPITRE III

APPLICATION AU CALCUL DE VALEURS PROPRES D'OPERATEURS COMPACTS AUTOADJOINT

INTRODUCTION :

Soit T un opérateur compact autoadjoint sur un espace de Hilbert séparable H. Supposons qu'on approche les valeurs propres de T par la méthode de Galerkin :

si E_n est une suite croissante de s.e.v. de H, de dimension finie n, cela revient à approcher les valeurs propres λ_i de T par celles $\lambda_i^{(n)}$ de la partie de T dans E_n .

Appelons F_n le supplémentaire orthogonal de E_n et notons :

$$T = \left(\begin{array}{c|c} T_{E_n} & U_n \\ \hline U_n^* & T_{F_n} \end{array} \right)$$

Nous allons utiliser les résultats de la première partie pour corriger les approximations $\lambda_i^{(n)}$.

La question que nous nous posons est la suivante : comment utiliser certaines informations contenues dans U_n et T_{F_n} pour améliorer l'approximation $\lambda_i^{(n)}$ de λ_i ? On pourrait bien entendu augmenter la dimension du problème approché, mais cela conduit à des résolutions de problèmes de valeurs propres de dimensions plus grandes et ce n'est donc pas très intéressant comme on peut le voir en faisant le décompte des opérations nécessaires.

Supposons par exemple que l'on veuille approcher les deux plus grandes valeurs propres de T et que l'on soit obligé pour avoir une précision convenable, de choisir la dimension n = 100.

On doit tridiagonaliser la matrice d'ordre 100 par la méthode de Householder ($\sim \frac{2}{3}$ n^3 multiplications) et appliquer QR à la matrice tridiagonale obtenue (nombre d'opérations négligeable par rapport à celui de la tridiagonalisation car n est grand).

On peut penser qu'il est sans doute plus intéressant de faire plusieurs fois un calcul de valeurs propres sur une matrice de dimension n' beaucoup plus petit (nombre de multiplications $\sim \frac{2}{3} n'^3$) que de procéder à la tridiagonalisation dela matrice d'ordre n.

Il se révèlera très intéressant lorsqu'on veut calculer quelques valeurs propres de T seulement, de corriger les valeurs propres obtenues par une itération du procédé d'Abramov appliqué à la matrice d'ordre n.

Nous commençons au paragraphe 1 par étudier le comportement asymptotique de la précision $\frac{\lambda_i - \lambda_i^{(n)}}{\lambda_i}$; cette étude sera utile pour établir des résultats sur la précision asymptotique des valeurs propres corrigées.

Au paragraphe 2, nous étudions le procédé de corrections et nous décrivons des expériences numériques au paragraphe 3.

1. NOTATIONS ET RESULTATS IMMEDIATS.

1.1.

Soit T un opérateur compact autoadjoint dans H.

Nous adoptetons les mêmes notations que celles de la première partie : si E est un sous-espace vectoriel de dimension finie n, et F son supplémentaire orthogonal, on notera :

$$T = \left(\begin{array}{c|c} T_E & U \\ \hline U^* & T_F \end{array}\right)$$

Nous ne parlerons que des valeurs propres positives de T que nous ordonnons en décroissant.

Pour les valeurs propres négatives, un travail similaire peut être fait mais on peut également se ramener au cas étudié en remarquant que les valeurs propres négatives de T sont les valeurs propres positives de -T.

On note pour $x \in \mathbb{R} \setminus \sigma(T_F)$:

$$S(x) = T_E - x1_E - U(T_F - x1_F)^{-1}U^*$$

Les n valeurs propres de S(x) ordonnées en décroissant sont notées $\Lambda_1(x), \Lambda_2(x), \ldots, \Lambda_n(x)$.

D'après la proposition 11 de la première partie, si $\Lambda_j(\lambda) = 0$, alors λ est une valeur propre de T.

Sous quelle condition a-t-on $\lambda = \lambda_j$?

Nous pouvons répondre grâce au résultat suivant qui est une généralisation du lemme 7 (chapitre I).

LEMME 1 :

1) Soit λ_i la $i^{ème}$ valeur propre de T et supposons que l'opérateur $\lambda_i 1_F - T_F$ soit défini positif. Alors :
pour $j \leq i$ on a :

$$\wedge_j(\lambda_j) = 0$$

2) Réciproquement, soit λ une valeur propre de T telle que $\lambda \geq \lambda_i$ et soit j un indice pour lequel $\wedge_j(\lambda) = 0$. Alors λ est la $j^{ème}$ valeur propre de T.

En traçant les courbes $\wedge_j(x)$ en fonction de x, pour j=1,2,...,i, le lemme 1 est illustré par la figure suivante :

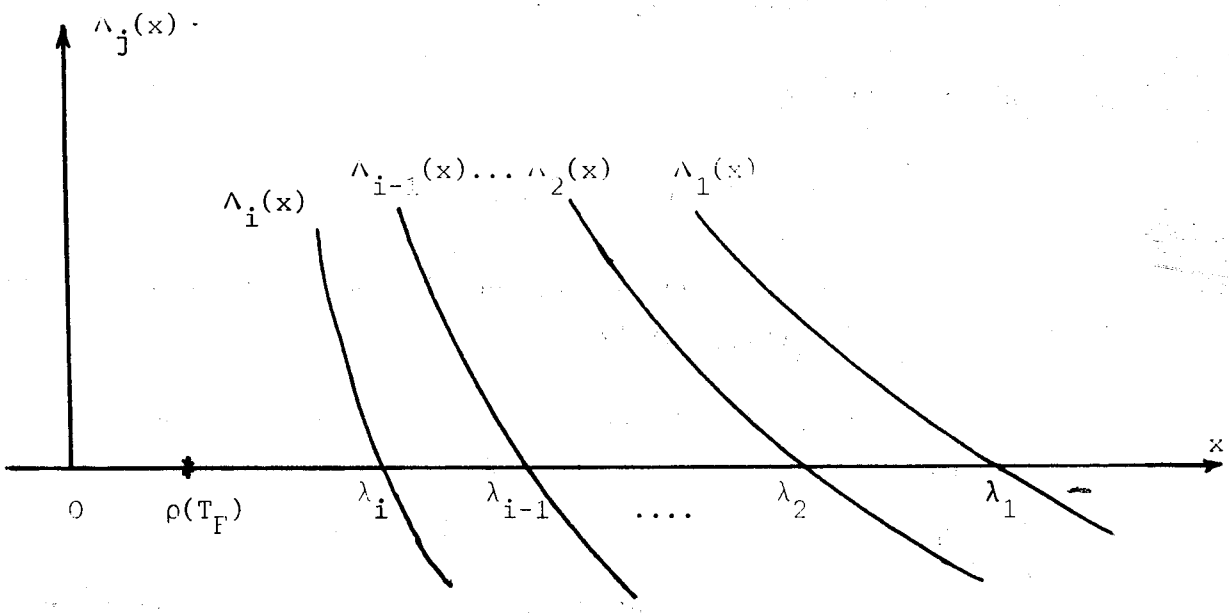


Figure 1. $\lambda_i > \rho(T_F)$

Les fonctions $\Lambda_i(x)$ sont décroissantes strictement :
en effet si $\varphi_i(x)$ est le vecteur propre normé associé à $\Lambda_i(x)$, donné par la proposition 11 de la première partie ,

$$\Lambda_i'(x) = - 1 - \left\| (T_F - x1_F)^{-1} U^* \varphi_i(x) \right\|^2$$

Les i fonctions $\Lambda_j(x)$, $j=1,2,\ldots,i$ admettent un zéro et un seul : λ_j .

D'après la figure 1, lorsque $\lambda_i 1_F - T_F$ est défini positif, il est clair que si $\lambda_i \leq x \leq \lambda_{i-1}$ alors :

$$\Lambda_{i-1}(x) \geq 0 \qquad \text{et} \qquad \Lambda_i(x) \leq 0$$

Soit x tel que $\lambda_i \leq x \leq \lambda_{i-1}$.

Si $\lambda_i 1_F - T_F$ défini positif alors :

$$\Lambda_n(x) \leq \Lambda_{n-1}(x) \leq \ldots \leq \Lambda_i(x) \leq 0 \leq \Lambda_{i-1}(x) \leq \ldots \leq \Lambda_1(x)$$

i.e. S(x) possède i-1 valeurs propres positives et
n-i+1 valeurs propres négatives.

1.2. CONVERGENCE ASYMPTOTIQUE DES VALEURS PROPRES APPROCHEES PAR LA METHODE DE GALERKIN.

Nous allons rappeler un résultat dont nous auront besoin au paragraphe 2.2. concernant l'ordre de convergence des valeurs propres.
On note Π_n la projection orthogonale sur E que nous supposons de dimension n.
On note λ_i et $\lambda_i^{(n)}$ les valeurs propres de T et T_E respectivement et on suppose qu'elles ne sont pas valeurs propres de T_F .
φ_i et $\varphi_i^{(n)}$ sont des vecteurs propres normalisés de T et T_E associés à λ_i et $\lambda_i^{(n)}$ respectivement.

Il est possible de démontrer le résultat suivant concernant la convergence des valeurs propres $\lambda_i^{(n)}$: [cf. 4, 5].

PROPOSITION 2 :

Soit λ_i une valeur propre de T, de multiplicité m, approchée par $\lambda_j^{(n)}$, $j=j_1,j_2,\ldots,j_m$.

Soit P_i la projection spectrale associée à λ_i et posons

$$\Delta = \max_{\substack{\varphi\in P_iH \\ \|\varphi\|=1}} \|(1-\Pi_n)\varphi\| \quad . \quad \text{Alors :}$$

$$\frac{\lambda_i-\lambda_j^{(n)}}{\lambda_i} \leq \rho_i^{(n)}\ \Delta^2 \qquad j=j_1,\ldots,j_m$$

où $\rho_i^{(n)}$ est un coefficient qui tend vers 1 lorsque n tend vers l'infini.

La démonstration est dans [4].

2. CORRECTIONS DES VALEURS PROPRES OBTENUES PAR LA METHODE DE GALERKIN.

Il est possible pour les matrices hermitiennes partitionnées ainsi $A = \left(\begin{array}{c|c} a & b \\ \hline b^H & d \end{array}\right)$, de corriger lorsque $\|b\|$ est assez petit, une valeur propre $\lambda_i^{(a)}$ de a qui est une approximation de la valeur propre λ_i de A : Il suffit de faire une itération de l'algorithme d'Abramov-Chichov c'est-à-dire de prendre comme valeur corrigée $\lambda_i^{(a)}$ + la $i^{\text{ème}}$ valeur propre de $a(\lambda_i^{(a)})$.

On peut étudier pour les opérateurs compacts autoadjoints le même procédé de correction sans perdre de vue que dans ce cas-là, le calcul de S(x) (qui correspond à a(x)), est nécessairement approché. Il faudre donc étudier également les corrections intermédiaires correspondant à l'approximation par troncature.

2.1. CORRECTION PAR UNE ITERATION DE LA METHODE D'ABRAMOV.

Sans nous soucier ici du calcul pratique, nous étudions le procédé qui consiste à approcher λ_i non plus par $\lambda_i^{(n)}$, obtenue par l'approximation de Galerkin, mais par la valeur μ_i suivante :

$$\mu_i = \lambda_i^{(n)} + \Lambda_i(\lambda_i^{(n)}) \qquad (5)$$

On dira que $\Lambda_i(\lambda_i^{(n)})$ est la correction (théorique) de $\lambda_i^{(n)}$ et que μ_i est la valeur corrigée de $\lambda_i^{(n)}$.

REMARQUE :

μ_i dépend de n : il aurait été plus cohérent de noter $\mu_i^{(n)}$ au lieu de μ_i. Cependant cela alourdit inutilement l'écriture car d'autres indices interviennent au paragraphe 2.2.

μ_i est obtenu par une itération de l'algorithme d'Abramov-Chichov.

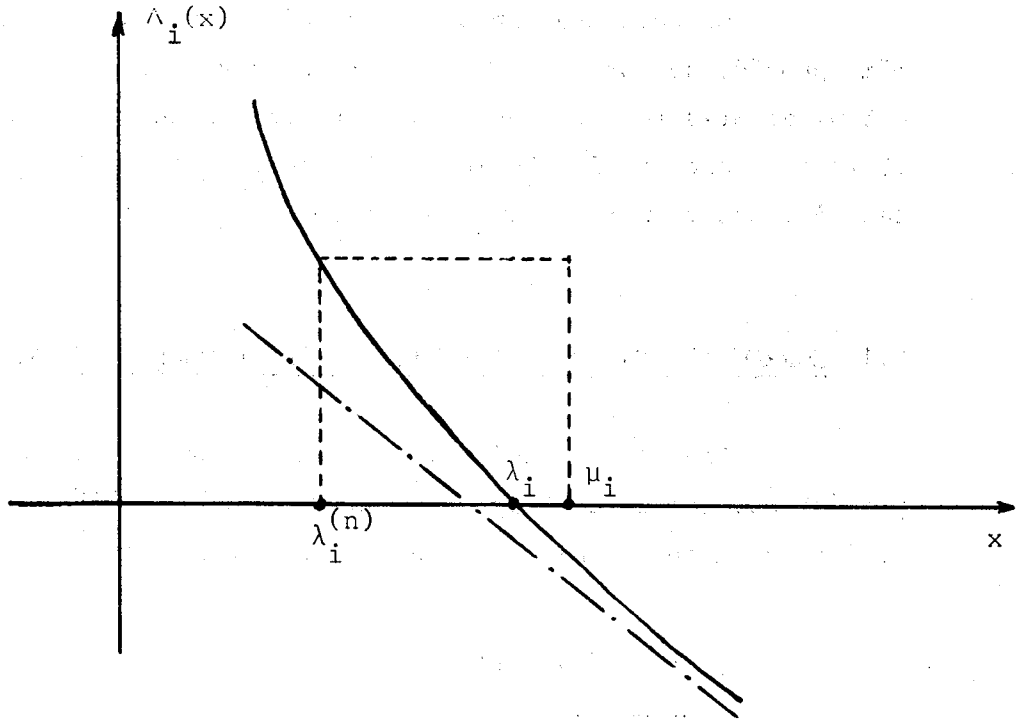


Figure 2.

C'est donc aussi la valeur obtenue en une itération de la méthode des approximations successives appliquée à la résolution de $\Lambda_i(x) + x = x$ à partir du point $\lambda_i^{(n)}$ [cf. figure 2].

PROPRIETE 3 :

Si $\lambda_i^{(n)} 1_F - T_F$ est défini positif alors :

$$\mu_i \geq \lambda_i$$

Cette propriété se démontre à l'aide du lemme suivant :

LEMME 4 :

Dans l'intervalle $\{x ; x \geq \rho(T_F)\}$, les fonctions $x \to \Lambda_i(x) + x$ sont décroissantes et $\mathcal{C}^\infty$.

La démonstration est une conséquence directe de la proposition 11 de la première partie.

DEMONSTRATION DE LA PROPRIETE 3.

Lorsque les valeurs propres positives de T sont ordonnées en décroissant, on a, d'après les résultats classiques sur la méthode de Galerkin ;

$$\lambda_i^{(n)} \leq \lambda_i \ .$$

$\Lambda_i(x)+x$ étant décroissante, on a :

$$\Lambda_i(\lambda_i^{(n)}) + \lambda_i^{(n)} \geq \Lambda_i(\lambda_i) + \lambda_i \ .$$

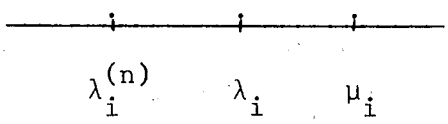
D'après le lemme 1, $\Lambda_i(\lambda_i) = 0$.

D'où le résultat :

$$\mu_i = \Lambda_i(\lambda_i^{(n)}) + \lambda_i^{(n)} \geq \lambda_i \ .$$

C.Q.F.D.

Nous savons maintenant que $\mu_i \geq \lambda_i$:



Il nous faut estimer $\frac{\mu_i - \lambda_i}{\lambda_i}$ pour avoir une idée de la précision de ce procédé.

PROPOSITION 5 :

Soit φ_i un vecteur propre de T associé à la valeur propre λ_i et soit $\hat{\varphi}_{i,n}$ un vecteur défini par :

$\Pi_n \hat{\varphi}_{i,n}$ = un vecteur propre de $S(\lambda_i^{(n)}) + \lambda_i^{(n)} 1_{E_n}$ associé à la valeur propre μ_i

$$(1-\Pi_n)\hat{\varphi}_{i,n} = -(T_F - \lambda_i^{(n)} 1_F)^{-1} U^* \Pi_n \hat{\varphi}_{i,n}$$

Alors si $(\varphi_i, \hat{\varphi}_{i,n}) \neq 0$:

$$\frac{\mu_i - \lambda_i}{\lambda_i} = \frac{\lambda_i - \lambda_i^{(n)}}{\lambda_i} \cdot \frac{((1-\Pi_n)\varphi_i, (1-\Pi_n)\hat{\varphi}_{i,n})}{(\varphi_i, \hat{\varphi}_{i,n})}$$

La démonstration est analogue à celle de la relation (1) du paragraphe 2.1. chapitre I ; proposée par Chichov [6].

On peut alors en déduire le résultat suivant :

COROLLAIRE 6 :

Il existe une constante K indépendante de n telle que :

$$\frac{\mu_i - \lambda_i}{\lambda_i} \leq K \Delta^4$$

DEMONSTRATION :

Faisons la démonstration dans le cas où μ_i est une valeur propre simple.

On a :

$$\frac{\mu_i-\lambda_i}{\lambda_i} = \frac{\lambda_i-\lambda_i^{(n)}}{\lambda_i} \cdot \frac{((1-\Pi_n)\varphi_i,(1-\Pi_n)\hat{\varphi}_{i,n})}{(\varphi_i,\hat{\varphi}_{i,n})}$$

* $\dfrac{\lambda_i-\lambda_i^{(n)}}{\lambda_i} \leq \rho_{i,n}\,\Delta^2$ où $\rho_{i,n} \to 1$ (proposition 2)

* $(\varphi_i,\hat{\varphi}_{i,n}) \to 1$ car $\hat{\varphi}_{i,n} \to \varphi_i$.

En effet $\Pi_n\hat{\varphi}_{i,n}$ = un vecteur propre de $T_E-U(T_F-\lambda_i^{(n)})^{-1}U^*$ associé à μ_i .

Or $\|U(T_F-\lambda_i^{(n)})^{-1}U^*\| \to 0$ et donc

$\Pi_n\hat{\varphi}_{i,n}$ et $\varphi_i^{(n)}$

ont la même limite φ_i qui est de norme égale à 1.

* Il reste à montrer que $((1-\Pi_n)\varphi_i,(1-\Pi_n)\hat{\varphi}_{i,n})\ /\Delta^2 \leq K$

On peut écrire que :

$$((1-\Pi_n)\varphi_i,(1-\Pi_n)\varphi_{i,n}) = \|(1-\pi_n)\varphi_i\|^2 + ((1-\Pi_n)\varphi_i,(1-\Pi_n)(\varphi_i-\hat{\varphi}_{i,n}))$$

$$(1) \qquad \frac{((1-\Pi_n)\varphi_i,(1-\Pi_n)\hat{\varphi}_{i,n})}{\Delta^2} \leq 1 + \left(\frac{(1-\Pi_n)\varphi_i}{\|(1-\Pi_n)\varphi_i\|}, \frac{(1-\Pi_n)(\varphi_i-\hat{\varphi}_{i,n})}{\|(1-\Pi_n)\varphi_i\|}\right)$$

Il est alors facile de voir que comme $\hat{\varphi}_{i,n} \to \varphi_i$,

$$\frac{\|(1-\Pi_n)(\varphi_i-\hat{\varphi}_{i,n})\|}{\|(1-\Pi_n)\varphi_i\|}$$

est, à la limite, borné par 2 par exemple, ce qui montre le résultat.

C.Q.F.D.

REMARQUE :

Il est assez facile de voir que d'après (1) on a en réalité :

$$\frac{\mu_i-\lambda_i}{\lambda_i} \leq \rho'_{i,n}\,\Delta^4 \quad \text{où} \quad \rho'_{i,n} \to 1 \quad \text{lorsque} \quad n \to \infty .$$

2.2. CORRECTIONS INTERMEDIAIRES.

En pratique la méthode de corrections étudiée en 2.1 est impossi à réaliser car elle fait intervenir l'opérateur

$$U\,(T_F - \lambda_i^{(n)} 1_F)^{-1} U^*$$

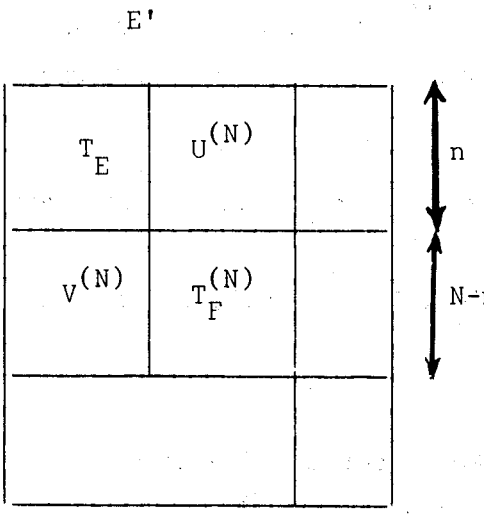
que nous ne savons pas calculer.

Une idée naturelle consiste à remplacer T par la partie de T dans un sous-espace E', où $E' \supset E$ et dim E' = N , N > n . (Cela revient à tronquer la matrice infinie qui représente T dans une base, à une dimension N).

2.2.1. NOTATIONS :

Dans ce qui suit nous reprenons les mêmes notations que précédemment : n , la dimension de E est supposée constante.

Notons T =



avec $V^{(N)} = (U^{(N)})^*$

$T^{(N)}$ désigne la partie de T dans E' représenté matriciellement par

$$T^{(N)} = \left(\begin{array}{c|c} T_E & U^{(N)} \\ \hline V^{(N)} & T_F^{(N)} \end{array} \right)$$

Si on appelle $F_{N,n}$ l'espace supplémentaire de E dans E' les opérateurs $U^{(N)}$, $T_F^{(N)}$, et $V^{(N)}$ sont donc définis par :

$$U^{(N)} = \Pi_E T_{F_{N,n}}$$

$$T_F^{(N)} = \Pi_{F_{N,n}} T_{|F_{N,n}}$$

$$V^{(N)} = \Pi_{F_{N,n}} T_{|E} \quad (=(U^{(N)})^*)$$

On pose :

$$S_N(x) = T_E - x1_E - U^{(N)}(T_F^{(N)} - x1_{F_{N,n}})^{-1}V^{(N)}$$

Autrement dit $S_N(x)$ est l'application linéaire, de E dans E, définie pour $T^{(N)}$ de la même façon que S(x) à été définie pour T.
Nous confondrons l'opérateur $S_N(x)$ et sa représentation matricielle.

REMARQUE :

Nous noterons souvent l'identité sur $F_{N,n}$ par 1_F, au lieu de 1_{F_N}, $\Lambda_i^{(N)}(x)$ désigne la $i^{ème}$ valeur propre de $S_N(x)$.

2.2.2.

Notre but est d'étudier le procédé de corrections intermédiaires qui consiste à approcher λ_i par

$$\mu_i^{(N)} = \lambda_i^{(n)} + \Lambda_i^{(N)}(\lambda_i^{(n)}) \qquad (7)$$

La correction intermédiaire $\Lambda_i^{(N)}(\lambda_i^{(n)})$ est ici aisément calculable contrairement à la correction théorique $\Lambda_i(\lambda_i^{(n)})$ intervenant dans la formule 5.

En effet :

$\Lambda_i^{(N)}(\lambda_i^{(n)})$ = la $i^{ème}$ valeur propre de

$$T_E - \lambda_i^{(n)}1_E - U^{(N)}(T_F^{(N)} - \lambda_i^{(n)}1_F)^{-1}V^{(N)}$$

et $U^{(N)}$, $T_F^{(N)}$, $V^{(N)}$ sont des matrices de dimensions finies.

On peut établir une égalité analogue à celle de la proposition 3 Mais le reste

$$R_N(x) = S_N(x) - S(x)$$

intervient dans la formule.

On peut montrer que :

$$\frac{\mu_i^{(N)}-\lambda_i}{\lambda_i} = -\frac{\lambda_i-\lambda_i^{(n)}}{\lambda_i} \cdot \frac{((1-\Pi_n)\varphi_i,(1-\Pi_n)\hat{\varphi}_{i,n})}{(\varphi_i,\hat{\varphi}_{i,n})}$$

$$-\frac{1}{\lambda_i} \frac{(R_N(\lambda_i^{(n)})\Pi_n\hat{\varphi}_{i,n},\Pi_n\varphi_i)}{(\hat{\varphi}_{i,n},\varphi_i)} \qquad (8)$$

L'égalité (8) est inutilisable car nous ne connaissons ni le signe, si une majoration du dernier terme du second membre.
Nous allons étudier autrement la précision $|\lambda_i-\mu_i^{(N)}|$.

2.2.3. ETUDE DE L'ERREUR $|\lambda_i-\mu_i^{(N)}| \,/\, |\lambda_i|$

Soit $\lambda_i^{(N)}$ la $i^{ème}$ valeur propre de $T^{(N)}$.

D 'après les propriétés de l'approximation de Galerkin on a :

$$\lambda_i^{(n)} \leq \lambda_i^{(N)} \leq \lambda_i \qquad (9)$$

De plus d'après la propriété 3 appliquée à T et à $T^{(N)}$:

$$\lambda_i \leq \mu_i \qquad (10)$$

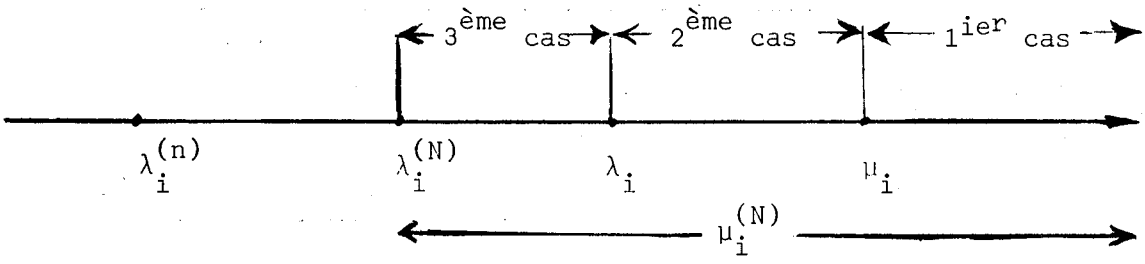
$$\lambda_i^{(N)} \leq \mu_i^{(N)} \qquad (11)$$

D'après (9), (10) et (11), il y a donc trois possibilités pour $\mu_i^{(N)}$:

1 - soit $\mu_i^{(N)} > \mu_i$

2 - soit $\lambda_i \leq \mu_i^{(N)} \leq \mu_i$

3 - soit $\lambda_i^{(N)} \leq \mu_i^{(N)} \leq \lambda_i$



Nous allons montrer qu'en fait la première possibilité n'a jamais lieu.

On en déduira le résultat suivant :

- Soit $\mu_i^{(N)}$ est une approximation de λ_i , meilleure que μ_i (possibilité 2
- Soit $\mu_i^{(N)}$ est une approximation de λ_i , meilleure que $\lambda_i^{(N)}$ (possibilité

Nous utiliserons le lemme :

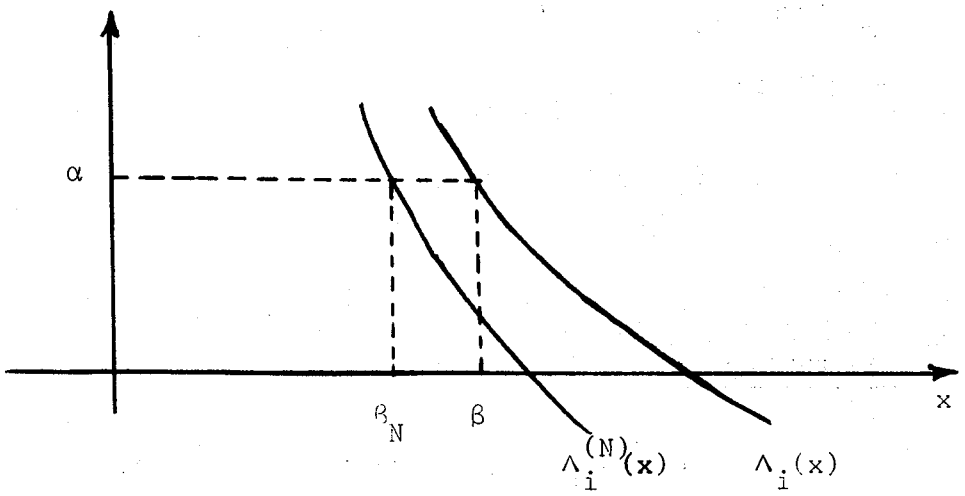
LEMME 7 :

Soit $x \in \mathbb{R}$ tel que l'opérateur $x1_F - T_F$ est défini positif. Alors :

$$\wedge_i^{(N)}(x) \leq \wedge_i(x)$$

DEMONSTRATION :

Pour montrer que $\Lambda_i^{(N)}(x) \leq \Lambda_i(x)$ il suffit de montrer que les deux solutions β_N et β des équations $\Lambda_i^N(x) = \alpha$ et $\Lambda_i(x) = \alpha$ respectivement vérifient $\beta_N \leq \beta$.



Or β_N est tel que la $i^{ème}$ valeur propre de

$$T_E - \beta_N 1_E - U^{(N)}(T_F^{(N)} - \beta_N 1_F,)^{-1} V^{(N)}$$

est α, i.e. la $i^{ème}$ valeur propre de

$$T_E - (\beta_N + \alpha) 1_F - U^{(N)}(T_F^{(N)} - \beta_N 1_F,)^{-1} V^{(N)}$$

est nulle.

Donc d'après le lemme 1 β_N est la $i^{ème}$ valeur propre de :

$$\tilde{T}^{(N)} = \left(\begin{array}{c|c} T_E - \alpha 1_E & U^{(N)} \\ \hline V^{(N)} & T_F^{(N)} \end{array} \right)$$

Le même raisonnement appliqué à Λ_i et β montre que β est la $i^{ème}$ valeur propre de :

$$\tilde{T} = \left(\begin{array}{c|c} T_E - \alpha 1_E & U \\ \hline V & T_F \end{array} \right)$$

Or $\tilde{T}^{(N)}$ n'est autre que la partie de $\tilde{T}$ dans E' et le théorème du mini-max montre que :

$$\beta_N \leq \beta \ .$$

C.Q.F.D.

PROPOSITION 8 :

Si $\lambda_i^{(n)} 1_F - T_F$ est défini positif alors :

- soit $\lambda_i^{(n)} \leq \lambda_i^{(N)} \leq \mu_i^{(N)} \leq \lambda_i \leq \mu_i$

- soit $\lambda_i^{(n)} \leq \lambda_i^{(N)} \leq \lambda_i \leq \mu_i^{(N)} \leq \mu_i$

D'où le résultat suivant sur la précision de $\mu_i^{(N)}$:

COROLLAIRE 9 :

Si $\lambda_i^{(n)} 1_F - T_F$ est défini positif alors :

* soit $\mu_i^{(N)} \leq \lambda_i$, et $\left| \frac{\lambda_i - \mu_i^{(N)}}{\lambda_i} \right| \leq \rho_N \| (1-\Pi_N)\varphi_i \|^2$

* soit $\mu_i^{(N)} \geq \lambda_i$, et $\left| \frac{\lambda_i - \mu_i^{(N)}}{\lambda_i} \right| \leq K \| (1-\Pi_n)\varphi_i \|^4$

où $\rho_{i,N}$ et K sont des coefficients , $\rho_{i,N}$ tendant vers 1 lorsque $N \to \infty$.

DEMONSTRATION DE LA PROPOSITION :

. $\mu_i^{(N)} \leq \mu$ d'après le lemme 7 . En effet :

$$\mu_i^{(N)} = \Lambda_i^{(N)}(\lambda_i^{(n)}) + \lambda_i^{(n)} \leq \Lambda_i(\lambda_i^{(n)}) + \lambda_i^{(n)} = \mu_i \quad .$$

. $\mu_i^{(N)} \geq \lambda_i^{(N)}$ car d'après le lemme 1, $\Lambda_i^{(N)}(\lambda_i^{(N)}) = 0$ et de plus $\Lambda_i^{(N)}(x)+x$ est décroissante. Donc comme $\lambda_i^{(n)} \leq \lambda_i^{(N)}$ on aura :

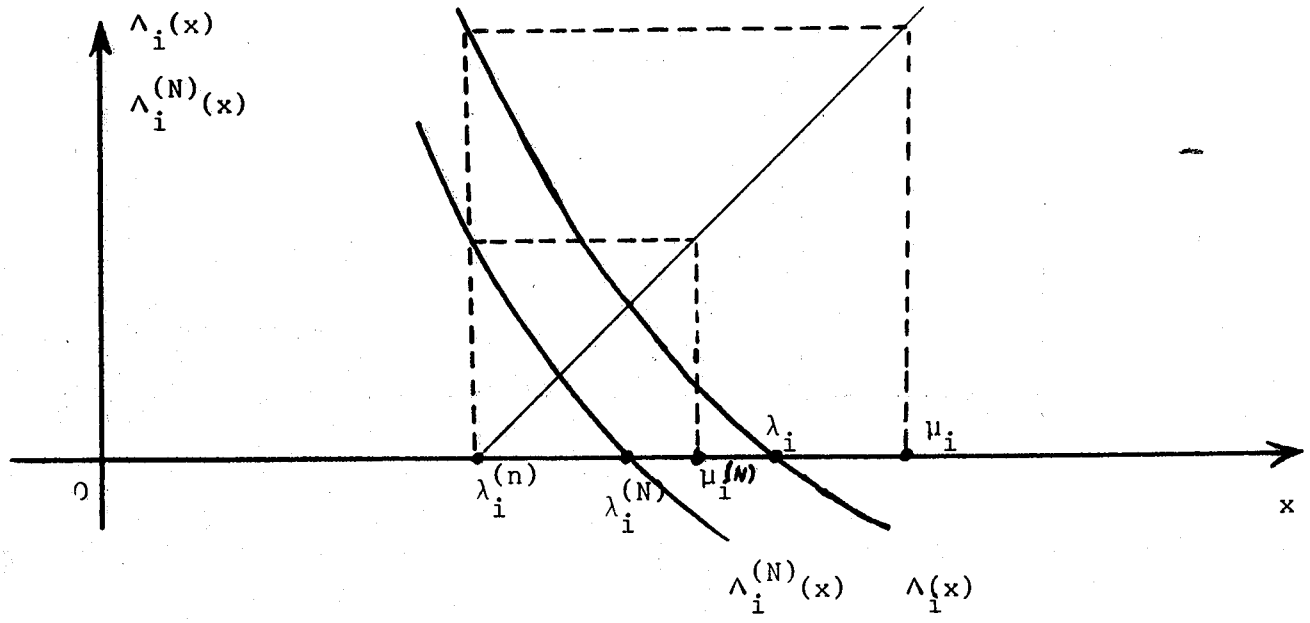
$$\mu_i^{(N)} = \Lambda_i^{(N)}(\lambda_i^{(n)}) + \lambda_i^{(n)} \geq \Lambda_i^{(N)}(\lambda_i^{(N)}) + \lambda_i^{(N)} \quad .$$

INTERPRETATION GEOMETRIQUE . COMPORTEMENT DE $\mu_i^{(N)}$ LORSQUE N VARIE (n fixé).

. $\mu_i^{(N)}$ est obtenue par une itération des approximations successives appliquée à l'équation $\Lambda_i^{(N)}(x)+x = x$ à partir de $x = \lambda_i^{(n)}$.

De même μ_i est obtenue par une itération du procédé des approximations successives appliquée à $\Lambda_i(x)+x = x$ à partir de $x = \lambda_i^{(n)}$.

Comme $\Lambda_i^{(N)}(x) \leq \Lambda_i(x)$ pour $x > \rho(T_F)$ on a le schéma suivant :



3. EXPERIENCES NUMERIQUES.

Soit T l'opérateur à noyau dans $\mathcal{L}^2[0,1]$ dont le noyau est

$$k(x,y) = \begin{cases} (1-x)y & \text{si } y \geq x \\ (1-y)x & \text{si } y < x \end{cases}$$

T est l'inverse de l'opérateur : $u \to -u''$ défini dans l'ensemble des fonctions de $\mathcal{C}^2[0,1]$ telles que $u(0) = u(1) = 0$.

Les valeurs propres de T sont connues et sont

$$\lambda_k = \frac{1}{k^2 \Pi^2} \qquad k=1,2,\ldots,n,\ldots$$

Comme base de $\mathcal{L}^2[0,1]$ on prend :

$$e_1(x) = 1$$

$$e_k(x) = \sqrt{2} \cos((k-1)\Pi x). \qquad k=2,3,\ldots,n.$$

Si E_N est le sous-espace vectoriel de dimension N engendré par $(e_1,e_2,\ldots,e_N)$, la matrice représentant matriciellement la partie de T dans E_N a pour éléments non nuls (*) :

$$t_{11} = \frac{1}{12}$$

$$t_{2p+1,1} = \frac{-1}{2\sqrt{2}p^2\pi^2} \qquad ,p=1,2,\ldots,\left[\frac{N}{2}\right]-1$$

$$t_{2p,2p} = \frac{1}{(2p-1)^2\pi^2} - \frac{8}{(2p-1)^4\pi^4} \qquad p=1,2,\ldots\left[\frac{N}{2}\right]$$

$$t_{2p+1,2p+1} = \frac{1}{(2p)^2\pi^2} \qquad p=1,2,\ldots\left[\frac{N}{2}\right]-1$$

$$t_{2p,2q} = \frac{8}{(2p-1)^2(2q-1)^2\pi^4} \qquad p\neq q \ , \ p=1,\ldots\left[\frac{N}{2}\right]$$

$$q=1,2,\ldots\left[\frac{N}{2}\right]$$

(*) Cet exemple est tiré de :
F. CHATELIN : "Méthodes numériques de calcul des valeurs propres et vecteurs propres d'un opérateur linéaire". Thèse U.S.M.G. (1971).

Dans les expériences qui suivent nous avons choisi différents n et N et fait calculer quelques valeurs propres.

Nous indiquons dans les tableaux de 1 à 9 les résultats obtenus. Ceux-ci comprennent :

1°) les valeurs propres exactes que l'on veut approcher

2°) les valeurs propres approchées $\lambda_i^{(n)}$ ainsi que les précisions

$$\frac{\lambda_i-\lambda_i^{(n)}}{\lambda_i}$$

3°) les valeurs corrigées $\mu_i^{(n)}$ et les précisions correspondantes

$$\frac{\lambda_i-\mu_i^{(n)}}{\lambda_i}$$

4°) Les valeurs propres $\lambda_i^{(N)}$ et les précisions

$$\frac{\lambda_i-\lambda_i^{(N)}}{\lambda_i}$$

Pour chacune des trois méthodes (méthode de Galerkin avec la dimension n, méthode des corrections, et méthode de Galerkin avec la dimension N) nous indiquons les temps d'exécutions.

Le caractère "n" ne pouvant pas être imprimé sur la console nous avons choisi de le remplacer par "No".

3.1. CALCUL D'UNE VALEUR PROPRE

TABLEAU 1

```
   NO   =              8     ,   N  =              80
 ------------------------------------------------------
CALCUL       DE   LA          PLUS GRANDE  VALEUR  PROPRE
 ------------------------------------------------------

          ** VALEUR  PROPRE  EXACTE  **
              +00 101321183642338


  ***  METHODE DE GALERKIN POUR NO=              8     ***
       * TEMPS D'EXECUTION:     0.07630998   SEC. *

 * I *        * LAMDA(I,NO) *         * PRECISION  *
   1       +00 101243472803320          0.0007669752


  ***  METHODE DES CORRECTIONS , N =             80     ***
       * TEMPS D'EXECUTION:       8.088234   SEC. *

 * I *          * MU(I,N) *          * PRECISION  *
   1       +00 101321188405729      -4.701279'-08


  ***    METHODE DE GALERKIN POUR N=              80     ***
       * TEMPS D'EXECUTION:       33.52561   SEC. *

 * I *        * LAMDA(I,N) *          * PRECISION  *
   1       +00 101321128129840          5.478864'-07
```

3.2. CALCUL DE DEUX VALEURS PROPRES

TABLEAU 2

```
   NO   =              6     ,   N  =              36
 ----------------------------------------------------
CALCUL DES             2  PLUS GRANDES VALEURS PROPRES
 ----------------------------------------------------

          ** VALEURS PROPRES EXACTES **
              +00 101321183642338
              -01 253302959105844


  ***  METHODE DE GALERKIN POUR NO=              6     ***
       * TEMPS D'EXECUTION:            0   SEC. *

* I *        * LAMDA(I,NO) *          * PRECISIONS *
  1       +00 101112825684357          0.002056411
  2       -01 252014439199838          0.005086873


  ***  METHODE DES CORRECTIONS , N =            36     ***
       * TEMPS D'EXECUTION:       1.596348   SEC. *

* I *          * MU(I,N) *          * PRECISIONS *
  1       +00 101320982374880       1.986430'-06
  2       -01 253304071184747      -4.390312'-06


  ***    METHODE DE GALERKIN POUR N=              36     ***
         * TEMPS D'EXECUTION:       3.279067   SEC. *

* I *        * LAMDA(I,N) *          * PRECISIONS *
  1       +00 101320545860985       6.294650'-06
  2       -01 253297089569380       2.317200'-05
```

TABLEAU 3

```
   N0   =              6    ,   N   =              50
-----------------------------------------------------
CALCUL DES             2  PLUS GRANDES VALEURS PROPRES
-----------------------------------------------------

          ** VALEURS PROPRES EXACTES **
              +00 1013211836423 38
              -01 2533029591058 44


   ***  METHODE DE GALERKIN POUR N0=              6     ***
        * TEMPS D'EXECUTION:   0.004887999   *

* I *        * LAMDA(I,N0) *          * PRECISIONS *
  1       +00 1011128256843 57          0.002056411
  2       -01 2520144391998 38          0.005086873


   ***  METHODE DES CORRECTIONS , N =             50     ***
        * TEMPS D'EXECUTION:    4.073965   *

* I *          * MU(I,N) *          * PRECISIONS *
  1       +00 1013213692865 80     -2.029627'-06
  2       -01 2533077891728 96     -1.900834'-05


   ***    METHODE DE GALERKIN POUR N=              50     ***
        * TEMPS D'EXECUTION:    8.414820   *

* I *        * LAMDA(I,N) *           * PRECISIONS *
  1       +00 1013209510856 80      2.295242'-06
  2       -01 2533007686643 45      8.647516'-06
```

TABLEAU 4

```
   NO   =              8    ,   N  =              70
-------------------------------------------------------
CALCUL DES             2  PLUS GRANDES VALEURS PROPRES
-------------------------------------------------------

        ** VALEURS PROPRES EXACTES **
           +00 101321183642338
           -01 25330295910584 4

  ***  METHODE DE GALERKIN POUR NO=              8     ***
       * TEMPS D'EXECUTION:     0.07524398   SEC. *

* I *       * LAMDA(I,NO) *         * PRECISIONS *
  1       +00 101243472803320        0.0007669752
  2       -01 252763795839930          0.002128531

  ***  METHODE DES CORRECTIONS , N =             70     ***
       * TEMPS D'EXECUTION:       10.81212   SEC. *

* I *         * MU(I,N) *          * PRECISIONS *
  1       +00 101321160565812       2.277562'-07
  2       -01 253303351620891      -1.549587'-06

  ***    METHODE DE GALERKIN POUR N=             70     ***
       * TEMPS D'EXECUTION:       22.15675   SEC. *

* I *       * LAMDA(I,N) *          * PRECISIONS *
  1       +00 101321100332824       8.222319'-07
  2       -01 253302160906013       3.151167'-06
```

TABLEAU 5

```
   NO   =              8     ,   N  =              80
 ----------------------------------------------------
CALCUL DES              2  PLUS GRANDES VALEURS PROPRES
 ----------------------------------------------------


        ** VALEURS PROPRES EXACTES **
            +00 101321183642338
            -01 253302959105844


 ***  METHODE DE GALERKIN POUR NO=              8     ***
      * TEMPS D'EXECUTION:     0.05743399   SEC. *

* I *       * LAMDA(I,NO) *         * PRECISIONS *
  1       +00 101243472803320          0.0007669752
  2       -01 252763795839930          0.002128531


 ***  METHODE DES CORRECTIONS , N =             80     ***
      * TEMPS D'EXECUTION:       16.18422   SEC. *

* I *         * MU(I,N) *          * PRECISIONS *
  1       +00 101321188405729      -4.701279'-08
  2       -01 253303616245043      -2.594282'-06


 ***    METHODE DE GALERKIN POUR N=              80     ***
     * TEMPS D'EXECUTION:       32.99917   SEC. *

* I *       * LAMDA(I,N) *          * PRECISIONS *
  1       +00 101321128129840        5.478864'-07
  2       -01 253302424384981        2.110993'-06
```

3.3. CALCUL DE TROIS VALEURS PROPRES

TABLEAU 6

```
   NO    =              10     ,    N   =              60
 ----------------------------------------------------------
CALCUL DES              3   PLUS GRANDES VALEURS PROPRES
 ----------------------------------------------------------


          ** VALEURS PROPRES EXACTES **
              +00 101321183642338
              -01 253302959105844
              -01 112579092935931


  ***  METHODE DE GALERKIN POUR NO=              10     ***
       * TEMPS D'EXECUTION:      0.1284920   SEC. *

* I *        * LAMDA(I,NO) *          * PRECISIONS *
  1        +00 101284246030883          0.0003645596
  2        -01 253027823980237          0.001086190
  3        -01 112185800752021          0.003493474


  ***  METHODE DES CORRECTIONS , N =              60     ***
       * TEMPS D'EXECUTION:      10.45005   SEC. *

* I *          * MU(I,N) *           * PRECISIONS *
  1        +00 101321063873022          1.182076'-06
  2        -01 253301995240164          3.805189'-06
  3        -01 112579226677767         -1.187981'-06


  ***    METHODE DE GALERKIN POUR N=              60     ***
         * TEMPS D'EXECUTION:      14.06438   SEC. *

* I *          * LAMDA(I,N) *           * PRECISIONS *
  1        +00 101321050400966          1.315040'-06
  2        -01 253301691554865          5.004091'-06
  3        -01 112577758672497          1.185179'-05
```

TABLEAU 7

```
   N0   =            10     ,   N  =             70
-----------------------------------------------------
CALCUL DES             3  PLUS GRANDES VALEURS PROPRES
-----------------------------------------------------

        ** VALEURS PROPRES EXACTES **
            +00 101321183642338
            -01 253302959105844
            -01 112579092935931


 ***  METHODE DE GALERKIN POUR N0=             10     ***
      * TEMPS D'EXECUTION:     0.1235780   SEC. *

* I *        * LAMDA(I,N0) *          * PRECISIONS *
  1       +00 101284246030883          0.0003645596
  2       -01 253027823980237          0.001086190
  3       -01 112185800752021          0.003493474


 ***  METHODE DES CORRECTIONS , N =            70     ***
      * TEMPS D'EXECUTION:     16.07496   SEC. *

* I *          * MU(I,N) *          * PRECISIONS *
  1       +00 101321113841334       6.889083'-07
  2       -01 253302465621404       1.948198'-06
  3       -01 112579730632287      -5.664430'-06


 ***    METHODE DE GALERKIN POUR N=             70     ***
      * TEMPS D'EXECUTION:     21.71717   SEC. *

* I *        * LAMDA(I,N) *           * PRECISIONS *
  1       +00 101321100332824       8.222319'-07
  2       -01 253302160906013       3.151167'-06
  3       -01 112578258996129       7.407590'-06
```

TABLEAU 8

```
  NU   =            10    ,   N  =            80
-------------------------------------------------
CALCUL DES             3  PLUS GRANDES VALEURS PROPRES
-------------------------------------------------

        ** VALEURS PROPRES EXACTES **
            +00 101321183642338
            -01 253302959105844
            -01 112579092935931


 ***  METHODE DE GALERKIN POUR NO=            10    ***
      * TEMPS D'EXECUTION:    0.09229998   SEC. *

* I *       * LAMDA(I,NO) *        * PRECISIONS *
  1      +00 101284246030883        0.0003645590
  2      -01 253027823980237        0.001086190
  3      -01 112185800752021        0.003493474


 ***  METHODE DES CORRECTIONS , N =           80    ***
      * TEMPS D'EXECUTION:     23.70508   SEC. *

* I *         * MU(I,N) *         * PRECISIONS *
  1      +00 101321141658665       4.143622'-07
  2      -01 253302729679329       9.057396'-07
  3      -01 112580011039739      -8.155189'-06


 ***    METHODE DE GALERKIN POUR N=            80    ***
     * TEMPS D'EXECUTION:     32.06280   SEC. *

* I *       * LAMDA(I,N) *         * PRECISIONS *
  1      +00 101321128129840       5.478864'-07
  2      -01 253302424364961       2.110995'-06
  3      -01 112578537361936       4.934768'-06
```

TABLEAU 9

N0 = 10 , N = 90

CALCUL DES 3 PLUS GRANDES VALEURS PROPRES

** VALEURS PROPRES EXACTES **
+00 101321183642338
-01 253302959105844
-01 112579092935931

*** METHODE DE GALERKIN POUR N0= 10 ***
* TEMPS D'EXECUTION: 0.1091220 SEC. *

I	LAMDA(I,N0)	PRECISIONS
1	+00 101284246030883	0.0003645596
2	-01 253027823980237	0.001086190
3	-01 112185800752021	0.003493474

*** METHODE DES CORRECTIONS , N = 90 ***
* TEMPS D'EXECUTION: 34.28071 SEC. *

I	MU(I,N)	PRECISIONS
1	+00 101321158357367	2.495527'-07
2	-01 253302889203750	2.759624'-07
3	-01 112580179311106	-9.649884'-06

*** METHODE DE GALERKIN POUR N= 90 ***
* TEMPS D'EXECUTION: 46.21850 SEC. *

I	LAMDA(I,N)	PRECISIONS
1	+00 101321144816340	3.831972'-07
2	-01 253302583559403	1.482598'-06
3	-01 112578704439687	3.450874'-06

3.4. CONCLUSION :

Nous allons comparer, à la lumière des expériences numériques précédentes les deux méthodes suivantes :

Méthode I :

On choisit $n << N$ et on approche la $i^{ème}$ valeur propre de T par $\mu_i^{(N)}$ = la $i^{ème}$ valeur propre de $S_N(\lambda_i^{(n)}) + \lambda_i^{(n)}$.

Méthode II :

On calcule la $i^{ème}$ valeur propre de T_{E_N} en tridiagonalisant la matrice représentant T_{E_N} dans une base et en appliquant QR à la matrice tridiagonale obtenue.

On constate que du point de vue de la précision, les méthodes I et II sont à peu près identiques, alors que le temps d'exécution de la méthode II est environ quatre fois plus long que celui de la méthode I. Ceci est valable pour le calcul d'une valeur propre seulement.
Pour chaque valeur propre supplémentaire calculée il faut recommencer la méthode I. Donc en particulier si on calcule deux valeurs propres par la méthode I, le temps d'exécution total pour obtenir les deux valeurs propres sera deux fois plus long que celui mis pour obtenir une valeur propre .

Ceci s'explique aisément si on fait le décompte des opérations nécessaires à l'exécution des méthodes I et II.
La méthode I exige le calcul de la $i^{ème}$ valeur propre de

$$T_E + U^{(N)}(\lambda_i^{(n)} - T_F^{(N)})^{-1} U^{(N)*}$$

On procède ainsi :

1. Décomposer selon la décomposition de Choleski :

$\lambda_i^{(n)} 1_{F'} - T_F^{(N)}$

Soit : $\lambda_i^{(n)} 1_{F'} - T_F^{(N)} = RR^T$

Nombre de multiplications nécessitées $\sim \frac{1}{6} (N-n)^3$

2. Faire $U^{(N)} (\lambda_i^{(n)} 1_{F'} - T_F^{(N)})^{-1} U^{(N)*} = (R^{-1} U^{(N)T})^T (R^{-1} U^{(N)T})$

Nombre d'opérations nécessitées : $\sim \frac{2}{3} (N-n)^2 \times n + n^2 (N-n)$

3. Calculer les valeurs propres de :

$$S_N(\lambda_i^{(n)}) + \lambda_i^{(n)} .$$

Nombre de multiplications nécessitées : $\sim O(n^2)$.

On voit donc que pour la méthode I le nombre total de multiplications est en $\frac{1}{6} N^3$.

Pour la méthode II, le nombre de multiplications est $\frac{2}{3} N^3$ pour la tridiagonalisation par la méthode de Householder. Le nombre de multiplications nécessitées pour la réalisation de l'algorithme QR est en N^2 et il est négligeable par rapport aux $\frac{2}{3} N^3$.

On retrouve bien ce que l'on a constaté expérimentalement : Si on calcule ℓ valeurs propres et si t est le temps mis pour exécuter une multiplication

- le temps d'exécution de la méthode I est de l'ordre de

$$\ell \times \frac{1}{6} N^3 \times t$$

- le temps d'exécution de la méthode II est de l'ordre de

$$\frac{2}{3} N^3 \times t$$

Ceci indique en particulier que le procédé I n'est intéressant que si le nombre de valeurs propres à calculer n'excède pas 4.

BIBLIOGRAPHIE POUR LA DEUXIEME PARTIE

[1] ABRAMOV, A. ; NEUHAS, M.

"Bemerkungen über Eingeinvertprobleme von Matrizen hoherer Ord nu
C.R. du Congrès Int. des Math. de l'Ingénieur, Mons et Bruxelles (1958).

[2] ABRAMOV, A.

"On the seperation of the principal part of some algebraic proble
Zh. Vych. Mat 2 : N° 1, 141-5, (1962).

[3] ABRAMOV, A.

"Remarks on finding the eingenvalues and eingenvectors of matric
which arise in the application of Ritz's method or in the difference method".
Zh. vych. Mat. Fiz. 7, 3, 644, 647, (1967).

[4] CHATELIN, F.

"Calcul numérique de valeurs propres d'opérateurs linéaires".
Cours D.E.A. Analyse Numérique (1973-1974).

[5] CHATELIN, F. ; LEMORDANT, J.

"La méthode de Rayleigh-Ritz appliquée à des opérateurs elliptiqu
Ordre de convergence des éléments propres".
Séminaire d'Analyse Numérique. Université Scientifique et Médical
de Grenoble, premier semestre 1973-1974. Grenoble.

[6] CHICHOV, V.S.

"A method for partitionning a high order matrix into blocks in order to find its eingenvalues".
Zh. vych. Mat. 1, N° 1, 169-173, (1961).

[7] CHICHOV, V.S.

"The determination of eingenvalues and eingenfunctions of a linear integral operator with a symmetric kernel by means of group elimination of unknowns".
Zh. vych. mat. 2, N° 3, 389-410, (1962).

[8] DEKKER, T.J. ; TRAUB, J.F.

"The shifted QR algorithm for hermitian matrices".
J. of lin. Alg. and its appl. 4, 137-154, (1971).

[9] GREGORY, R.T. ; KARNEY, D.A.

"A collection of matrices for testing computationnal algorithms"
John Wiley and Sons. New-York (1969).

[10] KATO, T.

"Perturbation theory for linear operators"
Springer Verlag, (1965).

[11] LEBAUD, C.

"L'algorithme double QR avec shift".
Numer. Mat. 16, 163-180, (1970).

[12] PARLETT, B.N. ;

"Présentation géométrique des méthodes de calcul des valeurs propres".
Numer. Mat. 21, 223-233, (1973).

[13] SAAD, Y.

"Quelques applications du partitionnement au calcul de valeurs propres de matrices hermitiennes".
Séminaire d'Analyse Numérique, Université Scientifique et Médica de Grenoble, premier semestre 1973-1974. Grenoble.

[14] WILKINSON, J.H.

The algebraic Eingenvalue Problem.

Clarendon Press. London (1965).

[15] WILKINSON, J.H. ; REINSCH, C.

Handbook for automatic computation.

Volum II ; linear Algebra

Springer Verlag (1971).

[16] WILKINSON, J.H.

Global convergence of QR algorithm.

Proceedings of the IFIP congress (1968) A22-A24.

TROISIEME PARTIE

SUR LE CALCUL DES ELEMENTS PROPRES DE TRES GRANDES MATRICES

INTRODUCTION :

Soit A une matrice symétrique d'ordre N, supposé grand, dont nous voulons calculer certaines valeurs propres, généralement les plus grandes en valeur absolue.

Les algorithmes classiques QR, LR, Jacobi, etc..., très performants pour les matrices de petites dimensions (en général $N \leq 80$) ne peuvent plus s'appliquer à A lorsque N est grand, pour deux raisons principales :

1°/ Ils sont trop coûteux en temps-machine : on peut calculer que, par exemple $N = 10^4$, pour réduire A à la forme tridiagonale par l'algorit de Householder il faut, en supposant qu'une multiplication coûte 10^{-6} seco au moins $\frac{2}{3} N^3 \times 10^{-6}$. sec $\sim$ 200 heures !

Nous avons négligé ici les problèmes de l'organisation des calcu ces algorithmes sont en général mal adaptés à l'organisation paginée de la mémoire centrale.

2°/ Le nombre d'opérations demandées étant très grand, l'accumulatio des erreurs d'arrondis est telle que le résultat est entièrement faussé.

Il est donc nécessaire de ramener le calcul des éléments propres de A à celui d'une matrice d'ordre n suffisamment petit (par exemple $n \leq 8$ pour que l'on puisse appliquer les algorithmes de haute précision.

Pour cela, certaines méthodes numériques consistent à appliquer la méthode de Galerkin à A : on approche certains éléments propres de A pa ceux de $\pi_n A \pi_n$ où π_n est une projection sur un sous-espace vectoriel de dimension finie n "presque invariant par A". cf [1,12,14,17].

Nous commençons au §1 par préciser la notion d'approximation d'un espace invariant et nous donnons les bornes d'erreurs qui en découlent pour tous les éléments propres. Cependant, on verra que ces bornes ne sont souvent que des sur-estimations très larges de l'erreur effective, parce que généralement les éléments propres de A ne sont pas approchés de façon uniforme : on peut avoir de bonnes approximations de certains éléments propres de A sans que E soit nécessairement un sous-espace invariant approché.

C'est précisémment le cas lorsque l'on prend pour E l'espace vectoriel engendré par la suite de Krylov $X_o, AX_o, A^2X_o, \ldots, A^{n-1}X_o$, où X_o est un vecteur initial quelconque. La méthode obtenue, qui s'appelle alors la méthode de Rayleigh généralisée (ou méthode de Lanczos), possède l'avantage très appréciable de n'utiliser que les produits $X \to AX$, qui sont bien adaptés à l'organisation en pages des mémoires des ordinateurs. Cette méthode est étudiée au §2.

Nous donnerons des bornes d'erreurs sur les vecteurs propres et nous complèterons les résultats partiels donnés par S. Kaniel, pour les valeurs propres.

Kaniel [2] a en effet démontré des résultats qui lient les erreurs sur les valeurs propres, sans pour autant donner des bornes sur les vecteurs propres.

Nous nous intéresserons davantage aux vecteurs propres qu'aux valeurs propres puisque l'erreur sur la valeur propre approchée $\lambda_i^{(n)}$, s'exprime aisément en fonction de la norme du résidu : $\|(A-\lambda_i^{(n)})\varphi_i^{(n)}\|$, où $\varphi_i^{(n)}$ est un vecteur propre approché, de norme égale à 1, associé à la valeur propre $\lambda_i^{(n)}$.

Afin de faciliter l'exposé, nous ne parlerons que de matrices symétriques réelles. Pour les matrices hermitiennes complexes, tous les résultats trouvés se généralisent de manière évidente.

1. APPROXIMATION D'ESPACES VECTORIELS INVARIANTS

1.1. DEFINITIONS ET PROPRIETES

Soit H un espace vectoriel de dimension finie N et E un sous-espace vectoriel de dimension n de H.

Soit T un opérateur linéaire sur H et π la projection orthogonale de H sur E.

Si E est invariant par T, i.e. $TE \subset E$, alors il est bien connu que les valeurs propres de T_E sont aussi des valeurs propres de T.

Beaucoup de méthodes numériques de calcul de valeurs propres construisent un sous-espace E voisin d'un certain sous-espace vectoriel E' invariant par T. cf [17, 14, 12, 1, ...] .

Afin de définir la notion de précision de l'approximation de E' par E, examinons d' abord le cas où E est invariant par T :

$$TE \subset E$$

s'écrit également :

$$\pi T x = T x \quad \text{pour tout} \quad x \text{ de } E$$

ou encore :

$$\pi T \pi y = T \pi y \quad \text{pour tout} \quad y$$

élément de H. C'est-à-dire :

$$(1-\pi) T \pi = 0 \qquad (1)$$

Une condition nécessaire et suffisante pour que E soit invar par T est donc que $(1-\pi)T\pi = 0$.

Lorsque E est quelconque un moyen immédiat de mesurer la var de E par T est de considérer une norme de $(1-\pi)T\pi$.

Nous prenons la définition proposée en [1] :

DEFINITION :

Soit T un opérateur sur H et E un sous-espace vectoriel de H de dimension n, muni de la base orthonormale $(f_1, f_2, \ldots, f_n)$.

1. On appelle variation de E par T , notée $V_T(E)$, le nombre pc suivant :

$$V_T(E) = \left(\sum_{i=1}^{n} \| (1-\pi)Tf_i \|^2 \right)^{1/2}$$

où $\| \ \|$ est la norme euclidienne sur H.

2. Si $V_T(E) = \varepsilon$, on dira que E est ε-invariant par T.

Il est a remarquer que $V_T(E)$ ne dépend pas de la base orthor choisie puisque $V_T(E)$ n'est autre que

$$\| (1-\pi)T\pi \|_2 = \left[\mathrm{tr}((1-\pi)T\pi)((1-\pi)T\pi)^* \right]^{1/2}$$

L'importance de cette définition provient du fait qu'elle dc lieu à des bornes d'erreurs uniformes sur les éléments propres : en ei nous allons voir que si E est ε-invariant par T tous les vecteurs proj et toutes les valeurs propres de T_E sont des éléments propres approché avec une précision inférieure ou égale à ε .

Nous aurons besoins des définitions suivantes :

DEFINITIONS :

1) On appelle quotient de Rayleigh d'un vecteur $y \neq 0$ de H, le nombre, noté $\mu(y)$, défini par :

$$\mu(y) = (Ty,y) / \|y\|^2$$

2) On appelle résidu d'un vecteur $y \neq 0$, le vecteur

$$r(y) = Ty - \mu(y).y$$

3) On appelle valeurs propres E-approchées de T, les valeurs propres de T_E et vecteurs propres E-approchés de T les vecte de H dont les projections sur E sont des vecteurs propres de et dont les projections sur l'orthogonal de E sont nulles.

1.2.

Soit y un vecteur quelconque de H et supposons que T soit autoadjoint. Les définitions précédentes sont utiles lorsqu'on veut estimer à postériori l'erreur que l'on fait en prenant $\mu(y)$ comme vale propre approchée de T.

PROPRIETE 1 :

Pour tout y de H, de norme 1, il existe une valeur propre λ telle que :

$$|\lambda - \mu(y)| \leq \|r(y)\| \qquad \text{(cf [10])}$$

PROPRIETE 2 :

Pour tout y de H, de norme 1, il existe une valeur propre λ telle que , si on désigne par d la distance de $\mu(y)$ aux vale propres de T autres que λ , on ait :

$$|\lambda - \mu(y)| \leq \frac{\|r(y)\|^2}{d} \qquad \text{(cf [10])}$$

PROPRIETE 3 :

Pour tout y de H, de norme 1, il existe un vecteur propre de T, de norme 1, associé à la valeur propre λ et tel que si θ désigne l'angle aigü entre φ et y on ait :

$$\sin \theta \leq \frac{\|r(y)\|}{d}$$

où d est le même que pour la propriété 2.

cf [15]

Ces inégalités sont également connues sous la forme équivalente consistant à remplacer $\|r(y)\|$ par le nombre

$$(\|Ty\|^2 - (\mu(y))^2)^{1/2} \qquad \text{cf [10]}$$

En effet,

$$\|r(y)\|^2 = \|Ty\|^2 - (Ty,y)^2$$

puisque r(y) est orthogonal à y et que l'on a la relation :

$$Ty = (T-\mu(y))y + \mu(y).y = r(y) + \mu(y)y.$$

Une autre propriété remarquable de $\mu(y)$ est que :

PROPRIETE 4 :

$$\min_{\alpha \in \mathbb{R}} \|(T-\alpha)y\| = \|(T-\mu(y))y\| . \qquad \text{cf [16]}$$

Les propriétés 2 et 3 montrent que l'un des moyens d'estimer l'erreur sur la valeur propre et sur le vecteur propre est de calculer $\|r(y)\|$.

Dans la plupart des cas pratiques c'est ce nombre qui est pris comme "l'erreur sur le vecteur propre" (tests d'arrêts, essais d'algorithmes, ...) (cf deuxième partie chapitre 2).

Que peut-on dire des résidus sur les vecteurs propres de T_E considérés comme des vecteurs propres approchés de T ?
La réponse est donnée par la proposition :

PROPOSITION 5 :

Supposons que T est autoadjoint et soit $(\tilde{\varphi}_i)_{i=1,\ldots,n}$ les vecteurs propres E-approchés de T.
Alors :

$$\left(\sum_{i=1}^{n} \|r(\tilde{\varphi}_i)\|^2\right)^{1/2} = V_T(E)$$

DEMONSTRATION :

Supposons que $\tilde{\varphi}_i$ soit associé à la valeur propre μ_i de T_E. Alors :

$$\pi T \tilde{\varphi}_i = \mu_i \tilde{\varphi}_i \qquad (1)$$

$\|r(\tilde{\varphi}_i)\| = \|T\tilde{\varphi}_i - (\pi\tilde{\varphi}_i, \tilde{\varphi}_i)\tilde{\varphi}_i\|$ $\qquad \tilde{\varphi}_i \in E$ donc :

$\|r(\tilde{\varphi}_i)\| = \|T\tilde{\varphi}_i - (T\tilde{\varphi}_i, \pi\tilde{\varphi}_i)\tilde{\varphi}_i\|$

$\|r(\tilde{\varphi}_i)\| = \|T\tilde{\varphi}_i - (\pi T\tilde{\varphi}_i, \tilde{\varphi}_i)\tilde{\varphi}_i\|$ $\qquad$ et d'après (1) :

$\|r(\tilde{\varphi}_i)\| = \|T\tilde{\varphi}_i - \mu_i\tilde{\varphi}_i\|$ $\qquad$ encore d'après (1) :

$\|r(\tilde{\varphi}_i)\| = \|T\tilde{\varphi}_i - \pi T\tilde{\varphi}_i\| = \|(1-\pi)T\tilde{\varphi}_i\|$

comme T_E est autoadjoint, $(\tilde{\varphi}_i)_{i=1,\ldots,n}$ forme une base orthonormale de E et donc d'après la définition de $V_T(E)$:

$$V_T(E) = \left(\sum_{i=1}^{n} \|r(\tilde{\varphi}_i)\|\right)^{1/2}$$

C.Q.F.D.

Une conséquence de la proposition est que si $V_T(E) = \varepsilon$, c'est-à-dire si E est ε-invariant par T, les valeurs propres E-approc $\tilde{\lambda}_i$, et les vecteurs propres E-approchés $\tilde{\varphi}_i$ vérifient d'après les prop 2 et 3 :

1 - Pour tout $\tilde{\lambda}_i$ il existe une valeur propre λ_i de T telle que

$$\sum_{i=1}^{n} |\lambda_i - \tilde{\lambda}_i| \leq \frac{\varepsilon^2}{\delta}$$

où $\delta = \min_{i=1,\ldots n} \min_{\lambda_j \neq \lambda_i} |\lambda_j - \tilde{\lambda}_i|$

2 - Pour tout vecteur propre E-approché $\tilde{\varphi}_i$, il existe un vecteu propre φ_i de T tel que les angles aigüs θ_i entre φ_i et $\tilde{\varphi}_i$, pour i=1,...,n vérifient :

$$\left(\sum_{i=1}^{n} \sin^2 \theta_i \right)^{1/2} \leq \frac{\varepsilon}{\delta}$$

comme $V_T(E)$ n'est pas aisémment calculable, ce genre de bor d'erreurs n'est intéressant que si l'on a une <u>estimation</u> de variation de E par T : c'est le cas dans la méthode de Baue [14] ,[12], et dans [1] , [17].

<u>REMARQUE</u> :

En [1] il a été démontré seulement l'inégalité $\|r(\tilde{\varphi}_i)\| \leq V_T$ La proposition 5 indique donc qu'en réalité :

$$\|r(\tilde{\varphi}_i)\| = \sqrt{(V_T(E))^2 - \sum_{j \neq i} \|r(\varphi_j)\|^2}$$

D'autre part, l'inégalité $|\lambda_i - \tilde{\lambda}_i| \leq V_T(E)$ proposée, (qui es conséquence de la propriété 1), est une majoration trop large d'après l'on vient de voir.

2. LA METHODE DE GALERKIN APPLIQUEE AU CALCUL DES ELEMENTS PROPRES D'OPERATEUR

APPLICATION AUX MATRICES .

2.1.

Dans ce paragraphe, H est un espace de Hilbert séparable de dimension infinie.

La méthode de Galerkin pour le calcul des éléments propres d'un opérateur compact T défini sur un espace de Hilbert séparable H, consiste, étant donnée une suite de projections π_n sur des sous-espace vectoriels en croissants, vérifiant $(1-\pi_n)x \to 0$ pour tout x de H, à approcher certains éléments de T par ceux de $\pi_n T \pi_n$ c'est-à-dire par le éléments propres E_n-approchés de T.

Nous allons rappeler brièvement des propriétés connues de l' ximation de Galerkin que nous utiliserons ultérieurement : elles conce les bornes d'erreurs asymptotiques dans l' approximation de Galerkin c opérateur compact.

Soient $\lambda_1, \lambda_2, \ldots, \lambda_N, \ldots$ les valeurs propres positives de T, supposons que λ_i soit de multiplicité m. Les valeurs propres positives sont ordonnées en décroissant :

$$\ldots \leq \lambda_N \leq \ldots \leq \lambda_{i+m} \leq \lambda_{i+m-1} \leq \lambda_{i+m-2} \leq \ldots \leq \lambda_1 \quad .$$

Soit φ_i un vecteur propre associé à la valeur propre λ_i et P_i la projection spectrale associée à λ_i , c'est-à-dire la projection orthogonale sur le sous-espace propre.

Soient $\lambda_j^{(n)}$ $j=1,2,\ldots,n$ les valeurs propres de T_{E_n} , comptées avec leurs ordres de multiplicité et ordonnées en décroissant, et $\varphi_j^{(n)}$ les vecteurs propres normalisés correspondants.

$P_j^{(n)}$ est la projection orthogonale sur le vecteur propre $\varphi_j^{(n)}$. On pose alors :

$$\hat{P}_{i,n} = \sum_{j=i}^{i+m-1} P_j^{(n)}$$

et on a le résultat suivant :

THEOREME 6 :

1) Pour tout vecteur propre φ_i , $\|\varphi_i\| = 1$, associée à la valeur propre λ_i ,

$$\| \varphi_i - \hat{P}_{i,n}\varphi_i\| = \rho_{i,n} \| (1-\pi_n) \varphi_i\| \quad \text{où} \quad \hat{P}_{i,n}$$

où $\hat{P}_{i,n}$ est la projection spectrale associée aux valeurs propres $\lambda_j^{(n)}$ $j=i,i+1,\ldots,i+m-1$, et où $\rho_{i,n}$ est un coefficient qui tend vers 1 lorsque n tend vers l'infini.

2) $$\frac{\lambda_i - \lambda_j^{(n)}}{\lambda_i} = \rho'_{i,n} \| (1-\pi_n) \varphi_i\|^2$$

pour $j=i,i+1,\ldots,i+m-1$ où $\rho'_{i,n}$ est un coefficient qui tend vers 1 lorsque n tend vers l'infini.

REMARQUE :

Dans le théorème, lorsque $m = 1$, $\| \varphi_i - \hat{P}_{i,n} \varphi_i \|$ représente le sinus de l'angle aigü formé par φ_i et $\varphi_i^{(n)}$.

Le théorème 6 indique l'importance du nombre $\|(1-\pi_n)\varphi_i\|$ pour les bornes d'erreurs sur les éléments propres approchés par la méthode de Galerkin.

Nous allons nous intéresser à la méthode de Galerkin appliquée aux matrices et chercher des bornes d'erreurs analogues à celles du théorème 6.

2.2. APPLICATION AUX MATRICES.

Dans ce qui suit $H = \mathbb{R}^N$.

La méthode de Galerkin peut, bien entendu, s'appliquer à une matrice A d'ordre N : on ramène alors le problème du calcul des éléments propres de A à celui du calcul des éléments propres d'une matrice d'ordre n.

Généralement N est grand et n très inférieur à N.

La méthode consiste à approcher les éléments propres de A par ceux de la partie de A dans un sous-espace vectoriel E_n de dimension n.

Lorsqu'on dispose d'une suite finie de sous-espaces vectoriels croissants E_n, on a une suite finie d'approximations des éléments propres de A.

Nous rappelons ici l'écriture du problème E_n-approché dans une base $Y_1, Y_2, \ldots, Y_n$, de H

PROPRIETE :

1) Soit Y une matrice dont les colonnes $Y_1,\ldots,Y_n$ forment une base orthonormale de $E_n(X_o)$. Alors la partie de A dans E_n est représenté relativement à cette base par :

$$Y^TAY$$

2) Soit Y une matrice dont les colonnes $Y_1,Y_2,\ldots,Y_n$ forment u base quelconque de E_n. Alors la partie A dans E_n est représ par la matrice

$$(Y^TY)^{-1}(Y^TAY)$$

DEMONSTRATION : [1]

Le 1°) est un cas particulier du 2°) que nous allons établi

Soit π_n la représentation matricielle de la projection orthogonale su Tout élément y de E_n s'écrit comme combinaison linéaire des Y_i pour $i=1,\ldots,n$. Donc en particulier on a :

$$\pi_n AY_j = \sum_{k=1}^{n} t_{kj} Y_k \qquad j=1,\ldots,n$$

La partie de A dans E est l'application linéaire qui à un vecteur $x = \sum_{k=1}^{n} x_k Y_k$ de E_n, associe le vecteur $\pi_n Ax$ de E_n et elle est donc re sentée par la matrice $\mathcal{C}$ d'éléments t_{kj} définis en (1). Calculons

De (1) on déduit $(\pi_n AY_j, Y_i) = \sum_{k=1}^{n} t_{kj}(Y_k, Y_i)$ pour $i,j = 1,\ldots,n$

On a $(\pi_n AY_j, Y_i) = (AY_j, \pi_n Y_i) = (AY_j, Y_i) = Y_i^T AY_j$

Appelons A_n la matrice d'élément général $a_{ij}^{(n)} = Y_i^T AY_j$

et B_n la matrice d'élément général $b_{ij}^{(n)} = Y_i^T Y_j$

La relation (2) s'écrit d'après (3), (4) et (5) :

$$a_{ij}^{(n)} = \sum_{k=1}^{n} b_{ik}^{(n)} t_{kj} \qquad i,j = 1,\ldots, n \qquad (6)$$

La relation (6) s'exprime matriciellement par :

$$A_n = B_n \mathcal{T}$$

d'où

$$\mathcal{T} = B_n^{-1} A_n \ .$$

C.Q.F.D.

REMARQUE :

Si A_n est la matrice $Y^T A Y$ et B_n la matrice $Y^T Y$, alors le problème approché à résoudre est donc le problème :

$$B_n^{-1} A_n \varphi^{(n)} = \lambda^{(n)} \varphi^{(n)}$$

ou encore :

$$A_n \varphi^{(n)} = \lambda^{(n)} B_n \varphi^{(n)}$$

On reconnait ici la méthode d'approximation de Rayleigh-Rit (cf [10]) appliquée à des matrices.

2.3. BORNES D'ERREURS.

Dans les bornes d'erreurs données au paragraphe 2.1 intervient dans la démonstration du théorème 6, le fait que lorsque $n \to \infty$, $\|(1-\pi_n)T\| \to 0$, car T est un opérateur compact. On ne peut pas appliquer directement ces bornes d'erreurs aux matrices puisque nous n'avons pas de notion de convergence.($n \leq N$ fini).

Cependant on peut obtenir un résultat analogue pour les matrices en faisant l'hypothèse que $\|(1-\pi_n)A\| \leq \varepsilon$, où ε est suffisamment petit (ceci remplacera "n suffisamment grand" dans le théorème 6).

Commençons par montrer le lemme suivant :

LEMME :

Soit $\lambda^{(n)}$ une valeur propre E_n-approchée de A et $\varphi^{(n)}$ un vecteur propre associé à $\lambda^{(n)}$. On suppose que $\|(1-\pi_n)A\| \leq \varepsilon$.
Alors il existe un vecteur propre φ de A associé à la valeur propre λ tel que si on pose

$$d = \min_{\lambda_j \neq \lambda} |\lambda_j - \lambda^{(n)}| \quad , \quad d' = d(1+\varepsilon^2/d^2) \quad , \quad d'' = |\lambda| \left(1 - \frac{\varepsilon}{d'}\right)$$

et si $\varepsilon < d'$, on ait

$$1°/ \qquad \|\varphi - \varphi^{(n)}\| \leq \frac{\varepsilon}{d'}$$

$$2°/ \qquad |\lambda - \lambda^{(n)}| \leq \frac{\varepsilon^2}{d''}$$

DEMONSTRATION :

1°) On a :

$$\|r(\varphi^{(n)})\| = \|(A-\lambda^{(n)})\varphi^{(n)}\| = \|\pi_n(A-\lambda^{(n)})\varphi^{(n)} + (1-\pi_n)(A-\lambda^{(n)})\varphi^{(n)}\|$$

$$\pi_n (A-\lambda^{(n)})\varphi^{(n)} = 0$$

par définition des éléments propres E_n-approchés.
D'où :

$$\|r(\varphi^{(n)})\| = \|(1-\pi_n)A\varphi^{(n)}\| \qquad (1)$$

$$\|r(\varphi^{(n)})\| = \|(1-\pi_n)A\varphi^{(n)}\| \leq \varepsilon$$

et d'après la propriété 3, il existe un vecteur propre φ de A tel que l'angle θ entre φ et $\varphi^{(n)}$ vérifie :

$$\sin\theta \leq \varepsilon / d \ .$$

Par un simple raisonnement géométrique on voit que

$$\|\varphi-\varphi^{(n)}\| = 2\sin\frac{\theta}{2} \leq \sin\theta\,(1+\sin^2\theta) \leq \frac{\varepsilon}{d}\left(1+\frac{\varepsilon^2}{d^2}\right)$$

$$\left(\sin\theta = \frac{2\mathrm{tg}\theta/2}{1+\mathrm{tg}^2\theta/2} \geq \frac{2\sin(\frac{\theta}{2})}{1+\mathrm{tg}^2\theta/2} \geq \frac{2\sin\theta/2}{1+\sin^2\theta}\right)$$

2°) $$|\lambda-\lambda^{(n)}| = \frac{((A-\lambda^{(n)})\varphi,\varphi^{(n)})}{(\varphi,\varphi^{(n)})} = \frac{(\varphi,(A-\lambda^{(n)})\varphi^{(n)})}{(\varphi,\varphi^{(n)})} \qquad (2)$$

$$= \frac{(\varphi,(1-\pi_n)A\varphi^{(n)})}{(\varphi,\varphi^{(n)})} \quad \text{d'après la relation (1)}$$

D'où :

$$|\lambda-\lambda^{(n)}| = \frac{|((1-\pi_n)\varphi,(1-\pi_n)A\varphi^{(n)}|}{|(\varphi,\varphi^{(n)})|}$$

$$= \frac{1}{|\lambda|}\,\frac{|((1-\pi_n)A\varphi,(1-\pi_n)A\varphi^{(n)}|}{|(\varphi,\varphi^{(n)})|}$$

$$\|(1-\pi_n)A\varphi\| \leq \varepsilon \ , \quad \|(1-\pi_n)A\varphi^{(n)}\| \leq \varepsilon$$

si $(\varphi,\varphi^{(n)}) > 0$, on a :

$$(\varphi,\varphi^{(n)}) = (\varphi,\varphi) - (\varphi,(\varphi-\varphi^{(n)}) \geq 1 - \frac{\varepsilon}{d}$$

d'où le résultat.

(Si $(\varphi,\varphi^{(n)}) < 0$ on échange φ en $-\varphi$).

C.Q.F.D.

Nous aurons besoin ultérieurement, dans la méthode de Rayle: généralisée , de majorer

$$\frac{|\lambda-\lambda^{(n)}|}{|\lambda|} \text{ et } \|r(\varphi^{(n)})\| .$$

Nous pouvons montrer le résultat suivant concernant ces deux nombres

PROPOSITION 6' :

Soit $\lambda^{(n)}$ une valeur propre E_n-approchée de A et $\varphi^{(n)}$ un ve propre associé à $\lambda^{(n)}$. On suppose que $\|(1-\pi_n)A\| \leq \varepsilon$.
Alors il existe un vecteur propre φ de A associé à la valeu propre λ tel que si on pose :

$$d = \min_{\lambda_j \neq \lambda} |\lambda_j - \lambda^{(n)}| \ , \ d' = d/(1+\frac{\varepsilon^2}{d^2})$$

et si $\varepsilon < d'$, on ait :

$$1°/ \qquad \frac{1}{|\lambda|}\|r(\varphi^{(n)})\| \leq \|(1-\pi_n)\,\varphi\| + \frac{\varepsilon^2}{|\lambda| d'}$$

$$2°/ \qquad \frac{|\lambda-\lambda^{(n)}|}{|\lambda|} \leq \frac{1}{1-\frac{\varepsilon}{d'}} \left(\|(1-\pi_n)\,\varphi\|^2 + \frac{\varepsilon^3}{\lambda^2 d'}\right)$$

DEMONSTRATION :

1°/ D'après la relation (1) de la démonstration du lemme :

$$\|r(\varphi^{(n)})\| = \|(1-\pi_n)A\varphi^{(n)}\|$$

$$= \|(1-\pi_n)A\varphi - (1-\pi_n)A(\varphi-\varphi^{(n)})\|$$

$$\leq \|(1-\pi_n)A\varphi\| + \|(1-\pi_n)A\|\|\varphi-\varphi^{(n)}\|$$

. $\|(1-\pi_n)A\varphi\| = |\lambda|\|(1-\pi_n)\varphi\|$

. $\|(1-\pi_n)A\| \ \|\varphi-\varphi^{(n)}\| \leq \varepsilon(\frac{\varepsilon}{d'})$

d'après le lemme. D'où le résultat.

2°/ De la relation (2) de la démonstration du lemme on tire :

$$(\lambda-\lambda^{(n)}) = \frac{(\varphi,(A-\lambda^{(n)})\varphi^{(n)})}{(\varphi,\varphi^{(n)})}$$

$$= \frac{(\varphi,(1-\pi_n)A\varphi^{(n)})}{(\varphi,\varphi^{(n)})} = \frac{((1-\pi_n)\varphi,(1-\pi_n)A\varphi^{(n)})}{(\varphi,\varphi^{(n)})}$$

On écrit que $\varphi^{(n)} = \varphi-(\varphi-\varphi^{(n)})$; d'où :

$$\lambda-\lambda^{(n)} = \frac{(1-\pi_n)\varphi,(1-\pi_n)A\varphi)}{(\varphi,\varphi^{(n)})} - \frac{((1-\pi_n)\varphi,(1-\pi_n)A(\varphi-\varphi^{(n)}))}{(\varphi,\varphi^{(n)})}$$

$$\lambda-\lambda^{(n)} = \lambda\frac{((1-\pi_n)\varphi,(1-\pi_n)\varphi)}{(\varphi,\varphi^{(n)})} - \frac{1}{\lambda}\frac{((1-\pi_n)A\varphi,(1-\pi_n)A(\varphi-\varphi^{(n)}))}{(\varphi,\varphi^{(n)})}$$

$$\frac{\lambda-\lambda^{(n)}}{\lambda} = \frac{1}{(\varphi,\varphi^{(n)})}\left[\|(1-\pi_n)\varphi\|^2 - \frac{1}{\lambda^2}((1-\pi_n)A\varphi,(1-\pi_n)A(\varphi-\varphi^{(n)}))\right]$$

$$\leq \frac{1}{|(\varphi,\varphi^{(n)})|}\left[\|(1-\pi_n)\varphi\|^2 + \frac{1}{\lambda^2}\|(1-\pi_n)A\|.\|(1-\pi_n)A\|\|\varphi-\varphi^{(n)}\|\right)$$

En appliquant alors le lemme on aboutit au résultat cherché.

REMARQUES :

1°) Les résultats du lemme et de la proposition 6 sont valables également pour les opérateurs compacts.

2°) L'inégalité 2) de la proposition peut également donner l'inégalité

$$\left|\frac{\lambda-\lambda^{(n)}}{\lambda}\right| \leq \|(1-\pi_n)\varphi\|^2 + 2\frac{\varepsilon^3}{\lambda^2 d'} + \frac{\varepsilon^4}{\lambda^2 d'^2}$$

3. LA METHODE DE RAYLEIGH GENERALISEE .

Soit A une matrice symétrique dont on recherche les éléments propres. Nous allons considérer une méthode de Galerkin particulière dans laquelle l'espace approximant E_n est le sous-espace vectoriel engendré par les vecteurs $X_o, AX_o, \ldots, A^{n-1}X_o$, où X_o est un vecteur initial quelconque.

Cette méthode connue sous des formes diverses ([1],[3],[4],...) est appelée méthode de Rayleigh généralisée. Elle donne de bonnes approximations des éléments propres correspondants aux valeurs propres situées aux extrémités du spectre.

Notre but est de donner une idée de la précision de l'approximation des éléments propres en proposant des bornes d'erreurs à priori : les expériences numériques montrent que les majorations théoriques obtenues sur les erreurs, évoluent de la même manière que les erreurs exactes, tout en restant assez larges.

Au paragraphe 3.1 nous commençons par décrire cette méthode et nous proposons en particulier une présentation différente de la méthode, qui fait intervenir un problème de meilleure approximation dans un espace de polynômes. Nous nous inspirons pour cela d'un résultat démontré par Vandergraft [1].

Au paragraphe 3.3, nous déduisons des résultats précédents, des bornes d'erreurs sur les vecteurs propres et sur les valeurs propres.

3.1. SUITE DE KRYLOV , ESPACES $E_n(X_o)$.

Dans tout ce qui suit on note par la même lettre A l'opérateur T et sa représentation matricielle A.

- On appelle suite de Krylov issue de X_o, la suite de vecteurs $X_o, AX_o, A^2X_o, \ldots, A^nX_o, \ldots$ on pose $X_i = A^iX_o$ pour tout entier $i \in \mathbb{N}$ $X_o \neq 0$ est le vecteur initial.

- On notera $E_n(X_o)$ (ou plus simplement E_n s'il n'y a pas d'ambigüité) l'espace vectoriel engendré par les n vecteurs $X_o, AX_o, \ldots, A^{n-1}X_o$.

- $\mathcal{P}_n^*$ désigne l'ensemble des polynômes de degré n, dont le coefficient de plus haut degré est 1

$$p \in \mathcal{P}_n^* \iff p(x) = x^n + a_{n-1}x^{n-1} + \ldots + a_o .$$

- $\mathcal{P}_n$ est l'ensemble des polynômes de degré inférieur ou égal à n. Comme le rang de A est fini ($\leq N$) il existe un premier vecteur X_s de la suite de Krylov qui s'exprime en une combinaison linéaire des précédents. C'est-à-dire que :

$$X_s = \sum_{i=0}^{s-1} \alpha_i X_i$$

ou encore :

$$(A^s - \sum_{i=0}^{s-1} \alpha_i A^i)X_o = 0$$

Donc il existe un polynôme p de $\mathcal{P}_s^*$ tel que $p(A)X_o = 0$ et s est le plus petit degré pour lequel cette propriété est vérifiée.

. Ce polynôme est appelé polynôme annihilateur de X_o, ou polynôme minimal pour X_o (cf [12], [13]).

a) Propriétés immédiates :

On se pose la question de savoir sous quelles conditions sur l'espace $E_n(X_o)$ est invariant par A, puisque dans ce cas les éléments propres E_n-approchés sont des éléments propres exacts de A. Les propriét suivantes sont très faciles à démontrer [12] :

1°) On a $E_1 \subset E_2 \subset \ldots \subset E_n \subset \ldots \subset H$

et $E = \bigcup_{i=1}^{N} E_i$

est un sous-espace vectoriel invariant par A.

2°) Soit n un entier inférieur à N, ordre de la matrice A. Alors le polynôme annihilateur de X_o est de degré n si et seulement si E_n est de dimension n et est invariant par A.

3°) Plus généralement E_n est invariant par A si et seulement si : degré m du polynôme annihilateur est inférieur ou égal à n et la dimension de E_n est alors égale à m.

b) Approximation de Galerkin sur les espaces $E_n(X_o)$: la méthode de Rayleigh généralisée.

Nous appellerons π_n la projection orthogonale sur l'espace E_n La méthode de Galerkin appliquée à A avec les projections π_n est connue dans la littérature sous le nom de méhtode de Rayleigh généralisée.

D'après la propriété a) ci-dessus, du fait que $\bigcup_{i=1}^{N} E_i = E$, alors si dénote la projection orthogonale sur E, les projections π_n sont statit naires à partir de $n \leq N$ et sont égales à π, et la méthode de Galerkir approche alors les éléments propres de la partie de A dans E.

Comme E est un espace invariant par A, les éléments propres de la partie de A dans E sont aussi des éléments propres de A.

En général, si X_o est quelconque $E = H$, mais il peut arrive que $E \subset H$.
$\neq$

La méthode Rayleigh généralisée est connue sous des variant diverses : Méthode de Lanczos [4] , méthode d'Erdelyi [3] , méthode de moments [2], etc...

Afin de regrouper toutes les formes connues de cette méthode nous allons en donner une caractérisation à l'aide des propriétés des polynômes caractéristiques associés aux problèmes approchés.

DEFINITION :

On appelle polynôme caractéristique E_n-approché de A le poly

$$P_A^{(n)}(x) = \det(x - Y^T A Y)$$

où Y est une matrice dont les colonnes forment une base orth normale de $E_n(X_o)$.

D'après la propriété vue au paragraphe 2.2., $P_A^{(n)}$ est exact le polynôme caractéristique de la partie de A dans $E_n(X_o)$.

3.2. AUTRE FORMULATION DE LA METHODE DE RAYLEIGH GENERALISEE

Les espaces $E_n(X_o)$ ont des liens évidents avec les espaces de polynômes de degré inférieur ou égal à n-1, puisque :

$$E_n(X_o) = \{X \in H \; ; \; \exists \; p \in \mathcal{P}_{n-1} \; ; \; X = p(A)X_o\} \quad .$$

Il est donc naturel de chercher à interprêter l'approximatic de Galerkin sur ces sous-espaces comme un problème d'approximation dar espaces de polynômes, plus manipulables que les $E_n(X_o)$.

Cette deuxième formulation qui découle d'un théorème démontré par Vandergraft [1] nous permettra de faciliter la recherche de bornes d'erreurs (cf paragraphe 4).

On considère dans toute la suite le problème approché associé à l'espace E_n où n est fixé.

p est le polynôme annihilateur de X_o. Soit m son degré.

Deux cas peuvent se présenter selon le degré de p

1°) $m \leq n$

Alors d'après la troisième propriété du 3.1.a) E_n est invariant par A et dim $(E_n) = m$.

2°) $m > n$

Alors dim $(E_n) = n$ et E_n n'est pas invariant par A.

Nous allons étudier les deux cas ci-dessus successivement :

a) $m \leq n$

L'espace E_n est de dimension m , il est invariant par A.

Les éléments propres E_n-approchés, qui sont des éléments propres exacts de A sont caractérisés par la propriété suivante :

PROPRIETE :

Si $m \leq n$, alors le polynôme caractéristique E_n-approché est exactement le polynôme annihilateur de X_o, les valeurs propres E_n-approchées sont les racines de ce polynôme et les vecteurs propres correspondants sont les $p_i(A)X_o$ où

$$p_i(x) = \frac{p(x)}{x-\lambda_i}$$

DEMONSTRATION :

En effet $p(A)X_o = 0$ s'écrit :

$$(A-\lambda_1)(A-\lambda_2)(\ldots)(A-\lambda_m)X_o = 0$$

où chaque λ_i est une racine de p. λ_i est une valeur propre de A avec le vecteur propre associé: $\varphi_i = p_i(A)X_o$ (qui est non nul sinon le degré de p serait inférieur strictement à n).

D'autre part $p_i(A)X_o \in E_n(X_o)$ et donc l'égalité

$$A\varphi_i = \lambda_i \varphi_i$$

s'écrit aussi

$$\pi_n A \varphi_i = \lambda_i \varphi_i \qquad (\varphi_i \in E_n)$$

i.e. λ_i est donc aussi une valeur propre E_n-approchée de A.

C.Q.F.D.

D'après cette propriété si le degré de p est inférieur ou égal à n, on peut calculer les éléments propres E_n-approchés en recherchant le polynôme p qui vérifie $p(A)X_o = 0$ ou encore

$$\|p(A)X_o\| = 0 .$$

REMARQUE :

Puisque la dimension de E_n est m $(m \leq n)$, il n'y aura que m éléments propres E_n-approchés (caractérisés par la propriété).

b) $m > n$

Il existe alors pas de polynôme de $\mathcal{P}_n^*$ tel que $p(A)X_o = 0$

Soit $\bar{p}$ le polynôme appartenant à $\mathcal{P}_n^*$ qui réalise le minimum de $\|p(A)X_o\|$ dans $\mathcal{P}_n^*$.

Ce minimum existe, puisque :

$$\min_{\substack{\alpha_i \in \mathbb{R} \\ i=0,\ldots,n-1}} \|A^n X_o - \alpha_{n-1} A^{n-1} X_o - \ldots - \alpha_o X_o\|$$

est un problème de meilleure approximation de $A^n X_o$ dans le sous-espace

Il a été démontré en [1] que :

THEOREME 7 :

Le polynôme caractéristique E_n-approché est le polynôme $\bar{p}$ de qui réalise le minimum de $\|p(A)X_o\|$ dans $\mathcal{P}_n^*$.

Les valeurs propres E_n-approchées $\lambda_i^{(n)}$ sont donc les racines polynôme et de plus les vecteurs propres sont, à une constant multiplicative près les $\bar{p}_i(A)X_o$ où

$$\bar{p}_i(x) = \frac{\bar{p}(x)}{x - \lambda_i^{(n)}}$$

REMARQUE :

Ce résultat a été démontré par Vandergraft en 1971.
C'est Erdelyi [3] qui a eu, en 1965, l'idée de calculer des valeurs prc de matrices en recherchant le polynôme de $\mathcal{P}_n^*$ qui minimise $\|p(A)X_o\|$. Vandergraft a montré que la méthode de Rayleigh et la méthode d'Erdelyi sont identiques, i.e. , elles calculent les mêmes éléments propres appr Cette propriété importante n'est pas mise en évidence par des auteurs t que Lanczos [4], Householder [12], Kaniel [2], Golub [7], Ruhe [5].

Ceux-ci étudient la méthode de Rayleigh pour n = N mais observ cependant que pour n << $\mathbb{N}$ on peut obtenir de bonnes approximations des éléments propres correspondants aux valeurs propres extrémales.

Le cas $m \leq n$ est un cas particulier, le cas le plus "probable" du point de vue pratique, si X_o est choisi quelconque, est celui où m > D'autre part si au cours de l'algorithme n = m , i.e. si E_n est invar par A, il n'est plus nécessaire de poursuivre l'algorithme car les éléme propres E_n-approchés sont alors des éléments propres exacts de A. (On re mence en fait le procédé en choisissant un autre vecteur initial X_o' orth gonal à l'espace $E_n(X_o)$ pour approcher les autres éléments propres).

Ceci nous amène à délaisser la possibilité $m \leq n$.
Nous ferons pour toute la suite l'hypothèse suivante :

(H_n) Le degré du polynôme annihilateur est strictement supérieur à m > n .

Cela entraîne que si pour un polynôme q on a $q(A)X_o = 0$;

Alors : - soit $d^\circ q > n$

- soit $q \equiv 0$

Le polynôme caractéristique E_n-approché $p^{(n)}(x)$ est tel que $p^{(n)}(A)X_o$ est le meilleur approximant de 0 dans l'ensemble des vecteurs de la forme $p(A)X_o$, où $p \in \mathcal{P}_n^*$.

Mais $p^{(n)}(x)$ peut-il être caractérisé comme le meilleur approxi de 0 dans $\mathcal{P}_n^*$ au sens d'une certaine norme ?

On peut répondre à cette question à l'aide du lemme suivant :

LEMME 8 :

Sous l'hypothèse (H_n), l'application $p \to \|p(A)X_o\|$ est une nor sur l'espace vectoriel des polynômes de degré inférieur ou égal à n .

NOTATION :

Nous noterons $\|.\|_{X_o}$ cette norme.

DEMONSTRATION :

.) L'inégalité triangulaire et la relation $\|\alpha p\|_{X_o} = |\alpha| \|p\|_{X_o}$ se déduisent aisément de la linéarité de l'application $p \to p(A)X_o$.

.) Supposons que $\|p\|_{X_o} = 0$, i.e. que $\|p(A)X_o\| = 0$, cela entraîne, d'après l'hypothèse (H_n), que :

soit $p \equiv 0$, soit $d^o p > n$. Si on suppose que $p \in \mathcal{P}_n$ le degré de p est $\leq n$ et p est donc nul.

C.Q.F.D.

Le théorème 7 s'énonce donc ainsi :

COROLLAIRE 8 :

Sous l'hypothèse (H_n), le polynôme caractéristique E_n-approché est l'élément $\bar{p}$ de $\mathcal{P}_n^*$ qui est le meilleur approximant du polynôme identiquement nul, au sens de la norme $\| \|_{X_o}$.

Les valeurs propres $\lambda_i^{(n)}$ E_n-approchées sont donc les racines de ce polynôme et les vecteurs propres sont les $\bar{p}_i(A)X_o$ où

$$\bar{p}_i(x) = \frac{\bar{p}(x)}{x-\lambda_i}$$

Si on connait $\bar{p}(x)$; le meilleur approximant dans $\mathcal{P}_n^*$ de 0 , au sens de la norme $\|.\|_{X_o}$, on sait calculer les valeurs propres et les vecteurs propres E_n-approchés.

Le corollaire 8 est donc une caractérisation de la méthode de Rayleigh généralisée.

3.3. UNE DES VARIANTES DE LA METHODE DE RAYLEIGH GENERALISEE :

L'ALGORITHME DE LANCZOS.

Une des formes les plus connues et les plus intéressantes de la méhtode de Rayleigh généralisée est l'algorithme de Lanczos.

Cet algorithme consistait à l'origine à mettre A sous la forme tridiagonale :

$$V^T AV = \left(\begin{array}{cc|c} \alpha_1 & \beta_2 & \\ \beta_2 & \ddots & \beta_{n+1} \\ & \beta_n & \alpha_n \\ \hline & \beta_{n+1} & \alpha_{n+1} \quad \ddots \\ & & \ddots \quad \beta_N \\ & & \beta_N \quad \alpha_N \end{array}\right)$$

On a constaté (Lanczos lui-même [4]) que l'on obtient de bonnes approximations des valeurs propres dominantes de A si l'on arrête le processus au $n^{ième}$ pas et que l'on prend comme valeurs propres approchées les valeurs propres de

$$A_n = \begin{pmatrix} \alpha_1 & \beta_2 & & \\ \beta_2 & \ddots & \ddots & \\ & \ddots & \ddots & \beta_n \\ & & \beta_n & \alpha_n \end{pmatrix}$$

Ceci est dû au fait que A_n n'est autre que la partie de A dans $E_n(X_o)$ écrite dans une base orthonormale $(v_1, v_2, \ldots, v_n)$ de $E_n(X_o)$.

L'algorithme pour la construction de A_n est le suivant :

. $v_1 = X_o/\|X_o\|$, $\alpha_1 = \mu(v_1) = (Av_1,v_1)$, $\beta_1 = 0$, $v_o =$

. de j = 1 , à n - 1 faire :

$$w_{j+1} = Av_j - \alpha_j v_j - \beta_j v_{j-1} \qquad (1)$$

$$v_{j+1} = \frac{w_{j+1}}{\|w_{j+1}\|} \qquad (2)$$

$$\beta_{j+1} = \|w_{j+1}\| \qquad (3)$$

$$\alpha_{j+1} = \mu(v_{j+1})$$

On peut alors démontrer que les (v_i) $i=1,2,\ldots,n$ forment un base orthonormale de $E_n(X_o)$ et qu'en raison des relations (1) et (2), on pose $V_n = (v_1,v_2,\ldots,v_n)$ on a :

$$V_n^T A V_n = A_n .$$

Ceci est bien l'écriture du problème approché par la méthode Galerkin (cf paragraphe 2.2.).

Cette écriture de la méthode de Rayleigh généralisée permet d'obtenir des renseignements précieux sur la façon dont les convergent ;

ceci grâce aux propriétés bien connues des matrices tridiagonales.

En effet, sous l'hypothèse (H_n), l'algorithme de Lanczos est constructible car les β_i sont non nuls. $i=2,\ldots,n$ (sinon d'après (3) (2) il existerait un polynôme q de degré $\leq n$ tel que $q(A)X_o = 0$).

Si on ordonne les valeurs propres en décroissant

on sait alors (Wilkinson [9] p. 300) que les valeurs propres vérifient la propriété d'alternance stricte :

$$\lambda_{i+1}^{(n-1)} < \lambda_{i+1}^{(n)} < \lambda_i^{(n-1)} < \lambda_i^{(n)} < \lambda_i$$

On en déduit que sous l'hypothèse (H_n) :

1°) Le problème approché associé à l'espace $E_n(X_o)$, n'admet aucu valeur propre multiple.

2°) Tant que l'hypothèse (H_n) est vérifiée, $\lambda_i^{(n)}$ est strictement croissante en fonction de n . Dès que $\lambda_i^{(n)} = \lambda_i^{(n-1)}$ alors l'hypothèse (H_n) n'est plus vérifiée et donc l'espace E_n est invariant par A.

On veut maintenant répondre à la question suivante :

L'espace E_n n'est pas invariant par A puisque le degré du polynôme annilateur est supérieur strictement à n.
Sous quelle condition E_n est-il ε-invariant par A ?

Cette question est justifiée par le fait que beaucoup d'auteu prennent E_n comme espace invariant approché.

En réalité :

PROPOSITION 9 :

Sous l'hypothèse (H_n) , on pose $\delta_i = \min_{p \in \mathcal{S}_i^*} \|p(A)X_o\|$ $i=n-1,n$

Alors E_n est ε-invariant par A si et seulement si :

$$\frac{\delta_n}{\delta_{n-1}} = \varepsilon$$

Pour que E_n soit ε-invariant par A il faut non seulement que $\delta_n = \min_{p \in \mathcal{S}_n^*} \|p(A)X_o\|$ soit petit mais que δ_{n-1} ne soit pas du même ordre.

Cela montre que même si $\min_{\mathcal{S}_n^*} \|p(A)X_o\|$ est très petit nous n'a pas forcément de bonnes approximations de tous les éléments propres E_n-approchés (cf. proposition 5).

DEMONSTRATION DE LA PROPOSITION 9 :

Elle découle du lemme suivant :

LEMME :

Les vecteurs v_i construits par l'algorithme de Lanczos sont les $\frac{p_i(A)X_o}{\|p_i(A)X_o\|}$ successifs où $p_i \in \mathcal{S}_i^*$ vérifie

$$\min_{p \in \mathcal{S}_i^*} \|p(A)X_o\| = \|p_i(A)X_o\| \quad .$$

De plus ce minimum est égal à $\prod_{j=2}^{i+1} \beta_j$.

DEMONSTRATION DU LEMME :

Soit $\tilde{v}_i = (\prod_{j=2}^{i+1} \beta_j) v_i$ on va montrer par réccurrence que

$$\tilde{v}_i = p_i(A)X_o$$

où $p_i(A)$ est le polynôme caractéristique de

$$A_i = \begin{pmatrix} \alpha_1 & \beta_2 & & \\ \beta_2 & & & \beta_i \\ & & & \\ & \beta_i & & \alpha_i \end{pmatrix}$$

Cela sera suffisant d'après le théorème 7.

. Le résultat est vrai pour i=1. i.e. si $\tilde{v}_2 = \beta_2 v_1$

$$\beta_2 v_2 = W_2 = Av_1 - \mu(v_1)v_1 = (A-\alpha_2)v_1$$

Or :

$$p_1(A) = A-\alpha_1$$

. Si le résultat est vrai pour i démontrons-le pour i+1.

D'après l'algorithme $\beta_{i+1}v_{i+1} = (A-\alpha_i)v_i-\beta_i v_{i-1}$ multiplions les deux membres par $\prod_{j=2}^{i}\beta_j$ alors :

$$\prod_{j=2}^{i+1}\beta_j v_{i+1} = (A-\alpha_i)(\prod_{j=2}^{i}\beta_j)v_i - (\prod_{j=2}^{i}\beta_j)\beta_i v_{i-1}$$

ou encore :

$$\tilde{v}_{i+1} = (A-\alpha_i)\tilde{v}_i - \beta_i^2\tilde{v}_i$$

d'après la récurrence

$$\tilde{v}_{i+1} = (A-\alpha_i)p_i(A)X_o - \beta_i^2 p_{i-1}(A)X_o$$

$$= ((A-\alpha_i)p_i(A) - \beta_i^2 p_{i-1}(A))X_o$$

d'où :

$$\tilde{v}_{i+1} = p_{i+1}(A)X_o$$

d'après les propriétés de polynôme caractéristiques de matrices tridiagonales.

Pour finir la démonstration du lemme il suffit d'ajouter que

$$\|\tilde{v}_i\| = \prod_{j=2}^{i+1}\beta_j \qquad \text{car} \qquad \|v_i\| = 1 .$$

DEMONSTRATION DE LA PROPOSITION :

$(1-\pi_n)A\pi_n$ est représenté par la matrice $\beta_{n+1}e_n e_{n+1}^T$ dont la norme $\|(1-\pi_n)A\pi_n\|_2$ est β_{n+1}

$$\beta_{n+1} = \frac{\|\tilde{v}_n\|}{\|\tilde{v}_{n-1}\|} = \frac{\delta_n}{\delta_{n-1}}$$

C.Q.F.D.

3.4. BORNES D'ERREURS SUR LES ELEMENTS PROPRES E_n-APPROCHES .

D'après la remarque qui suit la proposition 9 l'application des résultats du paragraphe 1 aboutirait à des majorations trop larges

$$V_A(E_n) = \|(1-\pi_n)A\pi_n\|_2$$

n'est pas en général "petit".

Nous allons donc plutôt chercher des résultats du même type que ceux du paragraphe 2.1 concernant les bornes d'erreurs asymptotique Ces résultats sont plus appropriés puisque l'on sait que la méthode de Rayleigh généralisée donne lieu, en pratique, à de bonnes approximatior sur les éléments propres correspondant aux valeurs propres dominantes. Nous ne parlerons que de l'approximation des plus grandes valeurs propr de A, les résultats étant identiques pour les plus petites.

On démontre (cf. Kaniel [2], Faddev , Faddeeva [19],...) que $\lambda_1^{(n)}$ est la plus grande valeur propre E_n-approchée et si on pose

$$\gamma_1 = 1 + 2\frac{\lambda_1-\lambda_2}{\lambda_2-\lambda_N}$$

et θ_1 = l'angle aigü formé par X_o et φ_1 , alors :

$$\lambda_1-\lambda_1^{(n)} \leq (\lambda_1-\lambda_N)\left(\frac{\operatorname{tg}\theta_1}{T_{n-1}(\gamma_1)}\right)^2$$

où $T_{n-1}(x)$ est le polynôme de Tchebichev de première espèce.

Kaniel [2] a tenté de généraliser l'inégalité ci-dessus aux autres valeurs propres, cependant sa formule contient les nombres $\|\varphi_k-\varphi_k^{(n)}\|$ dont il n'a pas donné une majoration. La formule proposée dans [2] est la suivante :

$$\lambda_i-\lambda_i^{(n)} \leq \frac{K_i}{[T_{n-i}(\gamma_i)]^2} + \sum_{k=1}^{i-1} \lambda_k \|\varphi_k-\varphi_k^{(n)}\|^2$$

où K_i et γ_i sont des constantes dépendant des valeurs propres exactes de A seulement.

Nous allons procéder de manière différente grâce à la remarq simple suivante :

D'après les propriétés (2) et (3) du paragraphe 1, il ressor que si l'on dispose d'une majoration de $\|(\varphi_i^{(n)})\|$ on peut en déduire d'u part une borne d'erreur sur le vecteur propre (propriété 3) et d'autre part une borne d'erreur sur la valeur propre (propriété 2).

Il faut commencer par majorer le nombre $\|(1-\pi_n)\varphi_i\|$ qui inter dans les bornes asymptotiques.

a) Majoration de l'angle entre l'espace E_n et le vecteur φ_i :

Nous supposerons comme en [2] que les valeurs propres de A s simples. Toutefois les résultats s'étendent sans problème au cas des valeurs propres multiples mais les notations se compliquent légèrement

- Nous appellerons θ_i l'angle aigü formé par X_o et le vecteur propre φ_i associé à la valeur propres λ_i

- Les valeurs propres sont supposées ordonnées en décroissant

  $\lambda_1 \geq \lambda_2 \geq \lambda_3 \cdots \geq \lambda_N$

  Pour tout $X \neq 0$ $\theta(X,E_n)$ désigne l'angle entre le vecteur X et l'espace $E_n(X_o)$, c'est-à-dire :

$$\theta(X,E_n) = \text{Arcsin}\left(\min_{Y\in E_n} \frac{\|X-Y\|}{\|X\|}\right) \qquad \text{(Ruhe [5])}$$

$$= \text{Arcsin}\,\frac{\|(1-\pi_n)X\|}{\|X\|}$$

- Les vecteurs propres exactes φ_i et E_n-approchés $\varphi_i^{(n)}$ sont de norme 1.

- On suppose que le vecteur initial X_o est normé et qu'il s'éc dans la base orthonormale des vecteurs propres de A :

$$X_o = \sum_{i=1}^{N} \alpha_i \varphi_i$$

En raison des notations nous avons les relations :

$$\cos \theta_i = |\alpha_i|$$

$$\sin \theta_i = \Big(\sum_{j \neq 1} \alpha_j^2 \Big)^{1/2}$$

$$\operatorname{tg}(\theta_i) = \Big(\sum_{j \neq i} \big(\frac{\alpha_j}{\alpha_i}\big)^2 \Big)^{1/2}$$

La proposition suivante donne l'angle aigü entre φ_i et l'espace E_n

PROPOSITION 10 :

On se place dans l'hypothèse (H_n) .

X_o est normé et s'écrit $X_o = \sum_{k=1}^{N} \alpha_k \varphi_k$ dans la base des vecteurs propres de A.

On suppose que φ_i n'est pas orthogonal à X_o $(\alpha_i \neq 0)$.

On pose :

$$\beta_j = \frac{|\alpha_j|}{\Big(\sum_{k \neq i} (\alpha_k)^2 \Big)^{1/2}} \qquad \text{pour } j=1,2,\ldots,N$$

et :

$$t_{i,n} = \inf_{\substack{p \in \mathcal{P}_{n-1} \\ p(\lambda_i)=1}} \Big(\sum_{j \neq i} (\beta_j p(\lambda_j))^2 \Big)^{1/2}$$

Alors :

$$\operatorname{tg} \theta(\varphi_i, E_n) = t_{i,n} \operatorname{tg} \theta_i$$

DEMONSTRATION :

Un vecteur X quelconque de E_n s'écrit :

$$X = p(A)X_o = \sum_{j=1}^{N} p(\lambda_j)\alpha_j\varphi_j = p(\lambda_i)\alpha_i\varphi_i + \sum_{j\neq i} p(\lambda_j)\alpha_j\varphi_j \qquad (1)$$

où p est un polynôme de $\mathcal{P}_{n-1}$.

Appelons γ l'angle entre X et φ_i . Ce que l'on cherche c'est le vecteur X tel que l'angle γ est soit minimum : le produit scalaire entre X et φ_i est d'après (1) :

$$p(\lambda_i)\alpha_i$$

Comme α_i est non nul l'angle entre X et φ_i est $\pi/2$ lorsque $p(\lambda_i) = 0$ et cet angle n'est pas minimum.

On peut donc se restreindre à considérer les $X \in E_n$ tels que $X = p(A)X_o$ et $p(\lambda_i) \neq 0$ alors nous avons :

$$tg^2\gamma = \sum_{j\neq i} \left(\frac{p(\lambda_j)\alpha_j}{p(\lambda_i)\alpha_i}\right)^2 = \sum_{j\neq i} \left(\frac{\alpha_j}{\alpha_i}\right)\left(\frac{p(\lambda_j)}{p(\lambda_i)}\right)^2 \qquad (2)$$

si

$$\beta_j^2 = \frac{\alpha_j^2}{\sum_{k\neq i} \alpha_k^2}$$

on aura :

$$\left(\frac{\alpha_j}{\alpha_i}\right)^2 = \beta_j^2 \sum_{k\neq i} \left(\frac{\alpha_k}{\alpha_i}\right)^2 = \beta_j^2 \, tg^2\,\theta_i \qquad (3)$$

d'où d'après (2) et (3) :

$$tg^2\gamma = \sum_{j\neq i} \beta_j^2 \left(\frac{p(\lambda_j)}{p(\lambda_i)}\right)^2 tg^2\,\theta_i \qquad (4)$$

l'angle est donc minimum lorsque :

$$\left(\sum_{j \neq i} \beta_j^2 \left(\frac{p(\lambda_j)}{p(\lambda_i)} \right)^2 \right)^{1/2}$$

est minimum sur l'ensemble des polynômes de degré inférieur ou égal à n-1 tels que $p(\lambda_i) \neq 0$, ou encore lorsque $\left(\sum_{j \neq i} \beta_j^2 (p(\lambda_j))^2 \right)^{1/2}$ est minimum sur l'ensemble des polynômes de $\mathcal{P}_{n-1}$ tels que $p(\lambda_i) = 1$, et la relation donne alors la valeur de la tg de cet angle minimum.

C.Q.F.D.

Majoration des $t_{i,n}$:

La proposition 10 ramène le problème de la majoration de $\theta(\varphi_i, E_n)$ à celui de la majoration de $t_{i,n}$.

Pour faciliter la résolution de ce problème de bornes, il faut remarquer que l'applation $p \to \left(\sum_{j \neq i} (\beta_j p(\lambda_j))^2 \right)^{1/2}$ est une norme sur l'espace $\mathcal{P}_{n-1}$ si l'hypothèse (H_n) est vérifiée .

En effet, si on pose

$$X_i' = \sum_{j \neq i} \beta_j \varphi_j$$

On constate que :

$$\left(\sum_{j \neq i} \beta_j^2 (p(\lambda_j))^2 \right)^{1/2} = \| p(A) X_i' \|$$

$$\| X_i' \| = 1 \quad \left(\sum_{j \neq i} \beta_j^2 = \sum_{j \neq i} \frac{\alpha_j^2}{\sum_{k \neq i} \alpha_k^2} = 1 \right)$$

D'autre part il est facile de voir que l'hypothèse (H_n) pour X_o entraine l'hypothèse (H_{n-1}) pour X_i' .

Tout cela permet de conclure, grâce au lemme 8, que :

L'application $p \to \|p(A)X_i'\| = \|p\|_{X_i'}$ est une norme sur $\mathcal{P}_{n-1}$.

On va majorer cette norme par la norme $\max_{j\neq i} |p(\lambda_j)|$:

PROPRIETE :

$$\|p\|_{X_i'} \leq \max_{j\neq i} |p(\lambda_j)|$$

En effet :

$$\|p\|^2_{X_i'} = \sum_{j\neq i} \beta_j^2 (p(\lambda_j))^2 \leq \sum_{j\neq i} \beta_j^2 \left(\max_{j\neq i}(p(\lambda_j))^2\right)$$

$$= \max_{j\neq i} (p(\lambda_j))^2 \sum_{j\neq i} \beta_j^2 = \max_{j\neq i} (p(\lambda_j))^2$$

puisque :

$$\sum_{j\neq i} \beta_j^2 = \|X_i'\|^2 = 1 .$$

REMARQUE :

Sous l'hypothèse (H_n) $\max_{j\neq i} |p(\lambda_j)|$ est une norme sur $\mathcal{P}_{n-1}$.

On va appliquer cette propriété simple pour majorer $t_{1,n}$ et retrouver l'inégalité classique sur $\lambda_1 - \lambda_1^{(n)}$.

PROPRIETE 11 :

$$\text{Soit } \gamma_1 = 1 + 2\,\frac{\lambda_1 - \lambda_2}{\lambda_1 - \lambda_N}$$

alors :

$$t_{1,n} \leq \frac{1}{T_{n-1}(\gamma_1)}$$

où $T_{n-1}(x)$ désigne le polynôme de Tchebichev de degré n-1 c'est-à-dire :

$$T_{n-1}(\gamma_1) = \mathrm{ch}((n-1)\mathrm{Argch}\ \gamma_1)$$

DEMONSTRAITON :

D'après la propriété précédente :

$$t_{1,n} = \min_{\substack{p \in \mathcal{P}_{n-1} \\ p(\lambda_1)=1}} \|p\|_{X_1'} \leq \min_{\substack{p \in \mathcal{P}_{n-1} \\ p(\lambda_1)=1}} \max_{j \neq 1} |p(\lambda_j)|$$

$\max_{j\neq 1} |p(\lambda_j)|$ est encore majoré par la norme $\max_{\lambda_N \leq x \leq \lambda_2} |p(x)|$.

On sait (cf P.J. Laurent [20]) que parmi les polynômes de $\mathcal{P}_{n-1}$ qui vérifient $p(a) = 1$ pour un point a tel que $|a| > 1$, c'est

$\dfrac{T_{n-1}(x)}{T_{n-1}(a)}$ qui rend minimum la norme uniforme sur l'intervalle $[-1,+1]$.

En faisant le changement de variable

$$x = \frac{\lambda_2 - \lambda_N}{2}\, y + \frac{\lambda_2 + \lambda_N}{2} \qquad \text{avec} \qquad -1 \leq y \leq 1$$

on voit que $\lambda_N \leq x \leq \lambda_2$ et le problème :

$$\min_{\substack{p\in\mathcal{P}_{n-1}\\ p(\lambda_1)=1}} \max_{N\leq x\leq\lambda_2} |p(x)|$$

devient :

$$\min_{\substack{p\in\mathcal{P}_{n-1}\\ p(\gamma_1)=1}} \max_{-1\leq y\leq 1} |p(y)| = \max_{-1\leq y\leq 1} \frac{T_{n-1}(y)}{T_{n-1}(\gamma_1)} = \frac{1}{T_{n-1}(\gamma_1)}$$

C.Q.F.D.

REMARQUES :

1) $T_{n-1}(x) = \cos((n-1)\text{Arcos } x)$ si $-1 \leq x \leq 1$

et $T_{n-1}(x) = \text{ch}((n-1)\text{Argch}(x))$ si $|x| > 1$

2) Le nombre γ_1 s'exprime en fonctions du conditionnement τ_1 de $[A-\lambda_1]_{[\varphi_1]^\perp}$:

$$\gamma_1 = \frac{\lambda_1-\lambda_N}{\lambda_1-\lambda_2} \qquad \tau_1 = \frac{\tau_1+1}{\tau_1-1}$$

On peut alors retrouver la majoration connue de $\lambda_1-\lambda_1^{(n)}$.

PROPRIETE :

$$\lambda_1-\lambda_1^{(n)} \leq (\lambda_1-\lambda_N) \frac{tg^2\theta_1}{(T_{n-1}(\gamma_1))^2}$$

DEMONSTRATION :

Il suffit de démontrer que :

$$\lambda_1 - \lambda_1^{(n)} \leq (\lambda_1 - \lambda_N)tg\theta(\varphi_1, E_n)$$

et le résultat suivrait en raison de la proposition 10 et de la majoration précédente de $t_{1,n}$.

On a :

$$\lambda_1^{(n)} = \max_{X \in E_n} \frac{(AX,X)}{\|X\|^2}$$

d'où :

$$\lambda_1 - \lambda_1^{(n)} - \frac{((\lambda_1 - A)\pi_n\varphi_1, \pi_n\varphi_1)}{\|\pi_n\varphi_1\|^2} = \frac{((\lambda_1 - A)(\varphi_1 - (1-\pi_n)\varphi_1), (\varphi_1 - (1-\pi_n)\varphi_1))}{\|\pi_n\varphi_1\|^2}$$

$$= \frac{((\lambda_1 - A)(1-\pi_n)\varphi_1, (1-\pi_n)\varphi_1)}{\|\pi_n\varphi_1\|^2} \leq \rho(\lambda_1 - A)\frac{\|(1-\pi_n)\varphi_1\|^2}{\|\pi_n\varphi_1\|^2} \quad .$$

$\rho(\lambda_1 - A)$ le rayon spectral de $\lambda_1 - A$ est égal à $\lambda_1 - \lambda_N$ et

$\frac{\|(1-\pi_n)\varphi_1\|}{\|\pi_n\varphi_1\|}$ n'est autre que $\operatorname{tg}\theta(\varphi_1, E_n)$.

C.Q.F.D.

REMARQUE :

La démonstration ci-dessus de l'inégalité sur $\lambda_1 - \lambda_1^{(n)}$ est plus simple que celle proposée par Kaniel en [2].

Majoration de $t_{i,n}$:

où i est quelconque $1 \leq i \leq n$.

PROPOSITION 12 :

Supposons que les valeurs propres $\lambda_1,\lambda_2,\ldots,\lambda_n$ de A soient simples et posons :

$$\gamma_i = 1 + 2\,\frac{\lambda_i-\lambda_{i-1}}{\lambda_i-\lambda_N} \quad ; \quad \begin{cases} K_i = \prod_{j=1}^{i-1} \dfrac{(\lambda_j-\lambda_N)}{(\lambda_j-\lambda_i)} & \text{si} \quad i\neq 1 \\ K_1 = 1 \end{cases}$$

Alors :

$$t_{i,n} \leq \frac{K_i}{T_{n-i}(\gamma_i)}$$

et

$$\operatorname{tg}\mathcal{L}(\varphi_i,E_n) \leq \frac{K_i \operatorname{tg}\theta_i}{T_{n-i}(\gamma_i)}$$

DEMONSTRATION :

Démontrons le résultat pour i=2.

Pour i quelconque la généralisation est immédiate :

$$t_{2,n} = \min_{\substack{p\in\mathcal{P}_{n-1}\\ p(\lambda_2)\neq 0}} \max_{j\neq 2} \left| \frac{p(\lambda_j)}{p(\lambda_2)} \right|$$

$$\leq \min_{\substack{p\in\mathcal{P}_{n-1}\\ p(\lambda_2)\neq 0\\ p(\lambda_1)=0}} \max_{j\neq 2} \left| \frac{p(\lambda_j)}{p(\lambda_2)} \right|$$

Or tout polynôme p de $\mathcal{P}_n$ tel que $p(\lambda_1) = 0$ s'écrit :

$$(\lambda_1 - x)p(x)$$

où $p \in \mathcal{P}_{n-2}$

d'où :

$$t_{2,n} \leq \min_{\substack{\tilde{p} \in \mathcal{P}_{n-2} \\ \tilde{p}(\lambda_2) \neq 0}} \max_{j \neq 2} \frac{|(\lambda_1 - \lambda_j)\tilde{p}(\lambda_j)|}{|(\lambda_1 - \lambda_2)\tilde{p}(\lambda_2)|}$$

$$= \frac{1}{\lambda_1 - \lambda_2} \min_{\substack{\tilde{p} \in \mathcal{P}_{n-2} \\ \tilde{p}(\lambda_2) \neq 0}} \max_{j \neq 2} |(\lambda_1 - \lambda_j)\tilde{p}(\lambda_j)|$$

$$\leq \frac{\lambda_1 - \lambda_N}{\lambda_1 - \lambda_2} \min_{\substack{\tilde{p} \in \mathcal{P}_{n-2} \\ \tilde{p}(\lambda_2) \neq 0}} \max_{j \neq 2} \frac{|\tilde{p}(\lambda_j)|}{|\tilde{p}(\lambda_2)|}$$

$\frac{\lambda_1 - \lambda_N}{\lambda_1 - \lambda_2} = K_2$. En poursuivant la démonstration de manière analogue à la propriété 11, on obtient :

$$\min_{\substack{\tilde{p} \in \mathcal{P}_{n-2} \\ \tilde{p}(\lambda_2) \neq 0}} \max_{j \neq 1} \frac{|\tilde{p}(\lambda_j)|}{|\tilde{p}(\lambda_2)|} \leq \min_{\substack{\tilde{p} \in \mathcal{P}_{n-2} \\ \tilde{p}(\lambda_2) \neq 0}} \max_{\lambda_N \leq x \leq \lambda_3} \frac{|\tilde{p}(x)|}{|\tilde{p}(\lambda_2)|} = \frac{1}{T_{n-2}(\gamma_2)}$$

C.Q.F.D.

b) Bornes d'erreurs sur les éléments propres :

Nous pouvons donner les bornes d'erreurs obtenues au paragraphe 2.2. sur les éléments propres.

Nous disposons d'une part, d'une majoration de $\|(1-\pi_n)\varphi\|$ et d'autre part des résultats de la proposition 6' :

$$\|(1-\pi_n)\varphi\| = \sin\theta(\varphi,E_n) = \frac{\operatorname{tg}\theta(\varphi,E_n)}{(1+\operatorname{tg}^2\theta(\varphi,E_n))^{1/2}} \leq \frac{K_i \operatorname{tg}\theta_i}{T_{n-i}(\gamma_i)}$$

D'où d'après la proposition 6', si $\|(1-\pi_n)A\| \leq \varepsilon$,

pour toute valeur propre E_n-approchée , il existe un vecteur propre φ de A associé à la valeur propre λ tel que si on pose

$$d = \min |\lambda_j - \lambda^{(n)}| \quad , \quad d' = d / (1+\varepsilon^2/d^2)$$

alors :

$$1°/ \quad \frac{1}{|\lambda|}\|r(\varphi_n)\| \leq \|(1-\pi_n)\varphi\| + \frac{\varepsilon^2}{|\lambda| d'}$$

$$2°/ \quad \left|\frac{\lambda-\lambda^{(n)}}{\lambda}\right| \leq \frac{1}{1-\frac{\varepsilon}{d'}}\left(\|(1-\pi_n)\varphi\|^2 + \frac{\varepsilon^2}{\lambda^2 d'}\right)$$

Ces deux inégalités s'écrivent aussi :

$$\frac{1}{|\lambda|}\|r(\varphi_n)\| \leq \alpha(\varepsilon)\|(1-\pi_n)\varphi\|$$

$$\left|\frac{\lambda-\lambda^{(n)}}{\lambda}\right| \leq \alpha'(\varepsilon)\|(1-\pi_n)\varphi\|^2$$

où $\alpha(\varepsilon)$ et $\alpha'(\varepsilon)$ sont des constantes d'autant plus voisines de 1 que ε est petit.

Les bornes d'erreurs ci-dessus ont l'inconvénient de ne pas être toujours réalistes : en effet $V_T(E_n)$ n'est pas toujours "petit" d'après la remarque qui suit la proposition 9. Nous devons donc compléter les bornes ci-dessus en des bornes qui soient spécifiques de la méthode.

D'après les propriétés 2 et 3 du paragraphe 1, il ressort que des bornes de $\|r(\varphi_i^{(n)})\|$ fourniront à la fois des bornes d'erreurs sur $\lambda_i - \lambda_i^{(n)} = \lambda_i - \mu(\varphi_i^{(n)})$ et sur l'angle entre φ_i et $\varphi_i^{(n)}$.

PROPOSITION 13 :

Supposons que la valeur propre de A, la plus voisine de $\lambda_i^{(n)}$ soit λ_i et qu'elle vérifie $\lambda_i < \lambda_{i-1}^{(n)}$, $i \neq 1$.

Supposons que, de plus, $\pi_n \varphi_i \neq \pi_{n-1} \varphi_i$.

Posons :

$$D_i = \max_j |\lambda_i - \lambda_j| \quad , \quad d_i = \min_{\lambda_j \neq \lambda_i} |\lambda_i^{(n)} - \lambda_j|$$

$$\varepsilon_{i,n} = \frac{K_i \operatorname{tg} \theta_i}{T_{n-i}(\gamma_i)}$$

où les notations sont les mêmes que pour la proposition 12.

Alors :

$$1°/ \quad \|r(\varphi_i^{(n)})\| \leq D_i \, \varepsilon_{i,n}$$

$$2°/ \quad \lambda_i - \lambda_i^{(n)} \leq \frac{1}{d_i} [D_i \varepsilon_{i,n}]^2$$

DEMONSTRATION :

Démontrons d'abord le lemme suivant :

LEMME 14 :

Soit I un intervalle fermé, contenant $\lambda_i^{(n)}$ et aucune autre valeur propre E_n-approchée.
Alors :

$$\min_{\substack{\mu\in I \\ Y\in E_n\setminus E_{n-1}}} \frac{\|(A-\mu)Y\|}{\|Y\|} = \|(A-\lambda_i^{(n)})\varphi_i^{(n)}\|$$

On a :

$$\varphi_i^{(n)} = \frac{\bar{p}_i(A)X_o}{\|\bar{p}_i(A)X_o\|}$$

avec, lorsque $\bar{p}(x)$ est le polynôme caractéristique E_n-approché,

$$\bar{p}_i(x) = \frac{\bar{p}(x)}{x-\lambda_i^{(n)}} \quad .$$

De plus, $\bar{p} \in \mathcal{P}_n^*$ vérifie d'après le théorème 7 :

$$\|\bar{p}(A)X_o\| = \min_{p\in\mathcal{P}_n^*} \|p(A)X_o\|$$

On a :

$$\bar{p}(x) = (x-\lambda_i^{(n)})\bar{p}_i(x)$$

et de plus :

$$\mathcal{P}_n^* = \{p \; ; \; p(x) = (x-\mu)\,q(x) \; ; \; \mu \in \mathbb{R} \, , \, q \in \mathcal{P}_{n-1}^*\} \; .$$

D'où :

$$\|(A-\lambda_i^{(n)})\bar{p}_i(A)X_o\| = \min_{\substack{q\in\mathcal{P}_{n-1}^* \\ \mu\in\mathbb{R}}} \|(A-\mu)q(A)X_o\| = \min_{\substack{q\in\mathcal{P}_{n-1}^* \\ \mu\in I}} \|(A-\mu)q(A)X_o\|$$

Cela entraine que le minimum de $\left\|(A-\mu)\dfrac{q(A)X_o}{\|\bar{p}_i(A)X_o\|}\right\|$ est atteint pour le vecteur $\bar{Y} = \dfrac{\bar{p}_i(A)X_o}{\|\bar{p}_i(A)X_o\|}$ de norme 1. D'où :

$$\|(A-\lambda_i^{(n)})\varphi_i^{(n)}\| = \min_{\mu\in I}\|(A-\mu)\| = \min_{\mu\in I}\|(A-\mu)Y\| = \min_{\mu\in I}\|(A-\mu)Y\|$$

$$\left\{ Y = \frac{q(A)}{\|\bar{p}_i(A)X_o\|} \right. \qquad \left\{ Y = \frac{q(A)}{\|\bar{p}_i(A)X_o\|},\ \|Y\|=1 \right. \qquad \left\{ Y = \frac{q(A)X_o}{\|q(A)X_o\|} \right.$$

$$\text{i.e. } \|(A-\lambda_i^{(n)})\varphi_i^{(n)}\| = \min_{\substack{\mu\in I \\ y\in E_n\setminus E_{n-1} \\ \|Y\|=1}}\|(A-\mu)Y\|$$

et le lemme est démontré.

DEMONTRONS LA PROPOSTION :

La deuxième inégalité est une conséquence de la première et de la propriété 2 du paragraphe 1.
Démontrons la première inégalité.

Soit I un intervalle contenant λ_i et $\lambda_i^{(n)}$ et autre valeur propre E_n-approchée de A. I existe en raison de l'hypothèse $\lambda_i < \lambda_{i-1}^{(n)}$ qui entraine $\lambda_{i+1}^{(n)} < \lambda_i^{(n)} \leq \lambda_i < \lambda_{i-1}^{(n)}$. Alors puisque $\pi_n\varphi_i \neq \pi_{n-1}\varphi_i$, $\pi_n\varphi_i \in E_n \setminus E_{n-1}$ et le lemme montre que :

$$\|(A-\lambda_i^{(n)})\varphi_i^{(n)}\| \leq \frac{\|(A-\lambda_i)\pi_n\varphi_i\|}{\|\pi_n\varphi_i\|} = \frac{\|(A-\lambda_i)\varphi_i - (A-\lambda_i)(1-\pi_n)\varphi_i\|}{\|\pi_n\varphi_i\|}$$

$$= \frac{\|(A-\lambda_i)(1-\pi_n)\varphi_i\|}{\|\pi_n\varphi_i\|} \leq D_i\frac{\|(1-\pi_n)\varphi_i\|}{\|\pi_n\varphi_i\|} = D_i \operatorname{tg}\theta(\varphi_i,E_n)$$

et la proposition 12 montre que $\operatorname{tg}\theta(\varphi_i,E_n) \leq \varepsilon_{i,n}$.

C.Q.F.D.

Il peut arriver (cf paragraphe 3.5, deuxième exemple) que la valeur propre de A la plus voisine de $\lambda_i^{(n)}$ ne soit pas λ_i mais une autre valeur propre $\lambda_\ell (\ell \neq i)$. Ceci dépend en particulier du choix du vecteur initial. Nous sommes donc amenés à généraliser la proposition 13 ainsi :

PROPOSITION 15 :

Supposons que la valeur propre de A, la plus voisine de $\lambda_i^{(n)}$ soit λ_ℓ et que λ_ℓ vérifie :

$$\lambda_{i+1}^{(n)} < \lambda_\ell < \lambda_{i-1}^{(n)}$$

Supposons de plus que $\pi_n \varphi_\ell \neq \pi_{n-1} \varphi_\ell$.
Soient D_ℓ, $d_\ell, \varepsilon_{\ell,n}$ définis de la même façon que pour la proposition 13. Alors :

1°/ $\quad \| r(\varphi_i^{(n)}) \| \leq D_\ell \ \varepsilon_{\ell,n}$

2°/ $\quad |\lambda_\ell - \lambda_i^{(n)}| \leq \frac{1}{d_\ell} [D_\ell \ \varepsilon_{\ell,n}]^2$

DEMONSTRATION :

La démonstration est analogue à celle de la proposition 13, seule l'inégalité (1) ci-dessus change et devient :

$$\| (A-\lambda_i^{(n)}) \varphi_i^{(n)} \| \leq \frac{\| (A-\lambda_\ell) \pi_n \varphi_i \|}{\| \pi_n \varphi_\ell \|} = \frac{\| (A-\lambda_\ell)(1-\pi_n) \varphi_\ell \|}{\| \pi_n \varphi_\ell \|} \leq D_\ell \ \mathrm{tg}\ \theta(\varphi_\ell, E_n).$$

C.Q.F.D.

REMARQUES SUR LES HYPOTHESES :

1°) D'après l'hypothèse $\lambda_{i+1}^{(n)} < \lambda_\ell < \lambda_{i-1}^{(n)}$ et puisque $\lambda_{i-1}^{(n)} < \lambda_{i-1}$ cela entraîne que $\lambda_\ell < \lambda_{i-1}$ et donc $\lambda_\ell < \lambda_i$ i.e. la valeur propre $\lambda_i^{(n)}$ approche une des valeurs $\lambda_i, \lambda_{i+1}, \ldots$
(cf. exemple pour vérification).

2°) L'hypothèse $\pi_n\varphi_i \neq \pi_{n-1}\varphi_i$ est équivalente d'après la proposition 10, à l'hypothèse suivante : $\min_{\substack{p\in\mathcal{P}_{n-1}\\ p(\lambda_i)=1}} \|p\|_{X_i'} \neq \min_{\substack{p\in\mathcal{P}_{n-2}\\ p(\lambda_i)=1}} \|p\|_{X_i'}$

comme en plus

$$\min_{\substack{p\in\mathcal{P}_{n-1}\\ p(\lambda_i)=1}} \|p\|_{X_i'} \leq \min_{\substack{\mathcal{P}_{n-2}\\ p(\lambda_i)=1}} \|p\|_{X_i'}$$

l'hypothèse $\pi_n\varphi_i \neq \pi_{n-1}\varphi_i$ est équivalente à

$$\min_{\substack{p\in\mathcal{P}_{n-1}\\ p(\lambda_i)=1}} \|p\|_{X_i'} < \min_{\substack{p\in\mathcal{P}_{n-2}\\ p(\lambda_i)=1}} \|p\|_{X_i'}$$

Cette propriété est vraie pour i=1, lorsqu'on remplace la norme $\|\ \|_{X_i'}$ par la norme uniforme sur $[\lambda_N,\lambda_2]$.

Nous ne savons pas si elle est vraie pour la norme $\|\ \|_{X_i'}$.

Cependant la proposition 12 montre que $\pi_n\varphi_i$ est une suite (finie croissante au sens large avec n .

3.5. EXPERIENCES NUMERIQUES.

Nous allons commencer par tester les bornes d'erreurs du paragraphe 3.4, sur un exemple simple, puis nous proposerons diverses expérienc sur des matrices de forme bande d'ordre N = 500 et N = 1000.

En 3.5.3, nous testons la méthode de Lanczos sur une matrice creuse, de dimension 10.000.

3.5.1.

Soit A la matrice de dimension N = 50 construite à l'aide de la formule :

$$A = (I-2uu^T)D(I-2uu^T)$$

où u^T est pris égal au vecteur $e^T = (1,1,1,\ldots,1)$ normalisé.

D est une matrice diagonale contenant les valeurs propres de A. Nous avons choisi de prendre pour valeurs propres

0.01;0.02;0.03;...;4,6;5;7;9;13;

Le vecteur initial X_o est le vecteur e. Ce vecteur fait avec les vecteurs propres φ_i de A le même angle θ_i vérifiant :

$$tg(\theta_i) = \sqrt{N-1} = 7 .$$

On a pris $n = 15$ i.e. $E_n(X_o) = [(X_o, AX_o, \ldots A^{14}X_o)]$

Pour i=1,2,3 nous avons fait calculer : (cf. tableau .1.)

1°) $\varepsilon_{i,n}$, majoration de $t_{i,n}$ données par la proposition 12

2°) $\varepsilon_{i,n}\ tg\theta_i$ majorations des $tg\theta_i(\varphi_i, E_n(X_o))$

3°) $tg\theta_i(\varphi_i, E_n)$

4°) $D_i\ \varepsilon_{i,n}\ tg\theta_i$ qui majorent $\|r(\varphi_i^{(n)})\|$ d'après la proposition 15

5°) $\|r(\varphi_i^{(n)})\|$.

REMARQUE 1 :

Il est nécessaire, si l'on veut comparer les erreurs de méthode seulement, de prendre des matrices de dimensions assez petites pour éviter l'accumulation des erreurs d'arrondis (Cf. remarque 2).

Tableau .1. N = 50 , n = 15

i	$\varepsilon_{i,n}$	$\varepsilon_{i,n} tg\theta_i \geq tg\theta(\varphi_i, E_n)$		$D_i \varepsilon_{i,n} tg\theta_i \geq \|r(\varphi_i^{(n)})\|$	
1	$4,959 \times 10^{-8}$	$3,471 \times 10^{-7}$	$1,186 \times 10^{-7}$	$4,509 \times 10^{-6}$	$3,887 \times 10^{-}$
2	$1,069 \times 10^{-5}$	$7,481 \times 10^{-5}$	$1,203 \times 10^{-7}$	$6,725 \times 10^{-4}$	$4,458 \times 10^{-}$
3	$1,166 \times 10^{-5}$	$8,163 \times 10^{-5}$	$9,020 \times 10^{-6}$	$5,706 \times 10^{-4}$	$3,728 \times 10^{-}$

Les résultats du tableau 1 indiquent que $D_i tg\theta_i \varepsilon_{i,n}$ est souvent une majoration assez large de $\|r(\varphi_i^{(n)})\|$.

Ceci pouvait être prévu en raison des nombreuses inégalités utilisées pour établir les résultats de la proposition 15.

Nous allons comparer expérimentalement, les évolutions de $\|r(\varphi_i^{(n)})\|$ et de $D_i \varepsilon_{i,n} tg\theta_i$ en fonction de n.

Pour cela, nous représentons graphiquement dans la figure 1, $\log(\|r(\varphi_2^{(n)})\|)$ et $\log(D_2, \varepsilon_{2,n}, tg\theta_2)$ en fonction de n.

- On <u>vérifie</u> que $\log(D_2, \varepsilon_{2,n}, tg\theta_2)$ décroît de manière presque linéaire.

  [En effet, si on pose $K_2' = D_2 K_2 tg\theta_2$ on a

$$\log(D_2 \varepsilon_{2,n}\ tg\theta_2) = \log\left(\frac{K_2'}{ch(n-2)\gamma_2}\right) = \log K_2' - \log\left(\frac{e^{(n-1)\gamma_2}}{2}(1+e^{-2(n-2)\gamma}\right)$$

  $\#$ $\text{Log } K_2' - (n-2)\gamma_2 \log \frac{e}{2}$]

- On observe pour $\log(\|r(\varphi_2^{(n)})\|)$ une décroissance qui semble également voisine d'une décroissance linéaire et de pente assez voisine de de $\log(D_2\ \varepsilon_{2,n}\ tg\theta_2)$.

<u>REMARQUE 2</u> :

Pour $n \geq 20$, l'effet, de l'accumulation des erreurs n'est plus négligeable et l'irrégularité de la courbe $\log(\|r(\varphi_2^{(n)})\|)$ est probablement dûe à cet effet.

Nous faisons les mêmes constatations pour i = 3 (cf. figure 2). Les deux courbes sont ici plus voisines que pour i = 2.

Toutes ces expériences "montrent" qu'il existe une constante K(i) indépendante de n telle que :

$$r(\varphi_i^{(n)}) \simeq \frac{K(i)}{ch(n-i)\gamma_i}$$

c'est cette constante K(i) qui a été majorée dans la proposition 15.

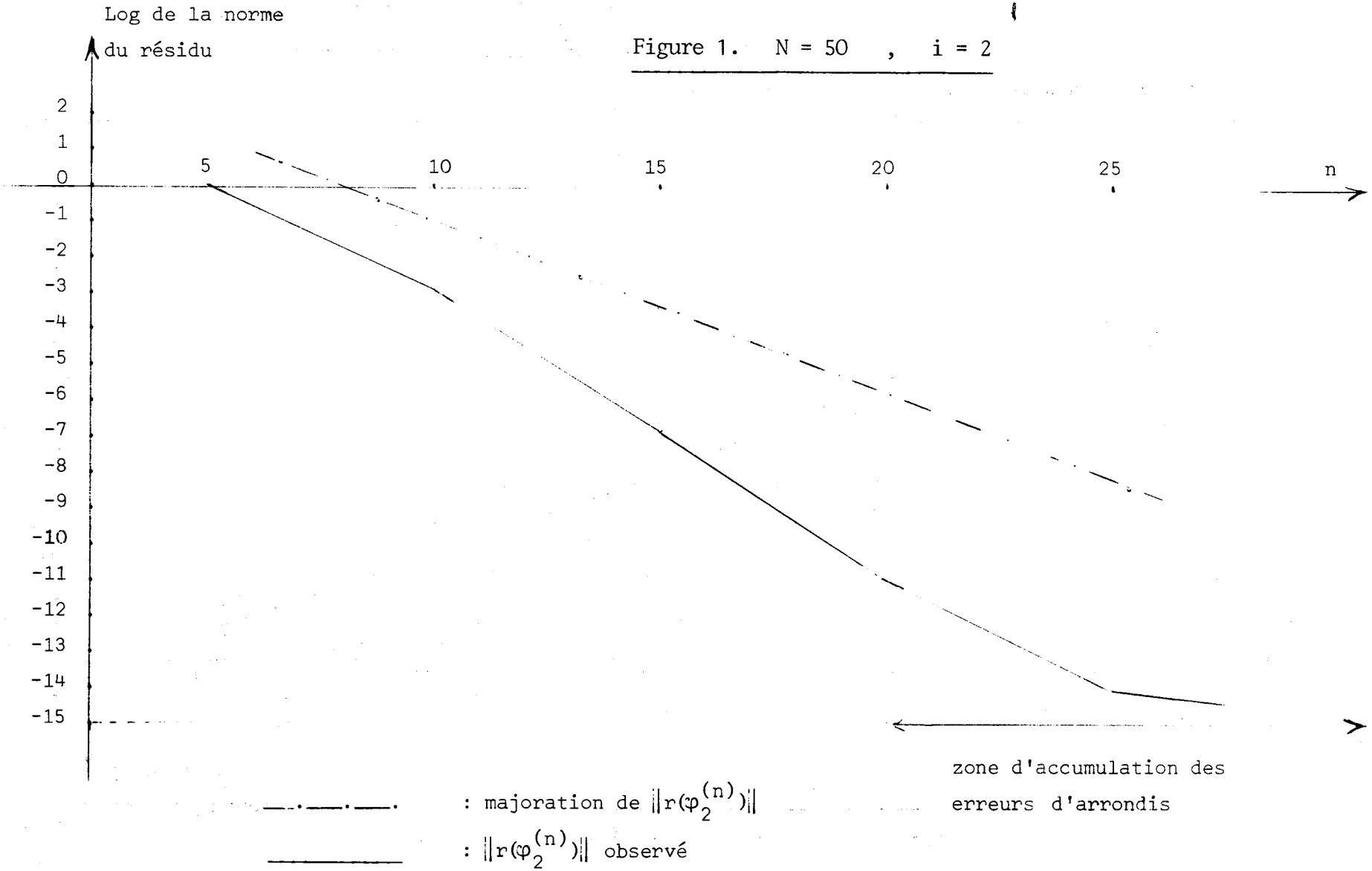
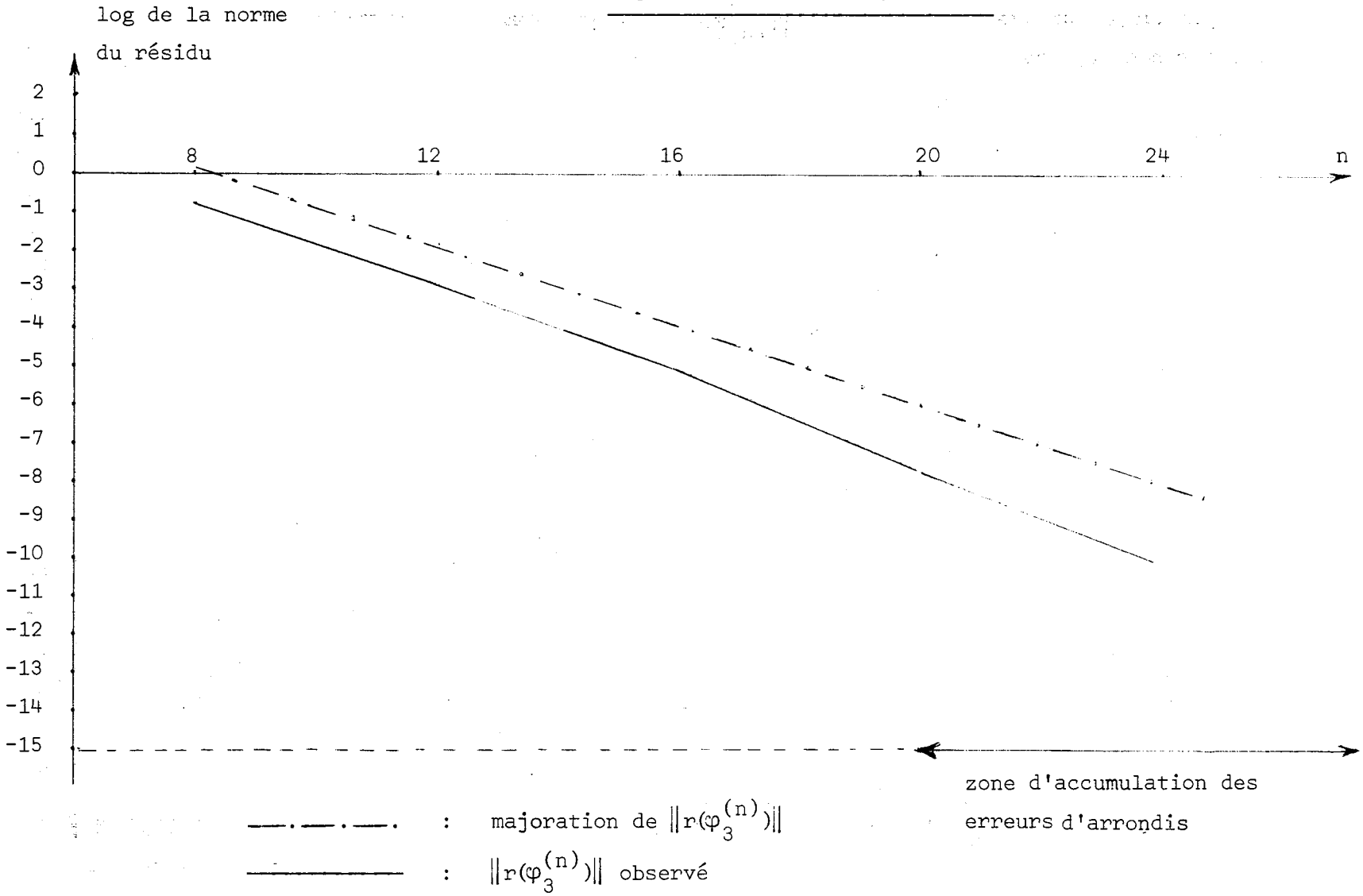


Figure 1. N = 50 , i = 2

Figure 2. N = 50 , i = 3



3.5.2.

En général $\lambda_i^{(n)}$ approche une valeur propre λ de A d'après la proposition 15, cependant λ n'est pas toujours la $i^{ème}$ valeur propre λ_i de A. C'est ce que montrent les deux exemples suivants :

Soit $A = \begin{pmatrix} B & -I & & & \\ -I & B & -I & & \\ & -I & B & \ddots & \\ & & \ddots & \ddots & -I \\ & & & -I & B \end{pmatrix}$

où

$$B = \begin{pmatrix} 4 & -1 & & & \\ -1 & \ddots & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & -1 \\ & & & -1 & 4 \end{pmatrix}$$

est de dimension 20 et A est de dimension N = 500.

On pose A' = A-4I.

Nous avons essayé la méthode de Lanczos (sans réorthogonalisation en prenant comme vecteur initial le vecteur

$$X_o = (A')^5 e \; / \; \|(A')^5 e\|$$

où

$$e^T = (1,1,1,\ldots,1)$$

La méthode a été effectuée sur A' au lieu de A afin d'alléger les calculs. Nous prenons :

$$n = 50 \quad \text{i.e.} \quad E_n(X_o) = [(A'^5 e,\ A'^6 e,\ldots,A'^{56} e)]$$

REMARQUE 3 :

L'espace $E_n(X_o) = [(X_o, A'X_o, \ldots, A'^{49}X_o)]$ étant identique à l'espace $[(X_o, AX_o, \ldots, A^{49}X_o)]$, les valeurs propres E_n-approchées de A' diffèrent des valeurs propres E_n-approchées de A par la constante 4.

Dans le tableau 2, nous indiquons :

1 - Pour i=1,2,3,4,5 , les valeurs $\lambda_i^{(n)}$ calculées par l'algorithme de Lanczos.

2 - Pour i=1,2,3,4,5 , le numéro ℓ de la valeur propre exacte λ_ℓ qui est approchée par $\lambda_i^{(n)}$ (i.e. telle que $|\lambda_i^{(n)}-\lambda_\ell| = \min_{\lambda\in\sigma(A)} |\lambda_i^{(n)}-\lambda|$) , ainsi que l'erreur $\delta_{i,n} = |\lambda_i^{(n)}-\lambda_\ell|$ et la précision : $\left|\frac{\lambda_i^{(n)}-\lambda_\ell}{\lambda_\ell}\right|$.

Tableau .2.

$N = 500$, $n = 50$, $X_o = (A')^5 e / \|(A')^5 e\|$

VALEURS PROPRES CALCULEES

* I *	* LAMBDA(I,I) *
1	+01 7[illegible]
2	+01 7761178[illegible]
3	+01 7637899[illegible]
4	+01 7557113[illegible]
5	+01 7522519[illegible]

I * I.P. EXACTE *	DELTA(I,I) *	PRECISION
3	1.7[illegible]'-14	2.221412'-15
7	1.7[illegible]'-14	0.078193'-15
12	2.1[illegible]'-10	2.878552'-11
16	7.[illegible]'-09	9.516825'-10
17	1.1[illegible]'-08	1.500059'-09

On constate donc que $\lambda_1^{(n)}$ approche λ_3 , $\lambda_2^{(n)}$ approche λ_7
$\lambda_3^{(n)}$ approche λ_{12} , $\lambda_4^{(n)}$ approche λ_{16}
et $\lambda_5^{(n)}$ approche λ_{17} .

b) Soit A la matrice symétrique définie de même façon qu'en a) mais de dimension N = 1000 : A est bloc-tridiagonale avec des blocs de dimension 20 .

De même, qu'en a), on pose A' = A - 4I .

Prenons également ici n = 50 et $X_o = (A')^5 e \,/\, \|(A')^5 e\|$.

Nous obtenons alors les résultats indiqués au tableau 3, où les notations sont les mêmes qu'en a).

Tableau .3.

$N = 1000$, $n = 50$, $X_o = (A')^5 e \,/\, \|(A')^5 e\|$

VALEURS PROPRES CALCULEES

* I *	* LAMBDA(I,N) *
1	+01 7[illegible]
-	+01 7[illegible]
3	+01 7[illegible]
4	+01 7[illegible]
5	+01 7[illegible]

N° DE LA V.P. EXACTE	* DELTA(I,N) *	PRECISION
1	0.2[illegible]-09	7.8[illegible]-10
3	8.0[illegible]-08	1.1[illegible]-08
12	2.1[illegible]-06	2.7[illegible]-07
16	0.0[illegible]	0.0[illegible]-05
17	0.00[illegible]	0.0[illegible]

c) Les tableaux 4, 5 et 6 suivants décrivent certaines expériences pour N = 500 et N = 1000 , avec différents choix du vecteur initial X_o et d la dimension n .

Tableau .4.

$$N = 1000 \quad , \quad n = 50 \quad , \quad X_o = (A')^5 e_1 \; / \; \|A'^5 e_1\|$$

VALEURS PROPRES CALCULEES

* I *	* LAMBDA(I,N) *
1	+01 7970230789679 20
2	+01 7947546586831 94
3	+01 7904517479055 28
4	+01 7858039413736 53
5	+01 7792547069127 582

NO DE LA V.P. EXACTE	* DELTA(I,N)	* PRECISION
1	0.003037520	0.0004561801
3	0.003938735	0.0004558370
5	0.0024347 80	0.0003079147
9	0.007299930	0.0009298398
15	0.002044505	0.0002622979

Tableau .5.

$$N = 500 \quad , \quad n = 70 \quad , \quad X_o = (A')^5 v \; / \; \|A'^5 v\|$$

où v a pour composantes $v_i = 1 + 10^3 \sin(3i)$, i=1,...,N

VALEURS PROPRES CALCULEES

* I *	* LAMBDA(I,N) *
1	+01 7963079400646 28
2	+01 7919545287263 92
3	+01 7896503311870 20
4	+01 7853028291082 70
5	+01 7847093659739 40

ECART A LA V.P. EXACTE *	DELTA(I,N) *	PRECISION
1	6.704149 '-14	1.093063 '-14
2	3.644169 '-11	4.854028 '-12
3	4.789812 '-08	6.065692 '-09
4	9.355417 '-07	1.210526 '-07
5	4.780817 '-07	6.092002 '-08

Tableau .6.

$$N = 1000 \quad , \quad n = 80 \quad , \quad X_o = A'^5 v \; / \; \|A'^5 v\|$$

où $v_i = 1 + 10^4 \sin(4i)$

VALEURS PROPRES CALCULEES

* I *	* LAMBDA(I,N) *
1	+01 797386612812070
2	+01 790245017610252
3	+01 794355718851701
4	+01 791564070740051
5	+01 788288059076275

NO DE LA V.P. EXACTE *	DELTA(I,N) *	PRECISION
1	2.181798'-06	2.730185'-07
2	5.249569'-05	6.592863'-06
3	5.006330'-05	6.377870'-06
4	0.001614757	0.0002039541
7	0.0006500625	8.321922'-05

Les expériences précédentes montrent que le choix du vecteur initial est très important si on veut éviter des cas particuliers tels que ceux des tableaux 2, 3, 4. Ce choix du vecteur initial est le problème le plus difficile du point de vue pratique : l'idéal serait de disposer d'un vecteur qui approche les vecteurs propres associés aux valeurs propres dominantes. Ceci arrive souvent dans les problèmes qui proviennent des approximations de phénomènes physiques (cf. 1 , 2). De manière générale

si on ne dispose d'aucune information initiale, il y a intérêt à choisir un vecteur initial qui soit le plus "aléatoire" possible.
Dans les exemples précédents, les matrices étudiées ont une structure très particulière par rapport à la base $e_1, e_2, \ldots, e_N$, et des choix tels que

$$X_o = A^p e_1 \quad \text{(tableau.4.)}$$

ou

$$X_o = A^p e \quad \text{(tableau .5.)}$$

sont a déconseiller.
Des vecteurs du type de ceux décrits dans les tableaux 5, 6 conviennent mieux.

Nos expériences numériques nous ont permis de faire les remarques suivantes :

1. On pourrait prendre un vecteur initial de la forme

$$X_o = \frac{A^p v}{\|A^p v\|}$$

où p est un entier choisi et v un vecteur quelconque.
Il est peu intéressant de prendre p trop grand car il en résulte une accumulation des erreurs d'arrondis trop importante (cf. [9], [6], [5]).

2. La réorthogonalisation des vecteurs v_i est inutile si on écrit des procédures efficaces pour le calcul de $X \to AX$ et $X, Y \to (X, Y)$.

3.5.3.

Nous allons tester la méthode de Lanczos sur la matrice :

$$A = \begin{pmatrix} B & -I & & & \\ -I & B & -I & & \\ & \ddots & \ddots & \ddots & \\ & -I & \ddots & \ddots & -I \\ & & -I & B & \end{pmatrix}$$

d'ordre N = 10.000 où :

$$B = \begin{pmatrix} 4 & -1 & & & \\ -1 & \ddots & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & -1 \\ & & & -1 & 4 \end{pmatrix}$$

est de dimension 80 .

Nous nous intéressons aux plus petites valeurs propres de A. Ce problème présente deux difficultés majeures :

- La dimension étant grande, l'accumulation des erreurs d'arrondis est importante.

- Les valeurs propres extrémales sont très voisines et les constan $\gamma_i = 1 + 2 \frac{\lambda_i - \lambda_{i-1}}{\lambda_i - \lambda_N}$ dont dépend la rapidité de la convergence (cf. proposition 12) sont très voisines de 1 (convergence lente)

Les résultats du tableau .7. montrent cependant que l'on obtient de bonnes approximations de certaines valeurs propres.

Tableau .7.

$$N = 10.000 \quad , \quad n = 170 \quad , \quad X_o = A'^5 v \,/\, \|A'^5 v\|$$

où $v_i = 1 + 100 \sin(4i)$

VALEURS PROPRES CALCULEES

* I *	* LAMBDA(I,N) *
1	-02 21257[illegible]9219315
2	-02 7[illegible]391537120
3	-01 14143378[illegible]11116
4	-01 17163755[illegible]79769
5	-01 19637775542760[illegible]4
6	-01 31103[illegible]3142778972
7	-01 37954056645[illegible]774
8	-01 45335354717150[illegible]
9	-01 52862245327070[illegible]4
10	-01 72971795357[illegible]9362

N° DE LA V.P. EXACTE *	DELTA(I,N) *	PRECISION
1	2.235[illegible]61'-12	1.[illegible]7532[illegible]'-[illegible]
4	3.2[illegible]61[illegible]2'-[illegible]	4.[illegible]3[illegible]721'-[illegible]
3	3.45[illegible]67'-[illegible]	[illegible].[illegible]244445[illegible]
65	[illegible].[illegible]13[illegible]741	[illegible].[illegible]81[illegible]54
51	[illegible].[illegible]572[illegible]94	[illegible].[illegible]2952774
97	[illegible].[illegible]7802747	[illegible].[illegible]244[illegible]877
34	[illegible].[illegible]1563531	[illegible].[illegible]411[illegible]444
114	[illegible].[illegible]3337373	[illegible].[illegible]73[illegible]8614
69	[illegible].[illegible]14872[illegible]2	[illegible].[illegible]28[illegible]5461
25	[illegible].[illegible]1,34	[illegible].[illegible]123[illegible]4[illegible]

BIBLIOGRAPHIE

[1] VANDERGRAFT J.S.
Generalized Rayleigh method with applications to finding eingenvalue of large matrices.
J. of linear algebra and its applications 4 (1971) 353-368.

[2] KANIEL, S.
Estimates for some computationnal techniques in linear algebra.
Math. of comp. 20 (1966) 369-378.

[3] ERDELYI, I.
An iterative least square algorithm suitable for computing partial eingensystems.
S.I.A.M. J. of numer. Anal. Ser B(3) 2 (1965).

[4] LANCZOS, C.
An iterative method for the solution of the eingenvalue problem of linear differential and integral operators.
J. res. Nat. Bur. Stand. 45 (1950).

[5] RUHE, A.
Iterative eingenvalue algorithms for large symmetric matrices.
Report UMINEF 31 (1972).

[6] PAIGE, C.C.
Practical use of the symmetric Lanczos process with reorthogonalization.
B.I.T. 10 (1970) 183-195.

[7] GOLUB, G.H.
Some uses of the Lanczos algorithm in numerical linear algebra.
Technical report STAN CS.72.302 . Computer Science department Stanford University (1972).

[8] WILKINSON, J.H. ; PETERS, G.
AX = λBX and the generalized eingenproblem.
S.I.A.M. J. of Num. Anal. Vol. 7 n° 4 (1970) pp.479-492.

[9] WILKINSON, J.H.
The algebraïc eingenvalue problem.
Clarendon Press (1965).

[10] CHATELIN, F.
Calcul numérique de valeurs propres d'opérateurs linéaires.
Cours D.E.A. Analyse Numérique. Université de Grenoble (1973-1974).

[11] CHATELIN, F.; LEMORDANT J.
La méthode de Rayleigh-Ritz appliquée à des opérateurs elliptiques.
Ordre de convergence des éléments propres.
Séminaire d'Analyse Numérique. Université de Grenoble,
premier semestre 1973-1974.

[12] HOUSEHOLDER, A.S.
The theory of matrices in numerical Analysis.
Blaisdell-Waltman Mass. (1964).

[13] GASTINEL, N.
Analyse Numérique Linéaire. Hermann (1966).

[14] RUTISHAUSER, H.
Computational aspect of F.L. Bauer's simultaneous iteration method.
Num. Math. 13 (4-13) (1969).

[15] DAVIS, C. ; KAHANE, W.H.
The rotation of eingenvectors by a perturbation.
S.I.A.M. J. Num. Analysis Vol. 7 - 1 pp.1-46 (1970).

[16] KELLER, H.B. ; ISAACSON, E.
Analysis of numerical methods.
Wiler, New-York (1966).

[17] JENSEN, P.S.
The solution of large symmetric eingenproblems by sectionning.
S.I.A.M. J. of Num. Analysis Vol.9 n° 4 (1972) 534-545.

[18] HOUSEHOLDER, A.S.
"Moments and characteristic roots II"
Num. Math. 11 (1968) pp.126-128.

[19] FADDEV, D.K. ; FADDEEVA,V.N.
Computational methods of linear algebra
W.H. Freeman and Co, San Francisco London (1961).

[20] LAURENT, P.J.
Approximation et optimisation.
Ed. Hermann (1973).

BIBLIOGRAPHIE GENERALE

ABRAMOV, A. :

"On the separation of the principal part of some algebraic proble
Zh. Vych. Mat. 2, N° 1, 141-5 (1962).

ABRAMOV, A. :

"Remarks on finding the eigenvalues and eingenvectors of matrices which arise in the application of Ritz's method or in the difference method".
Zh. vych. Mat. Fiz. 7, 3, 644-647.

ABRAMOV, A. ; NEUHAS, M. :

" Bemerkungen über Eingenwertprobleme von Matrizen hoherer Ordnung".
C.R. du Congrès Int. des Math. de l'ingénieur Mons et Bruxelles (

CHATELIN, F. :

"Méthodes numériques de calcul des valeurs propres et vecteurs propres d'un opérateur linéaire".
Thèse Université Scientifique et Médicale de Grenoble (1971).

CHATELIN, F. :

"Calcul numérique de valeurs propres d'opérateurs linéaires"
Cours D.E.A., Analyse numérique, Université Scientifique et Médicale de Grenoble (1973).

CHATELIN, F. ; LEMORDANT, J. :

"La méthode de Rayleigh-Ritz appliquée à des opérateurs elliptiqu
Ordre de convergence des éléments propres ".
Séminaire d'Analyse Numérique, Université Scientifique et Médical de Grenoble, premier semestre (1973-1974).

CHICHOV, V.S. :

"A method for partitionning a high order matrix into blocks in order to find its eingenvalues".
Zh. vych. Mat. 1, N°1, 169-173 (1961) .

CHICHOV, V.S. :

"The determination of eingenvalues and eingenfunctions of a linear integral operator with a symmetric kernel by means of groupe elimination of unknowns".
Zh. Vych. Mat. 2, N°3, 389-410 (1962).

DAVIS, C. ; KAHAN , W.M. :

"The rotation of eingenvectors by a perturbation".
SIAM, J. Num. Analys. Vol. 7 - 1, (1-46) (1970).

DEKKER T.J. ; TRAUB, J.F. :

"The shifted QR algorithm for hermitian matrices".
J. of lin. Alg. and its appl. 4, 137-154 (1971).

DURAND, E. :

"Solutions numériques des équations algébriques".
Tome II. Masson et Cie.

ERDELYI, I. :

"An itérative least square algorithm suitable for computing partial eingensystems".
SIAM J. Numer. Anal. B 3.2 (1965).

FADDEV, D.K. ; FADDEEVA, V.N. :

"Computational methods of linear algebra".
W.H. Freemann and Co. San Franciscon-London (1961).

GASTINEL, N. :

"Analyse numérique linéaire".
Hermann Ed. (1966).

GOLUB, G.H. :

"Some uses of the symmetric Lanczos process with reorthogonalisation in numerical linear algebra".
Technical report Stan. CS. 72-302. Computer Science Department
Stanford University (1972).

GOULD, S.H. :

"Variational methods for eingenvalue problems".
University of Torento Press (1957).

GREGORY, R.T. ; KARNEY, D.A. :

"A collection of matrices for testing computationnal algorithms".
John Wiley and Sons, New-York (1969).

HOUSEHOLDER, A.S. :

"The theory of matrices in numerical analysis.
Blaisdell-Walt mann-Mass (1964).

ISAACSON, E. ; KELLER, H.B. :

"Analysis of numerical methods".
Wiley, New-York (1966).

JENSEN, P.S. :

"The solution of large symmetric eingenproblems by sectionning".
SIAM J. Numer. Anal. Vol 9, N° 4 (1972) (534-545).

KANIEL, S. :

"Estimates for some computationnal techniques in linear algebra".
Math. of Comp. 20 (1966) (369-378).

KATO, T. :

"Perturbation theory for linear operators".
Springer Verlag (1965).

KRASNOSELSKII et Al. :

"Approximate solutions of operator equations".
Wolters-Noordhoff (1972).

LANCZOS, C. :

"An iterative method for the solution of the eingenvalue problem of linear differential and integral operators".
J. Res. Nat. Bur. Stan. 45 (1950).

LAURENT, P.J. :

"Approximation et Optimisation".
Ed. Hermann (1972).

LEBAUD, C. :

"L'algorithme double QR avec shift".
Numer. Mat. 16. 163-180 (1970).

PAIGE, C.C. :

"Practical use of the symmetric Lanczos process with reorthogonalisation".
B.I.T. 10 (1970) (183-195).

PARLETT, B.N. :

"Présentation géométrique des méthodes de calcul des valeurs propres".
Numer. Mat. 21-3 (1973) (223-233).

PETERS, G. ; WILKINSON, J.H. :

" $Ax = \lambda BX$ and the generalized eingenproblem".
SIAM J. of Num. Anal. Vol.7 N° 4, pp. 479-492.

REINSCH, C. ; WILKINSON, J.H. :

"Handbook for automatic computation".
Volum II ; linear algebra Springer-Verlag (1972).

RUHE, A. :

"Iterative eingenvalue algorithms for large symmetric matrices". Report UMINEF 31 (1972).

RUTISHAUSER, H. :

"Comutationnal aspects of F.L. Bauer's simultaneous iteration method".
Num. Mat. 13 (4-13) (1969).

SAAD, Y. :

"Quelques applications du partitionnement au calcul de valeurs propres de matrices hermitiennes".
Séminaire d'Analyse Numérique, premier semestre 1972-1973, Université Scientifique et Médicale de Grenoble.

SAAD, Y. :

"Etude des translations d'origine dans les algorithmes LR et QR".
C.R.A.S. n° 2, janvier 1974.

SAAD, Y. :

"Shiftsof origin for the QR algorithm".
I.F.I.P. Congress. Stockholm (1974) (à paraître).

STEINGER, W. ; WEINSTEIN, A. :

"Methods of intermediate problems for eingenvalues".
Academic Press (1972).

VANDERGRAFT, J.S. :

"Generalized Rayleigh methods with applications to finding eingenvalues of large matrices".
J. of lin. alg. and its appl. 4 (1971) 353-368.

WILKINSON, J.H. :

"The algebraic eingenvalue problem". Clarendon Press (1965).

WILKINSON, J.H. :

"Global convergence of the QR algorithm".
Proceedings of the IFIP Congress (1968) A22-A24.