Autoapprentissage d'une partition : application au classement itératif de données multidimensionnelles

Ramon Lopez de Mantaras Badia

N° d'ordre : 1998

Année 1977

THESE

présentée

DEVANT L'UNIVERSITE PAUL SABATIER DE TOULOUSE (SCIENCES)

en vue de l'obtention

du TITRE de DOCTEUR de SPECIALITE E.E.A.

Option Automatique

par

Ramón LOPEZ DE MANTARAS BADIA

Ingénieur EUIT de Mondragon, Espagne
Maître ès Sciences

AUTOAPPRENTISSAGE D'UNE PARTITION
APPLICATION AU CLASSEMENT ITERATIF DE DONNEES MULTIDIMENSIONNELLES

Soutenue le 21 juin 1977, devant la commission d'examen :

MM.	G. GRATELOUP	Président
	J. AGUILAR-MARTIN	Examinateurs
	E. DIDAY	
	G. GIRALT	
	D. NUALART	
	G. PERENNOU	

A la gent qu'estimo

AVANT-PROPOS

Le travail de recherche présenté dans ce mémoire a été effectué au Laboratoire d'Automatique et d'Analyse des Systèmes du C.N.R.S. dans l'équipe Aspects Stochastiques de l'Automatique animée par J. AGUILAR MARTIN, Maître de Recherche au C.N.R.S., qui m'a constamment encouragé et grâce à qui ce travail a pu voir le jour. Je lui exprime ici toute ma reconnaissance pour l'amicale et efficace manière dont il a dirigé mes travaux.

Je tiens à remercier Monsieur le Professeur G. GRATELOUP, qui a la charge de diriger le L.A.A.S., d'avoir bien voulu accepter de présider le jury de thèse.

Il m'est agréable d'exprimer ici ma reconnaissance à Monsieur E. DIDAY, Maître de Conférences à l'Université de Paris IX qui a bien voulu s'intéresser à ce travail, à Monsieur G. GIRALT, Directeur de Recherche au C.N.R.S., qui me fait l'honneur de participer au jury de cette thèse et à Monsieur le Professeur G. PERENNOU qui a bien voulu nous faire profiter de son expérience en acceptant d'être l'un des examinateurs de ce

jury.

Je veux dire ici à D. NUALART, Professeur à l'Université Polytechnique de Barcelone, combien ses remarques judicieuses et sa clairvoyance mathématique m'ont aidé dans la mise au point de ce mémoire et je lui exprime ma profonde gratitude pour l'aide apportée et pour sa participation au Jury de ma thèse.

Je suis redevable à Messieurs G. BANON, M. BRIOT et A. FOURNIE pour les discussions sur les problèmes touchant à ce travail.

Mes remerciements vont aussi à toutes les personnes qui on d'une façon ou d'une autre participé à la réalisation matérielle de ce mémoire.

Finalement je voudrais remercier le C.R.O.U.S. et en particulier Monsieur le Professeur H. MASCART qui par leur soutien matériel m'ont permis de me consacrer entièrement à ce travail de recherche.

INTRODUCTION

0)- Introduction

Dans le domaine de l'automatisation, le comportement humain est le modèle de référence le plus proche et celui dont il est le plus facile d'évaluer les défauts ou les qualités de façon qualitative. Une des caractéristiques principales que l'homme partage avec les ordinateurs est l'utilisation d'une mémoire. On peut constater que, bien que la mémoire de l'ordinateur soit d'accès plus rapide et réalise une accumulation complète et sans oubli, elle atteint rapidement sa saturation devant un environnement d'informations hétérogènes et non classées. Or, l'homme, dans ce même environnement qui lui est familier, avec une mémoire beaucoup moins exacte est capable de tirer des conséquences et de prendre des décisions admissibles.

Deux fonctions liées l'une à l'autre, semblent à la base de cette souplesse dans le comportement :

1. L'utilisation d'une capacité d'oubli relatif ou plutôt de résumer l'information sous forme de traits à retenir en oubliant les autres.
2. La possibilité d'établir des associations entre idées, en utilisant des relations de similitude.

Nous allons essayer, dans ce travail, de formuler une des manifestations essentielles de ce comportement : le classement, par autoapprentissage, en absence d'information initiale.

Grâce à la faculté d'oubli l'homme est capable d'ignorer certaines caractéristiques d'un objet et peut l'associer à un autre objet en raison des caractéristiques non oubliées. Ceci permet d'établir des pseudo-équivalences entre objets et, comme nous montrerons dans la suite de ce travail, d'effectuer des classements. Les mécanismes d'oubli et d'association peuvent se faire de façon adaptative en modifiant en permanence, suivant un algorithme, les critères de classification.

La recherche de tels algorithmes est dictée par le désir de prendre en charge certaines caractéristiques de l'opérateur humain dans la conduite de processus où des situations diverses sont possibles.

Dans la 1ère partie on traite quelques aspects théoriques sur l'autoapprentissage et on pose le problème du classement comme étant celui de l'estimation d'une partition.

Dans la 2ème partie nous développons deux indices permettant de comparer des partitions dont l'un remplit les axiomes de la distance, ces indices vont être très utiles au niveau de l'analyse des résultats puisqu'ils vont nous permettre de mesurer la différence entre les partitions obtenues par notre algorithme et la partition "vraie" ou "naturelle" donnée a priori.

Dans la 3ème partie nous présentons un algorithme de classement itératif par autoapprentissage basé sur le critère du maximum de vraisemblance comme règle de décision, les probabilités d'appartenance des éléments aux classes sont estimées

par comptage pondéré.

Dans la dernière partie enfin, on porte l'attention sur un algorithme de classement itératif par autoapprentissage de données gaussiennes, l'élément fondamental de la récursivité de cet algorithme est le filtrage linéaire adaptatif qui est utilisé pour estimer les densités de probabilité gaussiennes qui caractérisent chaque classe.

Nous avons appliqué ces algorithmes à la reconnaissance tactile de solides de forme géométrique régulière, en plus l'algorithme gaussien a été appliqué au problème de la reconnaissance des composants d'un mélange de densités de probabilité gaussiennes multidimensionnelles, de façon a être comparé avec la méthode des Nuées Dynamiques.

-:-

CHAPITRE I

ASPECTS THEORIQUES SUR L'AUTOAPPRENTISSAGE ET L'ESTIMATION D'UNE PARTITION

-:-

I-1. ASPECTS GENERAUX SUR L'AUTOAPPRENTISSAGE

Le jeune enfant entend des sons émis par son entourage, leur apparition n'est pas purement aléatoire et une partition se dégage dans l'ensemble des sons dont les classes sont les phonèmes du système linguistique de son milieu c'est l'apprentissage sans professeur et sans information a priori, apprentissage libre, ou à environnement passif ; ensuite l'enfant découvre la fonction de communication du langage et il entre dans un apprentissage avec environnement actif : la sanction étant donnée par la mesure de la réussite de cette fonction, plus tard les personnes chargées de l'éducation de cet enfant fourniront la "vérité" des phonèmes du système linguistique qui ne seraient pas correctement distingués par lui, il s'agit du type d'apprentissage avec professeur.

Nous essayons de donner ici un modèle mathématique de cet aspect de l'apprentissage libre ou avec environnement passif.

On peut remarquer que cet apprentissage peut être orienté par un "guide" qui sans être un professeur qui sanctionne peut sélectionner l'ordre des données et influer sur la période transitoire de l'apprentissage.

On constate que l'apprentissage libre, que nous appellerons autoapprentissage, est une situation théoriquement élémentaire à partir de laquelle nous pourrons introduire l'autoapprentissage orienté, l'apprentissage avec environnement actif et l'apprentissage avec professeur.

Nous postulons que si cette fonction d'autoapprentissage est totalement absente il n'y a pas de véritable apprentissage mais simple "conditionnement".

I-2. AUTOAPPRENTISSAGE ET ESTIMATION D'UNE PARTITION

Le processus d'autoapprentissage consiste à ranger les données passées dans une partition qui sera remaniée au fur et à mesure que de nouvelles données arrivent.

Appelons Ω un ensemble d'éléments présentant une structure d'espace mesurable (un cas particulier important étant un espace euclidien structuré par rapport à une mesure de probabilité sous forme de mélange de populations).

Une donnée issue de Ω consiste en l'observation d'un élément x_t de cet ensemble.

Un tirage séquentiel de données S_i est une suite de données $(x_{t_o},\ldots,x_{t_i})$ qui s'arrête à l'instant t_i.

Il est utile de distinguer S_i du sous ensemble observé $E_i \subset \Omega$ à l'instant t_i puisque E_i est l'ensemble des éléments qui apparaissent dans la suite S_i.

E_i peut être, dans certains cas, formé des mêmes éléments que S_i mais où l'ordre temporel est supprimé, dans d'autres cas on prendra pour E_i le plus petit convexe contenant les éléments de S_i (cas des espaces euclidiens).

Si nous considérons une suite S_i elle soustend une suite de suites $\left\{S_j\right\}_{j=0}^{i}$ telles que S_j est la suite des j premiers éléments de S_i.

Un algorithme d'autoapprentissage est la suite d'opérations suivante, qui se renouvelle par incrémentation de l'indice

i :

Sont données à l'instant t_i :

La suite S_i

Le sous ensemble E_i (directement construit à partir de S_i)

La partition de E_i : $\mathcal{P}_i$

ainsi qu'un nouvel élément $x_{i+1} \in \Omega$

A l'aide d'une règle de décision (ou reconnaissance) $\mathcal{R}$ on attribue x_{i+1} à une classe $C_\alpha \in \mathcal{P}_i$

Une transformation $\mathbb{A}$ est appliquée à E_i qui devient E_{i+1}.

La partition $\mathcal{P}_i$ de E_i devient la partition $\mathcal{P}_{i+1}$ de E_{i+1}.

Remarque 1.1 :

La représentation de la partition $\mathcal{P}_i$ peut être faite de trois façons :

. Non paramétrique sans résumé (catalogue de tous les éléments de S_i)

. Non paramétrique avec résumé (choix des éléments de S_i les plus utiles à la règle de décision)

. Paramétrique (ensemble de paramètres Θ_i qui représentera la partition et sera utilisé dans la règle de décision).

Remarque 1.2 :

Soit $\mathcal{P}_\Omega$ l'ensemble des partitions de Ω soit $\mathcal{P}_i$ une partition de $E_i \subset \Omega$ alors $\mathcal{P}_i^* = \mathcal{P}_i \cup \left\{ \complement_\Omega E_i \right\}$ est une partition de Ω et donc : $\mathcal{P}_i^* \in \mathcal{P}_\Omega$

Processus d'autoapprentissage ou estimation d'une partition

Le problème du classement peut être formulé comme étant le problème d'estimer une partition.

L'ensemble $\mathcal{P}_\Omega$ est alors muni d'une distance [Chap. II]: est un espace métrique.

Definissons sur Ω une mesure-probabilité caractérisée par une densité au sens large que nous noterons :

$$p[x|\Omega] \qquad \forall x \in \Omega$$

ainsi qu'une partition $\mathcal{P}_v$ dite "vraie" et une suite $\{p[x|C_K]\}_{C_K \in \mathcal{P}_v}$ telle que :

$$p[x|\Omega] = \frac{1}{Card(\mathcal{P}_v)} \sum_{C_K \in \mathcal{P}_v} p[x|C_K] \qquad \forall x \in \Omega$$

Alors, toute partition $\mathcal{P}_i^*$ construite à partir de la séquence S_i et étendue a Ω sera appelée un estimateur (plus correctement une statistique) de $\mathcal{P}_v$, sa distance $d(\mathcal{P}_i^*, \mathcal{P}_v)$ sera appelée erreur d'estimation.

Nous chercherons en particulier à construire des suites d'estimateurs $\{\mathcal{P}_i\}_{i=t_o}^{t}$ à partir de la suite $\{x_i\}_{i=t_o}^{t}$ telles que l'erreur d'estimation diminue lorsque t augmente. Nous l'appellerons : processus d'autoapprentissage.

I-3. PARAMETRISATION D'UNE PARTITION $\mathcal{P}$ DE Ω

a) Canonique : Une partition $\mathcal{P}$ induit une règle de décision ou de reconnaissance $\mathcal{R}$ qui est une application de Ω sur $\mathcal{P}$

$$\mathcal{R}: \Omega \longrightarrow \mathcal{P}$$

où

$$\mathcal{R}(x) = C_\alpha \iff x \in C_\alpha$$

L'équivalence E_q associée canoniquement à $\mathcal{P}$ s'exprime alors par :

$$x \; E_q \; y \iff \mathcal{R}(x) = \mathcal{R}(y)$$

Définition : Soit $\underline{\underline{=}} = \left\{ \chi_{C_\alpha}(\cdot) \right\}$ l'ensemble des fonctions indicatrices des classes $C_\alpha \in \mathcal{P}$, on constate qu'il y a équivalence entre se donner la partition $\mathcal{P}$ ou l'ensemble $\underline{\underline{=}}$ associé à la règle de décision suivante

$$\mathcal{R}(x) = C_\alpha$$

avec α tel que

$$\chi_{C_\alpha} = \max_1 \left[\chi_{C_1}(x) \right]$$

Nous appellerons $\underline{\underline{=}}$ paramétrisation de $\mathcal{P}$, c'est la paramétrisation canonique, et nous noterons cette règle $\mathcal{R}\left(x \mid \underline{\underline{=}}\right)$

b) Paramétrisation floue

Si nous remplaçons les fonctions caractéristiques $\chi_{C_\alpha}(x) \in \{0,1\}$ par des fonctions d'appartenances continues $\mu_{C_\alpha}(x) \in [0,1]$ et si nous appelons $M = \left\{ \mu_{C_\alpha}(\cdot) \right\}_{C_\alpha \in \mathcal{P}}$ nous définissons une nouvelle règle de reconnaissance :

$$\mathcal{R}(x \mid M) = C_\alpha$$

avec α tel que

$$\mu_{C_\alpha} = \max_1 \left[\mu_{C_1}(x) \right]$$

Il est clair que l'on peut toujours choisir M tel que la partition $\mathcal{P}_M$ induite par $\mathcal{R}(x|M)$ soit exactement $\mathcal{P}$, en effet, il existe au moins une solution qui consiste à faire $M = \underline{\underline{=}}$.
Inversement toute partition $\mathcal{P}_M$ induite par une règle de reconnaissance $\mathcal{R}(\cdot \mid M)$ possède une paramétrisation canonique.

c) Paramétrisation par noyaux

En vertu du principa d'oubli évoqué en introduction nous

recherchons pour chaque classe $C_\alpha \in \mathcal{P}$ un nombre n_α fini de nombres réels que nous rangerons sous forme de vecteur $\theta_\alpha \in \mathbb{R}^{n_\alpha}$ et tels que la classe C_α puisse être caractérisée entièrement par eux vis à vis de la règle de reconnaissance $\mathcal{R}$.

Appelons $\mathcal{M} = \{\theta_\alpha\}_{C_\alpha \in \mathcal{P}}$ l'ensemble des paramètres correspondant à toutes et chacune des classes $C_\alpha \in \mathcal{P}$, $\mathcal{M}$ est donc un ensemble fini de nombres réels, que nous appellerons ensemble des noyaux (la notion de noyau a été introduite en classification par E. DIDAY dans [1.2]).

On notera $\mathcal{R}(x|\mathcal{M})$ une règle de décision ou de reconnaissance dépendant uniquement de x et de l'ensemble $\mathcal{M}$.

Il est clair que si l'ensemble M de la paramétrisation floue peut être caractérisé de façon complète et équivalente par un ensemble de nombres réels $\mathcal{M}$, et que $\mathcal{M}$ peut être partionné en $\{\theta_\alpha\}_{C_\alpha \in \mathcal{P}}$ tel que θ_α détermine uniquement la fonction d'appartenance $\mu_{C_\alpha}(\cdot)$, alors $\mathcal{R}(x|\mathcal{M})$ est équivalent à $\mathcal{R}(x|M)$.

Définition : Si pour $\mathcal{P}$ donné nous savons que $\mathcal{R}(\cdot|M)$ est équivalent à $\mathcal{R}(\cdot|\mathcal{M})$ alors on dira que $\mathcal{P}$ est paramétrisable par noyaux.

Proposition :

Une partition $\mathcal{P}$ de Ω paramétrisable par noyaux est équivalente au doublet suivant :

$$\{\mathcal{M}, \mathcal{R}(\cdot|\mathcal{M})\}$$

où $\mathcal{M}$ est l'ensemble des noyaux et $\mathcal{R}(\cdot|\mathcal{M})$ est une règle de décision.

Démonstration : En effet, toute partition, nous l'avons vu, est équivalente au doublet

$$\{\Xi, \mathcal{R}(\cdot|\Xi)\}$$

il existe donc au moins une paramétrisation floue M telle que $\mathcal{P}$ est équivalente a

$$\{M, \mathcal{R}(\cdot|M)\}$$

et puisque nous avons posé que $\mathcal{R}(\cdot|M)$ est équivalent à $\mathcal{R}(\cdot|\textcircled{M})$ car il est paramétrisable, alors on déduit que $\mathcal{P}$ est équivalent à $\{\textcircled{M}|\mathcal{R}(\cdot|\textcircled{M})\}$

-:-

REFERENCES

/-1.1-/ BOURBAKI N.
"Eléments de Mathématiques"
Hermann

/-1.2-/ DIDAY E. (1972)
"Nouvelles méthodes et nouveaux concepts en classification automatique et reconnaissance des formes"
Thèse d'Etat ès Sciences Mathématiques, Université Paris VI.

/-1.3-/ MINSKY M. (1961)
"Steps toward artificial intelligence"
Proceedings of the IRE, January.

/-1.4-/ PARRY W. (1969)
"Entropy and generators in ergodic theory"
W.A. Benjamin, Inc.

/-1.5-/ SAMON J.W. (1970)
"Interactive pattern analysis and classification"
IEEE Trans. on Computers, Vol. C.19, n° 7, July.

/-1.6-/ SKLANSKY J. (1966)
"Learning systems for automatic control"
IEEE Trans on Aut. Control, AC-11, n° 1, January

/-1.7-/ TOMASSONE R. (1970)
"Analyse multidimensionnelle et classification"
Revue de Statistique Appliquée, vol. XVIII, n° 4.

/-1.8-/ TOU J.T. et GONZALEZ R.C. (1974)
"Pattern recognition principles"
Addison Wesley Pullishing company, Inc.

/-1.9-/ WIENER N. (1948)
"Cybernetics"
The M.I.T. Press.

/-1.10-/ ZADEH L.A. (1965)
"Fuzzy sets"
Information and Control, vol. 8, June.

CHAPITRE II

METRIQUES ENTRE PARTITIONS BASEES SUR LA THEORIE DE L'INFORMATION

-:-

II-1. INTRODUCTION

Nous allons présenter dans ce chapitre plusieurs indices de dissemblance entre partitions et on montrera que certains remplissent les axiomes de la distance.

Tous sont inspirés par la théorie mathématique de l'Information de Claude E. Shanon (1948) et, par l'extension de cette théorie, de la part du mathématicien soviétique A. Ja. Khintchine (1953), en utilisant la notion d'entropie dans la théorie des probabilités.

Ces indices de dissenblance vont nous apporter une aide importante pour analyser les résultats de nos algorithmes de classement, puisque ils permettront de comparer les résultats des algorithmes entre eux, ou avec un classement "naturel" ou "vrai" donné a priori.

II-2. RAPPELS SUR LA THEORIE DE L'INFORMATION [2.1]

II-2-1) Information apportée par une partition

Soit $(\Omega, \mathcal{A}, P)$ un espace de probabilité, et A un évènement de la σ-algèbre $\mathcal{A}$

On va donner un sens mathématique à la phrase "dire que ω appartient à A, c'est fournir une information sur ω".

Il faudra donc définir un nombre H(A) qui sera l'information apportée par l'assertion $\omega \in A$

Supposons H(A) déjà défini pour toute partie A de $\mathcal{A}$ ayant une probabilité non nulle, et étendons la notion

d'information au système complet d'évènements de probabilité non nulle, c'est-à-dire aux partitions finies vérifiant les deux conditions suivantes /⁻2.6_⁻/

. Soit (A_n) une famille fini d'éléments de $\mathcal{A}$ alors :

1) $\bigcup_n A_n = \Omega$

2) $A_n \cap A_m = \emptyset \qquad \forall\, n \neq m$

Maintenant, soit $\mathcal{P}_c$ une partition finie de Ω en parties mesurables C_i.

Pour chaque $\omega \in \Omega$, soit $C_i(\omega)$ la partie de $\mathcal{P}_c$ qui contient ω (elle est unique).

Notons par $H_{\mathcal{P}_c}$ la variable aléatoire telle que

$$H_{\mathcal{P}_c}(\omega) = H(C_i(\omega))$$

On peut construire le diagramme d'applications suivant :

$$\begin{array}{ccccc} \Omega & \longrightarrow & \mathcal{P}_c & \longrightarrow & R \\ (\omega) & & (C_i(\omega)) & & (H(C_i(\omega))) \end{array}$$

$H_{\mathcal{P}_c}$: $(\omega) \longmapsto (H(C_i(\omega)))$

Maintenant on associera à la partition $\mathcal{P}_c$ une quantité d'information qui sera l'espérance mathématique de la variable $H_{\mathcal{P}_c}$.

On écrira $$\bar{H}(\mathcal{P}_c) = \mathbb{E}(H_{\mathcal{P}_c})$$

c'est-à-dire :

$$\bar{H}(\mathcal{P}_c) = \sum_i \left\{ P(C_i)\, H(C_i) \;\middle|\; C_i(\omega) \in \mathcal{P}_c \right\}$$

L'information apportée par l'assertion $\omega \in C_i$ reçoit ici un poids P_i égal à la probabilité qu'a cette assertion d'être vraie.

Rappelons ce que sont deux partitions indépendantes. Soient $\mathcal{P}_c$ et $\mathcal{P}_F$ deux partitions de Ω ; on dit que les partitions $\mathcal{P}_c$ et $\mathcal{P}_F$ sont indépendantes si $\forall C_i \in \mathcal{P}_c$, $\forall F_j \in \mathcal{P}_F$ on a

$$P(C_i \cap F_j) = P(C_i) \cdot P(F_j)$$

en langage de la théorie de l'information on dit que l'assertion $\omega \in C_i$ ne nous apporte aucune information sur la partie $F_j(\omega)$ en ce sens que la probabilité relative pour que ω qui est dans C_i, soit aussi dans F_j ne diffère pas de la probabilité $P(F_j)$ pour qu'un élément, indéterminé a priori soit dans F_j :

$$P(\omega \in F_j \mid \omega \in C_i) = P(\omega \in F_j)$$

L'information H(A) est caractérisée par les trois conditions suivantes :

1) H(A) est une fonction réelle positive décroissante de la seule probabilité : $H(A) = f(P(A))$

2) Si les partitions $\mathcal{P}_C$ et $\mathcal{P}_F$ sont indépendantes on a :

$$\bar{H}(\mathcal{P}_C \times \mathcal{P}_F) = \bar{H}(\mathcal{P}_C) + \bar{H}(\mathcal{P}_F)$$

où on note $$\mathcal{P}_C \times \mathcal{P}_F = \left\{ C_i \cap F_j \mid C_i \in \mathcal{P}_C , F_j \in \mathcal{P}_F \right\}$$

avec $P(C_i \cap F_j) = 0$

3) Une assertion aléatoire dont la probabilité d'être vraie est 1/2, apporte une information unité : $f(1/2) = 1$.

Que f soit une fonction décroissante s'impose parce que plus A est petit plus ω est localisé ($\omega \in A$).

On suppose aussi $f(1) = 0$, car dire que $\omega \in \Omega \left(P(\Omega) = 1\right)$ équivaut à ne donner aucune information.

L'additivité des informations apportées par des partitions indépendantes correspond à ce que nous avons rappelé plus haut c'est-à-dire si $C_i(\omega)$ et $F_j(\omega)$ sont indépendants, affirmer $\omega \in C_i$, nous laisse quant à $F_j(\omega)$ dans la même expectative que

si l'on admettait seulement $\omega \in \Omega$

La condition $f(1/2) = 1$ est une condition de normalisation indispensable, puisque si une fonction H satisfait aux conditions 1) et 2), la fonction $KH (K>0)$ y satisfait aussi.

Rappelons maintenant comment des trois conditions postulées on déduit que :

$$\forall p \in]0,1[\quad , \quad f(p) = -\log_2 (p)$$

Remarque : Supposons que toutes les partitions de Ω utilisées dans les démonstrations existent.

Soit $\mathcal{P}_{C_m}$ une partition en m classes d'égales probabilités.

On a d'après 2)

$$\overline{H}(\mathcal{P}_{C_m}) = m \times \left(\frac{1}{m}\right) \times f\left(\frac{1}{m}\right) = f\left(\frac{1}{m}\right)$$

Soient donc $\mathcal{P}_{C_m}$ et $\mathcal{P}_{F_p}$ deux partitions indépendantes en classes d'égales probabilités ; on a

$$\overline{H}\left(\mathcal{P}_{C_m} \times \mathcal{P}_{F_p}\right) = f\left(\frac{1}{m}\right) + f\left(\frac{1}{p}\right) = f\left(\frac{1}{mp}\right)$$

De cette relation et de la condition de normalisation 3), on déduit que :

$$\forall k \in \mathbb{N} \; ; \; f\left(\frac{1}{2^k}\right) = k = -\log_2\left(\frac{1}{2^k}\right)$$

plus généralement soit Z un entier quelconque, on peut donner de $\log_2(Z)$ une approximation arbitrairement précise, en fraction rationnelle

$$\left(\frac{p}{q}\right) \leq \log_2(z) \leq \left(\frac{p+1}{q}\right)$$

d'où l'on en déduit $\quad 2^p \leq z^q \leq 2^{p+1}$

et puisque f est décroissante :

$$p \leq f\left(\frac{1}{z^q}\right) = q f\left(\frac{1}{z}\right) \leq p+1$$

ce qui suffit à établir que, quelque soit l'entier Z,

$$f\left(\frac{1}{Z}\right) = -\log_2\left(\frac{1}{Z}\right)$$

Reste à montrer que pour tout p rationnel $f(p) = -\log_2 p$; de là on passe aisèment à toute valeur réelle car la monotonicité entraîne ici la continuité.

Soit donc $p = \frac{m}{m+n}$; soit $\mathcal{P}_G$ une partition de Ω en n+1 classes, l'une de probabilité p et les n autres de probabilité $\frac{1}{m+n}$, on a : $\bar{H}(\mathcal{P}_G) = p f(p) + (1-p) f\left(\frac{1}{m+n}\right) =$

$$= p f(p) + (1-p)\log_2(m+n)$$

soit maintenant $\mathcal{P}_{C_m}$ une partition de Ω en m classes d'égale probabilité, que l'on suppose indépendante de $\mathcal{P}_G$; on a:

$$\bar{H}(\mathcal{P}_G \times \mathcal{P}_{C_m}) = p f\left(\frac{1}{m+n}\right) + (1-p) f\left(\frac{1}{(m+n)m}\right)$$

(la classe de $\mathcal{P}_G$ dont la mesure est p est divisée par $\mathcal{P}_{C_m}$ en m classes de mesure $\frac{1}{m+n}$, ...)

Mais on a aussi d'après l'hypothèse 2) :

$$\bar{H}(\mathcal{P}_G \times \mathcal{P}_{C_m}) = \bar{H}(\mathcal{P}_G) + \bar{H}(\mathcal{P}_{C_m})$$

d'où l'égalité :

$$f(p) = f\left(\frac{1}{m+n}\right) - f\left(\frac{1}{m}\right) = -\log_2 p$$

A ce moment là on peut écrire :

$$\bar{H}(\mathcal{P}_C) = -\sum_i \left\{ p(C_i)\log_2 p(C_i) \;\middle|\; C_i \in \mathcal{P}_C \right\}$$

et on l'appellera information moyenne apportée par la partition $\mathcal{P}_C$.

II-2-2) Information mutuelle et Information conditionnelle [2.4]

Considérons deux partitions de l'ensemble Ω :

La partition $\mathcal{P}_C$ (dont les classes sont notées C_1, C_2, C_3, ..., C_i, ..., C_p) et la partition $\mathcal{P}_F$ (dont les classes sont notées F_1, F_2, ..., F_j, ..., F_r).

Nous noterons de la façon suivante les probabilités correspondantes à chacune des classes :

$$P_i = p(C_i) \qquad (2.1)$$

$$P_j = p(F_j) \qquad (2.2)$$

$$P_{ij} = P(C_i \cap F_j) \qquad (2.3)$$

il est évident que l'on a les relations:

$$P_i = \sum_j P_{ij} \qquad (2.4)$$

$$P_j = \sum_i P_{ij} \qquad (2.5)$$

$$\sum_{ij} P_{ij} = 1 \qquad (2.6)$$

L'information moyenne apportée par $\mathcal{P}_C$ vaut:

$$\bar{H}(\mathcal{P}_C) = -\sum_i P_i \log_2 P_i \qquad (2.7)$$

de même :

$$\bar{H}(\mathcal{P}_F) = -\sum_j P_j \log_2 P_j \qquad (2.8)$$

L'information moyenne relative à la connaissance simultanée de $\mathcal{P}_C$ et $\mathcal{P}_F$ vaut :

$$\bar{H}(\mathcal{P}_C \times \mathcal{P}_F) = -\sum_{ij} P_{ij} \log_2 P_{ij}$$

et elle est appelée : Information mutuelle.

Si l'information telle qu'on la définit en mathématique, correspond bien à l'idée qu'on se fait de la notion d'information, on doit avoir :

$$\bar{H}(\mathcal{P}_C \times \mathcal{P}_F) \leq \bar{H}(\mathcal{P}_C) + \bar{H}(\mathcal{P}_F) \qquad (2.10)$$

En effet, mathématiquement, on a bien l'inégalité ci-dessus, que les propriétés de concavité de la fonction x log x permettent de démontrer /̄2.4_7̄ et /̄2.9_7̄.

La quantité

$$\bar{H}(\mathcal{P}_F/\mathcal{P}_C) = \bar{H}(\mathcal{P}_C \times \mathcal{P}_F) - \bar{H}(\mathcal{P}_C) \qquad (2.11)$$

est appelée information de $\mathcal{P}_F$ conditionnée par $\mathcal{P}_C$ ou information supplémentaire apportée par $\mathcal{P}_F$ quand $\mathcal{P}_C$ est réalisée.

Remarque : D'après (2.10), on a:

$$\bar{H}(\mathcal{P}_F/\mathcal{P}_C) \leq \bar{H}(\mathcal{P}_F) \qquad (2.12)$$

vérifions que cette définition est cohérente avec la définition des probabilités conditionnelles :

$$P(F_j/C_i) = P(F_j \cap C_i) / P(C_i) = P_{ij}/P_i \qquad (2.13)$$

Donc :

$$\bar{H}(\mathcal{P}_F/C_i) = -\sum_j \left(\frac{P_{ij}}{P_i}\right) \log_2\left(\frac{P_{ij}}{P_i}\right) \qquad (2.14)$$

L'information supplémentaire conditionnelle apportée par $\mathcal{P}_F$ à partir de $\mathcal{P}_C$ comme n'est autre que l'espérance mathématique connue par $\mathcal{P}_F$ pour chaque valeur particulière C_i de $\mathcal{P}_C$, c'est-à-dire l'espérance mathématique de la relation (2.14).

En effet :

$$\sum_i P_i \bar{H}(\mathcal{P}_F/C_i) = -\sum_i \sum_j P_{ij} \log_2\left(\frac{P_{ij}}{P_i}\right) =$$

$$= -\sum_i \sum_j P_{ij}\left(\log_2 P_{ij} - \log_2 P_i\right) = \bar{H}(\mathcal{P}_F \times \mathcal{P}_C) - \bar{H}(\mathcal{P}_C)$$

Donc :

$$\bar{H}\left(\mathcal{P}_F \mid \mathcal{P}_C\right) = -\sum_i \sum_j P_{ij} \log_2 \left(\frac{P_{ij}}{P_i}\right) \qquad (2.15)$$

II-3. L'INFORMATION CONDITIONNELLE APPLIQUEE A LA COMPARAISON ENTRE DEUX CLASSEMENTS : INDICE DE DISSEMBLANCE [2.5]

Un classement est une partition de l'ensemble Ω des observation à classer.

Exemple : soit $\Omega = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9\}$ et soit le classement représenté par la partition $\mathcal{P}_F = \{F_1, F_2, F_3, F_4\}$ de l'ensemble Ω , où les classes sont formées des éléments :

$F_1 = \{x_1, x_2\}$ $\quad F_2 = \{x_3, x_4, x_5, x_6\}$ $\quad F_3 = \{x_7, x_8\}$

$F_4 = \{x_9\}$

et soit un autre classement du même ensemble Ω représenté par

$\mathcal{P}_C = \{C_1, C_2, C_3, C_4, C_5, C_6\}$ où les classes sont formées des éléments :

$C_1 = \{x_9\}$ $\quad C_2 = \{x_8, x_7\}$ $\quad C_3 = \{x_3, x_4\}$ $\quad C_4 = \{x_5, x_6\}$

$C_5 = \{x_2\}$ $\quad C_6 = \{x_1\}$

Formons la matrice M(F/C), qu'on appellera matrice de contingence, de la façon suivante :

$$P_{ij} = \frac{\text{card}\left(F_j \cap C_i\right)}{\text{card}\left(\Omega\right)}$$

c'est-à-dire que P_{ij} est une estimation de la probabilité $P(C_i \cap F_j)$.

Dans notre cas on a la matrice :

$\mathcal{P}_C$ \ $\mathcal{P}_F$	F_1	F_2	F_3	F_4
C_1	0	0	0	1/9
C_2	0	0	2/9	0
C_3	0	2/9	0	0
C_4	0	2/9	0	0
C_5	1/9	0	0	0
C_6	1/9	0	0	0

on voit bien que l'on a la relation : $\sum_{i,j} P_{ij} = 1$

Remarquons que $M(C/F) = M^T(F/C)$

Remarque : Supposons que l'on ait un autre classement du même ensemble Ω représenté par la partition :

$$\mathcal{P}_G = \{G_1, G_2, G_3, G_4\}$$

où

$G_1 = \{x_7, x_8\} = F_3$

$G_2 = \{x_9\} = F_4$

$G_3 = \{x_3, x_4, x_5, x_6\} = F_2$

$G_4 = \{x_1, x_2\} = F_1$

Formons la matrice M(G/F) :

avec $P_{ij} = \dfrac{\text{card}(F_i \cap G_j)}{\text{card}(\Omega)}$

on a donc :

$\mathcal{P}_G$

$\mathcal{P}_F$: F_i \ G_j	G_1	G_2	G_3	G_4
F_1	0	0	0	2/9
F_2	0	0	4/9	0
F_3	2/9	0	0	0
F_4	0	1/9	0	0

on en déduit la propriété suivante :

Propriété : Si deux classements (partitions) sont égaux modulo une permutation d'indices, la matrice formée est carrée et c'est une matrice qui a un seul élément différent de zéro par ligne et par colonne; cette matrice sera appelée matrice "quasi-diagonale".

Remarque : Dans la comparaison entre $\mathcal{P}_C$ et $\mathcal{P}_F$ on constate que M(F/C) a la propriété suivante :

En ajoutant les lignes i=3 et i=4 ainsi que les lignes i=5 et i=6 on obtient une matrice "quasi-diagonale", cette matrice sera appelée "quasi-diagonalisable".

Définition : Une partition (classement) $\mathcal{P}_F$ comprenant moins de classes qu'une autre partition (classement) $\mathcal{P}_C$ est compatible avec $\mathcal{P}_C$, si et seulement si $\mathcal{P}_C$ est une partition plus fine que $\mathcal{P}_F$; on dit aussi que les partitions sont égales "modulo zéro" /¯2.7_/¯.

Dans ce cas la matrice formée, ayant ses lignes caractérisées par $\mathcal{P}_C$ et ses colonnes par $\mathcal{P}_F$, est

"quasi-diagonalisable".

Réciproque : Si une matrice est "quasi-diagonalisable" la partition $\mathcal{P}_F$ qui a moins de classes, est compatible avec la partition $\mathcal{P}_C$.

De ces propriétés on déduit que :

- Deux partitions sont égales modulo une permutation d'indices si la matrice qu'elles forment est "quasi-diagonale"
- Deux partitions sont compatibles si la matrice formée est "quasi-diagonalisable", dans ce cas la partition qui a le plus grand nombre de classes est plus fine.

Il s'agit maintenant de définir une mesure de la différence entre deux classements (partitions), c'est-à-dire une mesure sur l'ensemble des matrices de contingence; cette mesure doit être nulle si et seulement si, la matrice est "quasi-diagonale" ou bien "quasi-diagonalisable", c'est-à-dire quand les partitions sont soit égales soit compatibles ou égales (mod 0).

Nous proposons la mesure donnée par l'information conditionnelle.

On a : $\underset{(M(\cdot/\cdot))}{\mathcal{M}} \longrightarrow \underset{(r)}{R^+}$

où $\mathcal{M}$ est l'ensemble des matrices de contingence. On peut alors construire :

$$\bar{H}\left(\mathcal{P}_F/\mathcal{P}_C\right) = -\sum_i \sum_j P_{ij} \log_2 \left(\frac{P_{ij}}{P_i}\right) \qquad (2.16)$$

d'après la relation (2.15).

Nous appellerons cette mesure : Indice de dissemblance.

II-3-1) Propriétés métriques de l'indice de dissemblance /̄2.8_7

Nous allons montrer que la relation (2.16) que nous noterons $I_d(\mathcal{P}_F, \mathcal{P}_C)$ possède les propriétés métriques suivantes dans l'ensemble des partitions de Ω qu'on notera $\mathcal{P}_\Omega$

D1 (positivité):

$$\forall \mathcal{P}_F \in \mathcal{P}_\Omega \quad \text{et} \quad \forall \mathcal{P}_C \in \mathcal{P}_\Omega \quad \text{nous avons}$$

$$I_d(\mathcal{P}_F, \mathcal{P}_C) \geqslant 0 \qquad (2.17)$$

en effet, puisque $P_{ij} \leqslant P_i$ on a bien $I_d(\mathcal{P}_F, \mathcal{P}_C) \geqslant 0$

D2 (annulation):

$$\forall \mathcal{P}_F \in \mathcal{P}_\Omega \text{ et } \forall \mathcal{P}_C \in \mathcal{P}_\Omega \quad \text{alors :}$$

$$\mathcal{P}_F = \mathcal{P}_C \text{ ou } \mathcal{P}_C \subset \mathcal{P}_F \iff I_d(\mathcal{P}_F, \mathcal{P}_C) = 0 \qquad (2.18)$$

En effet, montrons d'abord que $\mathcal{P}_F = \mathcal{P}_C$ ou $\mathcal{P}_C \subset \mathcal{P}_F \Longrightarrow I_d(\mathcal{P}_F, \mathcal{P}_C) = 0$

1) $\mathcal{P}_F = \mathcal{P}_C \Longrightarrow \forall i \ (1 \leqslant i \leqslant p), \ \exists j \ (1 \leqslant j \leqslant r)$

avec p=r tels que $C_i = F_j$ avec $C_i \in \mathcal{P}_C$ et $F_j \in \mathcal{P}_F$, mais cela veut dire que $P(C_i \cap F_j) = P(C_i)$ donc $P_{ij} = P_i \ \forall i$ d'où : $I_d(\mathcal{P}_F, \mathcal{P}_C) = 0$

2) $\mathcal{P}_C \subset \mathcal{P}_F$, cela veut dire que la partition $\mathcal{P}_C$ est plus fine que la partition $\mathcal{P}_F$ c'est-à-dire que toute partie $F_j \in \mathcal{P}_F$ est union de parties de $\mathcal{P}_C$ dans ce cas on a donc $P(C_i \cap F_j) = P(C_i)$ d'où $I_d(\mathcal{P}_F, \mathcal{P}_C) = 0$

Maintenant, montrons la réciproque, c'est-à-dire que:

$I_d(\mathcal{P}_F, \mathcal{P}_C) = 0 \Longrightarrow \mathcal{P}_F = \mathcal{P}_C$ ou $\mathcal{P}_C \subset \mathcal{P}_F$

$$I_d(\mathcal{P}_F, \mathcal{P}_C) = 0 \Longrightarrow \begin{cases} P_{ij} = 0 \\ P_{ij} = P_i \end{cases}$$

mais puisque $\sum_j P_{ij} = P_i$ nous savons que :

$\forall i$, $\exists j$ tel que $P_{ij}=P_i$

Donc $\forall i$, $\exists j$ tel que $P(C_i \cap F_j) = P(C_i)$

Ceci implique que: $C_i \cap F_j = C_i$

c'est-à-dire $C_i \subset F_j$

alors, finalement nous avons : $\mathcal{P}_C \subseteq \mathcal{P}_F$

D3 (absence de symétrie)

En général nous avons :

$$I_d(\mathcal{P}_F, \mathcal{P}_C) \neq I_d(\mathcal{P}_C, \mathcal{P}_F) \qquad (2.19)$$

c'est évident puisque :

$$I_d(\mathcal{P}_F, \mathcal{P}_C) = \bar{H}(\mathcal{P}_F/\mathcal{P}_C) = -\sum_i \sum_j P_{ij} \log_2 \left(\frac{P_{ij}}{P_i}\right)$$

et

$$I_d(\mathcal{P}_C, \mathcal{P}_F) = \bar{H}(\mathcal{P}_C/\mathcal{P}_F) = -\sum_i \sum_j P_{ij} \log_2 \left(\frac{P_{ij}}{P_j}\right)$$

et en général $P_i \neq P_j$, alors :

$$I_d(\mathcal{P}_F, \mathcal{P}_C) \neq I_d(\mathcal{P}_C, \mathcal{P}_F)$$

D4 (Inégalité triangulaire) :

$\forall \mathcal{P}_F \in \mathcal{P}_\Omega$, $\forall \mathcal{P}_C \in \mathcal{P}_\Omega$, $\forall \mathcal{P}_G \in \mathcal{P}_\Omega$ nous avons

$$I_d(\mathcal{P}_F, \mathcal{P}_C) + I_d(\mathcal{P}_C, \mathcal{P}_G) \geqslant I_d(\mathcal{P}_F, \mathcal{P}_G) \qquad (2.20)$$

En effet, d'après (2.12) on peut écrire :

$$\bar{H}(\mathcal{P}_F/\mathcal{P}_C) + \bar{H}(\mathcal{P}_C/\mathcal{P}_G) \geqslant \bar{H}(\mathcal{P}_F/\mathcal{P}_C, \mathcal{P}_G) + \bar{H}(\mathcal{P}_C/\mathcal{P}_G) \qquad (2.21)$$

suivant (2.11) nous avons :

$$\bar{H}(\mathcal{P}_F/\mathcal{P}_C, \mathcal{P}_G) + \bar{H}(\mathcal{P}_C/\mathcal{P}_G) = \bar{H}(\mathcal{P}_F, \mathcal{P}_C/\mathcal{P}_G) \qquad (2.22)$$

mais puisque on sait que:

$$\bar{H}(\mathcal{P}_F, \mathcal{P}_C) \geqslant \bar{H}(\mathcal{P}_F)$$

nous pouvons écrire:

$$\bar{H}(\mathcal{P}_F, \mathcal{P}_C/\mathcal{P}_G) \geqslant \bar{H}(\mathcal{P}_F/\mathcal{P}_G) \qquad (2.23)$$

finalement en combinant (2.21), (2.22) et (2.23) on a :

$$\bar{H}(\mathcal{P}_F/\mathcal{P}_C)+\bar{H}(\mathcal{P}_C/\mathcal{P}_G)\geqslant\bar{H}(\mathcal{P}_F/\mathcal{P}_G) \qquad (2.24)$$

II-3-2) Normalisation de l'indice de dissemblance

Definissons $$I'_d(\mathcal{P}_F,\mathcal{P}_C)=\frac{\bar{H}(\mathcal{P}_F/\mathcal{P}_C)}{\bar{H}(\mathcal{P}_F,\mathcal{P}_C)} \qquad (2.25)$$

de telle façon que: $I'_d(\cdot,\cdot)\in[0,1]$

évidemment ce nouvel indice a les mêmes propriétés métriques que le précédent:

D'1 (positivité) : elle se déduit directement de D1

D'2 (annulation) : elle se déduit directement de D2

D'3 (absence de symétrie) : elle se déduit directement de D3

D'4 (inégalité triangulaire) :

Nous devons montrer que l'inégalité triangulaire est préservée :

D'après (2.11) nous avons :

$$\frac{\bar{H}(\mathcal{P}_F/\mathcal{P}_C)}{\bar{H}(\mathcal{P}_F,\mathcal{P}_C)}+\frac{\bar{H}(\mathcal{P}_C/\mathcal{P}_G)}{\bar{H}(\mathcal{P}_C,\mathcal{P}_G)}=\frac{\bar{H}(\mathcal{P}_F/\mathcal{P}_C)}{\bar{H}(\mathcal{P}_F/\mathcal{P}_C)+\bar{H}(\mathcal{P}_C)}+\frac{\bar{H}(\mathcal{P}_C/\mathcal{P}_G)}{\bar{H}(\mathcal{P}_C/\mathcal{P}_G)+\bar{H}(\mathcal{P}_G)}\geqslant$$

$$\geqslant\frac{\bar{H}(\mathcal{P}_F/\mathcal{P}_C)}{\bar{H}(\mathcal{P}_F/\mathcal{P}_C)+\bar{H}(\mathcal{P}_C/\mathcal{P}_G)+\bar{H}(\mathcal{P}_G)}+\frac{\bar{H}(\mathcal{P}_C/\mathcal{P}_G)}{\bar{H}(\mathcal{P}_F/\mathcal{P}_C)+\bar{H}(\mathcal{P}_C/\mathcal{P}_G)+\bar{H}(\mathcal{P}_G)}=$$

$$=\frac{\bar{H}(\mathcal{P}_F/\mathcal{P}_C)+H(\mathcal{P}_C/\mathcal{P}_G)}{\bar{H}(\mathcal{P}_F/\mathcal{P}_C)+\bar{H}(\mathcal{P}_C/\mathcal{P}_G)+\bar{H}(\mathcal{P}_G)}\geqslant\frac{\bar{H}(\mathcal{P}_F/\mathcal{P}_G)}{\bar{H}(\mathcal{P}_F/\mathcal{P}_G)+\bar{H}(\mathcal{P}_G)}=$$

$$=\frac{\bar{H}(\mathcal{P}_F/\mathcal{P}_G)}{\bar{H}(\mathcal{P}_F,\mathcal{P}_G)} \qquad (2.26)$$

donc :

$$I'_d(\mathcal{P}_F,\mathcal{P}_C)+I'_d(\mathcal{P}_C,\mathcal{P}_G)\geqslant I'_d(\mathcal{P}_F,\mathcal{P}_G)$$

II-4. DISTANCE ENTRE PARTITIONS

Nous allons définir maintenant, une mesure sur l'ensemble de matrices de contingence qui soit nulle si et seulement si la matrice est "quasi-diagonale" c'est-à-dire si les partitions sont égales modulo une permutation d'indices, par conséquent si l'une des deux partitions est plus fine que l'autre cette mesure ne doit pas être nulle.

Une telle mesure est directement obtenue en symétrisant l'indice de dissemblance définit dans la relation (2.16), c'est-à-dire :

$$d(\mathcal{P}_F,\mathcal{P}_C) = \bar{H}(\mathcal{P}_F/\mathcal{P}_C) + \bar{H}(\mathcal{P}_C/\mathcal{P}_F) \qquad (2.27)$$

nous allons montrer qu'elle satisfait les axiomes d'une distance

D"1 (positivité) : elle se déduit directement de D1

D"2 (annulation) :

$$\forall \mathcal{P}_F \in \mathcal{P}_\Omega \quad \text{et} \quad \forall \mathcal{P}_C \in \mathcal{P}_\Omega \;,\; \mathcal{P}_F = \mathcal{P}_C \iff d(\mathcal{P}_F,\mathcal{P}_C) = 0$$

en effet : $d(\mathcal{P}_F,\mathcal{P}_C) = 0 \iff \bar{H}(\mathcal{P}_F/\mathcal{P}_C) = \bar{H}(\mathcal{P}_C/\mathcal{P}_F) = 0$

mais $\bar{H}(\mathcal{P}_F/\mathcal{P}_C) = 0 \iff \mathcal{P}_C \subseteq \mathcal{P}_F$

et $\bar{H}(\mathcal{P}_C/\mathcal{P}_F) = 0 \iff \mathcal{P}_F \subseteq \mathcal{P}_C$

d'où on a : $\mathcal{P}_C = \mathcal{P}_F$

D"3 (symétrie) : évidente par construction

D"4 (inégalité triangulaire) :

Ecrivons a nouveau (2.24) :

$$\bar{H}(\mathcal{P}_F/\mathcal{P}_C) + \bar{H}(\mathcal{P}_C/\mathcal{P}_G) \geqslant \bar{H}(\mathcal{P}_F/\mathcal{P}_G)$$

en permuttant $\mathcal{P}_F$ avec $\mathcal{P}_G$ nous avons :

$$\bar{H}(\mathcal{P}_G/\mathcal{P}_C) + \bar{H}(\mathcal{P}_C/\mathcal{P}_F) \geqslant \bar{H}(\mathcal{P}_G/\mathcal{P}_F) \qquad (2.28)$$

faisons (2.24) + (2.28) :

$$\bar{H}(\mathcal{P}_F/\mathcal{P}_C) + \bar{H}(\mathcal{P}_C/\mathcal{P}_F) + \bar{H}(\mathcal{P}_C/\mathcal{P}_G) + \bar{H}(\mathcal{P}_G/\mathcal{P}_C) \geqslant$$

$$\geqslant \bar{H}(\mathcal{P}_F/\mathcal{P}_G) + \bar{H}(\mathcal{P}_G/\mathcal{P}_F) \qquad (2.29)$$

d'où le résultat recherché :

$$d(\mathcal{P}_F,\mathcal{P}_C) + d(\mathcal{P}_C,\mathcal{P}_G) \geqslant d(\mathcal{P}_F,\mathcal{P}_G) \qquad (2.30)$$

donc d(.,.) est une distance.

II-4-1) Normalisation de cette distance

De la même façon que nous avons normalisé l'indice de dissemblance défini dans (2.16), nous allons normaliser la distance entre partitions définie par (2.27) pour montrer que l'inégalité triangulaire est préservée, écrivons à nouveau la relation (2.26) :

$$\frac{\bar{H}(\mathcal{P}_F/\mathcal{P}_C)}{\bar{H}(\mathcal{P}_F,\mathcal{P}_C)} + \frac{\bar{H}(\mathcal{P}_C/\mathcal{P}_G)}{\bar{H}(\mathcal{P}_C,\mathcal{P}_G)} \geqslant \frac{\bar{H}(\mathcal{P}_F/\mathcal{P}_G)}{\bar{H}(\mathcal{P}_F,\mathcal{P}_G)}$$

Si dans cette dernière relation on permute $\mathcal{P}_F$ et $\mathcal{P}_G$ on obtient :

$$\frac{\bar{H}(\mathcal{P}_G/\mathcal{P}_C)}{\bar{H}(\mathcal{P}_G,\mathcal{P}_C)} + \frac{\bar{H}(\mathcal{P}_C/\mathcal{P}_F)}{\bar{H}(\mathcal{P}_C,\mathcal{P}_F)} \geqslant \frac{\bar{H}(\mathcal{P}_G/\mathcal{P}_F)}{\bar{H}(\mathcal{P}_G,\mathcal{P}_F)} \qquad (2.31)$$

si on additionne (2.26) et (2.31) on obtient :

$$\frac{\bar{H}(\mathcal{P}_F/\mathcal{P}_C)}{\bar{H}(\mathcal{P}_F,\mathcal{P}_C)} + \frac{\bar{H}(\mathcal{P}_C/\mathcal{P}_F)}{\bar{H}(\mathcal{P}_C,\mathcal{P}_F)} + \frac{\bar{H}(\mathcal{P}_C/\mathcal{P}_G)}{\bar{H}(\mathcal{P}_C,\mathcal{P}_G)} + \frac{\bar{H}(\mathcal{P}_G/\mathcal{P}_C)}{\bar{H}(\mathcal{P}_G,\mathcal{P}_C)} \geqslant$$

$$\geqslant \frac{\bar{H}(\mathcal{P}_F/\mathcal{P}_G)}{\bar{H}(\mathcal{P}_F,\mathcal{P}_G)} + \frac{\bar{H}(\mathcal{P}_G/\mathcal{P}_F)}{\bar{H}(\mathcal{P}_G,\mathcal{P}_F)}$$

c'est-à-dire : $\dfrac{\bar{H}(\mathcal{P}_F/\mathcal{P}_C) + \bar{H}(\mathcal{P}_C/\mathcal{P}_F)}{\bar{H}(\mathcal{P}_C,\mathcal{P}_F)} + \dfrac{\bar{H}(\mathcal{P}_C/\mathcal{P}_G) + \bar{H}(\mathcal{P}_G/\mathcal{P}_C)}{\bar{H}(\mathcal{P}_C,\mathcal{P}_G)} \geqslant$

$$\geqslant \frac{\bar{H}(\mathcal{P}_F/\mathcal{P}_G) + \bar{H}(\mathcal{P}_G/\mathcal{P}_F)}{\bar{H}(\mathcal{P}_F,\mathcal{P}_G)} \qquad (2.32)$$

On notera donc :

$$d'(\mathcal{P}_f, \mathcal{P}_c) = \frac{\bar{H}(\mathcal{P}_f/\mathcal{P}_c) + \bar{H}(\mathcal{P}_c/\mathcal{P}_f)}{\bar{H}(\mathcal{P}_f, \mathcal{P}_c)} \qquad (2.33)$$

d'(.,.) est donc une normalisation de d(.,.), et $d'(.,.) \in [0,1]$ elle est aussi une distance puisque l'inégalité triangulaire est préservée et les axiomes de positivité, d'annulation et de symétrie sont automatiquement satisfaits par construction.

-:-

REFERENCES

/ 2.1 / BENZECRI J.B. et Collaborateurs (1973)
"L'Analyse des données", Dunod.

/ 2.2 / BONET E. (1975)
"Espais de probabilitat finits", Teide.

/ 2.3 / DIDAY E. (1974)
"Recent progress in distance and similarity measures in pattern recognition"
Second International Joint Conference on Pattern Recognition, August 13-15, Lungby - Copenhagen.

/ 2.4 / GUISU S. et THEODORESCU R. (1971)
"Incertitude et Information"
Les Presses de l'Université de Laval, Québec.

/ 2.5 / LOPEZ DE MANTARAS R. et NUALART D. (1976)
"L'information conditionnelle comme mesure de dissemblance entre classifications"
Note interne LAAS 75I16, Mai.

/ 2.6 / NEVEU J. (1970)
"Bases mathématiques du calcul des probabilités",
Masson & Cie.

/ 2.7 / PARRY W. (1969)
"Entropy and generators in ergodic theory"
W.A. Benjamin, INC.

/ 2.8 / RAJSKI C. (1961)
"A metric space of discret probability distributions"
Information and Control 4, pp. 371-377.

/ 2.9 / SHANON C.E. et WEAVER W. (1948)
"The mathematical theory of communication"
The University of Illinois Press.

CHAPITRE III

ALGORITHME DE CLASSEMENT PAR AUTOAPPRENTISSAGE DE VARIABLES QUALITATIVES

-:-

III-1. INTRODUCTION

Dans ce chapitre nous allons présenter un algorithme de classement par autoapprentissage basé sur le critère du maximum de vraisemblance comme règle de décision.

Les probabilités d'appartenance des éléments aux classes seront estimées par comptage pondéré;ceci entraîne l'introduction implicite d'une mesure de resemblance entre variables qualitatives qui doit donner la possibilité de comparer (de manière pas très coûteuse) des groupes de variables qualitatives; ce point correspond à un courant de recherche en analyse des données appelé analyse factorielle des correspondances multiples qui se tourne vers l'analyse canonique généralisée.

Nous allons étudier une application à la reconnaissance tactile de solides de formes géométriques à l'aide,d'une part, des informations angulaires fournies par une pince symétrique et d'autre part à l'aide d'une main anthropomorphe munie de capteurs de pression; avec cette application on prétend seulement illustrer notre algorithme de classement mais sans nous pencher sur le problème de la perception puisque le lien entre les problèmes de perception et celui du classement se situe à un niveau beaucoup plus fondamental que celui abordé ici (codage, sélection de paramètres, représentation, etc ...).

Nous serons en mesure de comparer les résultats obtenus

avec le classement "naturel" ou "vrai", donné a priori, grâce aux métriques entre partitions définies dans le chapitre précédent.

III-2. DESCRIPTION DES FORMES A CLASSER

Les données issues des capteurs se présentent sous forme d'un vecteur à N_p composantes réelles.

Après une séquence d'essais bien choisis nous pouvons donc calculer les valeurs maximum et minimum prises par chaque composante. Ceci nous permet de les normaliser, après quoi nous quantifierons l'intervalle utile en N_m modalités.

A chaque mesure correspond donc un nouveau vecteur à N_p composantes dont chacune ne peut prendre que N_m valeurs.

L'ensemble des possibilités peut ainsi être représenté par un tableau à N_m lignes et N_p colonnes.

Une forme sera alors représentée par un tableau dont une et seulement une case par colonne est occupée.

Par exemple :

Soient :

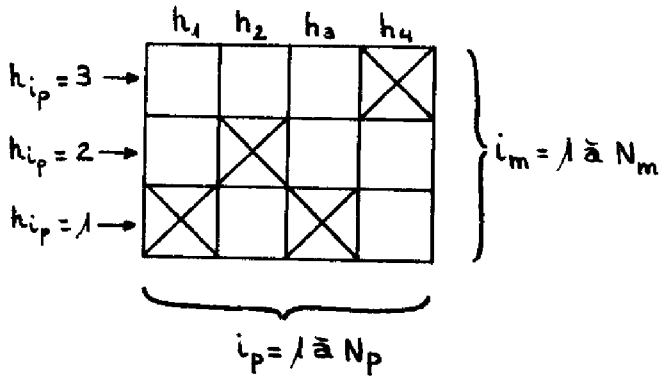
$$N_p = 4 \qquad X = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 3.2 \\ 7.0 \\ 1.3 \\ 4.2 \end{bmatrix}$$

$$X_{max} = \begin{bmatrix} 10.5 \\ 8.2 \\ 2.4 \\ 6.0 \end{bmatrix} \qquad X_{min} = \begin{bmatrix} 0.0 \\ 5.0 \\ 0.2 \\ 0.5 \end{bmatrix}$$

et si $N_M = 3$ nous avons :

$$H = X \text{ (quantifié)} = \begin{bmatrix} 1 \\ 2 \\ 1 \\ 3 \end{bmatrix} = \begin{bmatrix} h_1 \\ h_2 \\ h_3 \\ h_4 \end{bmatrix}$$

et le tableau représentant la forme est :



III-3. CARACTERISATION (PARAMETRISATION) ET AJUSTEMENT DES CLASSES

Une classe C_i sera caractérisée ou paramétrée (voir chapitre I), sous forme de matrice $F_i \in \mathcal{M}$ à $N_m \times N_p$ dimensions, par l'ensemble des probabilités d'occupation de chacune des cases possibles $f_i(i_m, i_p)$. Il est aisé de voir que la somme des probabilités se trouvant sur la même colonne est

$$\sum_{i_m=1}^{N_m} f_i(i_m, i_p) = 1$$

à cause du caractère unique et obligatoire de la case remplie.

III-4. NOTION DE "DISTANCE" D'UN ELEMENT A UNE CLASSE

En faisant l'hypothèse d'indépendance statistique entre les composantes du vecteur X nous pouvons calculer la probabilité pour que la classe C_i caractérisée par la matrice F_i, l'accepte.

$$P_i = P_r[X | X \in C_i] = \prod_{i_p=1}^{N_p} P_r[x_{i_p} | X \in C_i] = \prod_{i_p=1}^{N_p} f_i(h_{i_p}, i_p)$$

et suivant la méthode du maximum de vraisemblance, X sera attribué à la classe C_j telle que (règle de décision $\mathcal{R}$) :

$$P_j = \max_i [P_i]$$

P_i peut aussi être appelée fonction d'appartenance de X à C_i, remarquons que cette fonction d'appartenance est ici calculée à partir de probabilités. L'hypothèse d'indépendance n'est pas nécessaire pour sa définition mais elle est utile pour simplifier le calcul.

Il est possible, sans rien changer au reste du présent raisonnement, de remplacer cette fonction d'appartenance probabiliste par toute autre fonction d'appartenance satisfaisant aux propriétés des sous-ensembles flous.

La règle de décision sera alors appelé : du maximum d'appartenance.

III-5. ESTIMATION PAR AUTOAPPRENTISSAGE DES ELEMENTS DES MATRICES F_i

A partir d'ici nous allons indexer par t l'ensemble de valeurs F_i, P_i et X qui deviennent donc :

$$F_i(t) \ , \quad P_i(t) \ , \quad X(t)$$

ceci est nécessaire pour étudier le caractère évolutif de ce classement, c'est ce caractère évolutif qui nous intéresse en premier lieu.

Lorsqu'un élément X(t) a été attribué à la classe C_i

nous procédons à la modification ou ajustement de la matrice correspondante F_i de façon à tenir compte de cette décision. La loi de modification doit tendre à modifier le critère d'attribution dans le même sens que la dernière décision, elle consistera à incrémenter l'élément $f_i(h_{i_p}, i_p, t)$

Nous choisissons d'effectuer cette incrémentation à l'aide d'un paramètre α de la façon suivante :

$$f_i(h_{i_p}, i_p, t+1) = f_i(h_{i_p}, i_p, t) + \alpha$$

et de renormaliser la colonne i_p de façon à conserver:

$$\sum_{i_m=1}^{N_m} f_i(i_m, i_p, t) = \sum_{i_m=1}^{N_m} f_i(i_m, i_p, t+1) = 1$$

ceci fournit une estimation de la probabilité $Pr\left[x_{i_p}(t) \middle| X(t) \in C_i\right]$ à partir d'un comptage pondéré par α [3.5].

Le paramètre α peut être constant ou fonction de t, de C_i , de h_j et de i_p.

Il détermine la loi d'apprentissage.

Nous verrons plus loin plusieurs choix de α , ainsi que son influence sur les qualités de cette estimation.

III-6. MECANISME DE CROISSANCE DU NOMBRE DE CLASSES

L'attribution de X(t) à une classe selon le critère du maximum d'appartenance donne toujours un résultat. Il faut s'assurer que cette attribution est significative, par exemple qu'elle dépasse un certain seuil.

Pour ceci on prévoit une classe vide $C_{i_{max}}$ pour laquelle :

$$\forall i_m = 1 \text{ à } N_m, \forall i_p = 1 \text{ à } N_p : f_{i_{max}}(i_m, i_p, t) = \frac{1}{N_m}$$

Toutes les positions sont donc, a priori, équiprobables.

Lorsque $P_{i_{max}}$ est tel que $\forall i \quad P_i < P_{i_{max}}$ on attribue X(t) à la classe $C_{i_{max}}$ et celle-ci se trouve donc modifiée par l'incrémentation des éléments $f_{i_{max}}(h_{i_p}, i_p, t)$ de la matrice $F_{i_{max}}(t)$. A ce moment là on crée une nouvelle classe vide $C_{i_{max}+1}$ caractérisée par la matrice $F_{i_{max}+1}(t)$ dont les éléments valent : $f_{i_{max}+1}(i_m, i_p, t) = \frac{1}{N_m}$

On voit donc que l'attribution d'un élément à une classe non vide se fera toujours avec une probabilité (ou appartenance) supérieure à celle de la classe vide (dans le cas présent, supérieure à $\left(\frac{1}{N_m}\right)^{N_p}$)

III-7. INITIALISATION EN ABSENCE D'INFORMATION INITIALE

La présence de ce mécanisme de croissance du nombre de classes permet à l'algorithme de classement de débuter avec uniquement en mémoire une classe vide C_1 caractérisée par la matrice $F_1(t_o)$ telle que ses éléments sont :

$$f_1(i_m, i_p, t_o) = \frac{1}{N_m}$$

La procédure indiquée nous permet donc d'y attribuer le premier élément $X(t_1)$, de modifier $F_1(t_o)$ en $F_1(t_1)$ et de créer une nouvelle classe vide $F_2(t_1)$. On aperçoit à ce moment là que si les probabilités à estimer l'étaient par un comptage non pondéré on aurait :

$$f_i(i_m, i_p, t) = \frac{\text{nombre d'éléments de } C_i \text{ où } h_{i_p} = i_m}{\text{nombre total d'éléments de } C_i}$$

après le premier élément classé on ne pourrait avoir que deux valeurs possibles pour cette expression, soit 0, soit 1, à partir de là il serait donc impossible de produire une évolution du système.

Ceci est évité par :

1°/ L'existence d'une classe vide d'éléments mais ayant des probabilités non nulles d'acceptation (égales à $\frac{1}{N_m}$) que l'on appelle classe équiprobable.

2°/ La pondération par α du comptage des événements favorables /¯3.1_/.

III-8. DIFFERENTES FONCTIONS D'AUTOAPPRENTISSAGE

A) α subjectif

On attribue à α une valeur constante a priori et subjective, le nombre de classes créées dépend de cette valeur de α on constate que le nombre de classes augmente avec l'augmentation de α .

D'autre part l'estimation de la probabilité $P_r[x_{i_p}(t) \mid X(t) \in C_i]$ présente un biais indéterminé qui est fonction de la séquence $\{X(s)\}_{s=0}^{t}$ effectivement observée /¯3.6_/

B) α statistique

Nous avons vu précédemment que l'estimation de la

probabilité d'une case se modifie, en tenant compte de la normalisation, selon la loi :

$$f_i(h_{i_p}, i_p, t+1) = \frac{f_i(h_{i_p}, i_p, t) + \alpha}{1 + \alpha}$$

par ailleur nous avons initialisé la classe vide en faisant :

$$f_{i_{max}}(i_m, i_p, t) = \frac{1}{N_m}$$

Supposons qu'à un instant t une case quelconque de la matrice F_i ait été modifiée n fois sur N échantillons classés, alors nous aurons :

$$f_i(i_m, i_p, t) = \frac{n}{N+N_m}$$

Nous appellerons incrémentation "statistique" celle qui réalise, à partir de ce moment une estimation des $P_r[x_{i_p}(t) | X(t) \in C_i]$ basée sur la fréquence observée des évènements.

Si l'échantillon N+1 ième est tel que $f_i(i_m, i_p, t)$ doit être a nouveau modifiée, nous aurons :

$$f_i(i_m, i_p, t+1) = \frac{n+1}{N+N_m+1} = \frac{n/N+N_m + 1/N+N_m}{1 + 1/N+N_m} =$$

$$= \frac{f_i(i_m, i_p, t) + 1/N+N_m}{1 + 1/N+N_m}$$

Donc nous voyons que dans ce cas α est donné par

$$\alpha = \frac{1}{N+N_m}$$

III-9. ORGANIGRAMME

Cet algorithme a fourni la série de sous-programmes CLMV (Classement par Maximum de Vraisemblance) qui est l'outil de classement et d'apprentissage, il y a plusieurs versions du sous-programme :

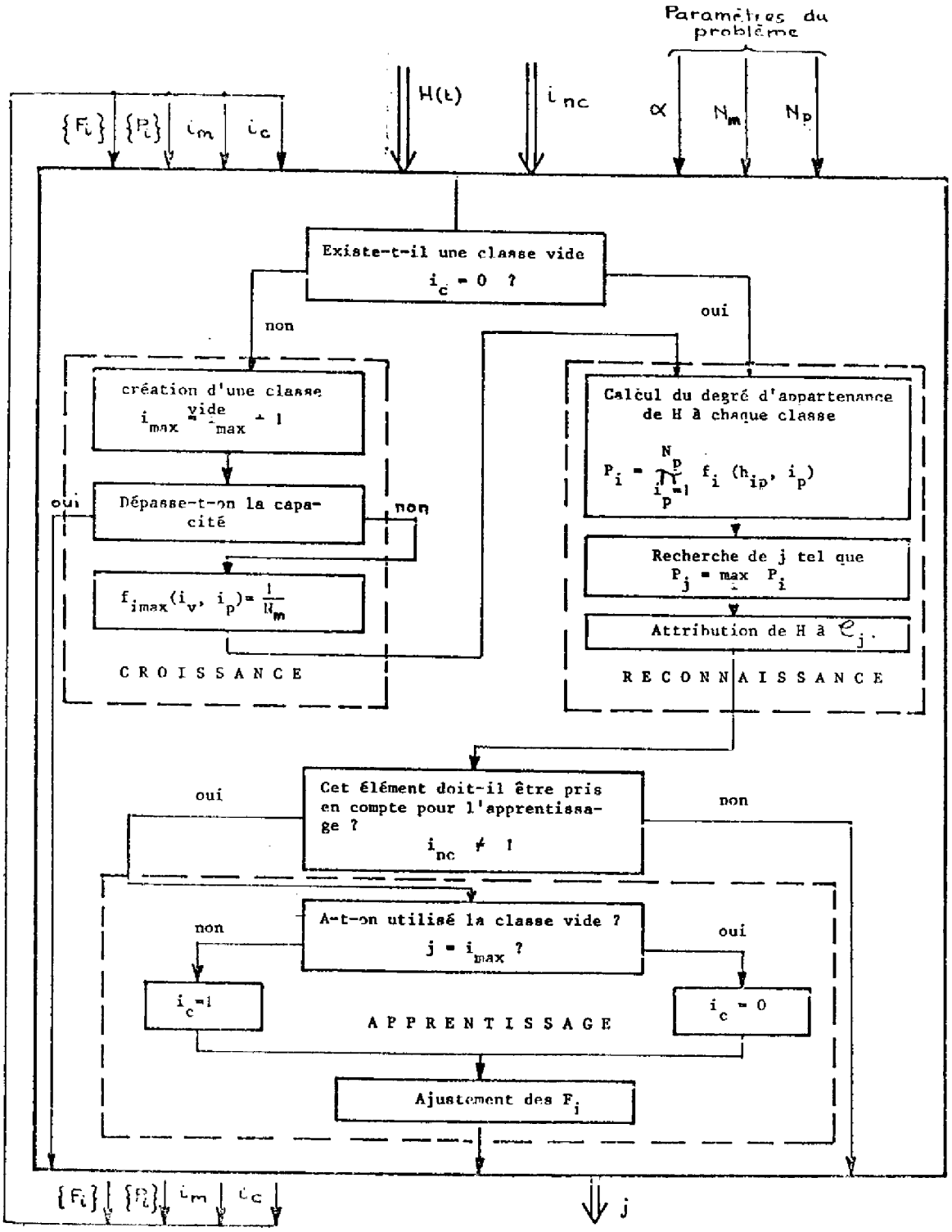
CLMV 1 (α subjectif) : il a comme entrées l'élément H à classer et un indice indiquant s'il doit être utilisé ou non l'apprentissage.

Il fournit en sortie le numéro de la classe à laquelle l'élément a été attribué.

Les matrices F_i, les vecteurs P_i, le nombre de classes existantes i_{max} et l'existence ou la non existence d'une classe vide (i_c=0 ou i_c=1), sont des variables de travail qui entrent et sortent du sous-programme, ou figurent dans une zone "common". Finalement, figurent aussi comme arguments de CLMV 1 les paramètres du problème : coefficient de pondération α (constant), nombre de modalités N_m et nombre de projecteurs N_p.

CLMV 2 (α statistique) : il a les mêmes arguments d'entrée et sortie que CLMV 1 sauf le coefficient de pondération α puisque dans ce cas la fonction d'autoapprentissage est calculée de façon statistique et il est donc nécessaire d'enregistrer à chaque pas le nombre d'échantillons absorbés par chaque classe.

ORGANIGRAMME DU SOUS PROGRAMME CLMV



III-10. APPLICATION A LA RECONNAISSANCE TACTILE DE SOLIDES

La reconnaissance naturelle tactile de formes solides doit utiliser les informations que les capteurs musculaires et sensoriels fournissent au système nerveux central.

Dans cette situation on peut grouper, d'une part les informations de pression et d'autre part les informations angulaires. Il est certain que le système nerveux central n'utilise pas simultanément et au même niveau d'importance l'énorme quantité d'informations qui lui parvient. Il procède à une agrégation, à un classement et à une hiérarchisation de celles-ci.

Dans le cadre des recherches en Robotique, nous nous sommes intéressés à la reconnaissance tactile de solides à l'aide d'organes de préhension artificiels pendant la phase de manipulation profitant ainsi du fait que les objets sont déjà empaumés par l'organe de préhension.

Nous nous proposons d'étudier les possibilités de reconnaissance avec autoapprentissage à l'aide d'une information réduite et partielle.

Ainsi nous procédons à l'analyse de la reconnaissance d'un certain nombre de formes géométriques solides, d'une part à l'aide d'informations uniquement angulaires et fournies par une pince symétrique, d'autre part à l'aide d'une main anthropomorphe munie de capteurs de pression uniquement.

Dans les deux cas nous étudions deux niveaux dans la richesse des informations : 2 ou 4 doigts actifs pour la pince,

3 ou 5 niveaux de pression pour la main.

La situation particulière de notre étude consiste à faire la reconnaissance par autoapprentissage ; c'est-à-dire que le système lui-même doit établir les classes et les différences entre la perception des objets qui lui sont proposés, aucune indication externe ne lui est donnée qui lui permettrait de discriminer les objets a priori.

III-10-1) Systèmes de préhension

Ce paragraphe est consacré à la description des deux systèmes de préhension que nous utilisons comme support aux expériences de classement d'objets :

A) La Main (fig 3.a)

La prothèse de main anthropomorphe que nous utilisons a été développée à l'Institut "Mihailo Pupin" de Belgrade.

Cette main a été équipée de capteurs du type "peau artificielle" (développée au L.A.A.S. par Monsieur J. CLOT) qui permettent d'avoir deux informations :

- une information de glissement
- une information de force de serrage (ou de pression)

La combinaison de ces deux informations permet d'avoir des possibilités d'actions réflexes [3.11].

La photographie de la figure (3.a') montre la disposition des capteurs sur la main

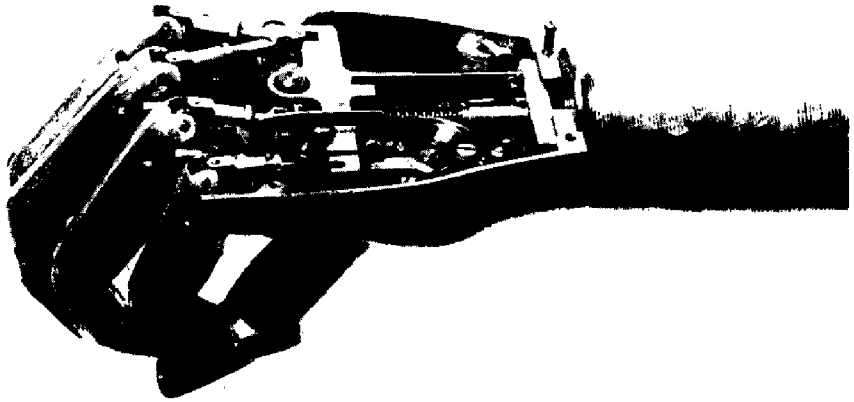


Figure 3.a - Détail de la commande mécanique

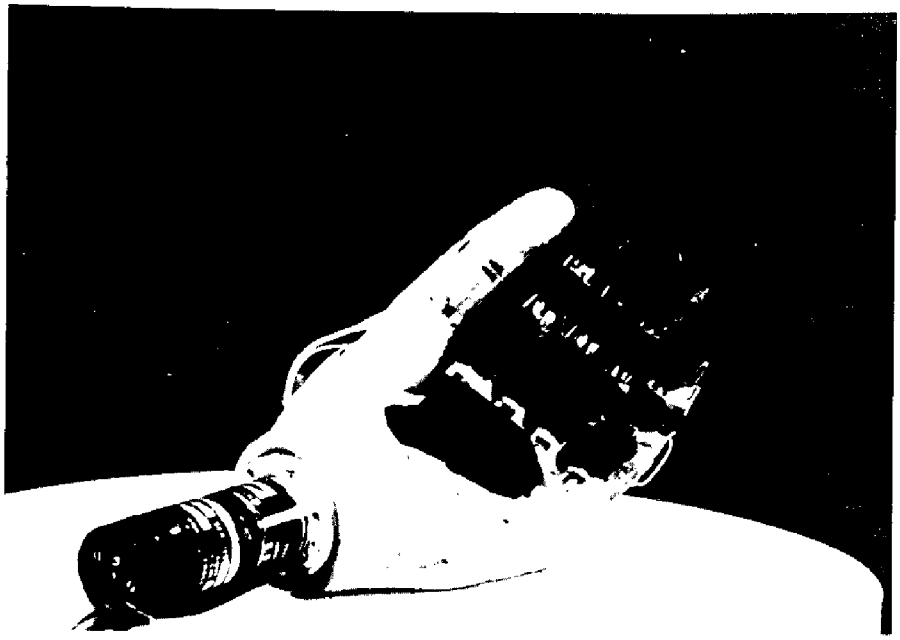


Figure 3.a' - Position des capteurs de pression sur la main

La peau artificielle qui recouvre cette prothèse comprend une pièce de polymère, de quelques millimètres d'épaisseur, chargée en carbonne ou en particules métalliques.

La résistance transversale de ce revêtement varie en fonction de la pression reçue. En mesurant le courant sur chacun des points, on obtient une information sur la pression ou,

après dérivation, sur la tendance au glissement.

Une électronique associée permet de convertir ces courants en tensions (fig. 3.a").

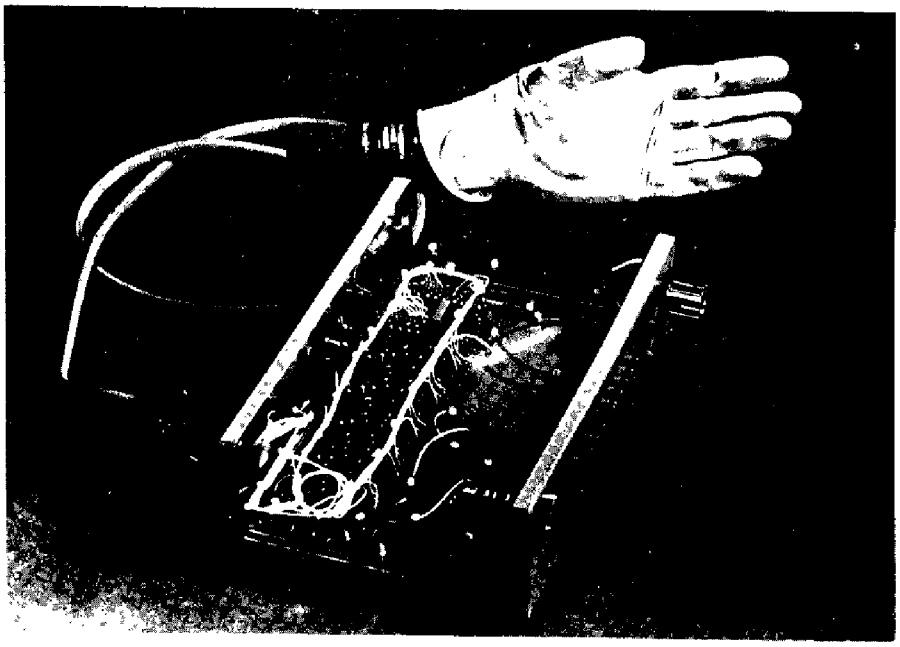


Figure 3.a"

Cette prothèse permet deux types de préhensions :

. entre le pouce et les autres doigts (pouce position basse)

. entre la paume et les doigts (pouce position haute)

B) La Pince

La pince que nous utilisons est constituée de quatre doigts symétriques ayant chacun trois phalanges. Chaque articulation est munie d'un capteur fournissant une information angulaire.

Les photographies de la figure 3.b représentent cette

pince serrant un cube et une sphère.

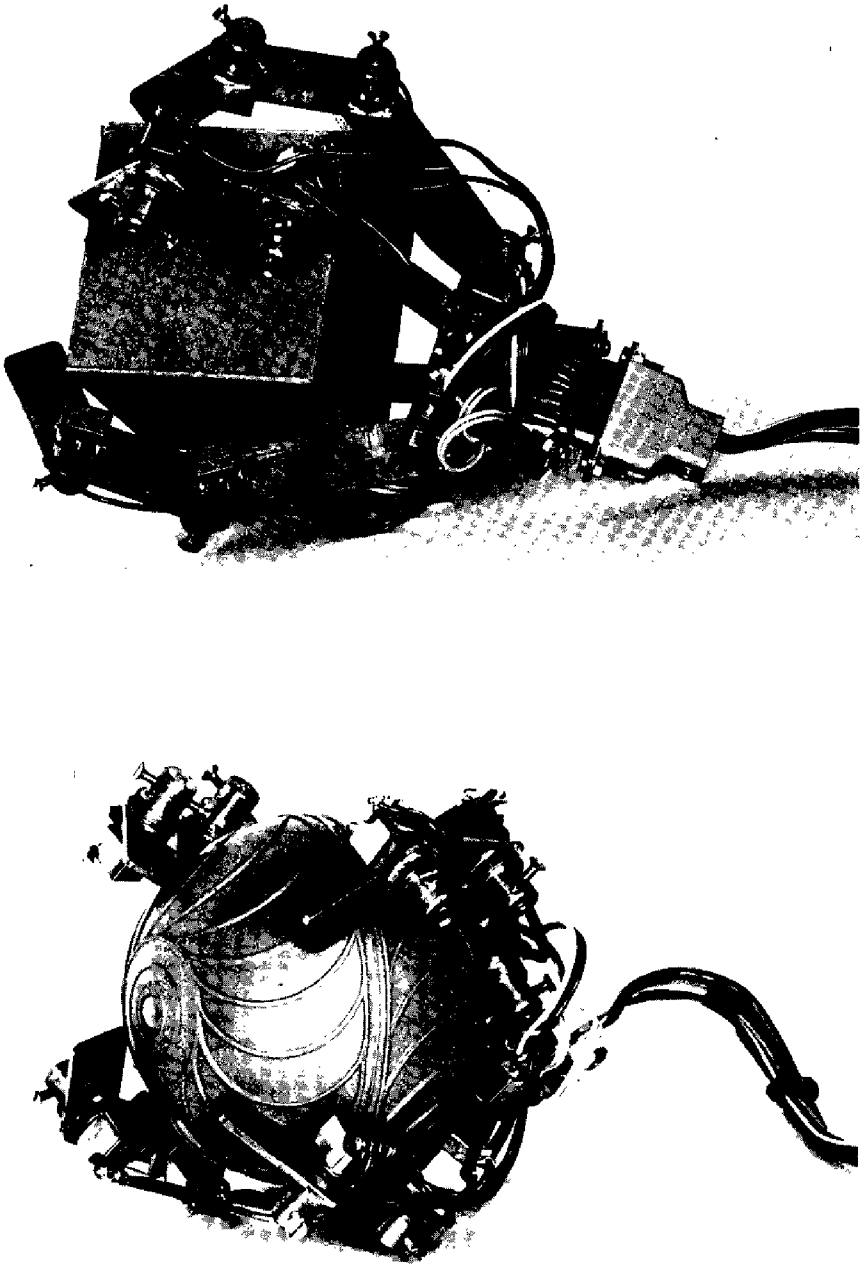


Figure 3.b

Les capteurs de position angulaire disposés sur chaque articulation sont constitués par des potentiomètres linéaires de gain : 0.037 V/degré.

De plus la valeur zéro pour un angle est obtenue lorsque les deux phalanges correspondantes sont alignées. En outre, il ne peut y avoir d'angles dont la valeur soit négative, c'est-à-dire qu'un doigt ne peut être plié à l'envers.

III-10-2) Acquisition des informations

A) Couplage des systèmes de préhension à un miniordinateur

Comme nous l'avons vu dans le paragraphe précédent, les informations délivrées par les capteurs, aussi bien de pression qu'angulaires, sont des tensions analogiques.

La prise en compte de ces différentes tensions est effectuée par une chaîne de conversion analogique-numérique qui sert d'interface entre le miniordinateur et le système de préhension (fig. 3.c).

Cette interface permet de saisir jusqu'à 16 mesures simultanées par objet.

Le codage des tensions analogiques se fait sur 16 bits et nous avons couplé les systèmes de préhension à un miniordinateur MITRA 15.

Les mesures sont enregistrées sur bande magnétique et constituent un fichier de données à partir duquel nous avons simulé l'apprentissage séquentiel. Chaque prise d'objet constitue une nouvelle série de mesures, sous forme d'un vecteur à N_p composantes.

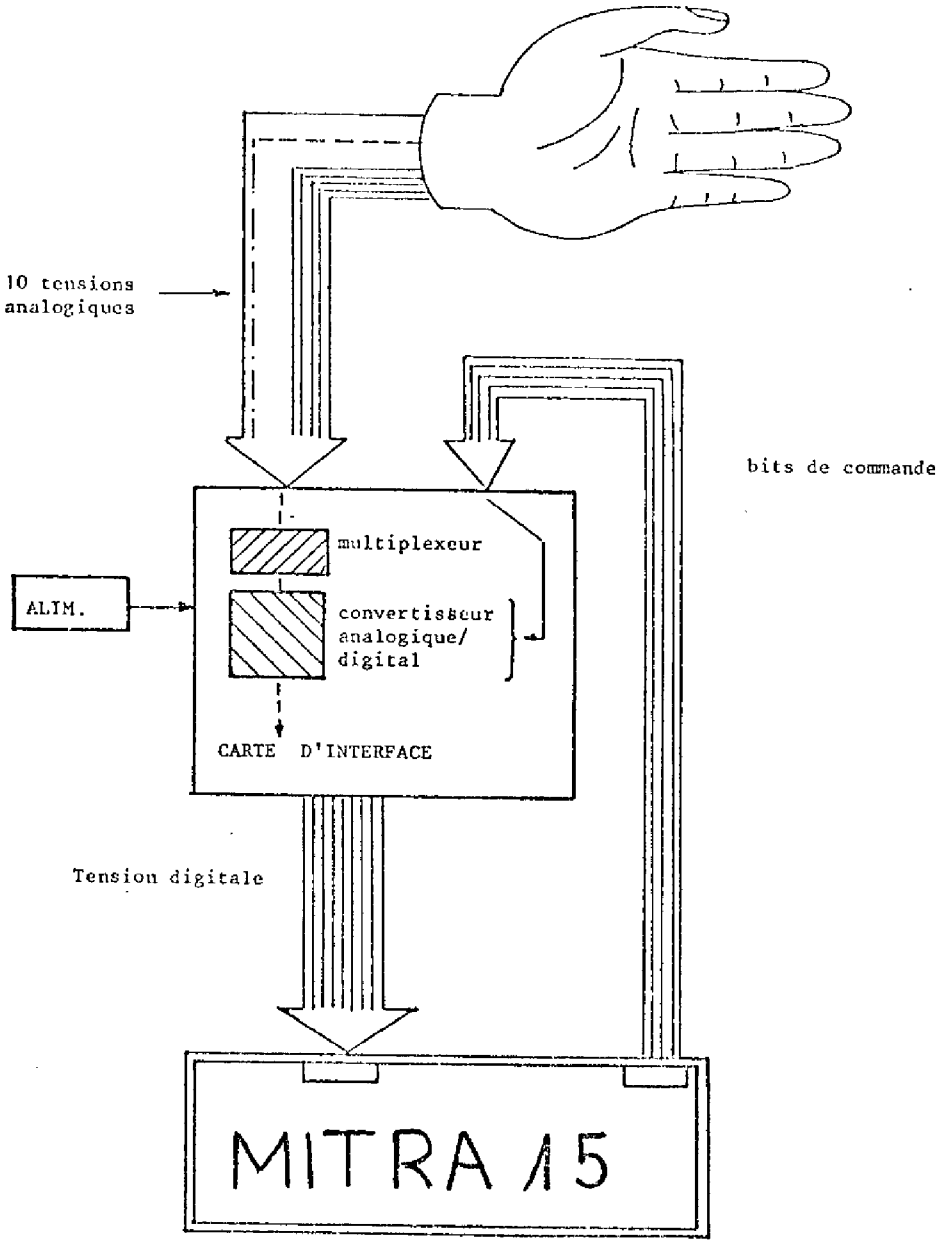


Figure 3.c - Schéma général Main-Interface-Calculateur

B) Conditions de mesures

Les différents objets considérés sont :

. un cylindre

. un parallélépipède rectangle

. un cube

. une sphère

. un tétraèdre régulier

Le choix de ces objets est arbitraire mais il faut que leurs dimensions ne soient pas trop grandes de manière à ce qu'ils soient englobés complètement (sphère, cube) ou partiellement (parallélépipède, cylindre) dans les systèmes de préhension considérés.

Dans le cas de la main, pour chaque objet il y a un certain nombre de positions possibles que nous groupons en deux types :

. pouce position haute

. pouce position basse

Les mesures pourront, pour une même position de l'objet, varier selon la force de serrage, nous avons particulièrement veillé à serrer l'objet toujours dans les mêmes conditions, c'est-à-dire avec la même intensité de courant dans le moteur d'asservissement lorsqu'il est bloqué.

Dans le cas de la pince chaque objet tend à être serré de façon que chaque phalange le touche,tout en respectant le fait qu'une articulation ne peut pas être pliée. Ceci peut entraîner que certaines phalanges n'aient pas de contact.

III-10-3) Analyse des résultats et conclusions

Nous avons procédé à des classements à l'aide de la main anthropomorphe,en utilisant uniquement les informations de pression, et de la pince munie de capteurs de position angulaire.

Pour la main les vecteurs à classer ont dix composantes distribuées selon le schéma de la figure 3.a' .

Pour la pince nous avons considéré soit deux doigts opposés, soit les quatre doigts ;chaque saisie fournit des vecteurs de 12 composantes si on a quatre doigts ou de 6 composantes si on a deux doigts.

L'environnement à classer est formé d'un cube, une sphère et un tétraèdre dans le cas de la pince et d'un cube, une sphère, un cylindre et un parallélépipède rectangle dans le cas de la main.

Dans le cas de la pince nous disposons de 30 mesures pour chacun des 3 objets, dans le cas de la main 22 mesures pour chaque objet. Nous avons utilisé l'indice de dissemblance normalisé basé sur l'information conditionnelle entre partitions décrit dans le chapitre précédent et dans [3.9] pour évaluer la différence D entre le classement obtenu et le classement "naturel" ou "vrai".

Le tableau (T-3.1) résume les résultats obtenus avec la fonction d'apprentissage caractérisée par le α subjectif, pour différents niveaux d'informations acquises :

- Main avec 3 niveaux de pression
- Main avec 5 niveaux de pression

- Pince avec 4 doigts actifs
- Pince avec 2 doigts actifs

Différentes valeurs du coefficient α subjectif ont été données : $\alpha = 0.2$; $\alpha = 0.5$; $\alpha = 0.7$ et pour chaque cas nous avons calculé la différence D (indice de dissemblance normalisé) entre le classement obtenu et le classement "naturel".

Des brefs commentaires dans le tableau (T-3.1) résument les résultats fournis par le programme dont un exemple de listing de résultats est donné dans le tableau (T-3.2).

Le tableau (T-3.3) résume les résultats obtenus avec la fonction d'apprentissage caractérisée par le α statistique pour les mêmes niveaux d'information que dans le cas du α subjectif c'est-à-dire :

- Main avec 3 niveaux de pression
- Main avec 5 niveaux de pression
- Pince avec 4 doigts actifs
- Pince avec 2 doigts actifs

Nous avons aussi calculé la différence D entre le classement obtenu et le classement "naturel" ou "vrai"

Conclusion

Avec cette application à la reconnaissance tactile de solides on prétendait seulement illustrer notre algorithme de classement sans nous pencher sur le problème de l'étude des différents systèmes de préhension, cependant il semble que les informations angulaires soient préférables, avec notre approche

par autoapprentissage, pour un traitement rapide et une discrimination correcte mais nous devons regretter le fait de ne pas avoir pu disposer de capteurs angulaires sur la main anthropomorphe ce qui nous aurait permis de faire la comparaison dans des conditions plus homogènes. D'autre part il faut aussi remarquer que la disposition et le nombre de capteurs de pression sur la main ne sont certainement pas les meilleurs et qu'une recherche dans ce domaine reste à faire, probablement l'utilisation d'un nombre réduit d'informations mixtes (angulaires et de pression) pourrait bien être la combinaison optimale dans le cas de notre algorithme. Signalons finalement que, dans le cas de la main, l'utilisation d'algorithmes qui traiteraient des données du type "empreinte" donneraient certainement des meilleurs résultats.

-:-

Expérience \ Subjectif α	0,2	0,5	0,7
Pince 4 doigts	D = 0,103 7 classes Très bonne reconnaissance après l'apprentissage complet	D = 0,027 14 classes Reconnaissance parfaite Amélioration par rapport à α = 0,2	D = 0,047 18 classes Très bonne reconnaissance mais un peu moins bonne que pour α = 0,5
Pince 2 doigts opposés	D = 0,184 8 classes La spère est reconnue parfaitement. Le tétraèdre est correctement reconnu mais il absorbe certaines positions du cube	D = 0,162 14 classes La spère est correctement reconnue Certaines positions du cube sont confondues avec le tétraèdre	D = 0,148 17 classes Légère amélioration dans la discrmination des cubes et tétraèdres par rapport au α = 0,2 et 0,5
Main 5 niveaux quantification des pressions	D = 0,905 2 classes Aucune recconaissance efficace (l'indice de dissemblance est proche de sont maximum = 1)	D = 0,827 5 classes Le cube et la sphère sont confondus, le cylindre se détâche en classe 2, qui n'accepte aucune position du parallélépipède La classe 3 est caractérisée surtout par le parallélépipède	D = 0,782 5 classes Le cylindre est la forme la mieux reconnue Une classe est formée presque exclusivement du parallélépipède et du cube. Le reste se trouvant dispersé !
Main 3 niveaux quantification des pressions	D = 0,8525 3 classes Il y a moins de classes créées que de classes initiales Aucune classe ne caractérise un objet particulier	D = 0,832 5 classes Seul le cylindre caractérise une classe bien que plus de la moitié de ses positions se distribuent dans d'autres classes	D = 0,777 5 classes Presque toutes les positions du cylindre se retrouvent en une classe. Les autres classes absorbent tout le reste à l'exclusion des cylindres.

TABLEAU T-3.1

PINCE 5 NIVEAUX $\alpha = 0,5$

```
ELEMENT 27     CLASSE    2  FORME= 1 2 0 3 0 3 3 0 0 2 3 0
ELEMENT 28     CLASSE    2  FORME= 0 2 0 3 0 3 1 1 0 0 3 0
ELEMENT 29     CLASSE    4  FORME= 0 3 0 2 2 3 2 2 1 0 4 0
ELEMENT 30     CLASSE    4  FORME= 2 2 0 2 2 3 1 2 2 0 4 1
ELEMENT 01     CLASSE    6  FORME= 4 2 1 1 4 0 2 3 2 4 3 0
ELEMENT 02     CLASSE    3  FORME= 4 0 4 0 4 1 1 3 1 4 2 0
ELEMENT 03     CLASSE    6  FORME= 4 1 4 1 4 0 2 4 1 3 2 0
ELEMENT 04     CLASSE    5  FORME= 2 4 3 2 3 0 3 2 1 4 3 2
ELEMENT 05     CLASSE    5  FORME= 2 3 0 4 2 0 3 1 0 3 4 0
ELEMENT 06     CLASSE    2  FORME= 2 2 0 3 2 0 3 3 0 3 3 0
ELEMENT 07     CLASSE    5  FORME= 1 4 0 4 0 4 4 2 2 3 3 2
ELEMENT 08     CLASSE    7  FORME= 1 4 1 4 1 3 4 0 3 3 2 3
ELEMENT 09     CLASSE    7  FORME= 1 2 3 3 2 2 4 0 3 4 2 2
ELEMENT 10     CLASSE    7  FORME= 3 0 2 3 1 3 3 0 3 4 2 2
ELEMENT 11     CLASSE    6  FORME= 3 3 0 3 1 0 2 4 1 3 4 1
ELEMENT 12     CLASSE    6  FORME= 4 1 3 3 1 4 2 3 2 4 2 2
ELEMENT 13     CLASSE    7  FORME= 1 2 4 3 2 3 4 0 4 2 2 3
ELEMENT 14     CLASSE    7  FORME= 2 1 1 3 1 3 3 0 4 4 0 3
ELEMENT 15     CLASSE    5  FORME= 2 2 3 4 2 3 3 2 2 4 0 4
 CLASSE  1
          0.0008 0.0008 0.0008 0.0008 0.0008 0.9309 0.0008 0.0008 0.0720
          0.0372 0.0008 0.4995
          0.0.99 0.0008 0.9966 0.0396 0.0008 0.0666 0.0695 0.0008 0.9255
          0.9603 0.0008 0.4580
          0.4865 0.1032 0.0008 0.5293 0.0021 0.0008 0.4768 0.0061 0.0008
          0.0008 0.0008 0.0008
          0.4924 0.0400 0.0008 0.4174 0.8678 0.0008 0.4519 0.9224 0.0008
          0.0008 0.1682 0.0008
          0.0008 0.0092 0.0008 0.0129 0.1284 0.0008 0.0008 0.0698 0.0008
          0.0008 0.8293 0.0008
CARACTERISATION DE LA CLASSE A L'AIDE DES  3  ELEMENTS LES + REPRESENTATIFS

RANG  1  VALEUR = 0.30E-01  ELEMENT 4 4 2 3 4 1 3 4 2 2 5 1
RANG  2  VALEUR = 0.30E-01  ELEMENT 4 4 2 3 4 1 3 4 2 2 5 2
RANG  3  VALEUR = 0.30E-01  ELEMENT 3 4 2 3 4 1 3 4 2 2 5 1
 LA CLASSE 1 EST FORMEE DE  30 ELEMENTS DONT
   30 DE L'ANCIENNE CLASSE 1
    0 DE L'ANCIENNE CLASSE 2
    0 DE L'ANCIENNE CLASSE 3
 LA CLASSE 2 EST FORMEE DE  20 ELEMENTS DONT
    0 DE L'ANCIENNE CLASSE 1
   19 DE L'ANCIENNE CLASSE 2
    1 DE L'ANCIENNE CLASSE 3
 LA CLASSE 3 EST FORMEE DE   7 ELEMENTS DONT
    0 DE L'ANCIENNE CLASSE 1
    5 DE L'ANCIENNE CLASSE 2
    2 DE L'ANCIENNE CLASSE 3
 LA CLASSE 4 EST FORMEE DE   5 ELEMENTS DONT
    0 DE L'ANCIENNE CLASSE 1
    5 DE L'ANCIENNE CLASSE 2
    0 DE L'ANCIENNE CLASSE 3
 LA CLASSE 5 EST FORMEE DE   9 ELEMENTS DONT
    0 DE L'ANCIENNE CLASSE 1
    1 DE L'ANCIENNE CLASSE 2
    8 DE L'ANCIENNE CLASSE 3
 LA CLASSE 6 EST FORMEE DE  10 ELEMENTS DONT
    0 DE L'ANCIENNE CLASSE 1
    0 DE L'ANCIENNE CLASSE 2
   10 DE L'ANCIENNE CLASSE 3
 LA CLASSE 7 EST FORMEE DE   8 ELEMENTS DONT
    0 DE L'ANCIENNE CLASSE 1
    0 DE L'ANCIENNE CLASSE 2
    8 DE L'ANCIENNE CLASSE 3
 89 ELEMENTS CLASSES
 ICAHF=  7ICARC=  3
          MATRICE DE CONTINGENCE
 [0.3371   0.0      0.0
  0.0      0.2135   0.0112
  0.0      0.0562   0.0225
  0.0      0.0562   0.0
  0.0      0.0112   0.0899
  0.0      0.0      0.1124
  0.0      0.0      0.0899]
          DIFFERENCE ENTRE LES DEUX CLASSIFICATIONS=  0.1033
```

TABLEAU T-3.2

Expérience \ α	α statistique
PINCE 2 DOIGTS OPPOSES	D = 0.1877 7 classes trés bonne reconnaissance : 14 éléments mal classés sur un total de 90. (4 tétraèdres dans la classe des sphères plus 10 tétraèdres dans les classes de cubes)
PINCE 4 DOIGTS	D = 0.1684 8 classes très bonne reconnnaissance : 13 éléments mal classés sur un total de 90. (3 tétraèdres dans la classe des sphères plus 10 tétraèdres répartis dans les 3 classes correspondantes aux cubes)
MAIN 5 NIVEAUX Quantification des pressions	D = 1 1 classe trés mauvaise reconnaissance : Les 88 éléments se trouvent tous dans la même classe.
MAIN 3 NIVEAUX Quantification des pressions	D = 0.7414 3 classes meilleure discrimination que dans le cas des 5 niveaux mais encore avec moins de classes créées que de classes initiales seul le cylindre caractérise bien une classe.

TABLEAU T-3.3

REFERENCES

/-3.1-/ AGUILAR MARTIN J. (1972)
"Algorithmes de classification itérative en l'absence d'information initiale"
Cybernética n° 4.

/-3.2-/ AGUILAR MARTIN J., BANON G., LOPEZ DE MANTARAS R. (1976)
"A self learning algorithm for the classification and recognition of vectorial patterns"
IEEE International Symposium on Information Theory
Ronneby, Sweden, June.

/-3.3-/ AGUILAR MARTIN J., BANON G., BRIOT M., LOPEZ DE MANTARAS R. (1976)
"Tentative de simulation de l'agrégation et du classement des informations dans la reconnaissance tactile de solides"
Colloque BIOMECA II, Toulouse, Novembre.

/-3.4-/ AGUILAR MARTIN J., BANON G., BRIOT M., RENAUD M. (1977)
"Utilisation en reconnaissance tactile de solides d'estimateurs de densités de probabilité basés sur la distance euclidienne dans $\mathbb{R}^m$ "
Colloque national sur le traitement du signal et ses applications. Nice, Avril.

/-3.5-/ BANON G. (1976)
"Jeu de hasard comportant une procédure d'apprentissage élémentaire"
R.A.I.R.O., pp. 111-117, Avril.

/-3.6-/ BRIOT M. (1977)
"Perception et identification tactile de solides, après apprentissage, en robotique"
Thèse d'Etat, Université Paul Sabatier (en préparation)

/-3.7-/ FU K.S. (1968)
"Sequential methods in pattern recognition and machine learning"
New York : Academic Press.

/-3.8-/ HO Y.C. et AGRAWALA A.K. (1968)
"On pattern classification algorithms : introduction and Survey"
IEEE Trans. Automatic Control, vol. AC-13, pp. 676-690, December.

/-3.9-/ LOPEZ DE MANTARAS R. et NUALART D. (1976)
"L'information conditionnelle comme mesure de dissemblance entre classifications"
Note interne LAAS, 76 I 16, Mai.

/-3.10-/ NILSSON N.J. (1965)
"Learning machines"
New York : Mc Graw Hill.

/-3.11-/ STOJILKOVICZ et CLOT J. (1977)
"Integrated behaviour of Artificial skin"
IEEE Transactions on Biomedical Engineering.

/-3.12-/ WATANABE S. (1969)
" Methodologies of pattern recognition"
New York : Academic Press.

CHAPITRE IV

ALGORITHME DE CLASSEMENT PAR AUTOAPPRENTISSAGE DANS LE CAS DE DONNEES GAUSSIENNES

-:-

IV-1. INTRODUCTION

Nous allons développer dans ce chapitre un algorithme de classement de données gaussiennes suivant l'approche de l'autoapprentissage.

Nous commençons par rappeler la théorie du filtrage linéaire optimal de Kalman-Bucy dans le cas discret qui est l'élément fondamental de la récursivité de cet algorithme.

Nous continuons ensuite avec l'établissement d'estimateurs séquentiels non biaisés des matrices de covariance des bruits de dynamique et d'observation ce qui nous amène à établir un filtre sous-optimal adaptatif, dans le paragraphe suivant nous décrivons l'algorithme qui utilise ce filtre sous-optimal pour estimer les densités de probabilité gaussiennes qui caractérisent chaque classe. Cet algorithme est illustré sur un exemple de classement de données multidimensionnelles fournies par la pince décrite dans le chapitre précédent en serrant des formes solides. On applique ensuite cet algorithme au problème de la reconnaissance des composants d'un mélange de densités de probabilité gaussiennes bidimensionnelles et une comparaison de notre algorithme avec celui des Nuées Dynamiques de E. DIDAY est alors faite.

IV-2. ESTIMATION RECURSIVE DE L'ETAT D'UN SYSTEME DYNAMIQUE DISCRET (Rappel du filtre de Kalman-Bucy)

L'élément fondamental de la récursivité de l'algorithme de classement que nous proposons dans ce chapitre est le filtre linéaire optimal discret, ou filtre de Kalman-Bucy. Nous rappelons ici les étapes du raisonnement qui établissent les résultats essentiels du filtrage récursif car l'originalité de ce travail réside dans leur utilisation dans le problème de regroupement de points dans l'espace à N dimensions (clustering).

La littérature traitant de ce dernier thème ignore en général cet aspect récursif et ne prend jamais à notre connaissance ce point de vue qui le rattache à l'estimation paramètrique récursive : [4.1], [4.8], [4.10], [4.13], [4.14], [4.15], [4.16], [4.24], [4.26], [4.29], [4.34].

Nous allons rappeler le problème d'estimer l'état d'un système dynamique discret, à partir de la connaissance de mesures bruitées de la sortie de ce système, quand des perturbations aléatoires agissent sur l'équation d'état du système et que l'état initial est lui même un vecteur aléatoire.

Nous avons donc le problème suivant :

Problème_1 : Considérons le système dynamique linéaire discret, dont l'état x_k à l'instant t_k est généré par l'équation aux différences suivantes :

$$x_{k+1} = \Phi x_k + \Gamma_k v_k \quad , \; x_k \in \mathbb{R}^n \qquad (4.1)$$

auquelle on associe une équation d'observation perturbée

$$y_K = H_K x_K + w_K \quad , \; y_K \in \mathbb{R}^m \qquad (4.2)$$

avec les hypothèses suivantes concernant le modèle et les perturbations :

Hypothèses :

* v_k et w_k sont des bruits blancs gaussiens discrets (séquences de variables gaussiennes indépendantes) tels que :

$$\mathbb{E}[v_K] = 0 \qquad \mathbb{E}[w_k] = 0$$

$$\mathbb{E}[v_K v_j^T] = Q\delta_{kj} \qquad \mathbb{E}[w_k w_j^T] = R\delta_{kj}$$

avec Q définie non-négative et R définie positive, et où δ_{kj} est le symbole de Kronecker défini par :

$$\delta_{kj} = \begin{cases} 1 & \text{si } K = j \\ 0 & \text{si } K \neq j \end{cases}$$

* Les bruits v_k et w_k ne sont pas correlés :

$$\mathbb{E}[v_k w_j^T] = 0 \qquad \forall K, \forall j$$

Cette hypothèse est compatible avec l'utilisation des résultats que nous comptons faire mais n'est pas indispensable, on peut traiter le cas plus général où les bruits de dynamique et d'observation sont correlés entre eux, mais cela compliquerait les équations du filtre correspondant.

* L'état initial x_o est distribué suivant une loi gaussienne supposée connue avec :

$$\mathbb{E}[x_o] = \mu_o \quad , \quad \mathbb{E}[(x_o - \mu_o)(x_o - \mu_o)^T] = P_o$$

* L'état initial x_o et les bruits v_k et w_k ne sont pas correlés :

$$\mathbb{E}\left[x_o v_k^T\right] = 0 \quad , \quad \mathbb{E}\left[x_o w_k^T\right] = 0 \qquad \forall k$$

* Les éléments des matrices ou vecteurs Φ_k, Γ_k et H_k (pouvant être fonction du temps dans le cas général) sont parfaitement connus à chaque instant et ils ne sont pas aléatoires.

Pour chaque K = 0,1,2, ... notons Y_K la séquence $\{y_j\}_{j=0}^{K}$ des vecteurs aléatoires sortie du système.

Alors le problème s'énonce sous la forme suivante : Déterminer, pour chaque K= 0,1,2, ... le meilleur estimateur de l'état du système x_k connaissant la suite de mesures $\{y_j\}_{j=0}^{K}$

Il faut donc préciser quel sera le critère d'optimalité choisi :

Critère : $$\text{Min } \mathbb{E}\left[(x_k - \hat{x}_k)^T S (x_k - \hat{x}_k)\right] \qquad (4.3)$$

où S est une matrice symétrique quelconque définie non négative et $\hat{x}_k$ est une fonction définie sur les observations : $\{y_j\}_{j=0}^{K}$

La solution au problème 1 est donnée par la proposition bien connue suivante [4.6].

Proposition 1 : L'estimateur optimal au sens du critère défini dans (4.3) est la moyenne conditionnelle, on le note :

$$\hat{x}_{k/k} = \mathbb{E}\left[x_k \,\middle|\, \{y_j\}_{j=0}^{k}\right] = \mathbb{E}\left[x_k \,\middle|\, Y_k\right] \qquad (4.4)$$

Remarque 1 : d'après les hypothèses faites, la densité de probabilité conditionnelle $p(x_k|Y_k)$ est gaussienne, et l'estimateur du maximum a posteriori est, dans cette situation, confondu avec celui de la moyenne conditionnelle.

Définition 1 : Le "prédicteur à un pas", est l'estimateur

de l'état x_k à l'instant t_k, connaissant la suite de mesures jusqu'à l'instant t_{k-1}, on le notera :

$$\hat{x}_{k/k-1} = \mathbb{E}\left[x_k \,\middle|\, \{y_j\}_{j=0}^{k-1}\right] = \mathbb{E}\left[x_k \,\middle|\, Y_{k-1}\right] \qquad (4.5)$$

Remarque 2 : On voit d'après cette définition, que l'on pourra être amené à traiter deux problèmes voisins de celui du filtrage :

- si $\hat{x}_{k/i}$ avec $i < k$ c'est le problème de la prédiction
- si $\hat{x}_{k/i}$ avec $i > k$ c'est le problème du lissage

Proposition 2 : Supposant connu l'estimateur $\hat{x}_{k-1/k-1}$ le prédicteur à un pas définit dans la définition I peut être exprimé sous la forme récursive suivante :

$$\hat{x}_{k/k-1} = \Phi_{k-1}\, \hat{x}_{k-1/k-1} \qquad (4.6)$$

et la covariance de l'erreur de prédiction à un pas s'écrit :

$$P_{k/k-1} = \Phi_{k-1} P_{k-1/k-1} \Phi_{k-1}^T + \Gamma_{k-1} Q\, \Gamma_{k-1}^T \qquad (4.7)$$

Démonstration : On considère l'équation dynamique

$$x_k = \Phi_{k-1} x_{k-1} + \Gamma_{k-1} v_{k-1}$$

en prenant l'espérance mathématique conditionnelle des deux membres on obtient

$$\mathbb{E}\left[x_k \,\middle|\, \{y_j\}_{j=0}^{k-1}\right] = \Phi_{k-1}\, \mathbb{E}\left[x_{k-1} \,\middle|\, \{y_j\}_{j=0}^{k-1}\right] + \Gamma_{k-1}\left[v_{k-1} \,\middle|\, \{y_j\}_{j=0}^{k-1}\right]$$

ce qui donne

$$\hat{x}_{k/k-1} = \Phi_{k-1}\, \hat{x}_{k-1/k-1} \qquad (4.8)$$

le deuxième terme est nul car v_{k-1} est indépendant des w_k, de v_{k-2}, v_{k-3}, ... v_o et de x_o, donc v_{k-1} est indépendant des y_o, y_1, ... y_{k-1}, on a alors

$$\mathbb{E}\left[v_{k-1} \,\middle|\, \{y_j\}_{j=0}^{k-1}\right] = \mathbb{E}\left[V_{k-1}\right] = 0$$

La matrice de covariance de l'erreur de prédiction à un pas est

$$P_{k/k-1} = \mathbb{E}\left[(x_k - \hat{x}_{k/k-1})(x_k - \hat{x}_{k/k-1})^T \mid \{y_i\}_{j=0}^{k-1}\right]$$

mais puisque $x_k - \hat{x}_{k/k-1}$ est indépendant de $\{y_i\}_{i=0}^{k-1}$ d'après la propriété I (voir Annexe I), la matrice de covariance sera donc aussi égale à :

$$P_{k/k-1} = \mathbb{E}\left[(x_k - \hat{x}_{k/k-1})(x_k - \hat{x}_{k/k-1})^T\right] \qquad (4.9)$$

L'erreur de prédiction à un pas s'écrit :

$$x_k - \hat{x}_{k/k-1} = \Phi_{k-1}\left(x_{k-1} - \hat{x}_{k-1/k-1}\right) + \Gamma_{k-1}\, v_{k-1}$$

et l'on obtient alors :

$$P_{k/k-1} = \Phi_{k-1}\, P_{k-1/k-1}\, \Phi_{k-1}^T + \Phi_{k-1}\, \mathbb{E}\left[(x_{k-1} - \hat{x}_{k-1/k-1}) v_{k-1}^T \mid Y_{k-1}\right] \Gamma_{k-1}^T +$$

$$+ \Gamma_{k-1}\, \mathbb{E}\left[v_{k-1}(x_{k-1} - \hat{x}_{k-1/k-1})^T \mid Y_{k-1}\right] \Phi_{k-1}^T + \Gamma_{k-1}\, Q\, \Gamma_{k-1}^T \qquad (4.10)$$

On remarquera que :

$\mathbb{E}\left[x_{k-1}\, v_{k-1}^T\right] = 0$ car x_{k-1} est fonction de x_o et de v_o, v_1, ... v_{k-2} ; d'après les hypothèses, x_{k-1} et v_{k-1} sont donc indépendants et de plus v_{k-1} est a moyenne nulle.

D'autre part :

$$\mathbb{E}\left[\hat{x}_{k-1/k-1}\, v_{k-1}^T \mid Y_{k-1}\right] = \hat{x}_{k-1/k-1}\, \mathbb{E}\left[v_{k-1}^T \mid Y_{k-1}\right] = \hat{x}_{k-1/k-1}\, \mathbb{E}\left[v_{k-1}^T\right] = 0$$

la matrice de covariance de l'erreur de prédiction à un pas sera donc déterminée à partir de la relation

$$P_{k/k-1} = \Phi_{k-1}\, P_{k-1/k-1}\, \Phi_{k-1}^T + \Gamma_{k-1}\, Q\, \Gamma_{k-1}^T \qquad (4.11)$$

Proposition 3 : L'estimateur optimal défini dans la proposition 1 peut être exprimé sous la forme récursive suivante, connaissant

le prédicteur à un pas et sa covariance d'erreur :

$$\hat{x}_{k/k} = \hat{x}_{k/k-1} + K_k z_k \qquad (4.12)$$

avec

$$z_k = y_k - H_k \hat{x}_{k/k-1} \qquad (4.13)$$

Z_k étant appelé : processus d'innovation, et avec

$$K_k = P_{k/k-1} H_k^T (H_k P_{k/k-1} H_k^T + R)^{-1} \qquad (4.14)$$

La matrice de covariance de l'erreur de l'estimateur optimal s'écrit :

$$P_{k/k} = P_{k/k-1} - K_k H_k P_{k/k-1} \qquad (4.15)$$

Démonstration :

$\hat{x}_{k/k} = \mathbb{E}[x_k \mid Y_{k-1}]$ peut être mis sous la forme :

$$\hat{x}_{k/k} = \mathbb{E}[x_k \mid Y_{k-1}, y_k] \qquad (4.16)$$

on cherche alors quelle information se trouve dans y_k qui n'était pas contenue dans Y_{k-1}.

Pour cela on utilise la propriété I (voir annexe 1) en formant le vecteur : $\begin{bmatrix} y_k \\ \hline Y_{k-1} \end{bmatrix}$

Nous savons alors que $y_k - \mathbb{E}[y_k \mid Y_{k-1}]$ est indépendant de Y_{k-1}, ce terme est appelé, on l'a vu, processus d'innovation z_k :

$$z_k = y_k - \mathbb{E}[y_k \mid Y_{k-1}] = y_k - H_k \hat{x}_{k/k-1}$$

toujours d'après la propriété I, on aura $\mathbb{E}[z_k] = 0$ d'autre part, on remarquera qu'il revient au même, d'écrire l'estimateur $\hat{x}_{k/k}$ sous la forme :

$$\hat{x}_{k/k} = \mathbb{E}[x_k \mid Y_{k-1}, z_k]$$

puisque dans l'expression de z_k, le terme $H_k \hat{x}_{k/k-1}$ est fonction de Y_{k-1}

D'après la propriété II (voir annexe I) on écrira alors :

$$\hat{x}_{k/k} = \mathbb{E}[x_k \mid Y_{k-1}] + \mathbb{E}[x_k \mid z_k] - \mathbb{E}[x_k]$$

$\begin{bmatrix} x_k \\ z_k \end{bmatrix}$ étant un vecteur gaussien (puisque d'après la propriété I $\hat{x}_{k/k-1}$ est une combinaison linéaire de y_o, y_1, ..., y_{k-1}) le terme $\mathbb{E}[x_k \mid z_k]$ s'écrira :

$$\mathbb{E}[x_k \mid z_k] = \mathbb{E}[x_k] + K_k z_k$$

d'où

$$\hat{x}_{k/k} = \hat{x}_{k/k-1} + K_k z_k \qquad (4.17)$$

Pour déterminer le coéfficient (matrice ou vecteur) K_k, on écrira d'abord l'équation relative à l'erreur d'estimation :

$$x_k - \hat{x}_{k/k} = x_k - \hat{x}_{k/k-1} - K_k z_k \qquad (4.18)$$

le terme du premier membre est $x_k - \mathbb{E}[x_k \mid Y_k]$ et d'après la propriété I et en formant le vecteur $\begin{bmatrix} x_k \\ \hline Y_k \end{bmatrix}$ on voit que $x_k - \mathbb{E}[x_k \mid Y_k]$ est indépendant de Y_k (ou de toute combinaison linéaire de $(y_o, y_1, \ldots, y_k)$.

Si l'on multiplie les deux membres de l'équation (4.18) par z_k^T et que l'on prenne l'espérance mathématique on aura pour le 1er membre :

$$\mathbb{E}[(x_k - \hat{x}_{k/k}) z_k^T] = \mathbb{E}[(x_k - \hat{x}_{k/k})] \cdot \mathbb{E}[z_k] = 0$$

(puisque z_k est une combinaison linéaire de y_o, y_1, ..., y_k).

Le 1er terme du second membre a pour expression :

$$\mathbb{E}[(x_k - \hat{x}_{k/k-1}) z_k^T] = \mathbb{E}[(x_k - \hat{x}_{k/k-1})(y_k - H_k \hat{x}_{k/k-1})^T] = P_{k/k-1} H_k^T$$

Pour le 2ème terme du second membre on obtient :

$$K_k \mathbb{E}[z_k z_k^T] = K_k \mathbb{E}[(H_k(x_k - \hat{x}_{k/k-1}) + w_k)(H_k(x_k - \hat{x}_{k/k-1}) + w_k)^T] = K_k[H_k P_{k/k-1} H_k^T + R]$$

(Compte tenu des remarques précédentes concernant l'indépendance de w_k et x_k, de w_k et $\hat{x}_{k/k-1}$)

Donc finalement nous avons

$$0 = P_{k/k-1} H_k^T - K_k (H_k P_{k/k-1} H_k^T + R)$$

d'où

$$K_k = P_{k/k-1} H_k^T (H_k P_{k/k-1} H_k^T + R)^{-1} \qquad (4.19)$$

Pour finir la démonstration nous allons calculer la matrice de covariance de l'erreur d'estimation :

$$P_{k/k} = \mathbb{E}\left[(x_k - \hat{x}_{k/k})(x_k - \hat{x}_{k/k})^T\right] = \mathbb{E}\left[(x_k - \hat{x}_{k/k-1} - K_k z_k)(x_k - \hat{x}_{k/k-1} - K_k z_k)^T\right]$$

$$= P_{k/k-1} - K_k \mathbb{E}\left[(H_k(x_k - \hat{x}_{k/k-1}) + w_k)(x_k - \hat{x}_{k/k-1})^T\right] + K_k \mathbb{E}\left[z z_k^T\right] K_k^T -$$

$$- \mathbb{E}\left[(x_k - \hat{x}_{k/k-1})(H_k(x_k - \hat{x}_{k/k-1}) + w_k)^T\right] K_k^T$$

d'où

$$P_{k/k} = P_{k/k-1} - K_k H_k P_{k/k-1} - P_{k/k-1} H_k^T K_k^T + K_k (H_k P_{k/k-1} H_k^T + R) K_k^T$$

en portant sur cette expression la valeur de K_k déterminée plus haut et après simplification on obtient :

$$P_{k/k} = P_{k/k-1} - P_{k/k-1} H_k^T (H_k P_{k/k-1} H_k^T + R)^{-1} H_k P_{k/k-1} \qquad (4.20)$$

on retrouve donc l'équation matricielle de Riccati donnée en (4.15).

L'ensemble des équations (4.6), (4.7), (4.12) et (4.15) constitue les équations du filtre linéaire optimal de Kalman-Bucy dans le cas discret.

IV-3. ESTIMATION DES STATISTIQUES DES BRUITS DE DYNAMIQUE ET D'OBSERVATION : Filtre sous-optimal adaptatif

Une limitation bien connue des applications du filtre Kalman-Bucy aux problèmes réels est l'hypothèse de la connaissance a priori des statistiques des bruits de dynamique et d'observation.

D'une façon générale ces statistiques sont obtenues au moyen d'une analyse préalable des données par simulation.

Cette approche conduit à un filtre non adaptatif. On appelle filtre adaptatif celui qui estime récursivement ces statistiques, ce qui devient un filtre non linéaire. Dans le problème de classement itératif par autoapprentissage nous devons estimer les statistiques des bruits de dynamique et d'observation de façon adaptative, puisqu'elles représentent les paramètres caractéristiques de chaque classe (paramétrisation par noyaux), et qu'ils sont donc inconnus a priori.

Nous allons décrire ici le filtre adaptatif sous-optimal que nous utiliserons ensuite.

Problème 2 : Considérons à nouveau le système dynamique linéaire discret défini dans le problème 1 :

$$\text{(4.21)} \qquad \begin{cases} x_{k+1} = \Phi \cdot x_k + v_k \quad , \quad x_k \in \mathbb{R}^n \\ y_k = H x_k + w_k \quad , \quad y_k \in \mathbb{R}^m \end{cases}$$

où Φ et H sont des matrices dont les éléments sont parfaitement connus et constants.

Nous avons les mêmes hypothèses que dans le problème 1 à savoir :

$$\mathbb{E}[v_k] = \mathbb{E}[w_k] = 0$$

$$\mathbb{E}[w_k w_j^T] = R\,\delta_{kj} \qquad \mathbb{E}[v_k v_j^T] = Q\,\delta_{kj}$$

avec Q définie non négative et R définie positive.

- Les bruits v_k et w_k ne sont pas correlés

$$\forall k, \forall j \qquad \mathbb{E}[v_k w_j^T] = 0$$

où Q et R sont les vraies covariances des bruits de dynamique et d'observation respectivement.

Récapitulons ici les résultats du paragraphe précédent en rappelant qu'ils s'appliqueraient ici si Q et R étaient connues :

- prédicteur à un pas :

$$\hat{x}_{k/k-1} = \phi\,\hat{x}_{k-1/k-1} \tag{4.22a}$$

- erreur de prédiction à un pas :

$$P_{k/k-1} = \phi\,P_{k-1/k-1}\,\phi^T + Q \tag{4.22b}$$

- processus d'innovation :

$$z_k = y_k - H\,\hat{x}_{k/k-1} \tag{4.22c}$$

- gain du filtre

$$K_k = P_{k/k-1}\,H^T\left(H\,P_{k/k-1}\,H^T + R\right)^{-1} \tag{4.22d}$$

- estimation de l'état

$$\hat{x}_{k/k} = \hat{x}_{k/k-1} + K_k\,z_k \tag{4.22e}$$

- matrice de covariance de l'erreur d'estimation

$$P_{k/k} = P_{k/k-1} - K_k\,H\,P_{k/k-1} \tag{4.22f}$$

Dans le problème de filtrage adaptatif qui nous intéresse l'ensemble de paramètres constants mais inconnus :

$$\Sigma \equiv \{Q, R\} \tag{4.23}$$

doit être estimé simultanément avec l'état du système et la covariance de l'erreur d'estimation.

Malheureusement l'estimateur optimal pour les paramètres de (4.23) est inaccessible par des équations récursives simples d'ordre fini. Plusieurs approches estiment de façon sous-optimale l'un ou les deux paramètres de l'ensemble Σ, simultanément avec l'état [4.3], [4.18], [4.23], G. SALUT [4.30] propose un algorithme optimal mais très compliqué à implémenter.

Nous avons choisi une approche partiellement heuristique basée sur les estimateurs ergodiques classiques que nous développons ici :

A) Estimation de la statistique du bruit d'observation :

Considérons l'équation d'observation du système à l'instant t_j :

$$y_j = H x_j + w_j \quad , \quad y_j \in \mathbb{R}^m \tag{4.24}$$

avec $w_j \in \mathcal{N}(0, R)$

La vraie valeur x_j de l'état est inconnue, donc W_j ne peut pas être déterminé ; mais une estimation intuitive de W_j est donnée par la quantité suivante :

$$r_j \equiv y_j - H \hat{x}_{j/j-1} \tag{4.25}$$

Puisque la séquence $\{r_j\}$ est supposée être représentative

de la séquence $\{w_j\}$.

On les considérera indépendantes et identiquement distribuées et nous appliquerons à $\{r_j\}$ l'estimateur ergodique classique [4.23].

Définissons une variable aléatoire de $\mathcal{R}$, caractérisée par la moyenne r_k et la covariance R_k, elle va être estimée d'après la séquence $\{r_j\}_{j=1}^{K}$ par l'estimateur ergodique qui est l'estimateur non biaisé

$$\hat{r}_K = \frac{1}{K} \sum_{j=1}^{K} r_j \qquad (4.26)$$

L'estimateur ergodique de R_k, la covariance de $\mathcal{R}$, est :

$$\hat{R}_K = \frac{1}{K-1} \sum_{j=1}^{K} (r_j - \hat{r}_K)(r_j - \hat{r}_K)^T \qquad (4.27)$$

or

$$\hat{R}_K = \frac{1}{K-1} \sum_{j=1}^{K} \left(r_j - \frac{1}{K}\sum_{j=1}^{K} r_j\right)\left(r_j - \frac{1}{K}\sum_{j=1}^{K} r_j\right)^T \qquad (4.28)$$

donc

$$\hat{R}_K = \frac{1}{K} \sum_{j=1}^{K} r_j r_j^T \qquad (4.29)$$

Nous utiliserons $\hat{R}_K$ comme estimation de la covariance R du bruit d'observation.

Pour estimer le bruit de dynamique considérons la relation :

$$x_{j+1} = \Phi x_{j-1} + v_j \quad , \quad x_j \in \mathbb{R}^n \qquad (4.30)$$

où $v_j \in \mathcal{N}(0, Q)$

les valeurs vraies de x_j et x_{j+1} sont inconnues donc v_j ne peut

pas être déterminé, mais une estimation intuitive de v_j est donnée par :

$$q_j \equiv \hat{x}_{j/j} - \Phi \hat{x}_{j-1/j-1} \qquad (4.31)$$

par hypothèse les v_j pour j = 1,2, ... K sont indépendants et Q est constante. En faisant le même raisonnement que précédemment, pour la suite $\{v_j\}$ la simple estimation ergodique peut être à nouveau appliquée. Définissons une variable aléatoire, $\mathcal{Q}$ sur l'espace $\Omega_{\mathcal{Q}}$ d'où les données q_j , j=1, ... K sont obtenues.

La distribution inconnue de $\mathcal{Q}$ caractérisée par une moyenne q_K et une covariance Q_k va être estimée d'après la séquence $\{q_j\}_{j=1}^{K}$ supposée indépendante.

Un estimateur non biaisé de q_k est :

$$\hat{q}_K = \frac{1}{K}\sum_{j=1}^{K} q_j \qquad (4.32)$$

Un estimateur de Q_k sera :

$$\hat{Q}_K = \frac{1}{K-1}\sum_{j=1}^{K}(q_j - \hat{q}_K)(q_j - \hat{q}_K)^T \qquad (4.33)$$

que l'on peut écrire :

$$\hat{Q}_K = \frac{1}{K-1}\sum_{j=1}^{K}\left(q_j - \frac{1}{K}\sum_{j=1}^{K} q_j\right)\left(q_j - \frac{1}{K}\sum_{j=1}^{K} q_j\right)^T$$

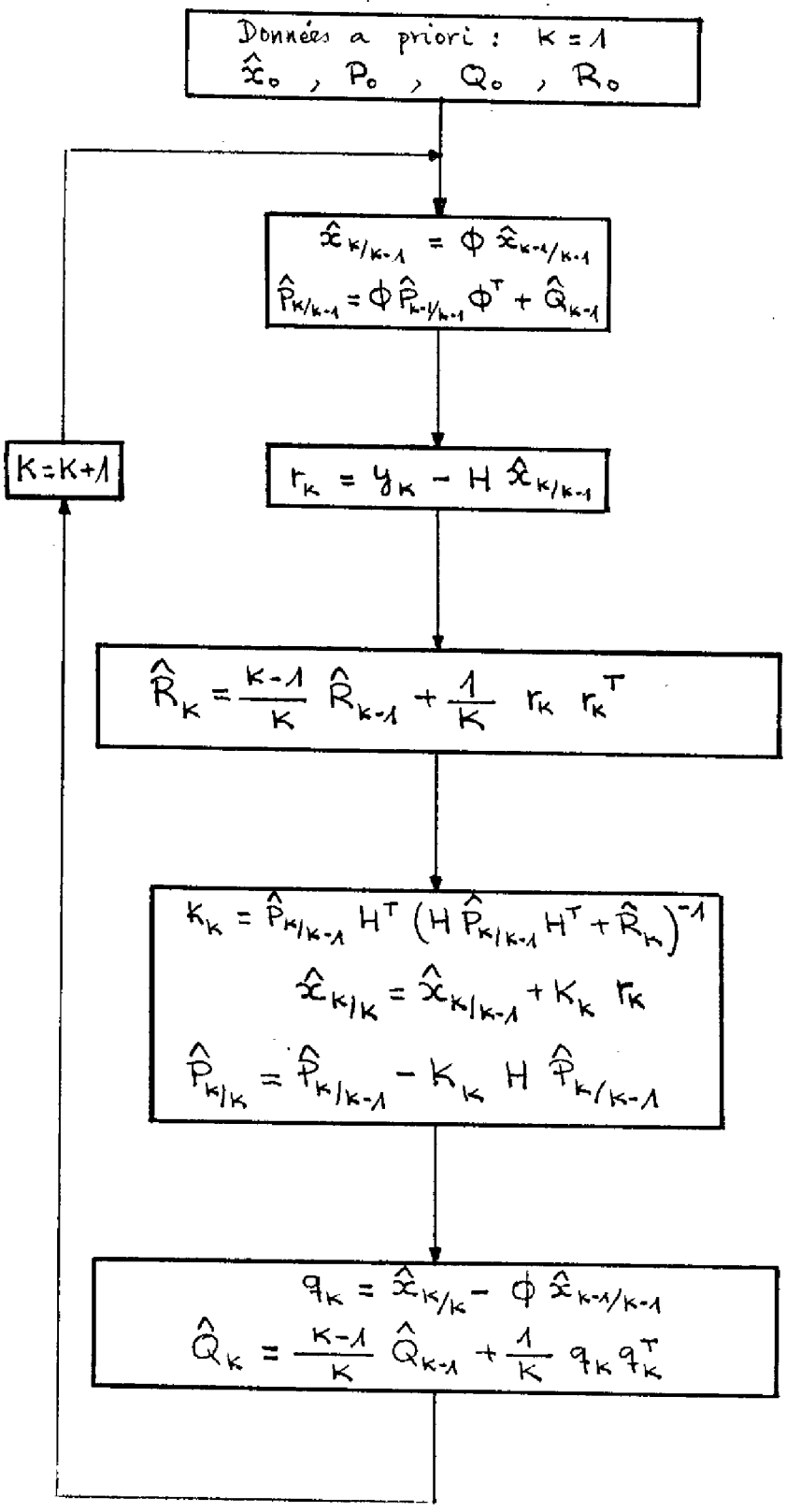
d'où

$$\hat{Q}_K = \frac{1}{K}\sum_{j=1}^{K} q_j q_j^T \qquad (4.34)$$

Nous utiliserons $\hat{Q}_K$ comme estimation de la covariance Q du bruit de dynamique.

Nous donnons ensuite l'organigramme de cet algorithme de filtrage sous-optimal adaptatif :

Algorithme du filtre sous-optimal adaptatif



IV-4. DESCRIPTION DE L'ALGORITHME D'AUTOAPPRENTISSAGE D'UN CLASSEMENT

Cet algorithme classe des données issues d'un environnement gaussien suivant une procédure d'autoapprentissage donc, le système lui même doit établir les classes et les différences entre les données qui lui sont présentées ; aucune indication externe ne lui est donnée qui lui permettrait de discriminer les données a priori, ni, non plus, le nombre de données à classer ainsi que le nombre de classes.

Les données sont assignées aux classes suivant le critère d'appartenance probabiliste du maximum de vraisemblance, quand la valeur de cette appartenance est inférieure à un certain seuil on crée une nouvelle classe au lieu d'affecter la donnée à la classe qui a donné l'appartenance maximum.

Pour calculer les appartenances de chaque donnée à chacune des classes, nous devons estimer la moyenne et la covariance qui forment le noyau de chaque classe (paramétrisation par noyau) en tenant compte de la donnée à classer,nous utilisons le filtre adaptatif sous-optimal défini dans le paragraphe précédent pour effectuer ces estimations après avoir modélisé les classes de la façon suivante.

IV-4.1) Modélisation des classes

Supposons qu'à un instant quelconque t_n l'algorithme ait créé n classes : $C_1, C_2, \ldots, C_N$ et qu'à l'instant t_{n+1} arrive une

nouvelle donnée, admettons que cette donnée soit attribuée à la classe C_i , de ce fait cette classe devra voir sa structure modifiée puisque le fait d'absorber un nouvel élément fait changer sa moyenne et sa covariance (noyau de la classe) ; on peut donc affirmer qu'il existe une dynamique associée à chaque classe et qui tient compte de son évolution au fur et à mesure que celle-ci s'enrichit avec des nouveaux éléments.

C'est pour cela que nous avons pensé à caractériser ou paramétriser les classes par leur moyenne ou centre de gravité $x^i_{K_i}$ et leur ellipsoide de dispersion ou d'inertie donné par la matrice de covariance que nous noterons $S^i_{k_i}$.

L'évolution du centre de gravité $x^i_{k_i}$ d'une classe quelconque C_i est définie par l'équation aux différences suivantes :

$$x^i_{K_i+1} = x^i_{K_i} + v^i_{K_i} \quad , \quad x^i_{K_i} \in \mathbb{R}^n \qquad (4.35)$$

où $x^i_{k_i}$ est la valeur de la moyenne de la classe C_i quand celle-ci a absorbé K_i éléments.

$x^i_{K_i+1}$ est la moyenne de cette même classe après avoir absorbé un élément de plus et $v^i_{K_i}$ est une séquence de variables aléatoires indépendantes avec :

$$\mathbb{E}[v^i_{K_i}] = 0 \qquad \mathbb{E}[v^i_{K_i} v^{i\,T}_{j_i}] = Q^i \delta_{K_i j_i}$$

La figure (4.a) donne une illustration graphique de l'évolution des classes estimées

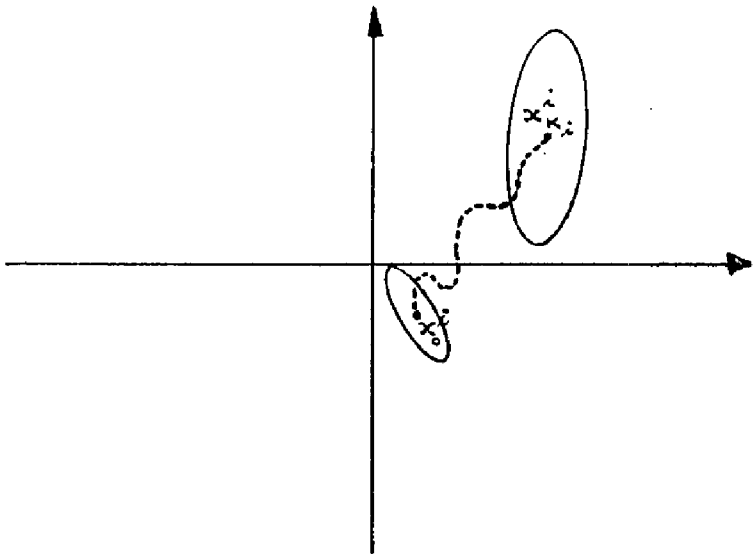


Fig 4.a - Evolution des classes dans $\mathbb{R}^2$

Nous faisons ensuite l'hypothèse que cette moyenne x_k est imparfaitement observée, et que la donnée obtenue à l'instant t_k correspond à une observation imparfaite de $x^i_{k_i}$ d'après l'équation :

$$y_t = x^i_{k_i} + w^i_{k_i} \quad , \quad y_t \in \mathbb{R}^n \qquad (4.36)$$

où W_K est une séquence de variables aléatoires gaussiennes indépendantes avec :

$$\mathbb{E}\left[w^i_{k_i}\right] = 0 \qquad \mathbb{E}\left[w^i_{k_i} w^{iT}_{j_i}\right] = R^i \delta_{k_i j_i}$$

et avec les hypothèses supplémentaires suivantes :

$$\forall k_i, \forall j_i \quad \mathbb{E}\left[v^i_{k_i} w^{iT}_{j_i}\right] = 0$$

de façon naturelle, et pour pouvoir établir un algorithme

récursif, on considère que l'état initial x_o^i est supposé distribué suivant une loi gaussienne avec :

$$\mathbb{E}[x_o^i] = 0 \qquad \mathbb{E}[x_o^i x_o^{iT}] = P_o$$

- l'état initial x_o^i et les bruits $v_{k_i}^i$ et $w_{k_i}^i$ ne sont pas correlés :

$$\mathbb{E}[x_o^i v_{k_i}^{iT}] = 0 \qquad \mathbb{E}[x_o^i w_{k_i}^{iT}] = 0 \qquad \forall k_i$$

De tout cela on déduit que y_t et $x_{k_i+1}^i$ sont aussi gaussiens parce qu'ils sont des combinaisons linéaires de bruits gaussiens.

IV-4.2) Caractérisation des classes à l'aide de densités de probabilité gaussiennes

D'après ce que nous avons vu dans le paragraphe précédent une classe est caractérisée par sa moyenne ou centre de gravité et sa matrice de covariance qui définit son ellipsoide de dispersion, puisqu'on travaille sous des hypothèses de normalité nous savons que ces deux premiers moments sont suffisants pour déterminer complètement une densité de probabilité gaussienne pour chaque classe, donc chaque classe va être caractérisée par cette densité de probabilité gaussienne (paramétrisation par noyaux).

Les données observées à classer sont la réalisation d'une variable aléatoire dont la loi de probabilité est un mélange de lois gaussiennes : $\mathcal{L}_i$. La moyenne ou centre d'inertie de chacune de ces lois est un vecteur constant inconnu : x^i.

Une difficulté surgit dans le choix de la matrice de covariance qui caractérise la dispersion dans chaque classe autour

de sa moyenne.

Nous allons faire le choix suivant, correspondant à des hypothèses sur les populations probabilisées qui correspondraient aux classes de la partition "vraie" inconnue: puisque la donnée observée à l'instant t répond à l'équation (4.36) que nous rappelons ici :

$$y_t = x^i_{k_i} + w^i_{k_i} \quad , \quad y_t \in \mathbb{R}^n$$

où $w^i_{k_i}$ est un bruit blanc gaussien avec :

$$\mathbb{E}[w^i_{k_i}] = 0 \qquad \mathbb{E}[w^i_{k_i} w^{i\,T}_{j_i}] = R^i \delta_{k_i j_i}$$

Nous dirons que la matrice de dispersion ou d'inertie de chacune des lois $\mathcal{L}_i$ est la covariance R^i.

L'ensemble $\Theta^i \equiv \{x^i, R^i\}$ qui caractérise chaque classe n'étant pas connu, nous devons l'estimer.

IV-4.3) Estimation des densités de probabilité gaussiennes caractérisant chaque classe

D'après le paragraphe précédent nous voyons que le problème d'estimation de la partition "vraie" est donc équivalent à l'estimation récurrente de l'ensemble de paramètres $\Theta^i \equiv \{x^i, R^i\}$.

Puisque nous avons modélisé les classes à l'aide des équations "dynamiques" fictives (4.35) et (4.36) du paragraphe IV-4.1., nous pouvons nous ramener à un problème de filtrage récursif pour estimer les paramètres de l'ensemble Θ^i .

Les équations (4.35) et (4.36) correspondent, pour chaque classe C_i , à un système dynamique stochastique discret où Q^i est aussi inconnu ; l'estimation de x^i, Q^i et R^i peut donc se faire par filtrage adaptatif sous-optimal (voir paragraphe IV-3).

Les particularités de ces filtres par rapport aux équations (4.22) sont les suivantes :

- $\phi = H = \mathbb{1}$

- la dimension de l'état est la même que celle des observations et les systèmes déterministes sous-jacents sont donc observables en un échantillon.

Remarque_3_ : A un instant quelconque du classement on prend comme centre de gravité de la classe C_i la valeur $\hat{x}^i_{K_i/K_i}$. La dispersion apparente qui caractérise la classe estimée sera donc, non plus la dispersion de y_t autour de la valeur vraie x^i mais autour de $\hat{x}^i_{K_i/K_i}$. Ceci entraîne que la classe est réellement caractérisée par

$$S^i_{K_i} = \mathbb{E}\left[(y_t - \hat{x}^i_{K_i/K_i})(y_t - \hat{x}^i_{K_i/K_i})^T\right]$$

en notant $\tilde{x}^i_{K_i/K_i} = x^i - \hat{x}^i_{K_i/K_i}$ on a :

$$S^i_{K_i} = \mathbb{E}\left[(w^i_{K_i} + \tilde{x}^i_{K_i/K_i})(w^i_{K_i} + \tilde{x}^i_{K_i/K_i})^T\right] =$$

$$= R^i_{K_i} + \mathbb{E}\left[w^i_{K_i}\tilde{x}^{iT}_{K_i/K_i}\right] + \mathbb{E}\left[\tilde{x}^{iT}_{K_i/K_i} w^{iT}_{K_i}\right] + P^i_{K_i/K_i}$$

mais $\mathbb{E}\left[w^i_{K_i}\tilde{x}^{iT}_{K_i/K_i}\right] = \mathbb{E}\left[\tilde{x}^i_{K_i/K_i} w^{iT}_{K_i}\right] = 0$ car

$\mathbb{E}\left[w^i_{K_i}\tilde{x}^{iT}_{K_i/K_i}\right] = \mathbb{E}\left[w^i_{K_i}(x^i - \hat{x}^i_{K_i/K_i})^T\right] = \mathbb{E}\left[w^i_{K_i} x^{iT}\right] = 0$

parce que $x^i_{K_i}$ est fonction de x^i_0 et de $w^i_0, w^i_1, \ldots, w^i_{K_i-1}$ d'après les hypothèses ; $x^i_{K_i}$ et $w^i_{K_i}$ sont donc indépendants, donc finalement:

$$S^i_{K_i} = R^i_{K_i} + P^i_{K_i/K_i} \qquad (4.37)$$

or nous ne disposons ni de $R^i_{K_i}$ ni de $P^i_{K_i/K_i}$ mais de $\hat{R}^i_{K_i}$ et

$\hat{P}^i_{K_i/K_i}$ ce qui nous entraîne donc à remplacer pratiquemment $S^i_{K_i}$ par $\hat{S}^i_{K_i} = \hat{R}^i_{K_i} + \hat{P}^i_{K_i/K_i}$

Remarque_4_: si $\lim_{K_i \to \infty} \hat{P}^i_{K_i/K_i} = 0$

alors $$\lim_{K_i \to \infty} \hat{S}^i_{K_i} = \lim_{K_i \to \infty} \hat{R}^i_{K_i}$$

ce qui est en accord avec nos hypothèses et justifie l'affirmation selon laquelle nous procédons à une estimation asymptotique de la partition "vraie". C'est pourquoi nous opérerons à la vérification de la convergence de $\hat{P}^i_{K_i/K_i}$ vers zéro dans les applications.

IV-4.4) Caractéristiques statistiques des estimateurs introduits

Les estimateurs des covariances des bruits de dynamique et d'observation s'écrivent dans notre cas particulier où $\Phi = H = \mathbb{1}$ de la façon suivante :

$$\hat{R}^i_{K_i} = \frac{1}{K_i} \sum_{j=1}^{K_i} \left[(y_j - \hat{x}^i_{j/j-1})(y_j - \hat{x}^i_{j/j-1})^T \right]$$

$$\hat{Q}^i_{K_i} = \frac{1}{K_i} \sum_{j=1}^{K_i} \left[(\hat{x}^i_{j/j} - \hat{x}^i_{j-1/j-1})(\hat{x}^i_{j/j} - \hat{x}^i_{j-1/j-1})^T \right]$$

Nous allons calculer le biais de ces estimateurs :

A) calculons l'espérance mathématique de $\hat{R}^i_{K_i}$

$$\mathbb{E}\left[\hat{R}^i_{K_i}\right] = \frac{1}{K_i} \sum_{j=1}^{K_i} \mathbb{E}\left[(y_j - \hat{x}^i_{j/j-1})(y_j - \hat{x}^i_{j/j-1})^T\right] = \frac{1}{K_i} \sum_{j=1}^{K_i} \mathbb{E}\left[w^i_j w^{iT}_j + \right.$$

$$\left. + w^i_j (x^i_j - \hat{x}^i_{j/j-1})^T + (x^i_j - \hat{x}^i_{j/j-1}) w^{iT}_j + (x^i_j - \hat{x}^i_{j/i-1})(x^i_j - \hat{x}^i_{j/i-1})^T \right] =$$

$$= \frac{1}{K_i} \sum_{j=1}^{K_i} \left[R^i + \mathbb{E}\left[w^i_j (x^i_j - \hat{x}^i_{j/j-1})^T\right] + \mathbb{E}\left[(x^i_j - \hat{x}^i_{j/i-1}) w^{iT}_j\right] + P^i_{j/i-1} \right]$$

mais puisque $\mathbb{E}[w_j^i] = 0$ et w_j^i est indépendant de x_j^i

nous avons :

$$\mathbb{E}\left[\hat{R}_{K_i}^i\right] = R^i + \frac{1}{K_i}\sum P_{j/j-1}^i$$

Donc nous voyons en particulier que si :

$$\lim_{K_i \to \infty} P_{K_i/K_i-1}^i = 0$$

ce qui sera le cas dans notre situation puisque $Q^i = 0$, alors:

$$\lim_{K_i \to \infty} \mathbb{E}\left[\hat{R}_{K_i}^i\right] = R^i$$

donc c'est un estimateur asymptotiquement non biaisé.

B) calculons l'espérance mathématique de $\hat{Q}_{K_i}^i$:

$$\mathbb{E}\left[\hat{Q}_{K_i}^i\right] = \frac{1}{K_i}\sum_{j=1}^{K_i} \mathbb{E}\left[(\hat{x}_{j/j}^i - \hat{x}_{j-1/j-1}^i)(\hat{x}_{j/j}^i - \hat{x}_{j-1/j-1}^i)^T\right] =$$

$$= \frac{1}{K_i}\sum_{j=1}^{K_i} \mathbb{E}\left[\left[(x_j^i - \tilde{x}_{j/j}^i) - (x_{j-1}^i - \tilde{x}_{j-1/j-1}^i)\right]\left[(x_j^i - \tilde{x}_{j/j}^i) - (x_{j-1}^i - \tilde{x}_{j-1/j-1}^i)\right]^T\right]$$

puisque $\hat{x}_{j/j}^i = x_j^i - \tilde{x}_{j/j}^i$

en développant la dernière expression on arrive à :

$$\mathbb{E}\left[\hat{Q}_{K_i}^i\right] = \frac{1}{K_i}\sum_{j=1}^{K_i}\left[Q^i - \mathbb{E}\left[v_{j-1}^i \tilde{x}_{j/j}^{iT}\right] - \mathbb{E}\left[\tilde{x}_{j/j}^i v_{j-1}^{iT}\right] + P_{j/j}^i - \right.$$

$$\left. - \mathbb{E}\left[\tilde{x}_{j/j}^i \tilde{x}_{j-1/j-1}^{iT}\right] - \mathbb{E}\left[\tilde{x}_{j-1/j-1}^i \tilde{x}_{j/j}^{iT}\right] + P_{j-1/j-1}^i\right]$$

d'où l'on obtient :

$$\mathbb{E}\left[\hat{Q}_{K_i}^i\right] = \frac{1}{K_i}\sum_{j=1}^{K_i}\left[Q^i - (P_{j/j}^i - P_{j-1/j-1}^i) - (P_{j/j}^i - P_{j-1/j-1}^i) + P_{j/j}^i - \right.$$

$$\left. - P_{j-1/j-1}^i - P_{j-1/j-1}^i + P_{j-1/j-1}^i\right] = Q^i + \frac{1}{K_i}\sum_{j=1}^{K_i}\left[P_{j-1/j-1}^i - P_{j/j}^i\right]$$

Donc on voit que si pour tout $\varepsilon > 0$ il existe un entier K_i tel que la relation $1 \geq K_i$ implique :

$$\left| P_{1-1/1-1}^i - P_{1/1}^i \right| < \varepsilon$$

alors : $\lim_{K_i \to \infty} \mathbb{E}[\hat{Q}^i_{K_i}] = Q^i$

donc $\hat{Q}^i_{K_i}$ est aussi asymptotiquement non biaisé.

Remarque :

$\hat{P}^i_{K_i/K_i}$ est une estimation de l'erreur d'estimation $P^i_{K_i/K_i}$, comme nous avons posé $Q^i = 0$ cette covariance $P^i_{K_i/K_i}$ doit tendre vers zéro et de même pour son estimateur $\hat{P}^i_{K_i/K_i}$, or ceci n'arrivera que si $\hat{Q}^i_{K_i}$ lui même tend vers zéro (puisque c'est un estimateur de $Q^i = 0$) d'après l'équation (4.20) que nous réécrivons ici avec $\Phi = H = \mathbb{1}$:

$$\hat{P}^i_{K_i/K_i} = \hat{P}^i_{K_i-1/K_i-1} + \hat{Q}^i_{K_i-1} - K_{K_i}\left(\hat{P}^i_{K_i-1/K_i-1} + \hat{Q}^i_{K_i-1}\right)$$

Nous pouvons donc vérifier dans les applications s'il en est ainsi, après l'avoir constaté nous pourrons déduire la convergence de $\hat{R}^i_{K_i}$ vers R^i et aussi de $\hat{x}^i_{K_i}$ vers x^i.

Nous allons finir ce paragraphe en faisant une récapitulation des équations des filtres sous-optimaux adaptatifs avec $\Phi = H = \mathbb{1}$

* Prédicteur à un pas :

$$\hat{x}^i_{K_i/K_i} = \hat{x}^i_{K_i-1/K_i-1}$$

* Erreur de prédiction à un pas :

$$\hat{P}^i_{K_i/K_i-1} = \hat{P}^i_{K_i-1/K_i-1} + \hat{Q}^i_{K_i-1}$$

* Processus d'innovation :

$$r^i_{K_i} = y_t - \hat{x}^i_{K_i/K_i-1}$$

* Estimateur du bruit d'observation :

$$\hat{R}^i_{K_i} = \frac{K_i - 1}{K_i} \hat{R}^i_{K_i - 1} + \frac{1}{K_i} \left[r^i_{K_i} r^{i\,T}_{K_i} \right]$$

* Gain du filtre :

$$K^i_{K_i} = \hat{P}^i_{K_i/K_i - 1} \left(\hat{P}^i_{K_i/K_i - 1} + \hat{R}^i_{K_i} \right)^{-1}$$

* Estimation de l'état :

$$\hat{x}^i_{K_i/K_i} = \hat{x}^i_{K_i/K_i - 1} + K^i_{K_i} r^i_{K_i}$$

* Matrice de covariance de l'erreur d'estimation :

$$\hat{P}^i_{K_i} = \hat{P}^i_{K_i/K_i - 1} - K_{K_i} \hat{P}^i_{K_i/K_i - 1}$$

* Estimateur du bruit de dynamique :

$$\hat{Q}^i_{K_i} = \frac{K_i - 1}{K_i} \hat{Q}^i_{K_i - 1} + \frac{1}{K_i} \left[q^i_{K_i} q^{i\,T}_{K_i} \right]$$

où

$$q^i_{K_i} = \hat{x}^i_{K_i/K_i} - \hat{x}^i_{K_i - 1/K_i - 1}$$

IV-4.5) Concept d'évolution du temps dans chaque classe

Nous avons vu que nous avons autant de filtres que de classes, mais les instants d'évolution de chaque filtre ne sont pas synchrones c'est-à-dire que, bien qu'étant modelisés de la même façon, chaque filtre n'effectue un pas que lorsque sa classe correspondante absorbe un nouvel élément c'est-à-dire que dans ce cas les équations (4.22e) et (4.22f) sont appliquées, dans ces conditions, l'équation dynamique associée à cette classe effectuerait aussi un pas, c'est-à-dire si la classe en question est C_i, dans l'équation :

$$x^i_{K_i + 1} = x^i_{K_i} + V^i_{K_i}$$

K_i est incrémenté d'une unité, on constate alors que pour chaque filtre le "temps" K_i correspond au nombre d'éléments réellement attribués à chacune des classes.

IV-4.6) Mesure de l'appartenance d'un élément à une classe. Règle de décision

La donnée observée à l'instant t répondra comme nous l'avons vu précédemment, à l'équation

$$y_t = x^i_{K_i} + w^i_{K_i}$$

où i est à déterminer parmi l'ensemble de classes existantes.

Le filtre i fournissant l'estimation de $x^i_{K_i}$ sera sollicité lorsque la décision prise sera celle de classer y_t dans la classe C_i, alors, comme nous avons vu dans le paragraphe précédent, K_i devient K_i+1.

Notre mesure de l'appartenance de l'élément à classer, représenté par y_t, à chacune des classes C_i dans le but de décider i sera la probabilité a posteriori suivante : $Pr(C_i|y_t)$ cette probabilité peut être calculée à partir de la probabilité a priori $Pr(C_i)$ de chaque classe et de la densité de probabilité conditionnelle $p(y_t|C_i)$ d'après le théorème suivant (règle de Bayes) :

$$Pr(C_i|y_t) = \frac{p(y_t|C_i) \cdot Pr(C_i)}{p(y_t)}$$

Si nous avons N classes existantes il faudra comparer tous les $Pr(C_i|y_t)$ pour i=1 à N, mais puisque $p(y_t)$ est commun aux N classes et si nous faisons l'hypothèse que les classes sont équiprobables alors on constate que comparer les probabilités $Pr(C_i|y_t)$ revient à comparer les densités de probabilité $p(y_t|C_i)$.

On utilise alors les équations du filtre adaptatif sous-optimal

décrit dans le paragraphe IV-3 pour estimer ces densités de probabilités, inconnues a priori, pour toutes les valeurs de i de 1 à N.

La règle de décision du maximum de vraisemblance est alors utilisée pour déterminer la classe C_j attribuée, c'est-à-dire :

$$y_t \in C_j \quad \text{si} \quad p(y_t|C_j) = \max_i \left[p(y_t|C_i)\right] \geqslant \alpha \qquad (4.38)$$

c'est-à-dire qu'une observation y_t sera d'autant plus proche à la classe C_j que la densité $p(\cdot|C_j)$ sera grande en y_t.

L'introduction du seuil α de l'équation (4.38) est fait pour éviter de fonder une décision de classement sur des probabilités trop faibles, le choix de ce seuil est subjectif et il a beaucoup d'influence sur le nombre de classes effectivement créées.

IV-4.7) Initialisation et croissance du nombre de classes

Nous avons choisi d'initialiser l'algorithme de la façon suivante, la 1ère observation y_1 est assignée à la classe C_1 et nous caractérisons ensuite cette classe par une moyenne ou centre de gravité égal à y_1 et une matrice de covariance S_1 qui définit son ellipsoide d'inertie égal à $\hat{R}_o$ (connu).

Nous augmentons le nombre de classes dans le cas envisagé dans le paragraphe précédent, c'est-à-dire si :

$$\max_i \left[p(y_t|C_i)\right] < \alpha$$

Une nouvelle classe C_{N+1} sera alors initialisée avec une moyenne égale à y_t et une matrice de covariance S_{N+1} égale à

$\hat{R}_o$ (connu).

La figure (4.b) donne une interprétation graphique du phénomène de croissance.

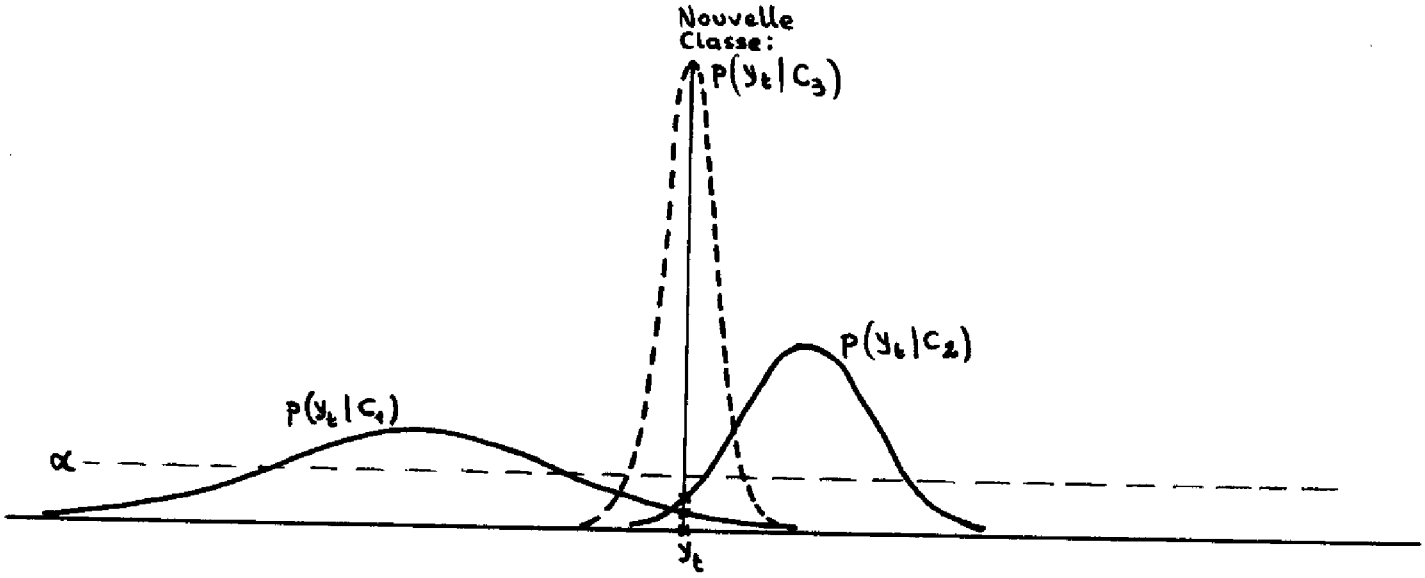
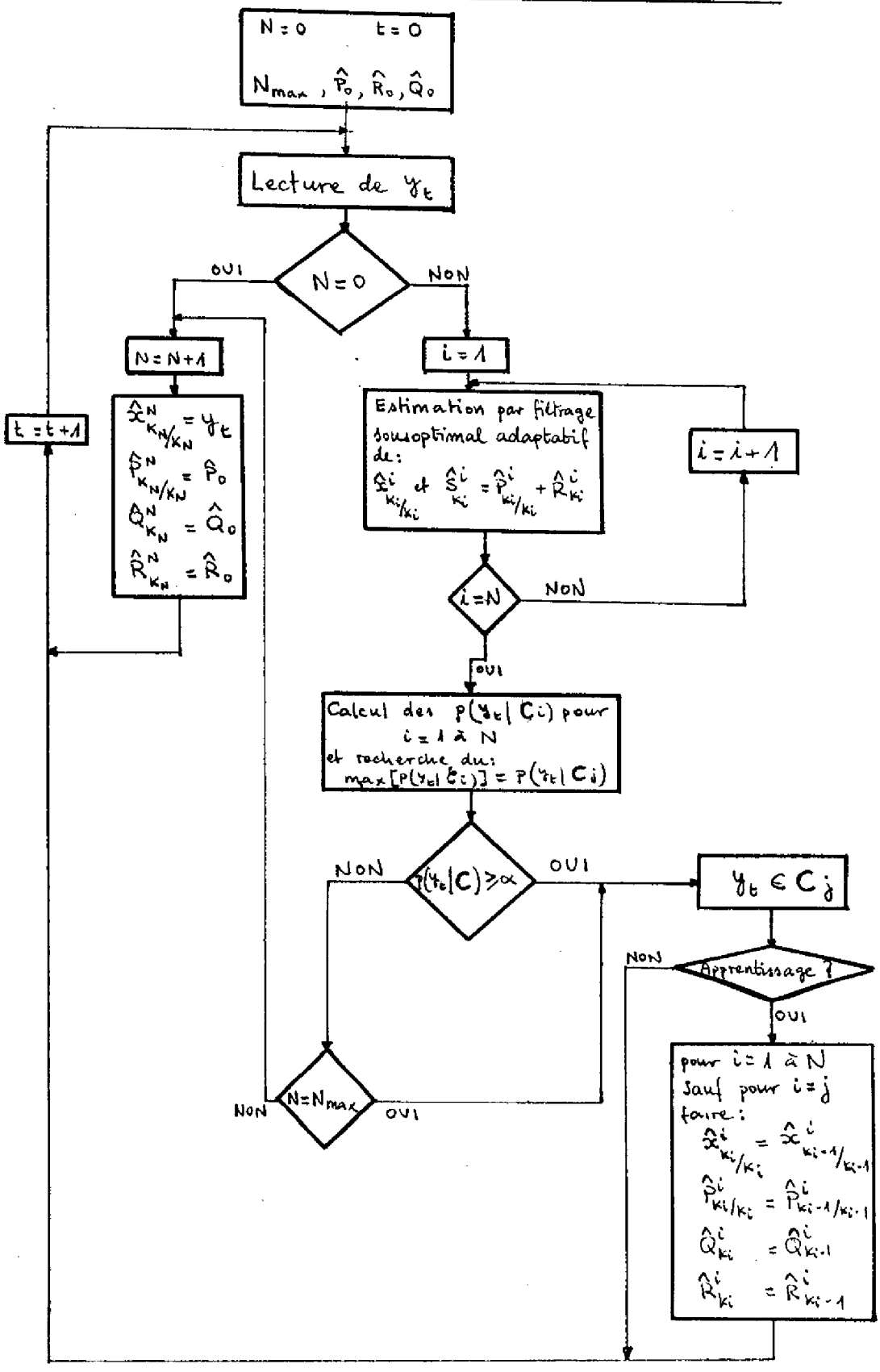


Fig. 4.b

IV-4.8) Organigramme de l'algorithme de classement



IV-4.9) Exemples d'application de données multidimensionnelles

Dans ce paragraphe nous allons traiter deux applications pour illustrer l'algorithme, la 1ère va être une application au problème de la reconnaissance des composants d'un mélange de densités de probabilité gaussiennes bidimensionnelles ; la 2ème va être une application à des données non gaussiennes dans le but de tester la robustesse de l'algorithme, il s'agit d'une application à la reconnaissance tactile de solides.

A) Application au problème de la reconnaissance des composants d'un mélange de densités de probabilité gaussiennes bidimensionnelles

Les méthodes d'Analyse des données donnent, à partir d'un échantillon multidimensionnel, une description de la population. Ces méthodes ne posent aucune hypothèse de distribution et ne font appel à aucun modèle probabiliste, bien que le problème peut être posé de façon à adapter par une technique convenable, un modèle stochastique à un phénomène observé.

Dans le problème de la reconnaissance des composants d'un mélange de densités de probabilité il s'agit de chercher à détecter dans un échantillon donné, l'existence possible de sous-ensembles qui seraient échantillons de lois de probabilité d'un type connu.

Plus précisément, étant donnée une distribution empirique on essayera de savoir si elle résulte des effets de plusieurs phénomènes aléatoires suivant différentes lois de distribution de type connu.

Si ces phénomènes peuvent être supposés exhaustifs et indépendants, la fonction de répartition globale F(x) aura la forme suivante :

$$\forall x \in \mathbb{R}^n \qquad F(x) = \sum_{j=1}^{K} P_j \cdot F_j(x) \qquad (4.39)$$

où - P_j : probabilité a priori de la j-ème composante

- $F_j(x)$: fonction de répartition de la j-ème composante

C'est le problème appelé "Mixtures résolution"

Il existe pas mal de techniques pour résoudre ce problème dues pour la plupart à des auteurs américains, par exemple : RAO (1948) [4.27], DAY (1969) [4.9].

La plupart s'appliquent aux mélanges gaussiens et sont très souvent restreintes aux distributions unidimensionnelles et/ou nécessitent un très grand nombre d'observations (BATTACHARYA 1967) [4.5].

Chez tous les auteurs, l'extension au cas multidimensionnel nécessite beaucoup d'hypothèses supplémentaires, par exemple le fait que les distributions ne différent que par leurs moyennes (cas qui devient extrêmement simple par la méthode que nous proposons).

Il existe d'autres méthodes qui procèdent par approximations successives (techniques de type bayesien, d'apprentissage, etc ...) pour estimer le modèle défini par la relation (4.39) à partir des observations, parmi les auteurs qui ont traité ces méthodes nous pouvons citer : PATRICK et HANCOCK (1966) [4.26], AGRAWALA (1970) [4.1], PATRICK et COSTELLO (1970) [4.25],

ces méthodes sont en général très restrictives et fixent a priori le nombre de composants du mélange.

Avec ces méthodes on peut formaliser le problème de la reconnaissance des composants d'un mélange en termes d'apprentissage avec ou sans maître. AGRAWALA (1970) [4.1], PATRICK (1972) [4.24], DUDA et HART (1973) [4.13].

Il existe aussi, dans le cas des mélanges gaussiens une méthode proposée par J.P. BENZECRI basée sur une série de déconvolutions successives [4.4].

A. SHROEDER (1974) [4.31] propose une approche qui cherche à reconnaître, dans une population observée la présence d'échantillons de lois de probabilité connues sans faire d'hypothèses au sujet de la distribution globale, cette approche est basée sur l'algorithme des Nuées Dynamiques de E. DIDAY (1972) [4.10].

L'algorithme des Nuées Dynamiques détecte parallèlement une partition en classes de la population et les traits caractéristiques de ces classes c'est-à-dire les distributions de probabilité, ce qui est le même principe que l'algorithme que nous proposons dans ce chapitre, l'algorithme des Nuées Dynamiques travaille avec un échantillon de dimension quelconque, un nombre quelconque de composants dans le mélange et des distributions de probababilité quelconques mais de type connu.

Sont requis en entrée :

. L'ensemble des données à analyser

. La forme de la famille des lois de probabilité cherchées (lois

gaussiennes, multinomiales, gamma)

. Le nombre de composants recherchés dans le mélange (nombre de classes)

. Une partition initiale des données à analyser.

Cette méthode traite de façon globale l'ensemble des données et itère sur les regroupements jusqu'au moment où une position d'équilibre (correspondant à un optimum local d'un critère lié à la dispersion dans chaque classe) est atteinte.

La méthode des Nuées Dynamiques peut être considérée comme une généralisation d'un ensemble de procédés appelés "Itération relocation procédure" par certains auteurs (Wishart (1971)), K-means par d'autres (Mc Queen (1967), Watanabe (1972)).

Si les lois de probabilité sont gaussiennes nous pouvons résoudre le problème de "mixture" à l'aide de l'algorithme de classement de données gaussiennes que nous avons présenté dans ce chapitre, notre approche ne nécessite ni la connaissance du nombre de composants du mélange ni une partition initiale.

Elle fournit, parallèlement, une partition en classes de l'ensemble de données et les noyaux (moyenne et covariance) de chacun des composants du mélange. De plus elle traite de façon itérative les données de telle façon qu'elle fournit à tout instant un classement des données déjà traitées.

* Exemples d'application à des données construites par simulation

a) un échantillon artificiel bidimensionnel de 200 points (voir fig. 4.c), tirés de façon aléatoire suivant une loi

uniforme, des quatre distributions gaussiennes suivantes :

$$x^1 = \begin{bmatrix} 3 \\ 3 \end{bmatrix} \qquad x^2 = \begin{bmatrix} 3 \\ -3 \end{bmatrix} \qquad x^3 = \begin{bmatrix} -3 \\ 3 \end{bmatrix} \qquad x^4 = \begin{bmatrix} -3 \\ -3 \end{bmatrix}$$

$$S^1 = \begin{bmatrix} 1.96 & 2 \\ 2 & 4 \end{bmatrix} \quad S^2 = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix} \quad S^3 = \begin{bmatrix} 2.25 & 2 \\ 2 & 3 \end{bmatrix} \quad S^4 = \begin{bmatrix} 0.25 & 0.25 \\ 0.25 & 1 \end{bmatrix}$$

La fig 4.d montre les ellipsoïdes d'équiprobabilité correspondantes.

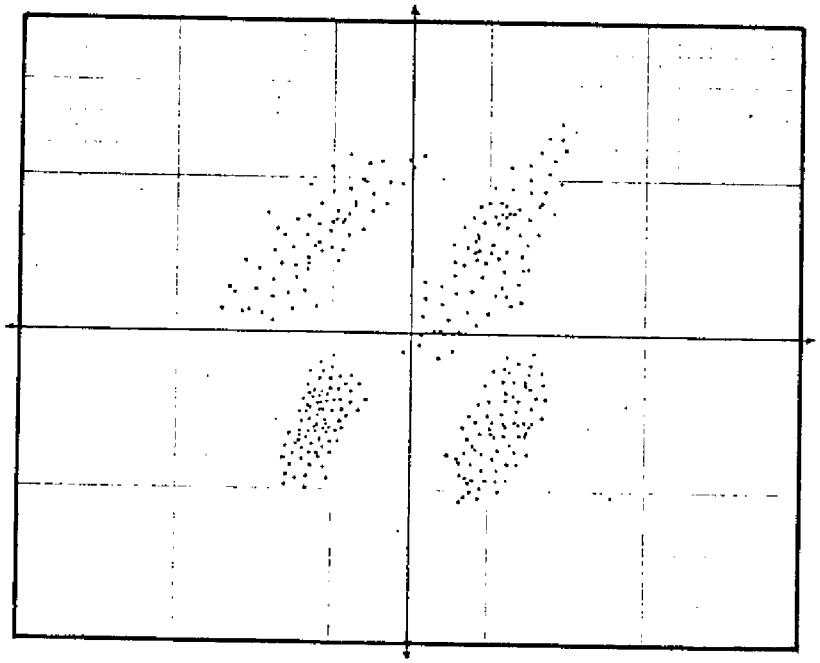


Fig. 4.c

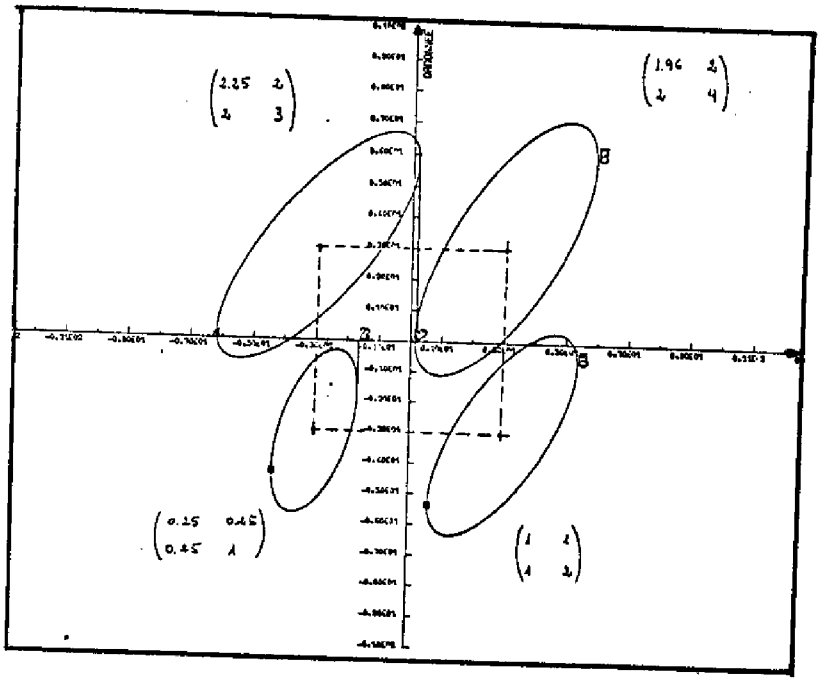


Fig. 4.d

D'après la façon dont nous avons modelisé les classes on rappelle qu'il faut que les $\hat{Q}^i_{K_i}$ tendent vers zéro pour tout i, puisque la vraie valeur des Q^i est zéro pout tout i. Suivant les équations du filtre sous-optimal on s'aperçoit que les $\hat{P}^i_{K_i/K_i}$ doivent tendre aussi vers zéro puisqu'ils sont

fonction des $\hat{Q}^i_{k_i}$

Dans l'illustration numérique de ce chapitre nous remarquons (fig. 4.f) qu'effectivement les $\hat{Q}^i_{k_i}$ tendent vers zéro.

La fig. 4.e ci-dessous montre les ellipses d'équiprobabilité obtenues avec notre algorithme, nous remarquons qu'on s'approche assez bien des ellipses d'équiprobabilité des données initiales représentées dans la figure 4.d.

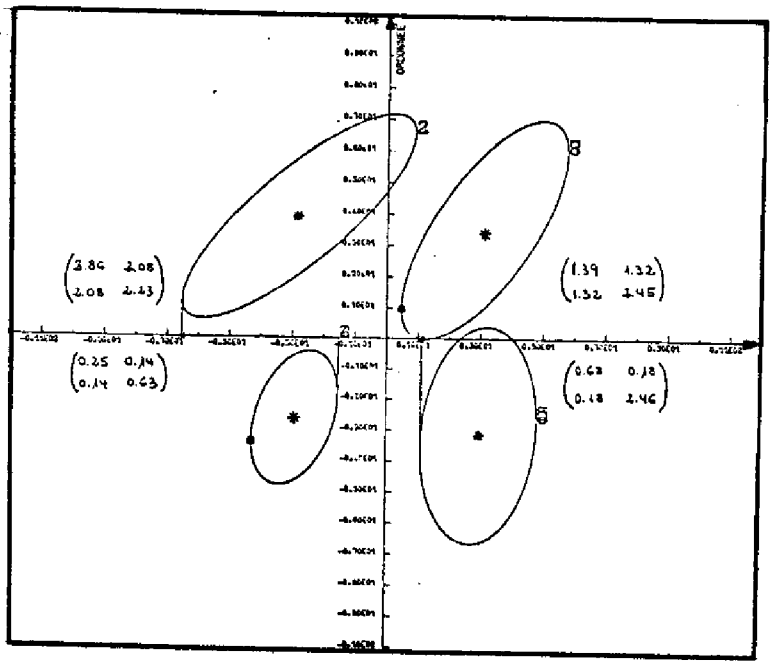
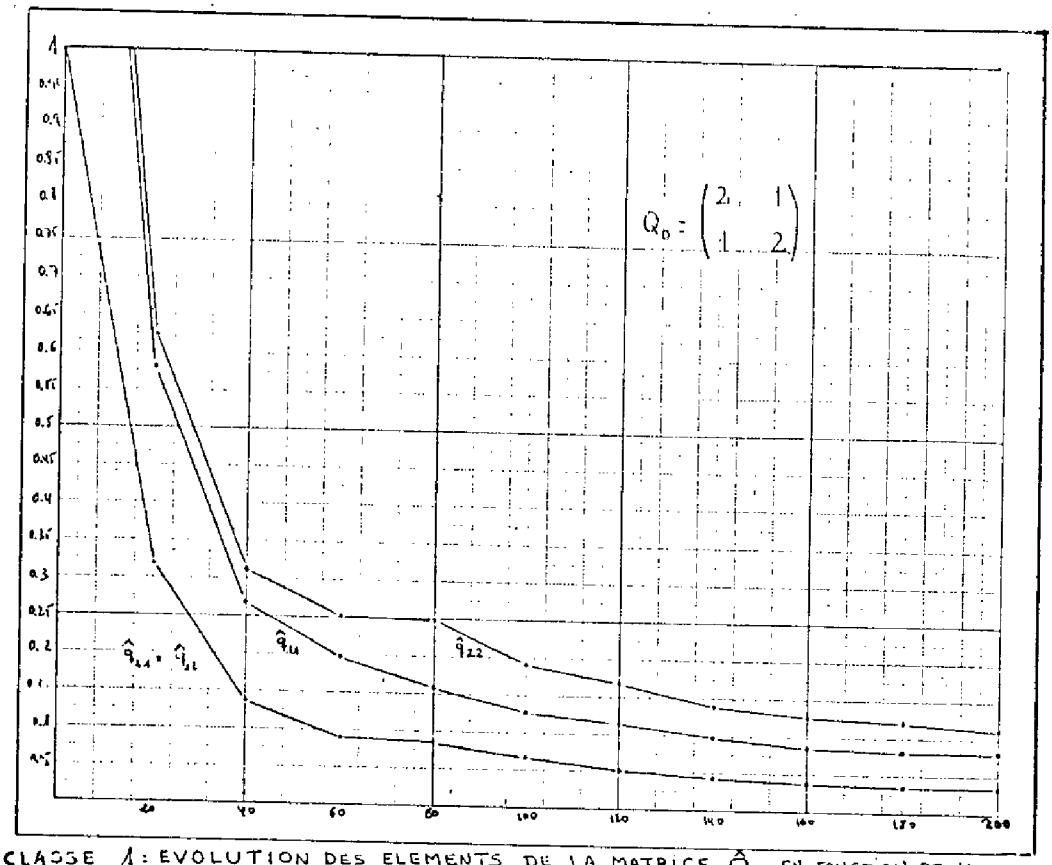
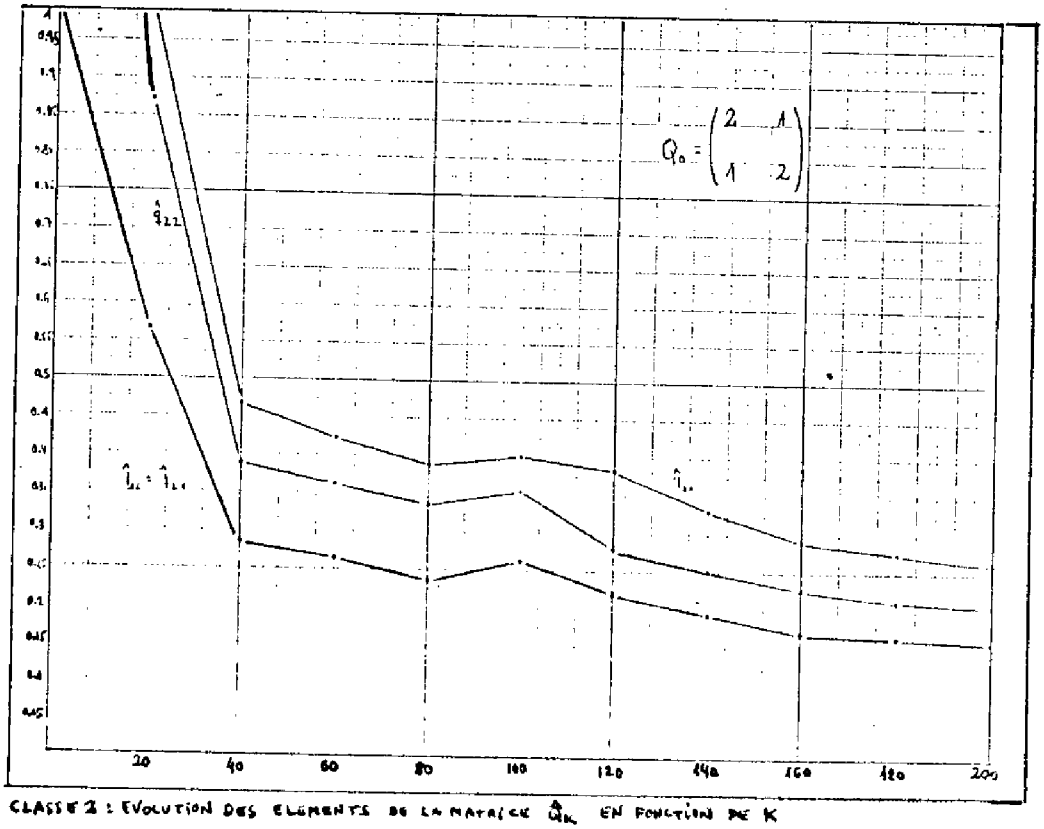


Fig. 4.e



CLASSE 1 : EVOLUTION DES ELEMENTS DE LA MATRICE $\hat{Q}_K$ EN FONCTION DE K



CLASSE 2 : EVOLUTION DES ELEMENTS DE LA MATRICE $\hat{Q}_K$ EN FONCTION DE K

FIGURE 4.f

Les résultats obtenus sont, donc, les suivants (voir fig. 4.e et tableau T.4-1)

$$\hat{X}^1 = \begin{bmatrix} 3.11 \\ 3.49 \end{bmatrix} \qquad \hat{X}^2 = \begin{bmatrix} 2.95 \\ -3.13 \end{bmatrix} \qquad \hat{X}^3 = \begin{bmatrix} -2.82 \\ 3.89 \end{bmatrix} \qquad \hat{X}^4 = \begin{bmatrix} -2.91 \\ -2.61 \end{bmatrix}$$

$$\hat{S}^1 = \begin{bmatrix} 1.39 & 1.32 \\ 1.32 & 2.45 \end{bmatrix} \qquad \hat{S}^2 = \begin{bmatrix} 0.68 & 0.18 \\ 0.18 & 2.46 \end{bmatrix} \qquad \hat{S}^3 = \begin{bmatrix} 2.86 & 2.08 \\ 2.08 & 3.23 \end{bmatrix} \qquad \hat{S}^4 = \begin{bmatrix} 0.25 & 0.14 \\ 0.14 & 0.63 \end{bmatrix}$$

Ces résultats ont été obtenus avec un seuil $\alpha = 0.03$. Sur les 200 éléments classés il n'y a eu aucune erreur de classement, c'est-à-dire que la discrimination est parfaite pour cet exemple, ceci est sans doute dû au fait que les classes sont assez séparées a priori (voir les valeurs obtenues concernant la différence entre le classement obtenu et le "vraie" ainsi que sa distance dans le tableau T.4-1).

La fig.4.g montre la convergence des $\hat{P}_{K_i}^i$ pour chacune des 4 classes obtenues par l'algorithme.

```
LA CLASSE 3 EST FORMEE DE  56 ELEMENTS DONT
  56 DE L'ANCIENNE CLASSE 1
   0 DE L'ANCIENNE CLASSE 2
   0 DE L'ANCIENNE CLASSE 3
   0 DE L'ANCIENNE CLASSE 4
LA CLASSE 4 EST FORMEE DE  50 ELEMENTS DONT
   0 DE L'ANCIENNE CLASSE 1
  50 DE L'ANCIENNE CLASSE 2
   0 DE L'ANCIENNE CLASSE 3
   0 DE L'ANCIENNE CLASSE 4
200      ELEMENTS CLASSES
ICARF=  4ICARC=  4
         MATRICE DE CONTINGENCE
 [0.0      0.0      0.0      0.2550]
 [0.0      0.0      0.2150   0.0   ]
 [0.2800   0.0      0.0      0.0   ]
 [0.0      0.2500   0.0      0.0   ]
         DIFFERENCE ENTRE LES DEUX CLASSIFICATIONS=  0.0
         DISTANCE ENTRE LES DEUX PARTITIONS=  0.0
•CLASSEMENT QUANTITATIF
DISTANCE GAUSSIENNE

ATTRIBUTS DE LA CLASSE     1
   VECTEUR DE MOYENNE ET MATRICE DE COVARIANCE
[-2.91]      [0.2510   0.1397]
[-2.61]      [0.1397   0.6351]
CLASSE     1
     MATRICES : Q̂  R̂  P̂
                                Q̂= [0.0624   0.0249]
                                   [0.0249   0.0970]
                                R̂= [0.1732   0.1007]
                                   [0.1007   0.4703]
                                P̂= [0.0778   0.0390]
                                   [0.0390   0.1648]
ATTRIBUTS DE LA CLASSE     2
   VECTEUR DE MOYENNE ET MATRICE DE COVARIANCE
[-2.82]      [2.8604   2.0852]
[ 3.89]      [2.0852   2.2287]
CLASSE     2
     MATRICES : Q̂  R̂  P̂
                                Q̂= [0.2293   0.1401]
                                   [0.1401   0.1744]
                                R̂= [2.2408   1.6542]
                                   [1.6542   1.7504]
                                P̂= [0.6196   0.4310]
                                   [0.4310   0.4783]
ATTRIBUTS DE LA CLASSE     3
   VECTEUR DE MOYENNE ET MATRICE DE COVARIANCE
[3.11]       [1.3935   1.3242]
[3.49]       [1.3242   2.4460]
CLASSE     3
     MATRICES : Q̂  R̂  P̂
                                Q̂= [0.1490   0.1078]
                                   [0.1078   0.1871]
                                R̂= [1.0624   1.0375]
                                   [1.0375   1.9284]
                                P̂= [0.3362   0.2967]
                                   [0.2967   0.5176]
ATTRIBUTS DE LA CLASSE     4
   VECTEUR DE MOYENNE ET MATRICE DE COVARIANCE

[ 2.95]      [0.4864   0.1779]
[-3.13]      [0.1779   2.4570]
CLASSE     4
     MATRICES : Q̂  R̂  P̂
                                Q̂= [0.0432   0.0334]
                                   [0.0334   0.0625]
                                R̂= [0.5364   0.1141]
                                   [0.1141   2.1369]
                                P̂= [0.1519   0.0638]
                                   [0.0638   0.3202]
```

TABLEAU T.4-1

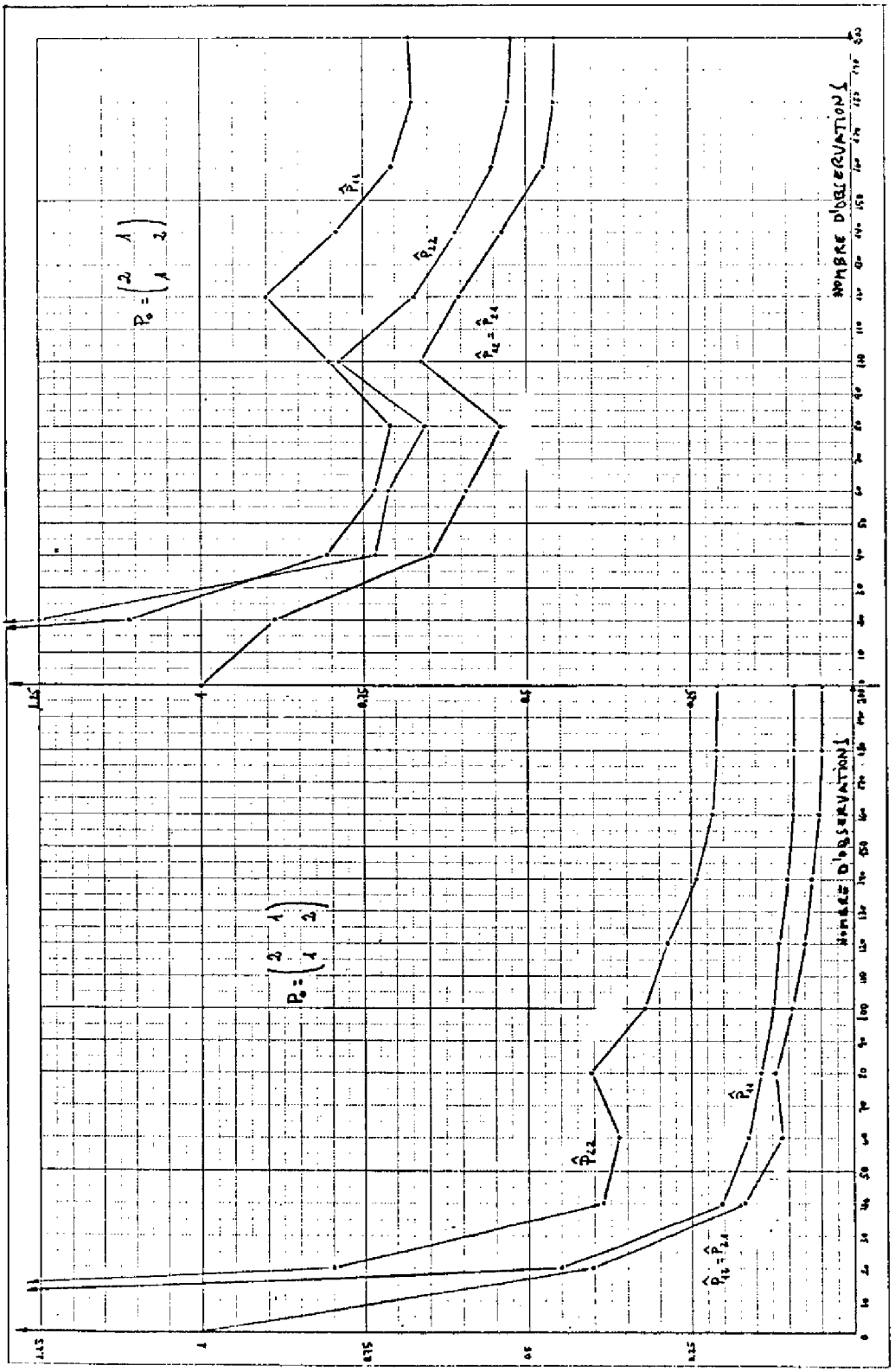

CLASSE 1 : CONVERGENCE DES ELEMENTS DE $\hat{P}^1_{k_2}$
CLASSE 2 : CONVERGENCE DES ELEMENTS DE $\hat{P}^2_{k_2}$
NOMBRE D'OBSERVATIONS
$P_0 = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$
$\hat{P}_{11}$
$\hat{P}_{12}$
$\hat{P}_{12} = \hat{P}_{21}$
$\hat{P}_{22} = \hat{P}_{21}$

FIGURE 4.g

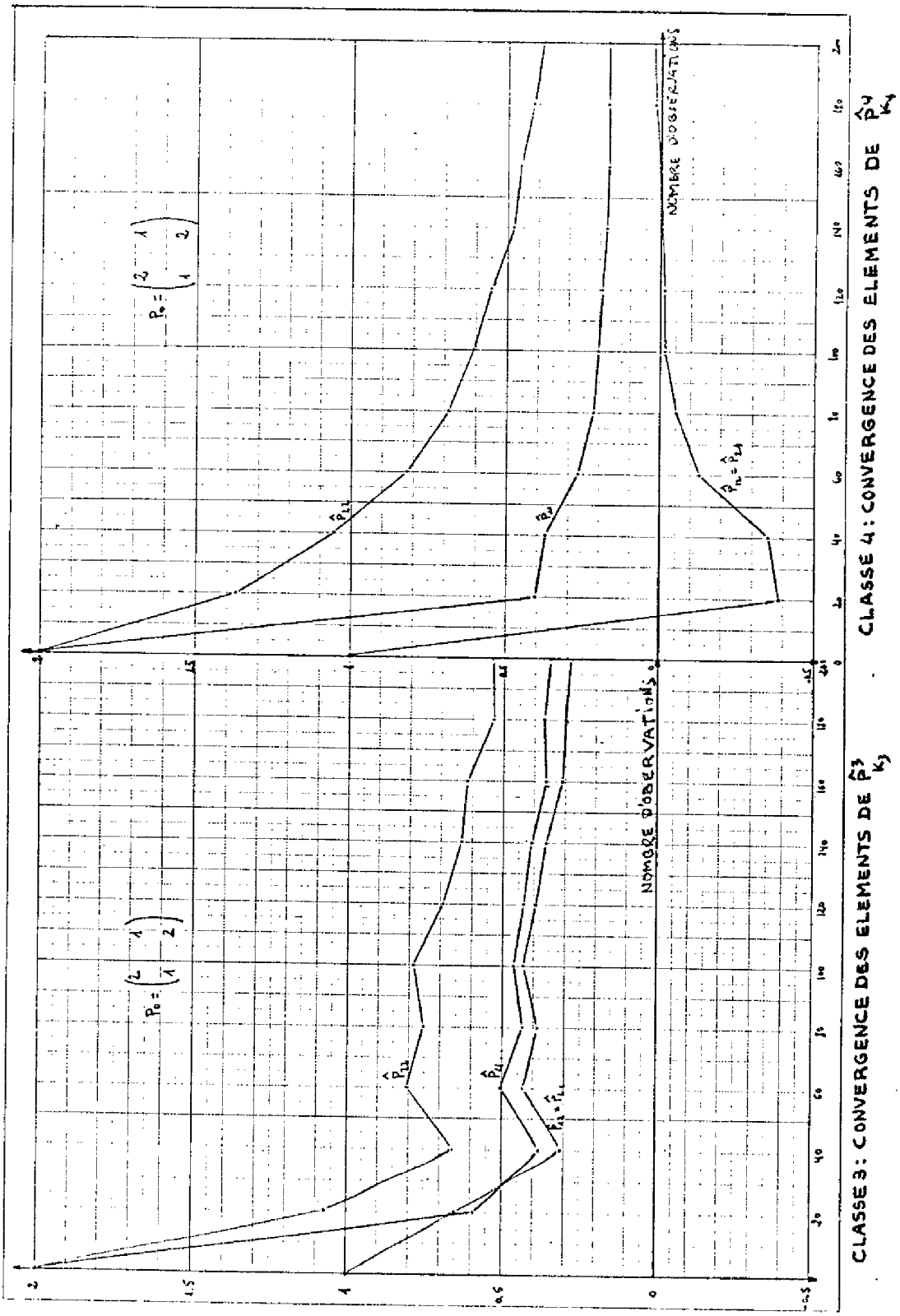


FIGURE 4.g (suite)

b) 320 points de $\mathbb{R}^2$ tirés de quatre distributions gaussiennes bidimensionnelles (voir fig. 4.h et fig. 4.i) de paramètres :

$$x^1 = \begin{bmatrix} 3 \\ 3 \end{bmatrix} \qquad x^2 = \begin{bmatrix} 3 \\ -3 \end{bmatrix} \qquad x^3 = \begin{bmatrix} -3 \\ 3 \end{bmatrix} \qquad x^4 = \begin{bmatrix} -3 \\ -3 \end{bmatrix}$$

$$S^1 = \begin{bmatrix} 4 & 4 \\ 4 & 6.25 \end{bmatrix} \qquad S^2 = \begin{bmatrix} 1 & 1 \\ 1 & 4 \end{bmatrix} \qquad S^3 = \begin{bmatrix} 4 & 4 \\ 4 & 9 \end{bmatrix} \qquad S^4 = \begin{bmatrix} 1/4 & 1/4 \\ 1/4 & 1 \end{bmatrix}$$

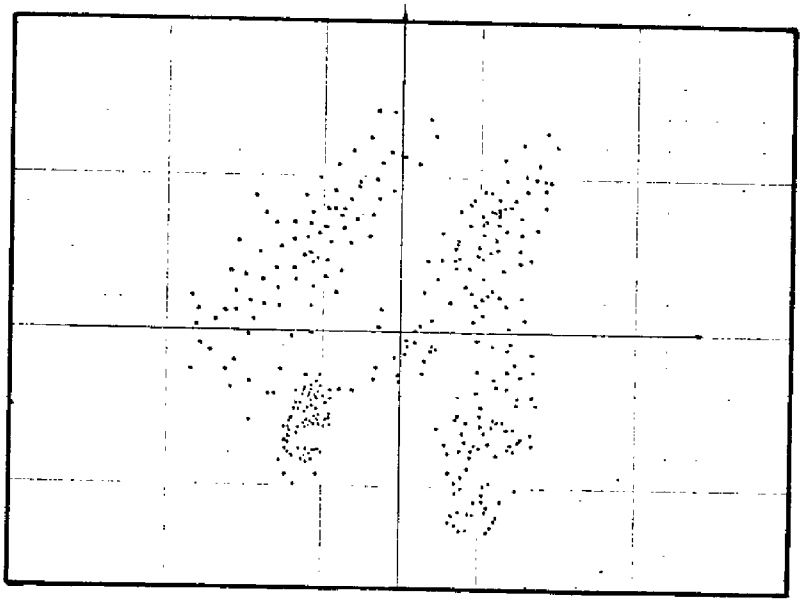


Figure 4.h

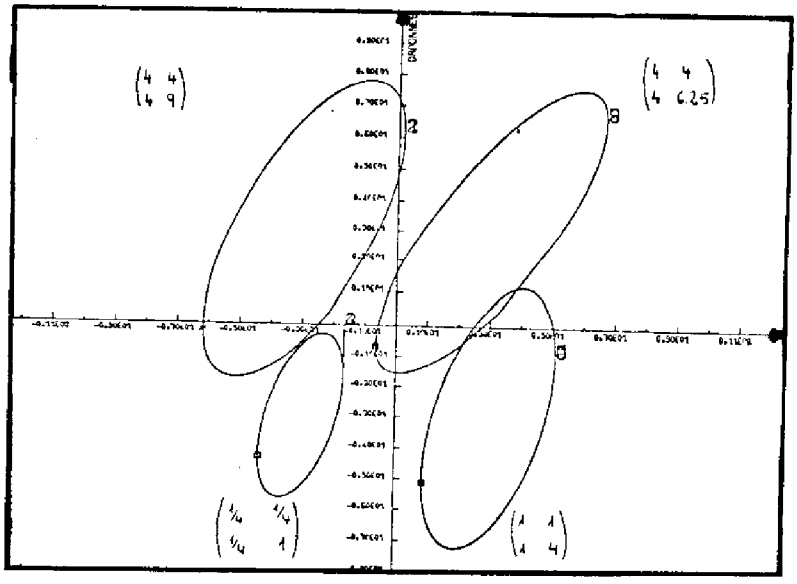


Figure 4.i

Nous avons obtenu les résultats suivants (voir fig. 4.j et Tableau T.4-2)

$$\hat{X}^1 = \begin{bmatrix} 2.6 \\ 2.87 \end{bmatrix} \quad \hat{X}^2 = \begin{bmatrix} 2.92 \\ -2.92 \end{bmatrix} \quad \hat{X}^3 = \begin{bmatrix} -2.9 \\ 4.2 \end{bmatrix} \quad \hat{X}^4 = \begin{bmatrix} -2.98 \\ -2.71 \end{bmatrix}$$

$$\hat{S}^1 = \begin{bmatrix} 3.4 & 3.6 \\ 3.6 & 4.9 \end{bmatrix} \quad \hat{S}^2 = \begin{bmatrix} 0.64 & 0.65 \\ 0.65 & 4.3 \end{bmatrix} \quad \hat{S}^3 = \begin{bmatrix} 5.4 & 4.3 \\ 4.3 & 5.98 \end{bmatrix} \quad \hat{S}^4 = \begin{bmatrix} 0.88 & 0.57 \\ 0.57 & 1.18 \end{bmatrix}$$

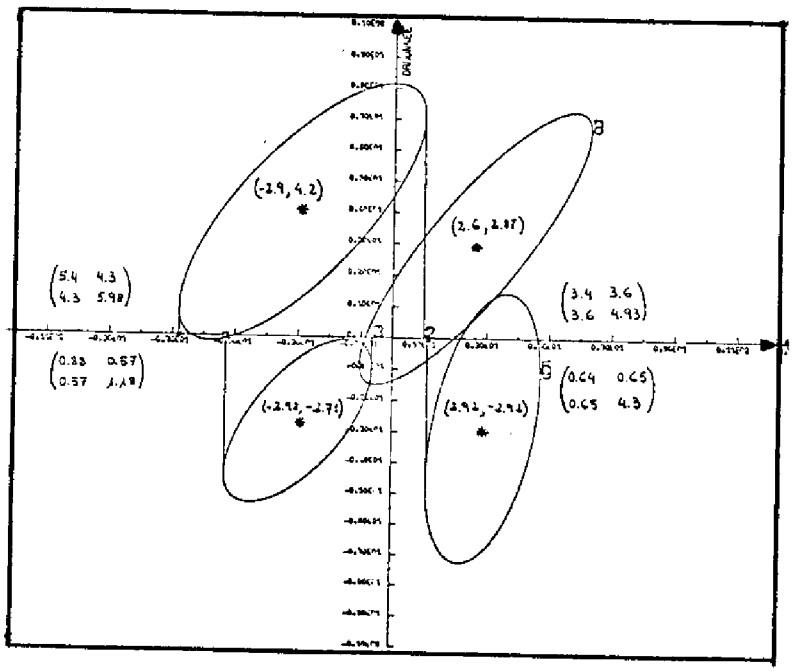


Fig. 4.j

Cet exemple montre que notre algorithme est assez efficace dans le cas où les classes ne sont pas nettement séparées a priori.

Le classement obtenu ne diffère pas trop du réel puisque sur 320 éléments il n'y a que 14 éléments mal classés (voir les valeurs prises par l'indice de dissemblance dans le tableau T.4-2 ainsi que la distance entre la classement "vrai" et l'obtenu).

La figure 4.k montre comment la différence entre le classement réel et le classement obtenu diminue en fonction du nombre d'observations à cause du phénomène d'autoapprentissage, nous avons considéré deux cas : le classement "historique" et le classement "instantané".

- Le classement "historique" est celui qui correspond à la situation rencontrée au fur et à mesure que les échantillons arrivent, la règle de décision dans ce cas est basée sur un ensemble de noyaux qui varie avec le temps. Le classement "historique" est un simple enregistrement de la suite des décisions.

- Le classement "instantané" est basé sur un ensemble de noyaux fixe qui est celui obtenu à l'instant t après la fin du classement "historique", c'est-à-dire qu'on fait un deuxième passage des données soumises toutes à la même règle de décision.

Les figures 4.l et 4.m que nous présentons par la suite, montrent respectivement la convergence vers zéro des $\hat{Q}^i_{K_i}$ et des $\hat{P}^i_{K_i}$.

ATTRIBUTS DE LA CLASSE 1
VECTEUR DE MOYENNE ET MATRICE DE COVARIANCE

$$\begin{bmatrix}-2.98\\-2.71\end{bmatrix} \qquad \begin{bmatrix}0.8563 & 0.5760\\0.5760 & 1.1831\end{bmatrix}$$

CLASSE 1
MATRICES : $\hat{Q}$ $\hat{R}$ $\hat{P}$

$$\hat{Q} = \begin{bmatrix}0.0764 & 0.0252\\0.0252 & 0.0793\end{bmatrix}$$

$$\hat{R} = \begin{bmatrix}0.6021 & 0.4714\\0.4714 & 0.9444\end{bmatrix}$$

$$\hat{P} = \begin{bmatrix}0.1962 & 0.1046\\0.1046 & 0.2387\end{bmatrix}$$

ATTRIBUTS DE LA CLASSE 2
VECTEUR DE MOYENNE ET MATRICE DE COVARIANCE

$$\begin{bmatrix}-2.91\\4.24\end{bmatrix} \qquad \begin{bmatrix}5.4323 & 4.3454\\4.3454 & 5.0815\end{bmatrix}$$

CLASSE 2
MATRICES : $\hat{Q}$ $\hat{R}$ $\hat{P}$

$$\hat{Q} = \begin{bmatrix}0.2650 & 0.1575\\0.1575 & 0.2558\end{bmatrix}$$

$$\hat{R} = \begin{bmatrix}4.4570 & 3.6332\\3.6332 & 4.0628\end{bmatrix}$$

$$\hat{P} = \begin{bmatrix}0.9754 & 0.7122\\0.7122 & 1.0186\end{bmatrix}$$

ATTRIBUTS DE LA CLASSE 3
VECTEUR DE MOYENNE ET MATRICE DE COVARIANCE

$$\begin{bmatrix}2.92\\-2.92\end{bmatrix} \qquad \begin{bmatrix}0.6416 & 0.6566\\0.6566 & 4.3000\end{bmatrix}$$

CLASSE 3
MATRICES : $\hat{Q}$ $\hat{R}$ $\hat{P}$

$$\hat{Q} = \begin{bmatrix}0.1005 & 0.0465\\0.0465 & 0.1646\end{bmatrix}$$

$$\hat{R} = \begin{bmatrix}0.4706 & 0.5321\\0.5321 & 3.6057\end{bmatrix}$$

$$\hat{P} = \begin{bmatrix}0.1710 & 0.1245\\0.1245 & 0.6943\end{bmatrix}$$

ATTRIBUTS DE LA CLASSE 4
VECTEUR DE MOYENNE ET MATRICE DE COVARIANCE

$$\begin{bmatrix}2.60\\2.87\end{bmatrix} \qquad \begin{bmatrix}3.3910 & 3.5862\\3.5862 & 4.0342\end{bmatrix}$$

CLASSE 4
MATRICES : $\hat{Q}$ $\hat{R}$ $\hat{P}$

$$\hat{Q} = \begin{bmatrix}0.0911 & 0.0655\\0.0655 & 0.1131\end{bmatrix}$$

$$\hat{R} = \begin{bmatrix}2.9263 & 3.1439\\3.1439 & 4.2927\end{bmatrix}$$

$$\hat{P} = \begin{bmatrix}0.4647 & 0.4423\\0.4423 & 0.6415\end{bmatrix}$$

LA CLASSE 1 EST FORMEE DE 92 ELEMENTS DONT
8 DE L'ANCIENNE CLASSE 1
0 DE L'ANCIENNE CLASSE 2
2 DE L'ANCIENNE CLASSE 3
82 DE L'ANCIENNE CLASSE 4
LA CLASSE 2 EST FORMEE DE 65 ELEMENTS DONT
0 DE L'ANCIENNE CLASSE 1
0 DE L'ANCIENNE CLASSE 2
65 DE L'ANCIENNE CLASSE 3
0 DE L'ANCIENNE CLASSE 4
LA CLASSE 3 EST FORMEE DE 87 ELEMENTS DONT
4 DE L'ANCIENNE CLASSE 1
83 DE L'ANCIENNE CLASSE 2
0 DE L'ANCIENNE CLASSE 3
0 DE L'ANCIENNE CLASSE 4
LA CLASSE 4 EST FORMEE DE 76 ELEMENTS DONT
76 DE L'ANCIENNE CLASSE 1
0 DE L'ANCIENNE CLASSE 2
0 DE L'ANCIENNE CLASSE 3
0 DE L'ANCIENNE CLASSE 4
320 ELEMENTS CLASSES

TABLEAU T.4-2

ICARE= 4ICARG= 4
MATRICE DE CONTINGENCE

0.0250	0.0	0.0062	0.2562
0.0	0.0	0.2031	0.0
0.0125	0.2594	0.0	0.0
0.2375	0.0	0.0	0.0

DIFFERENCE ENTRE LES DEUX CLASSIFICATIONS= 0.1071
DISTANCE ENTRE LES DEUX PARTITIONS= 0.1047

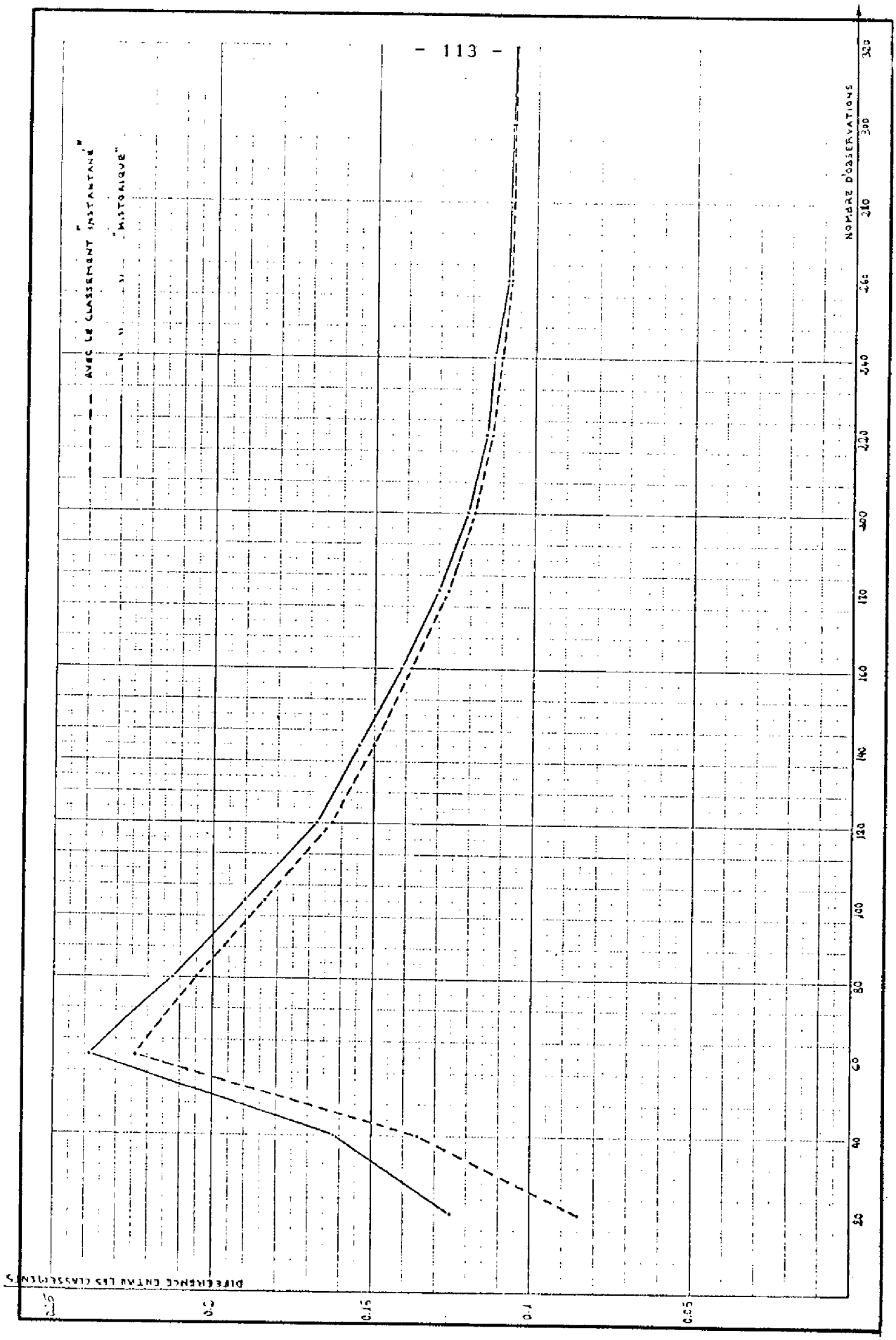
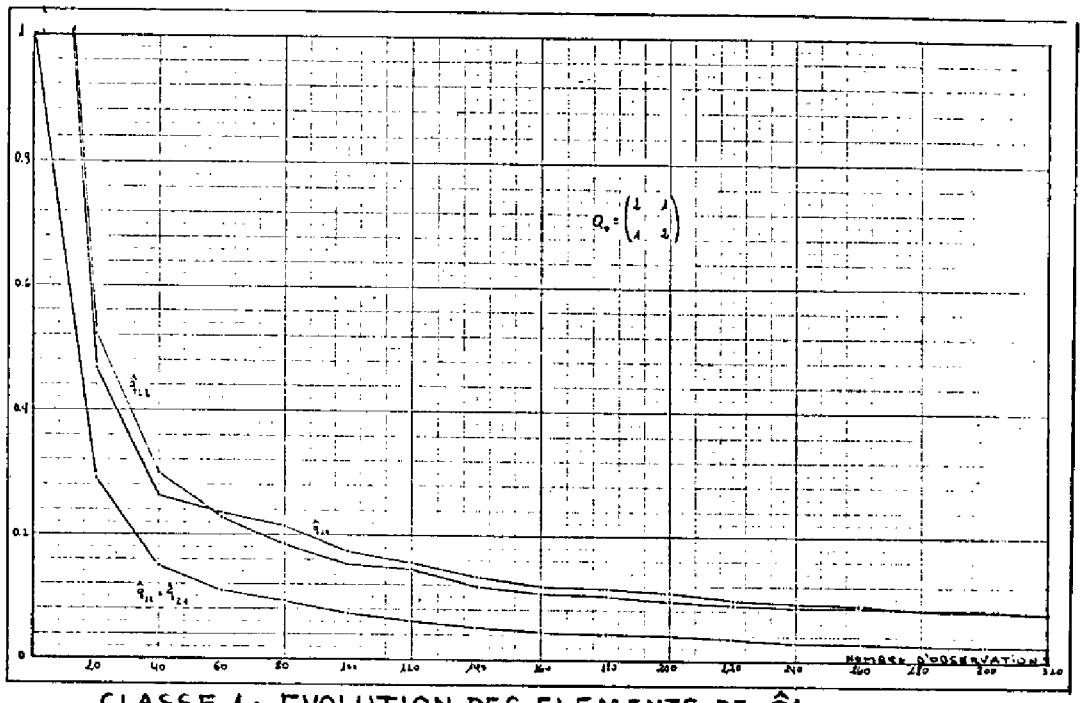
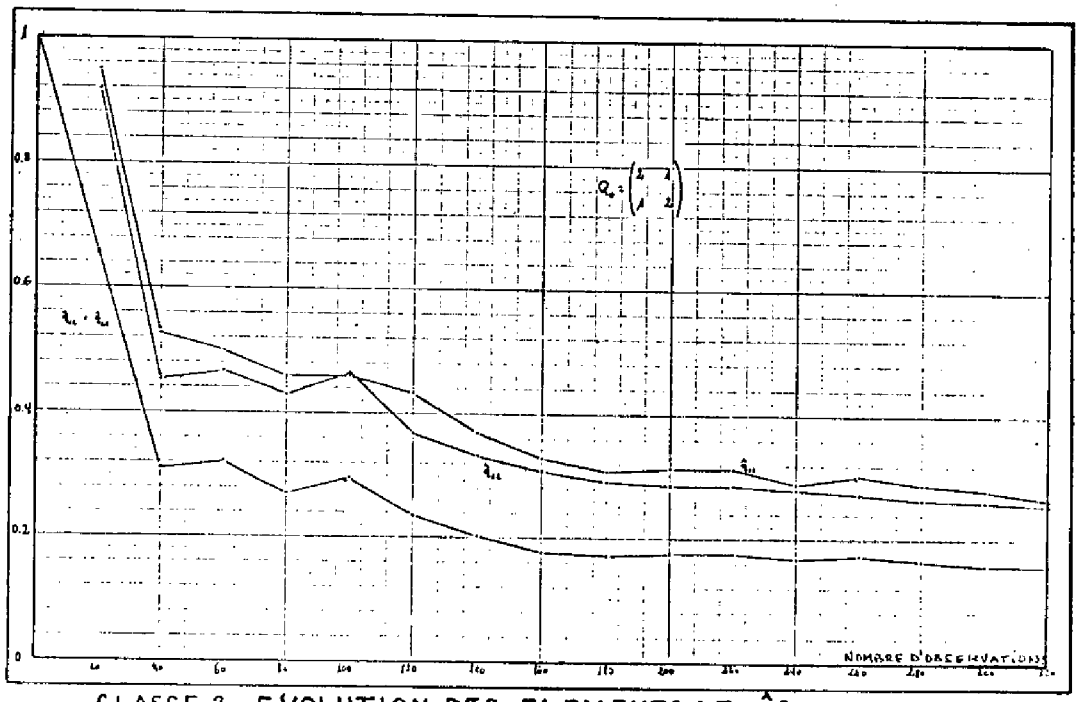

AVEC LE CLASSEMENT "INSTANTANE"
"HISTORIQUE"
NOMBRE D'OBSERVATIONS
DIFFERENCE ENTRE LES CLASSEMENTS
0.0
0.05
0.1
0.15

Fig. 4.k

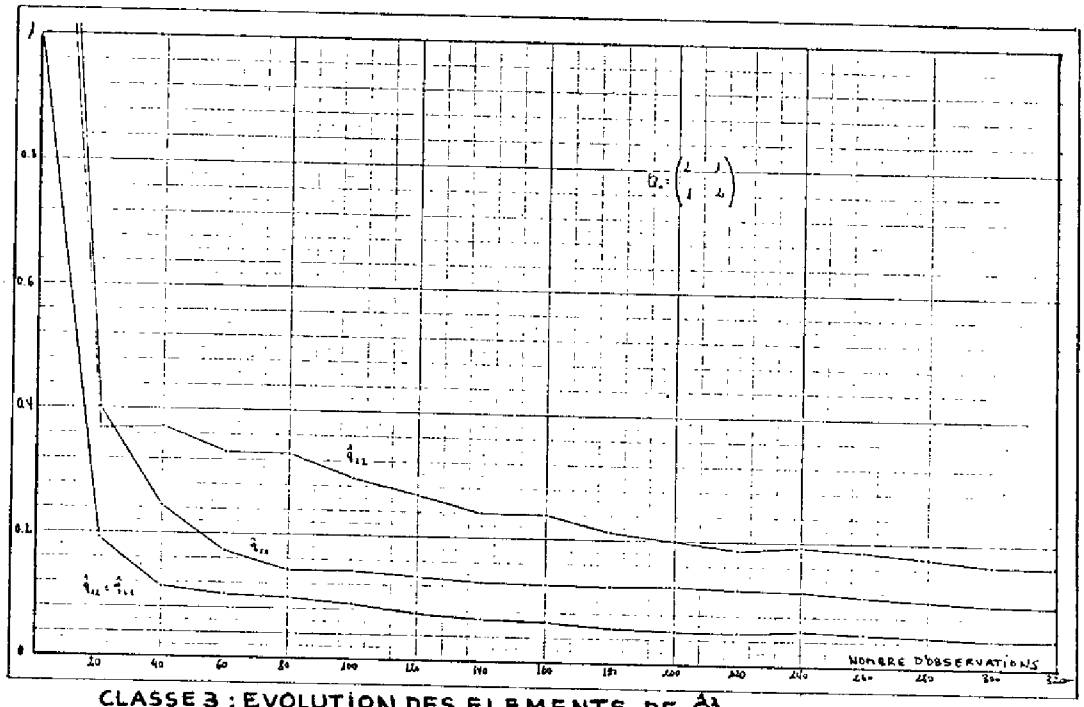


CLASSE 1 : EVOLUTION DES ELEMENTS DE $\hat{Q}^1_{K_1}$

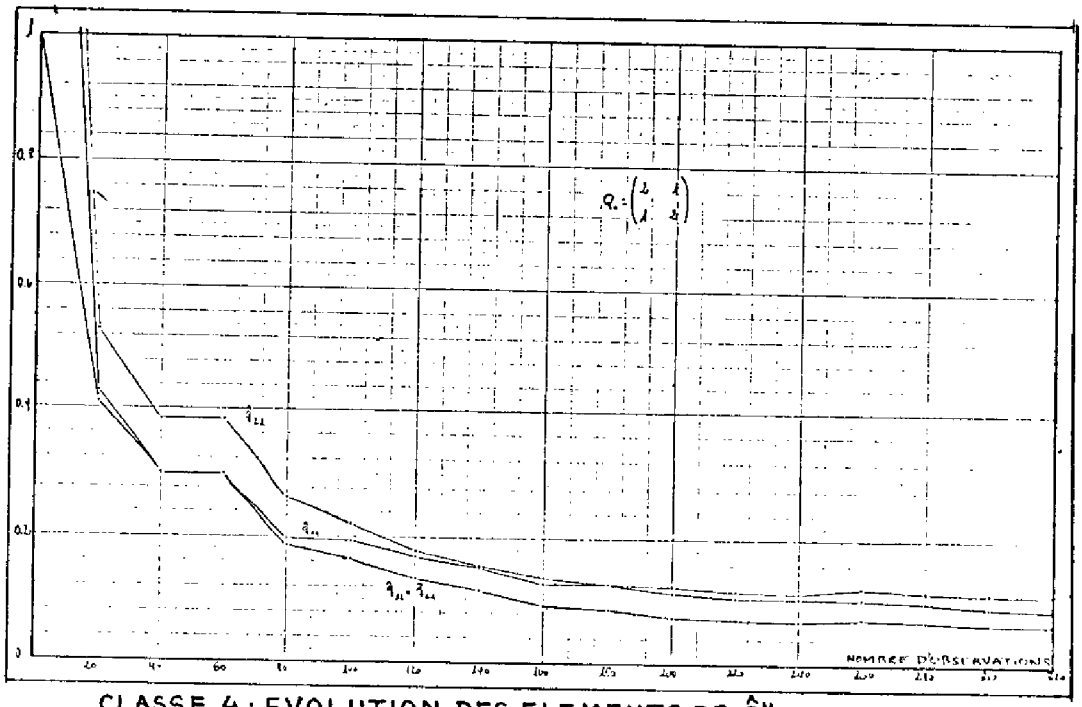


CLASSE 2 : EVOLUTION DES ELEMENTS DE $\hat{Q}^2_{K_2}$

Figure 4.1

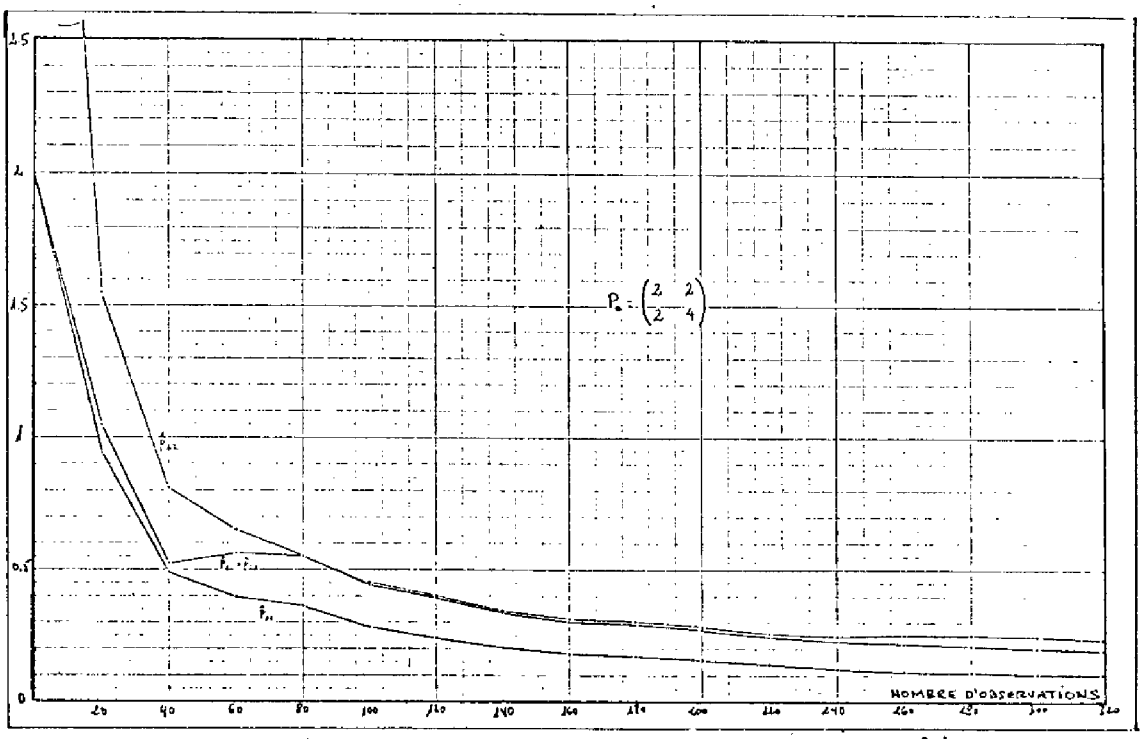


CLASSE 3 : EVOLUTION DES ELEMENTS DE $\hat{Q}^3_{K_3}$

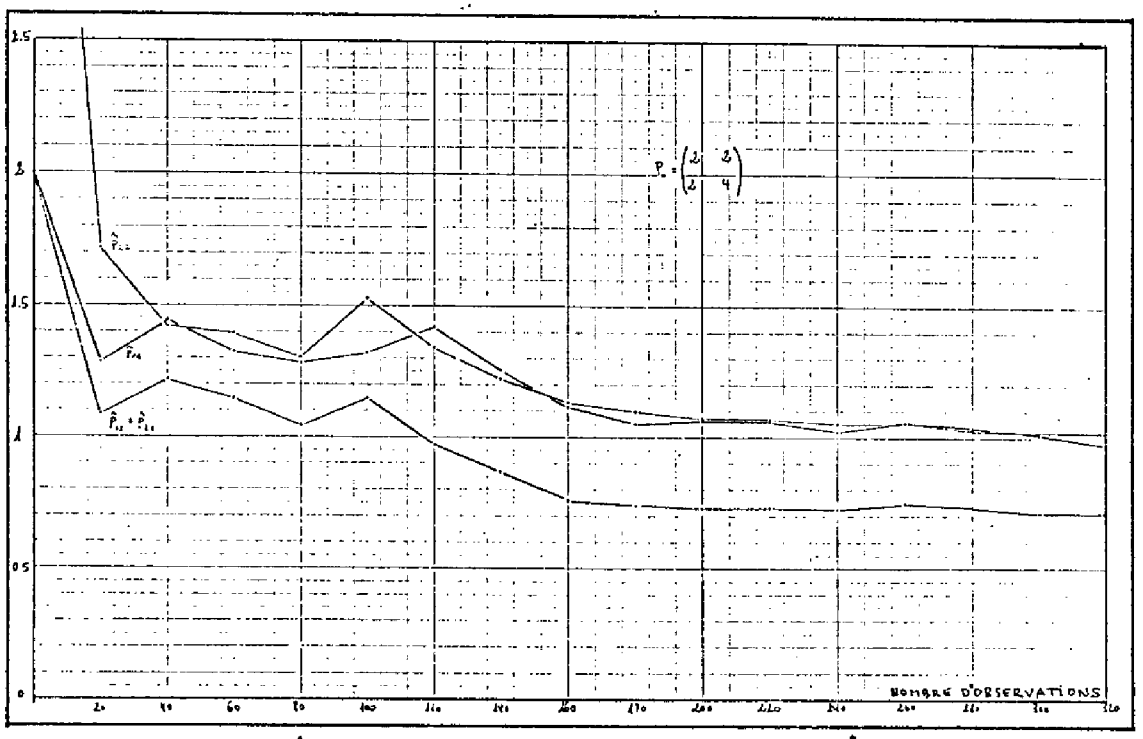


CLASSE 4 : EVOLUTION DES ELEMENTS DE $\hat{Q}^4_{K_4}$

Figure 4.1 (suite)

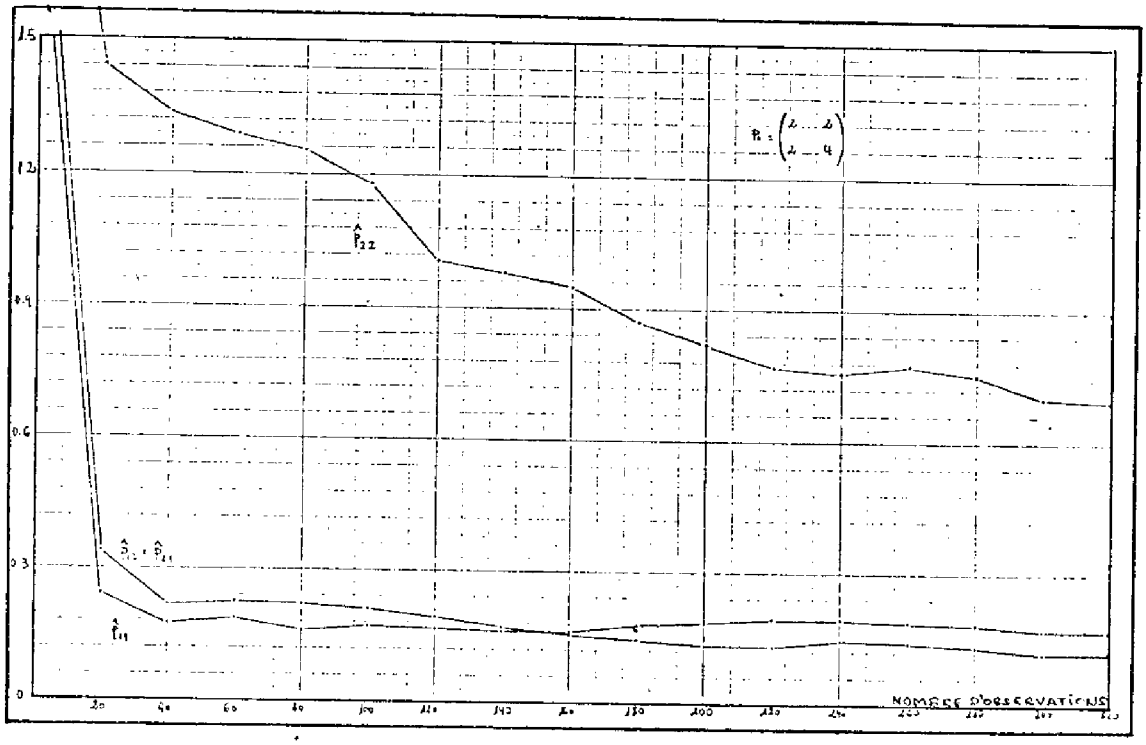


CLASSE 1 : CONVERGENCE DES ELEMENTS DE $\hat{P}^1_{K_1}$

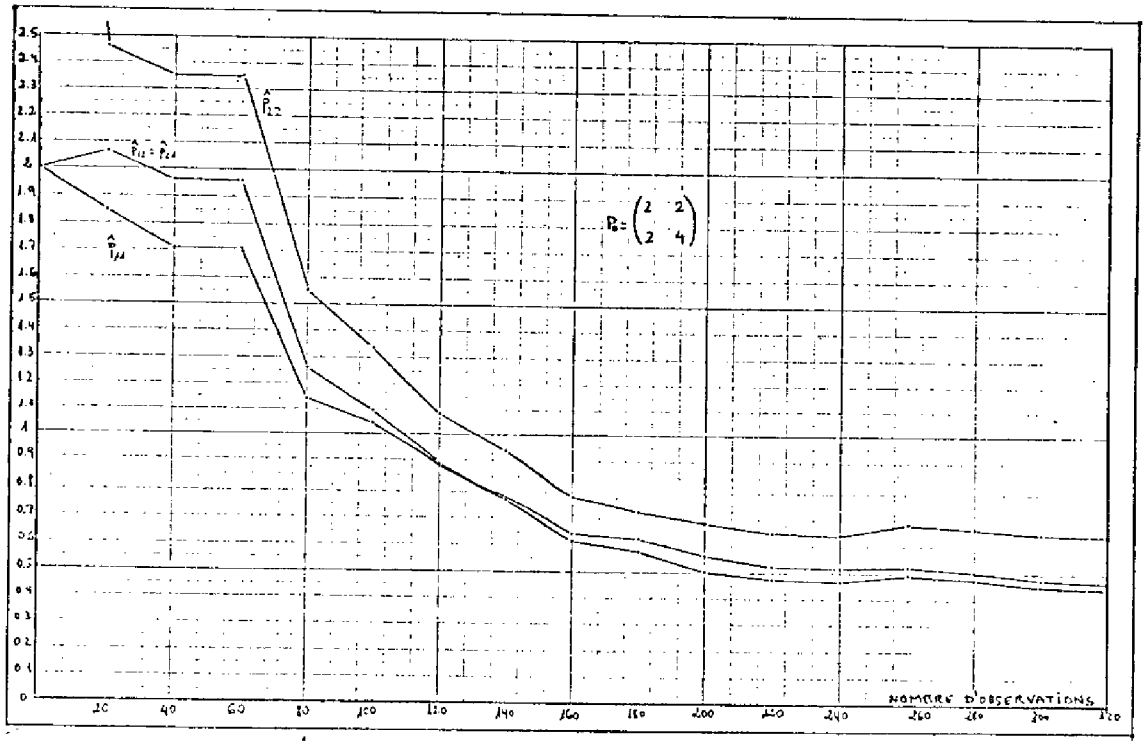


CLASSE 2 : CONVERGENCE DES ELEMENTS DE $\hat{P}^2_{K_2}$

Figure 4.m



CLASSE 3 : CONVERGENCE DES ELEMENTS DE $\hat{P}^3_{K_3}$



CLASSE 4 : CONVERGENCE DES ELEMENTS DE $\hat{P}^4_{K_4}$

Figure 4.m (suite)

c) Dans le but d'essayer de comparer notre méthode avec celle de A. SCHROEDER [4.31] basée sur l'algorithme des Nuées Dynamiques de E. DIDAY [4.10], nous allons traiter deux des exemples traités dans la thèse de A. SCHROEDER [4.31].

1er exemple :

150 points de $\mathbb{R}^2$ tirés de 3 distributions gaussiennes bidimensionnelles (voir fig. 4.n et fig. 4.o) de paramètres :

$$x^1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \qquad x^2 = \begin{bmatrix} 0 \\ 3 \end{bmatrix} \qquad x^3 = \begin{bmatrix} 4 \\ 3 \end{bmatrix}$$

$$S^1 = \begin{bmatrix} 4 & 1.7 \\ 1.7 & 1 \end{bmatrix} \qquad S^2 = \begin{bmatrix} 0.25 & 0 \\ 0 & 0.25 \end{bmatrix} \qquad S^3 = \begin{bmatrix} 4 & -1.7 \\ -1.7 & 1 \end{bmatrix}$$

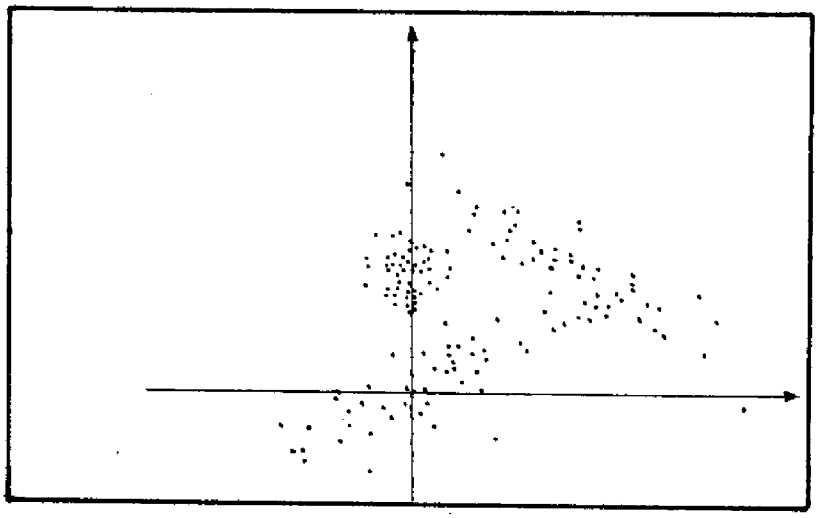


Fig. 4.n

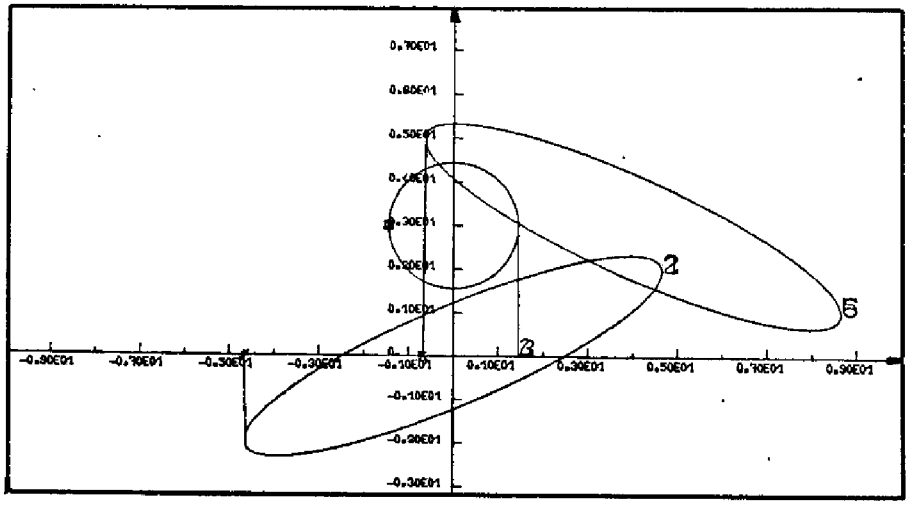


Fig. 4.o

Les résultats obtenus sont les suivants (voir fig. 4.p et Tableau T.4-3)

$$x^1 = \begin{bmatrix} -0.05 \\ 0.34 \end{bmatrix} \qquad x^2 = \begin{bmatrix} 0.07 \\ 3.39 \end{bmatrix} \qquad x^3 = \begin{bmatrix} 4.54 \\ 3.11 \end{bmatrix}$$

$$S^1 = \begin{bmatrix} 2.5 & 0.81 \\ 0.81 & 0.44 \end{bmatrix} \qquad S^2 = \begin{bmatrix} 0.61 & -0.10 \\ -0.10 & 0.23 \end{bmatrix} \qquad S^3 = \begin{bmatrix} 2.71 & -1.46 \\ -1.46 & 1.08 \end{bmatrix}$$

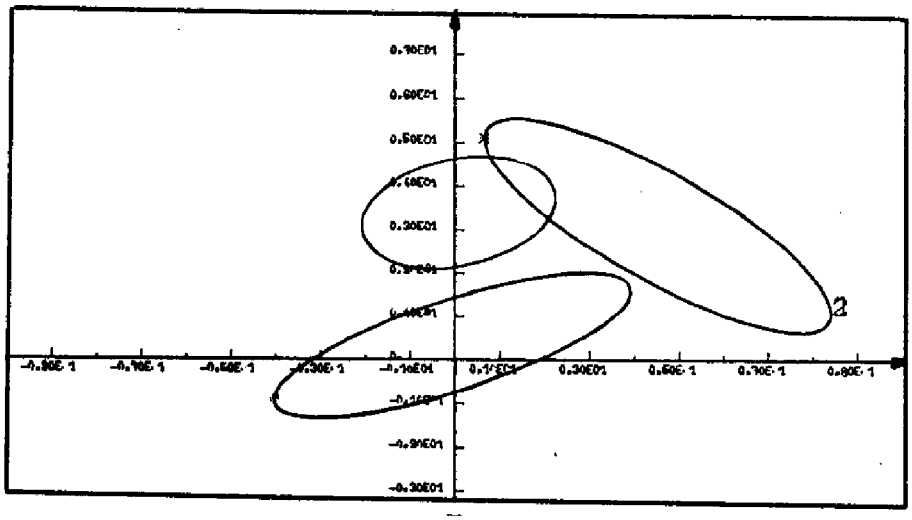


Fig. 4.p

Nous sommes dans un cas de discrimination parfaite puisque sur les 150 éléments classés il n'y a aucune erreur de classement (voir tableau T.4-3) et notamment les valeurs prises par l'indice de dissemblance dans : différence entre les classements, et la distance entre les classements "vrais" et obtenu).

```
ELEMENT  141 CLASSE  2 ANCIENNE CLASSE 1
   FORME= -4.37 -2.42
PY=  0.194E+00
ELEMENT  142 CLASSE  2 ANCIENNE CLASSE 1
   FORME= -3.44 -1.88
PY=  0.189E+00
ELEMENT  143 CLASSE  2 ANCIENNE CLASSE 1
   FORME= -3.56 -1.53
PY=  0.270E+00
ELEMENT  144 CLASSE  3 ANCIENNE CLASSE 2
   FORME= -0.57  2.81
PY=  0.229E+00
ELEMENT  145 CLASSE  2 ANCIENNE CLASSE 1
   FORME= -3.21 -1.61
PY=  0.233E+00
ELEMENT  146 CLASSE  2 ANCIENNE CLASSE 1
   FORME= -2.60 -1.31
PY=  0.236E+00
ELEMENT  147 CLASSE  2 ANCIENNE CLASSE 1
   FORME= -2.69 -1.21
PY=  0.195E+00
ELEMENT  148 CLASSE  2 ANCIENNE CLASSE 1
   FORME= -2.79 -0.79
PY=  0.173E+00
ELEMENT  149 CLASSE  3 ANCIENNE CLASSE 2
   FORME= -0.63  3.29
PY=  0.231E+00
ELEMENT  150 CLASSE  3 ANCIENNE CLASSE 2
   FORME= -0.49  3.36
•CLASSEMENT QUANTITATIF
DISTANCE GAUSSIENNE

ATTRIBUTS DE LA CLASSE     1
   VECTEUR DE MOYENNE ET MATRICE DE COVARIANCE
[4.54]      [2.7144    -1.4577]
[3.11]      [-1.4577    1.0791]
CLASSE     1
     MATRICES : Q̂  R̂  P̂

Q̂= [ 0.1590   -0.0497]
   [-0.0497    0.0388]
R̂= [ 2.2127   -1.2103]
   [-1.2103    0.8643]
P̂= [ 0.5017   -0.2474]
   [-0.2474    0.2148]

ATTRIBUTS DE LA CLASSE     2
   VECTEUR DE MOYENNE ET MATRICE DE COVARIANCE
[-0.05]     [2.5086    0.8191]
[ 0.34]     [0.8191    0.4441]
CLASSE     2
     MATRICES : Q̂  R̂  P̂

Q̂= [0.1710   0.0549]
   [0.0549   0.0578]
R̂= [2.0137   0.6532]
   [0.6532   0.3365]
P̂= [0.4449   0.1659]
   [0.1659   0.1076]

ATTRIBUTS DE LA CLASSE     3
   VECTEUR DE MOYENNE ET MATRICE DE COVARIANCE
[0.07]      [0.6198   -0.1053]
[3.39]      [-0.1053   0.2283]
CLASSE     3
     MATRICES : Q̂  R̂  P̂

Q̂= [ 0.0270   0.0019]
   [ 0.0019   0.0224]
R̂= [ 0.5142  -0.0954]
   [-0.0954   0.1778]
P̂= [ 0.1046  -0.0099]
   [-0.0099   0.0505]

LA CLASSE 1 EST FORMEE DE  46 ELEMENTS DONT
   0 DE L'ANCIENNE CLASSE 1
   0 DE L'ANCIENNE CLASSE 2
  46 DE L'ANCIENNE CLASSE 3
LA CLASSE 2 EST FORMEE DE  56 ELEMENTS DONT
  56 DE L'ANCIENNE CLASSE 1
   0 DE L'ANCIENNE CLASSE 2
   0 DE L'ANCIENNE CLASSE 3
LA CLASSE 3 EST FORMEE DE  48 ELEMENTS DONT
   0 DE L'ANCIENNE CLASSE 1
  48 DE L'ANCIENNE CLASSE 2
   0 DE L'ANCIENNE CLASSE 3
150      ELEMENTS CLASSES
ICARP=  3ICARC=  3
          MATRICE DE CONFUSION
[0.0        0.0        0.3167]
[0.3733     0.0        0.0   ]
[0.0        0.3200     0.0   ]
          DIFFERENCE ENTRE LES DEUX CLASSIFICATIONS=  0.0
          DISTANCE ENTRE LES DEUX CLASSIFICATIONS=  0.0
```

TABLEAU T.4-3

2ème exemple :

150 points de $\mathbb{R}^2$ tirés des 3 distributions gaussiennes à deux dimensions suivantes : (voir fig. 4.q et 4.r)

$$x^1 = \begin{bmatrix} 0 \\ 3 \end{bmatrix} \qquad x^2 = \begin{bmatrix} 3 \\ 0 \end{bmatrix} \qquad x^3 = \begin{bmatrix} 3 \\ 3 \end{bmatrix}$$

et

$$S^1 = S^2 = S^3 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

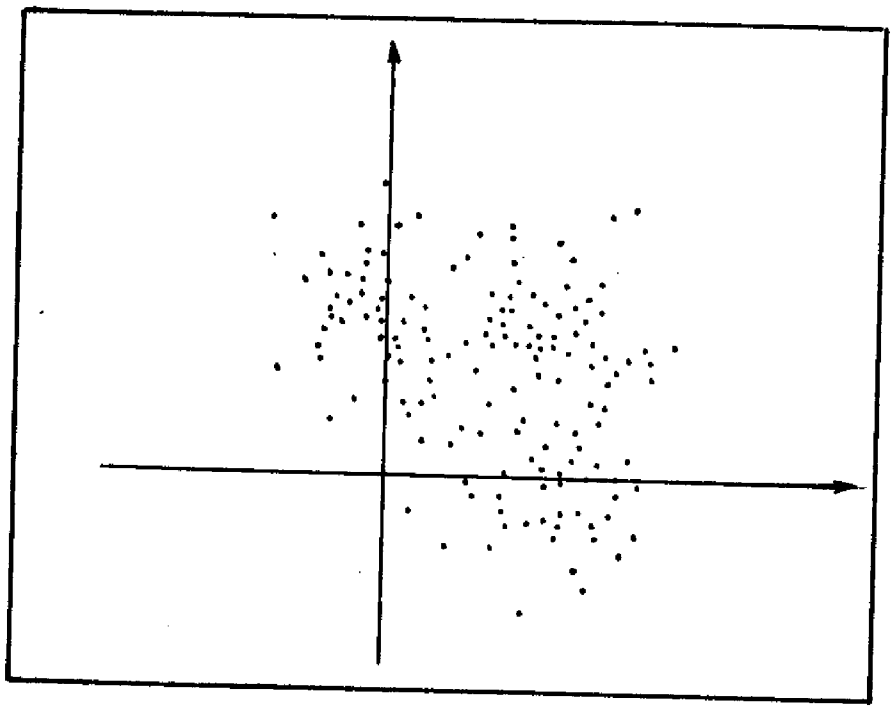


Fig. 4.q

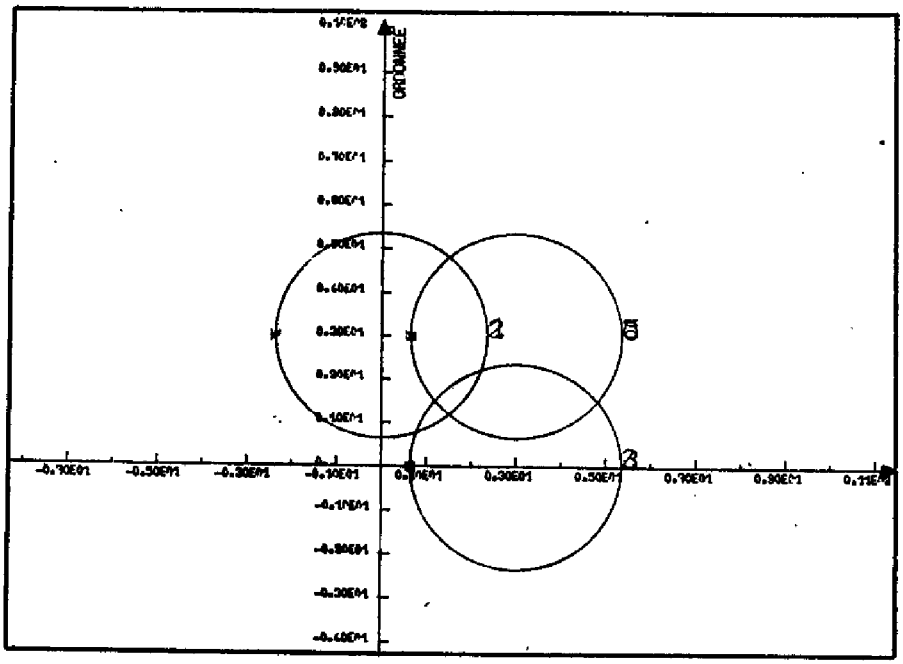


Fig. 4.r

Dans ce cas la discrimination a été assez bonne puisque sur 150 éléments classés on ne trouve que 6 éléments mal classés (voir tableau T.4-4 notamment les valeurs de la différence et la distance entre les classements).

Les résultats obtenus concernant l'estimation du noyau de chacune des classes sont les suivants : (voir tableau T.4-4 et fig. 4.s)

$$\hat{X}^1 = \begin{bmatrix} 0.24 \\ 3.29 \end{bmatrix} \qquad \hat{X}^2 = \begin{bmatrix} 3.42 \\ 0.38 \end{bmatrix} \qquad \hat{X}^3 = \begin{bmatrix} 3.16 \\ 3.68 \end{bmatrix}$$

$$\hat{S}^1 = \begin{bmatrix} 0.67 & -0.22 \\ -0.22 & 0.66 \end{bmatrix} \qquad \hat{S}^2 = \begin{bmatrix} 1.07 & -0.39 \\ -0.39 & 1.15 \end{bmatrix}$$

$$\hat{S}^3 = \begin{bmatrix} 0.64 & -0.03 \\ -0.03 & 1.23 \end{bmatrix}$$

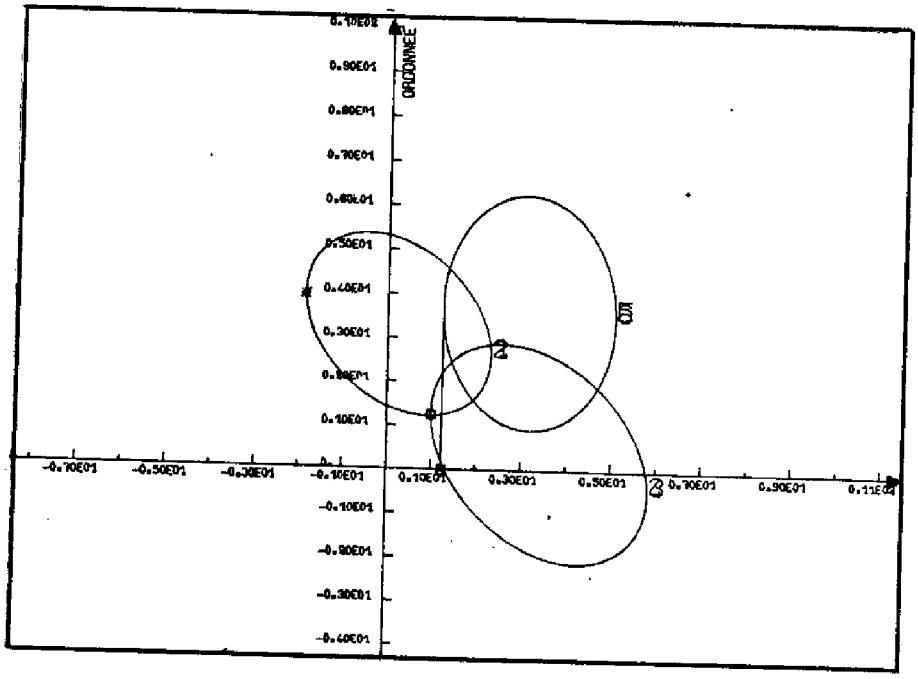


Fig. 4.s

ATTRIBUTS DE LA CLASSE 1

VECTEUR DE MOYENNE ET MATRICE DE COVARIANCE

$$\begin{bmatrix}3.16\\3.59\end{bmatrix} \qquad \begin{bmatrix}0.6400 & -0.0342\\-0.0342 & 1.2343\end{bmatrix}$$

CLASSE 1

MATRICES : $\hat{Q}$ $\hat{R}$ $\hat{P}$

$$\hat{Q}=\begin{bmatrix}0.0714 & -0.0026\\-0.0026 & 0.0836\end{bmatrix}$$

$$\hat{R}=\begin{bmatrix}0.4874 & -0.0291\\-0.0291 & 0.9842\end{bmatrix}$$

$$\hat{P}=\begin{bmatrix}0.1532 & -0.0051\\-0.0051 & 0.2501\end{bmatrix}$$

ATTRIBUTS DE LA CLASSE 2

VECTEUR DE MOYENNE ET MATRICE DE COVARIANCE

$$\begin{bmatrix}0.24\\3.29\end{bmatrix} \qquad \begin{bmatrix}0.6743 & -0.2268\\-0.2268 & 0.6654\end{bmatrix}$$

CLASSE 2

MATRICES : $\hat{Q}$ $\hat{R}$ $\hat{P}$

$$\hat{Q}=\begin{bmatrix}0.0646 & -0.0192\\-0.0192 & 0.0746\end{bmatrix}$$

$$\hat{R}=\begin{bmatrix}0.5265 & -0.1856\\-0.1856 & 0.5154\end{bmatrix}$$

$$\hat{P}=\begin{bmatrix}0.1478 & -0.0411\\-0.0411 & 0.1500\end{bmatrix}$$

ATTRIBUTS DE LA CLASSE 3

VECTEUR DE MOYENNE ET MATRICE DE COVARIANCE

$$\begin{bmatrix}3.42\\0.38\end{bmatrix} \qquad \begin{bmatrix}1.0771 & -0.3891\\-0.3891 & 1.1558\end{bmatrix}$$

CLASSE 3

MATRICES : $\hat{Q}$ $\hat{R}$ $\hat{P}$

$$\hat{Q}=\begin{bmatrix}0.0567 & 0.0031\\0.0031 & 0.0429\end{bmatrix}$$

$$\hat{R}=\begin{bmatrix}0.8837 & -0.3531\\-0.3531 & 0.9733\end{bmatrix}$$

$$\hat{P}=\begin{bmatrix}0.1934 & -0.0360\\-0.0360 & 0.1825\end{bmatrix}$$

LA CLASSE 1 EST FORMEE DE 52 ELEMENTS DONT
- 0 DE L'ANCIENNE CLASSE 1
- 6 DE L'ANCIENNE CLASSE 2
- 46 DE L'ANCIENNE CLASSE 3

LA CLASSE 2 EST FORMEE DE 56 ELEMENTS DONT
- 56 DE L'ANCIENNE CLASSE 1
- 0 DE L'ANCIENNE CLASSE 2
- 0 DE L'ANCIENNE CLASSE 3

LA CLASSE 3 EST FORMEE DE 42 ELEMENTS DONT
- 0 DE L'ANCIENNE CLASSE 1
- 42 DE L'ANCIENNE CLASSE 2
- 0 DE L'ANCIENNE CLASSE 3

150 ELEMENTS CLASSES

TABLEAU T.4-4

ICARF= 3 ICAMC= 3

MATRICE DE CONTINGENCE

$$\begin{bmatrix}0.0 & 0.0400 & 0.3067\\0.3733 & 0.0 & 0.0\\0.0 & 0.2600 & 0.0\end{bmatrix}$$

DIFFERENCE ENTRE LES DEUX CLASSIFICATIONS= 0.1020

DISTANCE ENTRE LES DEUX PARTITIONS= 0.3992

B) Application à la reconnaissance tactile de solides

Bien que notre algorithme travaille sous des hypothèses de normalité, nous avons voulu tester sa robustesse en traitant des données multidimensionnelles non gaussiennes.

Pour ce faire nous avons classé les données délivrées par la pince décrite dans le chapitre III, serrant les objets suivants :

- une sphère
- un cube
- un tétraèdre régulier

Nous avons fait 30 mesures pour chaque objet. Comme nous avons vu dans le chapitre III la pince est équipée de 12 optentiomètres, l'échantillon à classer est déterminé par le vecteur :

$$X^T = [\underbrace{x_1, x_2, x_3}_{\text{doigt 1}}, \underbrace{x_4, x_5, x_6}_{\text{doigt 2}}, \underbrace{x_7, x_8, x_9}_{\text{doigt 3}}, \underbrace{x_{10}, x_{11}, x_{12}}_{\text{doigt 4}}]$$

Nous avons considéré deux cas :

a) En tenant compte des mesures issues des potentiomètres des doigts 1 et 3

b) En tenant compte de tous les doigts.

Les résultats sont assez bons (voir tableau T.4-5 et T.4-6) ce qui nous fait penser que notre algorithme est assez robuste pour pouvoir traiter des données qui n'ont aucune raison particulière d'être gaussiennes.

Avec 4 doigts nous constatons que le nombre de classes créées est plus grand qu'avec 2 doigts, ce qui est normal puisque

le tétraèdre et le cube peuvent être positionnés de plusieurs manières dans la pince, donc on peut dire que les vraies classes correspondantes sont multimodales.

Comme nous pouvions l'espérer la classe des sphères est unique puisque la vraie classe correspondante est unimodale parce que sa forme est invariante par rotation.

Avec 2 doigts, le tétraèdre et le cube donnent aussi de vraies classes multimodales mais avec moins de modes que dans le cas de 4 doigts, (tableau T-4.6).

Pour le cas de 4 doigts nous avons 12 cubes absorbés par la classe des sphères, ainsi que 2 tétraèdres, par ailleurs 2 autres tétraèdres ont été classés avec les cubes ce qui fait 16 éléments mal classés sur un total de 90. (tableau T-4.6)

Dans le cas de 2 doigts nous remarquons qu'une dizaine de tétraèdres sont absorbés par la classe des sphères et trois autres tétraèdres ont été classés avec les cubes ce qui fait 13 éléments mal classés sur un total de 90 (tableau T-4.5)

Dans les deux cas nous avons calculé la différence (Dif) entre le classement obtenu et le classement réel ou "vrai" ainsi que la distance (Dist) qui résulte de la symétrisation de l'indice de dissemblance basé sur l'information conditionnelle (voir Chapitre II, paragraphe II.4.1).

```
PY=  0,920E-01
ELEMENT 25    CLASSE   1 ANCIENNE CLASSE 3
   FORME=  3,02   4,15   0,72   3,74   2,78   1,80
PY=  0,229E-01
ELEMENT 26    CLASSE   5 ANCIENNE CLASSE 3
   FORME=  1,94   3,61   2,48   3,95   1,99   3,60
PY=  0,256E-02
ELEMENT 27    CLASSE   1 ANCIENNE CLASSE 3
   FORME=  1,12   4,33   1,84   3,92   2,12   0,0
PY=  0,364E-02
ELEMENT 28    CLASSE   1 ANCIENNE CLASSE 3
   FORME=  2,89   4,25   1,81   1,43   5,00   0,09
PY=  0,626E-02
ELEMENT 29    CLASSE   1 ANCIENNE CLASSE 3
   FORME=  1,51   3,95   1,06   4,90   0,88   0,49
PY=  0,405E-02
ELEMENT 30    CLASSE   1 ANCIENNE CLASSE 3
   FORME=  0,43   5,00   0,0    5,00   1,52   0,39
*CLASSEMENT QUANTITATIF
DISTANCE GAUSSIENNE

ATTRIBUTS DE LA CLASSE     1
   VECTEUR DE MOYENNE ET MATRICE DE COVARIANCE
[2.03]   [ 0.7471  -0.2987   0.0277  -0.4029   0.2779   0.1678]
[4.16]   [-0.2987   0.3027  -0.0353   0.1662  -0.0454  -0.0699]
[0.86]   [ 0.0277  -0.0353   0.2349  -0.0457   0.1351  -0.0094]
[4.01]   [-0.4029   0.1662  -0.0457   0.5415  -0.4145   0.0077]
[2.73]   [ 0.2779  -0.0454   0.1351  -0.4145   0.8870  -0.0045]
[0.75]   [ 0.1678  -0.0699  -0.0094   0.0077  -0.0045   0.4186]
CLASSE     1
   MATRICES : Q̂  R̂  P̂

Q̂=
[ 0.0745   -0.0163    0.0001   -0.0210    0.0092    0.0044]
[-0.0163    0.0476   -0.0021    0.0081    0.0019   -0.0004]
[ 0.0001   -0.0021    0.0434   -0.0008    0.0055   -0.0007]
[-0.0210    0.0081   -0.0008    0.0629   -0.0209    0.0031]
[ 0.0092    0.0019    0.0055   -0.0209    0.0743   -0.0026]
[ 0.0044   -0.0004   -0.0007    0.0031   -0.0026    0.0456]

R̂=
[ 0.5942   -0.2445    0.0326   -0.3331    0.2450    0.1418]
[-0.2445    0.2275   -0.0316    0.1374   -0.0465   -0.0605]
[ 0.0326   -0.0316    0.1741   -0.0438    0.1164   -0.0062]
[-0.3331    0.1374   -0.0438    0.4213   -0.3458   -0.0004]
[ 0.2450   -0.0465    0.1164   -0.3458    0.7143    0.0014]
[ 0.1418   -0.0605   -0.0062   -0.0004    0.0014    0.3198]

P̂=
[ 0.1529   -0.0543   -0.0049   -0.0698    0.0329    0.0260]
[-0.0543    0.0752   -0.0037    0.0288    0.0011   -0.0095]
[-0.0049   -0.0037    0.0609   -0.0019    0.0186   -0.0032]
[-0.0698    0.0288   -0.0019    0.1202   -0.0687    0.0082]
[ 0.0329    0.0011    0.0186   -0.0687    0.1727   -0.0059]
[ 0.0260   -0.0095   -0.0032    0.0082   -0.0059    0.0988]

   LA CLASSE 1 EST FORMEE DE  40 ELEMENTS DONT
     30 DE L'ANCIENNE CLASSE 1
      0 DE L'ANCIENNE CLASSE 2
     10 DE L'ANCIENNE CLASSE 3
   LA CLASSE 2 EST FORMEE DE  17 ELEMENTS DONT
      0 DE L'ANCIENNE CLASSE 1
     17 DE L'ANCIENNE CLASSE 2
      0 DE L'ANCIENNE CLASSE 3
   LA CLASSE 3 EST FORMEE DE   9 ELEMENTS DONT
      0 DE L'ANCIENNE CLASSE 1
      7 DE L'ANCIENNE CLASSE 2
      2 DE L'ANCIENNE CLASSE 3
   LA CLASSE 4 EST FORMEE DE   7 ELEMENTS DONT
      0 DE L'ANCIENNE CLASSE 1
      6 DE L'ANCIENNE CLASSE 2
      1 DE L'ANCIENNE CLASSE 3
   LA CLASSE 5 EST FORMEE DE   6 ELEMENTS DONT
      0 DE L'ANCIENNE CLASSE 1
      0 DE L'ANCIENNE CLASSE 2
      6 DE L'ANCIENNE CLASSE 3
   LA CLASSE 6 EST FORMEE DE   3 ELEMENTS DONT
      0 DE L'ANCIENNE CLASSE 1
      0 DE L'ANCIENNE CLASSE 2
      3 DE L'ANCIENNE CLASSE 3
   LA CLASSE 7 EST FORMEE DE   1 ELEMENTS DONT
      0 DE L'ANCIENNE CLASSE 1
      0 DE L'ANCIENNE CLASSE 2
      1 DE L'ANCIENNE CLASSE 3
   LA CLASSE 8 EST FORMEE DE   6 ELEMENTS DONT
      0 DE L'ANCIENNE CLASSE 1
      0 DE L'ANCIENNE CLASSE 2
      6 DE L'ANCIENNE CLASSE 3
   LA CLASSE 9 EST FORMEE DE   1 ELEMENTS DONT
      0 DE L'ANCIENNE CLASSE 1
      0 DE L'ANCIENNE CLASSE 2
      1 DE L'ANCIENNE CLASSE 3
    90 ELEMENTS CLASSES
   ICARE=  3ICAR1=  9
          MATRICE DE CONTINGENCE
[0.3333   0.0      0.1111]
[0.0      0.1059   0.0   ]
[0.0      0.0778   0.0222]
[0.0      0.0667   0.0111]
[0.0      0.0      0.0667]
[0.0      0.0      0.0333]
[0.0      0.0      0.0111]
[0.0      0.0      0.0222]
[0.0      0.0      0.0111]
          DIFFERENCE ENTRE LES DEUX CLASSIFICATIONS=   0.1663
          DISTANCE ENTRE LES DEUX PARTITIONS=  0.2328
```

TABLEAU T-4.5

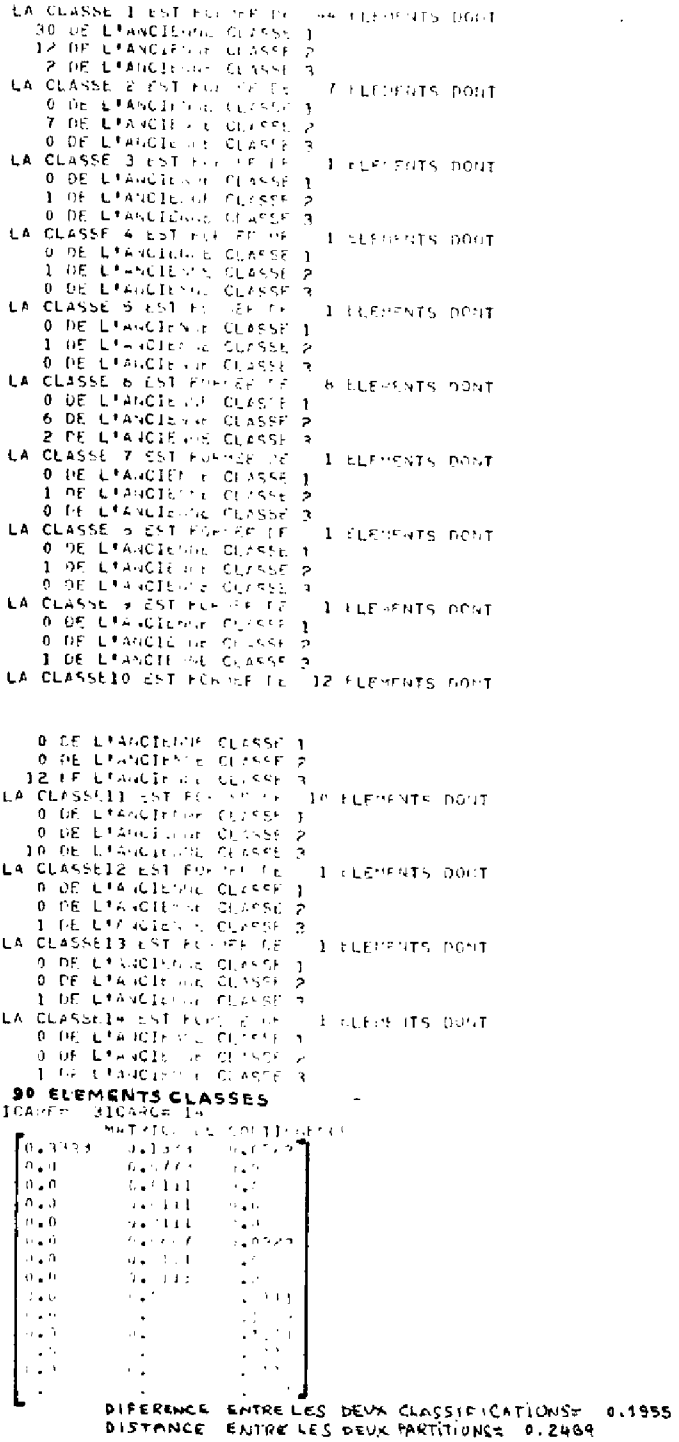

LA CLASSE 1 EST FORMEE DE 44 ELEMENTS DONT
30 DE L'ANCIENNE CLASSE 1
12 DE L'ANCIENNE CLASSE 2
2 DE L'ANCIENNE CLASSE 3
LA CLASSE 2 EST FORMEE DE 7 ELEMENTS DONT
0 DE L'ANCIENNE CLASSE 1
7 DE L'ANCIENNE CLASSE 2
0 DE L'ANCIENNE CLASSE 3
LA CLASSE 3 EST FORMEE DE 1 ELEMENTS DONT
0 DE L'ANCIENNE CLASSE 1
1 DE L'ANCIENNE CLASSE 2
0 DE L'ANCIENNE CLASSE 3
LA CLASSE 4 EST FORMEE DE 1 ELEMENTS DONT
0 DE L'ANCIENNE CLASSE 1
1 DE L'ANCIENNE CLASSE 2
0 DE L'ANCIENNE CLASSE 3
LA CLASSE 5 EST FORMEE DE 1 ELEMENTS DONT
0 DE L'ANCIENNE CLASSE 1
1 DE L'ANCIENNE CLASSE 2
0 DE L'ANCIENNE CLASSE 3
LA CLASSE 6 EST FORMEE DE 8 ELEMENTS DONT
0 DE L'ANCIENNE CLASSE 1
6 DE L'ANCIENNE CLASSE 2
2 DE L'ANCIENNE CLASSE 3
LA CLASSE 7 EST FORMEE DE 1 ELEMENTS DONT
0 DE L'ANCIENNE CLASSE 1
1 DE L'ANCIENNE CLASSE 2
0 DE L'ANCIENNE CLASSE 3
LA CLASSE 8 EST FORMEE DE 1 ELEMENTS DONT
0 DE L'ANCIENNE CLASSE 1
1 DE L'ANCIENNE CLASSE 2
0 DE L'ANCIENNE CLASSE 3
LA CLASSE 9 EST FORMEE DE 1 ELEMENTS DONT
0 DE L'ANCIENNE CLASSE 1
0 DE L'ANCIENNE CLASSE 2
1 DE L'ANCIENNE CLASSE 3
LA CLASSE10 EST FORMEE DE 12 ELEMENTS DONT
0 DE L'ANCIENNE CLASSE 1
0 DE L'ANCIENNE CLASSE 2
12 DE L'ANCIENNE CLASSE 3
LA CLASSE11 EST FORMEE DE 10 ELEMENTS DONT
0 DE L'ANCIENNE CLASSE 1
0 DE L'ANCIENNE CLASSE 2
10 DE L'ANCIENNE CLASSE 3
LA CLASSE12 EST FORMEE DE 1 ELEMENTS DONT
0 DE L'ANCIENNE CLASSE 1
0 DE L'ANCIENNE CLASSE 2
1 DE L'ANCIENNE CLASSE 3
LA CLASSE13 EST FORMEE DE 1 ELEMENTS DONT
0 DE L'ANCIENNE CLASSE 1
0 DE L'ANCIENNE CLASSE 2
1 DE L'ANCIENNE CLASSE 3
LA CLASSE14 EST FORMEE DE 3 ELEMENTS DONT
0 DE L'ANCIENNE CLASSE 1
0 DE L'ANCIENNE CLASSE 2
1 DE L'ANCIENNE CLASSE 3
90 ELEMENTS CLASSES
DIFFERENCE ENTRE LES DEUX CLASSIFICATIONS= 0.1955
DISTANCE ENTRE LES DEUX PARTITIONS= 0.2489

TABLEAU T-4.6

IV-4.10) Conclusion

D'après les résultats obtenus nous pouvons conclure que l'approche par autoapprentissage donne des résultats assez bons dans la résolution du problème de la reconnaissance des composants d'un mélange ; nous voulons remarquer que le principal avantage de notre approche est l'aspect globalement itératif du traitement des données et que, par conséquent, elle nous fournit à tout instant, une partition des données déjà traitées.

Nous constatons que les résultats les meilleurs sont ceux qui concernent le classement proprement dit ; l'estimation des noyaux (vecteur de moyenne et matrice de covariance) des classes est moins bonne qu'avec la méthode des Nuées Dynamiques, cela pouvait être prévu puisque la méthode des Nuées Dynamiques connait a priori l'ensemble des données à analyser, c'est-à-dire elle n'est pas séquentielle bien qu'il existe une variante appelée MNDS (Méthode des Nuées Dynamiques Séquentialisée) qui traite les données par paquets elle n'est donc pas globalement itérative, bien que si on a la possibilité de réduire les paquets à un individu elle devient globalement itérative ; de plus les méthodes des Nuées Dynamiques supposent connu le nombre de classes (composants du mélange).

Ces méthodes nécessitent une plus importante quantité d'information initiale que notre approche où l'information initiale est pratiquemment absente.

Un inconvénient de notre approche c'est qu'elle est restreinte au cas gaussien bien que l'algorithme ait démontré être

assez robuste pour pouvoir traiter, sans trop d'erreurs, des données qui n'ont aucune raison particulière d'être gaussiennes, comme c'était le cas dans l'application à la reconnaissance tactile de solides, un autre inconvénient c'est le choix du paramètre subjectif α qui définit le seuil.

En résumé nous dirons que dans le cas où on connait l'ensemble de données à traiter, a priori, ainsi que le nombre de classes, l'approche suivant la méthode des Nuées Dynamiques donne des meilleurs résultats que notre approche, mais dans le cas où l'on n'a pas d'information initiale et où l'acquisition des données est faite de façon séquentielle alors l'approche par autoapprentissage est peut être plus adaptée.

REFERENCES

/-4.1_7 AGRAWALA A.K. (1970)
"Learning with a probabilistic teacher"
IEEE Trans. on Information Theory - Vol. IT-16, n° 4, Mai.

/-4.2_7 AGUILAR MARTIN J. - BANON G. - LOPEZ DE MANTARAS R. (1976)
"A self learning algorithm for the classification and recognition of vectorial patterns"
IEEE International Symposium on Information Theory
Ronneby, Sweden, June.

/-4.3_7 ALSPACH D.L. (1974)
"A parallel filtering algorithm for linear systems with unknown time varying statistics"
IEEE Trans. on Automatic Control, Vol. AC-19.

/-4.4_7 BENZECRI J.P. (1972)
"La régression"
Laboratoire de Statistique Mathématique, Université de Paris VI.

/-4.5_7 BHATTACHARYA C.G. (1967)
"A simple method of resolution of a distribution into gaussian components"
Biometrics, March.

/-4.6_7 BUCY R.S. et JOSEPH P.D. (1968)
"Filtering for stochastic processes with applications to guidance"
John Wiley.

/-4.7_7 COOPER D.B. et COOPER P.W. (1964)
"Non supervised adaptative signal detection pattern recognition"
Information and Control, n° 7, pp. 416 à 444.

/-4.8_7 COVER T.M. et HART P.E. (1967)
"Nearest neighbor pattern classification"
IEEE Trans. on Information Theory, Vol. IT-13, n°1, January.

/-4.9_7 DAY N.E. (1969)
"Estimating the components of a mixture of normal distributions"
Biometrika, Vol. 56, n° 3, pp. 463-467.

/-4.10_7 DIDAY E. (1972)
"Nouvelles méthodes et nouveaux concepts en classification automatique et reconnaissance des formes"
Thèse d'Etat - Université Paris VI.

/-4.11_7 DIDAY E. et SCHROEDER A. (1974)
"A new approach in mixed distributions detection"
Rapport de recherche n° 52 - IRIA - LABORIA.
aussi dans : R.A.I.R.O. Recherche Opérationnelle, Vol. 10, n° 6, pp. 75-106, Juin.

/-4.12_7 DIDAY E., SCHROEDER A. et OK Y. (1974)
"The dynamic clusters method in pattern recognition"
Proceedings of IFIP Congress - Stockholm.

/-4.13_7 DUDA R.O. et HART R.E. (1973)
"Pattern classification and scene analysis"
Wiley N.Y.

/-4.14_7 FROMM F.R. et NORTHOUSE R.A. (1976)
"CLASS : A Nonparametric clustering algorithm"
Pattern recognition, Vol. 8, pp. 107-114, July.

/-4.15_7 FUKUNAGA K. (1972)
"Introduction to statiscal pattern recognition"
Academic Press.

/-4.16_7 FUKUNAGA K. et KOONTZ W.L.G. (1970)
"A criterion and an algorithm for grouping Data"
IEEE Trans on Computers, Vol. C-19, NO-10, October.

/-4.17_7 GOVAERT G. (1975)
"Classification automatique et distances adaptatives"
Thèse de 3ème cycle - Université Paris VI

/-4.18_7 HILBORN C.G. et LAINIOTIS D.G. (1969)
"Optimal estimation in the presence of unknown parameters"
IEEE Trans. Systems, Sciences, Cybernetics, Vol. SSC-1.

/-4.19_7 IMAI T. et SHIMURA M. (1976)
"Learning with probabilistic labeling"
Pattern Recognition, Vol. 8, pp. 5-10, January.

/-4.20_7 KITTLER J. (1976)
"A locally sensitive method for cluster Analysis"
Pattern Recognition, Vol. 8, pp. 23-33, January.

/-4.21_7 LECHEVALLIER Y. (1974)
"Optimisation de quelques critères en classification automatique et application à l'étude des modifications des protéines sériques en pathologie clinique"
Thèse 3ème cycle - Université Paris VI.

/-4.22_/ MICHEL M. et WEN-CHUN LIN (1973)
"Experimental study on information measure and Inter-Intra class distance ratios on feature selection and Orderings"
IEEE Trans. on Systems Man and Cybernetics, Vol. SMC-3, n° 2, March.

/-4.23_/ MYERS K.A. et TAPLEY B.D. (1976)
"Adaptative sequential estimation with unknown noise statistics"
IEEE Trans. Automatic control, August.

/-4.24_/ PATRICK E.A. (1972)
"Fundamentals of pattern recognition"
Prentice hall, Inc. - N.J.

/-4.25_/ PATRICK E.A. et COSTELLO J.P. (1970)
"On unsupervised estimation algorithms"
IEEE Transactions on Information theory, Vol. IT-16, n° 5, pp. 556-569, September.

/-4.26_/ PATRICK E.A. et HANCOCK J.C. (1966)
"Non supervised sequential classification and recognition of patterns"
IEEE Trans. on Inf. Theory, Vol. IT-12, n° 3, pp. 362-372, July.

/-4.27_/ RAO C.R. (1948)
"Utilisation of multiple measurement in problems of biological classification"
J.R. Statist. Soc. B. - Vol. X, n° 2, pp. 159-193.

/-4.28_/ RHODES I.B. (1976)
"A tutorial introduction to estimation and filtering"
IEEE Trans. on Aut. Control, Vol. AC-16.

/-4.29_/ RUSPINI E.H. (1969)
"A new approach to clustering"
Information and Control, Vol. 15, pp. 22-32.

/-4.30_/ SALUT G. (1976)
"Identification optimale des systèmes linéaires stochastiques"
Thèse de doctorat d'Etat, Université Paul Sabatier, Toulouse.

/-4.31_/ SCHROEDER A. (1974)
"Reconnaissance des composants d'un mélange"
Thèse de 3ème cycle - Université Paris VI.

/4.33/ SHAIDUL A.B. (1974)
"Compound sequential probability ratio test for the classification of statistically dependent patterns"
IEEE Trans. on computers, Vol. C-23, n° 4, April.

/4.34/ SPRAGINS J.S. (1966)
"Learning without a Teacher"
IEEE Transactions on Information Theory, Vol. IT-12, n° 2, April.

/4.35/ WHEELER S.G. et INGRAM D.S. (1976)
"Approximations for the probability of misclassification"
Pattern Recognition, Vol 8, pp. 119-126, July.

/4.36/ WOLVERTON C.T. et WAGNER T.J. (1969)
"Asymptotically optimal discriminant functions for pattern classification"
IEEE Trans on Inf. Theory, Vol. IT-15, n°2, March.

/4.37/ YOUNG T.Y. et CORALUPPI G. (1970)
"Stochastic estimation of a mixture of normal density functions using an information criterion"
IEEE Trans. on Inf. Theory, Vol. IT-16, n° 3, pp. 258-263, May.

CONCLUSION

V. CONCLUSION

Tout au long de ce travail nous avons surtout traité le problème du classement de données multidimensionnelles sous l'optique de l'estimation récursive d'une partition d'un ensemble de données à classer dont les éléments sont observés séquentiellement.

En fait ce travail laisse en suspens un certain nombre de questions ouvertes qui pourraient être approfondies aussi bien sur le plan mathématique que sur le plan des applications.

Dans la première partie on a essayé de formaliser mathématiquement le problème de l'estimation d'une partition par autoapprentissage, à ce niveau là apparaît la première question ouverte : trouver les conditions générales sous lesquelles on peut démontrer la convergence des partitions estimées vers la partition "vraie". La théorie de l'ergodicité semble fournir des résultats dans ce domaine mais il reste à savoir quelles hypothèses doit satisfaire la séquence des partitions estimées.

Dans la deuxième partie nous proposons plusieurs métriques entre partitions, basées sur la théorie de l'information qui sont particulièrement utiles quand il s'agit de comparer un classement quelconque d'un ensemble de données avec le classement "vrai" de ce même ensemble.

Dans la troisième partie nous avons étudié un algorithme

de classement par autoapprentissage de données qualitatives basé sur le critère du maximum de vraisemblance comme règle de décision, nous proposons, comme suite à cette étude, de remplacer la fonction d'appartenance probabiliste par une fonction d'appartenance suivant la théorie des sous-ensembles flous.

Nous avons illustré notre algorithme avec une application à la reconnaissance tactile de solides en utilisant pour cela deux systèmes de préhension : une pince de 4 doigts et une main anthropomorphe.

Nous voulons ici remarquer que la disposition et le nombre de capteurs de pression sur la main ne sont nécessairement pas les meilleurs et qu'une recherche dans ce domaine reste à faire, ainsi que sur la possibilité de disposer de capteurs angulaires sur la main puisque l'utilisation d'informations mixtes (angulaires et de pression) serait probablement la combinaison optimale.

Dans la quatrième partie nous avons présenté un algorithme de classement de données gaussiennes par autoapprentissage basé sur la théorie du filtrage linéaire adaptatif, dans le cas discret, qui est l'élément fondamental de la récursivité de cet algorithme, puisqu'on l'utilise pour estimer récursivement les densités de probabilité gaussiennes qui caractérisent chaque classe, à ce niveau là on peut se poser quelques questions Peut-on formaliser le choix du niveau α ? existe-t-il une relation explicite entre le niveau α et les matrices de covariance ?

Dans cette dernière partie nous avons traité un exemple de classement de données multidimensionnelles. Nous avons appliqué notre algorithme au problème de la reconnaissance des composants d'un mélange de densités de probabilité gaussiennes bidimensionnelles et nous l'avons comparé avec la méthode des Nuées Dynamiques de E. DIDAY ; nous avons constaté que notre approche est particulièrement intéressante dans le cas où l'information initiale est trés faible et où un traitement itératif des données doit être fait.

-:-

ANNEXE 1

RAPPEL SUR LES VARIABLE ALEATOIRES GAUSSIENNES MULTIDIMENSIONNELLES

-:-

ANNEXE 1

Rappel sur les variables aléatoires gaussiennes multidimensionnelles

Soient deux vecteurs aléatoires x et y gaussiens et tels que le vecteur $\begin{bmatrix} x \\ \hdashline y \end{bmatrix}$ soit gaussien, ils possèdent alors les propriétés suivantes :

Propriété I

L'espérance mathématique conditionnelle de x connaissant y est donnée par :

$$\mathbb{E}[x|y] = \mathbb{E}[x] + P_{xy} P_{yy}^{-1} [y - \mathbb{E}[y]]$$

avec $$P_{xy} = \mathbb{E}\left[(x - \mathbb{E}(x))(y - \mathbb{E}(y))^T\right]$$

et $$P_{yy} = \mathbb{E}\left[(y - \mathbb{E}(y))(y - \mathbb{E}(y))^T\right]$$

De plus $x - \mathbb{E}[x|y]$ est indépendant de y et a une moyenne nulle.

Propriété II

Soit un vecteur gaussien : $\begin{bmatrix} x \\ y_1 \\ y_2 \end{bmatrix}$

où y_1 et y_2 sont des vecteurs indépendants, l'espérance mathématique conditionnelle de x connaissant y_1 et y_2 a l'expression suivante :

$$\mathbb{E}[x|y_1, y_2] = \mathbb{E}[x|y_1] + \mathbb{E}[x|y_2] - \mathbb{E}[x]$$

Démonstration

On forme le vecteur $y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$

D'après l'hypothèse d'indépendance de y_1 et y_2 on a :

$$P_{yy} = \left[\begin{array}{c|c} P_{y_1 y_1} & \mathbb{0} \\ \hline \mathbb{0} & P_{y_2 y_2} \end{array}\right]$$

D'autre part, la matrice P_{xy} peut être partitionée :

$$P_{xy} = \left[\begin{array}{c|c} P_{xy_1} & P_{xy_2} \end{array}\right]$$

D'après la propriété I on aura :

$$\mathbb{E}\left[x \mid y_1, y_2\right] = \mathbb{E}[x] + \left[\begin{array}{c|c} P_{xy_1} & P_{xy_2} \end{array}\right] \begin{bmatrix} P_{y_1 y_1}^{-1} & \mathbb{0} \\ \mathbb{0} & P_{y_2 y_2}^{-1} \end{bmatrix} \begin{bmatrix} y_1 - \mathbb{E}[y_1] \\ y_2 - \mathbb{E}[y_2] \end{bmatrix}$$

d'où en développant :

$$\mathbb{E}\left[x \mid y_1, y_2\right] = \underbrace{\mathbb{E}[x] + P_{xy_1} P_{y_1 y_1}^{-1}\left[y_1 - \mathbb{E}[y_1]\right]}_{\mathbb{E}[x \mid y_1]} + \underbrace{P_{xy_2} P_{y_2 y_2}^{-1}\left[y_2 - \mathbb{E}[y_2]\right]}_{\mathbb{E}[x \mid y_2] - \mathbb{E}[x]}$$

Propriété III

La densité de probabilité conditionnelle du vecteur x connaissant le vecteur y est gaussienne de moyenne :

$$\mathbb{E}\left[x \mid y\right]$$

et de matrice de covariance : $P_{xx} - P_{xy} P_{yy}^{-1} P_{yx}$

avec
$$P_{xx} = \mathbb{E}\left[(x - \mathbb{E}(x))(x - \mathbb{E}(x))^T\right]$$
$$P_{xy} = \mathbb{E}\left[(x - \mathbb{E}(x))(y - \mathbb{E}(y))^T\right]$$
$$P_{yy} = \mathbb{E}\left[(y - \mathbb{E}(y))(y - \mathbb{E}(y))^T\right]$$

et
$$P_{yx} = \mathbb{E}\left[(y - \mathbb{E}(y))(x - \mathbb{E}(x))^T\right]$$

TABLE DES MATIERES