

50376
1988
251



503
198
25

THESE

présentée à

**L'UNIVERSITE DES SCIENCES ET TECHNIQUES DE LILLE
FLANDRES ARTOIS**

pour l'obtention du titre de

DOCTEUR DE L'UNIVERSITE

Spécialité : Automatique et Informatique industrielle

par

Serge OLEJNIK

**ANALYSE DE LA CONVEXITE DES FONCTIONS DE DENSITE
PAR DES TECHNIQUES D'ETIQUETAGE PROBABILISTE ITERATIF
APPLICATION EN CLASSIFICATION AUTOMATIQUE**

THESE

présentée à

L'UNIVERSITE DES SCIENCES ET TECHNIQUES DE LILLE
FLANDRES ARTOIS

pour l'obtention du titre de

DOCTEUR DE L'UNIVERSITE

Spécialité : Automatique et Informatique industrielle

par

Serge OLEJNIK

ANALYSE DE LA CONVEXITE DES FONCTIONS DE DENSITE
PAR DES TECHNIQUES D'ETIQUETAGE PROBABILISTE ITERATIF
APPLICATION EN CLASSIFICATION AUTOMATIQUE

THESE

présentée à

L'UNIVERSITE DES SCIENCES ET TECHNIQUES DE LILLE
FLANDRES ARTOIS

pour l'obtention du titre de

DOCTEUR DE L'UNIVERSITE

Spécialité : Automatique et Informatique industrielle

par

Serge OLEJNIK

ANALYSE DE LA CONVEXITE DES FONCTIONS DE DENSITE
PAR DES TECHNIQUES D'ETIQUETAGE PROBABILISTE ITERATIF
APPLICATION EN CLASSIFICATION AUTOMATIQUE

TABLE DES MATIERES.

I. CLASSIFICATION AUTOMATIQUE ET CONVEXITE DES FONCTIONS DE DENSITE.

<u>I.1 INTRODUCTION.</u>	PAGE 1
<u>I.2 APPROCHE METRIQUE.</u>	PAGE 2
<u>I.3 APPROCHE STATISTIQUE.</u>	PAGE 2
<u>I.4 CONVEXITE ET DETECTION DES MODES.</u>	PAGE 3
<u>I.5 CONVEXITE ET RELAXATION.</u>	PAGE 4
<u>I.6 RELAXATION ET DETECTION DES MODES.</u>	PAGE 5
<u>REFERENCES (chapitre I)</u>	PAGE 6

II. ANALYSE DE LA CONVEXITE DES FONCTIONS DE DENSITE DE PROBABILITE.

<u>II.1 INTRODUCTION.</u>	PAGE 11
<u>II.2 CONVEXITE D'UNE FONCTION DE DENSITE DE PROBABILITE MULTIVARIABLE.</u>	PAGE 11
<u>II.2.1 PRINCIPE D'ANALYSE.</u>	PAGE 11
<u>II.2.2 ESTIMATION DE LA VALEUR MOYENNE D'UNE FONCTION.</u>	PAGE 13
<u>II.3 TRANSFORMATION LINEAIRE DE L'ESPACE D'OBSERVATIONS.</u>	PAGE 19
<u>II.4 TEST RAPIDE DE CONVEXITE.</u>	PAGE 23
<u>II.5 PRE-TRAITEMENT DES DONNEES.</u>	PAGE 26
<u>II.6 CONCLUSION.</u>	PAGE 29
<u>REFERENCES (chapitre II)</u>	PAGE 32

III. CONVEXITE ET RELAXATION.

<u>III.1 INTRODUCTION.</u>	PAGE 35
<u>III.2 L'ETIQUETAGE PROBABILISTE OU RELAXATION.</u>	PAGE 35
<u>III.3 APPLICATION DES METHODES DE RELAXATION A L'ANALYSE DE LA CONVEXITE.</u>	PAGE 37
<u>III.4 DETERMINATION DES PROBABILITES INITIALES.</u>	PAGE 39
<u>III.5 DETERMINATION DES COEFFICIENTS DE COMPATIBILITE.</u>	PAGE 41
<u>III.6 ANALYSE DU COMPORTEMENT DE L'ALGORITHME DE RELAXATION.</u>	PAGE 45
<u>REFERENCES (chapitre III)</u>	PAGE 50

IV. RECONSTITUTION DES MODES ET CLASSIFICATION.

<u>IV.1 INTRODUCTION.</u>	PAGE 51
<u>IV.2 DETERMINATION DES DOMAINES MODAUX INITIAUX.</u>	PAGE 51
<u>IV.3 EXTENSIONS SEQUENTIELLES DES DOMAINES MODAUX INITIAUX.</u>	PAGE 55
<u>IV.4 CLASSIFICATION GLOBALE.</u>	PAGE 59
<u>IV.5 PROCEDURE D'ANALYSE LOCALE.</u>	PAGE 60
<u>IV.6 CLASSEMENT BAYESIEN.</u>	PAGE 61
<u>REFERENCES (chapitre IV)</u>	PAGE 66
<u>ANNEXE 1</u>	PAGE 67

V. IDENTIFICATION DES MELANGES GAUSSIENS.

<u>V.1 INTRODUCTION.</u>	PAGE 73
<u>V.2 EXEMPLE DE BASE A DEUX DIMENSIONS.</u>	PAGE 73
<u>V.3 EFFET DES PARAMETRES DE REGLAGE DE L'ALGORITHME DE CLASSIFICATION.</u>	PAGE 86
V.3.1 Effet du pas de discrétisation.	PAGE 86
V.3.2 Effet du seuil de rejet des modes représentatifs.	PAGE 86
<u>V.4 CLASSIFICATION DE DONNEES BIDIMENSIONNELLES.</u>	PAGE 88
V.4.1 Classes équiprobables.	PAGE 88
V.4.2 Classes non équiprobables.	PAGE 92
<u>V.5 ANALYSE GLOBALE ET ANALYSE LOCALE.</u>	PAGE 98
V.5.1 Analyse globale.	PAGE 98
V.5.2 Analyse locale.	PAGE 102
<u>V.6 CLASSIFICATION DE DONNEES TRIDIMENSIONNELLES.</u>	PAGE 104
<u>V.7 CLASSIFICATION DE DONNEES QUADRIDIMENSIONNELLES.</u>	PAGE 107
<u>REFERENCES (chapitre V)</u>	PAGE 109

VI. CONCLUSION GENERALE.

<u>REFERENCES (chapitre VI)</u>	PAGE 113
---------------------------------	----------

CHAPITRE I

**CLASSIFICATION AUTOMATIQUE ET CONVEXITE
DES FONCTIONS DE DENSITE**

I.1 INTRODUCTION.

Le but de toute méthode de classification est de définir des groupes (ou classes) dans un ensemble d'individus dont la structure est inconnue à priori. Ainsi la description complète et détaillée de chaque individu est remplacée par son appartenance à une classe bien définie. Chacun de ces individus étant représenté dans un espace multidimensionnel, l'automatisation de la classification constitue un pas décisif dans le processus d'analyse et de compréhension des données, le cerveau humain étant incapable d'accomplir une synthèse multidimensionnelle. Néanmoins, la classification automatique apparaît comme un arsenal de méthodes qui n'ont en commun que leur finalité et qui font appel à un grand nombre de notions mathématiques et de concepts scientifiques. Devant le foisonnement des techniques permettant d'identifier les classes en présence dans un ensemble de données multidimensionnelles, l'utilisateur le plus averti reste souvent perplexe. En effet, l'analyste doit choisir une stratégie parmi cet ensemble de méthodes très diverses et souvent peu comparables, chacune ayant ses points forts et ses faiblesses.

Ce problème de décision apparaît surtout lorsqu'il s'agit d'identifier les classes en présence dans un échantillon, uniquement à partir de la seule information qui peut être extraite des observations à classer. Il s'agit alors de problèmes de classification non supervisée, qui apparaissent comme une démarche exploratoire, créatrice d'hypothèses, car elle permet, dans un domaine inconnu, de découvrir des structures qui n'étaient pas visibles sur les données.

On peut dissocier deux grands types de méthodes de classification non supervisée : les méthodes basées sur des concepts de distance et celles se référant à des notions de statistique.

I.2 APPROCHE METRIQUE.

On recherche une partition de l'ensemble des observations telle que chaque individu de l'échantillon considéré ressemble plus aux individus intérieurs à son groupe qu'aux individus extérieurs. La mesure la plus naturelle de la ressemblance entre deux individus est la distance séparant les deux observations correspondantes /6/.

Si certains algorithmes sont totalement heuristiques /4/, d'autres se veulent optimaux en ce sens qu'ils minimisent soit un critère de dispersion intra-groupe de type quadratique /2-3-10-17-27-33-35/ soit un critère calculé à partir de la matrice de dispersion /15-37-38-42/.

Toujours dans le cadre des méthodes de recherche de groupements faisant référence à la notion de distance, nous pouvons distinguer celles qui opèrent par groupements ou séparations successifs des observations /7-24-26-40/ de celles qui utilisent la théorie des graphes /1-43/.

A l'exception d'une seule méthode fondée sur l'utilisation d'un critère invariant /15/, les autres sont sensibles à tout changement d'échelle sur les caractères. Certaines dépendent de l'ordre dans lequel les observations sont considérées /3-27/. De plus, pour la majorité d'entre elles, le nombre de classes est supposé connu à priori et, évidemment, les résultats dépendent de ce nombre imposé par l'analyste. Or lorsqu'on traite des échantillons inconnus, ce choix est difficile à justifier /11-41/. Pour éviter ces inconvénients, on peut faire appel à des concepts statistiques plutôt qu'au concept de distance.

I.3 APPROCHE STATISTIQUE.

Il existe beaucoup moins de méthodes statistiques que de méthodes métriques pour la recherche de groupements parmi des observations multidimensionnelles /5-39/. De plus certaines d'entre elles ne font en fait référence à des notions de corrélation ou de variance que pour définir un critère algébrique à extrémiser. Il ne s'agit en réalité que d'une généralisation de la notion de distance /15-27-37/.

On ne parle réellement de méthodes statistiques que lorsqu'il est fait appel à la notion de densité de probabilité /9-14/. Dans ce cas, les groupements d'observations correspondent aux modes de la fonction de densité /3-16-18-19-20-21-22-23-28/ et la classification se ramène essentiellement à un problème de détection des modes.

La recherche des maxima locaux de la fonction de densité sous-jacente est souvent obtenue par des techniques basées sur la notion de gradient qui permettent de remonter les pentes de cette fonction /9-13-18-28/. En général, ces méthodes sont très sensibles aux irrégularités de l'estimation de la fonction de densité elle-même et tendent à générer un nombre de modes excessif.

I.4 CONVEXITE ET DETECTION DES MODES.

Pour éviter les inconvénients liés à l'utilisation d'opérateurs différentiels du type gradient, tout en gardant les avantages liés à l'utilisation des méthodes statistiques, nous avons opté pour une méthode utilisant le concept de convexité.

Dans ce cas, les modes ne sont plus considérés comme des extréma locaux de la fonction de densité mais comme des domaines où cette fonction est localement convexe /36/.

Cette approche apporte une solution au problème classique de la recherche des groupements lorsqu'on n'émet aucune hypothèse particulière sur la nature statistique de la distribution des données à classer, mais elle permet également d'optimiser une classification par minimisation du taux d'erreur /31/.

Toutefois, nous constatons que la qualité des résultats obtenus par cette méthode de classification dépend énormément du soin apporté à l'ajustement du pas de discrétisation de l'espace de représentation des données multidimensionnelles. En général, l'ajustement de ce pas est basé sur une notion de stabilité des modes /12/.

Il faut ajouter à ce problème de discrétisation, le problème lié à toutes les méthodes d'estimation statistique, c'est-à-dire leur sensibilité aux irrégularités de la distribution des données. Ces irrégularités proviennent en général des erreurs introduites durant le processus d'acquisition de ces données, autrement dit du bruit de mesure.

Pour remédier à tous ces problèmes, nous allons faire appel à une approche itérative basée sur des techniques d'étiquetage probabiliste ou "relaxation" qui, jusqu'à présent, a surtout été employée dans le cadre du traitement des images numériques /8-25-32-34/.

I.5 CONVEXITE ET RELAXATION.

Pour appliquer la méthode de classification statistique basée sur le concept de convexité, on discrétise l'espace de représentation des données sur un réseau hypercubique et on attribue à chaque hypercube constituant ce réseau une étiquette "convexe" ou "concave". Les différents problèmes rencontrés lors de l'utilisation de cette méthode statistique résultent de la présence d'hypercubes convexes isolés entourés d'hypercubes concaves et vice-versa /29-30/. Nous montrerons que ces erreurs d'étiquetage peuvent être corrigées par l'application d'une méthode itérative basée sur des techniques d'étiquetage probabiliste. L'analyse des étiquettes assignées aux voisins de l'hypercube considéré permet, selon leur compatibilité, de renforcer ou diminuer la probabilité que cet hypercube soit correctement étiqueté. De plus, ce traitement des données a pour avantage supplémentaire d'être effectué en parallèle, c'est-à-dire sur tous les hypercubes en même temps, ce qui rend la méthode indépendante de l'ordre de traitement.

On constatera que cette méthode augmente la stabilité des modes en réduisant les ambiguïtés locales, ce qui donnera plus de souplesse à l'analyste pour la détermination du pas de discrétisation.

I.6 RELAXATION ET DETECTION DES MODES.

Après avoir brièvement rappelé les principes de la détection des modes par analyse de la convexité de la fonction de densité sous-jacente, on montre comment le test de convexité mis en oeuvre débouche sur la notion d'étiquetage des points de discrétisation de l'espace de représentation des données (Chapitre II).

Le troisième chapitre de ce mémoire est consacré à l'élaboration d'une procédure d'étiquetage probabiliste pour l'analyse de la convexité des fonctions de densité. En chaque point de discrétisation, on détermine une probabilité pour que la fonction de densité soit concave ou convexe. Ces probabilités sont ensuite réajustées par un schéma itératif et parallèle. Elles sont renforcées lorsqu'elles sont compatibles dans un voisinage donné et diminuées en cas d'incompatibilité.

Les résultats obtenus par cette procédure d'étiquetage probabiliste sont ensuite pris en compte pour reconstituer les modes de la fonction de densité sous-jacente. Le chapitre IV présente une méthode d'agrégation globale qui permet de mettre en évidence des domaines modaux pour chacune des classes en présence. Cette méthode a l'avantage de réactualiser la définition de chacun des modes pour l'affiner au fur et à mesure que se déroule le processus d'agrégation. Une analyse locale permet finalement de parfaire la classification lorsque des classes présentant un fort degré de chevauchement n'ont pas été dissociées par la procédure globale.

Les modes ainsi reconstitués permettent de proposer des fonctions de décision adaptées à la classification d'observations issues de mélanges gaussiens. Le dernier chapitre présente quelques exemples de classification pour des données normales générées artificiellement. Les résultats obtenus montrent la stabilité de l'algorithme par rapport à l'ajustement des paramètres de réglage. Sans aucune connaissance a priori sur les données à classer, la procédure permet de retrouver les classes en présence et de leur assigner les observations avec un taux d'erreur très proche du taux d'erreur théorique optimal défini par la théorie de la décision.

REFERENCES (chapitre I)

- /1/ J. G. AUGUSTSON, *An analysis of some graph theoretical cluster techniques*, J. ACM, Vol. 17, pp. 571-588, 1970.
- /2/ G. H. BALL, *Data analysis in the social sciences: what about the details*, Proc. FJCC, Spartan Books, Washington, pp. 533-560, 1965.
- /3/ G. H. BALL et D. J. HALL, *A clustering technique for summarizing multivariate data*, Behaviorial Sc., Vol. 12, pp. 153-155, 1967.
- /4/ B. G. BATCHELOR et B. R. WILKINS, *Methods for location of cluster of patterns to initialize a learning machine*, Electronics Letters, Vol. 5, n° 20, pp. 481-483, 1969.
- /5/ C. K. BAYNE, J. J. BEAUCHAMP, C. L. BEGOVICH et V. E. KANE, *Monte Carlo comparison of selected clustering procedures*, Pattern Recog., Vol. 12, pp. 51-62, 1980.
- /6/ M. BOURTON, *Contribution à l'analyse, le traitement et la reconnaissance des formes ponctuéés. Application à la classification de couches géologiques*, Thèse d'état, Université de Lille 1, 1975.
- /7/ F. CAILLEZ et J. P. PAGES, *Introduction à l'analyse des données*, SMASH Paris, pp. 544-580, 1976.
- /8/ A. J. DANKER et A. ROSENFELD, *Blob detection by relaxation*, IEEE Trans. Pattern Anal. Mach. Intell., PAMI-3, pp. 79-92, 1981.
- /9/ P. A. DEVIJVER et J. KITTLER, P. R: *A statistical approach*, Prentice Hall, Englewood cliffs, 1982.
- /10/ E. DIDAY, *Une nouvelle méthode en classification automatique et reconnaissance des formes: la méthode des nuées dynamiques*, Revue de Stat. Appl., Vol. 19, n° 2, pp. 20-33, 1971.

- /11/ R. DUBES et A. K. JAIN, *Validity studies in clustering methodologies*, Pattern Recog., Vol. 11, pp. 235-254, 1979.
- /12/ D. J. EIGEN , F. R. FROMM et R. A. NORTHOUSE, *Cluster analysis based on dimensional information with applications to feature selection and classification*, IEEE Trans. On Syst., Man. & Cyb., Vol. SMC-4, n° 3, pp. 284-294, 1974.
- /13/ B. EVERITT, *Cluster analysis*, Social Science Res. Courial, Holsted Press, G.B., 1980.
- /14/ J. L. FLEISS et J. ZUBIN, *On the method and theory of clustering*, Multivariate behavioral research, pp. 235-250, 1969.
- /15/ H. P. FRIEDMAN et J. ROBIN, *On some invariant criteria for grouping data*, J. Amer. Stat. Ass., Vol 62, pp. 1159-1178, 1967.
- /16/ K. FUKUNAGA et L. D. HOSTETLER, *The estimation of a density function with applications in pattern recognition*, IEEE Trans. on Inf. Theory, Vol IT-21, pp. 32-40, 1975.
- /17/ K. FUKUNAGA et W. L. G. KOONTZ, *A criterion and an algorith for grouping data*, IEEE Trans. on Computers, Vol C-19, pp. 917-923, 1970.
- /18/ I. GITMAN, *An algorith for nonsupervised pattern classification*, IEEE Trans. on Syst., Man. & Cyb., Vol. SMC-3, pp. 66-74, 1973.
- /19/ I. GITMAN et M. D. LEVINE, *An algorith for unimodal fuzy sets and its applications as a clustering technique*, IEEE Trans. on Computers, Vol. C-19, pp. 583-93, 1970.
- /20/ B. JOHNSTON, T. BAILEY et R. DUBES, *A variation on a nonparametric clustering method*, IEEE Trans. on Inf. Theory, Vol. IT-21, n° 1, pp. 32-40, 1975.

- /21/ J. O. KATZ et F. J. ROHLF, *Function point cluster analysis*, *Systematic Zoology*, Vol. 22, pp. 295-301, 1973.
- /22/ J. KITTLER, *A locally sensitive method for cluster analysis*, *Pattern Recog.*, Vol. 8, pp. 23-33, 1976.
- /23/ W. L. G. KOONTZ, P. M. NARANDRA et K. FUKUNAGA, *A graph theoretic approach to nonparametric cluster analysis*, *IEEE Trans. on Computers*, Vol. C-25, n° 9, pp. 936-944, 1976.
- /24/ G. N. LANE et W. T. WILLIAMS, *A general theory of classificatory sorting strategies 1*, *Hierarchical systems*, *Computer Jour.*, Vol. 9, pp. 973-980, 1967.
- /25/ A. LEV, S. W. ZUCKER and A. ROSENFELD, *Iterative enhancement of noisy images*, *IEEE Trans. Syst., Man. Cybern.*, Vol. SMC-7, pp; 435-442, June 1977.
- /26/ A. LUKASOVA, *Hierarchical agglomerative clustering procedure*, *Pattern Recog.*, Vol. 11, pp. 365-81, 1979.
- /27/ J. MAC QUEEN, *Some methods for classification and analysis of multivariate observations*, *Proc. 5th Berkeley Symp. on Math. Stat. and Prob.*, pp. 281-297, 1967.
- /28/ R. MIZOGUCHI et M. SHIMURA, *Non parametric learning without a teacher based on mode estimation*, *IEEE Trans. on Computers*, Vol. C-25, pp. 1109-1117, 1976.
- /29/ J. G. POSTAIRE, *Fonctions convexes et optimisation du processus de classification automatique. I . Identification des mélanges gaussiens par estimation de la convexité des fonctions de densité multivariables*, *R.A.I.R.O. Automatique/Systems Analysis and Control*, Vol. 16, n° 4, pp. 357-379, 1982.

- /30/ J. G. POSTAIRE, *Fonctions convexes et optimisation du processus de classification automatique. II. Identification de mélanges gaussiens à partir de petits échantillons par estimation de la convexité des fonctions de densité marginales*, R.A.I.R.O. Automatique/Systems Analysis and Control, Vol. 17, n° 1, pp. 39-59, 1983.
- /31/ J. G. POSTAIRE et C. VASSEUR, *An approximate solution of normal mixture identification with application to unsupervised pattern classification*, IEEE Trans. on Pattern Anal. & Machine Intelligence, Vol. PAMI-3, n° 2, pp. 163-179, 1981.
- /32/ A. ROSENFELD, R. A. HUMMEL et S. W. ZUCKER, *Scene labeling by relaxation operations*, IEEE Trans. Syst., Man. Cybern., Vol. SMC-6, pp. 420-433, 1976.
- /33/ G. S. SEBESTYEN, *Pattern recognition by an adaptive process of sample set construction*, IRE Trans. on Info. Theory, Vol. IT-8, pp. 82-91, 1962.
- /34/ R. SMITH et A. ROSENFELD, *Thresholding using relaxation*, IEEE Trans. Pattern Anal. Mach. Intell., PAMI-3, pp. 79-92, 1981.
- /35/ R. L. THORNDIKE, *Who belong in the family ?*, Psychometrica, Vol. 18, pp. 267-276, 1953.
- /36/ C. VASSEUR et J. G. POSTAIRE, *A convexity testing method for cluster analysis*, IEEE Trans. on Syst., Man. & Cyb., Vol. SMC-10, pp. 145-149, 1980.
- /37/ M. A. VOGEL et A. K. C. WONG, *PFS clustering method*, IEEE Trans. on Pattern Analysis & Machine Intel., Vol. PAMI-1, pp. 237-245, 1979.
- /38/ S. S. WILKS, *Multidimensional statistical scatter*, Contrib. to Proba. & Stat. I. Olkin, ed. Stanford, Calif., 1960.

- /39/ D. WISHART, *Estimating the modes of a multivariate sample density*, Proc. Joint Meeting of the Classification Soc. & British Pattern Recog. Ass. London, 1978.
- /40/ D. WISHART, *Mode analysis: a generalization of nearest neighbour which reduces chaining effects*, Numerical Taxinomy, pp. 282-308, A. J. Cole, ed. Academic Press New York, 1969.
- /41/ J. H. WOLFE, *Pattern clustering for multivariate mixture analysis*, Multiv. Behavioral Research, Vol. 5, pp. 329-350, 1970.
- /42/ S. S. YAU et S. C. CHANG, *A direct method for cluster analysis*, Pattern Recognition, Vol. 7, pp. 215-224, 1975.
- /43/ C. T. ZAHN, *Graph theoretical methods for detecting and describing gestalt clusters*, IEEE Trans. on Computers, Vol. C-20, pp. 68-86, 1971.

CHAPITRE II

**ANALYSE DE LA CONVEXITE DES FONCTIONS
DE DENSITE DE PROBABILITE**

II.1 INTRODUCTION.

Lorsque l'analyste ne dispose d'aucune autre information que celle qui peut être extraite de l'échantillon disponible, l'analyse de la convexité de la fonction de densité de probabilité sous-jacente constitue une approche bien adaptée au problème de la classification non supervisée. Nous nous proposons de faire quelques rappels sur l'étude de cette convexité afin d'en déduire ses possibilités d'application et ses limitations.

II.2 CONVEXITE D'UNE FONCTION DE DENSITE DE PROBABILITE MULTIVARIABLE.

II.2.1 Principe d'analyse.

Une fonction $f(X)$ est dite convexe sur un domaine convexe D de \mathbb{R}^N si :

$$f(X_1 + (1 - \mu)X_2) < f(X_1) + (1 - \mu)f(X_2)$$

$$\forall X_1, X_2 \in D \text{ et } \mu \in [0,1]$$

On pourrait envisager d'exploiter cette définition pour déterminer le sens de la convexité des fonctions de densité de probabilité. Une telle approche semble possible lorsqu'on dispose d'un modèle analytique de ces fonctions. Cependant, dans notre problème, on ne connaît pas l'expression analytique des fonctions de densité. En effet, ne disposant d'aucune autre information que celle apportée par les observations à classer, on ne peut qu'estimer point par point la fonction de densité de probabilité sous-jacente à partir des données en analysant leur distribution dans l'espace de représentation.

L'expression ci-dessus ne pouvant donc pas être utilisée pour l'étude de la convexité des fonctions de densité, nous allons faire appel à une méthode permettant de tester localement le sens de la convexité des fonctions multivariées /18/.

Le principe de cette méthode est le suivant :

Une fonction $f(X)$ est dite "localement convexe" au point X_0 s'il existe un voisinage de X_0 convexe sur lequel la fonction $f(X)$ est convexe /13/. Pour déterminer la convexité d'une fonction en un point X_0 , on analyse les variations de sa valeur moyenne :

$$\bar{f}[D(X_0, \alpha)] = \frac{\int_{D(X_0, \alpha)} f(X) dX}{V(D(X_0, \alpha))}$$

où $D(X_0, \alpha)$ est un domaine homothétique d'un domaine de référence $D_0(X_0)$, symétrique par rapport à son centre X_0 , de volume-unité, dans une homothétie de centre X_0 et de rapport α . L'ajustement du paramètre α permet de faire varier le volume du domaine $D(X_0, \alpha)$, appelé domaine d'observation (figure 1).

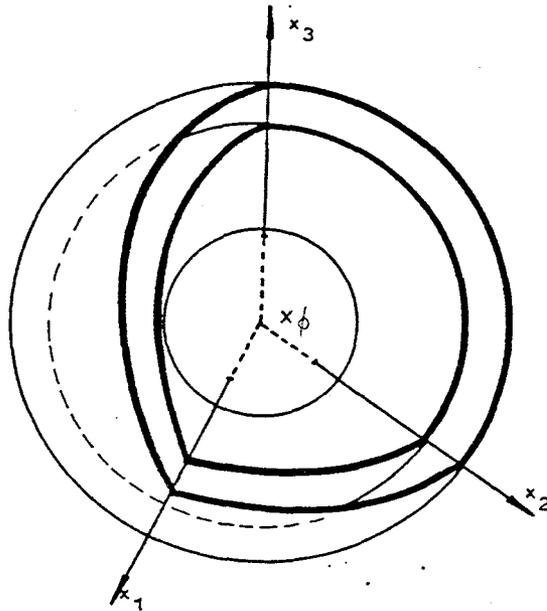


figure 1 : Domaines d'observation sphériques dans un espace à trois dimensions.

Il a été montré /19/ que si $f(X)$ est localement convexe, alors la valeur moyenne de $f(X)$ sur les domaines d'observations $D(X_0, \alpha)$ est une fonction décroissante de α , donc de la taille de ces domaines.

Inversement, si $f(X)$ est localement concave, alors la valeur moyenne de $f(X)$ sur les domaines d'observations $D(X_0, \alpha)$ est une fonction croissante de la taille de ces domaines .

On peut donc, à partir de ce résultat, déterminer facilement le sens de convexité des fonctions de densité en tout point X_0 de l'espace de représentation des données.

Pour cela, nous sommes amenés à utiliser une méthode d'estimation pour déterminer $\varrho [D(X_0, \alpha)]$.

II.2.2 ESTIMATION DE LA VALEUR MOYENNE D'UNE FONCTION.

Les principes de base de nombreuses méthodes d'estimation d'une fonction de densité de probabilité sont très simples, bien que les démonstrations rigoureuses de la convergence des estimateurs requièrent considérablement de soins /3-5-7/.

De nombreuses techniques d'estimation reposent sur le fait que la probabilité $P(X \in D(X_0, \alpha))$ pour qu'une observation X soit située dans le domaine $D(X_0, \alpha)$ est donnée par :

$$P(X \in D(X_0, \alpha)) = \int_{D(X_0, \alpha)} f(X') d(X')$$

En divisant $P(X \in D(X_0, \alpha))$ par $V(D(X_0, \alpha))$, on obtient la valeur moyenne de la fonction de densité $f(X)$ sur le domaine $D(X_0, \alpha)$ qui est souvent confondue avec l'estimation de la fonction de densité en X_0 /5/.

En supposant que l'on dispose d'un échantillon constitué de Q observations X_1, X_2, \dots, X_Q , indépendamment distribuées suivant la loi de probabilité sous-jacente $f(X)$ dans l'espace de représentation, la probabilité P_k pour que k de ces Q observations soient situées dans le domaine $D(X_0, \alpha)$ est donnée par la loi binomiale :

$$P_k = C_Q^k P(X \in D(X_0, \alpha))^k [1 - P(X \in D(X_0, \alpha))]^{Q-k}$$

pour laquelle l'espérance mathématique de k est :

$$\begin{aligned} E(P_k) &= \sum_{k=0}^Q k C_Q^k P(X \in D(X_0, \alpha))^k [1 - P(X \in D(X_0, \alpha))]^{Q-k} \\ &= Q \cdot P(X \in D(X_0, \alpha)) \end{aligned}$$

Ce résultat indique que k/Q est un estimateur non biaisé de $P(X \in D(X_0, \alpha))$.

On a donc :

$$\int_{D(X_0, \alpha)} f(X') dx' \approx k/Q$$

Dans ces conditions, on peut prendre

$$\hat{\xi} [D(X_0, \alpha)] = \frac{k/Q}{V(D(X_0, \alpha))}$$

pour estimateur de la valeur moyenne de $f(X)$ sur $D(X_0, \alpha)$.

Pour déterminer localement le sens de variation de la valeur moyenne de $f(X)$, c'est-à-dire le sens de la convexité de $f(X)$ en un point X_0 , il faut d'abord estimer la valeur moyenne de la fonction de densité de probabilité $f(X)$ sur un domaine $D(X_0, \alpha)$ aussi petit que possible.

Soit $\hat{e} [D(X_0, \alpha)]$ cet estimateur :

$$\hat{e} [D(X_0, \alpha)] = \frac{k/Q}{V(D(X_0, \alpha))}$$

Il faut ensuite estimer la valeur moyenne de $f(X)$ sur un domaine $D(X_0, \alpha')$ de taille légèrement supérieure à celle du domaine $D(X_0, \alpha)$:

$$\hat{e} [D(X_0, \alpha')] = \frac{k'/Q}{V(D(X_0, \alpha'))}$$

k' est le nombre d'observations contenues dans le domaine $D(X_0, \alpha')$, tel que $V(D(X_0, \alpha')) > V(D(X_0, \alpha))$

En comparant ces deux estimations, on détermine directement le sens de la convexité de $f(X)$ en X_0 .

Si $\hat{e} [D(X_0, \alpha)] < \hat{e} [D(X_0, \alpha')]$, la valeur moyenne de $f(X)$ est une fonction croissante du paramètre α : la fonction $f(X)$ est donc localement concave en X_0 .

Si $\hat{e} [D(X_0, \alpha)] > \hat{e} [D(X_0, \alpha')]$, la fonction est localement convexe en X_0 .

Enfin, dans le cas où $\hat{e} [D(X, \alpha)] = \hat{e} [D'(X, \alpha)]$, on ne peut conclure quant au sens de la convexité de la fonction $f(X)$.

En utilisant l'estimateur de Parzen-Rosenblatt :

$$\hat{f}(X_0) = \frac{1}{Q} \sum_{q=1}^Q \frac{1}{V(D(X_0, \alpha_q))} g \left[\frac{X_0 - X_q}{\alpha_q} \right]$$

avec le noyau cubique, qui est certainement l'un des plus utilisés dans la pratique /6-7-8/, défini par /1/:

$$g(Y) = \begin{cases} 1 & \text{si } |y_m| \leq 1/2, m=1, \dots, M \\ \emptyset & \text{autrement} \end{cases}$$

avec $Y = [y_1, \dots, y_n, \dots, y_N]^T$,

ou en faisant appel à l'estimateur des k plus proches voisins /19/, on constate que la valeur moyenne de la fonction f(X) estimée sur le domaine D(X₀, α) n'est autre que l'estimation usuelle de cette fonction en X₀.

On peut donc écrire :

$$\hat{f}(X_0) = \hat{E}[D(X_0, \alpha)] = \frac{k/Q}{V(D(X_0, \alpha))}$$

L'utilisation de ces estimateurs simples soulève plusieurs problèmes théoriques et pratiques.

Si nous fixons le volume V(D(X₀, α)) et que nous prenons de plus en plus d'observations, le rapport k/Q va converger en probabilité vers P(X ∈ D(X₀, α)). En divisant cet estimateur par V(D(X₀, α)), nous obtenons effectivement une estimation de la valeur moyenne de f(X) sur le domaine D(X₀, α) et non la valeur de f(X) en X₀.

Pour obtenir un estimateur de $f(X)$ en un point X_0 , il faudrait donc faire tendre le volume $V(D(X_0, \alpha))$ vers zéro, c'est-à-dire, faire tendre le rapport d'homothétie α vers zéro. Il importe donc d'étudier le comportement asymptotique de l'estimateur $\hat{e} [D(X_0, \alpha)]$ lorsque α tend vers zéro.

Pour assurer la convergence de l'estimateur vers $f(X_0)$, il a été montré que le volume du domaine intervenant dans l'expression de $\hat{e} [D(X_0, \alpha)]$ ainsi que le nombre d'observations k situées dans ce domaine doivent être liés au nombre total Q d'observations disponibles.

Plus précisément, soit α_Q la valeur de α et k_Q la valeur de k correspondant à un échantillon de Q observations. On démontre /21/ que $\hat{e} [D(X_0, \alpha_Q)]$ converge en moyenne quadratique vers $f(X_0)$ si :

$$(a) \quad \lim_{Q \rightarrow \infty} V(D(X_0, \alpha_Q)) = 0$$

$$(b) \quad \lim_{Q \rightarrow \infty} k_Q = \infty$$

$$(c) \quad \lim_{Q \rightarrow \infty} k_Q/Q = 0$$

Il y a deux façons d'obtenir des domaines $D(X_0, \alpha_Q)$ satisfaisant ces conditions.

La première est de réduire la taille du domaine $D(X_0, \alpha_Q)$ lorsque Q augmente en spécifiant le volume $V(D(X_0, \alpha_Q))$ comme une fonction de Q . Il a été montré qu'avec des expressions, telles que :

$$V(D(X_0, \alpha_Q)) = V_0 \cdot Q^{-b}, \quad b \in]0, 1[$$

ou

$$V(D(X_0, \alpha_Q)) = V_0 / \log Q,$$

où les variables aléatoires k_Q et k_Q/Q répondent aux conditions (b) et (c). C'est la méthode du noyau, initialement

proposée par Rosenblatt /15/ dans le cas monovariante, puis approfondie par Parzen /12/ avant d'être étendue au cas multivariante par Murthy /11/ et par Cacoullos /2/.

La deuxième méthode consiste à spécifier k_Q comme une fonction de Q , telle que $k_Q = k_0 \cdot \sqrt{Q}$. Ici, on augmente le volume $V(D(X_0, \alpha_Q))$ jusqu'à ce que celui-ci contienne les k_Q plus proches voisins de X_0 . C'est la méthode d'estimation des k_Q plus proches voisins /3-10/.

Les deux méthodes convergent et chacune d'elles présente des avantages spécifiques /17-20/.

Néanmoins quelle que soit la méthode choisie, il y aura un choix déterminant à faire:

- soit le choix de V_0 dans la méthode de Rosenblatt-Parzen,

- soit celui de k_0 dans la méthode des k_Q plus proches voisins,

chacun de ces choix ayant ses propres contraintes.

Lorsque le nombre Q d'observations disponibles tend vers l'infini, on sait que, théoriquement, quel que soit le choix effectué, il y a convergence asymptotique de l'estimateur. Or, dans la pratique, le nombre d'observations est toujours limité. Avec un ajustement de V_0 trop petit, on obtiendrait une estimation très bruitée de la densité de probabilité. En effet, l'estimateur de $f(X_0)$ serait nul dans presque tout l'espace. Seuls les domaines $D(X_0, \alpha_Q)$ centrés en des points X_0 proches des observations ne seraient pas vides et l'estimateur correspondant tendrait alors vers l'infini. A l'inverse, avec un volume V_0 trop grand, l'estimateur manquerait de résolution. Le volume $V(D(X_0, \alpha_Q))$ ne peut donc devenir ni arbitrairement petit, ni arbitrairement grand, ce qui explique que l'analyste utilise en fait toujours une estimation de la valeur moyenne plutôt qu'une estimation ponctuelle de $f(X)$. Ainsi, lorsqu'il s'agit d'estimer $f(X)$ en X_0 , on ajuste en général V_0 afin d'obtenir une estimation ni trop lissée, ni trop bruitée de la fonction de densité.

Dans le contexte de la détermination de la convexité de la fonction de densité $f(X)$, on ajuste la taille du domaine $D(X_0, \alpha)$ comme s'il s'agissait d'estimer $f(X)$ en X_0 , en fonction du nombre Q d'observations disponibles /13/. On prend ensuite pour $D(X_0, \alpha')$ un domaine de taille légèrement supérieure.

Si l'on désire décrire une fonction en termes de convexité sur tout son espace de définition, il faut appliquer le test de convexité point par point sur tout cet espace. Or le nombre de points nécessaires pour décrire correctement une fonction multivariable croît exponentiellement avec la dimension de l'espace de définition, ce qui entraîne des temps de calcul importants. De plus, pour déterminer le sens de la convexité de la fonction de densité de $f(X)$, il faut appliquer cette méthode d'estimation en chaque point par deux fois, la valeur moyenne de $f(X)$ étant d'abord estimée sur un domaine d'observation $D(X_0, \alpha_Q)$ aussi petit que possible puis sur un domaine d'observation $D(X_0, \alpha'_Q)$ de taille légèrement supérieure. Cette méthode de détermination du sens de la convexité de $f(X)$ ne pourra donc être employée efficacement qu'en faisant appel à une procédure permettant d'accélérer sa mise en oeuvre /14/. Cette procédure rapide nécessite une transformation préalable des données.

II.3 TRANSFORMATION LINEAIRE DE L'ESPACE D'OBSERVATIONS.

Pour permettre une implantation rapide de la procédure d'estimation, on normalise d'abord les plages de variations de chacune des composantes des observations de telle sorte que leurs valeurs extrêmes soient 0 et 1.

Soit $[X_q]$ l'ensemble des Q observations disponibles de la forme:

$$X_q = [X_{1,q}, \dots, X_{n,q}, \dots, X_{N,q}]^T, \quad q=1,2, \dots, Q$$

où N est la dimension de l'espace de représentation des données R^N .

Soient:

$$\begin{array}{l}
 X_n = \max_q X_{n,q} \quad q=1,2, \dots, Q \\
 \text{et} \\
 X_n = \min_q X_{n,q} \quad q=1,2, \dots, Q
 \end{array}
 \quad \left. \vphantom{\begin{array}{l} X_n \\ \text{et} \\ X_n \end{array}} \right\} n=1,2, \dots, N$$

les valeurs maximales et minimales des différentes composantes.

En appliquant une translation de vecteur: $T=[t_1, t_2, \dots, t_N]^T$

$$\text{avec } t_n = -X_n^m \quad n=1,2, \dots, N$$

suivie d'une transformation linéaire de matrice :

$$M = \begin{bmatrix} m_{11} & & & \emptyset \\ & & m_{nn} & \\ \emptyset & & & m_{NN} \end{bmatrix}$$

$$\text{avec } m_{nn} = \frac{1}{X_n^M - X_n^m} \quad n=1,2, \dots, N$$

on obtient de nouvelles composantes de la forme:

$$X' = M \cdot [X + T]$$

Dans ce nouvel espace noté R_1^N , les composantes des observations $X'_q = [X'_{1,q}, \dots, X'_{n,q}, \dots, X'_{N,q}]^T$ sont normalisées de sorte que:

$$\begin{array}{l}
 \emptyset \leq X'_{n,q} \leq 1 \quad n=1,2, \dots, N \\
 \quad \quad \quad \quad \quad \quad \quad q=1,2, \dots, Q
 \end{array}$$

si bien que toutes les observations sont situées dans un hypercube-unité (figures 2 et 3).

Espace R_0^N :

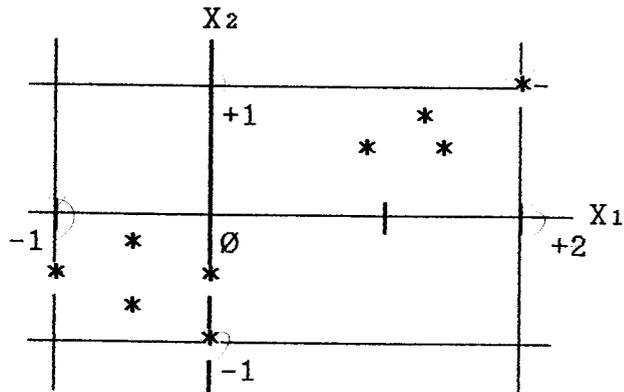


Figure 2: Données brutes.

Espace R_1^N :

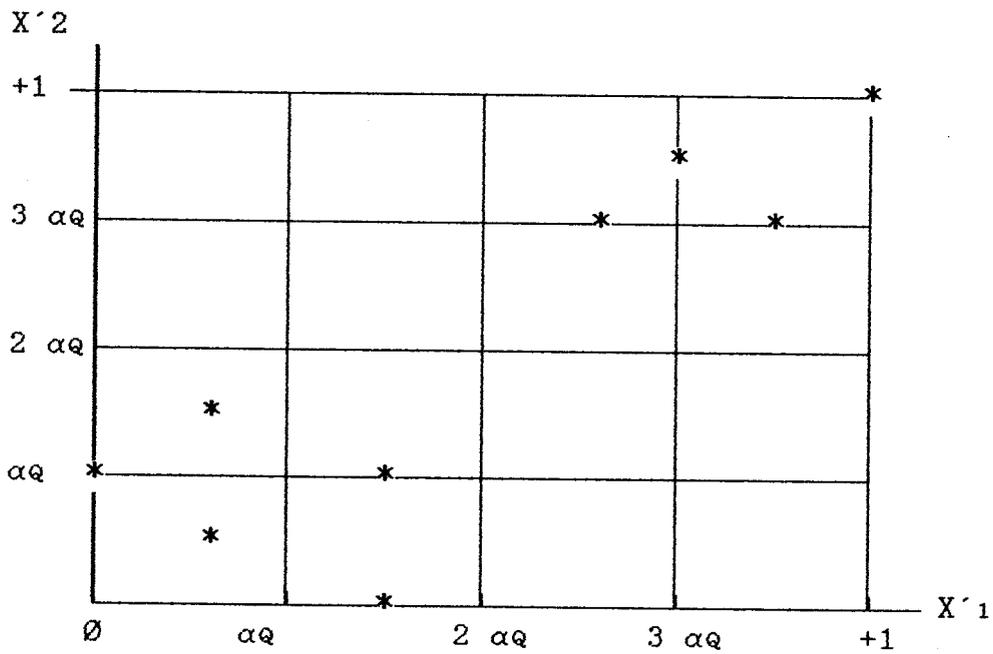


Figure 3 : Données normalisées par la transformation :

$$X' = \begin{bmatrix} X'_1 \\ X'_2 \end{bmatrix} = \begin{bmatrix} 1/3 & 0 \\ 0 & 1/2 \end{bmatrix} \cdot \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} - \begin{bmatrix} -1 \\ -1 \end{bmatrix}$$

Supposons que l'on désire estimer la fonction de densité de probabilité dans ce nouvel espace en utilisant des domaines d'observation hypercubiques de côté αq adjacents, comme représenté sur la figure 3.

Afin d'accélérer la procédure d'estimation, on multiplie préalablement chaque coordonnée de l'espace normalisé par $1/\alpha q = Nq$. Nq représente en fait le nombre de points de discrétisation sur chaque axe et αq le pas de cette discrétisation dans l'espace normalisé (figure 3).

Espace $R_{\alpha q}^N$:

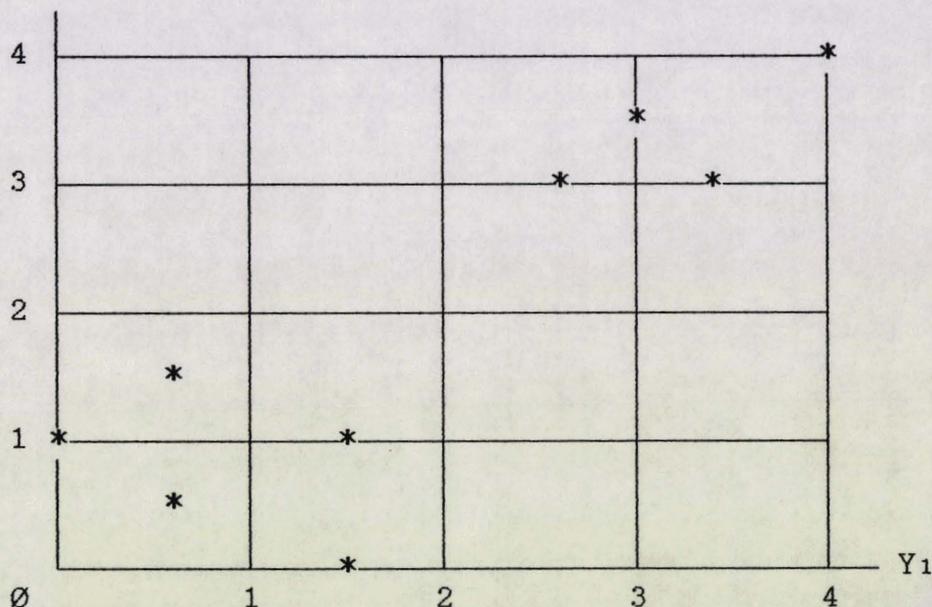


Figure 4 : Transformation de l'espace normalisé pour $Nq = 4$. On applique la transformation:

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} 4 & 0 \\ 0 & 4 \end{bmatrix} \cdot X'$$

Cette transformation conduit aux nouvelles coordonnées Y de la forme :

$$Y = A.M. [X+T] = A.X'$$

avec $A = 1/\alpha q.[IN] = Nq.[IN]$

où IN est la matrice-unité d'ordre N .

Dans ce nouvel espace, chaque domaine d'observation devient un hypercube de côté unité (Figure 4).

II.4 TEST RAPIDE DE CONVEXITE.

On constate que, dans le cas extrême où il n'y a pas plus qu'une observation par hypercube de discrétisation, le nombre d'hypercubes non vides est au plus égal au nombre Q d'observations. L'estimation de la fonction de densité $f(X)$ en utilisant les hypercubes unitaires pour domaines d'observation est donc nulle, sauf en un nombre limité de points au plus égal à Q . La procédure d'estimation de Rosenblatt-Parzen avec noyau cubique peut ainsi être accélérée en n'estimant $f(X)$ que dans les hypercubes non vides de ce nouvel espace, sous la forme:

$$\hat{f}(X_0) = \frac{\sum [D(X_0, \alpha_q)]}{V(D(X_0, \alpha_q))} = \frac{k/Q}{V(D(X_0, \alpha_q))}$$

où $V(D(X_0, \alpha_q))$ est le volume de l'hypercube non vide considéré de centre X_0 , dans lequel sont situées k des Q observations disponibles /14/.

Sans faire aucun calcul, on sait que si l'hypercube considéré est vide, l'estimateur associé de $f(X)$ est nul.

On peut exploiter cette remarque en extrayant directement la liste des hypercubes non vides de l'espace R_q^N de la liste des observations disponibles /14/.

En effet, dans l'espace transformé, une observation:

$$Y_q = [Y_{1,q}, Y_{2,q}, \dots, Y_{N,q}]^T$$

est située dans l'hypercube dont le centre a pour coordonnées:

$$\text{int}(Y_q) + 0.5 = [\text{int}(Y_{1,q}) + 0.5, \text{int}(Y_{2,q}) + 0.5, \dots, \text{int}(Y_{N,q}) + 0.5]^T$$

Cette procédure permet de dresser immédiatement la liste des hypercubes non vides à partir de la liste des observations et de déterminer le nombre d'observations situées dans chacun d'eux.

La valeur moyenne de $f(X)$ est ainsi estimée uniquement dans les hypercubes non vides $D(X_0, \alpha)$ de l'espace de départ, dont le nombre ne peut dépasser Q :

$$\hat{e} [D(X_0, \alpha)] = \frac{k/Q}{V(D(X_0, \alpha))}$$

On estime ensuite la valeur moyenne de $f(X)$ sur un hypercube également centré en X_0 , mais de côté triple de celui de $D(X_0, \alpha)$: $D(X_0, \alpha')$. Dans ces conditions, on a :

$$V(D(X_0, \alpha')) = 3^N \cdot V(D(X_0, \alpha))$$

et

$$\hat{e} [D(X_0, \alpha')] = \frac{k'/Q}{V(D(X_0, \alpha'))} = \frac{k'/Q}{3^N \cdot V(D(X_0, \alpha))}$$

où k' est le nombre d'observations situées dans l'hypercube $D(X_0, \alpha')$, obtenu par sommation des nombres d'observations situées dans les 3^N hypercubes élémentaires constituant $D(X_0, \alpha')$ (Figure 5).

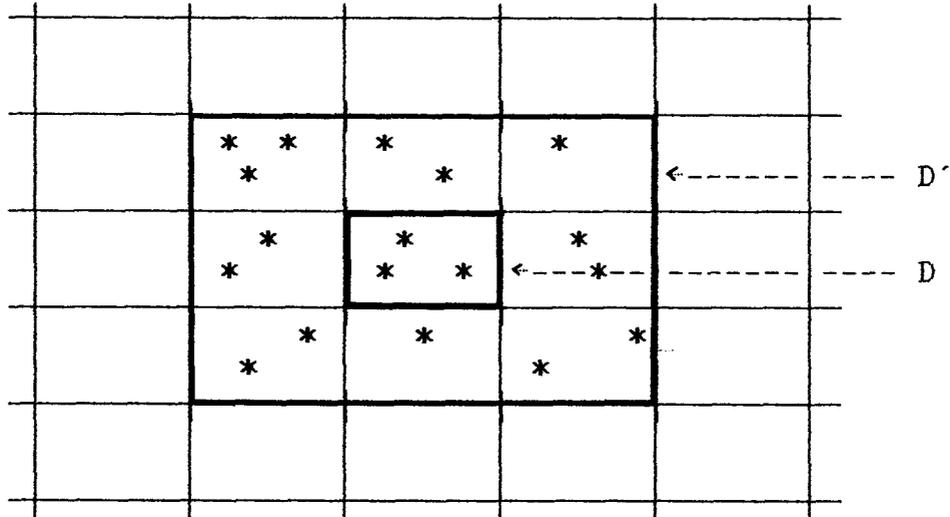


Figure 5 : Implantation du test de convexité dans un espace à deux dimensions.

La détermination du sens de la convexité de $f(X)$ en X_0 se fait alors aisément.

Si $\hat{\epsilon} [D(X_0, \alpha)] < \hat{\epsilon} [D'(X_0, \alpha)]$, c'est-à-dire

si $k/Q < k'/(3^N \cdot Q)$ ou encore

si $k < k'/3^N$

la fonction $f(X)$ est considérée comme localement concave dans l'hypercube $D(X_0, \alpha)$ auquel on affecte l'étiquette "concave".

Inversement,

si $k > k'/3^N$,

elle est considérée comme localement convexe dans l'hypercube $D(X_0, \alpha)$ auquel on affecte l'étiquette "convexe".

Il faut garder à l'esprit que le résultat de ce test n'a de sens que si la convexité de $f(X)$ est définie dans l'hypercube $D(X_0, \alpha')$. En effet, à l'encontre des fonctions monovariabiles, la convexité des fonctions multivariabiles peut ne pas être définie dans certaines régions de l'espace.

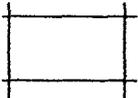
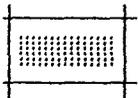
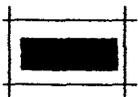
Dans de telles régions, on ne peut attribuer aucune valeur au résultat du test qui indique indifféremment l'un ou l'autre des deux sens de la convexité. Nous verrons par la suite comment éliminer les résultats erronés ainsi obtenus dans ces régions par un traitement approprié.

D'autre part, le fait de découper l'espace de représentation en hypercubes adjacents entraîne la création de cellules vides dans des régions de l'espace où la fonction de densité de probabilité n'est pas nulle. En effet, les utilisateurs de la méthode d'estimation avec noyau cubique ont souvent constaté que les résultats peuvent considérablement changer selon qu'une seule observation se situe d'un côté ou de l'autre d'une face d'un hypercube. C'est d'ailleurs une des motivations de l'utilisation des noyaux plus "doux", tels que les noyaux triangulaires ou gaussiens. Ce phénomène, si l'on ne prend pas garde, risque de perturber fortement le résultat final de la classification.

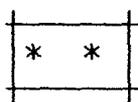
Pour remédier à la création de ces hypercubes vides perturbant la classification, on effectue un pré-traitement sur les données.

II.5 Pré-traitement des données.

A la suite de la discrétisation de l'espace de représentation des données, nous sommes en présence de 4 types d'hypercubes (figure 6):

- | | | |
|--------|---|--|
| type 1 |  | hypercube vide entouré uniquement d'hypercubes également vides |
| type 2 |  | hypercube vide ayant un voisinage constitué d'hypercubes vides et d'hypercubes non vides |
| type 3 |  | hypercube vide ayant essentiellement des hypercubes voisins non vides |

type 4



hypercube non vide

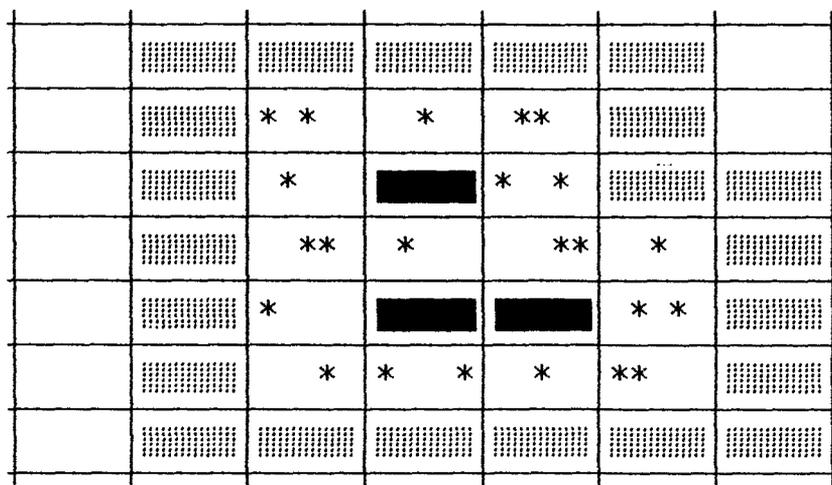


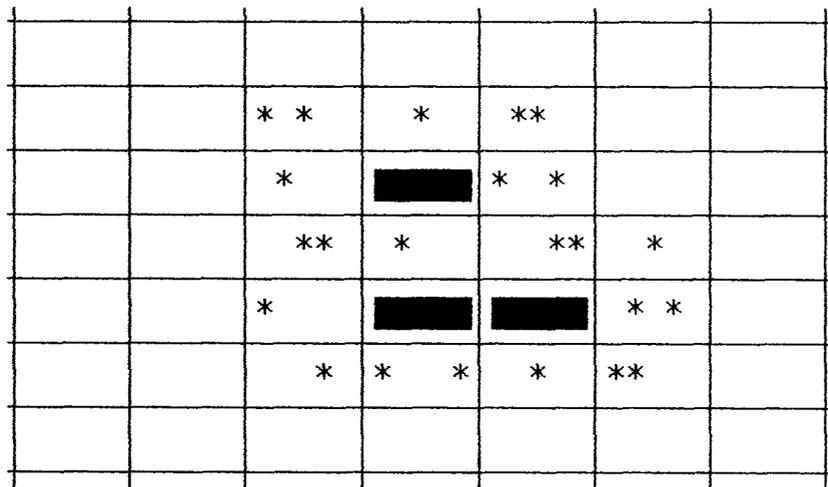
Figure 6 : Différents types d'hypercubes rencontrés après discrétisation de l'espace des données.

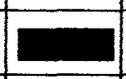
La liste des hypercubes non vides déterminée directement à partir de la liste des observations disponibles doit donc être complétée par la liste des hypercubes vides pouvant apparaître dans des régions de l'espace où la fonction de densité n'est pas uniformément nulle (type 3). L'apparition de tels hypercubes résulte souvent d'un pas de discrétisation trop petit.

Il est bien sûr hors de question de prendre en compte tous les hypercubes vides de type 1 et de type 2 de l'espace, ce qui reviendrait à abandonner la procédure d'estimation rapide. En effet, le nombre d'estimations locales de la fonction serait alors une fonction exponentielle de la dimension de l'espace de représentation des données et le temps d'exécution de la procédure deviendrait vite prohibitif /7/. Les seuls hypercubes vides à prendre en compte ne sont autres que ceux situés à l'intérieur des régions modales, c'est-à-dire les hypercubes de type 3.

Afin de les déterminer rapidement, on examine le voisinage de tous les hypercubes vides et on ne prend en considération que ceux dont le voisinage est constitué d'un nombre suffisant d'hyperscubes appartenant à la liste des hypercubes non vides. Ainsi, si ce nombre est nul, ce qui se produit dans le cas où on est en présence d'un hypercube vide situé dans une région de très faible densité, l'hyperscube est ignoré. Si ce nombre est inférieur à la moitié du nombre total d'hyperscubes voisins, ce qui est souvent le cas pour les hypercubes situés sur le pourtour d'un mode de la fonction de densité, cet hypercube est également ignoré.

Dans le cas de l'étude d'un espace de dimension N, ce nombre devra donc être au moins égal à $(3^N - 1)/2$ pour qu'un hypercube soit catalogué dans le type 3 (figure 7).




hypercube vide pris en compte

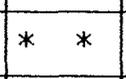

hypercube non vide

Figure 7 : Hypercubes pris effectivement en compte pour l'exemple bidimensionnel de la figure 6.

Après avoir complété la liste des hypercubes non vides par la liste des hypercubes vides devant être pris en compte, on obtient la liste des hypercubes sur lesquels on applique une procédure de lissage à l'aide d'une fenêtre glissante. Celle-ci consiste à attribuer à chaque hypercube de la liste un poids dont la valeur est égale à la moyenne, par hypercube, des nombres d'observations situées dans l'hypercube considéré et les hypercubes de son voisinage apparaissant dans la liste.

Ce lissage permet d'atténuer les problèmes résultant d'une discrétisation trop fine de l'espace de représentation.

II.6 CONCLUSION.

Le problème du choix du nombre de divisions N_q par composante reste primordial pour la suite des opérations et donc pour le résultat de la classification. Ce découpage de l'espace en domaines d'observation peut, s'il est trop fin, rendre la procédure trop sensible à de faibles variations locales de la fonction de densité de probabilité et nuire à la classification. Par contre, s'il est trop grossier, il peut faire disparaître des variations significatives de cette fonction, perturbant également la classification.

Ce choix dépend bien sûr de la dimension de l'espace, du nombre d'observations disponibles mais également de la répartition des observations dans l'espace.

Habituellement, l'ajustement de ce paramètre est laissé à l'initiative de l'analyste qui fait appel à son expérience, sinon à son intuition, pour trouver un compromis entre les deux situations extrêmes évoquées ci-dessus. Le concept de stabilité du nombre de modes détectés apporte un support heuristique pour guider cet ajustement qui reste cependant toujours délicat et primordial.

La procédure de classification proposée consiste à regrouper les hypercubes où la fonction de densité de $f(X)$ est considérée comme convexe afin de localiser les modes de la distribution des observations.

A ce niveau de la procédure des ambiguïtés peuvent exister :

- soit du fait que des hypercubes affectés de l'étiquette "concave" se trouvent situés au sein même des modes à détecter. Ce phénomène résulte en général d'une discrétisation trop fine de l'espace de représentation.

- soit du fait de l'apparition d'hypercubes portant l'étiquette "convexe" dans les vallées séparant les modes ou dans les zones où la convexité de la fonction de densité sous-jacente n'est pas définie. Une des causes principales de ces erreurs d'étiquetage est la présence d'observations isolées éloignées des régions modales.

Il s'avère donc nécessaire de rendre la procédure d'étiquetage des hypercubes plus robuste, afin que les résultats de la classification se révèlent moins sensibles à l'ajustement du pas de discrétisation de l'espace de représentation.

Plutôt que de limiter l'analyse de la convexité de la fonction de densité à l'exploitation directe des résultats locaux du test de convexité en chaque hypercube de l'espace, il semble intéressant de comparer au préalable ce résultat à ceux obtenus sur les hypercubes voisins.

En effet, si le sens de la convexité déterminé au niveau d'un hypercube est compatible avec les sens de la convexité déterminés dans son voisinage, on peut affirmer que l'étiquette assignée à cet hypercube a une forte probabilité d'être correcte. Par contre, si le sens de la convexité au niveau d'un hypercube apparaît comme incompatible avec les sens de la convexité au niveau des hypercubes voisins, il y a une forte probabilité pour que le résultat du test de convexité soit erroné et que l'étiquette affectée à l'hypercube soit à remettre en question.

En déterminant le degré de fiabilité du test de convexité en fonction des résultats de ce test dans le voisinage immédiat de chaque hypercube, on doit pouvoir améliorer la connaissance de la convexité de la fonction de densité de probabilité et obtenir des résultats moins sensibles à de petites irrégularités de la distribution des observations.

Pour améliorer cette détermination de la convexité, on fait appel à une nouvelle procédure itérative basée sur des techniques d'étiquetage probabilistique ou "relaxation" /4-9-16/.

REFERENCES (chapitre II)

- /1/ G. BANON, *Estimation non paramétrique de densité de probabilité pour les processus de Markov*, Thèse d'état, Univ. P. Sabatier, Toulouse, 1° partie, 1977.
- /2/ T. CACOULLOS, *Estimation of a multivariable density*, Ann. Inst. Stat. Math., Vol. 18, pp. 179-189, 1966.
- /3/ T. M. COVER et P. E. HART, *Nearest neighbour pattern classification*, IEEE Trans. Info. Theory, Vol. IT-13, pp 21-27, 1967.
- /4/ A. J. DANKER et A. ROSENFELD, *Blob detection by relaxation*, IEEE Trans. Pattern Anal. Mach. Intell., PAMI-3, pp. 79-92, 1981.
- /5/ R. O. DUDA et P. E. HART, *Pattern classification and scene analysis*, J. WILEY New York, Chap. 2, pp 10-39, 1973.
- /6/ W. GRABLICKI, *Pattern recognition procedure with non parametric density estimates*, IEEE Trans. on Syst. Man & Cyb., Vol. SMC-8, n°11, pp 809-812, 1978.
- /7/ J. KITTLER, *A locally sensitive method for cluster analysis*, Pattern Recog., Vol. 8, pp. 23-33, 1976.
- /8/ W. L. G. KOONTZ et K. FUKUNAGA, *Asymptotic analysis of a nonparametric clustering technic*, IEEE Trans. on Computers, Vol. C-21, n° 9, pp 967-974, 1972.
- /9/ A. LEV, S. W. ZUCKER et A. ROSENFELD, *Iterative enhancement of noisy images*, IEEE Trans. Syst., Man. Cybern., Vol. SMC-7, pp; 435-442, June 1977.
- /10/ D. O. LOFTSGAARDEN et C. P. QUESENBERNY, *A nonparametric estimate of a multivariable density function*, Ann. Math. Stat., Vol. 36, pp 049-1051, 1965.

- /11/ V. K. MURTHY, *Nonparametric estimation of multivariable densities with applications*, Multivariable analysis academic press, New York, pp 43-56, 1966.
- /12/ E. PARZEN, *On estimation of a probability density function and mode*, Ann. Math. Stat., Vol. 33, pp 1005-1076, 1962.
- /13/ J. G. POSTAIRE, *Optimisation du processus de classification automatique par analyse de la convexité des fonctions de densité de probabilité*, Thèse d'état, Université de Lille, 1981.
- /14/ J. G. POSTAIRE et C. VASSEUR, *A fast algorithm for nonparametric probability density estimation*, IEEE Trans. on Pattern analysis and machine Intelligence, Vol. PAMI-4 , n° 6, pp. 63-666, 1982.
- /15/ M. ROSENBLATT, *Remarks on some nonparametric estimates of a density function*, Ann Math. Stat. Ass., Vol 27, pp. 832-837, 1956.
- /16/ A. ROSENFELD, R. A. HUMMEL et S. W. ZUCKER, *Scene labeling by relaxation operations*, IEEE Trans. Syst., Man. Cybern., Vol. SMC-6, pp. 420-433, 1976.
- /17/ J. VAN RYZIN, *On strong consistency of density estimates*, Ann. Math. Stat., Vol. 40, pp 1765-1772, 1969.
- /18/ C. VASSEUR et J. G. POSTAIRE, *Convexité des fonctions de densité. Application à la détection des modes en reconnaissance des formes*, RAIRO Automatique, Vol. 13, n° 2, pp 171-188, 1979
- /19/ C. VASSEUR et J. G. POSTAIRE, *A convexity testing method for cluster analysis* , IEEE Trans. on Syst. Man & Cyb., Vol. SMC-10, pp. 145-149, 1980.

/20/ T. J. WAGNER, *Convergence of the nearest neighbour rule*,
IEEE Trans. on Inf. Theory, Vol. IT-17, pp. 566-
571, 1971.

/21/ E. J. WEGMAN, *Non parametric probability density
estimation. I. A summary of available methods*,
Technometrics, Vol. 14, pp 533-546, 1972.

CHAPITRE III

CONVEXITE ET RELAXATION

III.1 INTRODUCTION.

La description de la fonction de densité de probabilité sous-jacente à la distribution d'un ensemble d'observations en terme de convexité revient à assigner l'étiquette "convexe" à certains points de discrétisation de l'espace de représentation et l'étiquette "concave" à d'autres. Jusqu'à présent, la décision d'affecter à un point l'étiquette convexe ou concave résulte uniquement de l'analyse des propriétés de la fonction de densité de probabilité en ce point. La méthode introduit inévitablement des erreurs d'étiquetage, des points concaves apparaissant isolés dans des régions convexes et réciproquement.

Pour pallier ces faiblesses de la méthode, nous faisons appel aux méthodes d'étiquetage probabiliste, encore appelées, méthodes de "relaxation".

III.2 L'ETIQUETAGE PROBABILISTE OU RELAXATION.

Les méthodes de relaxation se basent sur un étiquetage probabiliste plutôt que sur des décisions immédiates, fermes et définitives. Cet étiquetage est itératif et est réactualisé en fonction des décisions prises à l'itération précédente. Cette mise à jour de l'étiquetage à une itération donnée est effectuée en parallèle, c'est-à-dire en même temps sur tous les objets à étiqueter /4/.

Chaque étiquette correspond en fait à l'une des classes à laquelle on doit affecter chaque objet à étiqueter. On émet alors implicitement l'hypothèse que l'étiquetage d'un objet est d'autant plus satisfaisant et sûr qu'il est cohérent avec l'étiquetage de ses voisins.

Soit $H = \{h_1, \dots, h_i, \dots, h_I\}$ l'ensemble des objets à étiqueter, c'est-à-dire à assigner à l'une des classes de l'ensemble $C = \{c_1, \dots, c_m, \dots, c_M\}$. On associe à chaque objet h_i un vecteur probabilité $[P_i(c_1), \dots, P_i(c_m), \dots, P_i(c_M)]^T$, où $P_i(c_m)$ est l'estimation de la probabilité que h_i appartienne à la classe c_m . Pour tout $h_i \in H$, on a :

M

$$\sum_{m=1} P_i(c_m) = 1$$

m=1

$$\text{et } \emptyset \leq P_i(c_m) \leq 1, \quad m=1,2,\dots,M$$

On commence par assigner à chaque objet h_i des probabilités initiales $P_i^0(c_m)$, $m=[1,\dots,M]$, d'appartenir à chacune des classes c_m , $m=1,\dots,M$, possibles. Puis, on examine les probabilités assignées aux objets voisins de h_i , augmentant $P_i^0(c_m)$ s'il existe de fortes probabilités que les voisins de h_i aient la même classification, ou réduisant $P_i^0(c_m)$ s'il existe de fortes probabilités que ses voisins soient classés différemment. La mise à jour est ensuite effectuée en parallèle pour tous les objets h_i et toutes les probabilités $P_i^0(c_m)$. La procédure est réitérée autant de fois qu'il est nécessaire pour obtenir une stabilisation des probabilités $P_i(c_m)$, $i=1,\dots,I$ et $m=1,\dots,M$.

Il y a deux modèles équivalents pour formaliser cette procédure /5//2/. Nous allons décrire brièvement un de ces modèles.

Pour chaque paire d'objets voisins $\{h_i, h_j\}$ et chaque paire de classes $\{c_m, c_n\}$, on définit une mesure de compatibilité /3/ entre le fait que l'objet h_i appartienne à la classe c_m alors que l'objet h_j appartient à la classe c_n . Cette compatibilité est donnée par la fonction $r_{i,j}(c_m, c_n)$, telle que

$$1) -1 \leq r_{i,j}(c_m, c_n) \leq +1$$

2) Si h_i appartenant à c_m est compatible avec h_j appartenant à c_n alors :

$$r_{i,j}(c_m, c_n) > \emptyset$$

3) Si h_i appartenant à c_m est incompatible avec h_j appartenant à c_n alors :

$$r_{i,j}(c_m, c_n) < \emptyset$$

4) S'il n'existe aucun lien entre ces deux assignations alors :

$$r_{i,j}(c_m, c_n) = \emptyset$$

5) La valeur de $r_{i,j}(c_m, c_n)$ représente le degré de compatibilité entre l'appartenance de h_i à c_m et l'appartenance de h_j à c_n .

La nouvelle estimation de la probabilité pour que l'objet h_i appartienne à la classe c_m à l'itération $(k+1)$ est déterminée en fonction de l'estimation de cette probabilité à l'itération k et de la contribution $Q_i^k(c_m)$ des estimations des probabilités calculées pour les objets voisins /2/ .

On a :

$$P_i^{k+1}(c_m) = \frac{P_i^k(c_m) \cdot (1 + Q_i^k(c_m))}{\sum_{\beta} P_i^k(\beta) \cdot (1 + Q_i^k(\beta))} , \forall c_m \in C$$

$$\text{où } Q_i^k(c_m) = \sum_j C_{i,j} \sum_{\beta} r_{i,j}(c, \beta) \cdot P_i^k(\beta)$$

\sum_j représentant la somme sur les voisins de l'objet h_i

Les coefficients $C_{i,j}$ sont introduits pour pondérer éventuellement l'importance des voisins et assurer que $Q_i^k(c_m)$ reste compris entre les valeurs -1 et +1.

III.3 APPLICATION DES METHODES DE RELAXATION A L'ANALYSE DE LA CONVEXITE.

A ce stade de l'étude, nous sommes en présence d'un ensemble d'hypercubes $H = \{h_1, \dots, h_i, \dots, h_I\}$. I est le nombre d'hypercubes pris en considération, c'est-à-dire contenus dans la liste des hypercubes élaborée précédemment.

Chacun de ces hypercubes appartient soit à la classe C des hypercubes considérés comme localement convexes, soit à la classe nC des hypercubes considérés comme localement concaves, autrement dit non convexes. Pour chaque hypercube h_i , le vecteur probabilité est de la forme $[P_i(C), P_i(nC)]^T$ avec $P_i(C) + P_i(nC) = 1$.

Pour notre problème, le principe de base est de faire croître, de manière itérative, la probabilité qu'un hypercube appartienne à la classe des hypercubes localement convexes s'il existe de fortes probabilités pour que ses voisins appartiennent à cette même classe. Ce cas correspond à un degré élevé de compatibilité entre la nature de l'hypercube considéré et celle de ses voisins. De manière analogue, on fait décroître cette probabilité s'il existe une forte probabilité pour que les hypercubes voisins appartiennent à la classes des hypercubes localement concaves. Ce cas correspond à un degré d'incompatibilité élevé entre la nature de l'hypercube considéré et celle de ses voisins.

La loi itérative de réactualisation des probabilités est définie par :

$$P_i^{k+1}(C) = \frac{P_i^k(C) \cdot (1 + Q_i^k(C))}{\sum_{\beta} P_i^k(\beta) \cdot (1 + Q_i^k(\beta))}$$

$P_i^k(C)$ est donc la probabilité à l'itération k que l'hypercube i appartienne à la classe C.

Les contributions de tous les hypercubes voisins devant, a priori, être prises en compte de la même manière, le terme $Q_i^k(C)$ est défini par :

$$Q_i^k(C) = 1/L \sum_j \sum_{\beta} r_{i,j}(C, \beta) P_j^k(\beta)$$

où $L = 3^N - 1$ est le nombre d'hypercubes voisins de l'hypercube h_i

\sum_j représente la somme sur les voisins de l'hypercube h_i

\sum_{β} représente la somme sur les deux classes possibles (concave et convexe)

$r_{i,j}(C,C)$ et $r_{i,j}(C,nC)$ sont les coefficients de compatibilité à déterminer

Pour pouvoir utiliser cette procédure, il faut commencer par déterminer les probabilités initiales qu'un hypercube h_i soit concave ou convexe, c'est-à-dire :

$$P_i^{\emptyset}(C) \text{ et } P_i^{\emptyset}(nC).$$

Pour cela, on étudie le voisinage de chaque hypercube en se basant sur les résultats du test de convexité de la fonction de densité de probabilité.

III.4 DETERMINATION DES PROBABILITES INITIALES.

D'après les définitions précédentes, on a :

$$P_i^{\emptyset}(C) + P_i^{\emptyset}(nC) = 1.$$

Il suffit donc de déterminer une seule de ces probabilités initiales $P_i^{\emptyset}(C)$ avec, pour seule contrainte :

$$0 \leq P_i^{\emptyset}(C) \leq 1$$

Le calcul de la seconde probabilité initiale $P_i^{\emptyset}(nC)$ est alors immédiat.

Afin de donner une valeur aussi proche de la réalité que possible à cette probabilité, on étudie le voisinage de l'hypercube h_i considéré. Rappelons que pour chacun des hypercubes voisins de h_i figurant dans la liste à traiter par la procédure, on a procédé au test de la convexité de la fonction $f(X)$. On a donc pour chacun d'eux le signe de $\Delta \mathcal{R}_i$ et la valeur effective de $\Delta \mathcal{R}_i$ où $\Delta \mathcal{R}_i = \mathcal{L}(D(X_0, \alpha)) - \mathcal{L}(D(X_0, \alpha'))$.

Posons :

$$P_i(C) = \frac{T' + L}{2L}$$

où L est le nombre d'hypercubes contenus dans la fenêtre d'observation, c'est-à-dire l'hypercube h_i considéré et ses hypercubes voisins pris en compte et où T' est défini par :

$$T' = T + 1 \text{ si } \sum_j \Delta \rho_j > 0$$

$$T' = T - 1 \text{ si } \sum_j \Delta \rho_j < 0$$

T est la différence entre le nombre d'hypercubes voisins localement convexes et le nombre d'hypercubes voisins localement concaves.

La somme \sum_j revient à additionner les différents $\Delta \rho_j$ calculés pour chacun des hypercubes voisins pris en compte.

Le choix de cette détermination des probabilités initiales a été guidé par le souci de donner à ces probabilités des valeurs représentatives du contexte dans lequel se situe l'hypercube traité. En effet, dans ces conditions, plus l'hypercube considéré est entouré d'hypercubes localement convexes, plus la probabilité initiale associée tend vers sa valeur maximale 1 qu'elle n'atteint que lorsque tous les hypercubes voisins sont localement convexes. Dans le cas contraire, cette probabilité tend vers sa valeur minimale 0 qu'elle n'atteint que lorsque tous les hypercubes voisins sont tous localement concaves.

Afin d'utiliser la procédure de relaxation, il faut également déterminer les différents coefficients de compatibilité $r_{i,j}(C,C)$, $r_{i,j}(C,nC)$, $r_{i,j}(nC,C)$ et $r_{i,j}(nC,nC)$ qui interviennent dans les expressions de $Q_i^k(C)$ et de $Q_i^k(nC)$.

III.5 DETERMINATION DES COEFFICIENTS DE COMPATIBILITE.

Les coefficients de compatibilité apparaissant dans l'expression de $Q_i^k(C)$ dépendent essentiellement de la configuration du voisinage de l'hypercube pour lequel ils sont calculés.

Nous les définissons par les tableaux suivants :

Définition de $r_{i,j}(C,C)$:

$r_{i,j}(C,C)$	$P_{j^k}(C)$
+1	> 0.5
-1	< 0.5
\emptyset	$= 0.5$

Le coefficient $r_{i,j}(C,C)$ prend la valeur +1 si l'hypercube h_j voisin de l'hypercube h_i localement convexe a une probabilité d'appartenir à la classe des hypercubes localement convexes supérieure à 0.5 . Dans ce cas de compatibilité, la probabilité $P_{j^k}(C)$ intervient ainsi positivement dans le calcul de la valeur de $Q_i^k(C)$. Par contre, dans le cas où cette probabilité est inférieure à 0.5, ce qui correspond à une incompatibilité, celle-ci intervient négativement dans le calcul de la valeur de $Q_i^k(C)$. Nous remarquons également que dans le cas où cette probabilité est égale à 0.5 , c'est-à-dire dans le cas où on ne peut conclure quant à son appartenance à l'une des classes "localement convexe" ou "localement concave" , elle n'intervient pas dans l'expression de $Q_i^k(C)$.

Définition de $r_{i,j}(C,nC)$:

$r_{i,j}(C,nC)$	$P_{j^k}(nC)$
-1	> 0.5
+1	< 0.5
\emptyset	$= 0.5$

Nous pouvons constater également dans ce tableau que si l'hypercube h_j voisin a une probabilité d'appartenir à la classe des hypercubes localement concaves supérieure à 0.5, celle-ci intervient négativement dans l'expression de $Q_i^k(C)$. Rappelons que nous considérons le voisinage d'un hypercube afin d'ajuster sa probabilité d'appartenir à la classe des hypercubes localement convexes. Ainsi, un voisin ayant une tendance à appartenir plutôt à l'autre classe ne peut que tendre à diminuer cette probabilité, ce qui explique qu'il intervient négativement dans l'expression de $Q_i^k(C)$.

Les coefficients $r_{i,j}(C,C)$ et $r_{i,j}(C,nC)$ interviennent dans l'expression de $Q_i^k(C)$.

Définition de $r_{i,j}(nC,C)$ et de $r_{i,j}(nC,nC)$:

$r_{i,j}(nC,C)$	$P_{j^k}(C)$
-1	> 0.5
+1	< 0.5
\emptyset	$= 0.5$

$r_{i,j}(nC,nC)$	$P_{j^k}(nC)$
+1	> 0.5
-1	< 0.5
\emptyset	$= 0.5$

Les coefficients $r_{i,j}(nC,C)$ et $r_{i,j}(nC,nC)$ interviennent dans l'expression de $Q_{i^k}(nC)$.

Si l'hypercube h_j voisin de l'hypercube h_i a une probabilité d'appartenir à la classe des hypercubes localement concaves supérieure à 0.5 pour $r_{i,j}(nC,nC)$ - ou une probabilité d'appartenir à la classe des hypercubes localement convexes inférieure à 0.5 pour $r_{i,j}(nC,C)$ - , alors la probabilité $P_{i^k}(nC)$ pour $r_{i,j}(nC,nC)$ - ou $P_{i^k}(C)$ pour $r_{i,j}(nC,C)$ - intervient positivement dans la valeur de $Q_{i^k}(nC)$.

Par contre, si cet hypercube h_j a une probabilité d'appartenir à la classe des hypercubes localement concaves inférieure à 0.5 pour $r_{i,j}(nC,nC)$ - ou une probabilité d'appartenir à la classe des hypercubes localement convexes supérieure à 0.5 pour $r_{i,j}(nC,C)$ - , alors elle intervient négativement dans la valeur de $Q_{i^k}(nC)$.

Enfin pour ces deux derniers tableaux, lorsque cette probabilité est égale à 0.5 , $P_{i^k}(nC)$ pour $r_{i,j}(nC,nC)$ - ou $P_{i^k}(C)$ pour $r_{i,j}(nC,C)$ - elle n'intervient pas dans la valeur de $Q_{i^k}(nC)$.

En examinant ces tableaux, nous constatons que :

$$\left. \begin{array}{l} r_{i,j}(C,C) = r_{i,j}(C,nC) \\ r_{i,j}(nC,C) = r_{i,j}(nC,nC) \end{array} \right\} \quad \forall i,j$$

ce qui permet de simplifier l'expression de $Q_{i^k}(C)$.

On a :

$$\begin{aligned}
 Q_i^k(C) &= 1/L \sum_j \sum_{\beta} r_{i,j}(C,\beta) P_j^k(\beta) \\
 &= 1/L \sum_j [r_{i,j}(C,C) P_j^k(C) + r_{i,j}(C,nC) P_j^k(nC)] \\
 &= 1/L \sum_j [P_j^k(C) \{r_{i,j}(C,C) - r_{i,j}(C,nC)\} + r_{i,j}(C,nC)]
 \end{aligned}$$

Comme $r_{i,j}(C,C) = r_{i,j}(C,nC)$, on obtient :

$$Q_i^k(C) = 1/L \sum_j r_{i,j}(C,nC)$$

Nous pouvons calculer, de la même manière, l'expression de $Q_i^k(nC)$:

$$\begin{aligned}
 Q_i^k(nC) &= 1/L \sum_j \sum_{\beta} r_{i,j}(nC,\beta) P_j^k(\beta) \\
 &= 1/L \sum_j r_{i,j}(nC,nC)
 \end{aligned}$$

Comme $r_{i,j}(nC,nC) = - r_{i,j}(C,nC)$, on conclut que :

$$Q_i^k(C) = - Q_i^k(nC)$$

ce qui simplifie grandement les calculs.

D'autre part, on remarque que les coefficients de compatibilité ne prennent que les valeurs +1 et -1. Nous constatons donc que pour déterminer la valeur de l'expression de $Q_i^k(C)$, il suffit de calculer la différence :

$$Q_i^k(C) = 1/L \cdot (NVS - NVI)$$

où NVS est le nombre de voisins pour lesquels $P_j^k(C) > 0.5$ et NVI est le nombre de voisins pour lesquels $P_j^k(C) < 0.5$

Les coefficients de compatibilité proposés nous permettent d'analyser le comportement de l'algorithme de relaxation.

III.6 ANALYSE DU COMPORTEMENT DE L'ALGORITHME DE RELAXATION.

La formulation particulière de cet algorithme de relaxation nous permet d'étudier l'évolution de la probabilité pour qu'un hypercube h_i appartienne à la classe des hypercubes localement convexes en fonction de son environnement.

L'expression :

$$P_i^{k+1}(C) = \frac{P_i^k(C) \cdot [1 + Q_i^k(C)]}{\sum_{\beta} P_i^k(\beta) \cdot [1 + Q_i^k(\beta)]}$$

devient :

$$P_i^{k+1}(C) = \frac{P_i^k(C) \cdot [1 + Q_i^k(C)]}{P_i^k(C) \cdot [1 + Q_i^k(C)] + P_i^k(nC) \cdot [1 + Q_i^k(nC)]}$$

Comme $P_i^k(C) = 1 - P_i^k(nC)$

$$P_i^{k+1}(C) = \frac{P_i^k(C) \cdot [1 + Q_i^k(C)]}{P_i^k(C) \cdot [Q_i^k(C) - Q_i^k(nC)] + 1 + Q_i^k(nC)}$$

En tenant compte du fait que $Q_i^k(C) = -Q_i^k(nC)$, on a :

$$P_i^{k+1}(C) = \frac{P_i^k(C) \cdot [1 + Q_i^k(C)]}{Q_i^k(C) \cdot [2P_i^k(C) - 1] + 1}$$

Pour étudier cette équation en fonction du terme $Q_i^k(C)$, examinons en premier lieu le domaine de définition de $P_i^{k+1}(C)$.

La fonction n'est pas définie si :

$$Q_i^k(C) \cdot [2P_i^k(C) - 1] + 1 = 0$$

Nous remarquons que pour $P_i^k(C) = 0.5$, cette équation ne peut être nulle, et ce, quelle que soit la valeur de $Q_i^k(C)$. Cette remarque nous amène à poser que si $P_i^k(C)$ est différent de 0.5, alors :

$$Q_i^k(C) \cdot [2P_i^k(C) - 1] + 1 = 0 \text{ si et seulement si :}$$

$$Q_i^k(C) = \frac{1}{1 - 2P_i^k(C)}$$

$$\text{Comme on a : } 0 \leq P_i^k(C) \leq 1$$

$$-1 \leq 1 - 2P_i^k(C) \leq +1$$

$$\text{alors : } \left| \frac{1}{1 - 2P_i^k(C)} \right| \geq 1$$

Or $Q_i^k(C)$ est compris entre -1 et +1 inclus. On en déduit que les seuls cas où la fonction $P_i^{k+1}(C)$ n'est pas définie sont :

$$- \text{ soit } Q_i^k(C) = +1, \text{ alors que } P_i^k(C) = 0$$

$$- \text{ soit } Q_i^k(C) = -1, \text{ alors que } P_i^k(C) = 1$$

Ces deux possibilités sont des cas extrêmes et rares. En effet, pour le premier cas, cela revient à trouver un hypercube ayant une probabilité nulle d'appartenir à la classe des hypercubes convexes alors qu'il est entouré d'hypercubes voisins ayant tous une probabilité supérieure à 0.5 d'appartenir à la classe des hypercubes convexes. Cette configuration est représentée par la figure ci-dessous pour un espace à deux dimensions.

> 0.5	> 0.5	> 0.5
> 0.5	0	> 0.5
> 0.5	> 0.5	> 0.5

Le second cas correspond, de manière évidente, à la configuration suivante :

< 0.5	< 0.5	< 0.5
< 0.5	1	< 0.5
< 0.5	< 0.5	< 0.5

Ces deux cas, peu fréquents, mais néanmoins possibles, nous amènent à poser les conditions :

$$\left. \begin{array}{l} \text{si } P_i^k(C) = 0 \text{ alors } P_i^{k+1} = 0 \\ \text{si } P_i^k(C) = 1 \text{ alors } P_i^{k+1} = 1 \end{array} \right\} \text{ quel que soit } Q_i^k(C)$$

Nous pouvons calculer ensuite la dérivée première de la fonction $P_i^{k+1}(C)$ par rapport à $Q_i^k(C)$.

$$\frac{d P_i^{k+1}(C)}{d Q_i^k(C)} = \frac{P_i^k(C) \{ Q_i^k(C) [2P_i^k(C) - 1] + 1 \} - [2P_i^k(C) - 1] P_i^k(C) [1 + Q_i^k(C)]}{\{ Q_i^k(C) [2P_i^k(C) - 1] + 1 \}^2}$$

$$= \frac{2P_i^k(C) \cdot [1 - P_i^k(C)]}{\{ Q_i^k(C) \cdot [2P_i^k(C) - 1] + 1 \}^2}$$

Cette dérivée permet de déterminer le sens de l'évolution des probabilités dans le schéma itératif. On constate d'abord que lorsque la probabilité $P_i^k(C)$ atteint les valeurs \emptyset ou 1 , elle demeure constante.

Dans ce cas, $P_i^{k+1}(C) = P_i^k(C)$, quel que soit le voisinage. Sinon la probabilité $P_i^{k+1}(C)$ demeure comprise entre \emptyset et 1 .

On a le tableau de variation suivant :

$Q_i^k(C)$	-1	\emptyset	+1
$\frac{d P_i^{k+1}(C)}{d Q_i^k(C)}$		+	
$P_i^{k+1}(C)$		$P_i^k(C)$	+1

Nous constatons que si l'expression de $Q_i^k(C)$ est nulle, c'est-à-dire qu'il y a autant d'hypercubes voisins dont la probabilité $P_j^k(C)$ est supérieure à $\emptyset.5$ que d'hypercubes voisins dont la probabilité $P_j^k(C)$ est inférieure à $\emptyset.5$, alors la probabilité à l'itération $k+1$ pour que l'hypercube considéré appartienne à la classe des hypercubes localement convexes reste inchangée vis à vis de cette probabilité à l'itération k . Nous remarquons que si la valeur de $Q_i^k(C)$ est positive, la probabilité $P_i^{k+1}(C)$ est supérieure à $P_i^k(C)$ et inversement que si la valeur de $Q_i^k(C)$ est négative, $P_i^{k+1}(C)$ est inférieure à $P_i^k(C)$.

Cette étude analytique permet de justifier, à posteriori, le choix des coefficients de compatibilité puisque l'algorithme de relaxation fait augmenter la probabilité d'appartenance à une classe s'il y a compatibilité avec l'environnement et la fait diminuer s'il y a incompatibilité.

Les méthodes de relaxation étant convergentes /6//1/, nous pourrions répéter le processus d'itérations jusqu'à obtenir un ensemble d'hypercubes dont les probabilités d'appartenir à l'une ou l'autre des classes seraient stables au-delà d'un certain nombre d'itérations. Mais ceci entraînerait des temps d'exécution prohibitifs. Il faut donc trouver un critère adéquat pour arrêter le processus itératif tout en tolérant un certain écart par rapport au résultat correspondant à la convergence complète. Expérimentalement, on constate, en calculant à chaque itération le nombre d'hypercubes dont la probabilité d'appartenir à la classe des hypercubes localement convexes est égale à un, que dès que ce nombre reste inchangé entre deux itérations successives, on peut arrêter la procédure itérative. On obtient ainsi un bon étiquetage probabiliste qui permet de décrire la fonction de densité sous-jacente en termes de convexité de manière satisfaisante.

Nous sommes ainsi en présence d'un ensemble d'hypercubes caractérisés par leurs probabilités d'être localement convexes : $P_i^k(C)$. La recherche de régions constituées d'hypercubes à forte probabilité d'être localement convexes permettra de détecter les différents modes de la fonction de densité de probabilité.

REFERENCES (chapitre III)

- /1/ EKLUNTH et A. ROSENFELD, *Convergence properties of relaxation*, Comput. Sci. Center Univ. Maryland, Rep. TR-701, Oct. 1978.
- /2/ PELEG, *A new probabilistic relaxation scheme*, IEEE Trans. Pattern Anal. Machine Intell., Vol. PAMI-2, pp 362-369, 1980.
- /3/ PELEG and A. ROSENFELD, *Determining compatibility coefficients for curve enhancement relaxation processes*, IEEE Trans. Syst. Man Cybern., Vol. SMC-8, pp 548-555, 1978.
- /4/ A. ROSENFELD, *Iterative methods in image analysis*, Proc. IEEE Conf. Pattern Recognition Image Processing, New York, pp 14-18, June 1977.
- /5/ A. ROSENFELD, R. A. HUMMEL et S. W. ZUCKER, *Scene labeling by relaxation operations*, IEEE Trans. Syst., Man. Cybern., Vol. SMC-6, pp. 420-433, 1976.
- /6/ S. W. ZUCKER, KRISHNAMURTHY and HAAR, *Relaxation processes for scene labeling convergence, speed and stability*, IEEE Trans. Syst. Man Cybern., Vol. SMC-8, pp 41-48, 1978.

CHAPITRE IV

REGONSTITUTION DES MODES ET GLASSIFICATION

IV.1 INTRODUCTION.

A la fin de la procédure d'étiquetage probabiliste, la seule information dont on dispose pour effectuer la classification des données, est constituée par le réseau des hypercubes de discrétisation, avec pour chacun d'eux une probabilité d'appartenir à la classe des hypercubes convexes. Il s'agit alors de mettre en évidence des ensembles d'hypercubes à forte probabilité d'être convexes. Sous l'hypothèse non paramétrique, ces ensembles matérialiseront les domaines de convexité ou domaines modaux de la fonction de densité sous-jacente à la distribution des observations à classer, c'est-à-dire les modes autour desquels on regroupera les observations par classe /6/. Sous l'hypothèse paramétrique normale, ces domaines modaux pourront être modélisés par des domaines hyperellipsoïdaux à partir desquels il sera possible d'élaborer une classification optimale /5/.

IV.2 DETERMINATION DES DOMAINES MODAUX INITIAUX.

Afin d'exposer la procédure permettant de détecter les modes des fonctions de densité, nous allons illustrer les différentes étapes de l'estimation de la fonction de densité en prenant pour exemple une distribution bidimensionnelle quelconque constituée de deux modes (figure 1) :

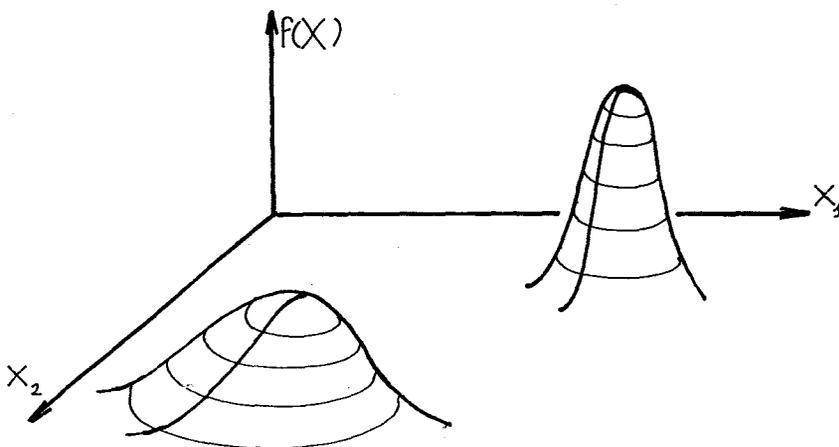


Figure 1 : Représentation graphiques des observations de l'exemple 1.

Après avoir appliqué aux données de l'espace de représentation (X_1, X_2) les différentes procédures exposées au chapitre II, on obtient le réseau hypercubique de la figure 2 :

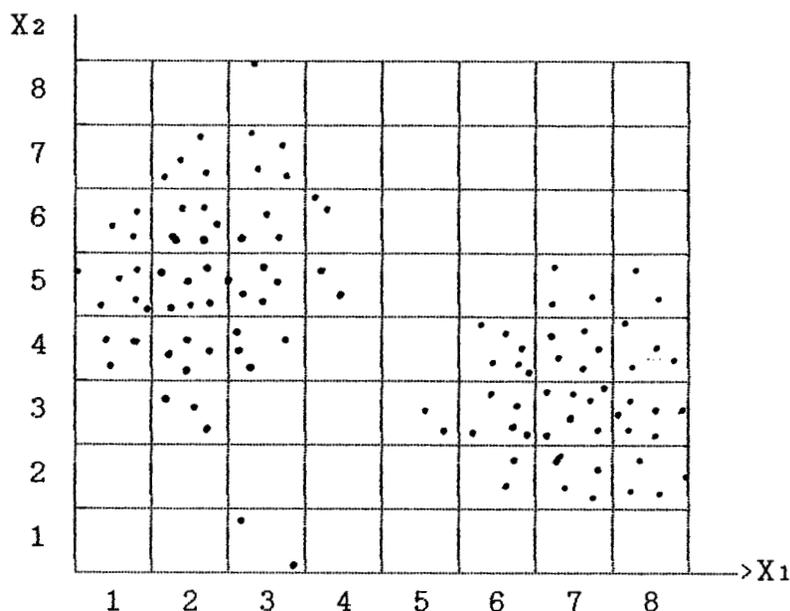


Figure 2 : Représentation des observations de la figure 1 après transformation de l'espace de représentation.

La procédure d'étiquetage probabiliste permet d'associer à chaque hypercube constituant le réseau une probabilité d'être convexe. Par exemple, sur la figure 3, l'hypercube de coordonnées (1,4) a une probabilité d'être convexe égale à 0,6.

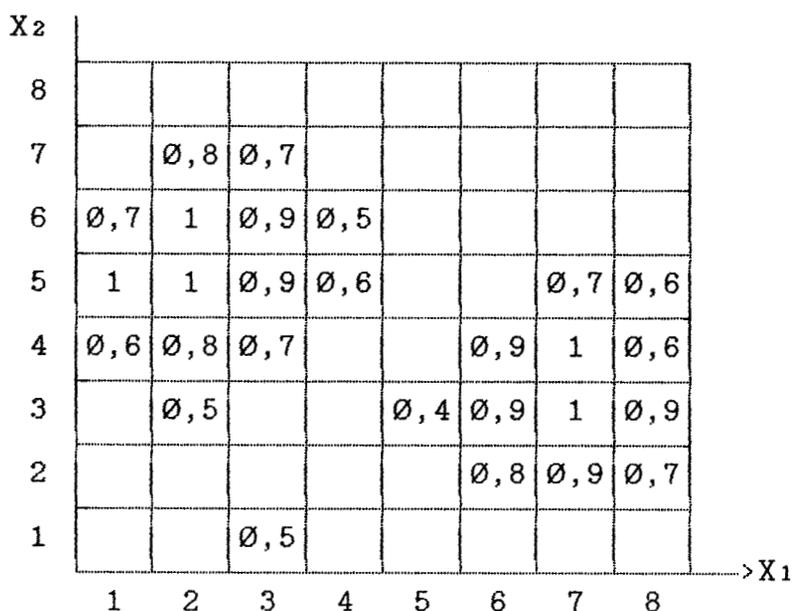
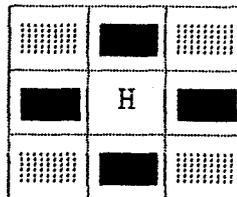


Figure 3 : Etiquetage probabiliste des hypercubes de la figure 1.

Afin d'initialiser la procédure d'agrégation qui permettra de mettre en évidence les domaines modaux, on

recherche en premier lieu un hypercube dont la probabilité d'être convexe soit supérieure à un certain seuil. Le seuil associé à ce critère est habituellement fixé à 0.9, afin de ne sélectionner que des hypercubes à très forte probabilité d'être convexe. Dans notre exemple, cet hypercube est l'hypercube de coordonnées (2,5).

On examine ensuite le voisinage direct de l'hypercube sélectionné, c'est-à-dire qu'on recherche parmi les hypercubes ayant soit un côté, soit un sommet en commun avec cet hypercube, ceux pour lesquels la probabilité d'être convexes dépasse le seuil et qui n'ont pas encore été sélectionnés.



Dans l'espace à deux dimensions, le voisinage direct d'un hypercube H est composé :

- d'hypercubes ayant un côté en commun avec l'hypercube

considéré  ,

- et d'hypercubes ayant un sommet en commun avec

l'hypercube considéré  .

L'étude de ce voisinage portera, dans un espace à N dimensions, au maximum sur $(3^N - 1)$ hypercubes. Dans l'exemple bidimensionnel ci-dessus, le voisinage de l'hypercube de coordonnées (2,5) est constitué des 8 hypercubes de coordonnées (1,6), (2,6), (3,6), (1,5), (3,5), (1,4), (2,4), (3,4). On constate que parmi ces différents hypercubes, ceux de coordonnées (1,6), (1,4), (2,4) et (3,4) ne répondent pas au critère de seuil et ne sont donc pas sélectionnés. Pour chacun des hypercubes restants ainsi sélectionnés, on étudie de la même façon leur voisinage direct afin de constituer une procédure de chaînage. On recommence ces différentes opérations jusqu'au moment où l'on ne détecte plus d'hypercubes voisins dont la probabilité associée soit

supérieure ou égale au seuil. Ce premier ensemble d'hypercubes ainsi déterminé matérialise le premier "domaine modal" de la fonction de densité.

Dans notre exemple, le premier domaine modal est représenté par les hypercubes hachurés de la figure 4.

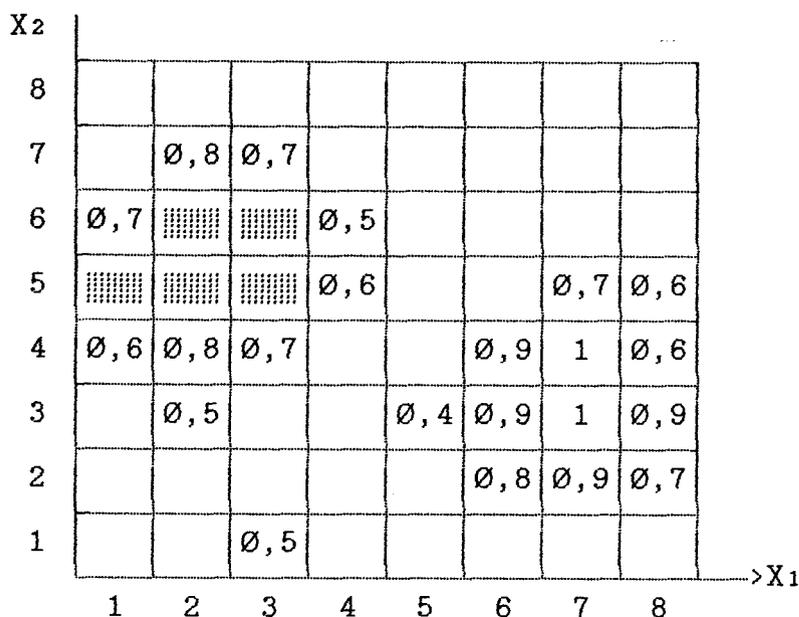


Figure 4 : Détection du premier domaine modal de l'exemple de la figure 1.

On examine ensuite, de la même manière, les hypercubes répondant au même critère et ne faisant pas encore partie d'un "domaine modal", afin de mettre en évidence d'autres domaines modaux. Ainsi, en réitérant cette procédure d'agrégation, on détermine les domaines modaux initiaux associés aux classes en présence. Cette procédure peut néanmoins générer des domaines non significatifs qu'il s'agit de rejeter en contrôlant le nombre d'observations contenues dans les domaines obtenus. Si ce nombre est trop faible, le mode concerné est considéré comme non représentatif et donc ignoré (dans le chapitre 5, ce nombre sera pris supérieur à 10% du nombre total d'observations à traiter). Pour notre exemple, on détecte deux domaines modaux initiaux D1 et D2 (figure 5) qui ont alors besoin d'être affinés.

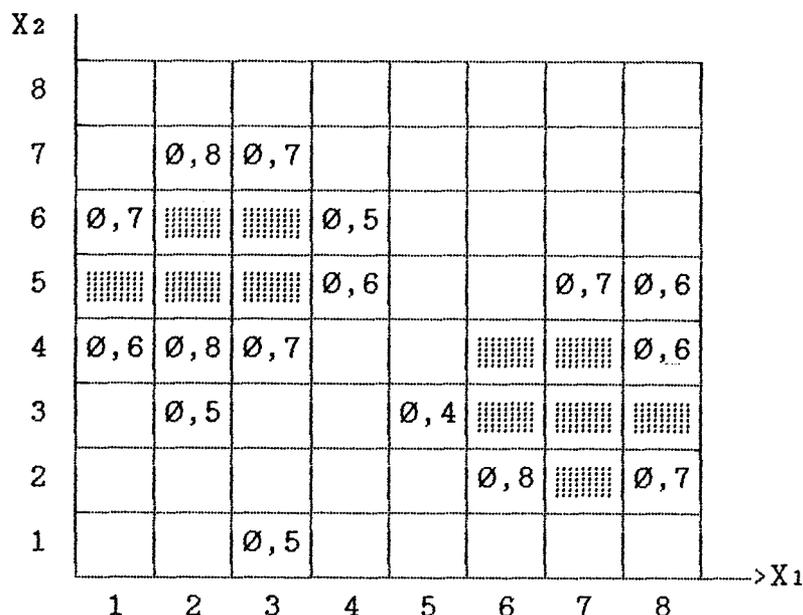


Figure 5 : Détection du deuxième domaine modal de l'exemple de la figure 1.

IV.3 EXTENSIONS SEQUENTIELLES DES DOMAINES MODAUX INITIAUX.

Connaissant le contenu des hypercubes constituant les domaines associés aux différents modes ainsi détectés, on obtient directement la liste des observations situées dans chacun d'eux.

Pour pouvoir assigner les observations contenues dans les hypercubes situés à l'extérieur des domaines modaux associés à chacune des classes mises en évidence, nous devons utiliser une procédure de classement.

Celle-ci peut être implantée sous la forme de fonctions de décision qui indiqueront à quelles classes attribuer les observations.

Le choix de cette procédure de classement dépend essentiellement du type de données à traiter. En effet, nous n'utiliserons pas les mêmes fonctions de décisions lorsque la distribution des observations correspond à un mélange gaussien que lorsque l'on n'émet aucune hypothèse de type paramétrique.

Néanmoins, quelle que soit la procédure de classement choisie, on procède de manière à ce que l'étude des hypercubes non sélectionnés s'effectue dans un ordre précis.

En effet, nous testons en premier lieu le voisinage direct des domaines modaux, c'est-à-dire que l'on prend en compte les hypercubes non sélectionnés ayant un côté ou un sommet en commun avec au moins un des hypercubes constituant le domaine modal examiné. En appliquant ce principe de recherche, on ne sélectionne que les hypercubes formant en quelque sorte une "hyperenveloppe" autour du domaine modal. Dans l'exemple, si on considère le premier domaine modal détecté, cette hyperenveloppe est constituée des hypercubes de coordonnées (1,4), (2,4), (3,4), (4,5), (1,6), (4,6), (2,7) et (3,7).

Cette recherche ne s'effectuant que sur le voisinage direct du domaine modal, l'hypercube de coordonnées (2,3) n'est momentanément pas pris en compte.

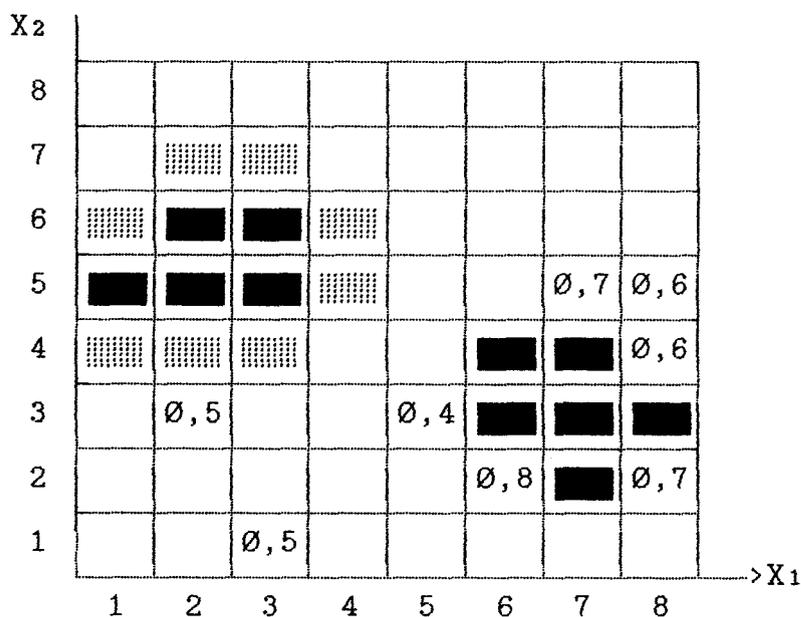


Figure 6 : Détermination de l'hyperenveloppe du premier domaine modal de l'exemple de la figure 1.

De plus, il importe d'établir un ordre de prise en compte de ces domaines modaux. En effet, en fonction du nombre d'observations contenues dans chacun de ces domaines modaux et du nombre total d'observations situées dans l'ensemble des domaines détectés, il est aisé de déduire, pour chaque mode en présence, une approximation de son poids dans l'échantillon analysé. Par exemple, si après la détection de domaines modaux, on constate qu'il y a k observations au total

dans ces domaines dont k_1 sont situées dans le premier domaine, k_2 dans le second, ... , k_n dans le dernier, on peut considérer que k_1/k est le poids du premier mode, k_2/k celui du second, ... , k_n/k celui du dernier. La somme de ces poids est bien égale à 1.

L'expérience montre que l'on a intérêt à initialiser la procédure de constitution des hyperenveloppes en commençant par le mode de poids le plus faible, ceci afin d'éviter que les classes avoisinantes ayant une importance plus élevée ne viennent englober les observations situées dans le voisinage proche d'une classe de poids plus faible.

Il importe de remarquer que toutes les observations contenues dans les hypercubes ainsi rattachés à une hyperenveloppe n'appartiennent pas nécessairement à la même classe. En effet, pour assigner les observations contenues dans l'hyperenveloppe, on utilise les fonctions de décisions associées à la règle de classement choisie.

Dès qu'une observation est ainsi assignée à une classe, les fonctions de décisions sont aussitôt réactualisées. En effet, cette nouvelle observation modifie les caractéristiques de la classe à laquelle elle est assignée et il importe de tenir compte de cette évolution avant de poursuivre le classement des observations restantes.

Ainsi, au cours de la constitution des hyperenveloppes, les modes évoluent au fur et à mesure vers leur forme finale.

La figure 7 indique les deux hyperenveloppes correspondant aux deux modes de l'exemple de la figure 1.

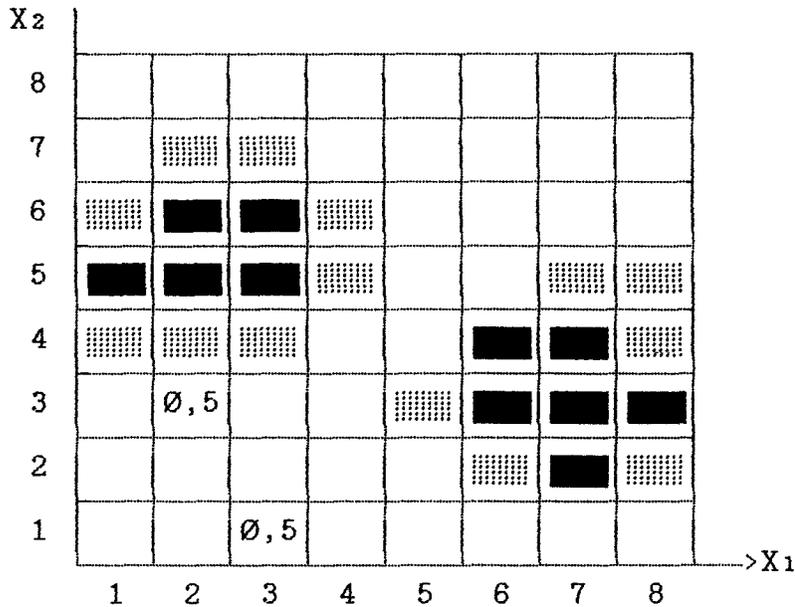


Figure 7 : Détermination de l'hyperenveloppe du deuxième domaine modal de l'exemple de la figure 1.

Les domaines constitués d'hypercubes représentés par



ou par



, c'est-à dire constitués des domaines

modaux initiaux et de leurs hyperenveloppes, peuvent être considérés comme les nouveaux domaines modaux. On applique alors ces différentes opérations de manière itérative jusqu'à ce que les domaines modaux aient atteint leur taille maximale, la procédure s'arrêtant d'elle-même lorsqu'il n'y a plus de voisins immédiats disponibles.

Ainsi dans notre exemple, c'est en examinant le voisinage direct de ces nouveaux domaines modaux que l'on prendra en compte l'hypercube de coordonnées (2,3). La figure 8 indique les 2 domaines modaux D1' et D2' ainsi obtenus.

Le soin apporté à cette procédure séquentielle d'extension des domaines modaux permet d'éviter des fusionnements ou des disparitions intempestifs de modes. Les domaines modaux étendus ainsi obtenus peuvent alors être utilisés pour classer les observations non encore prises en compte.

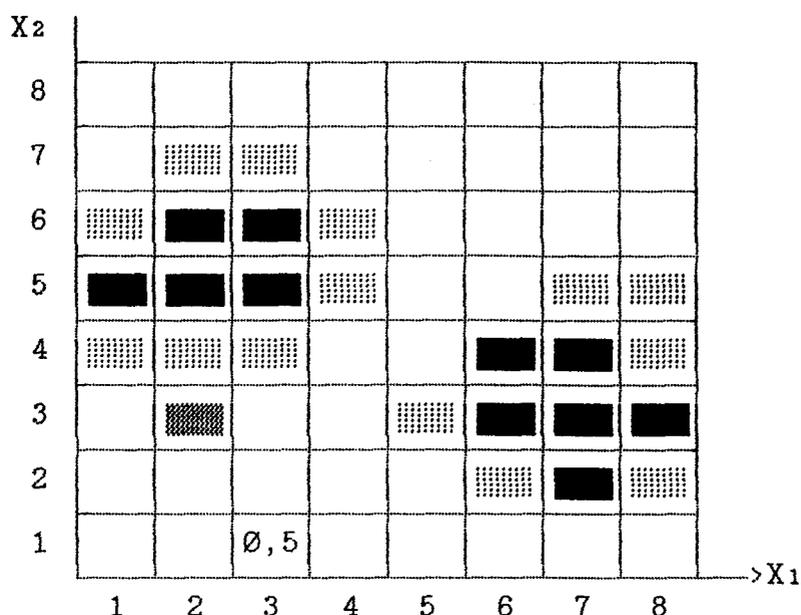


Figure 8 : Domaines modaux étendus $D1'$ et $D2'$.

IV.4 CLASSIFICATION GLOBALE

Afin d'achever la classification de toutes les observations, il faut considérer les hypercubes non encore pris en compte, c'est-à-dire les hypercubes isolés qui ne font partie d'aucun des domaines modaux étendus mis en évidence.

Dans notre exemple, cette ultime recherche d'hypercubes prendra donc en compte l'hypercube de coordonnées (3,1).

De la même manière, chacune des observations contenues dans ces hypercubes est directement assignée à une classe par le jeu des fonctions de décision qui sont réactualisées immédiatement.

Après cette première étude du réseau des hypercubes, on examine séparément les ensembles d'observations constituant chacune des classes détectées. Cette précaution permet de vérifier que les classes mises en évidence par cette première étude globale ne sont pas constituées, en fait, par plusieurs classes qui, à la suite d'une discrétisation trop grossière de l'espace, se seraient fusionnées.

IV.5 PROCEDURE D'ANALYSE LOCALE.

On sait que le choix d'un pas de discrétisation spatiale trop important peut conduire à masquer la présence de plusieurs classes dans certaines régions de l'espace de représentation des données. Pour s'assurer qu'un mode n'est passé inaperçu par suite d'un mauvais ajustement de ce pas, réputé délicat, on procède à une analyse locale de chacun des groupements mis en évidence. Pour ce faire, on utilise un pas de discrétisation plus fin, ce qui revient à examiner "à la loupe" les groupements mis en évidence par la procédure globale.

Pour chaque groupement obtenu par la procédure globale, cette procédure d'analyse locale est en tout point analogue à celle développée pour analyser l'ensemble des données. Si de nouveaux groupements sont détectés sous la forme de nouveaux domaines modaux, il faut alors reprendre intégralement les procédures d'extension séquentielle de ces domaines modaux initiaux pour assigner les observations aux nouvelles classes.

Cette procédure d'analyse locale a été systématiquement appliquée à tous les exemples présentés au chapitre suivant. Chaque groupement obtenu par l'analyse globale a été alors analysé, après normalisation, avec le même pas que celui utilisé pour l'ensemble des observations afin d'accroître substantiellement la finesse de l'analyse.

Il va sans dire que cette procédure locale est sans objet lorsque l'analyse globale ne permet de distinguer qu'une seule classe!

A la suite de cette procédure locale, nous disposons d'un nouvel ensemble de domaines modaux. Il s'agit alors de déterminer leur hyperenveloppe selon une procédure en tout point analogue à celle développée lors de l'analyse globale.

Les observations sont assignées aux différentes classes de manière séquentielle, les fonctions de décision évoluant ainsi au cours de la procédure.

Quant au choix de la procédure de classement à employer, nous avons l'embarras du choix, tant la littérature sur ce problème est riche et dense /1/. Pour pouvoir expérimenter notre procédure de classification, nous nous sommes placés sous l'hypothèse paramétrique normale qui nous a permis d'utiliser la procédure de classement bayésien, c'est-à-dire, la procédure de classement optimal.

IV.6 CLASSEMENT BAYESIEN.

Considérons un échantillon dont les observations proviennent de K classes gaussiennes M_k , $k=1,2, \dots, K$. La distribution au sein de chacune de ces classes est caractérisée par une fonction de densité conditionnelle normale :

$$p(X/M_k) = \frac{1}{(2)^{n/2} |\Sigma_k|^{1/2}} \exp\{-1/2(X-\bar{X}_k)^T \Sigma_k^{-1} (X-\bar{X}_k)\}$$

où \bar{X}_k est le vecteur-moyenne de la classe M_k
 et Σ_k est la matrice de covariance de la classe M_k

La densité de probabilité des observations pour des individus tirés aléatoirement et indépendamment de l'ensemble des classes M_k est un mélange des lois de probabilité relatives à ces différentes classes. La fonction de densité de ce mélange est de la forme :

$$f(X) = \sum_{k=1}^K p(X/M_k) \cdot P(M_k)$$

où $P(M_k)$ est la probabilité à priori de la classe M_k de telle sorte que :

$$\sum_{k=1}^K P(M_k) = 1$$

Lorsque le modèle probabiliste du mélange est connu, on peut calculer la probabilité pour qu'une observation X appartienne à la classe M_k . En appliquant le théorème de Bayés /3/, cette

probabilité, appelée probabilité à posteriori de la classe M_k , est donnée par la relation :

$$P(M_k/X) = \frac{P(X/M_k) \cdot P(M_k)}{f(X)}$$

Le risque d'erreur lié au classement de l'observation X dans la classe M_k est alors :

$$R(M_k/X) = 1 - P(M_k/X)$$

Pour obtenir un classement optimal, nous devons affecter une observation X à la classe M_k pour laquelle le risque d'erreur $R(M_k/X)$ est minimum.

Le problème du classement optimal revient à déterminer un ensemble de fonctions de décision $g_k(X)$, $k=1,2, \dots, K$, telles que l'observation X est attribuée à la classe M_k si et seulement si :

$$g_k(X) \geq g_i(X) ; i=1,2, \dots, K ; i \neq k$$

Nous choisissons les fonctions de décisions qui optimisent la procédure de classement de la forme /2/ :

$$g_k(X) = P(X/M_k)P(M_k)$$

$P(X/M_k)$ étant de type exponentielle, nous utilisons le logarithme de ces fonctions de décisions :

$$\begin{aligned} g_k'(X) &= \log g_k(X) = \log[P(X/M_k) \cdot P(M_k)] \\ &= -n/2 \log 2 - 1/2 \log |\Sigma_k| - 1/2 (X - \bar{X}_k)^T \Sigma_k^{-1} (X - \bar{X}_k) + \log P(M_k) \end{aligned}$$

Après suppression du terme constant $n/2 \log(2)$, les fonctions de décisions $g_k'(X)$ se ramènent aux expressions suivantes :

$$g_k'(X) = - 1/2 \log|\Sigma_k| - 1/2(X-\bar{X}_k)^T \Sigma_k^{-1} (X-\bar{X}_k) + \log P(M_k)$$

$k=1,2, \dots ,K$

Il s'agit alors de remplacer dans l'expression de $g_k'(X)$ les valeurs du vecteur-moyenne \bar{X}_k , de la matrice de covariance Σ_k et de la probabilité à priori de la classe M_k par des valeurs approchées obtenues à partir de la procédure d'analyse des données.

a) Approximation du vecteur moyenne.

Une approximation du vecteur-moyenne \bar{X}_k est obtenue à partir des observations situées dans le domaine modal D_k associé à la classe C_k sous la forme :

$$\bar{X}_k = \frac{1}{Q_k} \sum_{X_q \in M_k} X_q$$

où Q_k est le nombre d'observations situées dans le domaine D_k

Cette approximation sera actualisée lors de l'extension des domaines modaux.

b) Approximation de la matrice de covariance.

On peut montrer qu'une fonction de densité normale est convexe à l'intérieur de son domaine modal qui peut être modélisé par un hyperellipsoïde d'équation (voir annexe 1) /5/ :

$$(X-\bar{X}_k)^T \Sigma_k^{-1} (X-\bar{X}_k) = 1$$

et que sa convexité n'est pas définie à l'extérieur de celui-ci /5/.

Les vecteurs propres de l'inverse de la matrice de covariance Σ_k ont la direction des axes principaux de cet hyperellipsoïde. De plus, ces valeurs propres sont égales aux demi-longueurs de ces axes (voir figure 9). Ces propriétés peuvent être utilisées pour déterminer la matrice de covariance Σ_k à partir d'une analyse géométrique du domaine D_k présentée en annexe.

En fait, dans le cas d'indépendance statistique des variables, nous utiliserons une version simplifiée de cette procédure. Une simple estimation de la variance des composantes, par la méthode du maximum de ressemblance, portant sur les observations situées à l'intérieur du domaine modal constitue, dans ce cas, une approximation initiale satisfaisante de la dispersion de la classe correspondante.

L'approximation de la matrice de covariance ainsi obtenue peut alors être actualisée de manière séquentielle lors de l'extension des domaines modaux. En effet, toute adjonction des nouveaux éléments à la classe considérée se traduit par une modification de la matrice de covariance qui peut être réactualisée par un algorithme itératif /4/.

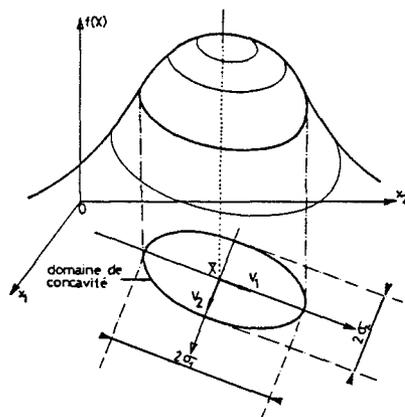


Figure 9 : Propriétés géométriques du "domaine de concavité" de la fonction de densité d'une distribution normale à deux dimensions.

c) Détermination des probabilités à priori des classes M_k .

On peut montrer que la probabilité pour qu'une observation d'un échantillon distribué selon la fonction $f(X)$ provienne de la classe M_k et appartienne en même temps au

domaine modal D_k de la fonction de densité $P(X/M_k)$ est proportionnel à $P(M_k) / 5/$.

Soit Q_k le nombre d'observations appartenant au domaine modal D_k , le rapport Q_k/Q est un estimateur classique de la probabilité $P(X/M_k) /1/$.

On peut alors écrire :

$$\beta P(M_k) \# Q_k/Q, \quad k=1,2,\dots,K$$

où β est une constante indépendante des caractéristiques statistiques de la classe M_k .

Donc, $\frac{P(M_k)}{Q_k} \# C$, C étant une constante.

Pour obtenir les valeurs approchées $\hat{P}(M_k)$ des probabilités à priori $P(M_k)$, il suffit de résoudre le système d'équations :

$$\frac{\hat{P}(M_1)}{Q_1} = \frac{\hat{P}(M_2)}{Q_2} = \dots = \frac{\hat{P}(M_k)}{Q_k}$$

$$\hat{P}(M_1) + \hat{P}(M_2) + \dots + \hat{P}(M_k) = 1$$

Les approximations ainsi obtenues seront actualisées lors de l'extension des domaines modaux.

Nous allons à présent expérimenter sur quelques exemples cette procédure de classification afin de suivre le déroulement de cette méthode justifiant ainsi le bien-fondé de notre étude. L'utilisation d'exemples artificiels permettra de mieux juger l'efficacité de cette nouvelle procédure de classification.

REFERENCES (chapitre IV)

- /1/ R. O. DUDA et P. E. HART, *Pattern classification and Scene Analysis*, J. Wiley New York, Ch. 2, pp. 10-39, 1973.
- /2/ T. MARILL et D. M. GREEN, *Statistical recognition fonctions and the design of patterns recognizers*, IRE Trans. Elec. Comp., Vol. EC-9, pp 472-477, 1960.
- /3/ A. PAPOULIS, *Probability, Random Variables and Stochastic Processes*, Mc Graw-Hill, New York, 1965.
- /4/ J. G. POSTAIRE, *De l'image à la décision*, Dunod, Paris, 1987.
- /5/ J. G. POSTAIRE et C. VASSEUR, *An approximate solution of normal mixture identification with application to unsupervised pattern classification*, IEEE Trans. on Pattern Anal. & Machine Intelligence, Vol. PAMI-3, n° 2, pp. 163-179, 1981.
- /6/ C. VASSEUR et J. G. POSTAIRE, *A convexity testing method for cluster analysis*, IEEE Trans. on Syst., Man. & Cyb., Vol. SMC-10, pp. 145-149, 1980.

ANNEXE 1

Détermination de la matrice de covariance.

Pour simplifier la présentation des calculs, mais sans aucune perte de généralité, nous considérons une fonction de densité normale à vecteur moyenne nul.

$$\begin{aligned} p(X/M_k) &= \frac{1}{(2)^{n/2} |\Sigma_k|^{1/2}} \exp\{-1/2 X^T \Sigma_k^{-1} X\} \\ &= A \exp\{-1/2 X^T \Sigma_k^{-1} X\} \end{aligned}$$

avec :

$$A = \frac{1}{(2)^{n/2} |\Sigma_k|^{1/2}}$$

Dans le cas où le vecteur moyenne n'est pas nul, une simple translation de l'origine, de vecteur X_k , permet de nous ramener à la formule ci-dessus.

Considérons la forme quadratique réelle :

$$Q = X^T \Sigma_k^{-1} X$$

Comme la matrice Σ_k est réelle et symétrique, elle peut être diagonalisée par une transformation orthogonale /4,IV/. En choisissant les vecteurs propres de la matrice Σ_k^{-1} comme nouveaux vecteurs de base, le changement de base :

$$X = M Y$$

permet d'écrire Q sous la forme :

$$Q = Y^T M^T \Sigma_k^{-1} M Y$$

$$= Y^T \Lambda Y$$

où :

$$\Lambda = \begin{bmatrix} \Phi_1 & & & \\ & \ddots & & \\ & & \Phi_i & \\ & & & \ddots \\ & & & & \Phi_n \end{bmatrix}$$

est une matrice diagonale.

Dans cette nouvelle base, la fonction de densité considérée prend la forme particulière :

$$p(Y) = A \exp(-1/2 Y^T \Lambda Y)$$

qui se prête particulièrement bien à l'analyse de sa convexité.

En effet, une condition nécessaire et suffisante pour qu'une fonction admettant des dérivées partielles d'ordre 2, continues sur un domaine convexe ouvert, soit concave sur ce domaine est que son hessien soit semi-défini négatif pour tout point de ce domaine /18/.

Comme les dérivées partielles du second ordre de $p(Y)$ sont continues quelque soit Y , la recherche du plus grand domaine D à l'intérieur duquel la fonction $p(Y)$ est concave se ramène à la détermination du plus grand domaine convexe où la forme quadratique :

$$F = Y^T H\{p(Y)\} Y$$

est semi-définie négative. Le hessien de $p(Y)$ s'écrit :

$$\begin{aligned} H\{p(Y)\} &= H\{A \cdot \exp[-1/2 Y^T \wedge Y]\} \\ &= A \cdot H\{\exp[-1/2 Y^T \wedge Y]\} \end{aligned}$$

La constante A étant positive, le signe de la forme quadratique F est le même que celui de la forme :

$$F' = Y^T \cdot H\{\exp[-1/2 Y^T \wedge Y]\} \cdot Y$$

qui peut être évalué grâce aux relations :

$$\frac{d^2 \exp[-1/2 Y^T \wedge Y]}{dy_i^2} = \left[\left[\frac{dY^T}{dy_i} \wedge Y \right]^2 - \frac{dY^T}{dy_i} \wedge \frac{dY}{dy_i} \right] \exp[-1/2 Y^T \wedge Y]$$

et :

$$\frac{d^2 \exp[-1/2 Y^T \wedge Y]}{dy_i^2} = \left[\left[\frac{dY^T}{dy_i} \wedge Y \right] \left[\frac{dY^T}{dy_j} \wedge Y \right] - \frac{dY^T}{dy_i} \wedge \frac{dY}{dy_j} \right] \exp[-1/2 Y^T \wedge Y]$$

où y_i , $i = 1, 2, \dots, N$ sont les composantes du vecteur Y .

Le hessien $H\{\exp[-1/2 Y^T Y]\}$ prend alors la forme :

$$H\{\exp[-1/2 Y^T Y]\} = (H_1 + H_2) \cdot \exp[-1/2 Y^T Y]$$

avec :

$$H_1 = \left[\begin{array}{cc} \left[\frac{dY^T}{dy_1} \wedge Y \right]^2 & \left[\frac{dY^T}{dy_1} \wedge Y \right] \left[\frac{dY^T}{dy_n} \wedge Y \right] \\ \left[\frac{dY^T}{dy_1} \wedge Y \right]^2 & \\ \left[\frac{dY^T}{dy_n} \wedge Y \right] \left[\frac{dY^T}{dy_1} \wedge Y \right] & \left[\frac{dY^T}{dy_n} \wedge Y \right]^2 \end{array} \right]$$

$$H_2 = \left[\begin{array}{cc} \frac{dY^T}{dy_1} \wedge \frac{dY}{dy_1} & \frac{dY^T}{y_1} \wedge \frac{dY}{dy_n} \\ \frac{dY^T}{dy_1} \wedge \frac{dY}{dy_1} & \\ \frac{dY^T}{dy_n} \wedge \frac{dY}{dy_1} & \frac{Y^T}{dy_n} \wedge \frac{dY}{dy_n} \end{array} \right]$$

En introduisant la matrice :

$$\left[\frac{dY^T}{dy_1}, \dots, \frac{dY^T}{dy_i}, \dots, \frac{dY^T}{dy_n} \right]^T = I_n$$

(I_n : matrice unité d'ordre n), H_1 peut s'écrire :

$$\begin{aligned} H_1 &= (I_n \wedge Y)(I_n \wedge Y)^T \\ &= (\wedge Y)(\wedge Y)^T. \end{aligned}$$

De manière analogue :

$$\begin{aligned} H_2 &= I_n \wedge I_n^T \\ &= \wedge. \end{aligned}$$

Ainsi :

$$H\{\exp[-1/2 Y^T \wedge Y]\} = \{(\wedge Y)(\wedge Y)^T - \wedge\} \exp[-1/2 Y^T \wedge Y].$$

La détermination de la convexité de la fonction $p(Y)$ se ramène finalement à l'étude du signe de la forme quadratique :

$$F'' = Y^T \cdot \{(\wedge Y)(\wedge Y)^T - \wedge\} \cdot Y.$$

Soient \wedge_i , $i = 1, 2, \dots, n$, les mineurs principaux de la matrice :

$$H_r = (\wedge Y)(\wedge Y)^T - \wedge$$

tels que :

$$\wedge_i = \begin{bmatrix} \overset{2}{\Phi_1} \overset{2}{y_1} - \Phi_1 & \Phi_1 \Phi_2 y_1 y_2 & \dots & \Phi_1 \Phi_i y_1 y_i \\ \Phi_2 \Phi_1 y_2 y_1 & \overset{2}{\Phi_2} \overset{2}{y_2} - \Phi_2 & & \\ & & & \\ \Phi_i \Phi_1 y_i y_1 & & & \overset{2}{\Phi_i} \overset{2}{y_i} - \Phi_i \end{bmatrix}$$

Une condition nécessaire et suffisante pour que la forme quadratique F'' soit définie négative est :

$$\left. \begin{array}{l} \Delta_i < 0 \text{ si } i \text{ impair} \\ \Delta_i > 0 \text{ si } i \text{ pair} \end{array} \right\} i=1, 2, \dots, n.$$

Après une série de manipulations sur les lignes et les colonnes de Δ_i , on montre que ce déterminant a le même signe que :

$$\Delta_i = (-1)^i (1 - \Phi_{1y_1}^2 - \Phi_{2y_2}^2 - \dots - \Phi_{ny_n}^2).$$

La forme quadratique F'' est donc semi-définie négative si et seulement si :

$$1 - \sum_{j=1}^i \Phi_{jy_j}^2 \geq 0, \quad i=1, 2, \dots, n.$$

-1

Compte tenu du signe des valeurs propres de la matrice Σ_k , cet ensemble d'inégalités est vérifié si et seulement si :

$$Y^T \Lambda Y \leq 1$$

La fonction de densité $p(Y)$ est donc concave dans le domaine convexe D intérieur à la surface Γ d'équation :

$$Y^T \Lambda Y = 1$$

CHAPITRE V

IDENTIFICATION DES MELANGES GAUSSIENS

V.1 INTRODUCTION.

Afin de montrer l'intérêt des procédures de relaxation présentées dans cette thèse, nous appliquons la procédure de classification proposée au problème de l'analyse des mélanges gaussiens générés artificiellement sous l'hypothèse d'indépendance statistique des composantes des observations. Le fait d'utiliser de telles données permet de contrôler leur structure et d'analyser l'effet des différents paramètres sur les résultats de la classification.

Ce cadre d'expérimentation permet de bien focaliser l'attention du lecteur sur des points précis et définis rigoureusement. Il va sans dire que la méthode peut être appliquée à des exemples présentant moins de contraintes (composantes statistiquement dépendantes les unes des autres, distribution non paramétrique, etc ...) mais l'interprétation des résultats fonction de très nombreux paramètres, serait plus délicate.

Nous présentons d'abord des exemples bidimensionnels dans le but de faciliter la compréhension du processus de classification et de permettre de visualiser simplement les résultats.

Pour la création des exemples artificiels, nous choisissons le nombre de classes que l'on désire avoir en présence et, pour chacune d'elles, le nombre d'observations, les coordonnées du vecteur-moyenne ainsi que les éléments de la matrice de covariance.

V.2 Exemple de base à deux dimensions.

Le premier exemple présenté est constitué de 300 points répartis en trois classes équiprobables de vecteurs moyennes :

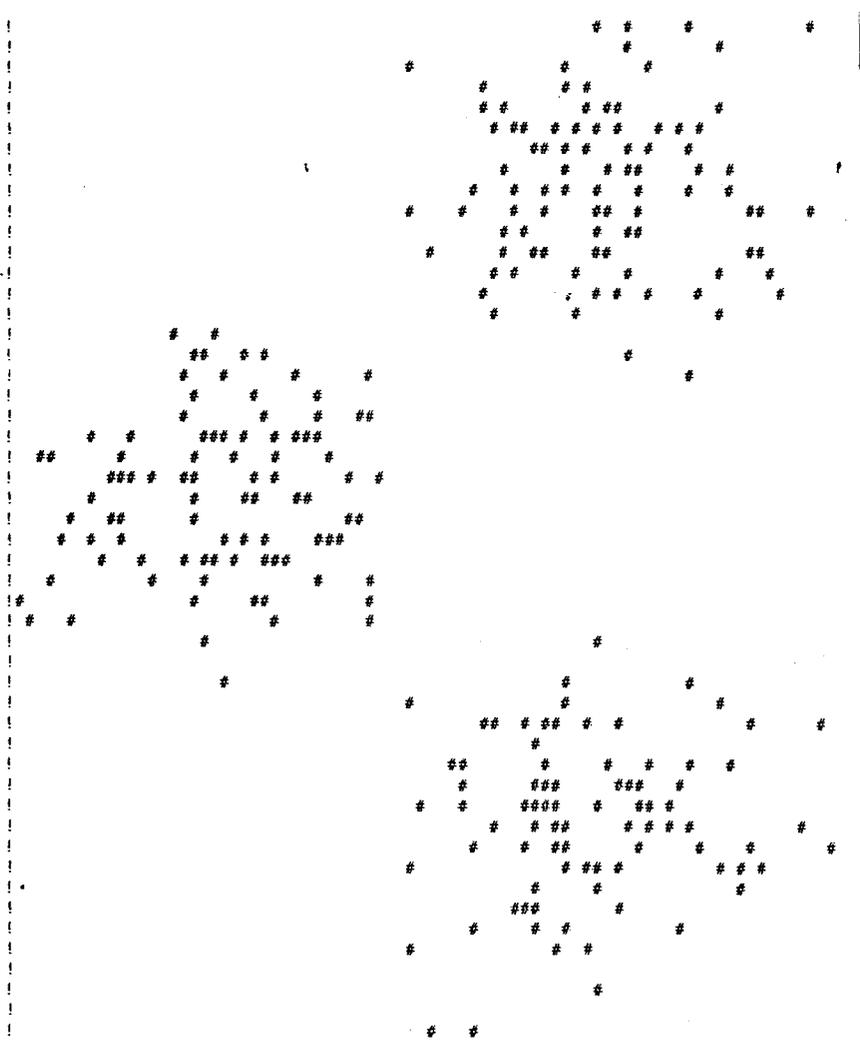
$$X_1 = (8,8)^T, X_2 = (14,14)^T, X_3 = (14,2)^T$$

et de matrices de covariance :

$$\Sigma_1 = \Sigma_2 = \Sigma_3 = 2.I_2, I_2 \text{ étant la matrice-unité d'ordre } 2.$$

Cet exemple numéro 1 nous sert de support pour montrer le comportement de la procédure aux différentes étapes de l'analyse des données et pour analyser l'effet du réglage des paramètres de l'algorithme.

Nous remarquons en examinant la représentation graphique des observations constituant cet exemple (figure 1) la présence de "trous", c'est-à-dire de zones vides d'observations, aux environs des valeurs moyennes de chacune des classes. Ces zones simulent des accidents dans la distribution des observations, assimilés à du bruit. Ils sont destinés à montrer l'efficacité et la robustesse de la procédure de classification.



représente les observations de la classe 1.
\$ représente les observations de la classe 2.
% représente les observations de la classe 3.

Figure 1 : Représentation graphique des 300 observations de l'exemple 1.

Après normalisation et discrétisation de l'espace de représentation des données de la figure 1, on obtient l'ensemble des hypercubes non vides de la figure 2 (cf chapitre II.3). Ceux-ci sont numérotés de 1 à 126 pour faciliter leur identification.

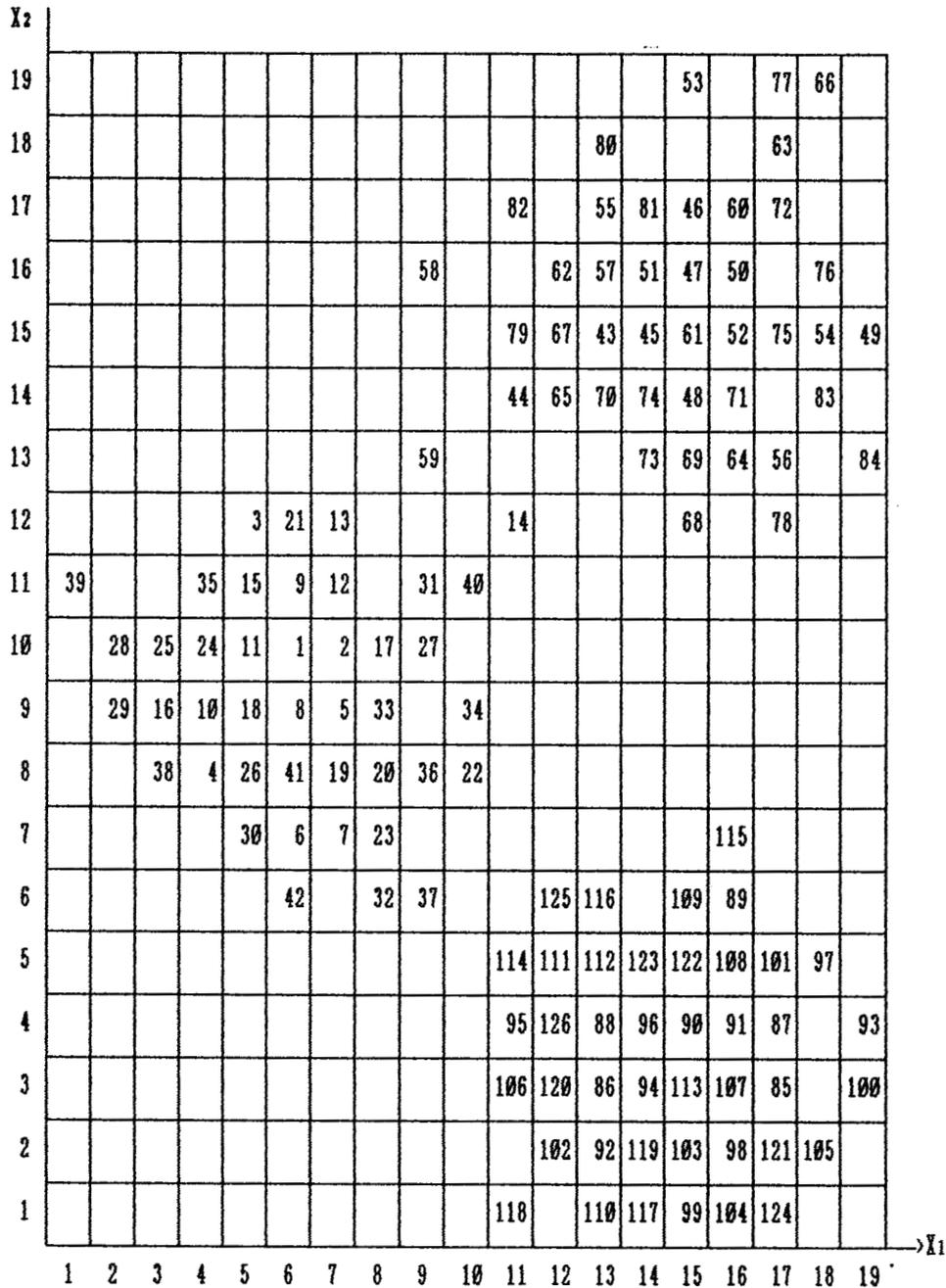


Figure 2 : Emplacement et numérotation des hypercubes non vides de l'exemple 1.

Sur la figure 3, on indique le nombre d'observations situées dans chaque hypercube non vide.

19																		1		1	1	
18													2							1		
17										1		3	1	2	2	1						
16								1			3	6	4	4	7		2					
15										1	2	4	4	8	6	2	2	1				
14										2	2	2	3	4	3		1					
13								1						2	1	1	2				1	
12				1	2	2				1					1		1					
11	1			2	3	6	1		2	1												
10		1	1	2	3	8	3	4	4													
9		2	3	4	6	3	6	3		1												
8			1	2	1	1	3	4	1	1												
7				1	2	1	3													1		
6					1		1	1			1	3			2	3						
5										1	1	2	2	1	2	1	2					
4										2	1	5	5	3	8	1						2
3										3	1	4	5	5	2	4						2
2												4	2	1	4	3	1	1				
1										1		1	1	3	1	2						

Figure 3 : Nombre d'observations par hypercube non vide de l'exemple 1.

Cet ensemble d'hypercubes est complété par l'adjonction de 28 hypercubes vides de type 3 (cf chapitre II.5). La figure 4 représente la totalité des 154 hypercubes qui seront pris en compte dans la suite de la procédure.

19															53		77	66	
18												80	151	152	153	63	154		
17										82	150	55	81	46	60	72			
16								58		148	62	57	51	47	50	149	76		
15										79	67	43	45	61	52	75	54	49	
14										44	65	70	74	48	71	146	83	147	
13								59		143	144	73	69	64	56	145	84		
12				3	21	13	140		141	14				68	142	78			
11	39		138	35	15	9	12	139	31	40									
10		28	25	24	11	1	2	17	27	137									
9		29	16	10	18	8	5	33	136	34									
8			38	4	26	41	19	20	36	22									
7				134	30	6	7	23	135						115				
6						42	131	32	37		125	116	132	109	89	133			
5											114	111	112	123	122	108	101	97	
4											95	126	88	96	90	91	87	130	93
3											106	120	86	94	113	107	85	129	100
2											128	102	92	119	103	98	121	105	
1											118	127	110	117	99	104	124		

Figure 4 : Emplacement et numérotation des hypercubes à prendre en compte dans l'exemple 1.

On étudie alors le sens de la convexité au niveau de chacun des hypercubes de la figure 4 (cf chapitre I.4). Les hypercubes de convexité positive sont repérés par un signe + alors que ceux de convexité négative sont identifiés par le signe - (cf figure 5). Les zéros représentent des hypercubes à convexité nulle, c'est-à-dire ni positive, ni négative.

19																	-	-	-	
18												-	-	-	-	-	-			
17												-	+	+	+	-	-	-		
16									∅			-	+	+	+	+	+	+	-	
15												-	+	+	+	+	+	+	-	+
14												-	-	-	+	+	+	+	-	-
13									-			-	-	-	-	-	-	-	-	-
12						+	-	-	-		+	-					-	-	-	
11	-		-	-	+	+	+	-	-	+										
10		-	+	+	+	+	+	-	-	-										
9		-	-	-	+	+	+	+	+	-										
8			+	+	-	+	+	+	-	-										
7				-	-	-	+	-	+									∅		
6						-	+	-	-			-	-	-	-	-	-			
5												-	+	-	+	+	+	+	-	
4												-	+	+	+	+	+	+	-	-
3												-	+	+	+	+	+	+	-	-
2												-	-	-	+	-	+	-	+	
1												-	-	-	-	-	+	-		
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19

Figure 5 : Résultat du test de convexité appliqué à l'exemple 1.

Si la procédure s'arrêtait à cette étape, il serait délicat d'obtenir une classification fiable /1/. La figure 6 marque l'étape de la relaxation en spécifiant pour chaque hypercube la valeur de sa probabilité initiale d'appartenir à la classe des hypercubes à convexité positive (cf chapitre III.4).

19															.28		.22	.28
18											.67	.33	.22	.06	.11	.22		
17										.5	.44	.72	.44	.56	.44	.28		
16								.5	.5	.56	1	.89	.78	.67	.39	.61		
15									.39	.33	.78	.89	1	1	.67	.61	.22	
14									.33	.28	.56	.56	.67	.67	.33	.33	.28	
13								.44		.17	.28	.33	.39	.33	.28	.22	.28	
12				.44	.72	.39	.28	.33	.61				.22	.17	.22			
11	.39	.44	.78	.72	.78	.33	.28	.28	.28									
10	.28	.17	.44	.78	1	.78	.67	.44	.39									
9	.44	.61	.78	.78	.89	.89	.67	.22	.28									
8		.28	.28	.44	.67	.78	.78	.5	.44									
7			.44	.33	.5	.44	.56	.22							.28			
6				.44	.28	.5	.39		.33	.39	.39	.5	.61	.44				
5									.44	.28	.56	.56	.56	.67	.39	.5		
4									.5	.44	.89	.89	1	1	.67	.39	.22	
3									.39	.33	.78	.78	.89	.78	.67	.39	.33	
2									.28	.22	.56	.44	.78	.56	.72	.28		
1									.28	.17	.28	.28	.61	.28	.56			

Figure 6 : Probabilités initiales calculées à partir de l'exemple 1.

Ces probabilités initiales sont alors remises à jour par itérations successives (cf chapitre III.5). Seules les itérations n° 1 et 2 sont représentées sur les figures 7 et 8 avant l'itération finale (figure 9) obtenue lorsqu'il n'y a plus de modifications substantielles des probabilités (cf chapitre III.6).

19															.15	.09	.15			
18												.54	.33	.09	.01	.02	.09			
17										.5	.71	.77	.71	.42	.32	.15				
16								.5		.38	.56	1	.98	.91	.77	.58	.55			
15										.28	.39	.91	1	1	1	.77	.42	.22		
14										.14	.15	.43	.68	.77	.67	.33	.14	.15		
13								.32			.13	.23	.33	.33	.14	.08	.04	.15		
12				.71	.67	.33	.09		.23	.42				.09	.04	.09				
11	.33		.32	.68	.85	.85	.45	.15	.09	.15										
10		.15	.13	.71	.96	1	.91	.55	.21	.13										
9		.32	.27	.68	.85	.98	1	.81	.18	.11										
8			.23	.23	.51	.72	.94	.82	.44	.27										
7				.21	.19	.31	.57	.43	.22						.28					
6					.27	.23	.31	.39			.23	.33	.51	.63	.55	.44				
5										.27	.19	.43	.73	.84	.81	.51	.31			
4										.19	.38	.93	1	1	1	.72	.28	.11		
3										.18	.28	.78	.96	.98	1	.77	.33	.14		
2										.08	.09	.29	.71	.85	.90	.77	.33			
1										.15	.08	.15	.33	.55	.63	.56				
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19

Figure 7 : Probabilités à la première itération de l'exemple 1.

19																	.07	.03	.07	
18											.67	.33	.03	0	.02	.03				
17										.56	.84	.94	.80	.43	.22	.12				
16								.5		.38	.62	1	.99	.97	.85	.64	.49			
15										.15	.28	.95	1	1	1	.85	.36	.15		
14										.05	.38	.31	.68	.85	.67	.23	.02	.04		
13								.22			.03	.12	.23	.28	.05	.02	.01	.07		
12					.88	.72	.28	.02	.09	.25				.03	.01	.03				
11	.28		.32	.78	.99	.95	.58	.07	.02	.04										
10		.04	.11	.80	1	1	.97	.42	.08	.03										
9		.14	.14	.68	.98	1	1	.88	.12	.03										
8			.12	.19	.51	.89	.98	.82	.26	.11										
7				.14	.12	.37	.44	.31	.09						.33					
6					.18	.12	.17	.22			.09	.28	.58	.83	.67	.51				
5										.11	.05	.43	.89	1	.97	.58	.26			
4										.05	.17	.93	1	1	1	.81	.23	.04		
3										.05	.11	.78	.99	1	1	.85	.28	.05		
2										.02	.01	.20	.80	.98	1	.88	.39			
1										.07	.02	.07	.39	.73	.88	.68				
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	→X ₁

Figure 8 : Probabilités à la seconde itération de l'exemple 1.

Entre les figures 7 et 8, on constate l'augmentation des probabilités au centre des classes et leur diminution sur leur pourtour et dans les zones inter-classes.

19																	0		0	0
18											.98	.33	0	0	0	0				
17										.97	.99	.99	.99	.43	0	0				
16								.50		.73	.99	1	1	.99	.96	.28	.04			
15										.01	.08	.99	1	1	1	.55	0	0		
14										0	0	.02	.68	.99	.67	.01	0	0		
13								.01			0	0	.01	.08	0	0	0	0		
12				.99	.99	.63	0		0	0					0	0	0			
11	.08		.32	.99	1	1	.79	0	0	0										
10		0	.03	.99	1	1	.99	.42	0	0										
9		0	0	.68	.99	1	1	.61	0	0										
8			0	.05	.51	.99	.99	.17	0	0										
7				.01	.01	.11	.01	0	0							.98				
6					0	0	0	0			0	.08	.86	.99	.99	.99				
5										0	0	.43	.99	1	1	.99	.61			
4										0	0	.93	1	1	1	.99	.10	0		
3										0	0	.78	.99	1	1	.99	.39	0		
2										0	0	.08	.99	1	1	.99	.74			
1										0	0	.01	.74	.99	.99	.99				

→X₁

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19

Figure 9 : Probabilités à la dernière itération de l'exemple 1.

La figure 10 représente les domaines modaux initiaux constitués des hypercubes convexe dont la probabilité d'être à convexité positive est supérieure à 0.5 .

19																																									
18														2	0	0	0	0	0	0																					
17														2	2	2	2	0	0	0																					
16									0					2	2	2	2	2	2	0	0																				
15														0	0	2	2	2	2	2	0	0																			
14														0	0	0	2	2	2	0	0	0																			
13									0					0	0	0	0	0	0	0	0	0																			
12					1	1	1	0		0	0								0	0	0	0																			
11	0		0	1	1	1	1	0	0	0																															
10		0	0	1	1	1	1	0	0	0																															
9		0	0	1	1	1	1	1	0	0																															
8			0	0	1	1	1	0	0	0																															
7				0	0	0	0	0	0																																
6						0	0	0	0					0	0	3	3	3	3																						
5														0	0	0	3	3	3	3																					
4														0	0	3	3	3	3	3	0	0																			
3														0	0	3	3	3	3	3	0	0																			
2														0	0	0	3	3	3	3	3																				
1														0	0	0	3	3	3	3																					

Figure 10 : Domaines modaux initiaux de l'exemple 1.

Ces domaines modaux initiaux sont repérés par le numéro affecté à la classe correspondante.

Les hypercubes repérés par des zéros ne sont pas encore assignés à des classes. Ces hypercubes sont caractérisés par une probabilité d'être à convexité négative supérieure à 0.5 . Remarquons également qu'un ensemble d'hyperscubes à probabilité d'être à convexité positive supérieure à 0.5 mais ne constituant pas des domaines modaux

représentatifs du fait du nombre réduit d'observations contenues dans ces domaines serait étiqueté avec des zéros (cf chapitre IV.2). Dans cet exemple, on considère qu'un domaine modal est représentatif que s'il contient une fraction non négligeable du nombre total d'observations à classer. L'ajustement de ce paramètre de rejet sera discuté ultérieurement. Il est ici fixé à un minimum de 15 observations par domaine modal .

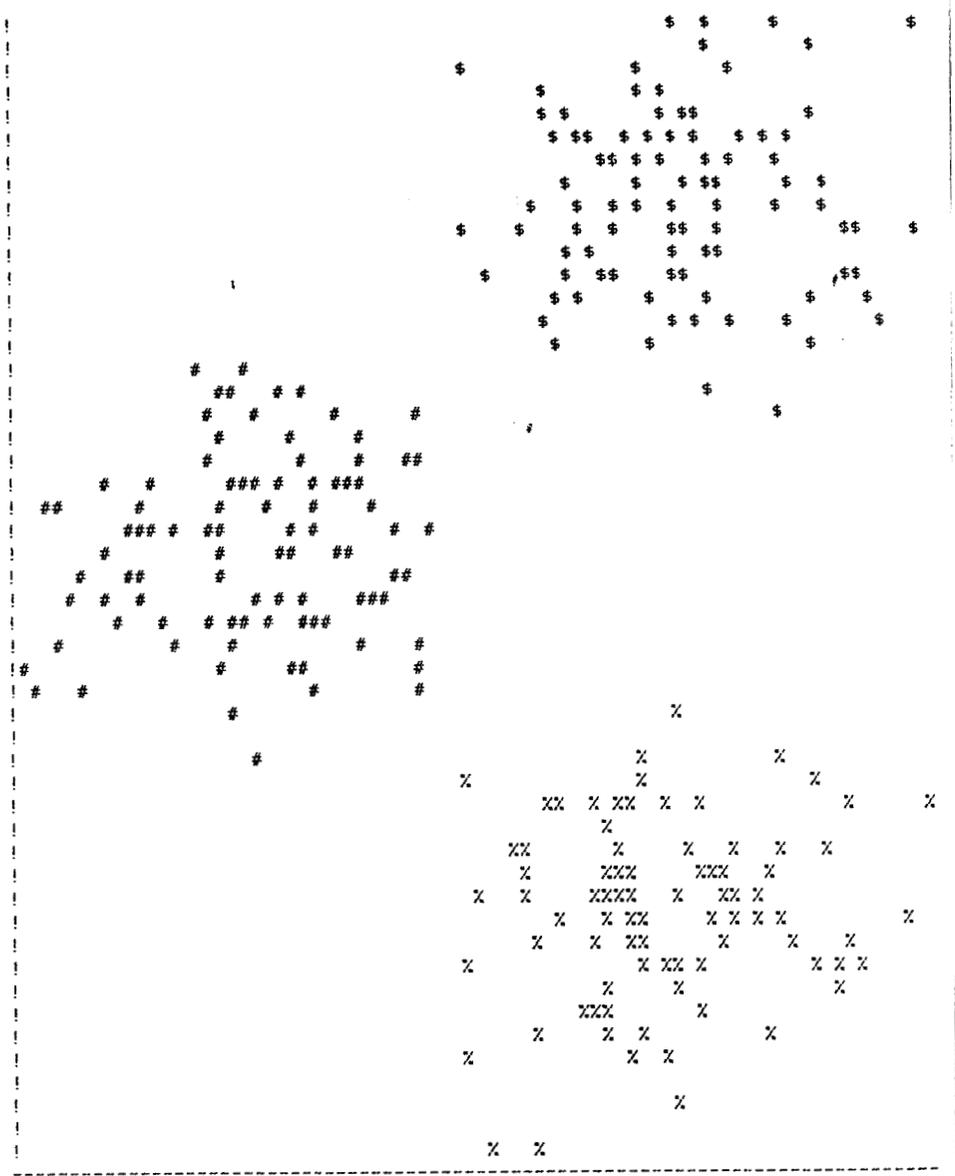
Nous analyserons ultérieurement l'effet de ce seuil de rejet R.

19																0		2	2	
18													2	0	0	0		2	0	
17												2	2	2	2	2	2	2		
16										0		2	2	2	2	2	2	0	2	
15												2	2	2	2	2	2	2	2	2
14												2	2	2	2	2	2	0	2	0
13											0		0	0	2	2	2	2	0	2
12					1	1	1	0		0	1					2	0	2		
11	1		0	1	1	1	1	0	1	1										
10		1	1	1	1	1	1	1	1	0										
9		1	1	1	1	1	1	1	0	1										
8			1	1	1	1	1	1	1	1										
7				0	1	1	1	1	0								3			
6						1	0	1	1			3	3	3	3	3	3	3		
5												3	3	3	3	3	3	3	3	3
4												3	3	3	3	3	3	3	0	3
3												3	3	3	3	3	3	3	0	3
2												0	3	3	3	3	3	3	3	3
1												3	0	3	3	3	3	3		

Figure 11 : Domaines modaux après extension séquentielle de l'exemple 1.

La figure 11 représente le résultat de l'extension séquentielle des domaines modaux initiaux (cf chapitre IV.3).

Les hypercubes repérés par un zéro sont soit les hypercubes vides introduits pour compléter la liste des hypercubes non vides (cf chapitre II.5), soit des hypercubes isolés. La figure 12 représente l'étendue des classes obtenues à la fin de la procédure.



représente les observations de la classe 1.
 \$ représente les observations de la classe 2.
 % représente les observations de la classe 3.
 & représente les observations mal classées.

Figure 12 : Etendue des classes de l'exemple 1 en fin de procédure.

V.3 EFFET DES PARAMETRES DE REGLAGE DE L'ALGORITHME DE CLASSIFICATION.

V.3.1 Effet du pas de discrétisation.

Le pas αq introduit pour discrétiser l'espace de représentation (cf chapitre II.6) influe sur les résultats de la classification.

La figure 13 représente les variations du nombre de modes détectés en fonction de ce pas, le seuil de rejet des modes initiaux non représentatifs étant fixé à un minimum de 15 observations par domaine modal .

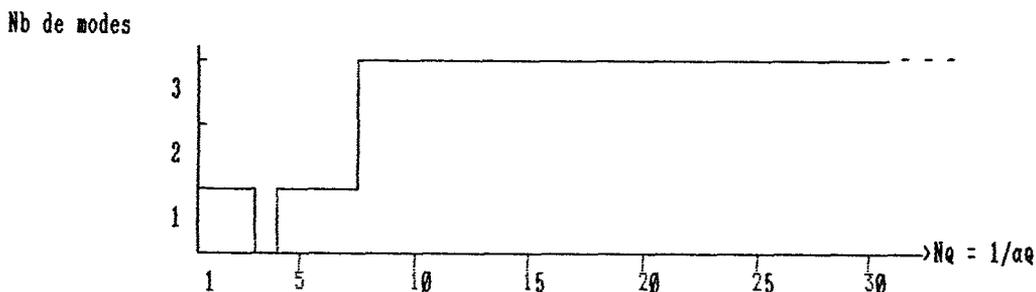


Figure 13 : Nombre de modes détectés en fonction du pas de discrétisation αq .

On constate sur cette figure que le nombre de classes détectées présente une plage de stabilité de longueur maximale pour Nq variant de 8 à plus de 30 , valeur maximale utilisée pour cette expérimentation.

Le concept de stabilité du nombre de modes détectés nous pousse à ajuster Nq au milieu de cette plage, soit $Nq = 19$

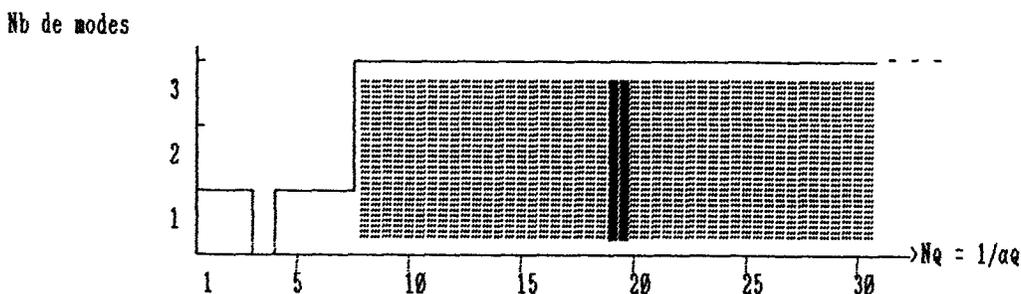
V.3.2 Effet du seuil de rejet des modes représentatifs.

Il est également important d'analyser l'effet de l'ajustement d'un second paramètre : le seuil de rejet des modes, c'est-à-dire le nombre d'observations minimum contenues dans un domaine modal initial nécessaire pour que ce dernier soit retenu.

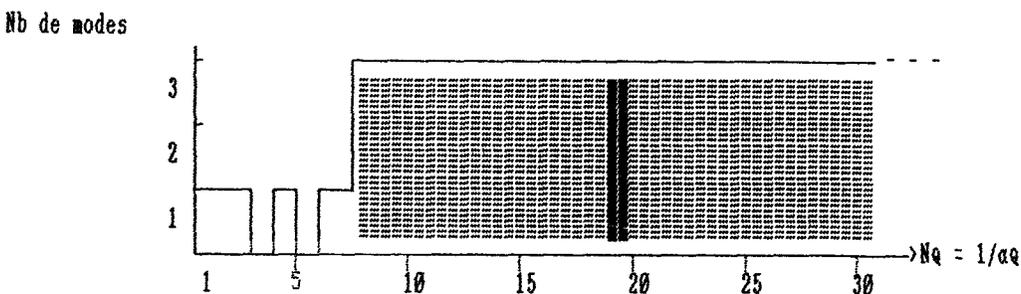
La figure 14 représente l'évolution du nombre de modes détectés en fonction du pas de discrétisation Nq pour trois valeurs de ce seuil de rejet R . Nous constatons que la plus grande plage de variation de Nq correspondant au même nombre de modes détectés présente les mêmes caractéristiques pour ces trois réglages du seuil de rejet R , elle s'étend de $Nq = 8$ à $Nq = 30$, valeur limite testée.

C'est donc cette plage que nous retiendrons, ajustant Nq en son milieu à la valeur 19.

Seuil de rejet fixé à un minimum de 15 observations par domaine modal :



Seuil de rejet fixé à un minimum de 20 observations par domaine modal :



Seuil de rejet fixé à un minimum de 25 observations par domaine modal :

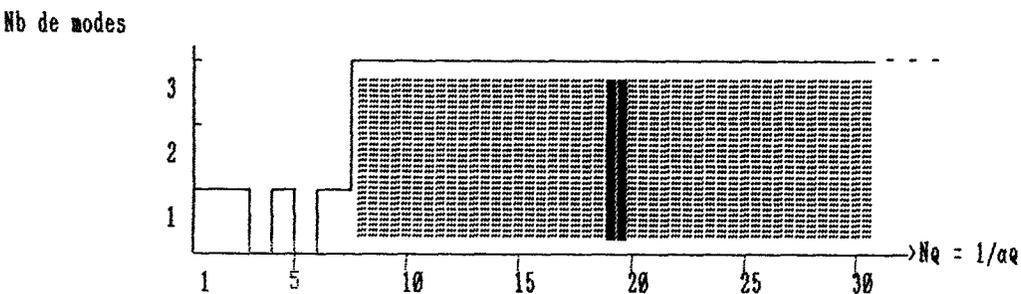


Figure 14 : Nombre de modes détectés en fonction du pas de discrétisation ae pour différentes valeurs du seuil de rejet R .

Cette comparaison de résultats pour différentes valeurs du seuil de rejet R met en évidence une plage importante pour l'ajustement de ce seuil, qui n'est donc pas critique.

Cette stabilité des résultats constitue l'un des apports fondamentaux de la méthode de relaxation par rapport à l'utilisation directe du concept de convexité. En effet, les algorithmes initiaux, sans étiquetage probabiliste itératif, sont connus pour leur grande sensibilité par rapport au réglage des paramètres /2-3/.

V.4 CLASSIFICATION DE DONNEES BIDIMENSIONNELLES.

V.4.1 Classes équiprobables.

L'exemple de base numéro 1, présenté au chapitre V.2 est constitué de classes équiprobables ne présentant qu'un faible degré de chevauchement.

En modifiant les valeurs des éléments des matrices de covariance des trois classes gaussiennes en présence, il est possible d'analyser l'effet du chevauchement des classes sur les performances de l'algorithme.

Les trois exemples dont les paramètres statistiques sont consignés dans le tableau 1 montrent que pour les exemples 1 et 2, la procédure de classification donne des résultats satisfaisants. Dans l'exemple 3, le degré de chevauchement est trop important pour que les classes puissent être différenciées sur la base du concept de stabilité des nombres de modes détectés.

Il peut paraître paradoxal que le taux d'erreur réel de la classification soit inférieur au taux d'erreur optimal en dessous duquel il est théoriquement impossible de descendre.

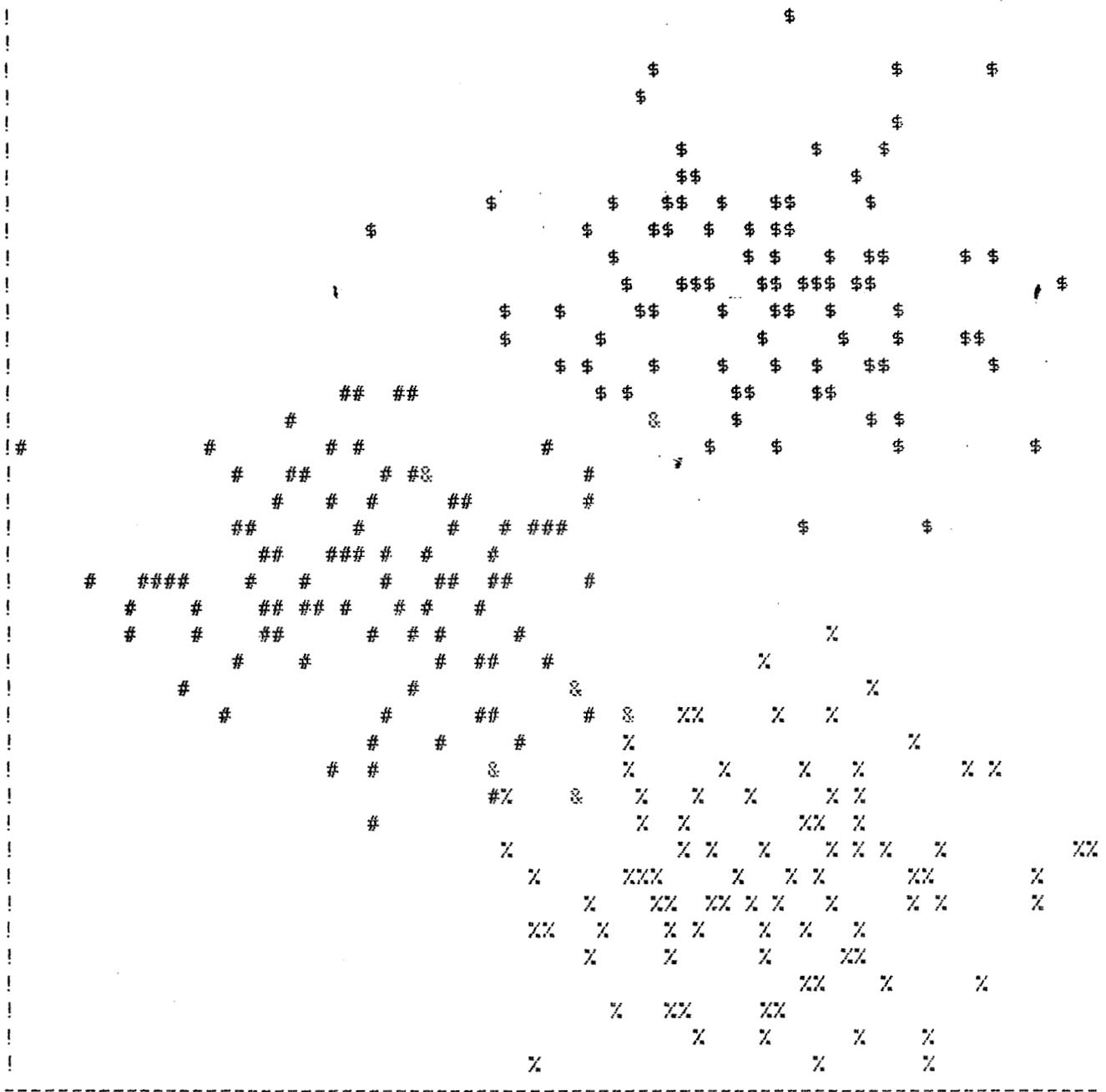
En fait, l'estimation du taux d'erreur réel, calculé comme le rapport entre le nombre d'observations mal classées au nombre d'observations total, peut être biaisé lorsque le nombre d'erreurs de classification est faible.

	VECTEURS MOYENNE	MATRICES DE COVARIANCE	TAUX D'ERREUR THEORIQUE MINI.	TAUX D'ERREUR DE LA CLASSIFICATION
Exemple 1	$\bar{X}_1 = \begin{bmatrix} 8 \\ 8 \end{bmatrix}$ $\bar{X}_2 = \begin{bmatrix} 14 \\ 14 \end{bmatrix}$ $\bar{X}_3 = \begin{bmatrix} 14 \\ 2 \end{bmatrix}$	$\Sigma_k = 2I_2, k=1,2,3$	1.2 %	0.33 %
Exemple 2	$\bar{X}_1 = \begin{bmatrix} 8 \\ 8 \end{bmatrix}$ $\bar{X}_2 = \begin{bmatrix} 14 \\ 14 \end{bmatrix}$ $\bar{X}_3 = \begin{bmatrix} 14 \\ 2 \end{bmatrix}$	$\Sigma_k = 4I_2, k=1,2,3$	3.33 %	2.00 %
Exemple 3	$\bar{X}_1 = \begin{bmatrix} 8 \\ 8 \end{bmatrix}$ $\bar{X}_2 = \begin{bmatrix} 14 \\ 14 \end{bmatrix}$ $\bar{X}_3 = \begin{bmatrix} 14 \\ 2 \end{bmatrix}$	$\Sigma_k = 8I_2, k=1,2,3$	10.10 %	11.33 %

Tableau 1 : Paramètres statistiques des distributions bidimensionnelles et performances de la procédure de classification.

Le taux d'erreur indiqué pour l'exemple 3 a été trouvé pour $N_q = 23$. Cet ajustement a été effectué manuellement du fait du manque de stabilité du nombre de modes détectés.

La figure 15 indique les résultats de la classification pour l'exemple 2 avec un ajustement de N_q à la valeur 20 (cf figure 16), un coefficient de rejet étant fixé à un minimum de 20 observations par domaine modal.



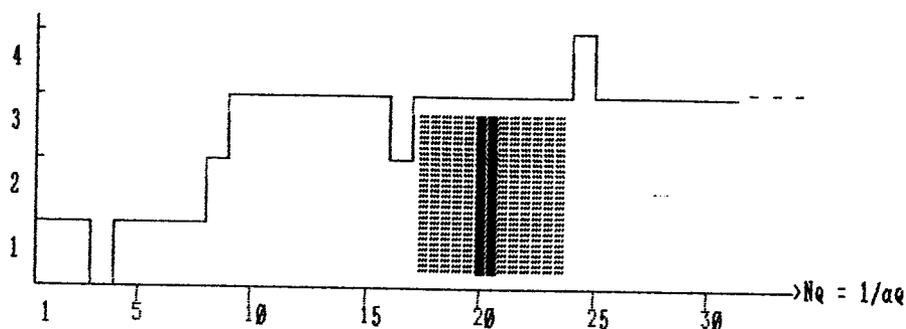
représente les observations de la classe 1.
 \$ représente les observations de la classe 2.
 % représente les observations de la classe 3.
 & représente les observations mal classées.

Figure 15 : Etendue des classes de l'exemple 2 en fin de procédure.

Cette représentation montre la difficultés de dissocier les classes en présence, celles-ci présentant un chevauchement non négligeable.

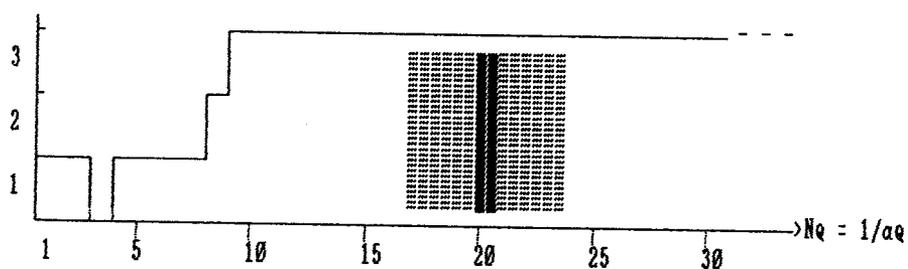
Seuil de rejet fixé à un minimum de 15 observations par domaine modal :

Nb de modes



Seuil de rejet fixé à un minimum de 20 observations par domaine modal :

Nb de modes



Seuil de rejet fixé à un minimum de 25 observations par domaine modal :

Nb de modes

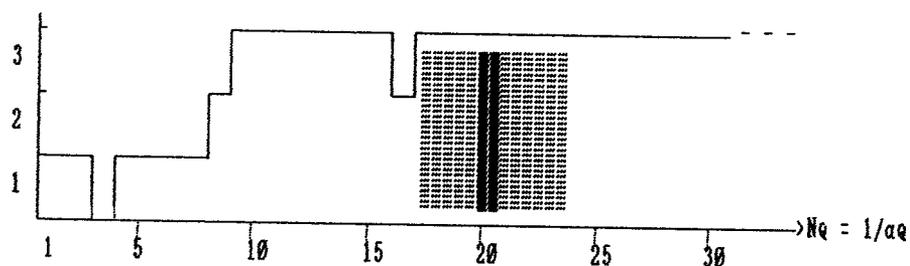


Figure 16 : Nombre de modes détectés pour l'exemple 2 en fonction du pas de discrétisation aq pour différentes valeurs du seuil de rejet R .

La figure 17 montre qu'il n'y a pas de plage stable du nombre de modes détectés en fonction de Nq pour l'exemple 3.

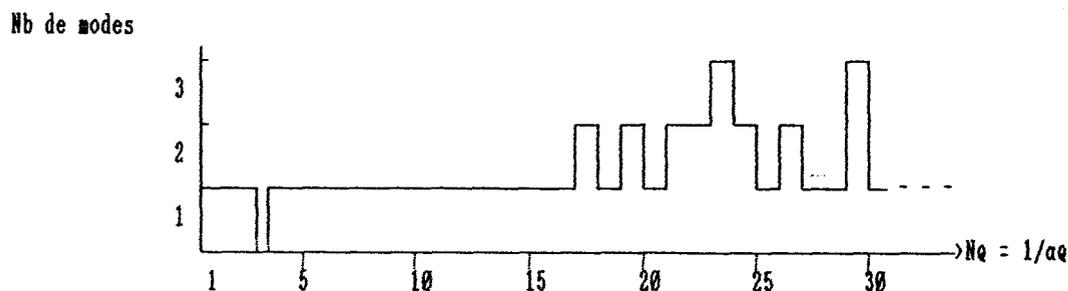


Figure 17 : Nombre de modes détectés pour l'exemple 3 en fonction du pas de discrétisation Δq .

V.4.2 Classes non équiprobables.

Un des intérêts majeurs de la procédure de classification est de pouvoir traiter des données constituées de classes non équiprobables.

En effet, on sait que la majorité des procédures de classification automatique opèrent d'autant mieux que les classes en présence présentent des poids comparables. La méthode utilisée étant basée sur une procédure d'estimation non paramétrique, la détermination de la convexité ne peut être obtenue que grâce à un compromis de telle sorte que l'algorithme donne des résultats corrects pour des classes de densités différentes. Le réglage des paramètres ne peut, en fait, être optimisé simultanément pour les trois classes.

L'exemple 4 (figure 18), dont les paramètres statistiques sont donnés dans le tableau 2, est utilisé pour montrer l'évolution des probabilités au cours de la phase de relaxation. Celles-ci tendent à devenir de plus en plus homogènes dans chaque domaine de convexité, renforçant ainsi la robustesse de la méthode.

	VECTEURS MOYENNE	VARIANCES DES COMPOSANTES	PROBABILITES A PRIORI
Classe 1	$\bar{X}_1 = \begin{bmatrix} 8 \\ 8 \end{bmatrix}$	$\sigma_{X1} = 2$ $\sigma_{X2} = 2$	$\sigma_{X1} = 2,01$ $\sigma_{X2} = 1,98$ $P(C_1)=0.20$ $P(C_1)=0.20$
Classe 2	$\bar{X}_2 = \begin{bmatrix} 14 \\ 14 \end{bmatrix}$	$\sigma_{X1} = 2$ $\sigma_{X2} = 2$	$\sigma_{X1} = 2,01$ $\sigma_{X2} = 2,00$ $P(C_2)=0.33$ $P(C_2)=0.33$
Classe 3	$\bar{X}_3 = \begin{bmatrix} 14 \\ 2 \end{bmatrix}$	$\sigma_{X1} = 2$ $\sigma_{X2} = 2$	$\sigma_{X1} = 1,99$ $\sigma_{X2} = 2,01$ $P(C_3)=0.47$ $P(C_3)=0.47$
TAUX D'ERREUR THEORIQUE MINIMUM = 0,16 % TAUX D'ERREUR DE LA CLASSIFICATION = 0 %			

Tableau 2 : Paramètres statistiques de l'exemple 4 et performances de la procédure de classification.

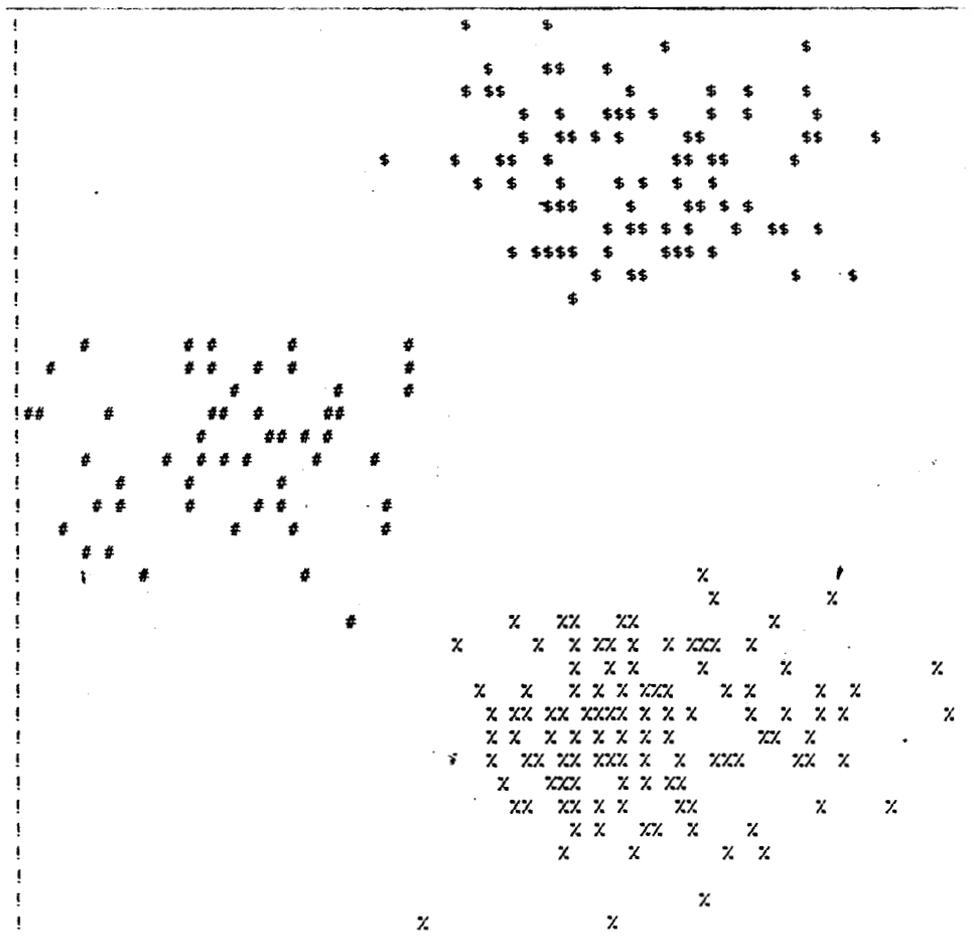


Figure 18 : Représentation graphique des 300 observations de l'exemple 4.

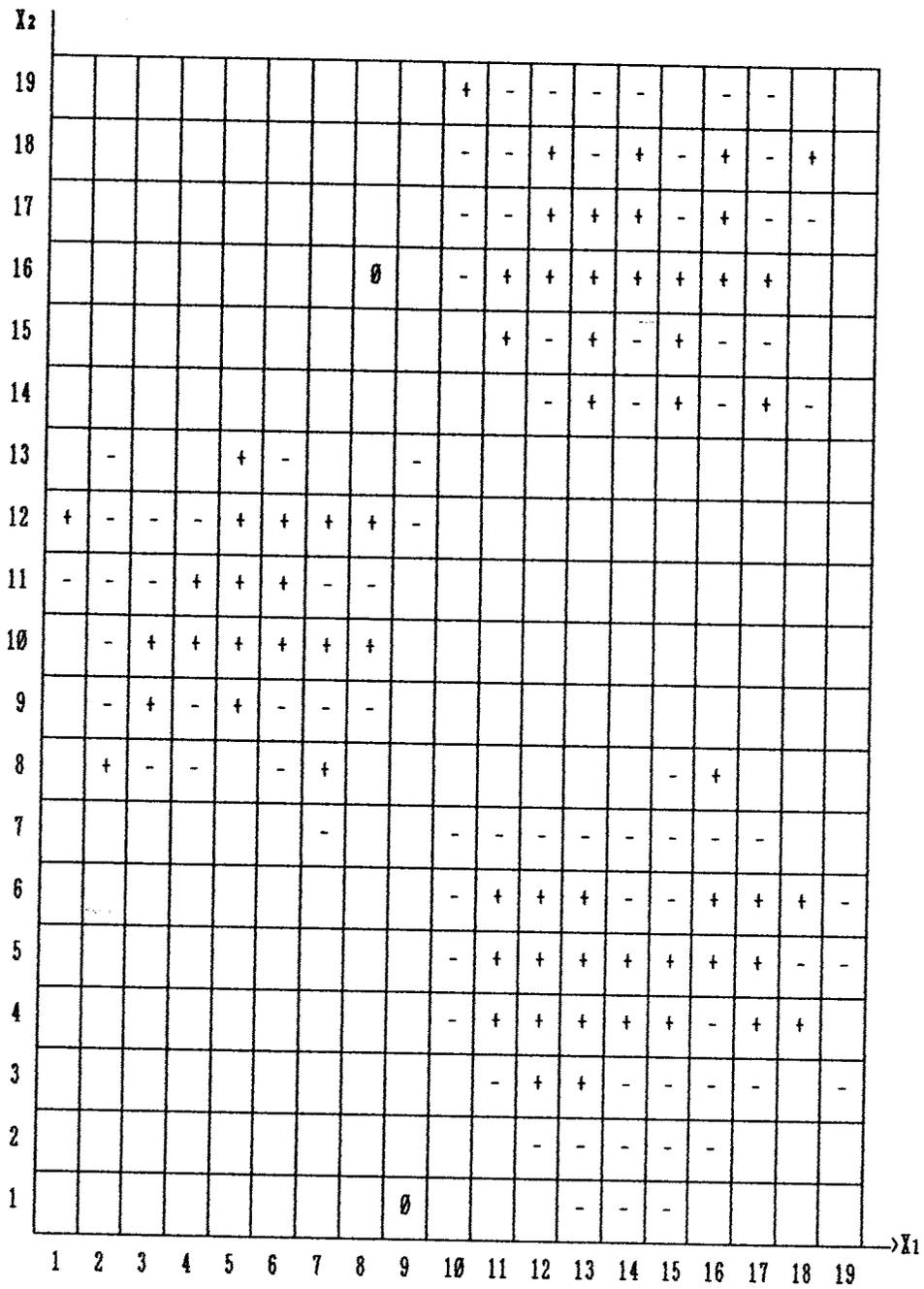


Figure 19 : Résultat du test de convexité appliqué à l'exemple 4.

Les figures 19-20-21 représentent différentes étapes de la procédure de classification.

19										.28	.39	.28	.39	.33		.33	.56		
18										.28	.33	.22	.56	.28	.50	.17	.39	.22	
17										.28	.44	.67	.89	.67	.89	.56	.72	.44	
16								.50		.56	.39	.78	.78	.78	.67	.44	.33		
15										.39	.83	.56	.89	.56	.78	.44			
14											.67	.28	.72	.28	.50	.22	.44		
13		.39			.44	.78		.44											
12	.22	.22	.22	.67	.72	.72	.56	.28	.39										
11	.33	.28	.33	.67	.89	.89	.89	.56											
10		.33	.33	.78	.78	.67	.33	.28											
9		.61	.33	.72	.39	.72	.50	.67											
8		.39	.39	.56		.39	.17								.33	.22			
7						.44			.39	.39	.61	.50	.22	.28	.39	.72			
6									.39	.33	.67	.56	.56	.44	.44	.50	.33	.39	
5									.61	.67	1	.89	.78	.67	.78	.78	.61	.44	
4									.44	.61	.89	.89	.78	.56	.67	.39	.33		
3										.61	.61	.56	.56	.22	.28	.39		.61	
2											.39	.28	.11	.06	.17				
1								.50				.22	.17	.22					

Figure 20 : Probabilités initiales de l'exemple 4.

19										0	0	0	0	0		0	0			
18										0	0	0	.03	.12	.01	0	0	0		
17										0	.01	.55	.99	.99	.91	.04	0	0		
16							.50			.20	.79	.99	1	1	.99	.71	.02			
15										.99	.99	1	1	1	.94	.15				
14											.99	.99	.99	.99	.78	.09	.02			
13	0				.99	.99		.02												
12	0	0	.01	.98	1	1	.99	.10	0											
11	0	0	.01	.99	1	1	.99	.43												
10	0	0	.94	1	.99	.68	.12													
9	0	0	.31	.98	.88	.15	.04													
8	0	0	.14		.08	0									0	0				
7						.02				.10	.70	.99	.99	.74	.02	0	.01			
6										.70	.99	1	1	.99	.97	.32	.05	0	0	
5										.99	1	1	1	1	1	.99	.13	0	0	
4										.99	1	1	1	.99	.99	.42	.06	0		
3											.99	.99	.99	.56	.22	0	0		.21	
2											.79	.45	0	0	0					
1							.50					0	0	0						
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19

Figure 21 : Probabilités à la dernière itération de l'exemple 4.

L'exemple 5 (figure 22) correspond à des classes non équiprobables présentant un chevauchement plus important que dans l'exemple 4.

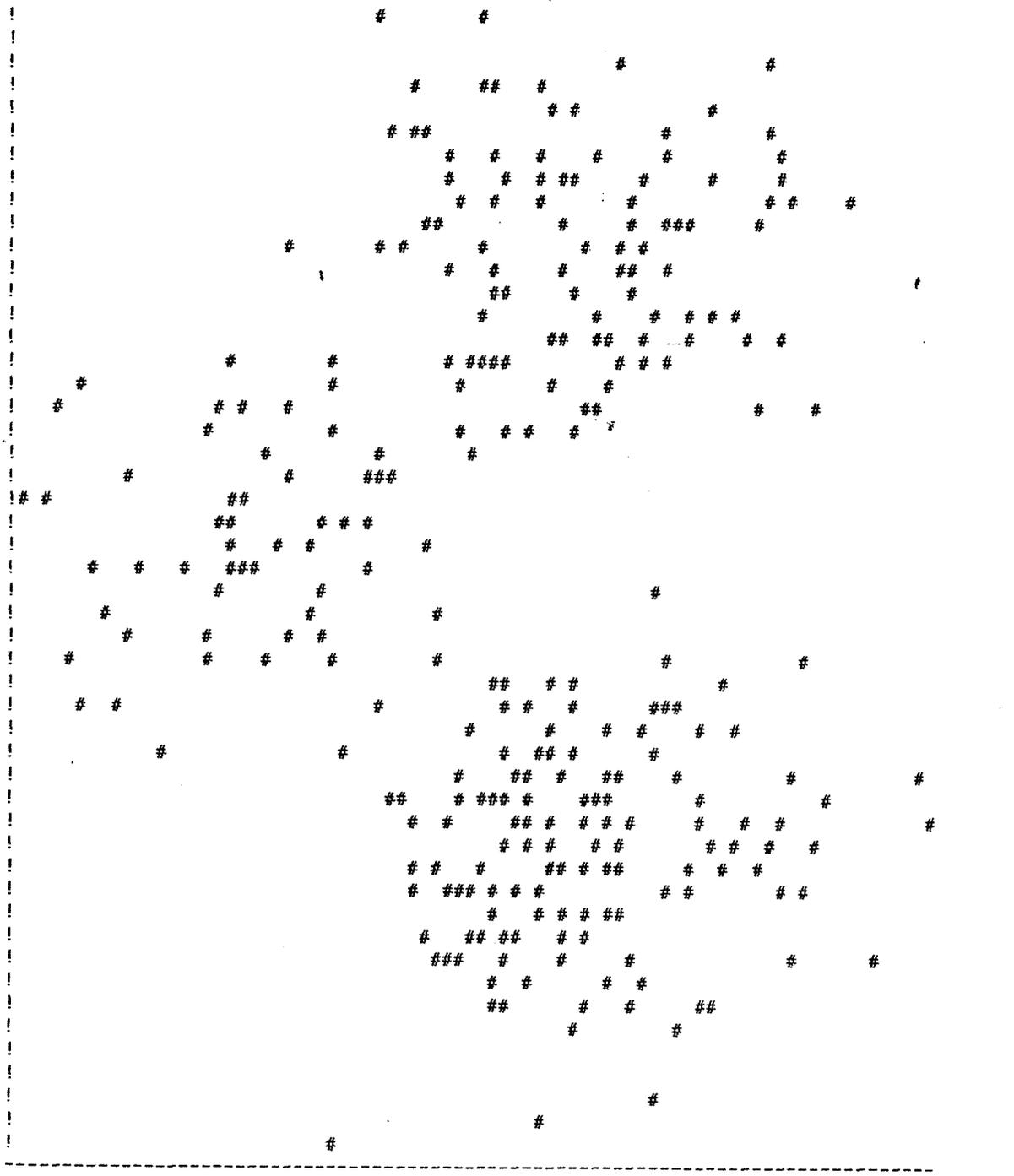


Figure 22 : Représentation graphique des 300 observations de l'exemple 5.

Ici également, la procédure d'étiquetage probabiliste améliore la détection des modes de la fonction de densité sous-jacente et le taux d'erreur réel reste encore très proche du taux d'erreur théorique minimum (cf tableau 3).

	VECTEURS MOYENNE		VARIANCES DES COMPOSANTES		PROBABILITES A PRIORI	
Classe 1	$\bar{X}_1 = \begin{bmatrix} 8 \\ 8 \end{bmatrix}$	$\bar{X}_1 = \begin{bmatrix} 8 \\ 7.8 \end{bmatrix}$	$\sigma_{X1} = 4$ $\sigma_{X2} = 4$	$\sigma_{X1} = 3,80$ $\sigma_{X2} = 4,01$	$P(C_1)=0.20$	$P(C_1)=0.20$
Classe 2	$\bar{X}_2 = \begin{bmatrix} 14 \\ 14 \end{bmatrix}$	$\bar{X}_2 = \begin{bmatrix} 14 \\ 14 \end{bmatrix}$	$\sigma_{X1} = 4$ $\sigma_{X2} = 4$	$\sigma_{X1} = 4,01$ $\sigma_{X2} = 4,10$	$P(C_2)=0.33$	$P(C_2)=0.34$
Classe 3	$\bar{X}_3 = \begin{bmatrix} 14 \\ 2 \end{bmatrix}$	$\bar{X}_3 = \begin{bmatrix} 14 \\ 2 \end{bmatrix}$	$\sigma_{X1} = 4$ $\sigma_{X2} = 4$	$\sigma_{X1} = 3,97$ $\sigma_{X2} = 3,95$	$P(C_3)=0.47$	$P(C_3)=0.46$
TAUX D'ERREUR THEORIQUE MINIMUM = 2 %			TAUX D'ERREUR DE LA CLASSIFICATION = 1 %			

Tableau 3 : Paramètres statistiques de l'exemple 5 et performances de la procédure de classification.

V.5 ANALYSE GLOBALE ET ANALYSE LOCALE.

L'exemple 2 présenté au paragraphe V.2 a en fait nécessité la mise en oeuvre de la procédure d'analyse globale puis de l'analyse locale.

Nous allons, dans ce paragraphe, montrer comment la procédure d'analyse locale a permis de mettre en évidence des groupements qui n'étaient pas détectés par la procédure d'analyse globale (cf chapitre IV.5).

V.5.1 Analyse globale.

La figure 23 indique les résultats du test de convexité de la fonction de densité sous-jacente.

16												-	-	+	
15									-	-	-	-	-	-	
14								+	-	+	-	-	-	-	
13					-	-	-	-	+	+	+	+	+	-	
12						+	-	+	+	+	+	+	+	-	+
11					+	-	-	-	-	+	+	+	+	+	-
10	0			-	-	+	-	+	-	+	-	-	-	-	-
9			-	+	+	+	-	+	-	-	-	-	-	-	
8		-	-	+	+	+	+	+	-				-		
7		+	+	+	+	+	+	+	-	-	-	-	-		
6			-	-	-	-	+	+	+	-	-	-	-	+	
5					-	-	+	-	+	+	+	+	-	+	-
4						-	+	-	+	+	+	+	+	+	-
3								-	-	+	+	+	+	-	-
2									+	+	-	+	-	+	-
1								-		-	-	-	-	-	

→X1

Figure 23 : Résultat du test de convexité appliqué à l'exemple 2.

Sur la figure 24, les probabilités initiales laissent à penser que les 3 classes peuvent être détectées. Cependant, à la fin du processus d'étiquetage itératif, (figure 25) 2 classes ont fusionné.

16												.28		.33	.28		
15									.44	.22	.22	.06	.17	.28			
14								.17	.39	.33	.56	.44	.22	.28			
13					.61		.61	.39	.78	.67	.78	.67	.33	.44			
12						.11	.28	.22	.67	.89	1	1	.67	.61	.22		
11				.33	.44	.39	.33	.67	.67	.78	.67	.67	.33	.33	.28		
10	.50			.61	.72	.33	.44	.11	.44	.22	.44	.33	.44	.28	.22	.28	
9		.39	.39	.78	.67	.89	.44	.61	.22	.28	.11	.11	.17				
8	.44	.72	.78	1	.89	.89	.56	.39					.17				
7	.33	.39	.44	.67	.78	.89	.78	.50	.22	.17	.11	.28					
6		.67	.44	.39	.67	.56	.78	.44	.44	.33	.22	.50	.28				
5				.17	.39	.33	.89	.56	.78	.67	.56	.78	.39	.61			
4					.44	.22	.50	.44	.89	1	.89	.78	.33	.28	.22		
3							.67	.56	.78	.78	.89	.67	.56	.17	.22		
2							.33	.28	.61	.44	.67	.22	.28	.17			
1							.67		.44	.28	.39	.28	.44				
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16

Figure 24 : Probabilités initiales de l'exemple 2.

16												.01		0	.01		
15									0	0	0	0	0	0	0		
14								0	0	.02	.09	.01	0	0			
13					.31		0	0	.21	.94	.99	.55	.02	0			
12						0	0	0	.86	1	1	1	.21	0	0		
11				.33	0	0	0	0	.21	.94	.96	.42	.02	0	0		
10	.50		.94	.85	.33	0	0	0	0	.02	.04	0	0	0	.01		
9		.69	.99	.99	.99	.89	.09	0	0	0	0	0	0				
8	.03	.18	.96	1	1	.99	.91	.10					0				
7	.01	0	.15	.98	.99	1	.99	.78	.01	0	0	0					
6		.01	.01	.23	.90	.99	1	.99	.69	.06	.01	0	0				
5				0	.06	.83	.99	1	.99	.98	.84	.26	.01	0			
4					.02	.27	.99	1	1	1	1	.97	.03	0	0		
3							.72	.99	1	1	.99	.77	.01	0	0		
2							.58	.58	.91	.91	.55	.04	0	0			
1							.30		.57	.45	.10	.01	0				
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16

Figure 25 : Probabilités à la dernière itération de l'exemple 2.

La figure 26 représente une classe repérée par le numéro 1 et constituée en fait des classes 1 et 3 du tableau 1, ainsi qu'une autre classe repérée par le numéro 2 qui correspond à la classe 2 du même tableau. On remarque que, du fait du poids très important du regroupement des classes 1 et 3 du tableau 1, la classe 2 se trouve amoindrie par le jeu des fonctions de décision qui privilégient les classes à forte probabilité a priori.

X ₂																
16											0		2	2		
15									2	0	0	0	2	0		
14								0	0	2	2	2	2	0		
13					0		1	2	2	2	2	2	2	2		
12						0	1	2	2	2	2	2	2	2	2	
11				0	1	1	1	1	1	2	2	2	2	2	0	
10	0			1	1	1	1	0	1	1	1	1	1	0	1	
9			1	1	1	1	1	1	0	0	1	0	1			
8		1	1	1	1	1	1	1				0				
7		1	1	1	1	1	1	1	0	0	1	1				
6			1	1	0	1	1	1	1	1	1	1	0			
5				1	1	1	1	1	1	1	1	1	1	1		
4					1	0	1	1	1	1	1	1	0	1		
3							1	1	1	1	1	1	0	1		
2							1	1	1	1	1	1	1			
1							1		1	1	1	1				
																X ₁
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16

Figure 26 : Répartition des classes de l'exemple 2 après l'analyse globale.

V.5.2 Analyse locale.

L'analyse locale de deux groupements obtenus par l'analyse globale va permettre de retrouver la structure de l'échantillon.

Le résultat de l'analyse de la convexité de la fonction de densité de probabilité associée au groupement numéro 1 résultant de l'analyse globale est représenté figure 27.

X ₂																	
16						-		+									
15								+	-								
14								-	-	+							
13						+	-	-	-	-	+						
12	0			-	+	+	-	+	-	+	-	-	-	+		0	
11			-	-	+	+	+	+	-	+		+	+				
10		+	-	+	+	+	+	+	-			-		-			
9		+	-	+	-	+	+	+	-			-	-				
8			+	+	-	-	+	-	-	-	-	-	-	-			
7					-	-	-	+	+	+	-	-	-	-			
6						+	-	+	-	+	-	+	+	-	+	-	
5							-	+	-	-	+	+	+	+	+	+	
4								-	-	-	+	+	+	+	-	-	
3									-	+	+	+	-	+	-	-	
2										+	-	-	-	+	-	-	
1									-			+	-	+	+		
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
																	X ₁

Figure 27 : Résultat du test de convexité appliqué au groupement 1 détecté par l'analyse globale de l'exemple 2.

Cette analyse fine de la convexité ne permet pas encore de mettre en évidence les deux classes constituant le groupement numéro 1. Par contre, à l'issue de la procédure itérative d'étiquetage probabiliste (cf figure 28), la vallée séparant les deux classes 1 et 3 du tableau 1 apparaissent sans aucune ambiguïté.

X ₂																
16						.31		.10								
15							0	0								
14							0	0	0							
13						.97	.86	.04	0	0	0					
12	.50			.96	.99	.99	.85	.22	0	0	0	.02	.05	.21	.50	
11			.45	.96	1	1	.99	.99	.43	.15		0	0			
10		.02	.26	.96	1	1	1	1	.98			.01		.09		
9		.04	.06	.37	.94	.99	.99	.99	.87			0	0			
8			.06	.04	.03	.44	.85	.78	.15	0	0	0	0	0		
7				0	0	0	.01	.01	0	0	0	.01	0	0		
6					0	0	0	0	0	.13	.78	.85	.21	.01	0	
5						0	0	0	.01	.85	1	1	.98	.05	0	0
4							0	0	.18	.99	1	1	.91	.02	0	0
3								.03	.63	.99	.99	.99	.12	0	0	
2									.89	.98	.84	.16	0	0	0	
1								.77			.27	.02	0	0		
																X ₁
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16

Figure 28 : Probabilités à la dernière itération pour le groupement 1 de l'exemple 2.

V.6 CLASSIFICATION DE DONNEES TRIDIMENSIONNELLES.

En projetant les données sur des plans, on peut représenter des échantillons tridimensionnels. La figure 29 représente les projections d'un ensemble de trois classes d'observations tridimensionnelles sur les plans définis par les vecteurs de base de l'espace de représentation pris 2 à 2.

Le tableau 4 indique les caractéristiques statistiques de cet échantillon.

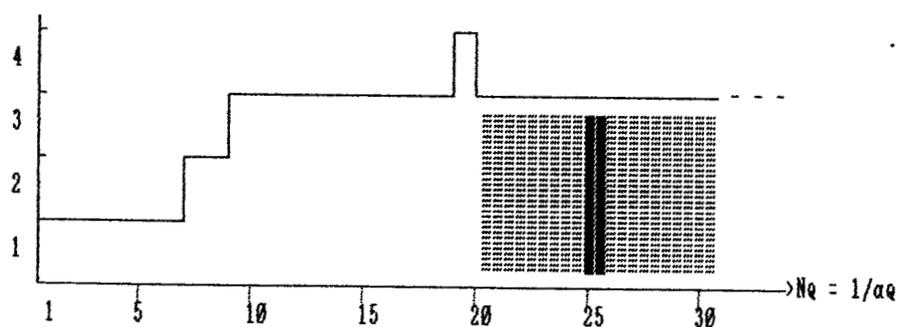
	VECTEURS MOYENNE		VARIANCES DES COMPOSANTES		PROBABILITES A PRIORI	
Classe 1	$\bar{X}_1 = \begin{bmatrix} 20 \\ 20 \\ 20 \end{bmatrix}$	$\bar{X}_1 = \begin{bmatrix} 20 \\ 20 \\ 20 \end{bmatrix}$	$\sigma_{X1} = 2$ $\sigma_{X1} = 2$ $\sigma_{X1} = 2$	$\sigma_{X1} = 1,97$ $\sigma_{X1} = 2,00$ $\sigma_{X1} = 1,99$	$P(C_1)=0.33$	$P(C_1)=0.33$
Classe 2	$\bar{X}_2 = \begin{bmatrix} 26 \\ 26 \\ 20 \end{bmatrix}$	$\bar{X}_2 = \begin{bmatrix} 26 \\ 26 \\ 20 \end{bmatrix}$	$\sigma_{X1} = 2$ $\sigma_{X1} = 2$ $\sigma_{X1} = 2$	$\sigma_{X1} = 1,96$ $\sigma_{X1} = 1,97$ $\sigma_{X1} = 2,02$	$P(C_2)=0.33$	$P(C_2)=0.33$
Classe 3	$\bar{X}_3 = \begin{bmatrix} 20 \\ 26 \\ 26 \end{bmatrix}$	$\bar{X}_3 = \begin{bmatrix} 20 \\ 26 \\ 26 \end{bmatrix}$	$\sigma_{X1} = 2$ $\sigma_{X1} = 2$ $\sigma_{X1} = 2$	$\sigma_{X1} = 2,00$ $\sigma_{X1} = 2,01$ $\sigma_{X1} = 1,98$	$P(C_3)=0.33$	$P(C_3)=0.33$
TAUX D'ERREUR THEORIQUE MINIMUM = 0,01 % TAUX D'ERREUR DE LA CLASSIFICATION = 0 %						

Tableau 4 : Caractéristiques statistiques de l'échantillon tridimensionnel.

La figure 30 représente le nombre de modes détectés en fonction du pas de discrétisation pour différents seuils de rejet.

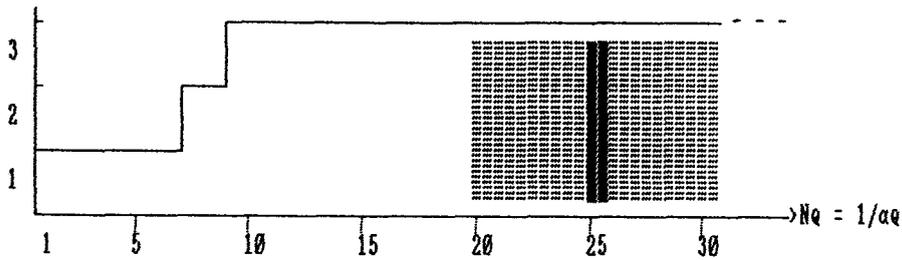
Seuil de rejet fixé à un minimum de 50 observations par domaine modal :

Nb de modes



Seuil de rejet fixé à un minimum de 60 observations par domaine modal :

Nb de modes



Seuil de rejet fixé à un minimum de 75 observations par domaine modal :

Nb de modes

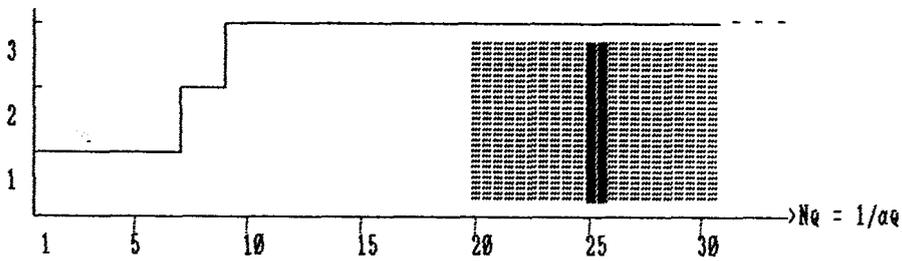


Figure 30 : Nombre de modes détectés en fonction du pas de discrétisation pour différentes valeurs du seuil de rejet.

On constate que la plage de variation de N_q commune à différents réglages du seuil de rejet correspond à bonne une stabilité du nombre de modes détectés.

En ajustant N_q à la valeur 25, on obtient un taux d'erreur qui reste très voisin du taux d'erreur théorique minimum.

V.7 CLASSIFICATION DE DONNEES QUADRIDIMENSIONNELLES.

Le dernier exemple artificiel utilisé pour tester la procédure de classification est un ensemble de trois classes d'observations quadridimensionnelles. Les caractéristiques de celles-ci sont indiquées sur le tableau 5.

	VECTEURS MOYENNE		VARIANCES DES COMPOSANTES		PROBABILITES A PRIORI
Classe 1	$\bar{X}_1 = \begin{bmatrix} 14 \\ 14 \\ 14 \\ 14 \end{bmatrix}$	$\bar{X}_1 = \begin{bmatrix} 14 \\ 14 \\ 14 \\ 14 \end{bmatrix}$	$\sigma_{X1} = 2$ $\sigma_{X2} = 2$ $\sigma_{X3} = 2$ $\sigma_{X4} = 2$	$\sigma_{X1} = 1,96$ $\sigma_{X2} = 2,01$ $\sigma_{X3} = 1,98$ $\sigma_{X4} = 2,01$	$P(C_1)=0.33$ $P(C_1)=0.33$
Classe 2	$\bar{X}_2 = \begin{bmatrix} 14 \\ 14 \\ 2 \\ 2 \end{bmatrix}$	$\bar{X}_2 = \begin{bmatrix} 14 \\ 14 \\ 2 \\ 2 \end{bmatrix}$	$\sigma_{X1} = 2$ $\sigma_{X2} = 2$ $\sigma_{X3} = 2$ $\sigma_{X4} = 2$	$\sigma_{X1} = 1,99$ $\sigma_{X2} = 2,00$ $\sigma_{X3} = 2,02$ $\sigma_{X4} = 1,97$	$P(C_2)=0.33$ $P(C_2)=0.33$
Classe 3	$\bar{X}_3 = \begin{bmatrix} 2 \\ 2 \\ 14 \\ 14 \end{bmatrix}$	$\bar{X}_3 = \begin{bmatrix} 2 \\ 2 \\ 14 \\ 14 \end{bmatrix}$	$\sigma_{X1} = 2$ $\sigma_{X2} = 2$ $\sigma_{X3} = 2$ $\sigma_{X4} = 2$	$\sigma_{X1} = 1,98$ $\sigma_{X2} = 2,01$ $\sigma_{X3} = 2,00$ $\sigma_{X4} = 2,01$	$P(C_3)=0.33$ $P(C_3)=0.33$
TAUX D'ERREUR THEORIQUE MINIMUM = 1,01 % TAUX D'ERREUR DE LA CLASSIFICATION = 0 %					

Tableau 5 : Caractéristiques statistiques de l'échantillon quadridimensionnel.

En représentant le nombre de modes détectés en fonction du pas de discrétisation pour différentes valeurs du seuil de rejet, une large plage de stabilité du nombre de modes est mise en évidence. Le milieu de cette plage correspond à une valeur du pas de discrétisation égale à 1/16. A cette valeur,

le résultat de la procédure de classification est très satisfaisant, le taux d'erreur étant nul.

REFERENCES (chapitre V)

- /1/ J. G. POSTAIRE, *Optimisation du processus de classification automatique par analyse de la convexité des fonctions de densité de probabilité*, Thèse d'état, Université de Lille, 1981.
- /2/ J. G. POSTAIRE, *Fonctions convexes et optimisation du processus de classification automatique. I. Identification des mélanges gaussiens par estimation de la convexité des fonctions de densité multivariées*, R.A.I.R.O. Automatique/Systems Analysis and Control, Vol. 16, n° 4, pp. 357-379, 1982.
- /3/ J. G. POSTAIRE, *Fonctions convexes et optimisation du processus de classification automatique. II. Identification de mélanges gaussiens à partir de petits échantillons par estimation de la convexité des fonctions de densité marginales*, R.A.I.R.O. Automatique/Systems Analysis and Control, Vol. 17, n° 1, pp. 39-59, 1983.

CONCLUSION

CONCLUSION GENERALE.

Nous avons montré comment l'utilisation des techniques d'étiquetage itératif permet d'améliorer les procédures de classification automatique basées sur l'analyse de la convexité des fonctions de densité de probabilité.

En effet, ces techniques d'analyse de données sont très sensibles aux bruits qui perturbent les observations à classer. Si elles ont conduit, par le passé, à des procédures originales permettant d'optimiser les résultats de la classification dans un environnement non supervisé, les résultats s'étaient révélés trop sensibles à l'ajustement des paramètres de réglage des algorithmes.

Ces techniques sont basées sur un test qui détermine la convexité locale des fonctions de densité sous-jacentes à la distribution des observations à classer. La prise en compte du voisinage de chaque point du test permet d'affirmer ou d'infirmer les résultats du test selon leurs cohérences avec les résultats obtenus dans ce voisinage.

Les résultats présentés, sous l'hypothèse paramétrique et sous l'hypothèse d'indépendance statistique des composantes des observations, permettent d'accéder à des procédures de classification optimale extrêmement robustes par rapport aux réglages des paramètres des algorithmes.

Partant d'un échantillon totalement inconnu, il est possible, sous l'hypothèse normale, de déterminer le nombre de classes en présence. Pour chacune d'elles, des valeurs approchées du vecteur-moyenne, de la matrice de covariance et de la probabilité à priori conduisent finalement à l'implantation d'une stratégie de classification optimale au sens de la théorie de la décision.

Le fait de retrouver, sans ambiguïté et avec une bonne précision, la structure d'un mélange à partir de la seule information apportée par les observations à classer elles-mêmes, permet d'assurer que les procédures proposées peuvent être appliquées sous l'hypothèse non paramétrique, c'est-à-dire sans modèle analytique pour les distributions en

présence. Dans de telles situations, on ne dispose pas de moyens objectifs pour juger les résultats de la classification. C'est la raison pour laquelle nous n'avons pas présenté de simulations sous cette hypothèse.

En présentant les résultats obtenus sur différents échantillons, nous avons montré la robustesse de l'algorithme par rapport à l'ajustement du pas de discrétisation de l'espace de représentation des données et du nombre minimum d'observations par classes.

Les diagrammes présentés guident, sans ambiguïté, l'analyste vers le bon réglage de ces paramètres.

Il serait intéressant de proposer, à terme, des procédures permettant d'analyser de tels diagrammes pour obtenir un réglage automatique de ces paramètres. De telles procédures devraient intégrer l'expertise de l'analyste. On touche ici au domaine de l'intelligence artificielle.

Les nouveaux coefficients de compatibilité proposés dans la procédure d'étiquetage itératif probabilistes devront également être adaptés à d'autres procédures pouvant bénéficier de cette technique.

Il serait intéressant de les utiliser pour l'extraction des contours de fonction de densité de probabilité qui a conduit des chercheurs du Centre d'Automatique à proposer des procédures de classification automatique paramétriques et non paramétriques dans un environnement non supervisé. /1-2/

Une nouvelle procédure de détection directe des modes par étiquetage itératif, également développé au Centre d'Automatique /3-4/ pourrait bénéficier de ces nouveaux coefficients.

Enfin, le traitement des images numériques, qui est à l'origine du concept d'étiquetage probabiliste itératif, pourrait également tirer parti de ces coefficients pour l'extraction des régions et des contours. La boucle serait ainsi fermée, puisque la classification automatique, qui a

emprunté certains concepts à l'analyse des images, reviendrait
apporter à ce domaine de nouveaux outils.

REFERENCES

- /1/ J. G. POSTAIRE et A. TOUZANI, *Mode boundary detection by relaxation with application to cluster analysis*, Pattern Recognition, (à paraître) 1988.
- /2/ J. G. POSTAIRE et A. TOUZANI, *Détection des modes par recherche des contours en classification automatique*, A.P.I.I., (à paraître) 1988.
- /3/ A. TOUZANI et J. G. POSTAIRE, *Clustering by mode boundary detection*, Pattern Recognition Letters, (à paraître) 1988.
- /4/ A. TOUZANI et J. G. POSTAIRE, *Mode detection by relaxation*, I.E.E.E. Trans on PAMI, (à paraître) 1988.