N° d'ordre : 1031

présentée à

L'UNIVERSITE DES SCIENCES ET TECHNOLOGIES DE LILLE

Pour obtenir le titre de

DOCTEUR

en Productique : Automatique et Informatique Industrielle

par

Eric CZESNALOWICZ

BIBLIOTHEQUE INTERUNIVERSITAIRE
SECTION DE SCIENCES
LILLE

APPLICATIONS DE L'ESTIMATEUR NON PARAMETRIQUE DES K PLUS PROCHES VOISINS EN CLASSIFICATION AUTOMATIQUE MULTIDIMENSIONNELLE

Soutenue le 16 décembre 1992 devant la commission d'examen :

MM.

P. Vidal	Président	Professeur à l'USTL
J.-G. Postaire	Directeur de recherche	Professeur à l'USTL
M. Bourton	Rapporteur	Professeur à L'Université de Valenciennes
H. Emptoz	Rapporteur	Professeur à l'INSA de Lyon
C. Vasseur	Examinateur	Professeur à l'USTL, Directeur de l'ENSAIT

AVANT PROPOS

Le travail présenté dans ce mémoire a été effectué au Centre d'Automatique de l'Université des Sciences et Technologies de Lille, dirigé par Monsieur le Professeur Pierre Vidal. Je le remercie de l'accueil qu'il m'a réservé au sein de son laboratoire et de l'honneur qu'il me fait en acceptant la présidence du jury de thèse.

J'adresse mes sincères remerciements à Monsieur Jack-Gérard Postaire, Professeur à l'Université des Sciences et Technologies de Lille pour son excellent encadrement tout au long de mes travaux, Je tiens à lui exprimer toute ma gratitude pour son dynamisme et ses conseils permanents.

Que Monsieur Michel Bourton, Professeur à l'Université de Valenciennes et du Hainaut-Cambrésis, trouve ici toute ma reconnaissance pour avoir accepté de juger mon travail.

J'exprime également mes sincères remerciements à Monsieur Hubert Emptoz, Professeur à l'INSA de Lyon, pour l'intérêt qu'il a porté à ce travail en acceptant d'être rapporteur de cette thèse.

Que Monsieur Christian Vasseur, Professeur à l'Université des Sciences et Technologies de Lille et Directeur de l'ENSAIT de Roubaix, trouve ici l'expression de ma considération pour sa participation au jury de thèse.

Je ne saurais terminer cet avant-propos sans adresser mes remerciements les plus sincères à tout ceux qui, de loin ou de près, m'ont aidé par leur compétence et leur amitié dans l'élaboration de ce travail.

SOMMAIRE

CHAPITRE I
LA CLASSIFICATION AUTOMATIQUE

CHAPITRE II
ESTIMATION DE LA FONCTION DE DENSITE DE PROBABILITE

CHAPITRE III

DETECTION DES MODES DE LA FONCTION DE DENSITE DE PROBABILITE

CHAPITRE IV

DETECTION DES CONTOURS DES MODES

CHAPITRE V

RESULTATS EXPERIMENTAUX

CHAPITRE I

LA CLASSIFICATION AUTOMATIQUE

CHAPITRE I

LA CLASSIFICATION AUTOMATIQUE

DE L'OBSERVATION DE L'ENVIRONNEMENT...

De tout temps, l'homme a cherché à connaître le milieu dans lequel il vit. Il essaie de comprendre les différents mécanismes qui régissent le fonctionnement de la nature. Cette compréhension, même si elle n'est que partielle, permet néanmoins de prévoir, par exemple, les phénomènes météorologiques à plus ou moins long terme. Bien entendu, il existe bien d'autres exemples.

L'exploitation des cultures nécessite également un effort de connaissance des principes de développement des plantes, ceci dans un but d'optimisation des récoltes, tant sur le plan du rendement que sur le plan financier.

Mais le milieu naturel n'est pas le seul sujet de préoccupation de l'homme. L'industrie nécessite également un investissement important en connaissance afin d'optimiser les processus de fabrication. Depuis des temps plus proches, les sciences humaines cherchent à définir un comportement sous-jacent, et donc une connaissance.

L'analyse de tous ces phénomènes, naturels ou non, demande de prendre en compte des données multiples si on veut en tirer des résultats exploitables et précis. Ici apparaît immédiatement une limitation de l'homme dans son aptitude à

gérer des masses de données. En effet, si on conçoit qu'il est assez facile d'extraire une structure d'un petit nombre de données sur lesquelles on connaît un faible nombre de caractéristiques, il est beaucoup moins concevable de dépouiller une multitude de données sans le concours d'un calculateur.

Lorsqu'on veut réaliser une étude statistique des données, on cherche à structurer celles-ci en groupes, ou classes. Plus concrètement, si on considère qu'à chaque donnée on associe un vecteur dont chaque composante en est une caractéristique, ou attribut, on peut représenter les données dans un repère euclidien où, à chaque donnée, correspond un point de l'espace de représentation. L'analyse visuelle de cette représentation peut être envisagée lorsqu'il s'agit de données définies par deux ou trois caractéristiques, donc représentables dans un repère bidimensionnel ou tridimensionnel. Néanmoins, il est très fréquent que la description des données nécessite la prise en compte d'un nombre plus important d'attributs. On se heurte alors à un problème de représentation car de telles données multidimensionnelles ne sont pas assimilables par l'homme. Des techniques automatiques sont donc nécessaire pour mener à bien tout processus de classification qui fera émerger une structure sous-jacente aux données.

...A LA CLASSIFICATION AUTOMATIQUE MULTIDIMENSIONNELLE.

Les différentes techniques de classification automatique visent toutes un même objectif. On cherche à décomposer un échantillon de données en classes, de telle sorte que les données d'une même classe soient plus semblables que celles appartenant à des classes différentes.

Sans toutefois prétendre passer en revue de manière exhaustive toutes les méthodes de classification automatique existantes, nous pouvons néanmoins distinguer, en première approche, deux grandes catégories parmi les méthodes de classification, d'une part les méthodes métriques, d'autre part les méthodes statistiques.

I. - LES METHODES METRIQUES

Les méthodes métriques font appel à des notions de similarité entre les individus d'une même classe [SOK63][BAL65]. Ces techniques cherchent en général à optimiser un critère qui maximise la dispersion interclasse tout en minimisant la dispersion intra-classe [FRI67][JON68][FUK70][BAL67][MAC67][JAI88].

Parmi ces méthodes métriques, nous avons les méthodes hiérarchiques. Ce moyen de classification des objets d'un échantillon est une approche basée sur le partitionnement de l'ensemble des observations disponibles. Dans le mode ascendant, cette approche consiste à considérer chaque observation comme formant une classe et ensuite, à chaque étape, à regrouper les classes deux à deux en maximisant un critère de similarité. A l'inverse, dans le mode descendant, elle consiste à considérer l'ensemble des observations comme constituant une seule classe et à procéder à des divisions successives en maximisant un critère de dissimilarité pour diviser chaque classe en deux. Le processus de division ou d'agrégation est terminé lorsqu'on atteint un critère prédéfini, qui peut être, par exemple, le nombre de classes désirées [LAN67][LUK79][BAY88].

On peut rechercher les classes en assignants les observations à des "noyaux" en cherchant à minimiser un critère de distance entre les noyaux et les observations. Les noyaux peuvent être considérés comme des observations représentatives de chaque classe, ou plus simplement les centres de gravité des classes initiales. Ces techniques sont utilisées dans l'algorithme Isodata [BAL67][FOR74] ou l'algorithme des nuées dynamiques proposé par Diday [DID71][DID82]. Dans cet algorithme, les classes initiales, ou plus précisément les noyaux qui les représentent, peuvent être fixés par l'analyste qui recherche une certaine structure dans l'échantillon qui lui est soumis. Elles peuvent être également issues d'une première phase de classification basée par exemple sur l'exploitation de la méthode hiérarchique. La méthode des nuées dynamiques consiste alors, tout en minimisant un critère, à recalculer de nouveaux noyaux à chaque itération du processus. L'algorithme est terminé lorsqu'un critère de stabilité des noyaux est atteint, ou lorsque le nombre de classes que l'on s'est proposé d'obtenir est trouvé.

Un axe de la recherche en analyse de données s'est orienté vers l'exploitation des réseaux neuromimétiques [LIP89]. Dans ces méthodes, on part d'un ensemble d'apprentissage qui permet de déterminer les paramètres du réseau de neurones, et ainsi obtenir une représentation des informations les plus significatives de l'échantillon analysé. Le réseau ainsi défini permet la classification des données après cette phase d'apprentissage qui revient à ajuster les poids des connexions en fonction de la structure des données.

On peut encore, parmi les méthodes métriques de classification, exploiter les notions de morphologie mathématique. Dans cette approche, on essaie de déterminer une classification des données en recherchant une certaine structure géométrique des classes en présence dans l'échantillon.

Proposée par C. Botte-Lecocq [BOT91], cette technique est appliquée en discrétisant l'espace de représentation des données par une grille régulière. Un élément structurant est alors mis en correspondance avec les différents hypercubes de cette grille, où on ne prend en compte que la présence ou non d'observations dans chacun de ces hypercubes. Ce traitement "tout ou rien" est mis à profit pour mettre en oeuvre une technique de classification faisant appel à la morphologie binaire.

I. - LES METHODES STATISTIQUES

L'ensemble des méthodes statistiques constitue un autre courant de la classification automatique multidimensionnelle . Ces méthodes font en général appel à l'analyse de la fonction de densité de probabilité sous-jacente à la distribution des observations disponibles. On admet alors qu'à chaque mode correspond une classe, ce qui permet de les rechercher par détection des modes de cette fonction [ASS89].

Dans cette approche statistique de la classification, on peut distinguer les méthodes paramétriques, qui font appel à un modèle probabiliste de la structure des données de l'échantillon analysé, et les méthodes non paramétriques, qui ne font appel à aucun modèle.

II. 1. - APPROCHE PARAMETRIQUE

Il existe un mode de classification supervisée, où une première étape consiste en un apprentissage. Celui-ci permet d'extraire de l'échantillon analysé des prototypes des classes en présence. L'étape suivante consiste alors à établir une classification des observations en se référant à ces prototypes.

Dans le mode non supervisé, il n'existe pas de prototypes des classes. On cherche seulement à respecter un modèle probabiliste pour classifier les données d'un échantillon. La fonction de densité de probabilité de l'échantillon peut être considérée comme un ensemble de fonction de densité correspondant à chaque classe en présence dans l'échantillon analysé. Daly [DAL62], puis Hillborn [HIL68] donnent une formulation Bayessienne de l'apprentissage des paramètres d'un mélange. Des résultats semblables sont obtenus en utilisant des techniques d'estimation par maximum de vraisemblance [HAS66][DAY69]. Ces techniques exigent néanmoins des hypothèses souvent restrictives comme la connaissance du nombre de classes ou leurs probabilités *a priori* [SCH76][MAK77][WOL70].

II. 2. - APPROCHE NON PARAMETRIQUE

L'approche non paramétrique se distingue de la méthode paramétrique dans le sens où on ne fait appel à aucun modèle pour effectuer la classification. Cette approche ne nécessite aucune connaissance *a priori* sur la structure de la distribution des données à analyser. La seule information que l'on possède dans ce cas est celle que l'on peut extraire des données elles-mêmes.

Dans une approche totalement non supervisé, on recherche les modes de la fonction de densité de probabilité sous-jacente à la distribution des observations disponibles. La fonction de densité de probabilité peut être estimée par la méthode de Parzen-Rosenblatt [PAR62] pour laquelle on se fixe un domaine d'estimation. Cette fonction peut aussi être estimée par la méthode de Cover et Hart [COV67] qui consiste, à l'opposé de la précédante, à fixer un nombre de voisins au point où on désire estimer la fonction de densité de probabilité et à rechercher un domaine qui englobe ces voisins.

On peut détecter les modes en remontant les pentes de la fonction de densité de probabilité [KOO76]. Une autre technique consiste à calculer directement le gradient à partir des observations [FUK75].

D'autres auteurs analysent la convexité de la fonction de densité de probabilité [VAS80][POS82b]. L'approche proposée par J.-G. Postaire [POS82b] est fondée sur la recherche des régions convexes de la fonction de densité de probabilité estimée, par la méthode du noyau de Parzen-Rosenblatt, sur une grille hypercubique régulière. Grâce à un test de convexité chaque hypercube est étiqueté "convexe" ou "concave". Les zones convexes sont alors assimilées aux modes de la fonction de densité de probabilité. S. Olejnik [OLE88] améliore cet étiquetage grâce à une procédure itérative d'étiquetage probabiliste qui permet de corriger les erreurs résultant de l'application du test de convexité. En effet, l'estimation de la fonction de densité de probabilité par la méthode du noyau est sensible aux irrégularités de la distribution des observations disponibles, ce qui a pour effet d'attribuer des étiquettes "convexe" à des hypercubes entouré d'hypercubes situés dans des régions concaves et inversement.

Plutôt que de chercher à détecter les modes comme des maxima locaux de la fonction de densité de probabilité sous-jacente, A. Touzani recherche leurs contours [TOU87]. Une procédure de détection des contours basée sur l'utilisation d'opérateurs différentiels permet d'attribuer à chaque hypercube d'un réseau de discrétisation défini sur tout l'espace de représentation des données, soit l'étiquette "contour", soit l'étiquette "non-contour". Les irrégularités de la réponse de l'opérateur différentiel amènent des erreurs d'étiquetage qui rendent parfois le suivi des contours problématique. Une procédure itérative d'étiquetage probabiliste est alors appliquée afin d'éliminer les erreurs d'étiquetages initiaux. Une autre approche consiste à appliquer une procédure de relaxation directement sur la fonction de densité de probabilité estimée pour faire ressortir des modes de cette fonction [TOU88].

Ce bref rappel bibliographique montre que la plupart des méthodes de détection des modes existantes font appel à une technique de discrétisation de

l'espace de représentation des données sous la forme d'un réseau régulier hypercubique.

En fait, de nombreuses procédures tirent profit d'une technique d'estimation rapide de la fonction de densité de probabilité sous-jacente à une distribution d'observations proposée par J.-G. Postaire et C. Vasseur [POS82]. En effet, l'un des inconvénients majeurs de ce type de discrétisation est l'explosion exponentielle du nombre d'hypercubes nécessaires à cette discrétisation dès que la dimension des données s'élève.

En proposant une technique d'estimation basée sur la méthode du noyau [PAR62], dont le coût en calcul est proportionnel au nombre d'observations disponibles, J.-G. Postaire et C. Vasseur ont redonné un nouvel essor à cette technique classique et favorisé toute une série de travaux utilisant cette discrétisation particulière.

C'est ainsi que l'approche morphologique [BOT91], l'approche par analyse de convexité [POS82], la recherche des contours des modes [TOU87] et les techniques d'étiquetage probabiliste itératif [OLE88] s'appuient, soit explicitement, soit implicitement, sur l'existence de ce réseau d'hypercubes adjacents qui forment un maillage sur tout l'espace, de manière assez comparable aux réseaux de pixels sur lesquels on définit une image numérique.

C'est d'ailleurs ce parallélisme qui a été à l'origine de l'adaptation de nombreux outils de l'analyse des images à la classification automatique.

Dans le travail présenté dans ce mémoire, on s'interroge sur les possibilités de s'affranchir de ce maillage, et par la même occasion, de l'estimateur de Parzen-Rosenblatt, dont on connaît les limitations. En effet, dans le cas d'échantillons constitués de classes de probabilité *a priori* très différentes, l'ajustement de la taille de la fenêtre pose un problème épineux, quasiment impossible à résoudre. En effet, une fenêtre de taille adaptée à une certaine concentration d'observations dans une région de l'espace de représentation des données, ne l'est pas pour une concentration différente dans une autre région du même espace.

C'est donc vers l'exploitation de l'estimateur des k plus proches voisins, calculé non pas aux centres des hypercubes d'un réseau de discrétisation régulier, mais uniquement au niveau de chacune des observations disponibles que nous nous dirigeons.

Nous nous proposons de reprendre différents travaux sur la détection des modes, en suivant cette nouvelle approche qui vise à remplacer la notion de voisinage fixe défini sur une grille hypercubique par un voisinage de taille variable, défini par le domaine qui contient les k plus proches voisins de chaque point où la fonction de densité de probabilité sous-jacente est estimée.

III. - EXPLOITATION DE L'ESTIMATEUR DES K PLUS PROCHES VOISINS

Dans le chapitre II, après avoir brièvement rappeler le principe d'estimation de la fonction de densité de probabilité sous-jacente à la distribution des observations disponibles, nous montrons que le réglage du paramètre k de voisins à prendre en compte dans l'estimation de la fonction de densité de probabilité par la méthode des k plus proches voisins est beaucoup moins critique que ne l'est le réglage de la taille du noyau, dans la méthode du noyau de Parzen-Rosenblatt.

La décomposition de l'algorithme d'estimation en deux étapes permet un gain de temps appréciable tant au niveau de l'estimation elle-même, qu'au niveau des différents traitements qui suivent la phase d'estimation de la fonction de densité de probabilité pour mener à la classification finale.

La recherche des noyaux des modes de la fonction de densité de probabilité est, dans le chapitre III, réalisé par l'intermédiaire d'un filtre non linéaire sur cette fonction. Une technique d'étiquetage itérative exploitant la notion de voisinage de taille variable permet d'attribuer à chaque noyau des modes une étiquette différente. La classification des observations consiste alors à assigner celles-ci au noyau le plus proche. Cette assignation est réalisée par agrégation itérative des observations autour des noyaux précédemment étiquetés.

Le chapitre IV présente une autre approche au problème de détection des modes. On recherche non plus le noyau des modes mais leur contour. Un nouvel opérateur de détection de contour permet de repérer les lieux où la fonction de densité de probabilité présente les plus fortes variations locales. Cette détection est suivie d'un algorithme d'étiquetage des contours qui attribue une étiquette différente à chaque contour, puis les observations sont classées par assignation au contour le plus proche.

Pour illustrer la technique de classification proposée, on applique cette méthode, dans le chapitre V, sur différents échantillons bidimensionnels et tridimensionnels générés artificiellement. Afin d'évaluer les résultats, la classification est comparée à celle obtenue par l'algorithme Isodata. Sans aucune connaissance a priori sur les données à analyser, le taux d'erreur de classification reste très faible, en particulier pour les mélanges non gaussiens, surtout si on le compare à celui obtenu par la méthode classique Isodata.

CHAPITRE II

ESTIMATION DE LA FONCTION

DE DENSITE DE PROBABILITE

CHAPITRE II

ESTIMATION DE LA FONCTION DE DENSITE DE PROBABILITE

I. - METHODES CLASSIQUES D'ESTIMATION NON PARAMETRIQUE

A chaque élément de l'échantillon disponible est associé un vecteur $X=(x_1,...,x_i,...,x_n)$ dont les composantes correspondent aux divers attributs utilisés pour caractériser l'élément considéré. Chaque observation multidimensionnelle ainsi définie est représentée par un point dans un espace euclidien dont chaque axe correspond à une composante du vecteur X. En classification automatique classique, le premier pas des procédures non paramétriques consiste à estimer la fonction de densité de probabilité sous-jacente à la distribution des observations disponibles. C'est l'exploitation de cette fonction qui permet ensuite de découvrir l'organisation des données. Rappelons le principe d'estimation.

Soit un échantillon $\chi=\{X_1,X_2,...,X_Q\}$ de Q observations indépendantes et identiquement distribuées suivant une loi de densité de probabilité p(X). La probabilité pour qu'une observation X soit située à l'intérieur d'un domaine $D(X_0)$ centré en un point X_0 de l'espace de représentation des observations, est donnée par :

$$P = \int_{D(X_0)} p(X)dX$$

La probabilité P_q pour que q observations parmi les Q observations disponibles soient situées à l'intérieur du domaine $D(X_0)$ est donnée par la loi binomiale :

$$P_q = C_Q^q \, P^q \, (1-P)^{Q-q}$$

L'espérance mathématique de q est :

$$E(P_q) = \sum_{q=0}^{Q} q.P_q$$

$$= Q.P$$

ce qui permet d'affirmer que q/Q est un estimateur non biaisé de P_q.

Sous l'hypothèse où p(X) ne présente pas de variations importantes à l'intérieur du domaine $D(X_0)$, la probabilité P_q peut se mettre sous la forme :

$$P_q = \int_{D(X_0)} p(X)\, dX$$

$$\approx p(X).V[D(X_0)]$$

où $V[D(X_0)]$ représente le volume du domaine $D(X_0)$.

L'estimation de la fonction de densité de probabilité p(X) en X_0 peut alors être obtenue sous la forme :

$$\hat{p}(X) = \frac{q/Q}{V[D(X_0)]}$$

Si le volume $V[D(X_0)]$ reste fixe lorsque le nombre Q d'observations tend vers l'infini, le rapport q/Q converge vers P, et l'estimateur $\hat{p}(X_0)$ tend vers la valeur

moyenne de p(X) en X_0. Mais, dans la pratique, le nombre d'observations disponibles est limité, et l'ajustement de la taille du domaine $D(X_0)$ est un point important pour estimer la fonction p(X). Si le domaine $D(X_0)$ associé à chaque point X_0 où on désire estimer $p(X_0)$ est trop petit, l'estimateur est nul dans presque tout l'espace de représentation des données, sauf au voisinage immédiat des observations où il prend des valeurs très élevées. L'estimateur présente alors des valeurs très irrégulières. A l'opposé, si le domaine $D(X_0)$ est trop grand, on obtient un phénomène de lissage, ce qui conduit à un estimateur qui manque de résolution.

On montre que $\hat{p}(X_0)$ converge vers $p(X_0)$ si les conditions suivantes sont respectées :

$$1) \quad \lim_{Q\to\infty} V\left[D(X_0)\right] = 0$$

$$2) \quad \lim_{Q\to\infty} q = \infty$$

$$3) \quad \lim_{Q\to\infty} \frac{q}{Q} = 0$$

Il existe deux méthodes non paramétriques d'estimation de la fonction de densité de probabilité sous-jacente à la distribution des observations respectant ces contraintes : la méthode du noyau de Parzen-Rosenblatt, et celle des k plus proches voisins de Cover et Hart. Examinons en détail ces deux principes d'estimation.

I.1. - METHODE DU NOYAU

I.1.1. - Principe d'estimation

La méthode du noyau, proposée initialement par Parzen et Rosenblatt [ROS56][PAR62], a été étendue au cas multidimensionnel par Cacoullos et Murthy [CAC62][MUR66]. L'estimateur s'écrit sous la forme :

$$\hat{p}(X) = \frac{1}{Q} \sum_{q=1}^{Q} \frac{1}{V[D(X_0)]} \Phi\left(\frac{X - X_q}{h_Q}\right)$$

où Φ est une fonction appelée noyau ou fenêtre.

La convergence de l'estimateur $\hat{p}(X)$ ainsi défini vers $p(X)$ est assurée sous les conditions suivantes :

$$\forall X, \quad \exists M \quad / \quad 0 \leq \Phi(X) \leq M$$

$$\lim_{\|X\| \to \infty} \quad \Phi(X).\|X\|^N = 0$$

$$\lim_{Q \to \infty} \quad V[D(X_0)] = 0$$

$$\lim_{Q \to \infty} \quad Q.V[D(X_0)] = \infty$$

Plusieurs types de fonctions Φ peuvent être choisie parmi différents types qui satisfont les conditions de convergence. On peut citer :

- Le noyau cubique

$$\begin{cases} \Phi(X) = 1 & \text{si} \quad |X| \leq \frac{1}{2} \\ \Phi(X) = 0 & \text{sinon} \end{cases}$$

- Le noyau triangulaire

$$\begin{cases} \Phi(X) = 1 - |X| & \text{si} \quad |X| \leq 1 \\ \Phi(X) = 0 & \text{sinon} \end{cases}$$

- Le noyau de Cauchy

$$\Phi(X) = \frac{1}{\pi}(1+X^2)^{-1}$$

- Le noyau gaussien

$$\Phi(X) = \frac{1}{\sqrt{2\pi}} e^{(-\frac{1}{2}X^2)}$$

Dans la pratique, le noyau le plus utilisé est le noyau cubique. Des considérations de facilité de programmation sur calculateur numérique motivent ce choix. Il est en effet très simple de compter le nombre d'observations situées à l'intérieur d'un hypercube de côté h_Q. La largeur h_Q est ajustée en fonction du nombre Q d'observations disponibles de la manière suivante :

$$h_Q = h_0\sqrt{Q}$$

$$\text{ou} \quad h_Q = h_0.\log Q$$

Une remarque peut être émise quant au choix des points où la fonction de densité de probabilité est estimée. Ce choix doit permettre une exploitation aisée de l'estimation de la fonction de densité de probabilité. Il doit donc tenir compte des traitements qui suivront la phase d'estimation. De nombreux algorithmes de classification passent par une discrétisation de l'espace de représentation des données sur un réseau hypercubique obtenu en divisant chaque axe de l'espace en intervalles égaux et réguliers de largeur h_Q. La fonction de densité de probabilité est alors estimée aux centres des hypercubes ainsi définis. Cette discrétisation de l'espace de représentation des données permet la mise en oeuvre simple de traitements de bas niveau sur les données qui nécessitent ce type de grille. De nombreuses techniques de filtrage [HAM77], l'implantation d'opérateurs différentiels, l'exploitation de la morphologie mathématique [SER82][STE86], les méthodes de relaxation [TOU88][OLE88] constituent un éventail de techniques de traitement qui ont été implantées sur des grilles hypercubiques et qui utilisent l'estimateur de Parzen-Rosenblatt.

Cependant, le simple dénombrement des observations situées dans chaque hypercube amène des calculs prohibitifs lorsque la dimension de l'espace augmente. En effet, plus la dimension de l'espace augmente, plus le nombre d'hypercubes définis par cette technique de discrétisation de l'espace de représentation devient important. Aussi, pour pallier ce problème qui pénalise les procédures au niveau des temps de calcul, un algorithme rapide d'estimation a été développé par Postaire et Vasseur [POS82].

I.1.2. - Problèmes posés par l'estimateur du noyau

La méthode du noyau, qui consiste à travailler avec une fenêtre de taille fixe, soulève plusieurs problèmes.

Si la taille de cette fenêtre est trop petite par rapport au nombre d'observations disponibles, l'estimateur présente de fortes irrégularités. En effet, l'estimateur ne prend des valeurs différentes de zéro qu'au voisinage des observations. Au contraire, si la taille de la fenêtre est trop grande, un phénomène de lissage apparaît. Dans ces deux cas, il est difficile d'obtenir une estimation de la fonction de densité de probabilité représentative de la structure réelle des données. Lorsque aucune connaissance *a priori* sur cette structure n'est disponible, c'est l'expérience de l'analyste qui guide l'ajustement de la taille de la fenêtre de Parzen pour l'adapter aux observations disponibles. Une fois h_0 fixé, les équations $h_Q=h_0\sqrt{Q}$ et $h_Q=h_0 \log Q$ ne permettent que de modifier h_Q si le nombre d'observations disponibles varie.

D'autre part, il se peut que la distribution des observations comporte des régions de densités très inégales. Il faudra donc trouver une taille de fenêtre adaptée pour que l'estimation de la fonction de densité de probabilité soit représentative de la structure de la distribution dans toutes les régions de l'espace de représentation des données. Ici également, l'expérience de l'analyste est de la toute première importance pour ajuster la taille de la fenêtre de Parzen. Il existe néanmoins des cas où l'échantillon comporte des densités locales si différentes qu'il est impossible de trouver une taille de fenêtre qui soit appropriée pour toutes les régions de l'espace.

Ces remarques amènent deux suggestions. D'une part, pour éviter de calculer tous les estimateurs nuls correspondant à des fenêtres vides, il serait plus judicieux d'estimer la fonction de densité de probabilité uniquement là où se trouvent les observations, c'est à dire plus précisément sur les observations elles-mêmes. Mais cette manière de procéder ne permettrait plus d'appliquer les traitements qui nécessitent une grille hypercubique de discrétisation. D'autre part, le fait de garder la même fenêtre de taille fixe dans les régions à faible densité comme dans les régions à forte densité d'observations compromet la recherche d'un estimateur représentatif de la structure sous-jacente à la distribution des observations disponibles. Compte-tenu de ces limites de la technique d'estimation par la fenêtre de Parzen sur une grille hypercubique, il serait intéressant d'étendre le champ d'application des travaux effectués avec cet estimateur, en particulier les techniques de filtrage non linéaire et de détection des contours des modes [TOU89] à des échantillons dont on estimerait la fonction de densité de probabilité sous-jacente par des procédures plus performantes.

C'est dans ce sens que nous présentons maintenant une seconde approche pour estimer les fonctions de densité de probabilité, celle des k plus proches voisins [COV67]. Elle consiste non plus à fixer la taille de la fenêtre d'estimation mais à déterminer le domaine qui englobe un nombre k, fixé, de voisins autour du point où on désire estimer la fonction de densité de probabilité sous-jacente. Cette méthode, sans résoudre totalement les problèmes soulevés précédemment par la méthode du noyau, permet une adaptation de la taille du domaine à la densité locale de la distribution des observations.

I.2. - METHODE DES K PLUS PROCHES VOISINS

I.2.1. - Principe de la méthode

Pour obtenir une estimation de la fonction de densité de probabilité sous-jacente à la distribution des observations disponibles, la méthode des k plus proches voisins consiste à rechercher un domaine, centré au point où on désire estimer cette fonction, qui englobe un nombre k fixé de voisins. Ces k voisins sont les plus proches du point d'estimation, d'où le nom qui est donné à cette méthode.

Cette technique évite l'apparition d'estimations nulles, puisque, quelque soit l'endroit où on estime la fonction de densité de probabilité, la taille du domaine est ajustée de telle sorte qu'un certain nombre de voisins du point d'estimation soient pris en compte. Cette constatation peut paraître triviale, mais elle justifie en partie la supériorité de cette méthode d'estimation. D'autre part, comme c'est le nombre k de voisins qui est fixé, la même quantité d'information est prise en compte quelque soit le point où on estime la fonction de densité de probabilité.

L'estimateur s'écrit sous la forme :

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

où $V[D(X_0)]$ est le volume du plus petit domaine $D(X_0)$ centré en X_0 qui contient les k plus proches voisins du point X_0. Généralement, la forme de ce domaine est liée à la norme définie sur l'espace de représentation des données. Nous préciserons plus loin la norme que nous avons retenue.

Par ce formalisme, nous voyons que l'estimateur de la fonction de densité de probabilité se présente de la même manière que celui utilisé pour la méthode du noyau. Toutefois, la différence fondamentale réside dans le fait que, pour la méthode des k plus proches voisins, ce n'est plus la taille du domaine $D(X_0)$ qui est fixée, mais le nombre k de voisins de X_0 à prendre en compte. La taille du domaine $D(X_0)$ est ajustée afin de satisfaire la contrainte imposée sur ce nombre k. Cette différence essentielle par rapport à la méthode du noyau permet ainsi à l'estimateur des k plus proches voisins de s'adapter à la densité locale de la distribution des observations. En effet, dans les zones à forte densité d'observations, le domaine a tendance à être petit alors qu'il s'agrandit dans les zones à faible densité d'observations.

Nous avons vu que, dans la pratique, la méthode d'estimation par la méthode du noyau s'accommodait très bien d'une discrétisation de l'espace sur une grille hypercubique et permettait ainsi la mise en oeuvre aisée de différents traitements très performants. En particulier, le fait de pouvoir ignorer les hypercubes vides permettait d'accélérer considérablement la procédure, et l'isotropie du maillage

se prêtait parfaitement à l'implantation de filtres non linéaires et d'opérateurs différentiels. Par la méthode des k plus proches voisins, l'estimation sur une grille de ce type amènerait des calculs extrêmement importants. En effet quelque soit le point d'estimation, il existe toujours des voisins, et aucun estimateur n'est strictement nul. L'algorithme proposé par Postaire et Vasseur [POS82] pour accélérer le processus d'estimation n'apporte plus d'aide dans la méthode des k plus proches voisins. Il est donc beaucoup plus intéressant d'estimer la fonction de densité de probabilité sur les observations elles-mêmes, ce qui résout simultanément plusieurs problèmes. Lorsque l'estimation est effectuée sur les observations elles-mêmes, la procédure évite de calculer l'estimateur dans les régions vides d'observations et ne nécessite aucune recherche de points particuliers d'estimation. De plus, en limitant l'estimation aux seules observations, on évite l'explosion exponentielle du nombre de calculs dans un espace de dimension élevée. Abordons maintenant le problème de réglage des paramètres.

Dans la méthode du noyau, deux paramètres devaient être ajustés par l'analyste : la taille h_Q du domaine d'observation et la fonction Φ qui définit la manière de prendre en compte les observations associées au domaine $D(X_0)$. Dans le cas de la technique des k plus proches voisins, il existe également deux paramètres à ajuster, à savoir le nombre k de voisins et la forme du domaine $D(X_0)$. Voyons comment répondre à la question du choix de ces paramètres.

I.2.2. - Ajustement du nombre k de plus proches voisins

L'ajustement du nombre k doit être fonction de la taille Q de l'échantillon disponible afin de respecter les contraintes qui assurent la convergence de l'estimateur. Pour un nombre Q d'observations, le nombre k peut être calculé ainsi :

$$k = k_0 \sqrt{Q}$$

$$\text{ou:} \quad k = k_0 . \log Q$$

En respectant ces règles d'ajustement, on est certain que l'estimateur converge quand le nombre Q augmente indéfiniment, quelque soit la valeur de k_0. Cependant, pour un échantillon fini, le choix de k_0 demeure très délicat. Cet ajustement est en général laissé à l'initiative de l'analyste. Il n'existe pas, en effet, de

règle objective pour fixer la valeur du coefficient k_0, et donc, à une constante multiplicative près, du nombre k. Les équations ci-dessus permettent seulement d'ajuster le nombre k de voisins à prendre en compte lorsque plusieurs échantillons de tailles différentes et provenant des mêmes sources aléatoires, sont soumis à l'analyse.

Un nombre k petit a tendance à rendre l'estimateur très sensible aux irrégularités de la distribution des observations, alors qu'un nombre k élevé conduit à une version lissée de la fonction de densité de probabilité. Cependant, dans le cas d'échantillons présentant de fortes disparités de densité, l'ajustement de k est beaucoup moins critique que celui de la largeur h_Q du noyau de la méthode précédente. En effet, la taille du domaine $D(X_0)$ s'adapte à la concentration locale des observations, de telle sorte que les fortes densités sont estimées avec de petits domaines alors que les faibles le sont avec de grands domaines.

Afin de montrer l'effet du nombre k de voisins pris en compte pour estimer la fonction de densité de probabilité, nous avons calculer l'estimateur en faisant varier le nombre k de 1 à 40, sur deux échantillons bidimensionnels générés artificiellement (ces deux échantillons seront repris plus loin dans ce chapitre). Sur les figures 1.2.2.a et 1.2.2.b on a représenté l'évolution de l'estimateur en chaque observation, en fonction du nombre k de voisin pris en compte pour estimer p(X). Sur l'axe X figure le numéro de l'observation, sur Y le nombre k, et sur Z l'estimateur. Il faut remarquer que l'estimateur ne varie guère avec le nombre k, seuls les estimateurs correspondant à $k < 5$, présentent des variations relativement plus importantes. L'information prise en compte dans ces cas, est en fait beaucoup plus locale que pour des valeurs de k plus grand, et les irrégularités de la distribution des observations ont alors une influence importante sur l'estimation de la fonction de densité de probabilité.

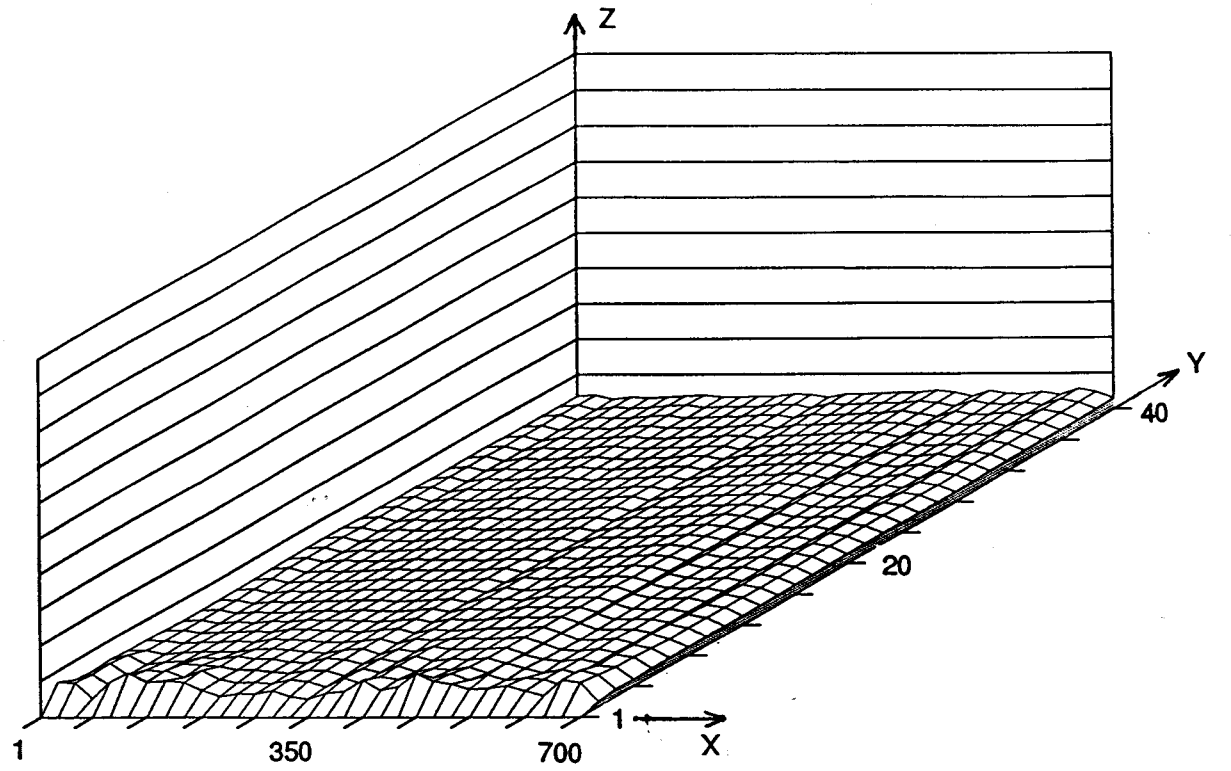


Figure 1.2.2.a : Evolution de l'estimateur en fonction de k (exemple 1).

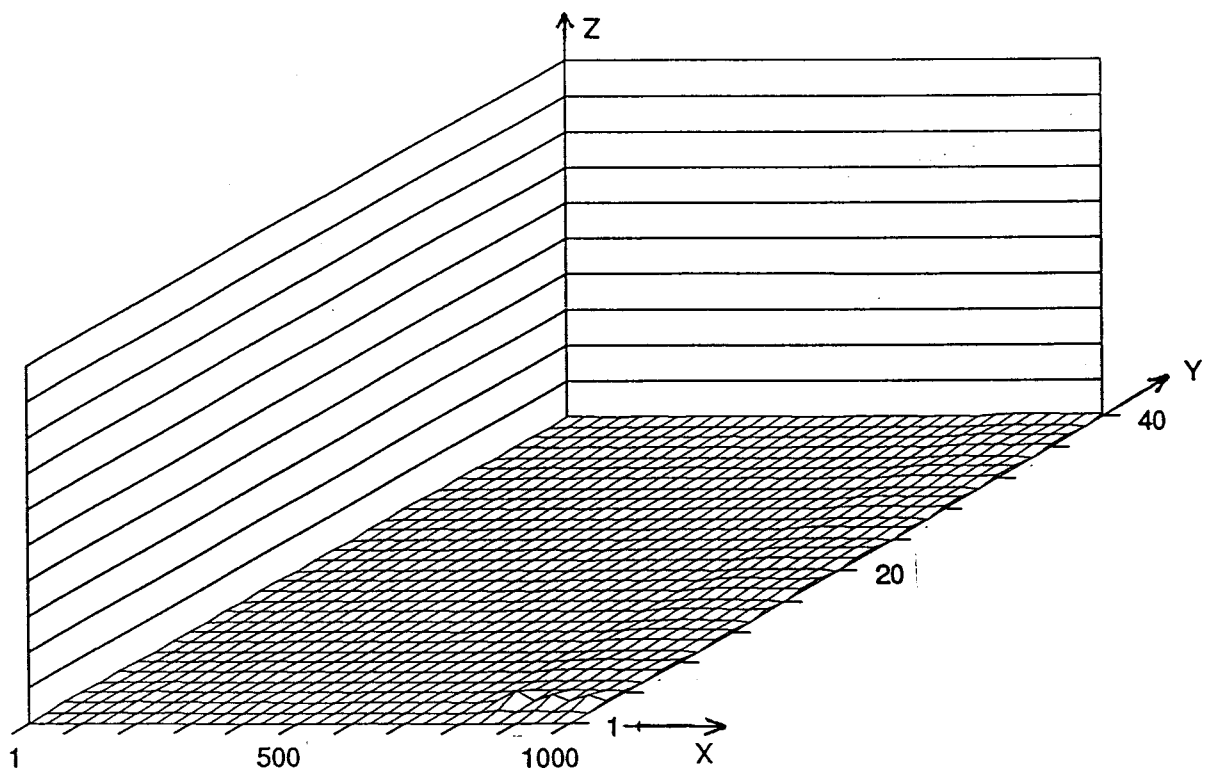


Figure 1.2.2.b : Evolution de l'estimateur en fonction de k (exemple 2).

1.2.3. - Choix de la forme du domaine

La procédure d'estimation consiste à centrer un domaine $D(X_0)$ de taille variable au point X_0 où on désire estimer la fonction de densité de probabilité, et à ajuster cette taille de telle sorte que le domaine $D(X_0)$ contienne k voisins du point X_0. Le choix de la forme de ce domaine est important si on veut extraire une information reflétant la structure sous-jacente à la distribution de manière significative. Ainsi, pour prendre en compte de manière isotropique tous les voisins, ou plus précisément les k plus proches voisins du point considéré, nous avons opté pour un domaine hypersphérique. Le choix d'un tel domaine n'est pas sans incidence sur le temps de calcul puisqu'il nécessitera la détermination du rayon, puis du volume de la plus petite hypersphère contenant les k plus proches voisins de X_0. Bien que ce qui importe le plus dans les procédures de classification est la mise en évidence, d'une manière aussi fidèle que possible, de la structure de l'échantillon soumis à l'analyse, nous ne délaissons pas le problème du temps de calcul, celui-ci pouvant devenir très important dans le cas d'un échantillon de grande dimension. Nous proposerons une technique qui permet d'accélérer de manière significative le processus d'estimation et, surtout, les traitements qui le suivent pour exploiter la fonction de densité de probabilité.

La forme du domaine d'observation étant maintenant arrêtée, nous pouvons réécrire l'estimateur de manière plus complète :

$$\hat{p}(X_0) = \frac{k/Q}{\frac{\pi^{n/2}}{\Gamma\left(\frac{n}{2}+1\right)}\|X_k - X_0\|^n}$$

où $\Gamma\left(\frac{n}{2}+1\right)$ est la fonction gamma telle que :

$$\Gamma\left(\frac{n}{2}+1\right) = \int_0^\infty t^{(n/2+1)} e^{-t}\, dt$$

et $\|X_k - X_0\|$: la distance euclidienne entre le point X_0 et le $k^{ième}$ voisin noté X_k.

Le calcul du volume du domaine hypersphérique peut être simplifié. Sachant que la dimension de l'espace de représentation a une valeur n entière, on peut écrire :

_ Pour une dimension n pair :

$$\Gamma\left(\frac{n}{2}+1\right)=\left(\frac{n}{2}\right)!$$

d'où :
$$V[D(X_0)]=\frac{\pi^{n/2}}{\left(\frac{n}{2}\right)!}\|X_k-X_0\|^n$$

_ Pour un nombre n de dimensions impair :

$$\Gamma\left(\frac{n}{2}+1\right)=\frac{(n+1)!\ \sqrt{\pi}}{2^{(n+1)}\cdot\left(\frac{n+1}{2}\right)!}$$

d'où :
$$V[D(X_0)]=\frac{2^{(n+1)}\cdot\left(\frac{n+1}{2}\right)!\ \ \pi^{\left(\frac{n-1}{2}\right)}}{(n+1)!}\|X_k-X_0\|^n$$

I.2.4. - Remarques sur la méthode des k plus proches voisins

Nous avons vu que, dans la méthode du noyau, en dehors du cas où on choisit de prendre pour fonction Φ le noyau cubique, les autres types de fonction pondèrent la contribution des observations en fonction de leur éloignement par rapport au point où on estime la fonction de densité de probabilité. Cette idée de pondérer la contribution des voisins de l'observation pourrait être adaptée à la méthode des k plus proches voisins. Il suffirait pour cela de modifier légèrement le calcul de l'estimateur en introduisant une fonction de pondération.

L'estimateur s'écrirait sous la forme :

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

où α serait une fonction décroissante de la contribution des voisins.

Intuitivement, l'introduction d'une fonction de pondération permettrait d'obtenir une meilleure estimation de la fonction de densité de probabilité, puisque le $k^{ième}$ voisin, c'est à dire le plus éloigné, apporterait une contribution moindre que le plus proche des voisins de l'observation où on estime la fonction de densité de probabilité. Mais il a été montré que la méthode des k plus proches voisins pondérés ne donne pas toujours de meilleurs résultats que la méthode classique des k plus proches voisins [MOR81][MAC87]. Nous utiliserons donc la méthode classique d'estimation.

En optant pour la méthode d'estimation des k plus proches voisins, il reste encore le problème important du temps de calcul lié au choix même de la forme du domaine d'observation. Ce problème, s'il n'est pas crucial, doit tout de même être pris en considération car l'analyste éprouve souvent le besoin de relancer plusieurs fois les procédures de classification pour ajuster les paramètres, sélectionner les attributs des observations les plus discriminants et évaluer les résultats.

Ceci nous a amenés à diviser l'algorithme des k plus proches voisins en deux étapes distinctes. La première consiste en un prétraitement des données se limitant à ordonner les voisins selon leurs distances euclidiennes par rapport à l'observation considérée. La seconde concerne le calcul de l'estimateur de la fonction de densité de probabilité proprement dit, au niveau de cette observation.

II. - ACCELERATION DE L'ALGORITHME D'ESTIMATION

Un des inconvénients majeurs de la méthode des k plus proches voisins est son temps d'exécution. Pour évaluer le volume du domaine $D(X_0)$ centré au point X_0 où on estime la fonction de densité de probabilité qui englobe les k plus proches voisins de X_0, il faut d'abord calculer toutes les distances entre le point X_0, qui est

une des observations disponibles, et les autres observations de l'échantillon. Il s'agit ensuite de trier ces distances pour identifier les k plus proches voisins du point X_0. Ce processus doit être répété pour toutes les observations, entraînant, dans le cas d'échantillons de taille et de dimension élevées, un temps de calcul important. Cette limitation pourrait inciter l'opérateur à choisir une autre stratégie d'estimation.

Afin de lever cet obstacle, nous proposons d'effectuer tout d'abord un prétraitement sur les échantillons qui consiste à ordonner les voisins de chaque observation en fonction de leurs éloignements par rapport à celle-ci. Cet ordonnancement, conservé en mémoire, permettra d'estimer la fonction de densité de probabilité sous-jacente à la distribution en réduisant sensiblement le temps d'exécution. D'autre part, cette technique permettra de relancer le processus d'estimation pour différentes valeurs de k, sans pour cela augmenter le temps de calcul de manière prohibitive.

Nous verrons ultérieurement que le fait de conserver en mémoire les voisins ordonnés selon leurs distances permettra d'accélérer également les différentes phases de traitement qui mèneront à la classification finale des observations disponibles. En effet, les étapes successives du processus de classification présenté dans la suite de ce mémoire ont pour point de départ l'estimation de la fonction de densité de probabilité sous-jacente à la distribution des observations disponibles. Plusieurs traitements s'avéreront indispensables afin d'extraire de cette fonction estimée les informations nécessaires à la classification des données analysées. Ces traitements font en général appel à la notion de voisinage du point d'estimation. Ils nécessitent la connaissance, soit de la position d'un voisin particulier du point d'estimation X_0, soit de la distance entre ce point X_0 et un voisin, soit encore des positions ou des distances des k plus proches voisins. En chaque point d'estimation X_0, c'est à dire en chaque observation dans notre cas, il est particulièrement intéressant de disposer immédiatement de l'information que l'on cherche, en l'occurrence la position relative d'un voisin, de la manière la plus rapide et la plus simple possible.

II.1. - ORDONNANCEMENT DES VOISINS

Au départ, l'analyste ne possède que la liste des observations à classer qui est constituée des seules valeurs des attributs sélectionnés pour caractériser les

éléments de l'échantillon soumis à l'analyse. A chaque élément correspond une observation, donc un point dans l'espace de représentation auquel on assigne un numéro, ou indice. Nous allons dresser un tableau dans lequel nous ferons correspondre à chaque observation, les indices de ses K plus proches voisines, ordonnées selon leurs distances euclidiennes croissantes par rapport à cette observation. Le nombre K est choisi suffisamment grand pour que le nombre k de voisins nécessaires à l'estimation, soit toujours inférieur à K.

Décomposons l'algorithme d'ordonnancement :

a) Pour le point $X_0=[x_1,...,x_i,...,x_n]^T$ correspondant à l'observation d'indice 0, nous déterminons le vecteur D_0 qui a pour composantes les distances euclidiennes entre le point X_0 et les autres points de l'échantillon :

$$D_0 = \begin{pmatrix} \left[\sum_{j=1}^{n}\left(x_{j,1}-x_{j,0}\right)^2\right]^{1/2} \\ \vdots \\ \left[\sum_{j=1}^{n}\left(x_{j,i}-x_{j,0}\right)^2\right]^{1/2} \\ \vdots \\ \left[\sum_{j=1}^{n}\left(x_{j,Q-1}-x_{j,0}\right)^2\right]^{1/2} \end{pmatrix}$$

où $x_{j,i}$ correspond à l'attribut j de l'observation X_i

b) Nous trions ensuite les composantes de D_0 par ordre croissant en arrêtant le processus dès que les K plus petites distances sont trouvées. Parallèlement à ce tri, nous faisons correspondre les indices des points voisins sur un vecteur I_0 initialisé comme suit :

$$I_0 = \begin{pmatrix} 1 \\ \vdots \\ Q \end{pmatrix}$$

Ce tri achevé, le vecteur I_0 se présente sous la forme :

$$I_0 = \begin{pmatrix} i_1 \\ \vdots \\ i_K \\ \vdots \\ i_Q \end{pmatrix}$$

où $i_1, i_2, .., i_K$ sont les indices des K plus proches voisins de X_0, ordonnés par ordre croissant des distances euclidiennes qui les séparent du point X_0.

c) Les indices des K plus proches voisins du point X_0 constituent ensuite la première ligne d'un tableau T, appelé tableau des indices des voisins de chaque point de l'échantillon.

d) En itérant la procédure de tri sur toutes les observations disponibles, on construit pas à pas le tableau des indices des voisins, chaque nouvelle ligne indiquant les voisins de l'observation correspondante.

$$T = \underbrace{\left.\begin{bmatrix} i_{1,1} & i_{1,2} & \cdots & i_{1,K} \\ i_{2,1} & i_{2,2} & \cdots & i_{2,K} \\ & & \vdots & \\ i_{Q,1} & i_{Q,2} & \cdots & i_{Q,K} \end{bmatrix}\right\}}_{K \text{ colonnes}} Q \text{ lignes}$$

où $i_{j,k}$ est l'indice du $k^{ième}$ voisin de l'observation X_j.

Nous obtenons finalement un tableau à Q lignes et K colonnes. Chaque ligne est constituée des indices des K plus proches voisins de l'observation correspondante. Les observations sont traitées dans l'ordre dans lequel elles se présentent dans la liste des données brutes. On évite ainsi la mise en place d'un quelconque artifice de mise en correspondance entre les points qui se trouvent dans le tableau T et les données.

II.2. - CALCUL DE L'ESTIMATEUR

Maintenant que nous disposons d'un tableau dans lequel figurent les indices des voisins ordonnés de chaque observation, il est très aisé de calculer l'estimateur $\hat{p}(X)$. Il suffit pour cela, une fois fixé le nombre k de voisins que l'on désire prendre en compte dans la procédure d'estimation, d'extraire du tableau T, l'indice du $k^{ième}$ voisin et d'évaluer la grandeur $\hat{p}(X)$.

$$\hat{p}(X) = \frac{k/Q}{V[D(X)]} \qquad ; k < K$$

On calcule V[D(X)] simplement en recherchant l'indice i de l'observation correspondant au $k^{ième}$ voisin dans le tableau T, puis en prenant les coordonnées de l'observation d'indice i.

Il faut signaler que le calcul de $\hat{p}(X)$ est effectué sur un nombre exact de voisins, et on a la certitude que le domaine D(X) est le plus petit domaine qui contient les k plus proches voisins de X.

II.3. - PERFORMANCE DE L'ALGORITHME

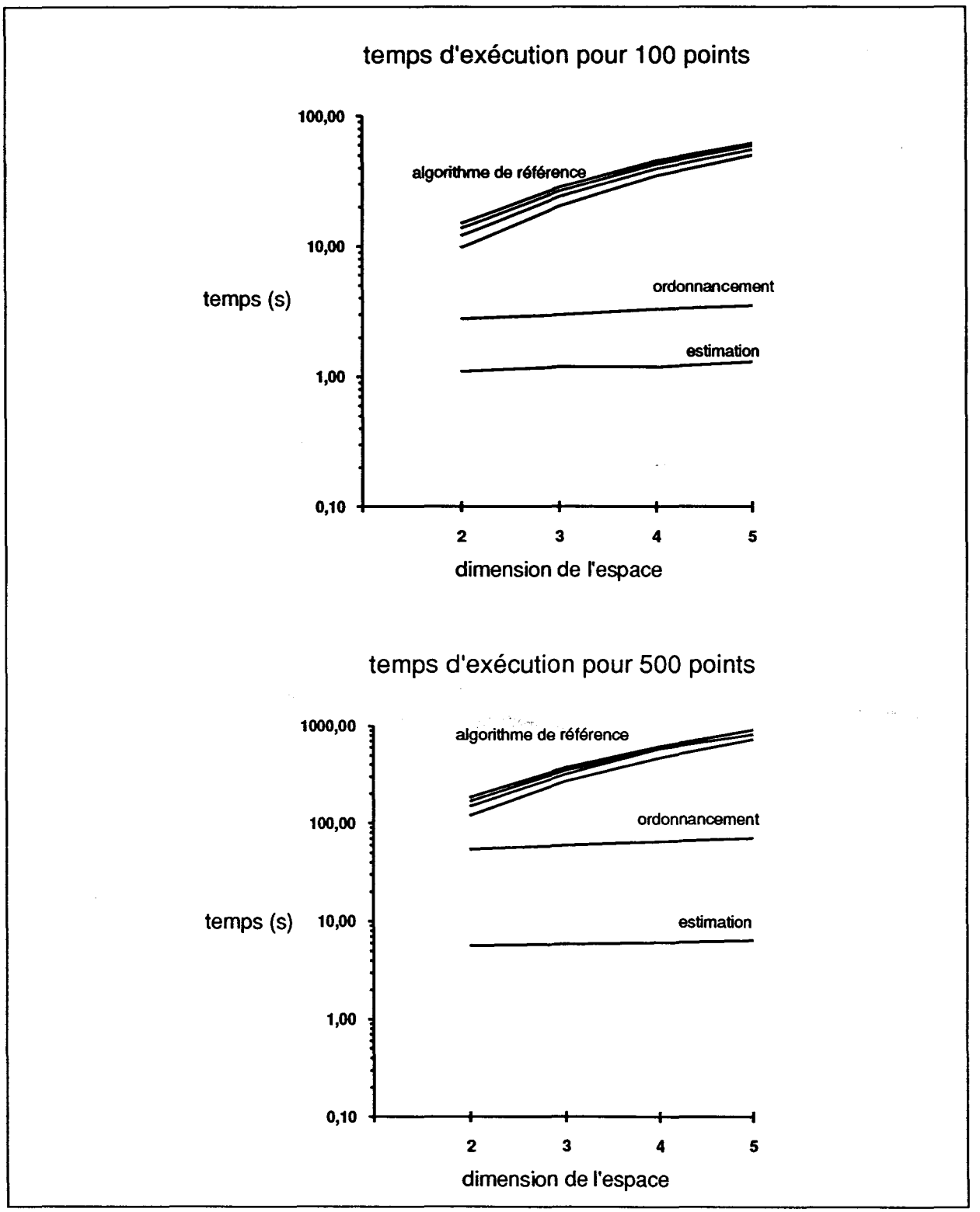
Pour illustrer l'intérêt de décomposer l'algorithme en deux étapes distinctes, nous nous proposons d'évaluer ses performances en l'appliquant à différents échantillons.

Chaque échantillon retenu pour cette analyse est composé d'une seule classe gaussienne. Différents échantillons ont été générés artificiellement pour des

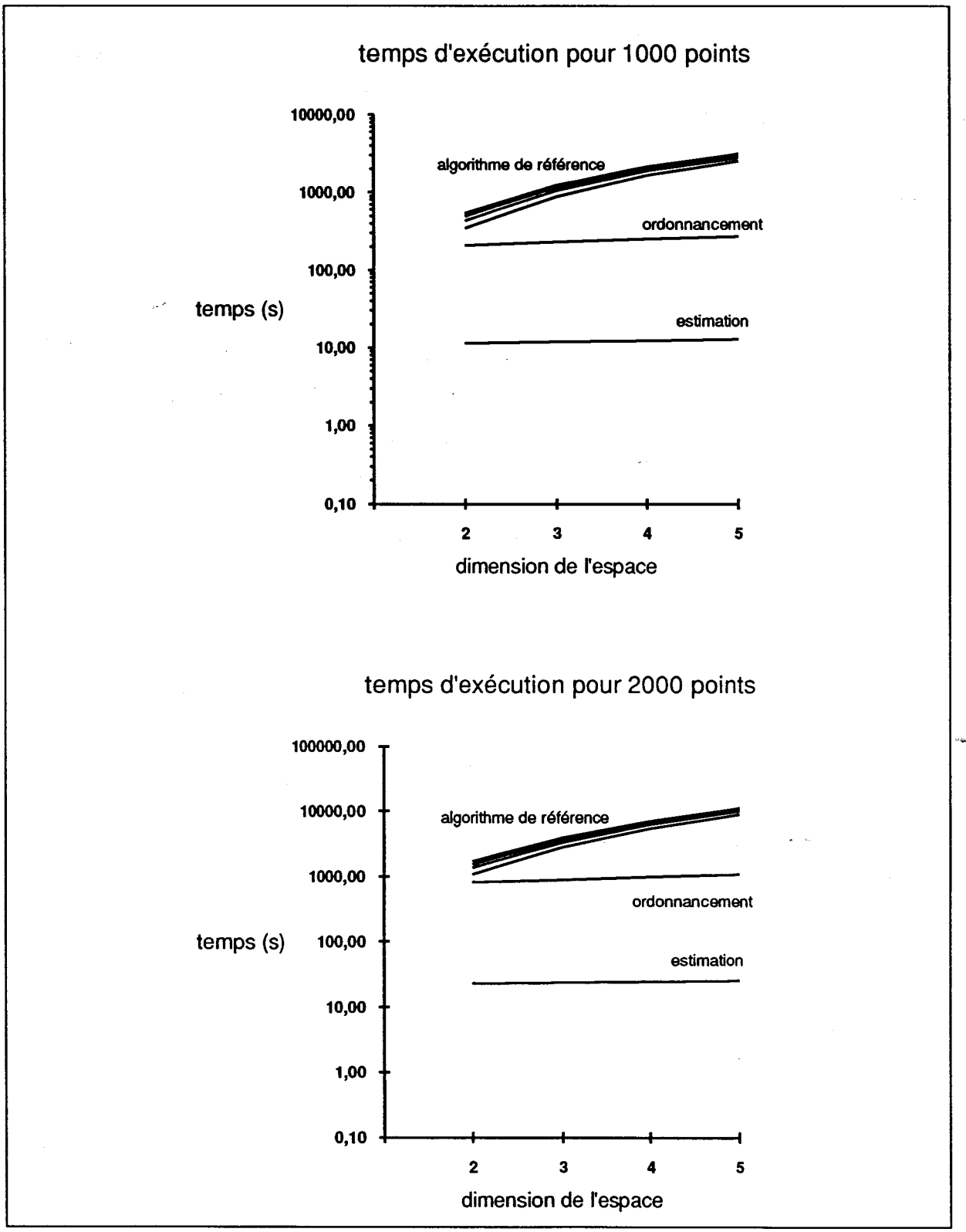
dimensions allant de 2 à 5, et des nombres d'observations de 100, 500, 1000 et 2000. Sur les figures II.3.a,b,c et d, sont reportés les temps d'exécution, sur une échelle logarithmique, en fonction de la dimension de l'échantillon, et en prenant 10, 20, 30 et 40 voisins pour estimer la fonction de densité de probabilité.

Pour un échantillon donné, l'ordonnancement de K plus proches voisins se fait une seule fois en construisant le tableau T des indices des K plus proches voisins. Nous pouvons aisément remarquer que le temps de calcul nécessaire à l'estimation elle-même est totalement indépendant du nombre k de voisins pris en compte. Les tracés pour k = 10, 20, 30 et 40 sont confondus. En effet, le calcul de l'estimateur en un point ne réclame que l'évaluation du volume de la plus petite hypersphère qui contient les k plus proches voisins de ce point. Or, le tableau T permet de connaître immédiatement l'indice du $k^{ième}$ voisin dont les coordonnées sont immédiatement disponibles, ce qui permet le calcul du volume de l'hypersphère. Le nombre d'attributs de l'échantillon n'influe que très peu sur ce temps d'exécution. En fait, il n'intervient que sur le calcul de la distance euclidienne qui sépare le point d'estimation et son $k^{ième}$ voisin. Le seul paramètre qui intervient sensiblement est le nombre Q d'observations disponibles. D'autre part, lorsque l'analyste doit relancer plusieurs fois le processus d'estimation, celui-ci prendra un temps très court puisque seul le temps de calcul de l'estimation proprement dite est à prendre en considération.

A titre de comparaison nous avons fait figurer sur les graphiques les temps de calcul correspondant à un algorithme de référence. Il consiste, en chaque observation, à se fixer une taille de domaine initial, à dénombrer les points voisins inclus dans ce domaine, et, si le nombre de voisins est inférieur au nombre k fixé, faire croître le domaine pas à pas jusqu'à englober le nombre de voisins désiré. Le temps de calcul par cet algorithme dépend du nombre de voisins puisqu'il consiste à faire croître le domaine d'autant plus que le nombre k est élevé.



Figures II.3. a et b : Temps d'exécution en fonction de la dimension de l'espace.



Figures II.3. c et d : Temps d'exécution en fonction de la dimension de l'espace.

III. - EXEMPLE D'ESTIMATION

Afin d'illustrer l'estimation de la fonction de densité de probabilité, nous appliquons l'algorithme présenté dans ce chapitre sur deux échantillons bidimensionnels générés artificiellement.

Exemple 1

L'échantillon est constitué de 3 classes gaussiennes dont les caractéristiques statistiques sont indiquées dans le tableau III.1. La représentation de cet ensemble de données dans un repère orthogonal est reproduite sur la figure III.1.a.

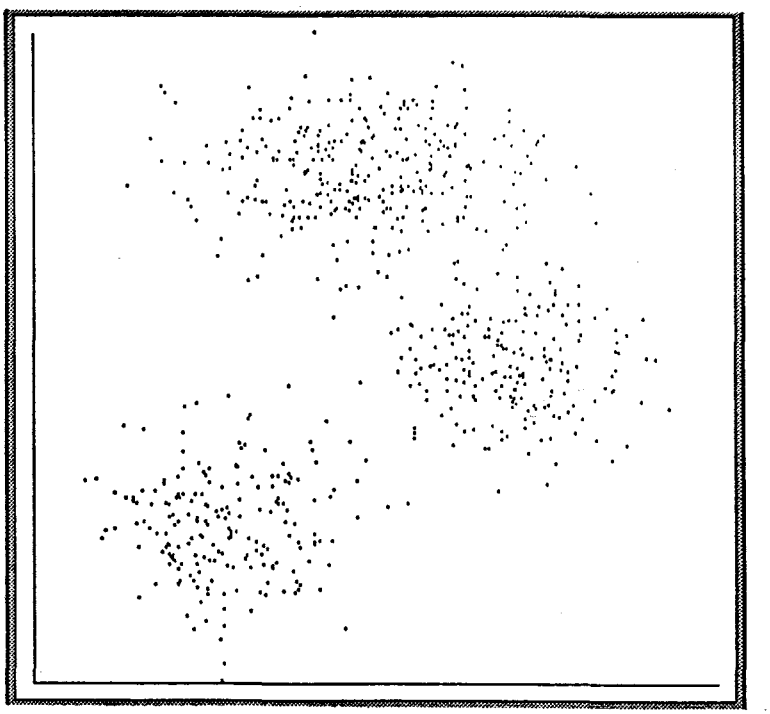


Figure III.1.a Représentation de l'ensemble des données de l'exemple 1

	classe 1	classe 2	classe 3
nombre de points	200	200	300
vecteur moyenne	$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{pmatrix}7\\4\end{pmatrix}$	$\begin{pmatrix}3\\9\end{pmatrix}$
matrice de covariance	$\begin{bmatrix}2 & 0\\0 & 2\end{bmatrix}$	$\begin{bmatrix}2 & 0\\0 & 2\end{bmatrix}$	$\begin{bmatrix}5 & 0\\0 & 1\end{bmatrix}$

Tableau III.1. Caractéristiques statistiques de l'exemple 1.

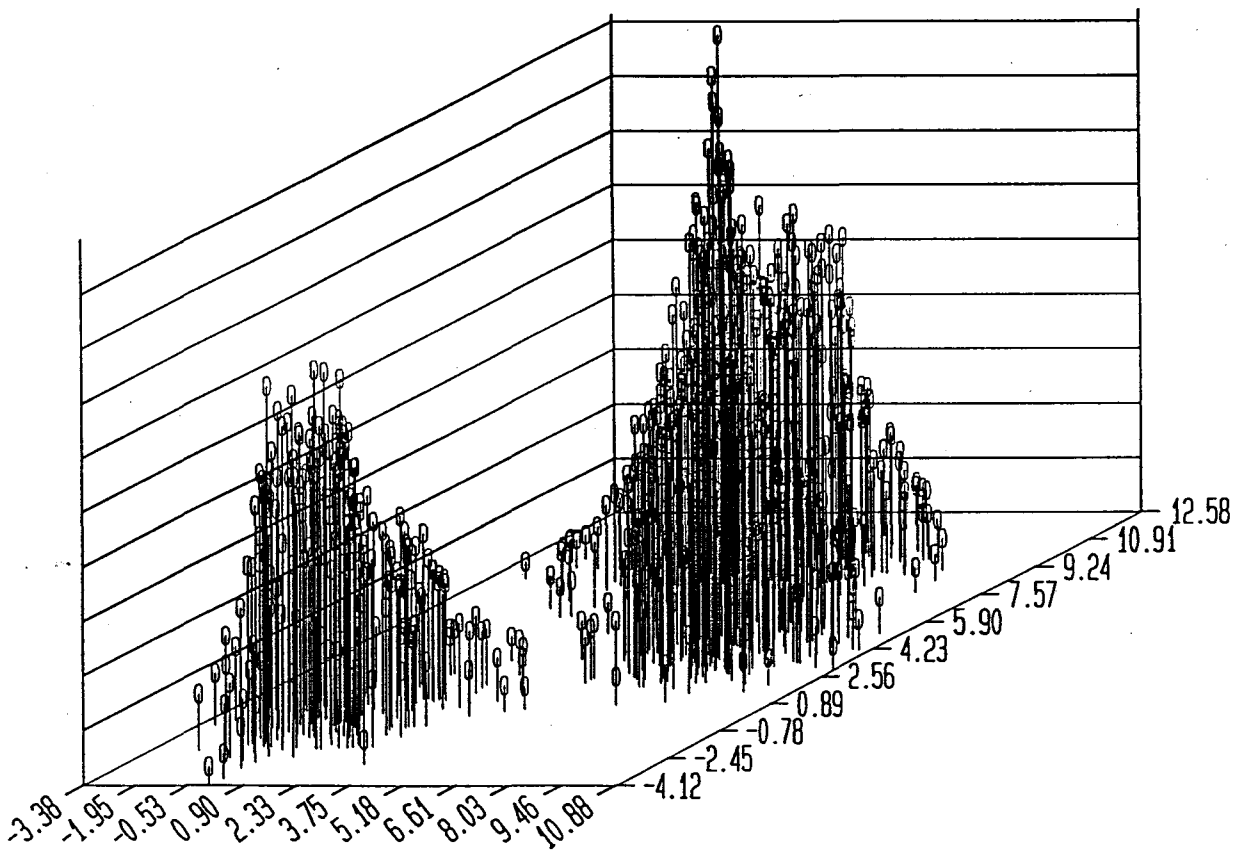


Figure III.1.b. Estimation de la fonction de densité de probabilité de l'exemple 1 (k=20)

Sur la figure III.1.b., on représente les estimateurs en chaque point de l'échantillon. La hauteur de chaque trait correspond à la valeur de l'estimateur calculé au niveau d'une observation donnée. On remarque qu'une telle présentation des résultats ne permet pas de distinguer les différentes valeurs des estimateurs au niveau des observations car on obtient un amalgame de traits.

On choisit alors une représentation des estimateurs plus lisible en effectuant un maillage hypercubique de l'espace de visualisation (Cf. figure III.1.c). Dans chaque maille, on calcule la moyenne des estimateurs des observations situées dans celle-ci. Ces moyennes permettent alors de représenter la fonction de densité de probabilité estimée à l'aide d'une structure maillée. Il est bien entendu que ce maillage a comme seul but de faciliter la lecture du graphique, les procédures d'estimations et les traitements qui seront présentés plus loin n'utilisent aucun maillage de l'espace de représentation des données disponibles lors des phases de traitements successifs.

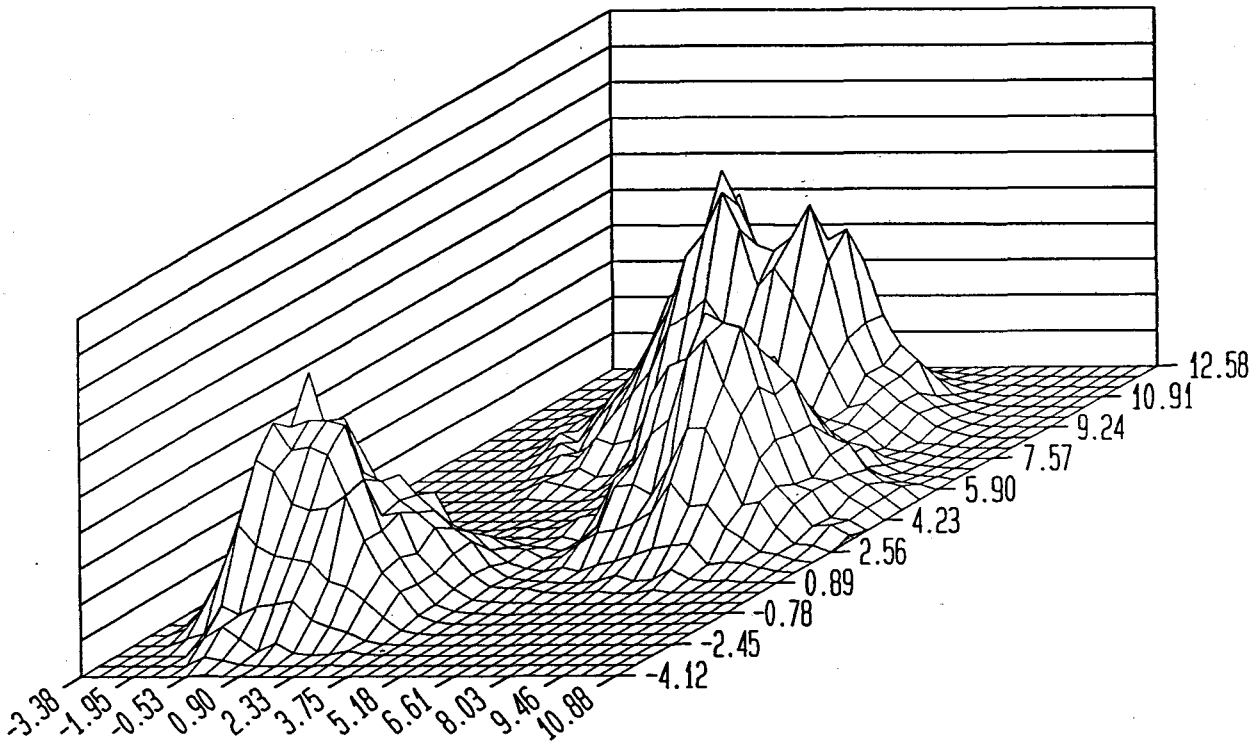


Figure III.1.c. Estimation de la fonction de densité de probabilité de l'exemple 1.

Exemple 2.

Cet exemple est constitué de classes non gaussiennes dont les attributs décrivant chaque observation sont définis par :

$$x_1 = a_1 \cos\Theta + b_1$$

$$x_2 = a_2 \cos\Theta + b_2$$

où - Θ est une variable aléatoire normale de moyenne m et de variance s
- b_1 et b_2 sont des variables aléatoires normales de moyennes μ et de variance σ^2 [FUK84].

Les caractéristiques statistiques sont indiquées dans le tableau III.2.

	classe 1	Classe 2	Classe 3
nombre de points	400	400	200
Θ	$m = 0°$ $s = 25°$	$m = 90°$ $s = 45°$	$m = 0°$ $s = 2°$
a_1	15	10	0
a_2	15	10	0
b_1	$\mu = 15$ $\sigma^2 = 2$	$\mu = 5$ $\sigma^2 = 2$	$\mu = 5$ $\sigma^2 = 2$
b_2	$\mu = 15$ $\sigma^2 = 2$	$\mu = 15$ $\sigma^2 = 2$	$\mu = 10$ $\sigma^2 = 2$

Tableau III.2. Caractéristiques statistiques de l'exemple 2.

Les données sont représentées figure III.2.a.

L'estimateur de la fonction de densité de probabilité est présenté sur la figure III.2.b.

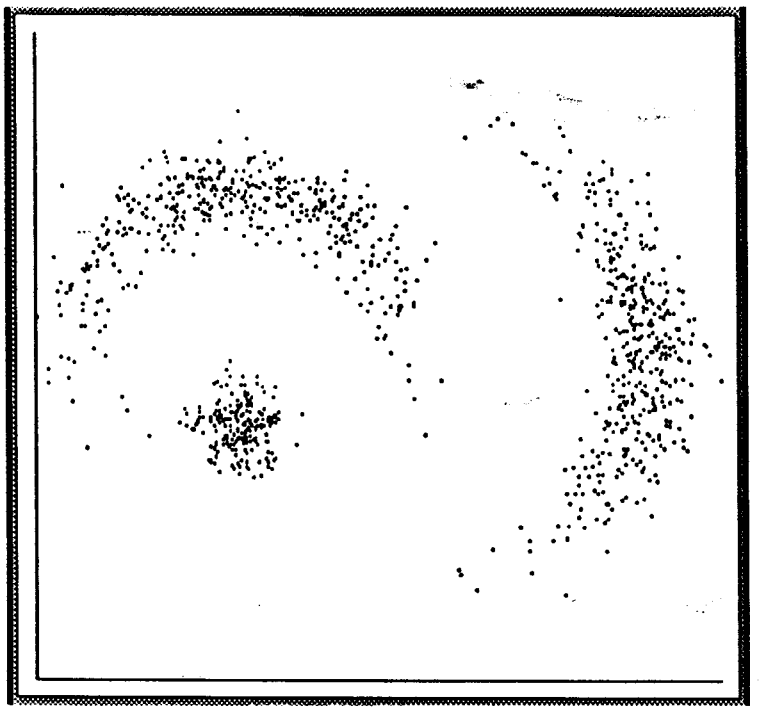


Figure III.2.a. Représentation des données de l'exemple 2.

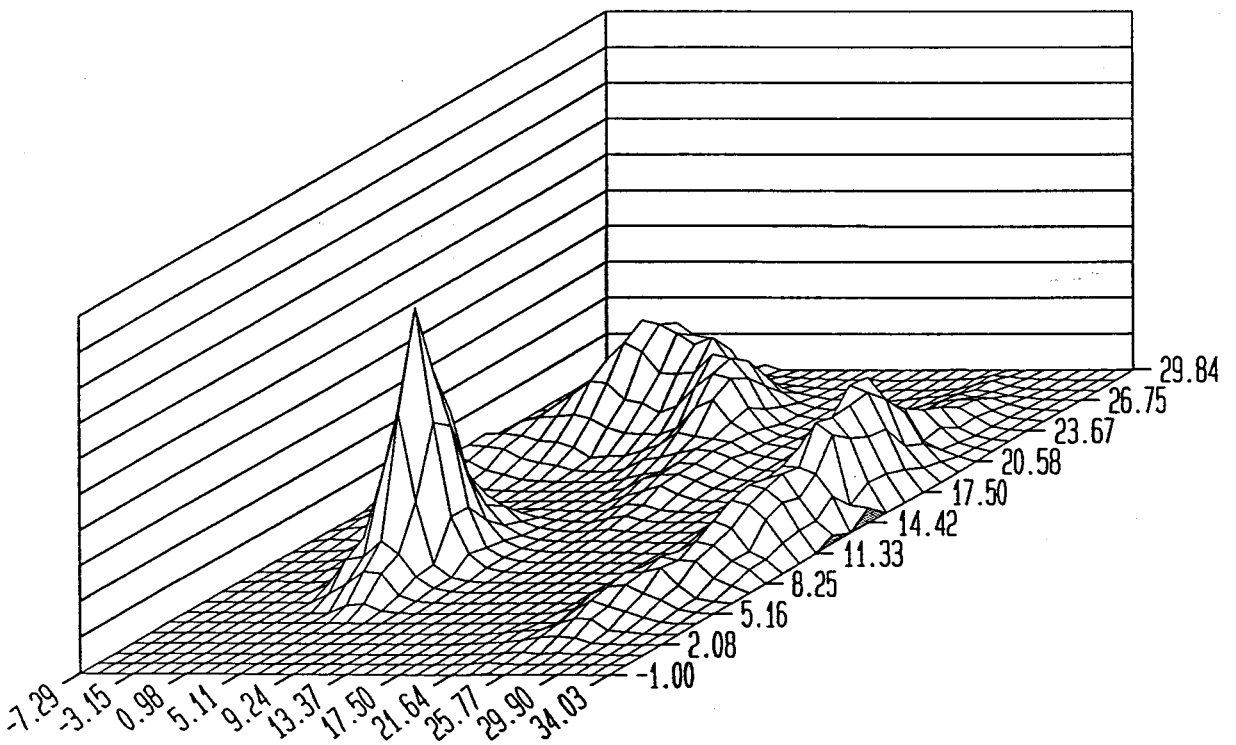


Figure III.2.b. Estimation de la fonction de densité de probabilité de l'exemple 2.

Remarque sur les exemples :

Sur les représentations des estimations de la fonction de densité de probabilité (fig. III.1.b et fig. III.2.a), on peut distinguer les modes correspondants aux différentes classes en présence dans les échantillons. Cependant, cette estimation est encore trop bruitée pour discerner correctement les différentes classes de manière automatique. Un filtrage de la fonction de densité de probabilité s'avère nécessaire.

IV. - CONCLUSION

Dans ce chapitre a été exposée la première étape qui doit mener à la classification des données soumises à l'analyse. Le choix de l'estimateur des k plus proches voisins permet d'adapter l'estimateur à la densité locale de la distribution. Un autre avantage de cette méthode d'estimation est que le réglage du paramètre k n'est pas critique pour obtenir un estimateur représentatif de la distribution des observations.

La séparation de l'algorithme en deux étapes distinctes autorise l'analyste à relancer le processus d'estimation pour un nombre k différent sans craindre de voir le temps de traitement croître de manière importante. La mise en mémoire des indices des K plus proches voisins accélère le processus d'estimation, mais sera également exploitée par les différents traitements qui suivront la phase d'estimation de la fonction de densité de probabilité, jusqu'à la classification finale.

CHAPITRE III

DETECTION DES MODES
DE LA FONCTION DE DENSITE DE PROBABILITE

CHAPITRE III

DETECTION DES MODES DE LA FONCTION DE DENSITE DE PROBABILITE

I. - FILTRAGE DE LA FONCTION DE DENSITE DE PROBABILITE

De nombreuses approches au problème de la classification automatique basées sur l'estimation de la fonction de densité de probabilité consistent à détecter les modes de cette fonction [ASS89]. Sous l'hypothèse communément admise que chaque classe de l'échantillon soumis à l'analyse donne naissance à un mode, il est ainsi possible de découvrir la structure des données en regroupant les observations autour de ces modes pour reconstituer les classes en présence.

Il existe de nombreuses procédures de détections des modes qui font appel à différentes propriétés de la fonction de densité de probabilité estimée. Les modes peuvent être considérés comme des maxima locaux de la fonction de densité sous-jacente. Ils sont alors identifiés par des techniques classiques de recherche des maxima basées sur des opérateurs différentiels qui indiquent la pente de la fonction. Une variante de cette approche consiste à estimer directement le gradient de cette fonction à partir des observations disponibles plutôt que d'estimer la fonction de densité de probabilité, puis son gradient [FUK75].

Comme toutes les techniques à base d'opérateurs différentiels, ces procédures sont sensibles aux bruits et aux irrégularités de la distribution des observations. L'introduction du concept de convexité permet de considérer les

modes comme des régions convexes de la fonction de densité de probabilité. Un test de convexité, basé sur des opérateurs plus robustes que les opérateurs différentiels [VAS80], permet de décrire la fonction de densité comme un ensemble de points à convexité positive, qui correspondent aux modes, noyés dans des points à convexité négative, qui correspondent aux vallées. La robustesse de cette approche peut être améliorée par l'utilisation de techniques d'étiquetage probabiliste itératif [OLE88].

Plus récemment, les modes ont été décrits par leurs contours, définis comme des régions de l'espace où la fonction de densité de probabilité présente de fortes variations spatiales. La détection de ces contours [TOU89], qui peut être améliorée par des techniques d'étiquetage probabiliste itératif [TOU87], fournit une description fine de la structure des données soumises à l'analyse.

Parmi les nombreuses techniques de détection des modes, il convient également de citer l'utilisation des outils de la morphologie mathématique, soit binaire [MAT75][SER82], soit multiniveaux [STE86]. Cette nouvelle approche semble présenter un potentiel d'application très important, mais nécessite une discrétisation très soignée de l'espace puisqu'elle repose sur la théorie des ensembles [BOT91].

Il peut paraître paradoxal que les chercheurs aient poussé si loin l'investigation de nouveaux outils et aient eu tendance à délaisser une technique simple et rapide à mettre en oeuvre : le seuillage de la fonction de densité de probabilité.

Il est cependant évident qu'un seuillage direct de cette fonction de densité, destiné à isoler les régions où elle présente des valeurs élevées dues à la présence de groupements significatifs d'observations, serait très souvent voué à l'échec. En effet, il est très difficile d'obtenir un estimateur idéal, constitué de modes bien marqués séparés par de profondes vallées. En général, et malgré toutes les précautions prises par l'analyste pour ajuster les paramètres de réglages de ces procédures, l'estimateur est soit hératique car trop sensible aux irrégularités de la distribution des observations, soit trop lissé, de telle sorte que les modes sont empâtés et mal différenciés.

Dans ce chapitre nous allons tirer profit de la qualité de l'estimateur des k plus proches voisins, qui est en général supérieure à celle de l'estimateur de Parzen-Rosenblatt. Il n'est cependant pas possible d'exploiter directement les résultats de l'estimation de la fonction de densité de probabilité, car celle-ci présente encore trop d'irrégularités. Nous allons donc soumettre l'estimateur à une procédure de filtrage multidimensionnel, de telle sorte qu'il sera possible de détecter les noyaux des modes par simple seuillage de la fonction estimée et filtrée.

La procédure retenue dans cette étude est basée sur l'extraction des noyaux des modes. Le noyau peut être défini comme étant constitué des observations qui correspondent à la partie convexe du mode [VAS80]. Pour que la classification obtenue soit représentative de la structure de l'échantillon, les noyaux extraits de la fonction de densité de probabilité doivent répondre à certaines contraintes. De toute évidence, la position des noyaux dans l'espace de représentation des données doit refléter celles des modes. Mais il faut également que la forme des noyaux, ou plus exactement celles de leurs contours géométriques, soit la plus proche possible de celle des modes, afin d'assurer la qualité de la classification finale. En effet, après une phase d'extraction et d'étiquetage des noyaux, nous utiliserons un algorithme de classification qui consiste à assigner les observations au noyau le plus proche, ce qui impose que les noyaux soient représentatifs des modes. Si les noyaux sont mal définis, la classification sera entachée d'erreurs plus ou moins importantes.

II - UTILISATION D'UN FILTRE DE TYPE MEDIAN

L'objectif de la procédure de filtrage de la fonction de densité de probabilité est de réduire, sinon d'éliminer, les pics dans les vallées de la fonction et les trous dans ses modes, ce qui permet de faire ressortir les modes de cette fonction.

L'utilisation d'un filtre moyenneur a été vite rejetée, ce dernier ayant tendance à écraser les modes et à combler les vallées. C'est la raison pour laquelle nous nous sommes tournés vers l'utilisation d'un filtre de type médian. Très utilisé en traitement du signal [RAB75] et de l'image [DAV88], ce filtre est connu pour ses capacités à réduire les irrégularités locales des fonctions tout en préservant leurs contours [JUS78].

Ces propriétés ont déjà été utilisées en analyse de données [TOU87], mais ce filtre a, jusqu'à présent, toujours été utilisé pour améliorer les qualités de l'estimateur de Parzen-Rosenblatt, calculé sur des grilles hypercubiques.

Nous nous proposons d'adapter cette technique de filtrage à la fonction de densité de probabilité estimée par la méthode des k plus proches voisins, telle quelle est exposée au chapitre précédent.

Tel qu'il sera utilisé pour filtrer une fonction de densité de probabilité, le filtrage médian consiste à remplacer la valeur de la fonction en un point par la valeur médiane de la fonction dans le voisinage du point considéré [JUS81].

Avec l'estimation calculée sur une grille hypercubique, les v voisins de X_0 sont en général les points d'échantillonnage de la fonction de densité situés dans un voisinage hypercubique centré en X_0.

Dans le cas de l'estimateur des k plus proches voisins, disponible uniquement au niveau des observations (Cf. chapitre II), cette notion de voisinage doit être adaptée à celle utilisée pour l'estimation elle-même.

Pour être plus précis, soient X_1, X_2, ..., X_v les v plus proches points voisins du point X_0 où on estime $p(X_0)$ par $\hat{p}(X_0)$. *A priori*, le nombre v de voisins pris en compte pour effectuer le filtrage n'est pas obligatoirement égal au nombre k de voisins pris en compte pour estimer la fonction de densité de probabilité. Néanmoins, en prenant v=k, les résultats obtenus sur les exemples traités dans ce travail nous donnent entièrement satisfaction. De plus, le fait de prendre v=k réduit le nombre de paramètres à régler lors de la classification d'un échantillon. On note M, l'ensemble des estimateurs des v observations voisines auxquelles on ajoute $\hat{p}(X_0)$.

La valeur filtrée de $\hat{p}(X_0)$, notée $\hat{p}^*(X_0)$, prend la forme :

$$\hat{p}^*(X_0) = \text{MEDIAN}\left\{\hat{p}(X_0),\hat{p}(X_1),\hat{p}(X_2),...,\hat{p}(X_v)\right\}$$

Cette opération consiste donc à ordonner les valeurs de $\hat{p}(X_0)$ à $\hat{p}(X_v)$ par ordre croissant ou décroissant, et à remplacer $\hat{p}(X_0)$ par la valeur située au milieu de cette suite ordonnée.

Si le nombre d'éléments dans M est impair :

$$\hat{p}^*(X_0) = \hat{p}\left(X_{\frac{v+1}{2}}\right)$$

Si le nombre d'éléments dans M est pair :

$$\hat{p}^*(X_0) = \frac{\hat{p}\left(X_{\frac{v+1}{2}}\right) + \hat{p}\left(X_{\frac{v}{2}}\right)}{2}$$

Les v voisins X_1, X_2, ..., X_v utilisés sont, dans l'approche proposée, les v voisins de X_0. Aucune autre notion de voisinage n'est *a priori* disponible. Lors de la phase de filtrage, on utilise la même notion de voisinage géométriquement variable que celle sur laquelle est basée l'estimation de la fonction de densité de probabilité.

Cette manière de procéder permet une adaptation du filtre à la densité locale de la distribution, comme c'est le cas pour l'estimation de la fonction de densité de probabilité. Cependant, l'application directe de cette technique de filtrage ne donne pas toute satisfaction car, si ce filtre permet d'obtenir des résultats corrects quant à l'élimination du bruit là où la répartition des observations est relativement homogène, c'est à dire dans les modes, cette procédure ne convient pas dans les régions presque vides d'observations de l'espace de représentation des données.

En effet, il apparaît immédiatement que les valeurs de l'estimateur associées aux observations isolées se trouvent augmentées, alors que nous recherchons l'effet opposé (Cf. fig. II.a.). Ce phénomène s'explique aisément car la constitution de l'ensemble des v voisins d'un point X_0 situé dans une telle région conduit à inclure des observations situées dans des régions modales, ou qui en sont proches, là où les estimateurs sont relativement plus grands que la valeur de l'estimateur associé à l'observation isolée considérée. Ceci revient à dire que la plupart des valeurs de

l'ensemble $\{\hat{p}(X_i), i = 1,...,v\}$ sont supérieures à la valeur de l'estimateur au point X_0. La procédure de calcul de la valeur médiane fournit donc un résultat supérieur, et même très supérieur dans le cas d'une observation très éloignée d'un mode, à la valeur de l'estimateur originel.

Il nous faut donc apporter un élément correcteur dans le calcul de la réponse du filtre pour les observations isolées.

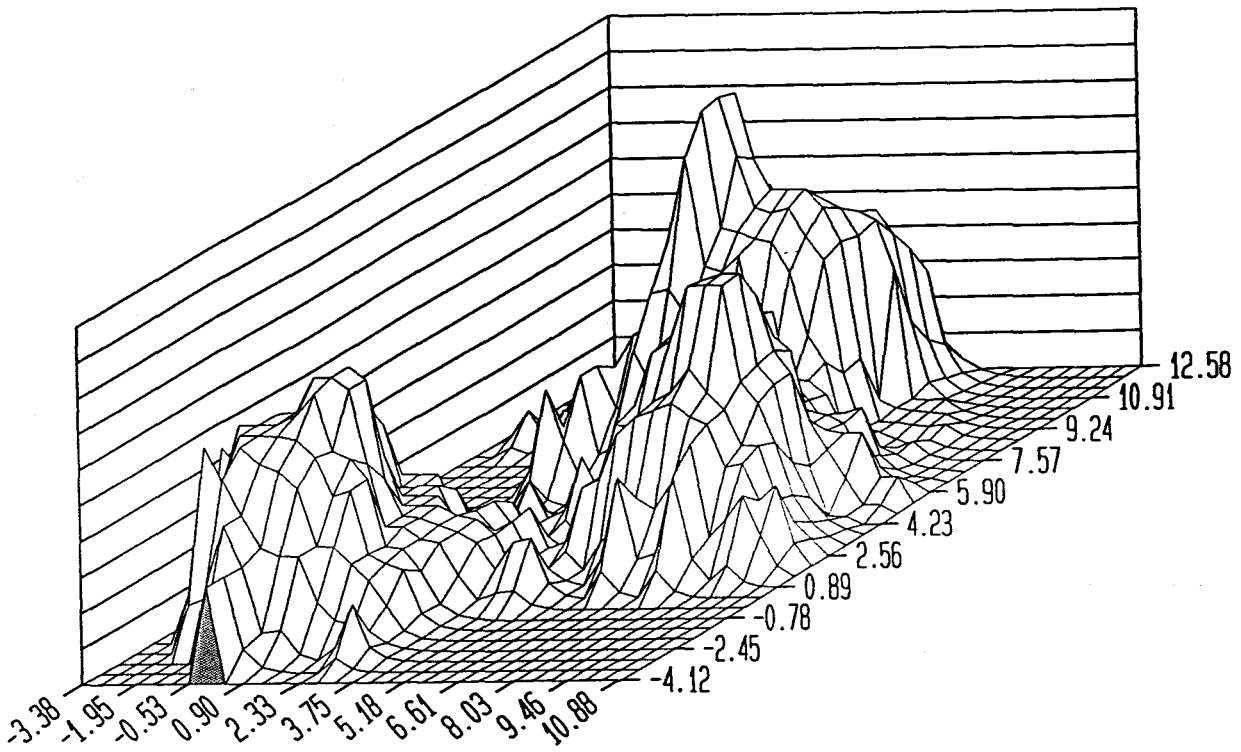


Figure II.a. Application d'un filtre médian classique sur l'estimation de la fonction de densité de probabilité de l'exemple 1 du chapitre II.

III. - FILTRE MEDIAN A PONDERATION BINAIRE

III. 1. - POSITION DU PROBLEME

Une première solution pour corriger la procédure de filtrage afin de réduire le phénomène exposé ci-dessus, pourrait être une pondération des estimations intervenant dans le calcul de la réponse du filtre médian. Néanmoins, après plusieurs essais avec différentes fonctions de pondération, une telle solution ne nous

a pas donné satisfaction. Sur la figure III.1.a. on a représenté l'effet du filtrage en prenant une fonction de la forme :

$$\hat{p}^{*}(X_0) = \text{MEDIAN}\{\hat{p}(X_0),\hat{p}(X_1),\hat{p}(X_2),...,\hat{p}(X_v)\}\cdot\hat{p}(X_0)$$

On constate que les valeurs des estimateurs des observations situées dans les vallées de la fonction de densité de probabilité sont bien atténuées; mais les modes ne sont pas suffisament mis en évidence pour permettre une exploitation facile.

Les résultats obtenus en pondérant directement les estimateurs intervenant dans le calcul du médian amenaient des problèmes de divers types qui peuvent être répartis en deux catégories :

– Soit la fonction de pondération permet de séparer correctement les modes en ne laissant apparaître que leur noyaux. Mais d'une part, elle ne conserve pas suffisamment la forme des contours pour que les noyaux puissent être considérés comme représentatifs des modes, et donc des classes, et, d'autre part, les classes à petit nombre d'observations sont pratiquement effacées de la distribution.

– Soit la fonction de pondération donne satisfaction quant à la conservation des classes de petites tailles, mais la séparation des modes est insuffisante pour permettre une classification correcte, surtout dans le cas de classes à chevauchements relativement importants.

Ces constatations montrent qu'il faut trouver une solution qui permet de "creuser" les vallées de la fonction de densité de probabilité sans pour autant réduire de façon significative la taille des noyaux des modes. De plus, le filtre utilisé doit amener l'estimateur associé aux observations isolées à une valeur très faible pour ne pas perturber le processus de classification.

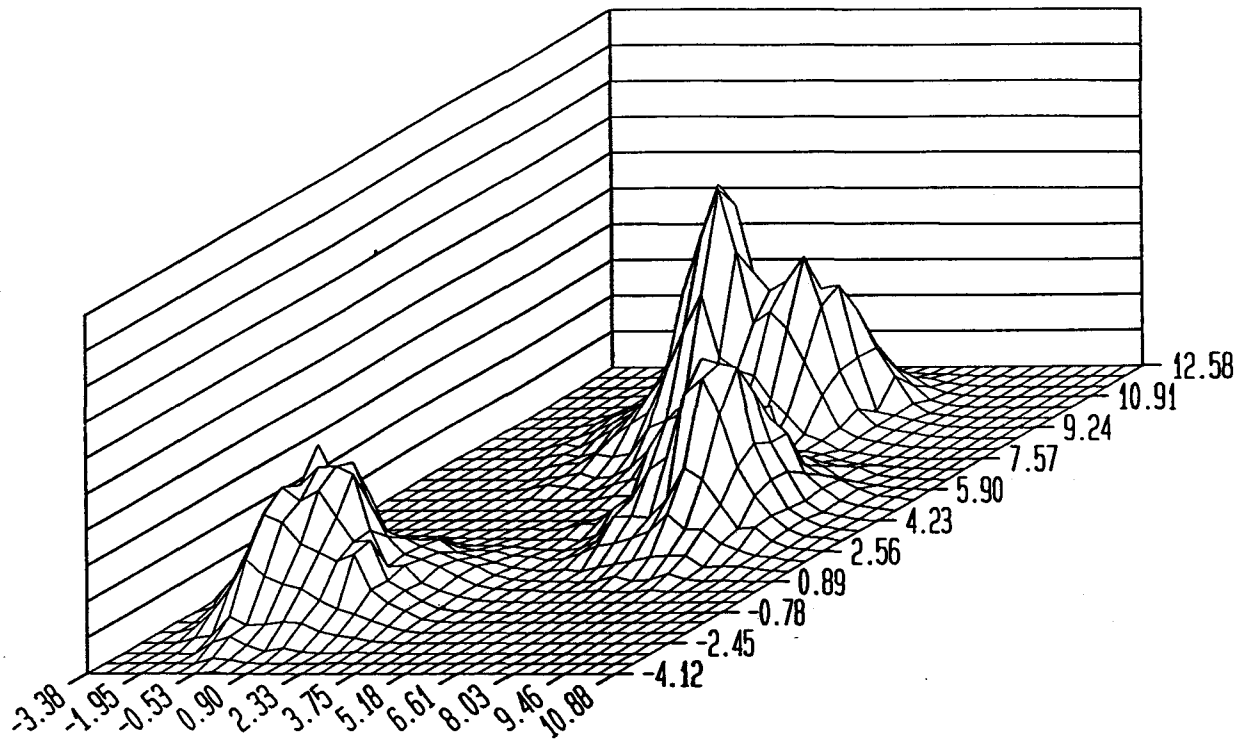


Figure III.1.a. Utilisation d'une pondération dans le calcul du médian

III.2. - ALGORITHME ADOPTE

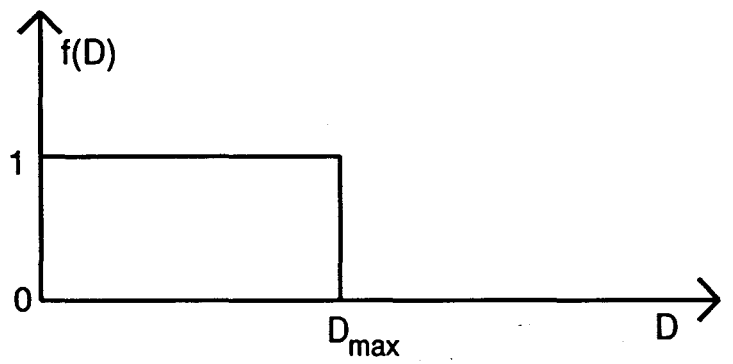
Les essais effectués avec des fonctions pour pondérer les estimateurs des k plus proches voisins entrant dans le calcul de la réponse du filtre médian amènent la constatation suivante : pour réussir à séparer les modes les uns des autres sans éliminer en même temps les classes à faible nombre d'observations, il est impératif d'adopter une stratégie de pondération à plusieurs paliers pour faire intervenir les estimateurs de manière différente selon leur éloignement par rapport au point où on effectue le filtrage. L'algorithme adopté utilise une pondération à deux niveaux, "0" ou "1".

Parmi les v plus proches voisins, les estimateurs associés à ceux qui sont situés à une distance inférieure à une valeur D_{max}, qui sera précisée ultérieurement, conserveront leurs valeurs initiales, tandis que les estimateurs correspondant aux voisins éloignés d'une distance supérieure à D_{max} seront considérés à valeurs nulles dans le calcul du médian.

Ceci peut être écrit sous la forme :

$$\hat{p}^{*}(X_0) = \text{MEDIAN}\{\hat{p}(X_0)\cdot f(D),\ldots,\hat{p}(X_v)\cdot f(D)\}$$

avec f(D) de la forme :



Cette pondération binaire permet, dans le cas d'observations isolées, de ne pas prendre en compte les estimateurs associés aux voisins éloignés qui perturberaient le processus de calcul du médian. En effet, les voisins d'une observation isolée sont généralement relativement éloignés de cette observation et ont une forte probabilité d'être proche d'un mode. Les valeurs des estimateurs qui leurs sont associés sont alors plus élevées que celle correspondant à l'observation isolée elle-même, et auraient pour effet de relever la valeur filtrée de l'estimateur au niveau de l'observation isolée.

La distance D_{max} qui intervient dans le processus d'évaluation du médian est calculée à partir des coordonnées extrêmes de l'espace de représentation des données. Cette distance D_{max} est évaluée comme une fraction de la distance euclidienne D_e qui sépare les deux observations les plus éloignées l'une de l'autre, parmi toutes celles disponibles. On prend : $D_{max} = \alpha\ D_e$, avec $\alpha < 1$. Il n'existe pas de méthode objective pour ajuster le coefficient α, mais l'expérience de l'analyste permet un ajustement satisfaisant car cette valeur n'est pas critique eu regard de l'échantillon qui lui est soumis.

Le calcul de la distance D_e nécessiterait *a priori* le calcul de toutes les distances entres les observations prises deux à deux, pour ne retenir que la valeur la plus élevée. Cette manière de procéder engendrerait des temps de calcul prohibitifs,

surtout dans le cas d'échantillons de grande taille et de grande dimension. Le problème du calcul de la distance D_e peut cependant être résolu de manière beaucoup plus intéressante en ne considérant que les coordonnées minimales et maximales suivant chaque axe de l'espace de représentation des données, tout en gardant en mémoire les indices des observations pour lesquelles ces coordonnées sont déterminées. L'algorithme de calcul de la distance D_e peut être décrit ainsi :

1) Recherche des coordonnées minimales et maximales $C_{min,i}$ et $C_{max,i}$, respectivement coordonnées suivant chaque axe, i=1,...,N, où N est la dimension de l'espace de représentation. Simultanément on mémorise les indices des observations correspondant à ces coordonnées minimales et maximales.

2) Calcul de toutes les distances euclidiennes entre les observations précédemment mémorisées.

3) Détermination de la distance D_e la plus élevée.

Lors du calcul du médian, l'ensemble M ordonné est constitué ainsi : $M=\{0,...,0,\hat{p}(X_i),...,\hat{p}(X_j)\}$ avec $j \leq v$. Le sous-ensemble $\{\hat{p}(X_i),...,\hat{p}(X_j)\}$ correspond aux estimateurs des observations voisines du point d'étude X_0, situées à une distance inférieure à D_{max} de ce point. Il se peut, et c'est le cas vers le centre d'un mode, que les k plus proches voisins de l'observation X_0 soient situés à une distance inférieure à D_{max}. Dans ce cas, l'ensemble M est égal à l'ensemble $\{\hat{p}(X_i),...,\hat{p}(X_j)\}$. A l'opposé, quand X_0 est totalement isolé, l'ensemble M ordonné s'écrit $\{0,0,...,0,\hat{p}(X_0)\}$ où $\hat{p}(X_0)$ est l'estimateur associé à X_0.

Les figures III.2.a. et b. montrent que l'introduction d'une pondération binaire prenant en compte la distance entre les voisins et le point où on filtre la fonction de densité de probabilité, permet de mettre correctement en évidence les noyaux des modes.

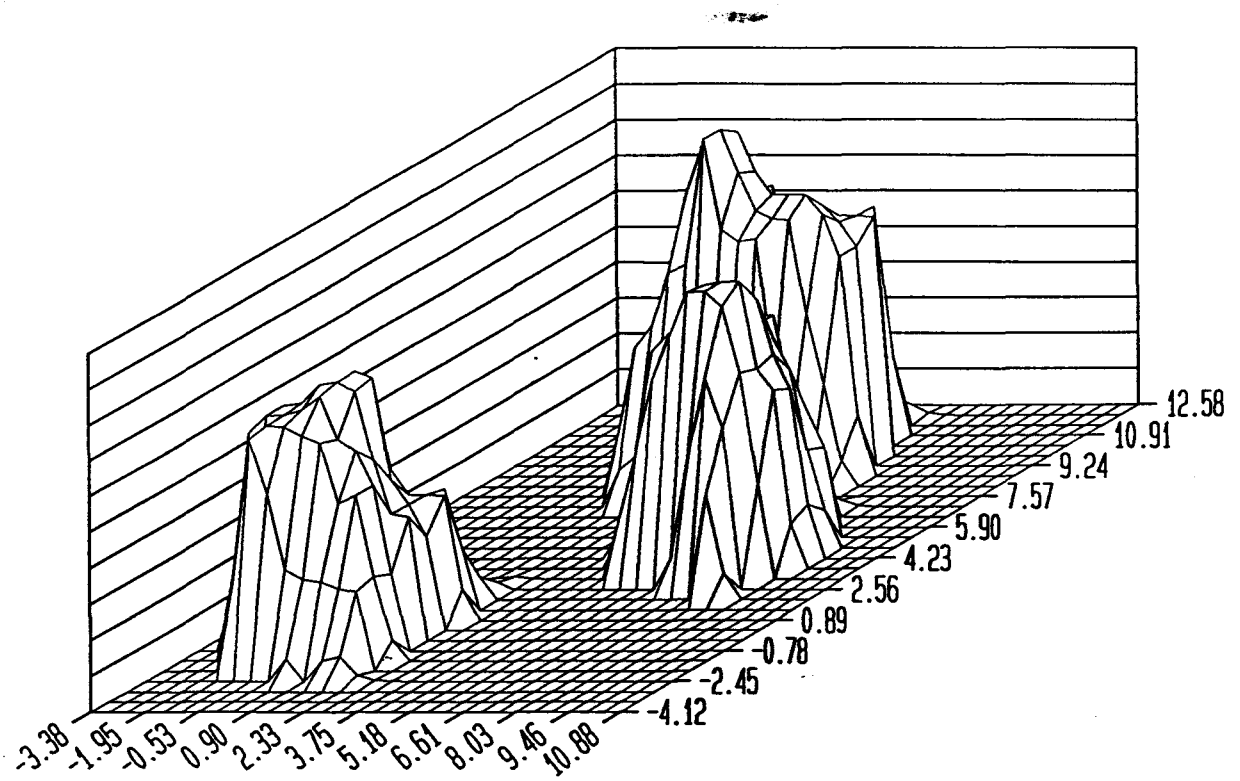


Figure III.2.a. Effet du filtrage à pondération binaire sur l'estimation de la fonction de densité de probabilité de l'exemple 1 du chapitre II.

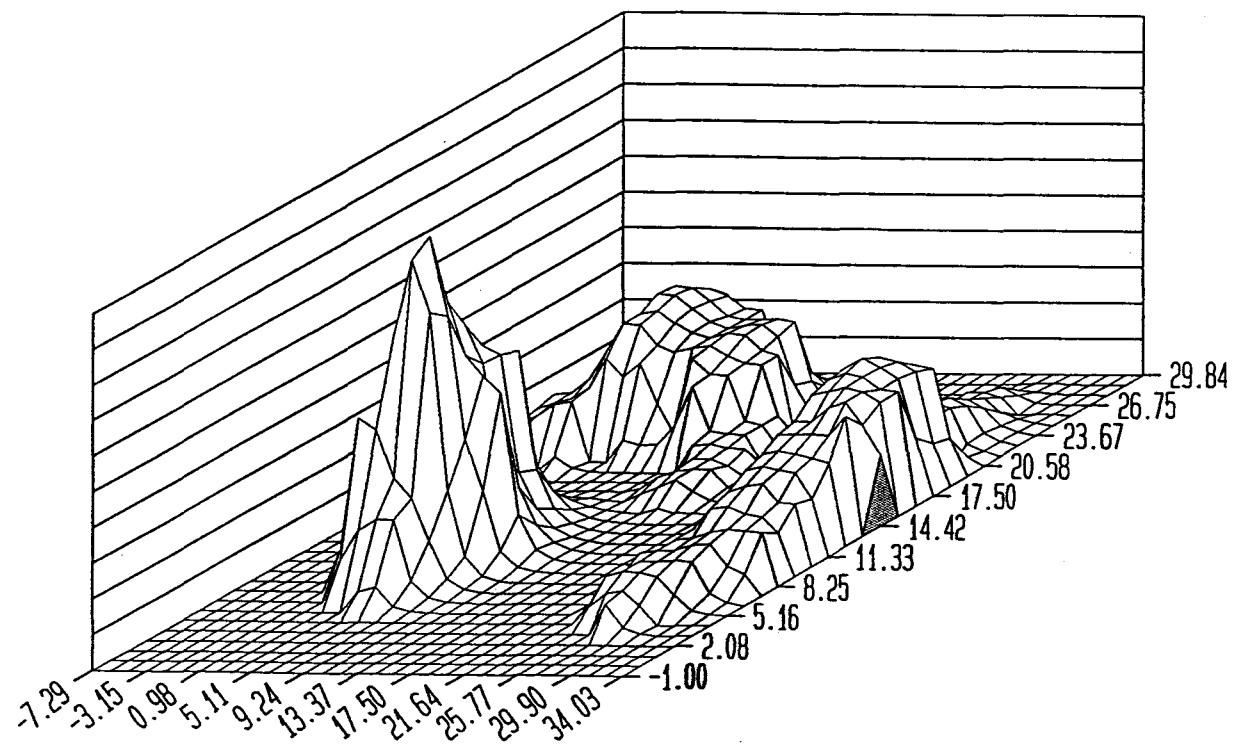


Figure III.2.b. Effet du filtrage à pondération binaire sur l'estimation de la fonction de densité de probabilité de l'exemple 2 du chapitre II.

III.3. - FILTRE MEDIAN ADAPTE A LA DENSITE LOCALE

Une amélioration peut être apportée au niveau de la détermination de la distance D_{max} calculée comme une fraction de la distance D_e séparant les deux observations les plus éloignées l'une de l'autre dans l'espace de représentation des données. En effet, dans le cas de fortes disparités de concentration des observations entre les différentes classes en présence dans l'échantillon, ou dans le cas de classes à fortes densités d'observations qui se trouvent très éloignées les unes des autres, une distance D_{max} calculée comme précédemment, ne permet pas toujours de mettre en évidence de façon correcte les noyaux des modes. En effet, cette distance D_{max} est identique en tous les points où on filtre l'estimateur, et elle ne prend en compte que la dispersion globale des observations disponibles dans l'espace de représentation. Afin d'apporter une amélioration dans le choix de D_{max}, il serait intéressant de prendre en considération la densité locale de la distribution et non plus sa dispersion globale. La densité locale moyenne peut être appréciée en calculant l'écart type des distances entre le point d'étude X_0 et ses v plus proches voisins. Cette valeur peut alors apporter une notion d'adaptation du filtre au point où on effectue le filtrage.

La distance D_{max} peut être évaluée sous la forme :

$$D_{max} = D_v - \frac{\sigma}{\hat{p}(X_0)}$$

où D_v est la distance entre l'observation X_0 et le $v^{ième}$ voisin
σ est l'écart-type des distances entre X_0 et ses v plus proches voisins.

De cette manière, lorsqu'on se trouve en un point isolé, l'écart-type des distances est assez proche de la distance d_v, ce qui a pour effet d'inclure beaucoup de valeurs nulles dans la constitution de l'ensemble M et induit donc une valeur nulle pour l'estimateur filtré en ce point. D'autre part, plus on se rapproche du centre d'un mode, plus la distance D_{max} est proche de d_v, ce qui a pour effet, cette fois, de prendre en compte la quasi-totalité des estimateurs associés aux v plus proches voisins. De plus, la distance D_{max} nouvellement définie ne comportant aucun coefficient arbitraire, elle ne requiert plus l'expérience de l'analyste pour son calcul,

et la phase de filtrage s'en trouve alors d'autant plus facile à mettre en oeuvre de manière automatique.

IV. - DETECTION ET ETIQUETAGE DES NOYAUX

IV.1. - SEUILLAGE DE LA FONCTION DE DENSITE DE PROBABILITE

La procédure de filtrage permet de renforcer les modes de la fonction de densité de probabilité. Ce renforcement peut se justifier à plusieurs titres. D'une part, les valeurs filtrées des estimateurs dans les régions modales sont plus élevées que les estimateurs bruts de la fonction de densité de probabilité, et la fonction de densité filtrée présente également une allure plus lissée. D'autre part, les différents noyaux se trouvent suffisamment séparés les uns des autres par le fait que les estimations associées aux observations qui se situent dans les vallées de la fonction de densité de probabilité voient leurs valeurs ramenées vers zéro. La faculté de préserver les contours des noyaux des modes conforte d'autre part le choix d'un filtrage de type médian.

L'étape suivante du processus de classification consiste à identifier les modes en présence dans la distribution et à leur affecter des étiquettes différentes. Il est alors utile d'éliminer les estimations de faible amplitude qui peuvent être considérées comme non significatives, et qui pourraient perturber la phase de détection et d'étiquetage des noyaux. C'est le cas notamment des estimations associées aux observations situées entre deux noyaux relativement proches l'un de l'autre. Ceux-ci se verraient éventuellement réunis alors qu'ils ne le devraient pas.

Toutes les estimations inférieures à la valeur du seuil sont ramenées à zéro, alors que celles supérieures ou égales au seuil sont amenées à la valeur "1". Ce seuillage permet de distinguer deux types d'observations. Celles marquées de la valeur "1", que nous appelons "observations modales" car elles ont de fortes chances d'appartenir à un mode, et celles marquées de la valeur "0", que nous dénommerons "observations non modales" car elles ont de fortes chances d'appartenir à des vallées situées entre les modes.

Comme dans tout problème de seuillage, l'ajustement de la valeur du seuil demeure un problème délicat. Dans notre cas, le réglage de ce paramètre n'est pas

critique. En effet, en observant les figures III.2.a. et b. représentant des fonctions de densité de probabilité filtrée, on s'aperçoit que les valeurs des estimations sont généralement assez élevées pour permettre à l'analyste de fixer sans grande précaution une valeur au seuil sans risquer d'engendrer de difficultés pour les phases suivantes de la classification. Sur les figures IV.1.a à h nous avons représenté, pour les exemples 1 et 2, l'estimation de la fonction de densité de probabilité filtrée et binarisée pour différentes valeurs du seuil. Ce seuil est pris comme une fraction de la valeur de l'esimateur le plus élevé.

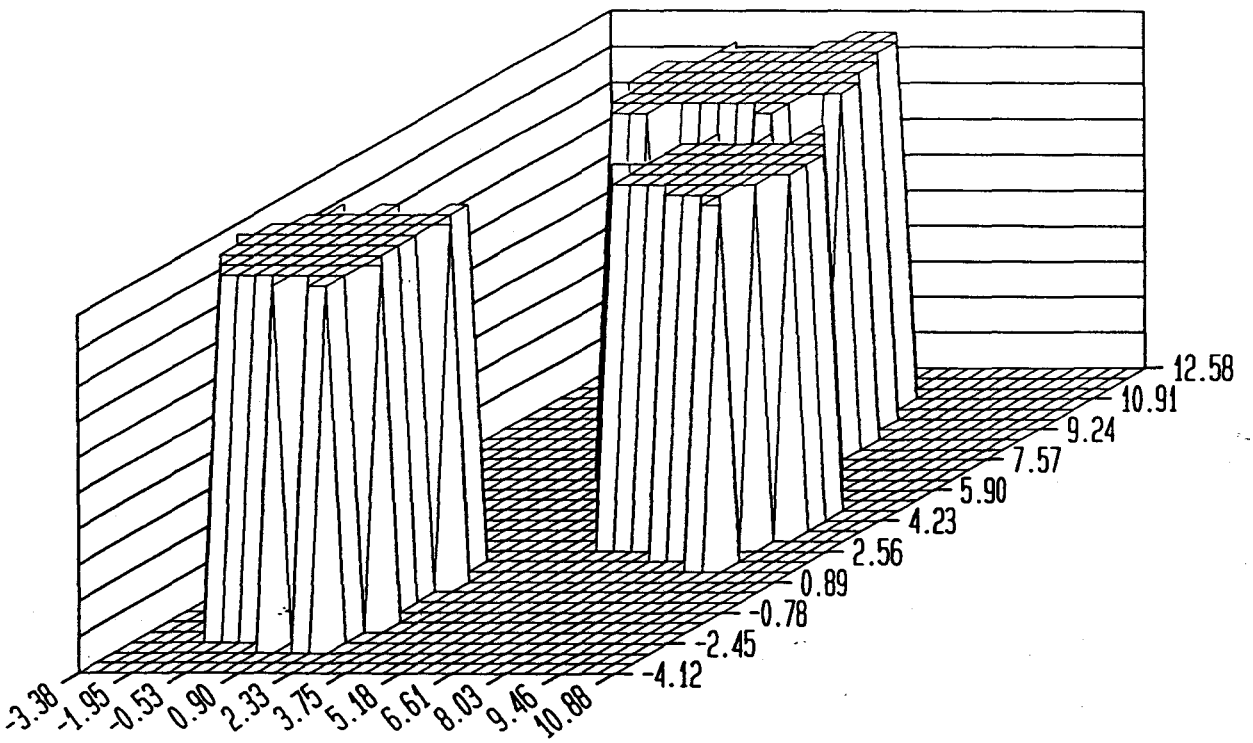


Figure IV.1.a. Binarisation de la f.d.p. filtrée avec un seuil de 5% (exemple 1).

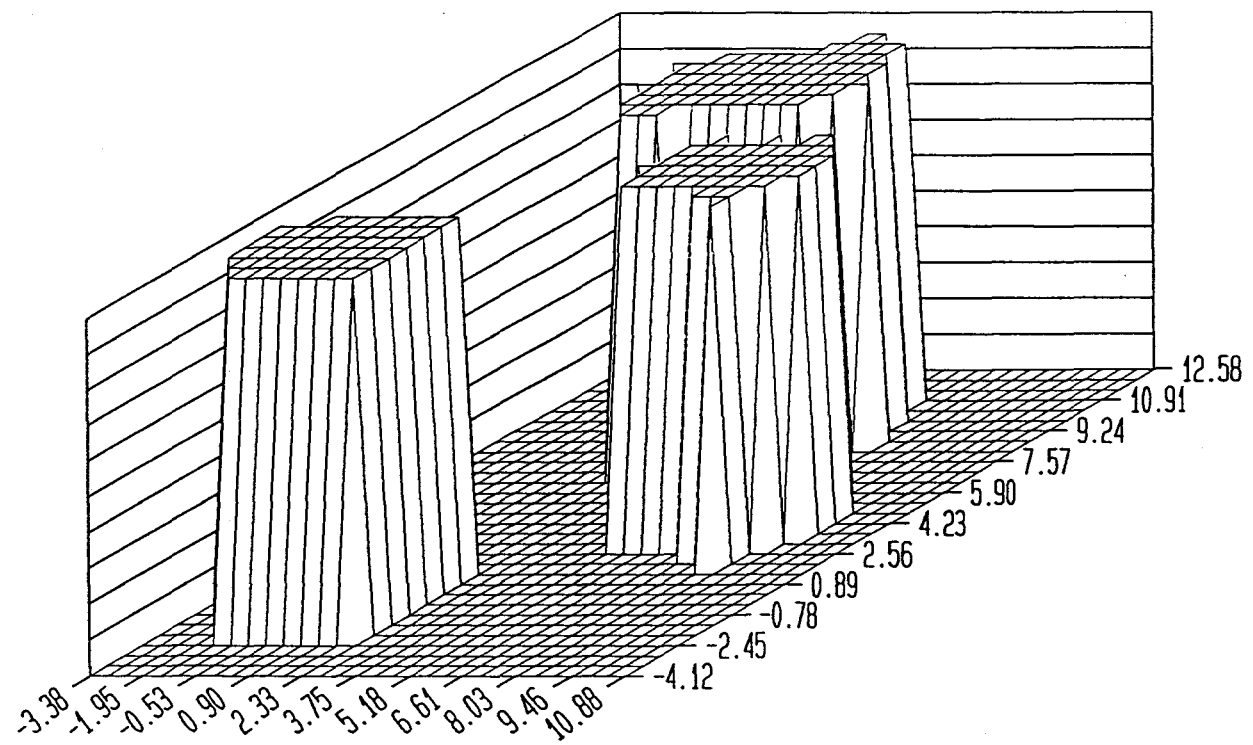


Figure IV.1.b. Binarisation de la f.d.p. filtrée avec un seuil de 10% (exemple 1).

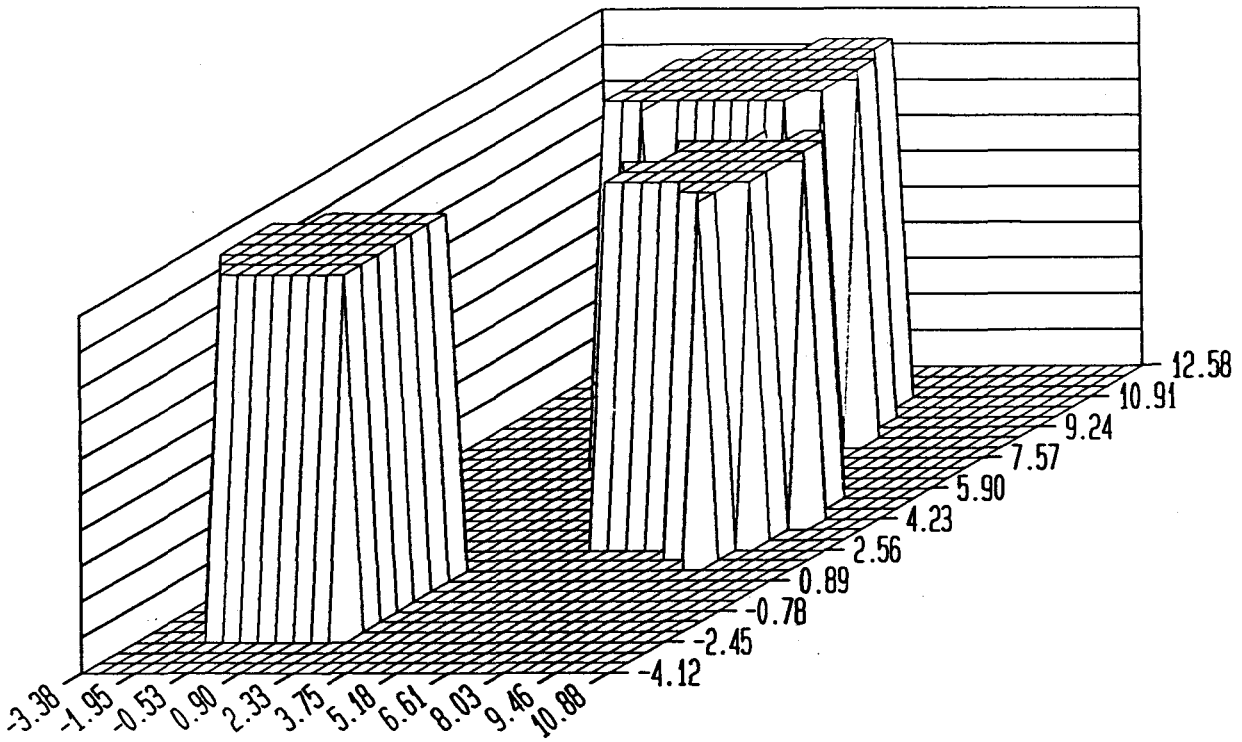


Figure IV.1.c. Binarisation de la f.d.p. filtrée avec un seuil de 15% (exemple 1).

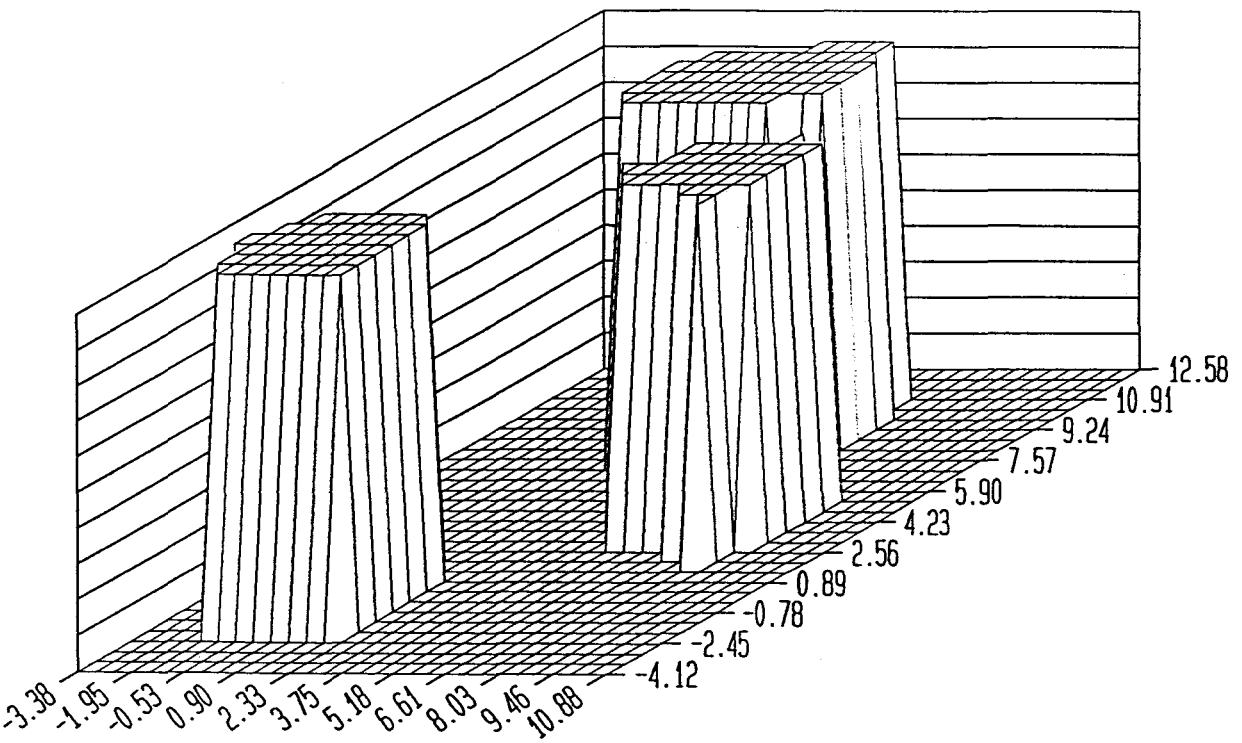


Figure IV.1.d. Binarisation de la f.d.p. filtrée avec un seuil de 20% (exemple 1).

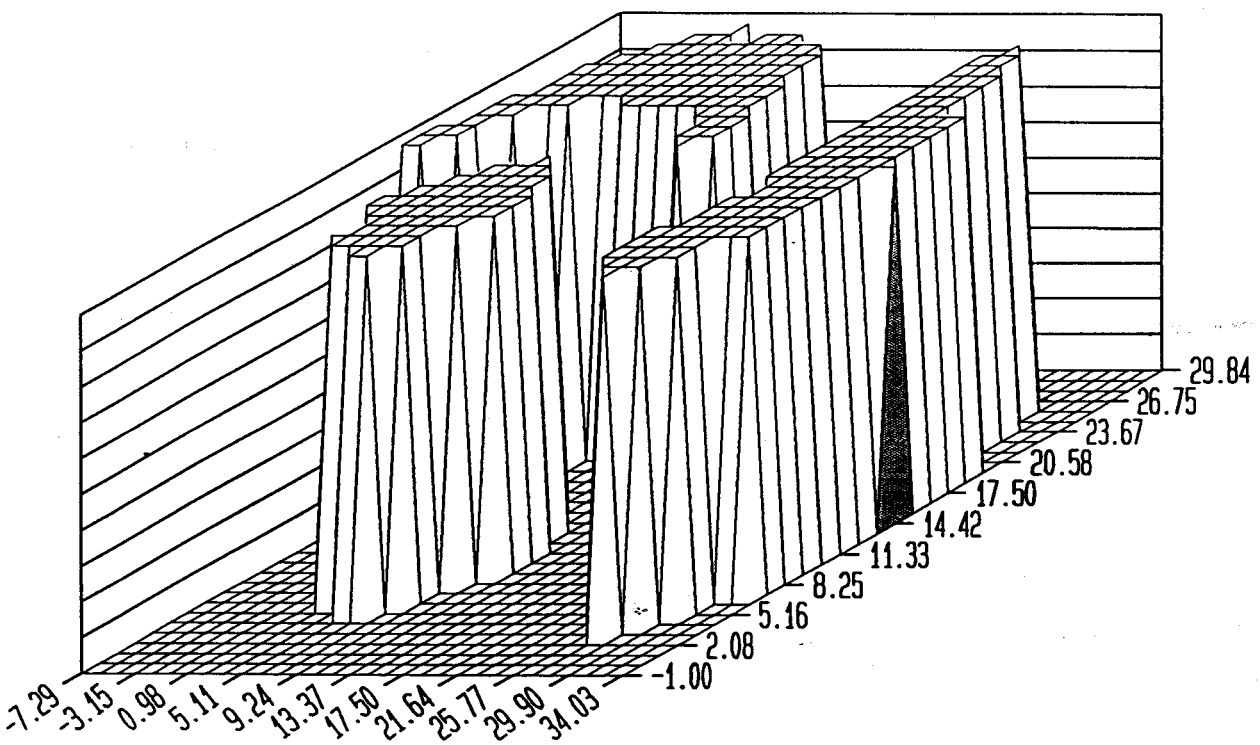


Figure IV.1.e. Binarisation de la f.d.p. filtrée avec un seuil de 5% (exemple 2).

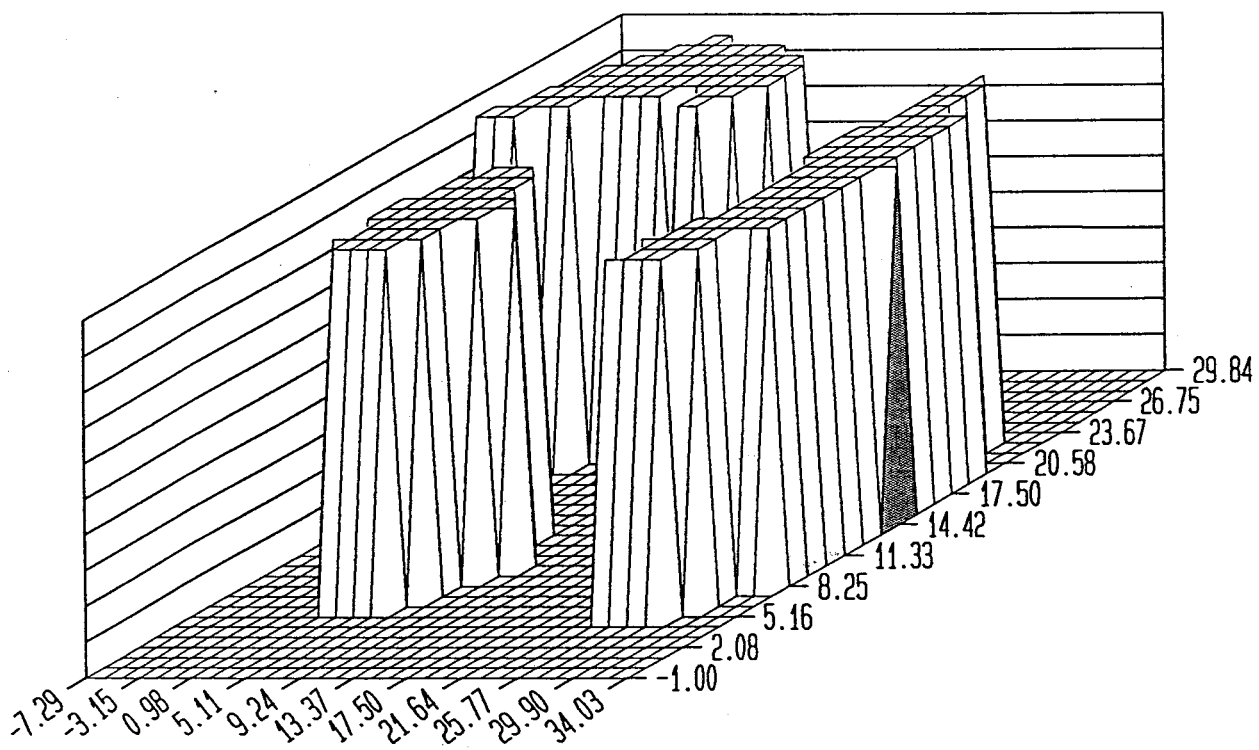


Figure IV.1.f. Binarisation de la f.d.p. filtrée avec un seuil de 10% (exemple 2).

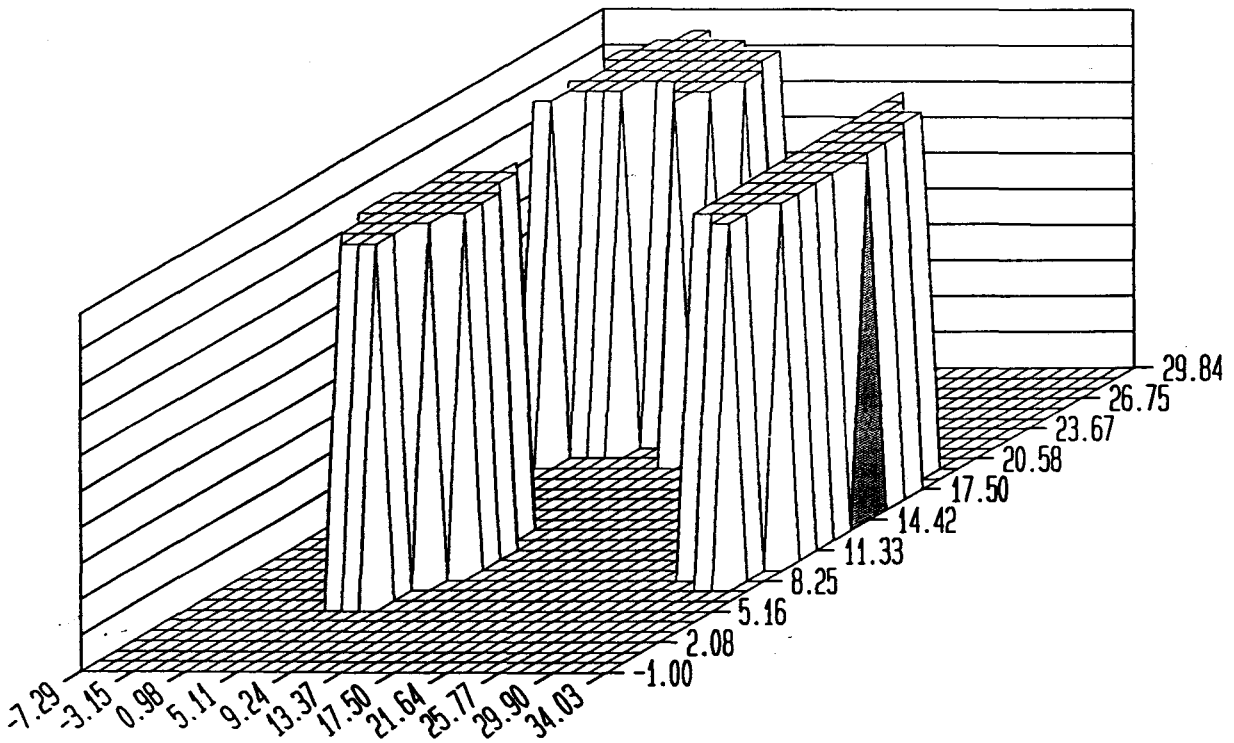


Figure IV.1.g. Binarisation de la f.d.p. filtrée avec un seuil de 15% (exemple 2).

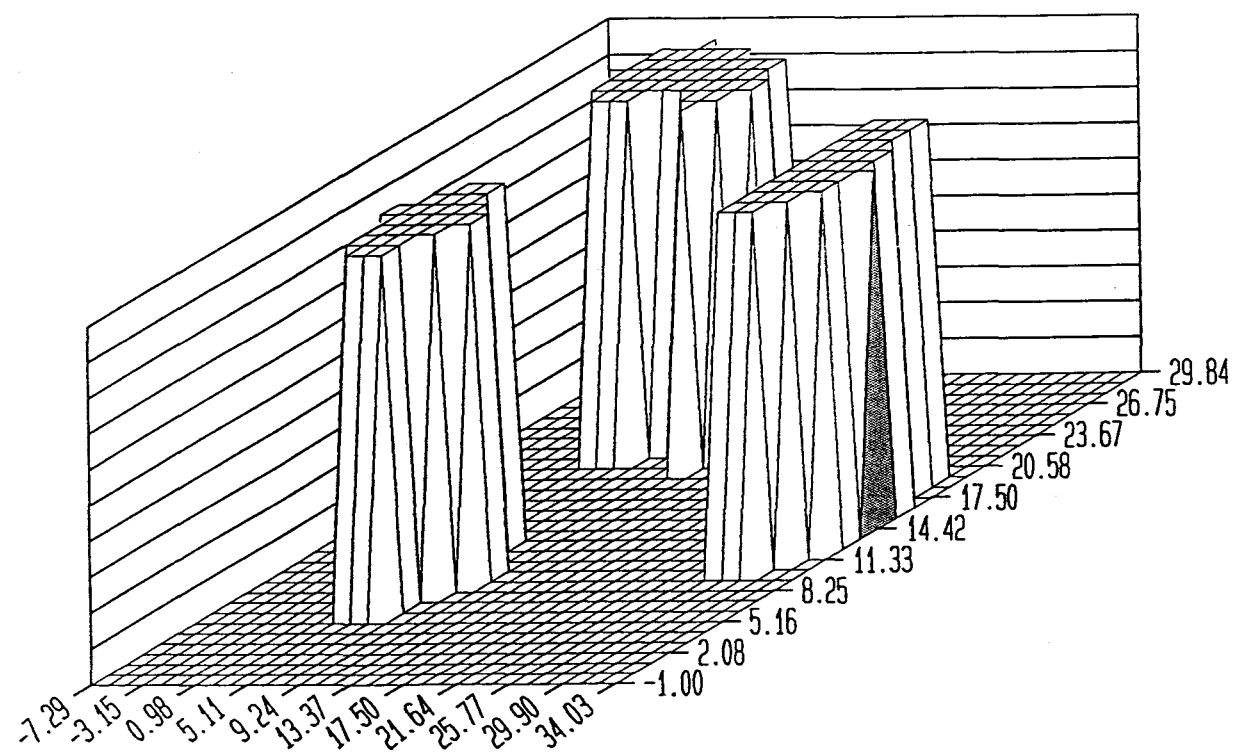


Figure IV.1.h. Binarisation de la f.d.p. filtrée avec un seuil de 20% (exemple 2).

IV.2. - DETECTION ET ETIQUETAGE DES NOYAUX

La procédure employée permet d'effectuer simultanément les opérations de détection des noyaux et leur étiquetage. Les points où la fonction de densité de probabilité est estimée étant les observations elles-mêmes, on recherche dans la liste des observations disponibles la première observation modale et on lui assigne l'étiquette "1". Cette observation ainsi étiquetée constitue le point de départ pour la formation du premier noyau.

La constitution du noyau complet exploite la notion de voisinage du point étiqueté. Parmi les m plus proches voisins de l'observation étiquetée, on attribue la même étiquette aux observations "modales", les observations "non modales" appartenant à cet ensemble des m plus proches voisins sont ignorées. Ici également, on peut prendre pour m la valeur attribuée à k lors de l'estimation de la fonction de densité de probabilité. Ces observations nouvellement étiquetées deviennent alors autant de points de départ pour la suite de l'opération d'étiquetage. L'attribution d'une étiquette à partir de ces nouveaux points se fait de la même manière en considérant le voisinage de ces points. Ainsi, de proche en proche, on agrège les observations "modales" qui forment un même noyau. La procédure d'agrégation se termine dès que dans les voisinages des observations étiquetées il ne reste plus d'observations "modales" non encore étiquetées. Le premier noyau se trouve ainsi identifié et les observations qui le constituent portent une même étiquette. L'algorithme décrit est réinitialisé à partir d'une observation "modale" non encore étiquetée. Les noyaux sont ainsi formés les uns après les autres.

Cet algorithme permet, par une simple procédure d'agrégation et d'étiquetage, de détecter les noyaux de la fonction de densité de probabilité. Les figures IV.2.a. et b. montrent les noyaux étiquetés des échantillons correspondant aux exemples 1 et 2. Ces noyaux ne peuvent néanmoins être formés de façon correcte que si les observations qui se trouvent dans les "vallées" ont vu leurs estimations ramenées à une valeur nulle. La phase de filtrage est donc importante pour mener à bien l'étape de détection des noyaux. Le type de filtre choisi permet d'obtenir une démarcation importante entre les différents noyaux et assure une bonne détection des noyaux des modes de la fonction de densité de probabilité estimée, sous-jacente à la distribution des observations disponibles.

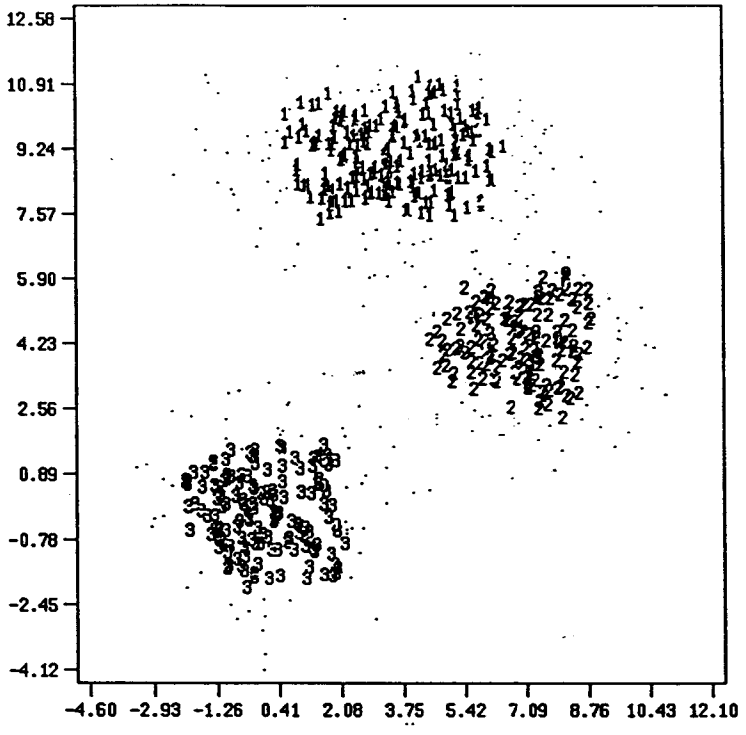


Figure IV.2.a. Noyaux étiquetés pour l'échantillon de l'exemple 1.

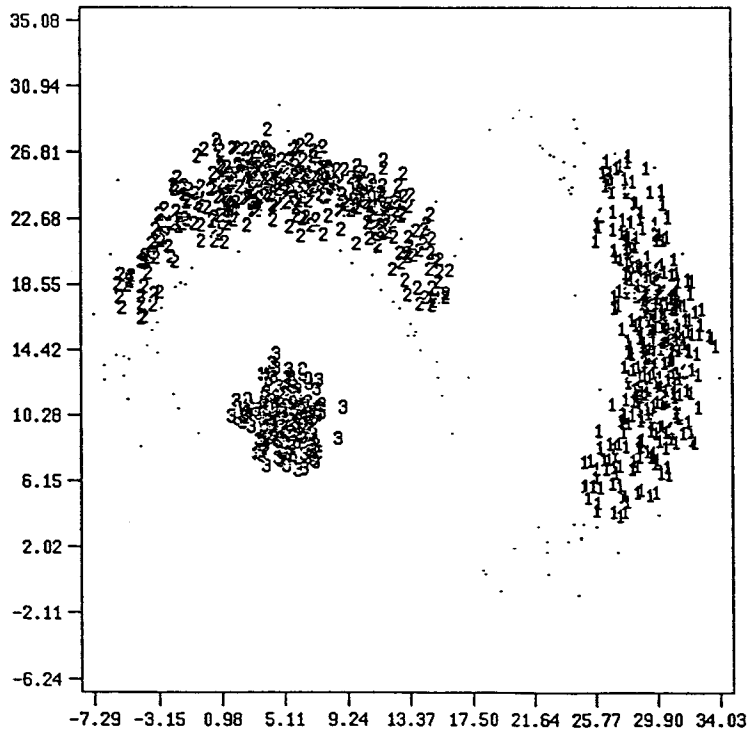


Figure IV.2.a. Noyaux étiquetés pour l'échantillon de l'exemple 2.

V. - ASSIGNATION DES OBSERVATIONS "NON MODALES"

La procédure d'étiquetage des "observations modales" a permis de distinguer les différents noyaux des modes présents dans la distribution des observations disponibles. Il reste maintenant à donner une étiquette aux observations qui n'ont pas encore été étiquetées par la procédure précédente. Il s'agit en fait des "observations non modales" qu'il faut assigner à un noyau, de manière à obtenir la classification finale de toutes les observations de l'échantillon analysé. On obtient alors les différentes classes constituant l'échantillon. L'algorithme d'assignation peut être décrit de la manière suivante :

On cherche parmi les observations disponibles, la première qui ne soit pas encore étiquetée. On recherche ensuite le plus proche voisin de celle-ci qui porte une étiquette "observation modale", à condition néanmoins que ce voisin soit inclus dans les m plus proches voisins de l'observation considérée, et on assigne à l'observation la même étiquette que "l'observation modale" voisine. La procédure est ainsi poursuivie pour toutes les observations non encore assignées. Dans le cas où parmi les m voisins d'une observation, il n'existe pas d'observation "modale", l'observation considérée est provisoirement non assignée. Pour les observations qui se trouvent dans ce cas, on réitère la procédure d'assignation en intégrant les observations déjà assignées dans celles notées "observations modales". On classe ainsi de proche en proche les observations disponibles dans l'échantillon analysé.

Les figures V.a. et b. montrent le résultat des assignations des "observations non modales" au noyau le plus proche pour les exemples 1 et 2.

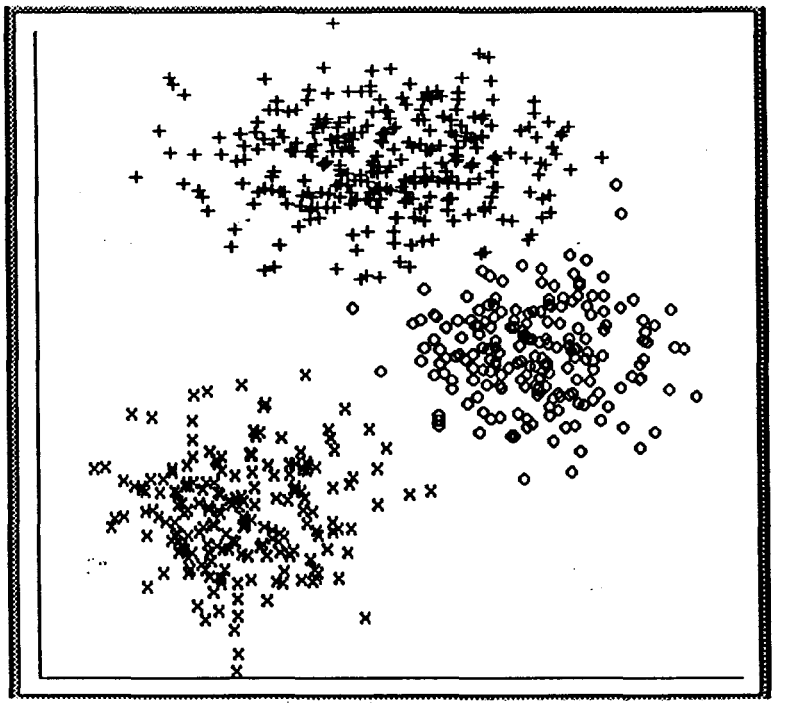


Figure V.a. Assignations des "observations non modales" , exemple 1 (m=20).

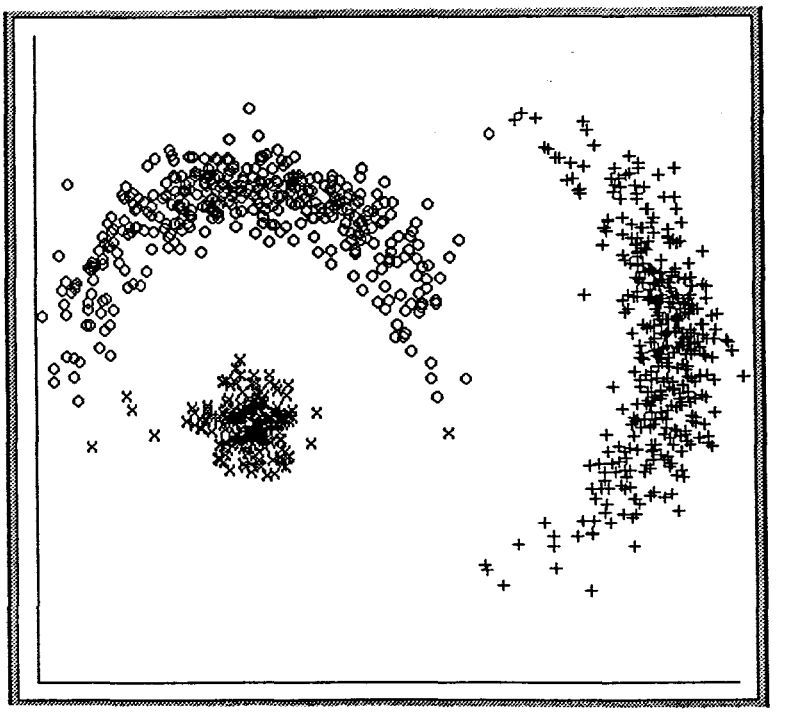


Figure IV.a. Assignations des "observations non modales", exemple 2 (m=20).

VI - CONCLUSION

Dans ce chapitre, nous avons vu que la mise en évidence des noyaux des modes de la fonction de densité de probabilité a été possible grâce à une opération de filtrage suivi d'un seuillage. Les problèmes posés par l'utilisation directe d'un filtre médian ont été levés avec succés par une pondération binaire de la contribution des estimateurs des k plus proches voisins du point où on estime la fonction de densité de probabilité.

Une procédure d'étiquetage itératif a ensuite permis d'attribuer une étiquette différente à chaque noyau, les observations constituant un même noyau portant alors la même étiquette. Finalement, en assignant chaque observation au noyau le plus proche, on obtient la classification de toutes les observations.

CHAPITRE IV

DETECTION DES CONTOURS DES MODES

DE LA FONCTION DE DENSITE DE PROBABILITE

CHAPITRE IV

DETECTION DES CONTOURS DES MODES

I. - GRADIENT DE LA FONCTION DE DENSITE DE PROBABILITE

I.1. - INTERET D'UN OPERATEUR DIFFERENTIEL

La mise en évidence des modes de la fonction de densité de probabilité sous-jacente à la distribution des données disponibles, a été, dans le chapitre précédent, basée sur l'extraction des noyaux de ces modes. Une autre approche, faisant appel à l'estimation de la fonction de densité de probabilité consiste, non plus à rechercher les noyaux des modes de cette fonction, mais le contour de ceux-ci.

En examinant la forme des fonctions de densité de probabilité lissée par le filtre à pondération binaire exposé dans le chapitre précédent, on peut remarquer aisément que la pente de la fonction est toujours importante au voisinage du contour du noyau des modes. Ces régions de la distribution où la fonction de densité de probabilité posséde un gradient important peuvent être mises à profit pour développer un processus de classification basé sur la détection des contours des modes.

Après la phase de filtrage de la fonction de densité de probabilité, la suite des opérations consiste à appliquer un opérateur différentiel sur cette fonction, pour mettre en évidence les régions où celle-ci présente de fortes variations locales. Le domaine du traitement des images numériques est riche d'opérateurs différentiels.

Parmi les plus connus on peut citer les opérateurs de Prewitt, Roberts, Sobel, le laplacien... [ROB65][PRE70][TAN79][MOD77]. Quelque soit celui que l'on choisit, le but à atteindre consiste toujours à augmenter le contraste de l'image afin de mettre en évidence, ou simplement renforcer, les contours des différentes régions significatives de la scène analysée afin d'en extraire les éléments utiles au traitement. Ces opérateurs ont chacun leur spécificité et s'appliquent le plus souvent sur des données bidimensionnelles, c'est à dire la fonction luminance de l'image numérique. Pour être employés en analyse de données, il est nécessaire de généraliser ces opérateurs aux cas multidimensionnels. Des travaux sur l'adaptation des opérateurs de Roberts et de Prewitt à une fonction de densité de probabilité estimée sur une grille hypercubique ont permis d'obtenir des résultats tout à fait satisfaisants [TOU87]. Dans notre cas, l'estimation de cette fonction ne met en jeu aucun maillage de l'espace de représentation des données. Une utilisation directe des travaux sur l'adaptation des opérateurs cités précédemment ne peut alors être envisagée.

Comme dans les phases de classification vues dans les chapitres précédents, nous nous dirigeons vers l'emploi d'un opérateur différentiel qui exploite les estimations des observations situées dans le voisinage du point X_0 où on désire évaluer le gradient de la fonction de densité de probabilité. Cette notion de voisinage est isotrope au sens de l'espace entourant le point d'étude X_0, c'est à dire que l'on tient compte de toutes les directions de l'espace autour de X_0, mais il ne l'est pas non au sens des distances entre ce point X_0 et ses voisins. L'absence de maillage de l'espace de représentation des données ne permet pas l'emploi d'un opérateur différentiel classique. En effet, tous les opérateurs qui consistent à estimer la valeur du gradient d'une fonction en un point, utilisent une discrétisation de cette fonction sur une grille régulière. La pente, ou gradient, de la fonction est ensuite estimée ou approchée à partir des valeurs que prend la fonction dans les points d'échantillonnage définis sur un voisinage formé par le maillage entourant le point d'étude.

L'exploitation directe des opérateurs différentiels existants ne peuvent pas être appliquée dans notre cas car on n'utilise pas de discrétisation de l'espace de représentation des données. Nous proposons alors un opérateur de détection de contours capable d'être appliqué sur une fonction de densité de probabilité estimée

par la méthode présentée au chapitre II, en s'affranchissant d'une grille de discrétisation.

II. - OPERATEUR DE DETECTION DES CONTOURS.

Nous proposons de rechercher les contours des modes de la fonction de densité de probabilité en mettant en évidence les points de l'espace où cette fonction présente les plus grandes variations sur les plus petites distances.

Soit X_0 le point où on désire quantifier ces variations. En comparant les variations de la fonction de densité de probabilité entre X_0 et un point X_i voisin, à la distance séparant ces deux points, exprimée comme une fraction de la taille du voisinage de X_0 défini par ses c plus proches voisins, on s'affranchit des difficultés levées par la taille variable de ce voisinage.

Pour être plus explicite, on calcule, pour chaque voisin X_i de X_0 la quantité :

$$\hat{g}(X_i) = \frac{\left|\hat{p}(X_i) - \hat{p}(X_0)\right|}{\frac{\|x_i - x_0\|}{\|x_c - x_0\|}}$$

où $\hat{p}(X_0)$ et $\hat{p}(X_i)$ sont les estimations de p(X) en X_0 et X_i.

Dans le cas où on utilise la norme euclidienne :

$$\|X_i - X_0\| = \left[\sum_{j=1}^{n} \left(x_{j,i} - x_{j,0}\right)^2\right]^{1/2}$$

et

$$\|X_c - X_0\| = \left[\sum_{j=1}^{n} \left(x_{j,c} - x_{j,0}\right)^2\right]^{1/2}$$

où X_c est le voisin le plus éloigné de X_0 parmi ses c plus proches voisins.

L'opérateur de détection de contour en X_0 prend alors la forme :

$$\hat{G}(X_0) = \frac{1}{c}\sum_{i=1}^{c}\hat{g}(X_i)$$

Dans l'application de cet opérateur de détection de contours, il faut en général prendre comme nombre c de voisin, une valeur inférieur au nombre k de voisins pris en compte lors des deux premières phases du traitement. C'est à dire : l'estimation de la fonction de densité de probabilité et son filtrage. En effet, l'opérateur proposé vise à détecter les plus grandes variations locales de la fonction de densité de probabilité, et donc un nombre c de voisins trop grand conduirait à ne pas prendre une information suffisamment locale.

La réponse de cet opérateur différentiel sur la fonction de densité de probabilité filtrée est cependant trop bruitée pour autoriser son exploitation directe (Cf. fig. IV.d.). Afin de réduire le bruit, il suffit d'appliquer un filtre, mais qui puisse néanmoins sauvegarder la forme des contours. Le filtre à pondération binaire présenté dans le chapitre précédent remplira tout à fait ce rôle. En effet, ce filtre a été proposé pour détecter les noyaux des modes car il présente l'avantage de conserver les contours tout en éliminant les irrégularités de la fonction de densité de probabilité. Ce filtre est donc tout à fait adapté pour réduire les irrégularités de la réponse de l'opérateur différentiel appliqué sur la fonction de densité de probabilité filtrée. Il ne faudrait pas effectivement annuler la réponse de cet opérateur par un filtre qui aurait tendance à écraser cette.

III. - CHAINAGE ET ETIQUETAGE DES CONTOURS

III.1. - SEUILLAGE DU GRADIENT DE LA FONCTION DE DENSITE DE PROBABILITE

L'application d'un opérateur différentiel sur la fonction de densité de probabilité filtrée a permis de démarquer les variations locales de cette fonction. Puis les irrégularités de la réponse du gradient ont été atténuées par l'adjonction d'une phase de filtrage. Il reste maintenant à identifier les contours et à leur attribuer une

étiquette différente. On effectue alors, un seuillage de la réponse de l'opérateur de détection de contours. Ainsi, les estimateurs inférieurs à un seuil prédéterminé sont amenés à la valeur "0", tandis que les autres prennent la valeur "1". Nous appelons alors "observations contours", les observations étiquetées "1", et "observations non-contours", celles étiquetées "0".

III.2. - CHAINAGE ET ETIQUETAGE DES CONTOURS

Comme dans le cas de la détection des noyaux du chapitre précédent, la procédure employée permet d'effectuer simultanément les opérations de détection des contours et leur étiquetage. Les points où la fonction de densité de probabilité est estimée étant les observations elles-mêmes, on recherche dans la liste des observations disponibles la première pour laquelle l'estimateur correspondant n'est pas nul, et on lui attribue l'étiquette "1". Cette observation ainsi étiquetée constitue le point de départ pour la formation du premier contour. La constitution du contour complet quant à elle, exploite la notion de voisinage du point étiqueté. Parmi les c plus proches voisins de l'observation étiquetée, on attribue la même étiquette aux observations "contours", les observations "non-contours" appartenant à cet ensemble des c plus proches voisins sont ignorées. Ces observations nouvellement étiquetées deviennent alors autant de point de départ pour la suite de l'opération d'étiquetage. L'attribution d'une étiquette à partir de ces nouveaux points se fait de la même manière en considérant le voisinage de ces points. Ainsi, de proche en proche, on agrège les observations "contours" qui forment un même contour. La procédure d'agrégation se termine dès que dans les voisinages des observations étiquetées il ne reste plus d'observations "contours" non encore étiquetées. Le premier contour se trouve ainsi identifié et les observations qui le constituent portent une même étiquette. L'algorithme décrit est réinitialisé à partir d'une observation "contours" non encore étiquetée. Les contours sont ainsi formés les uns après les autres.

IV. - ASSIGNATION DES OBSERVATIONS "NON CONTOURS"

La procédure d'étiquetage des "observations contours" a permis de distinguer les différents contours des modes présents dans la distribution des observations disponibles. Il reste maintenant à donner une étiquette aux "observations non-contours", ce qui permet alors de former les différentes classes

sous-jacentes à l'échantillon. Pour effectuer cette phase finale d'étiquetage, nous allons assigner les "observations non-contours" au contour le plus proche. L'algorithme peut alors être décrit de la manière suivante :

On cherche parmi les observations disponibles, la première qui ne soit pas encore étiquetée. On recherche ensuite le premier plus proche voisin de celle-ci qui porte une étiquette "contours", à condition néanmoins que ce voisin soit inclus dans les c plus proches voisins de l'observation considérée, et on assigne à l'observation la même étiquette que le voisin "contour". La procédure est ainsi poursuivie pour toutes les observations non encore assignées. Dans le cas où parmi les c voisins d'une observation, il n'existe pas d'observation "contour", l'observation considérée est provisoirement non assignée. Pour les observations qui se trouvent dans ce cas, on réitère la procédure d'assignation en intégrant les observations déjà assignées dans celles notées "observations contours". On classe ainsi, de proche en proche, les observations disponibles dans l'échantillon analysé.

V. - EXEMPLE

Pour illustrer la procédure de détection des contours, nous allons appliquer l'opérateur décrit précédemment sur un échantillon bidimensionnel généré artificiellement. Cet échantillon est constitué de deux classes gaussiennes dont les paramètres statistiques sont indiqués dans le tableau IV.a.

Les figures IV. a à f présentent les différentes phases de la classification de l'échantillon bidimensionnel correspondant au tableau IV.a.

Après l'estimation de la fonction de densité de probabilité sous-jacente à la distribution, on applique le filtre non linéaire à pondération binaire décrit au chapitre III. Aussi bien pour estimer la fonction de densité de probabilité que pour filtrer celle-ci, on prend un nombre k de voisins égal à 20. On peut alors voir que la pente de la fonction filtrée (Cf. fig. IV.b.) est suffisamment importante pour permettre de passer l'opérateur de contour et obtenir une réponse significative. Cet opérateur est appliqué en prenant en compte 15 voisins de manière à respecter le plus possible l'information locale qui existe dans la fonction de densité de probabilité estimée filtrée.

La réponse de l'opérateur de détection de contour est ensuite filtrée pour en atténuer les plus importantes irrégularités (Cf. fig. IV.c.). Les phases de seuillage (Cf. fig. IV.d.) et d'étiquetage (Cf. fig. IV.e.) permettent alors de dégager deux contours. Il ne reste plus qu'à assigner les observations restantes au contour le plus proche (Cf. fig. IV.f.).

	classe 1	classe 2
nombre de points	400	500
vecteur moyenne	$\begin{pmatrix}1\\1\end{pmatrix}$	$\begin{pmatrix}1\\9\end{pmatrix}$
matrice de covariance	$\begin{bmatrix}2 & 0\\0 & 2\end{bmatrix}$	$\begin{bmatrix}4 & 0\\0 & 2\end{bmatrix}$

Tableau IV.a. Paramètres statistiques de l'échantillon.

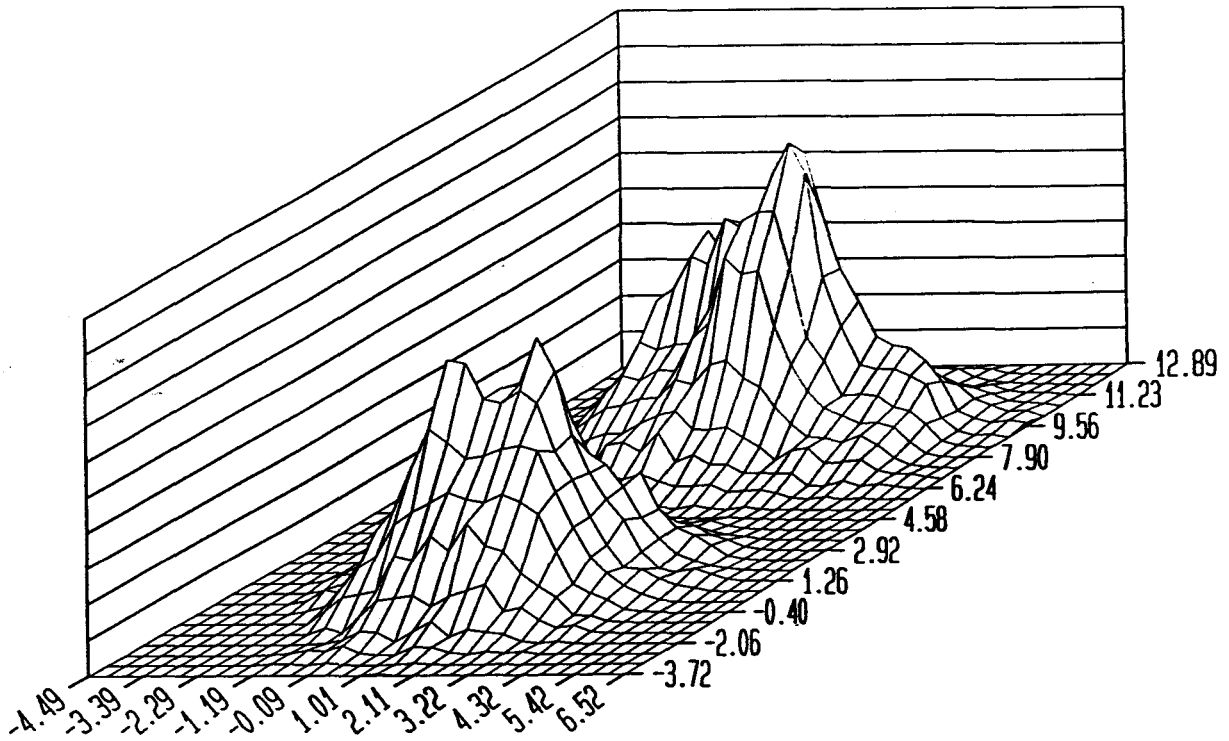


Figure IV.a. Estimation de la fonction de densité de probabilité (k=20)

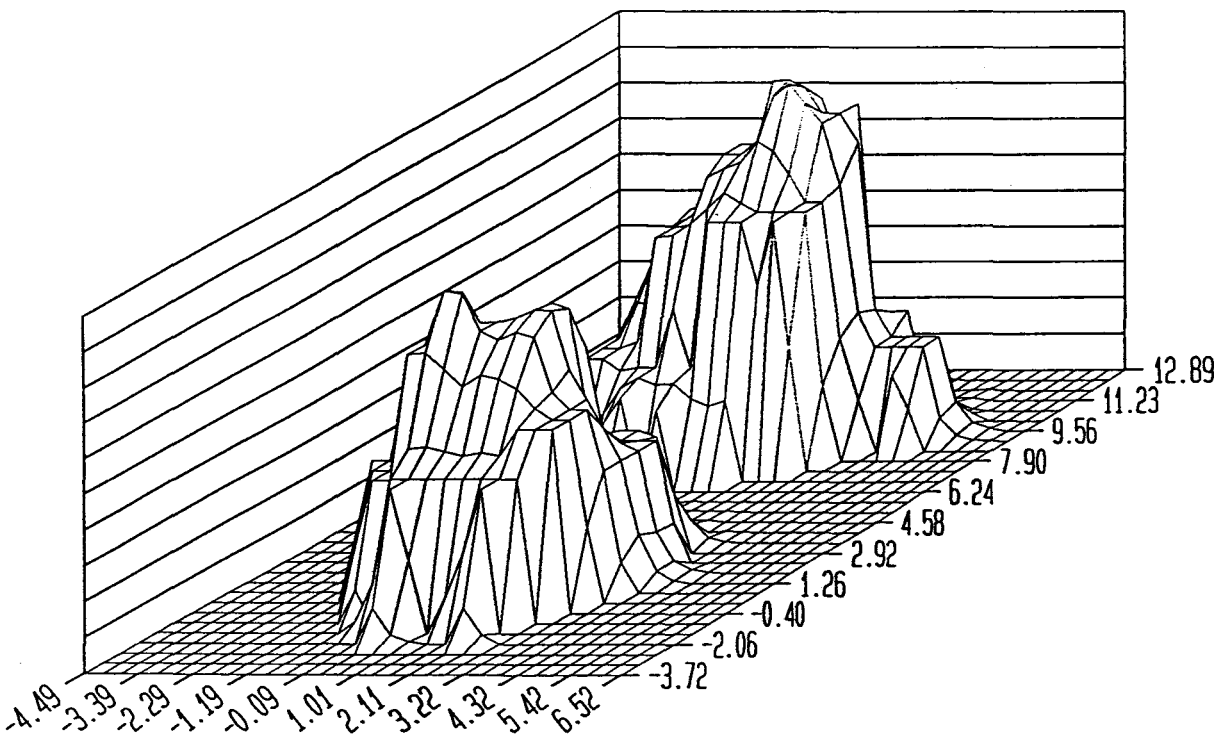


Figure IV.b. Fonction de densité de probabilité filtrée (k=20, 3 itérations).

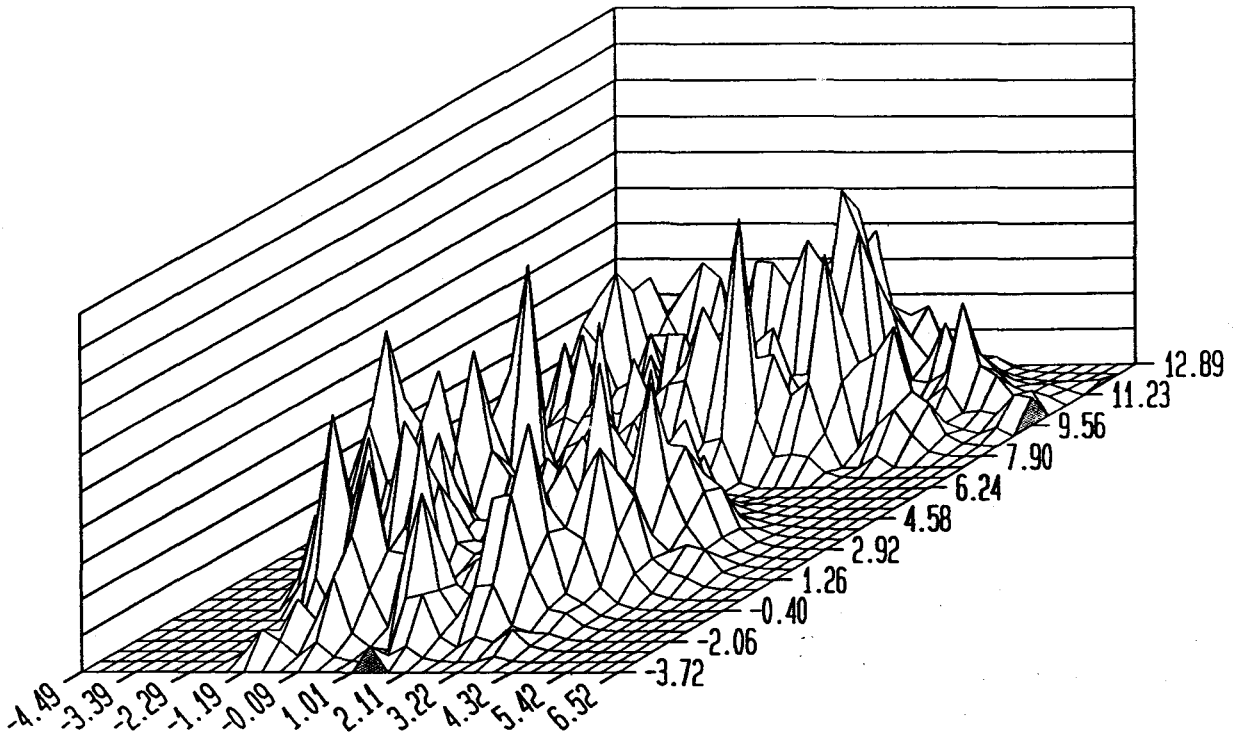


Figure IV.c. Passage de l'opérateur de détection de contours (c=15).

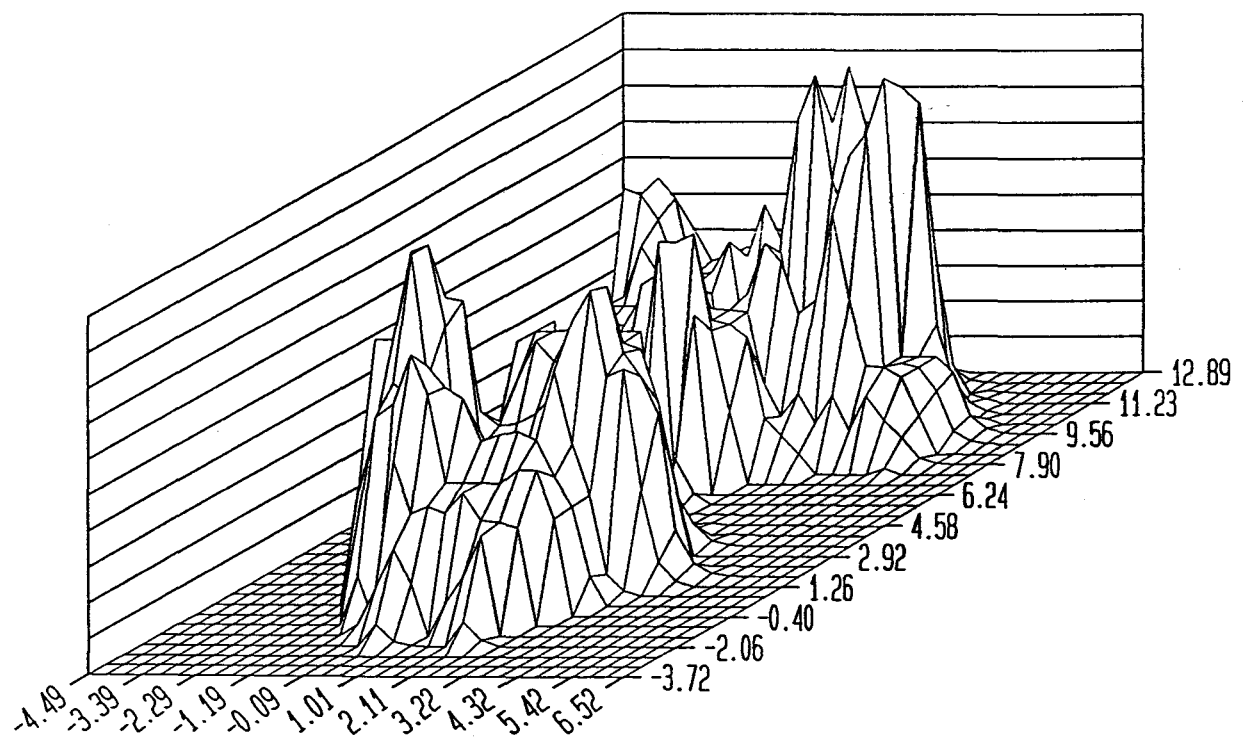


Figure IV.d. Filtrage de la réponse à l'opérateur de détection de contours (c=15, 3 itérations).

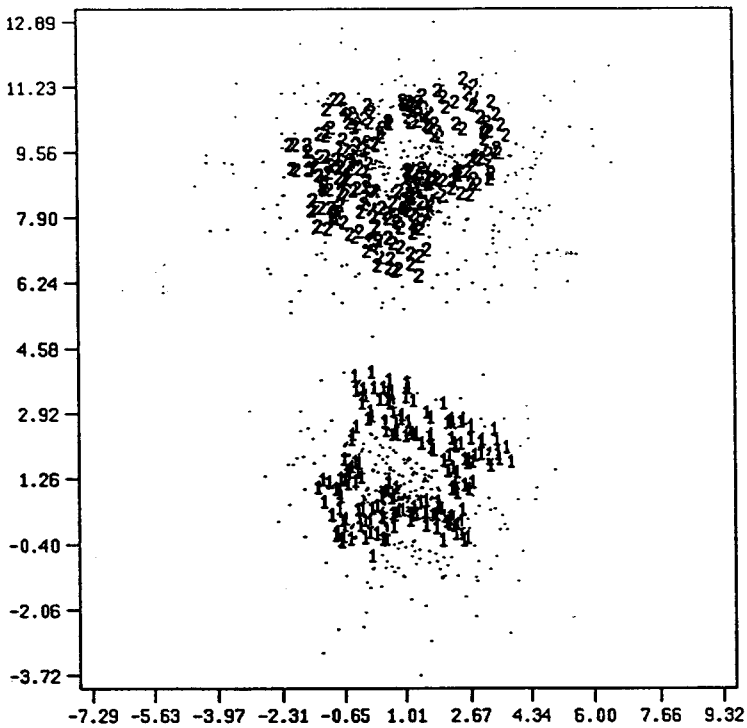


Figure IV.e. Etiquetage des contours.

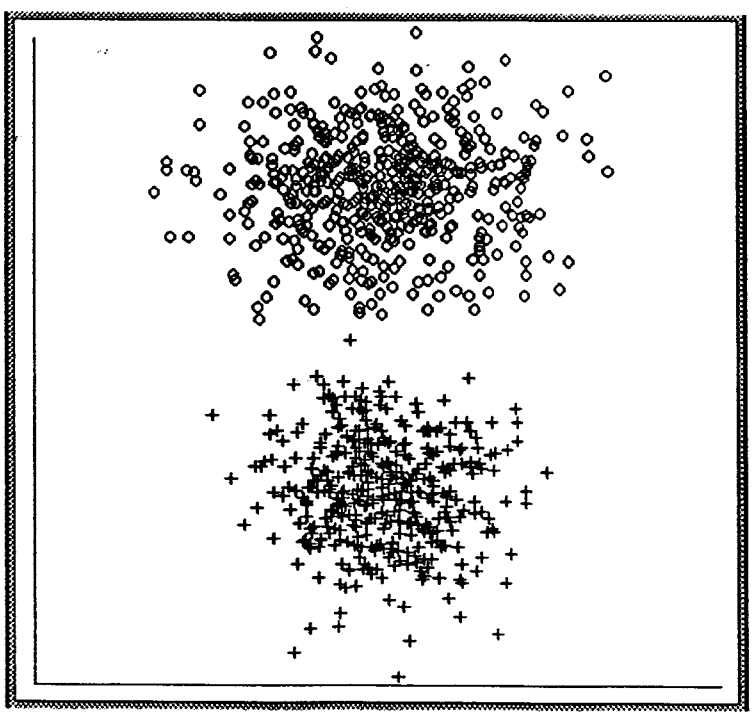


Figure IV.f. Assignation des observations aux contours.

VI - CONCLUSION

Dans ce chapitre une autre approche au problème de détection des modes de la fonction de densité de probabilité estimée a consisté à détecter le contour des modes et non plus leurs noyaux. Cette technique différente est néanmoins plus sensible aux irrégularités de la fonction de densité. En effet, il s'agit par cette méthode de prendre en considération une information bien plus locale que celle que l'on prend lorsqu'on recherche les noyaux des modes (Cf. Chapitre III).

Bien que cette technique soit plus sensible aux irrégularités de la fonction de densité de probabilité, on peut l'utiliser lorsque l'échantillon à analyser comporte des classes de densité plus ou moins régulière mais qui se situent relativement proche l'une de l'autre, trop proche pour que la phase de filtrage de la fonction de densité puisse creuser suffisamment les vallées. En effet, la procédure de détection des noyaux (Cf. chapitre III) exige la présence d'observations à estimateur nul dans les vallées.

CHAPITRE V

RESULTATS EXPERIMENTAUX

CHAPITRE V

RESULTATS EXPERIMENTAUX

I. - METHODE D'EVALUATION

Afin d'évaluer les performances de la méthode de classification décrite dans les chapitres précédents, nous allons l'appliquer sur différents types d'échantillons bidimensionnels et tridimensionnels générés artificiellement. Pour mettre en évidence l'intérêt de la méthode proposée, nous comparons les résultats obtenus avec ceux résultant de l'utilisation d'une méthode classique de classification, à savoir la méthode Isodata [BAL67][FOR74].

I.1. - LA METHODE ISODATA

La méthode Isodata est basée sur une technique de partitionnement de l'échantillon. On choisit tout d'abord t représentants $R_1, R_2, ..., R_t$ de t classes $C_1, C_2, ..., C_t$ définissant ainsi une représentation R et une partition C de l'échantillon telles que :

$$R = \{ R_1, R_2, ..., R_t \}$$

et

$$C = \{ C_1, C_2, ..., C_t \}$$

Une mesure de similarité permet alors d'assigner les observations à classer à l'une des classes C_i, entre chaque représentant R_i et l'observation considérée. La partition C est ensuite utilisée pour affiner une nouvelle représentation R. On affine

ainsi la représentation R en itérant la procédure jusqu'à ce qu'un critère de similitude entre la représentation R et la partition C soit optimisé.

Le représentant R_i de chaque classe C_i est défini par son centre C_i^*, et la mesure de similarité employée est la distance euclidienne qui sépare chaque observation des différents centres C_1^*,..., C_t^*.

L'algorithme général de la méthode Isodata nécessite le réglage de différents paramètres comme le nombre de classes, les coordonnées initiales des centres de ces classes, les différents seuils pour lesquels on décide de regrouper deux classes ou au contraire de scinder une classe en deux. Dans cette étude comparative nous allons simplifier la procédure en fixant le nombre de classes et donner comme centres de celles-ci des valeurs proches des centres réels des classes.

L'algorithme simplifié peut être décrit ainsi :

- classement des observations de l'échantillon en les assignant à la classe associée au centre le plus proche

- Recalcul des centres C_t^* des classes C_t en tenant compte de la nouvelle partition.

- Si l'un au moins des centres a changé, on recommence le classement, sinon on conserve la partition.

II - EXEMPLE 1

Dans ce premier exemple, on considère un échantillon bidimensionnel constitué de trois classes gaussiennes auxquelles on a ajouté du bruit. Le bruit est simulé par des observations ajoutées à l'échantillon, réparties de façon non uniforme sur tout l'espace de représentation des données. Plus précisément, elles sont réparties de manière uniforme si on se déplace suivant un des axes de l'espace de représentation, mais leur densité est croissante, par palier, suivant le deuxième axe. L'algorithme de génération du bruit peut être décrit ainsi :

- On se fixe le nombre d'observations qui constitueront le bruit

- On choisit l'axe, et le sens, suivant lequel on désire une variation du nombre d'observations considérées comme bruit

- Suivant cet axe on divise l'espace en un certain nombre de divisions et on fixe la valeur d'un coefficient α indiquant le rapport du nombre de points de bruits entre deux divisions adjacentes.

- En fonction du coefficient α, on réajuste éventuellement le nombre total de points de bruit, pour que celui-ci ait une valeur entière. Le nombre de points fixé au départ est alors considéré comme un maximum.

- On génère les points de bruit à l'aide un générateur aléatoire uniforme.

Le tableau II.a. reprend les caractéristiques statistiques des trois classes gaussiennes et le tableau II.b. les paramètres utilisés pour générer le bruit.

	classe 1	classe 2	classe 3
nombre de points	200	200	300
vecteur moyenne	$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{pmatrix}7\\4\end{pmatrix}$	$\begin{pmatrix}3\\9\end{pmatrix}$
matrice de covariance	$\begin{bmatrix}2 & 0\\0 & 2\end{bmatrix}$	$\begin{bmatrix}2 & 0\\0 & 2\end{bmatrix}$	$\begin{bmatrix}5 & 0\\0 & 1\end{bmatrix}$

Tableau II.a. Caractéristiques statistiques de l'exemple 1.

nombre de points générés	264
axe de la répartition par palier	$2^{ième}$ attribut
sens de variation	décroissant
nombre de divisions	5
coefficient α	1,5

Tableau II.b. Paramètres du bruit de l'exemple 1.

La distribution obtenue est représentée figure II.a.

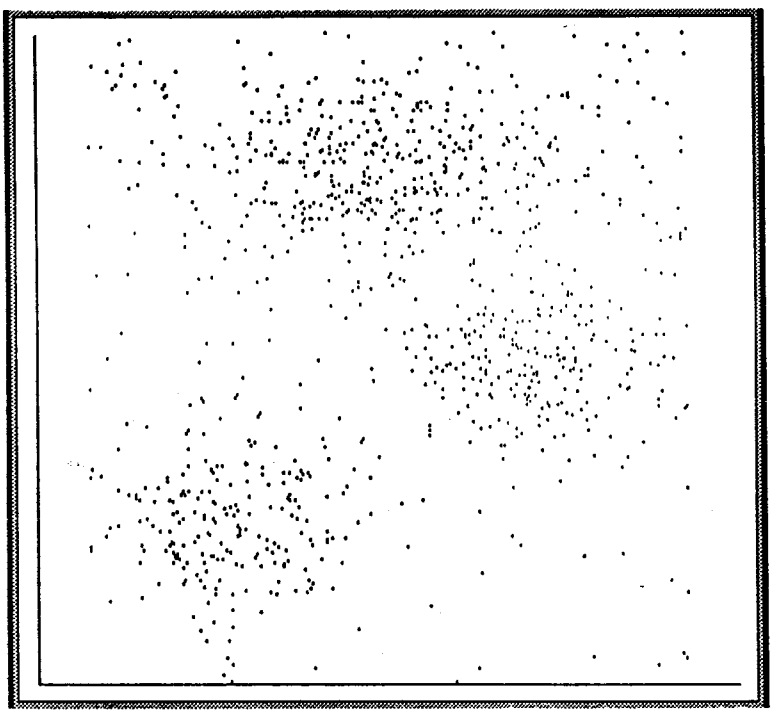


Figure II.a. Représentation de l'échantillon de l'exemple 1.

L'estimation de la fonction de densité de probabilité sous-jacente à la distribution est présentée figure II.b. Le nombre k de voisins a été pris égal à 20. La mise en évidence des noyaux des modes, après 3 itérations du filtre non linéaire à pondération binaire, proposé au chapitre III, de la fonction de densité est représentée figure II.c.

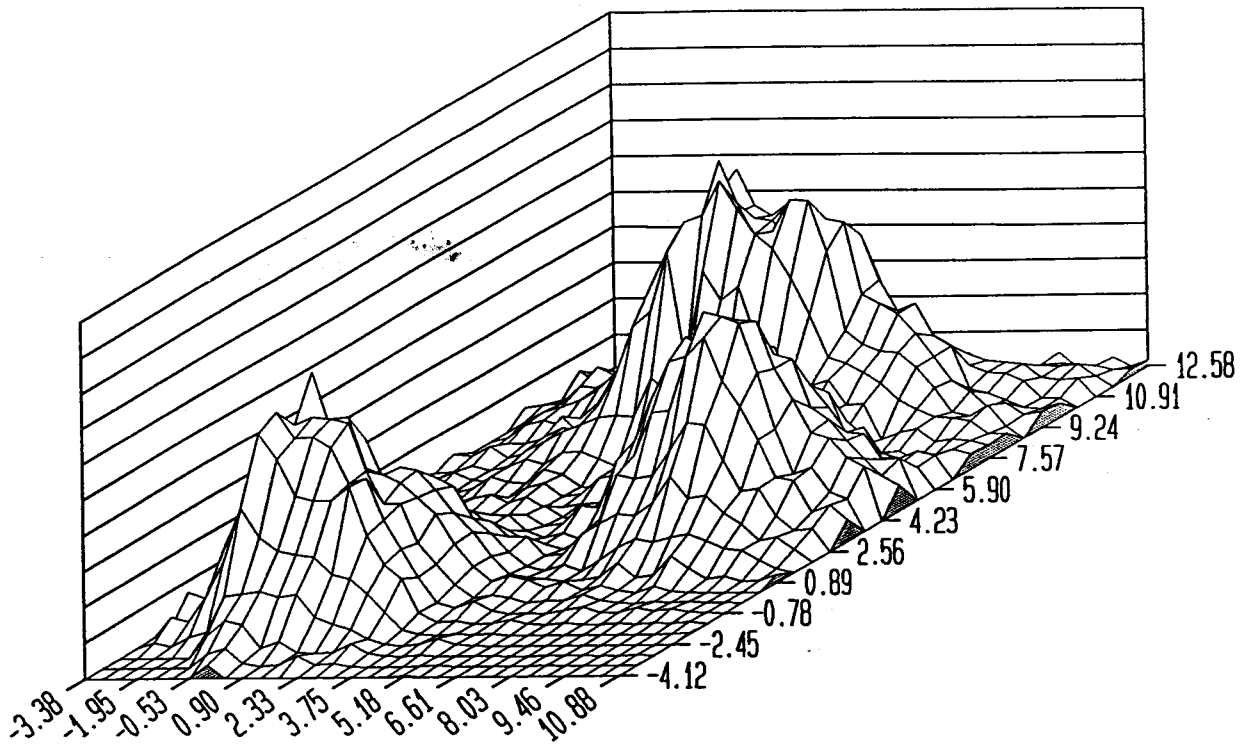


Figure II.b. Estimation de la fonction de densité de probabilité.

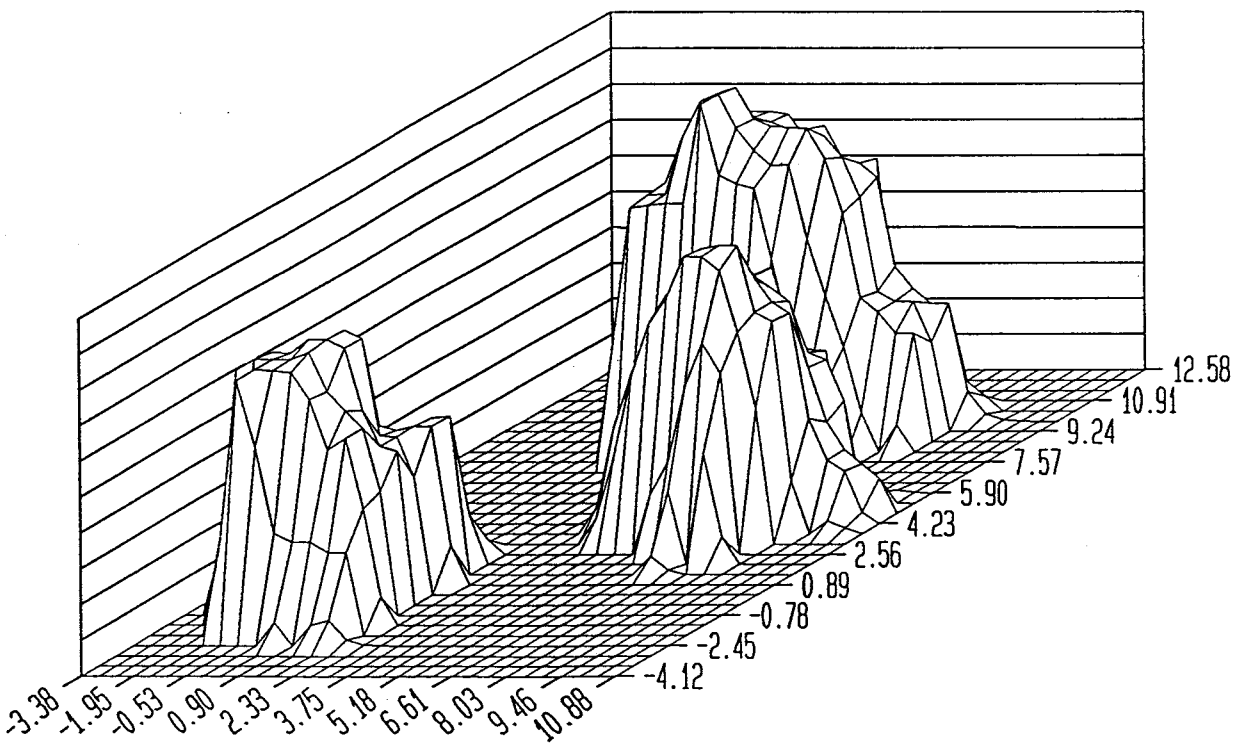


Figure II.c. Fonction de densité de probabilité filtrée (k=20, 3 itérations).

En prenant également k=20, on étiquette les noyaux des modes (Cf. Fig. II.d.). Puis on obtient la classification finale par assignation des observations restantes au noyau le plus proches (Cf. fig. II.e.). Les observations qui étaient au départ générées comme du bruit ne sont bien sûr pas différenciées des autres observations appartenant aux classes gaussiennes. Les observations dites de bruit ont été ajoutées à l'échantillon, mais pour la classification elles ne peuvent en aucun cas être considérées différentes des autres observations. En effet, le processus de classification n'exploite que les deux attributs qui caractérisent une observation.

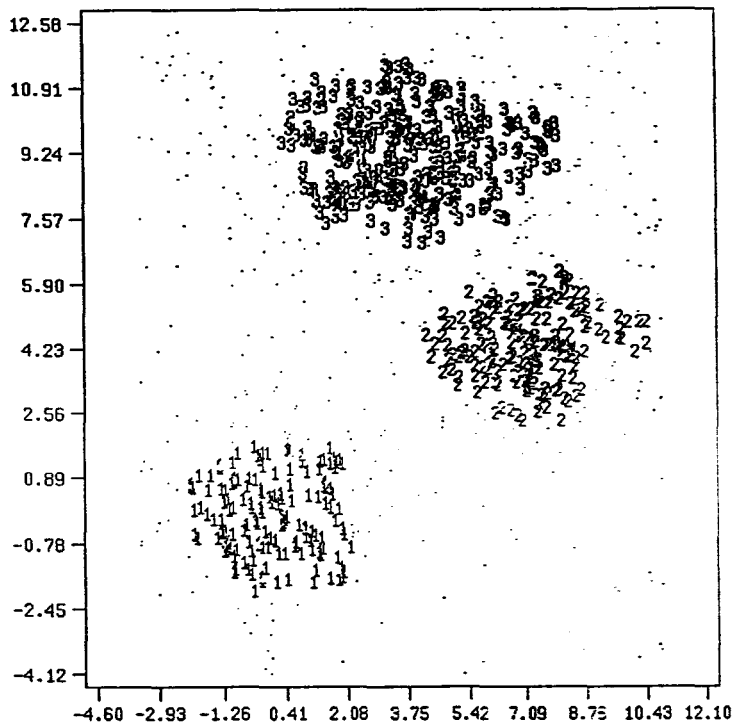


Figure II.d. Noyaux étiquetés des modes.

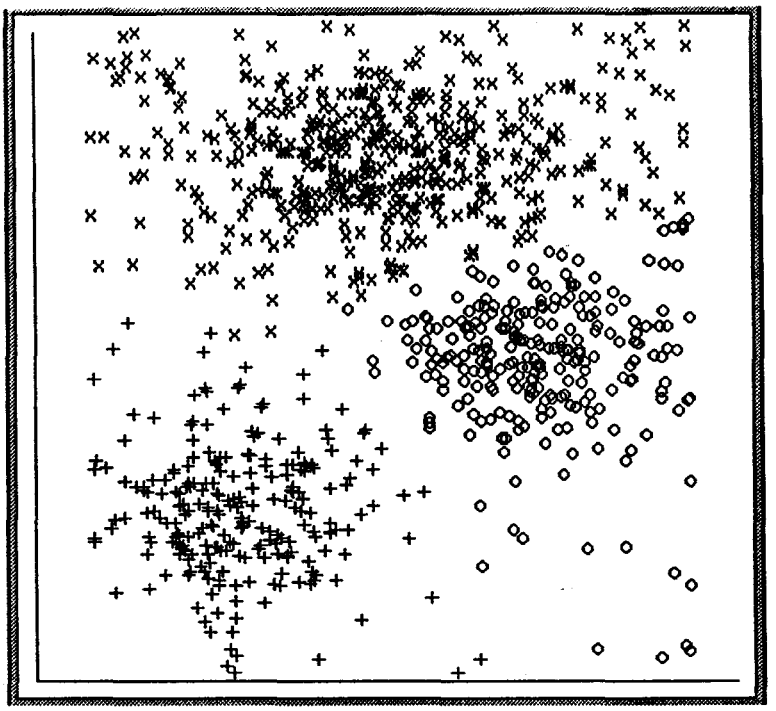


Figure II.e. Observations assignées aux noyaux.

La classification obtenue par l'algorithme Isodata est représentée figure II.f. Les matrices de confusion sont présentées sur le tableau II.c.

Les matrices de confusion donnent le nombre d'observations qui se trouvent assignées à une classe C_i alors qu'elles appartiennent à une classe C_j. Dans cet exemple le bruit est constitué d'observations qui ont été ajoutées à un échantillon qui comportait trois classes gaussiennes, la 4^{ième} ligne de la matrice de confusion correspond aux observations "bruits" assignées aux trois classes détectées. Si on ne prend en compte que les trois premières lignes de la matrice, on remarque aisément que l'erreur commise pour classer les observations des trois classes gaussiennes originelles reste très faible (1%) même en présence de bruit.

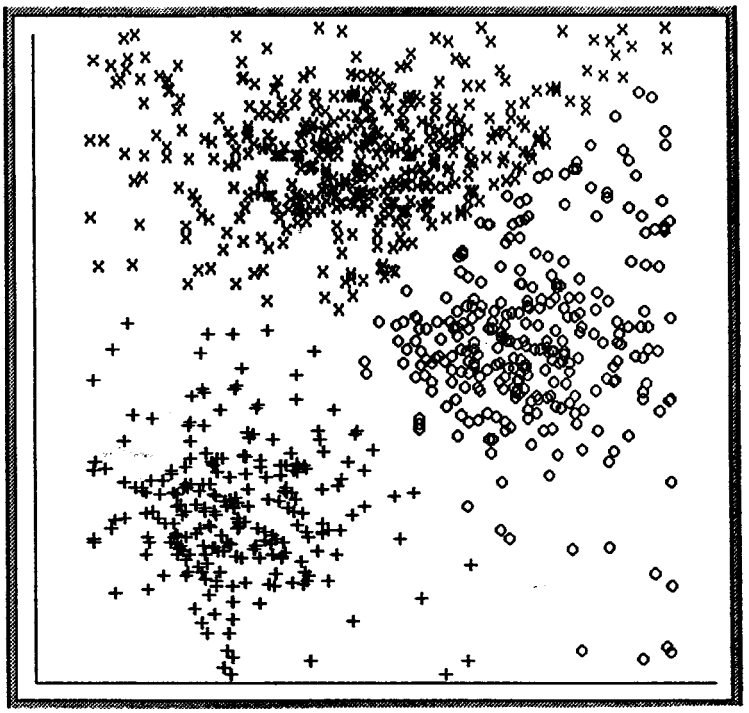


Figure II.f. Classification obtenue par l'algorithme Isodata.

	Algorithme proposé	Algorithme Isodata
Matrice de confusion	$\begin{bmatrix} 200 & 0 & 0 & 0 \\ 1 & 195 & 4 & 0 \\ 0 & 2 & 298 & 0 \\ 38 & 54 & 172 & 0 \end{bmatrix}$	$\begin{bmatrix} 200 & 0 & 0 & 0 \\ 1 & 197 & 2 & 0 \\ 0 & 14 & 286 & 0 \end{bmatrix}$

Tableau II.c. Matrices de confusion.

III - EXEMPLE 2

L'échantillon de l'exemple 2 est constitué de trois classes gaussiennes tridimensionnelles. Les caractéristiques statistiques correspondantes sont présentées dans le tableau III.a. Sur la figure III.a. sont représentées les observations.

	classe 1	classe 2	classe 3
nombre de points	200	200	300
vecteur moyenne	$\begin{pmatrix} 5 \\ 2 \\ 0 \end{pmatrix}$	$\begin{pmatrix} 2 \\ 6 \\ 4 \end{pmatrix}$	$\begin{pmatrix} 3 \\ 1 \\ 10 \end{pmatrix}$
matrice de covariance	$\begin{bmatrix} 1.5 & 0 & 0 \\ 0 & 1.5 & 0 \\ 0 & 0 & 1.5 \end{bmatrix}$	$\begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix}$	$\begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}$

Tableau III.a. Caractéristiques statistiques de l'échantillon de l'exemple 2.

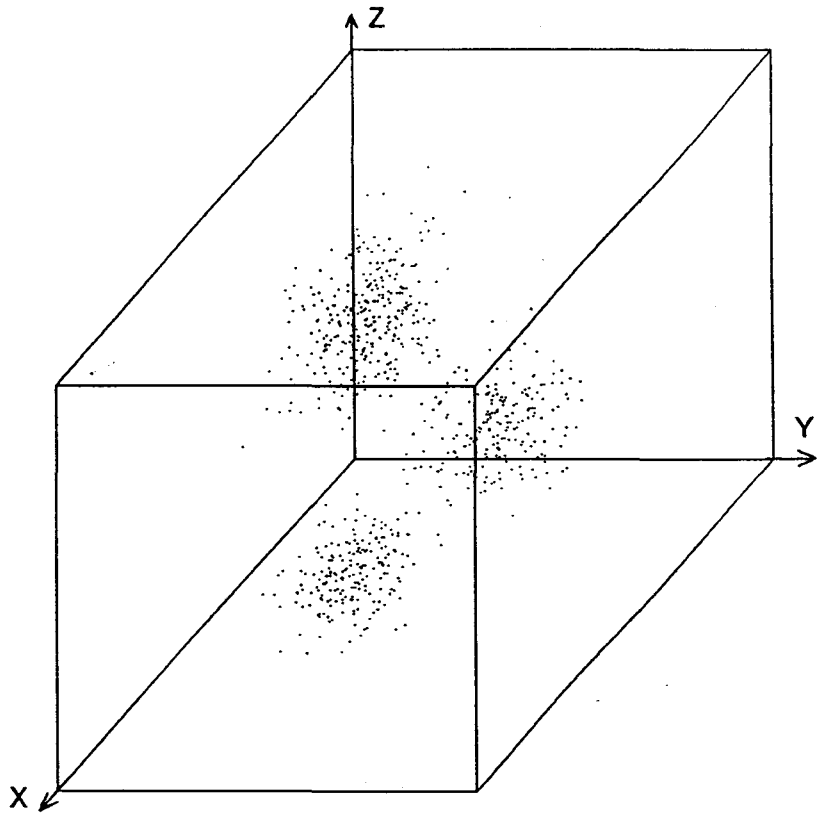
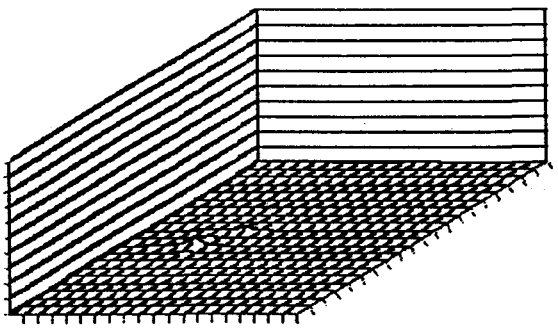
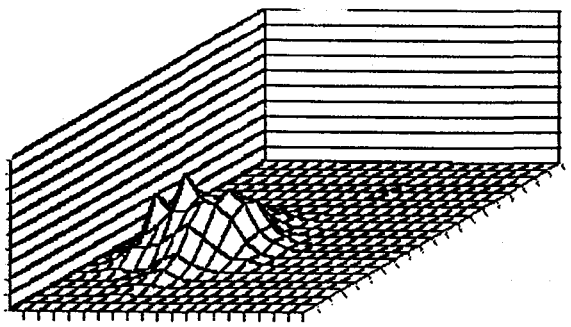


Figure III.a. Représentation des observations de l'échantillon.

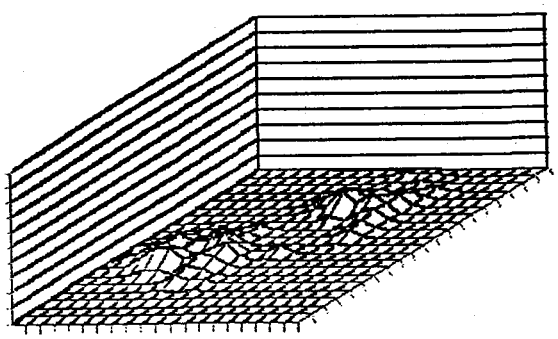
L'estimation de la fonction de densité de probabilité est calculée en prenant un nombre de voisins k égal à 20. La représentation de l'estimateur est donnée sur les figures III.b.1 à 9, sous la forme de plans parallèles et équidistants perpendiculaires à l'axe Z. On ne représente, pour chaque plan, que la valeur de l'estimateur des observations situées dans une plage comprise entre -1 et 1 autour de la valeur en Z du plan considéré. En prenant tous les plans, on retrouve ainsi toutes les observations.



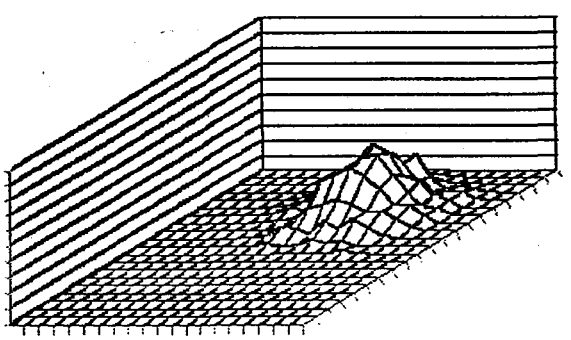
Z = -2



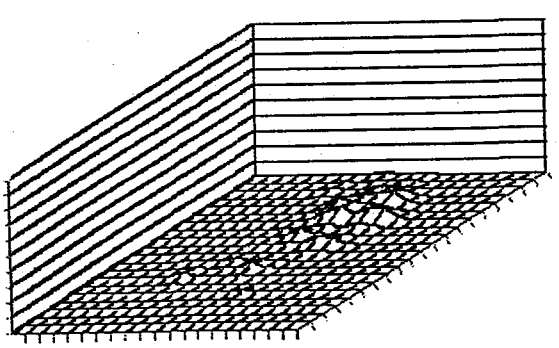
Z = 0



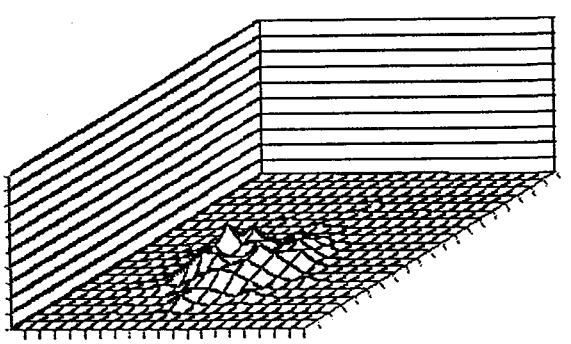
Z = 2



Z = 4

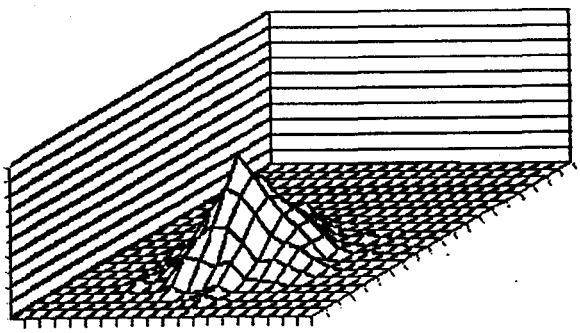


Z = 6

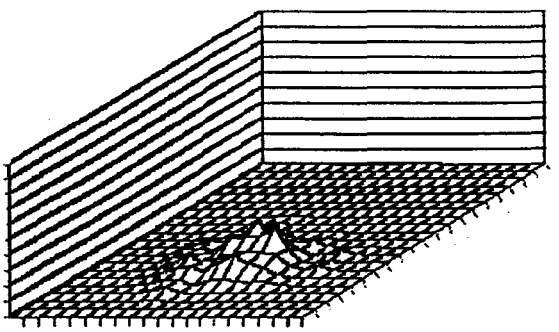


Z = 8

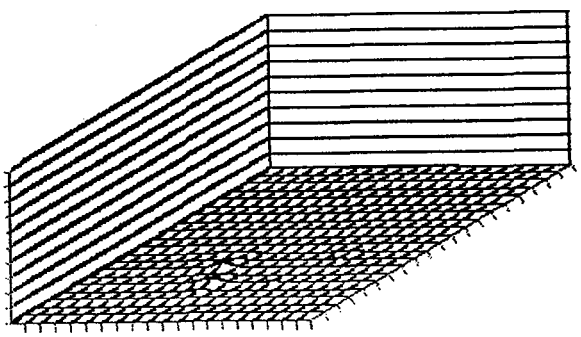
Figures III.b.1 à 6 Estimation de la fonction de densité de probabilité.



Z = 10



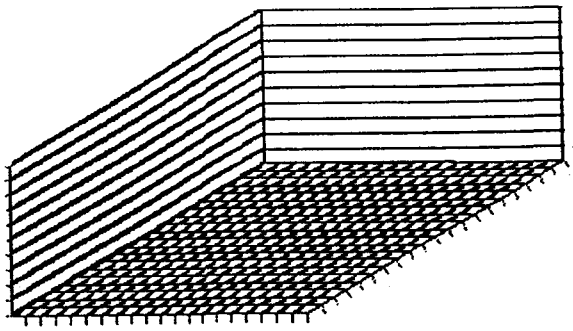
Z = 12



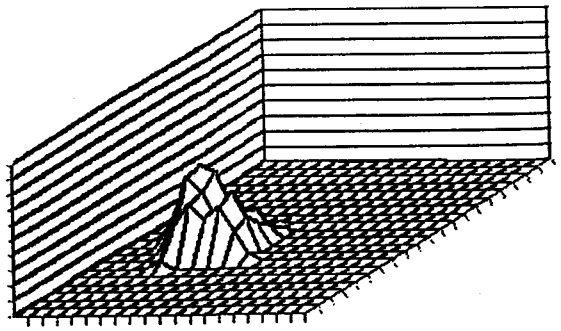
Z = 14

Figures III.b.7 à 9 Estimation de la fonction de densité de probabilité.

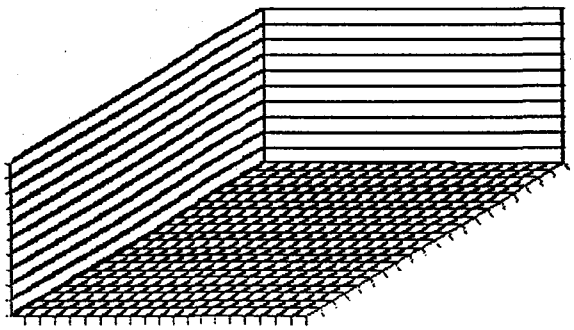
Le filtrage de la fonction de densité de probabilité après 3 itérations et en prenant 20 voisins est représenté sur les figures III.c.1 à 9. On adopte le même type de représentation que pour l'estimateur.



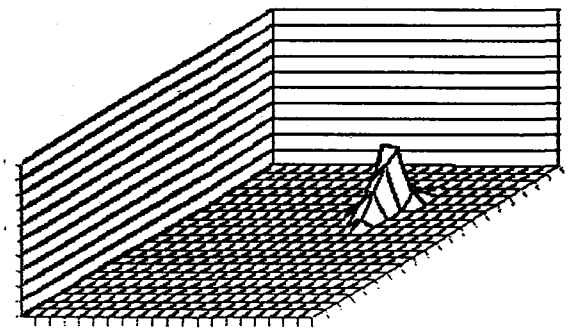
Z = -2



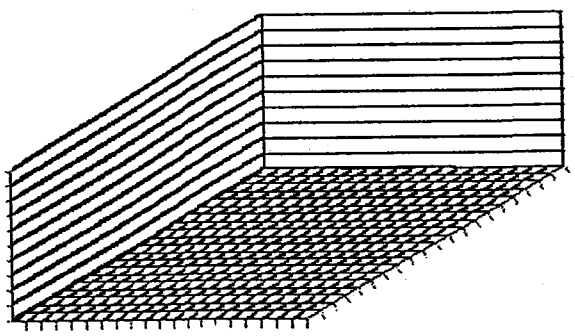
Z = 0



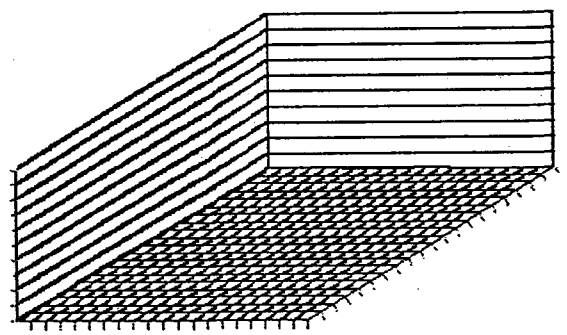
Z = 2



Z = 4

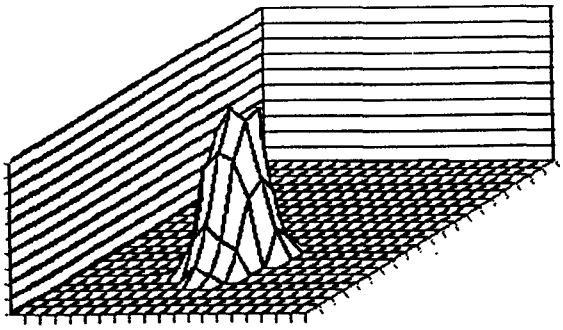


Z = 6

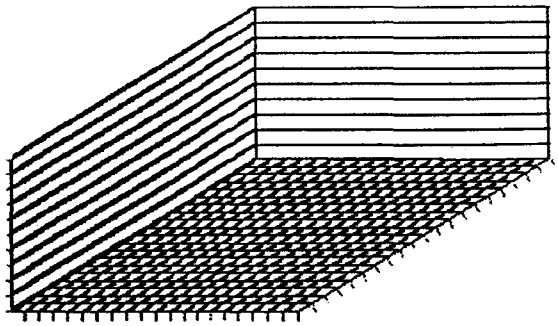


Z = 8

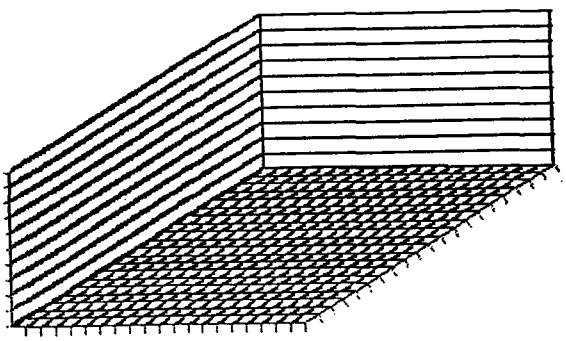
Figures III.c.1 à 6 Filtrage de la fonction de densité de probabilité.



Z = 10



Z = 12



Z = 14

Figure III.c.7 à 9 Filtrage de la fonction de densité de probabilité

Le filtrage permet de mettre en évidence les noyaux des modes (Cf. fig. III.d.) qui sont ensuite étiquetés. En assignant les observations restantes au noyau le plus proche, on obtient la classification finale. Le tableau III.b. donne les caractéristiques des classes obtenues, les matrices de confusion et le taux d'erreur, pour la méthode de classification proposée et pour l'algorithme Isodata. La classification des observations de l'échantillon par l'algorithme proposé donne un taux d'erreur supérieur à l'algorithme Isodata mais reste parfaitement acceptable. En effet, la différence de taux d'erreur obtenu entre l'algorithme proposé et l'algorithme Isodata n'est provoquée que par 4 observations, c'est à dire 0,5% du nombre total d'observations dans l'échantillon.

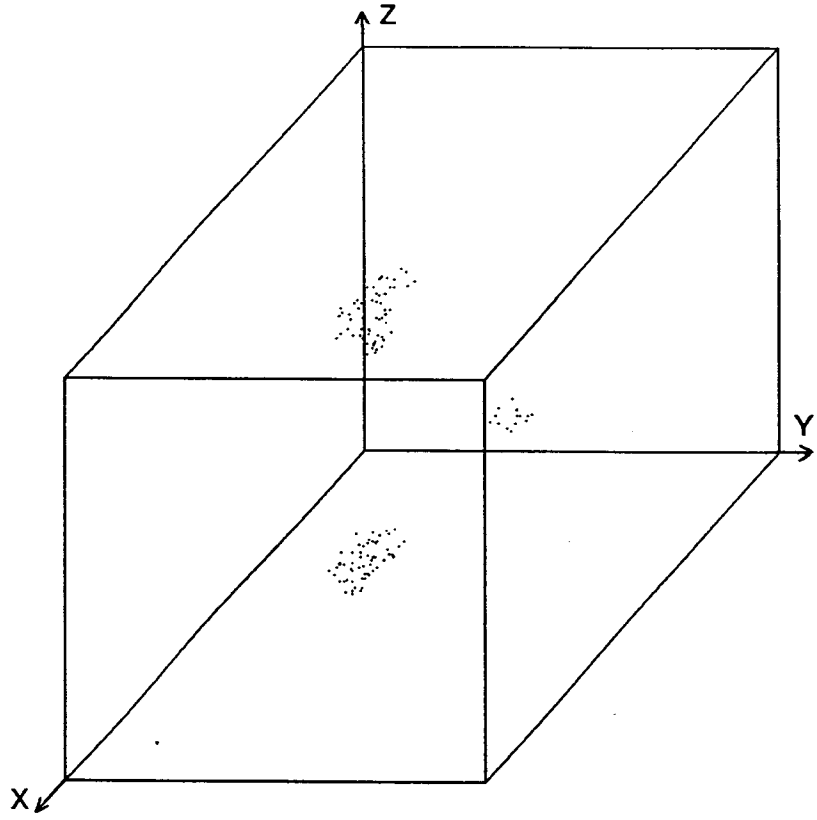


Figure III.d. Noyaux des modes de l'échantillon de l'exemple 2.

	classe 1	classe 2	classe 3
nombre de points	203	196	301
vecteur moyenne	$\begin{pmatrix} 5.038 \\ 1.958 \\ 0.102 \end{pmatrix}$	$\begin{pmatrix} 1.695 \\ 6.096 \\ 3.988 \end{pmatrix}$	$\begin{pmatrix} 2.993 \\ 1.155 \\ 9.870 \end{pmatrix}$
matrice de covariance	$\begin{bmatrix} 1.631 & -0.015 & 0.046 \\ -0.015 & 1.422 & 0.117 \\ 0.046 & 0.117 & 1.46 \end{bmatrix}$	$\begin{bmatrix} 1.702 & 0.277 & 0.065 \\ 0.27 & 2.073 & -0.055 \\ 0.065 & -0.055 & 1.680 \end{bmatrix}$	$\begin{bmatrix} 2.268 & -0.005 & -0.272 \\ -0.005 & 1.980 & -0.145 \\ -0.272 & -0.145 & 2.633 \end{bmatrix}$
matrice de confusion	$\begin{bmatrix} 200 & 0 & 0 \\ 3 & 193 & 4 \\ 0 & 3 & 297 \end{bmatrix}$		
taux d'erreur	1,43 %		

Tableau III.b. Caractéristiques de l'échantillon après classification.

	classe 1	classe 2	classe 3
nombre de points	200	200	300
vecteur moyenne	$\begin{pmatrix} 5.082 \\ 1.910 \\ 0.040 \end{pmatrix}$	$\begin{pmatrix} 1.785 \\ 6.005 \\ 4.021 \end{pmatrix}$	$\begin{pmatrix} 2.980 \\ 1.098 \\ 9.921 \end{pmatrix}$
matrice de covariance	$\begin{bmatrix} 1.565 & 0.042 & 0.156 \\ 0.042 & 1.321 & 00.0034 \\ 0.156 & 0.0034 & 1.46 \end{bmatrix}$	$\begin{bmatrix} 1.820 & 0.131 & 0.146 \\ 0.131 & 2.207 & -0.070 \\ 0.146 & -0.070 & 1.680 \end{bmatrix}$	$\begin{bmatrix} 2.288 & -0.051 & -0.233 \\ -0.051 & 1.801 & 0.022 \\ -0.233 & 0.022 & 2.633 \end{bmatrix}$
matrice de confusion	$\begin{bmatrix} 198 & 0 & 0 \\ 2 & 200 & 4 \\ 0 & 0 & 296 \end{bmatrix}$		
taux d'erreur	0,86 %		

Tableau III.b. Caractéristiques de l'échantillon après classification par l'algorithme Isodata.

IV - EXEMPLE 3

L'échantillon de cet exemple est constitué de deux classes tridimensionnelles en forme de tore. L'ensemble des données est présenté figure IV.a.

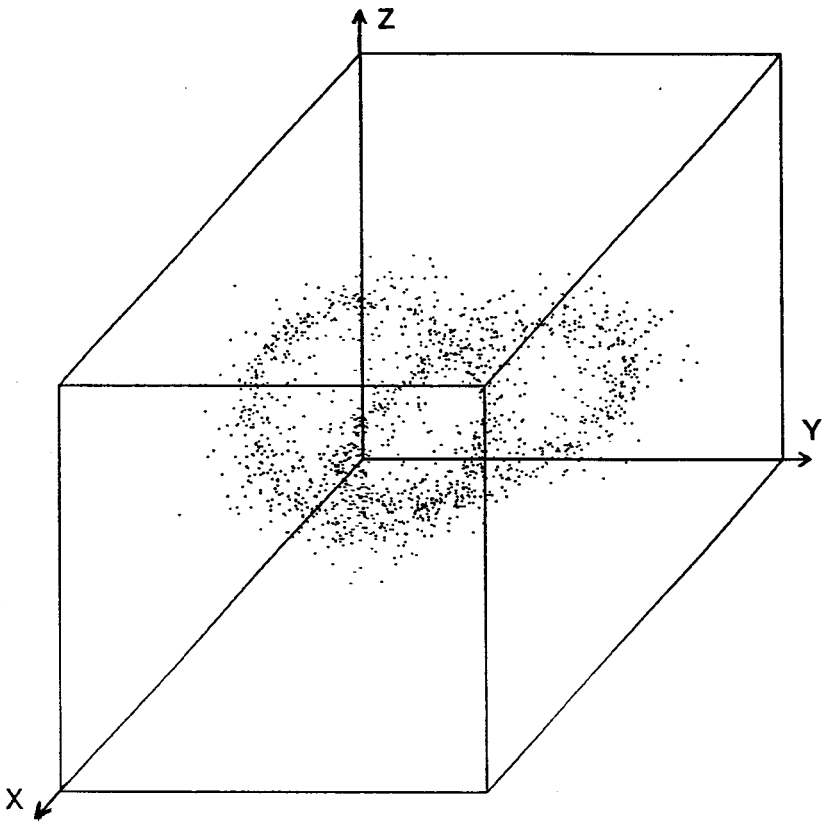
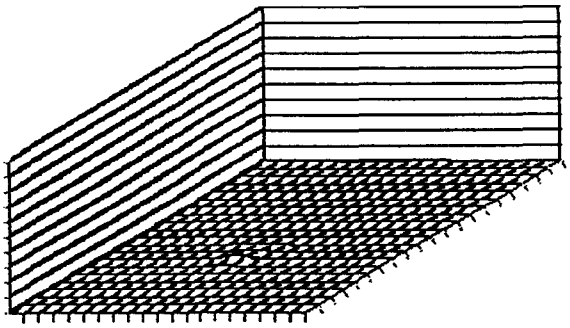
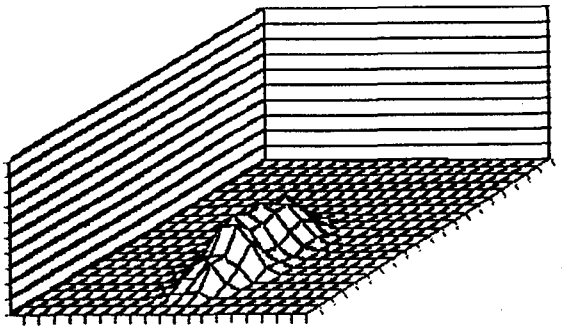


Figure IV.a. Représentation des observations de l'échantillon.

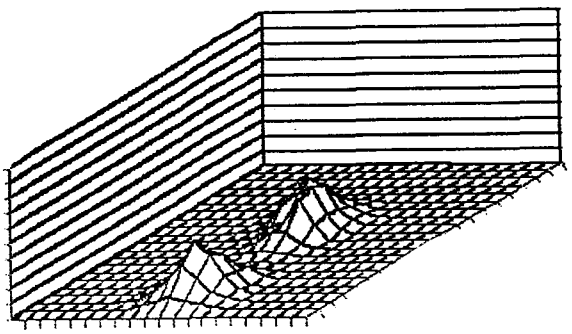
L'estimation de la fonction de densité de probabilité est calculée en prenant un nombre k de voisins égal à 20. La représentation de l'estimateur est donné sur les figures IV.b.1. à 7 sous la forme d'une série de plans parallèles et équidistants perpendiculaires à l'axe Z, comme dans l'exemple précédent.



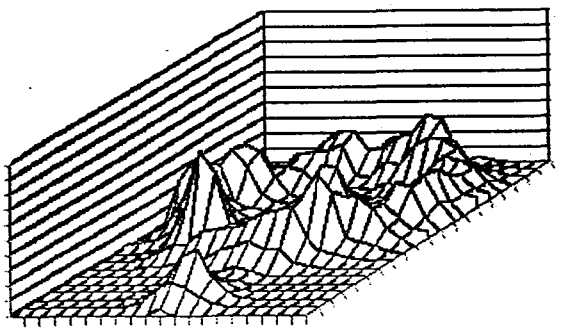
Z = -6



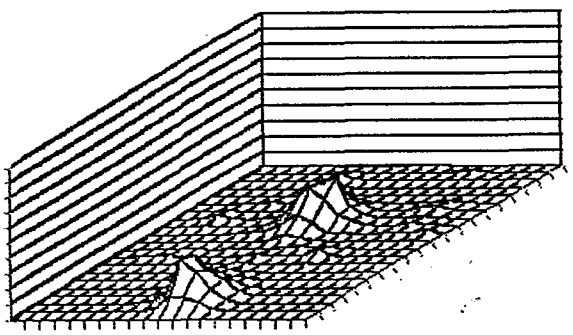
Z = -4



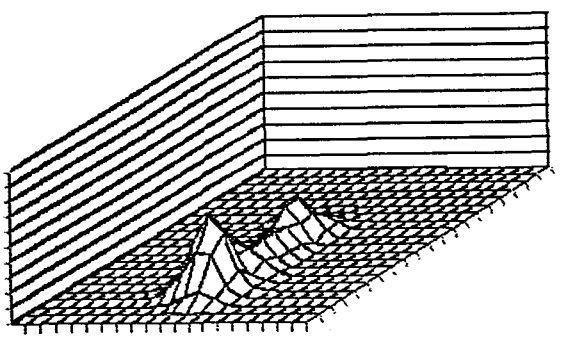
Z = -2



Z = 0

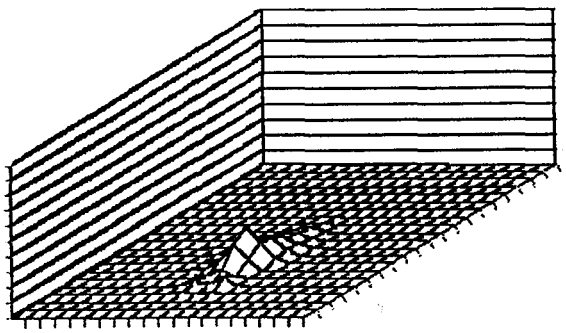


Z = 2



Z = 4

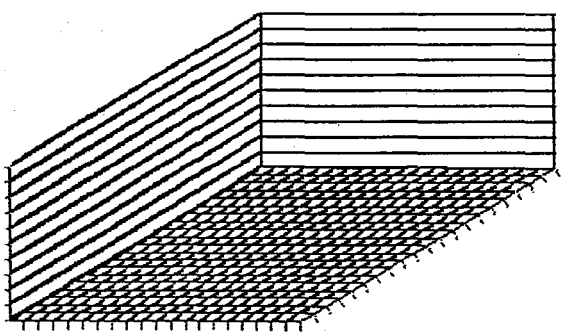
Figures IV.b.1 à 6. Estimation de la fonction de densité de probabilité.



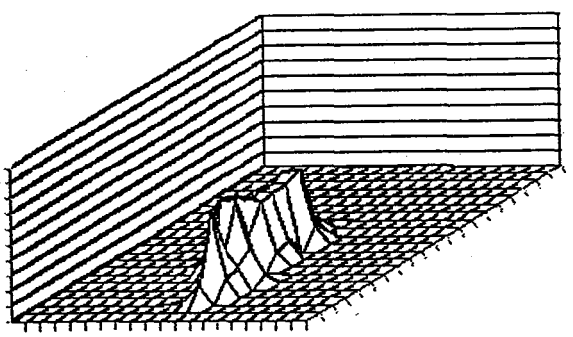
Z = 6

Figure IV.b.7. Estimation de la fonction de densité de probabilité.

La fonction de densité de probabilité estimée est ensuite filtrée par l'intermédiaire du filtre non linéaire proposé au chapitre III en prenant 20 voisins. Après 3 itérations le résultat de ce filtrage est présenté figure IV.c.1. à 7.

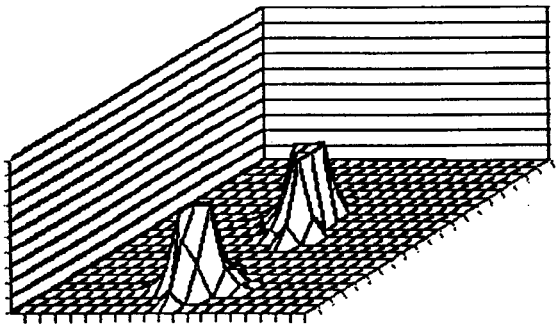


Z = -6

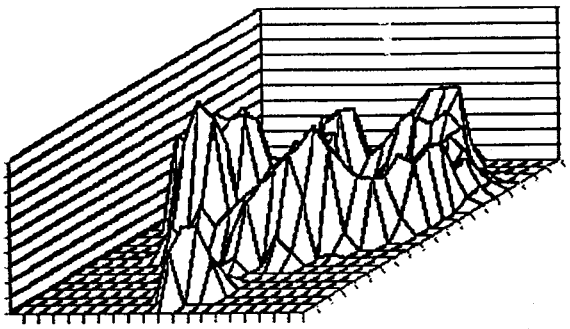


Z = -4

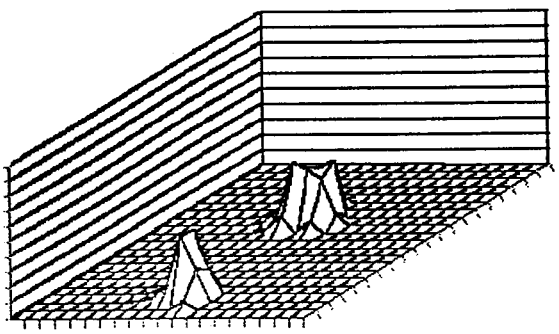
Figures IV.c.1 à 2 Fonction de densité de probabilité filtrée (k=20, 3 itérations.)



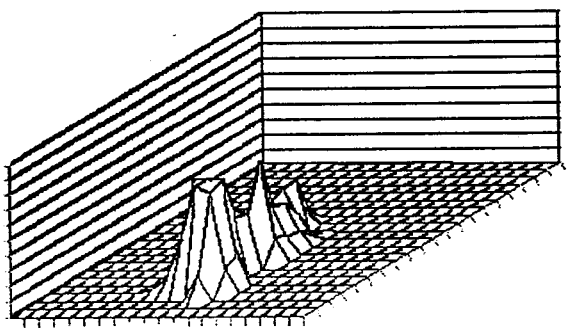
Z = -2



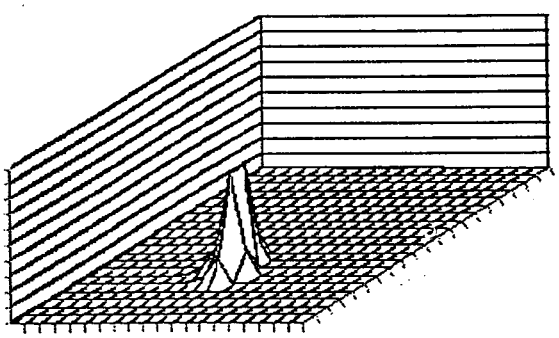
Z = 0



Z = 2



Z = 4



Z = 6

Figures IV.c.3 à 7 Fonction de densité de probabilité filtrée (k=20, 3 itérations).

On représente sur la figure IV.d. les noyaux des modes de la fonction de densité sous-jacente à la distribution des observations. Ces noyaux sont ensuite étiquetés, et les observations sont assignées au noyau le plus proche. Le tableau

IV.b. donne les matrices de confusion et le taux d'erreur de la classification, par la méthode de classification proposée et par l'algorithme Isodata.

On remarque immédiatement que sur un tel échantillon la méthode Isodata commet une erreur très importante. En effet, la méthode Isodata utilise un critère de distance par rapport au centre de la classe, ce qui implique de faire l'hypothèse que les observations sont plus nombreuses au centre des classes que sur les bords. Par contre l'échantillon analysé dans cet exemple comporte des classes dont les observations ne sont pas concentrées autour du barycentre de la classe, d'où le taux d'erreur important obtenu par cet algorithme. Quant à la technique de classification proposé dans cette étude, elle ne fait aucune hypothèse sur la géométrie des classes et explique ainsi le très faible taux d'erreur de classification obtenu. En effet, la procédure d'étiquetage des noyaux et celle d'assignation des observations aux noyaux le plus proche ne fait appel qu'à une notion de voisinage autour de chaque observation, et permet ainsi le suivi de la géométrie de chaque classe.

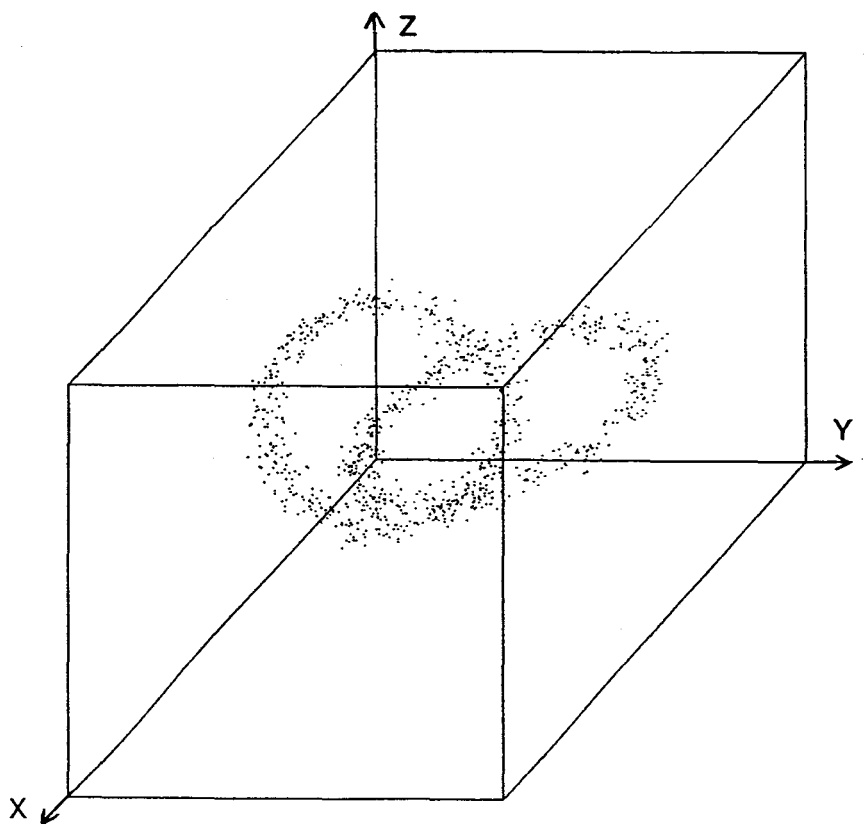


Figure IV.d. Noyaux des modes de la fonction de densité de probabilité.

	méthode proposée	algorithme Isodata
matrice de confusion	$\begin{bmatrix} 747 & 3 \\ 3 & 747 \end{bmatrix}$	$\begin{bmatrix} 507 & 243 \\ 273 & 477 \end{bmatrix}$
taux d'erreur	0.40%	34.40%

Tableau IV.b. Matrice de confusion et taux d'erreur.

V - LIMITE DE LA METHODE

Pour déterminer les limites de la méthode de classification proposée, nous allons rapprocher les classes les unes des autres. Reprenons l'exemple 1 composé de 3 classes gaussiennes bidimensionnelles, mais en modifiant le vecteur moyenne de la classe "2" afin de la rapprocher de la classe "1". et sans y adjoindre de bruit. Ceci permet ainsi de ne pas perturber le processus de classification et de ne s'attacher qu'aux seules observations constituant une classe. Les caractéristiques statistiques correspondant à cet échantillon sont reportées dans le tableau V.a.

	classe 1	classe 2	classe 3
nombre de points	200	200	300
vecteur moyenne	$\begin{pmatrix} 0 \\ 0 \end{pmatrix}$	$\begin{pmatrix} 7 \\ 5 \end{pmatrix}$	$\begin{pmatrix} 3 \\ 9 \end{pmatrix}$
matrice de covariance	$\begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$	$\begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$	$\begin{bmatrix} 5 & 0 \\ 0 & 1 \end{bmatrix}$

Tableau V.a. Caractéristiques statistiques de l'échantillon.

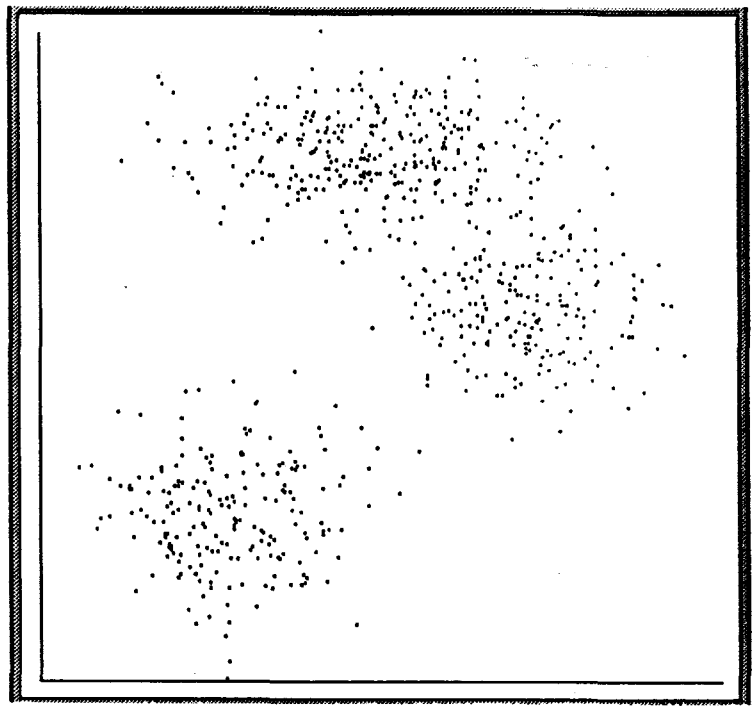


Figure V.a. Représentation des données de l'échantillon.

Les observations qui constituent l'échantillon sont représentées sur la figure V.a. En prenant le nombre k de voisins égal à 20, on estime la fonction de densité de probabilité sous-jacente à cette distribution (Cf. fig. V.b.). Le filtrage de la fonction de densité de probabilité est effectué pour k=20 et en itérant 3 fois la procédure (Cf. fig. V.c.).

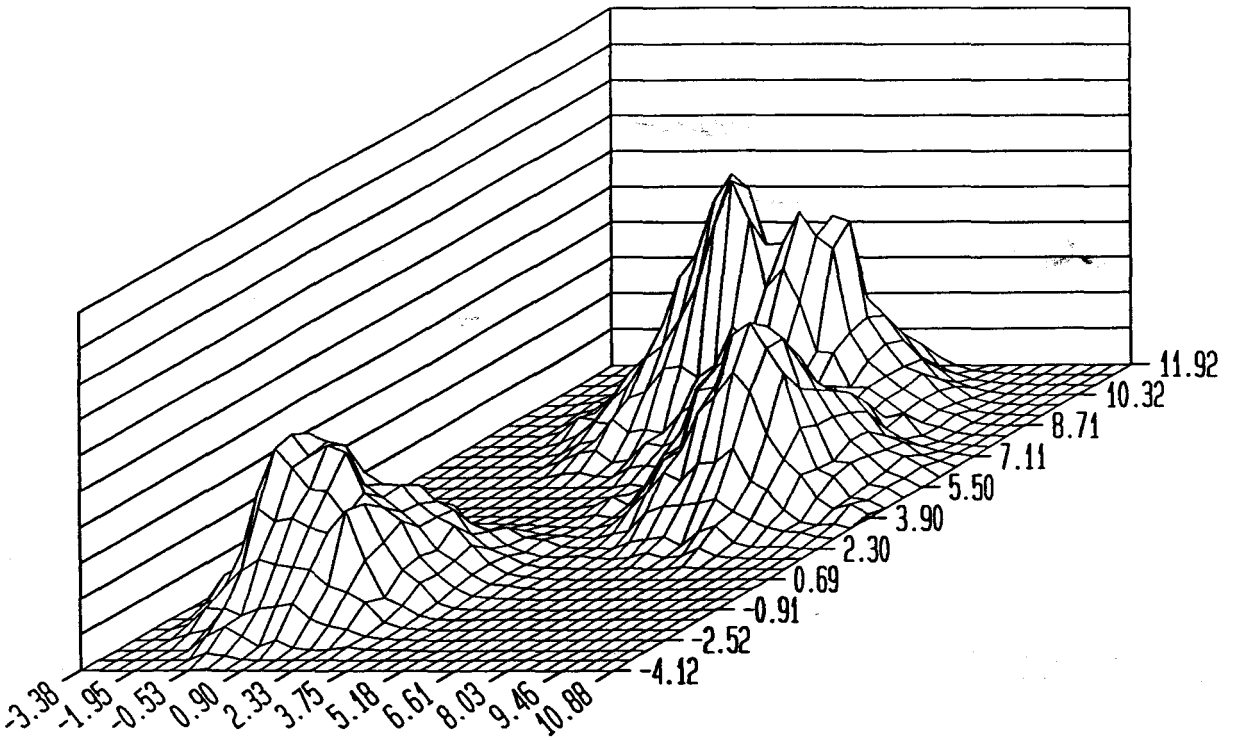


Figure V.b. Estimation de la fonction de densité de probabilité (k=20).

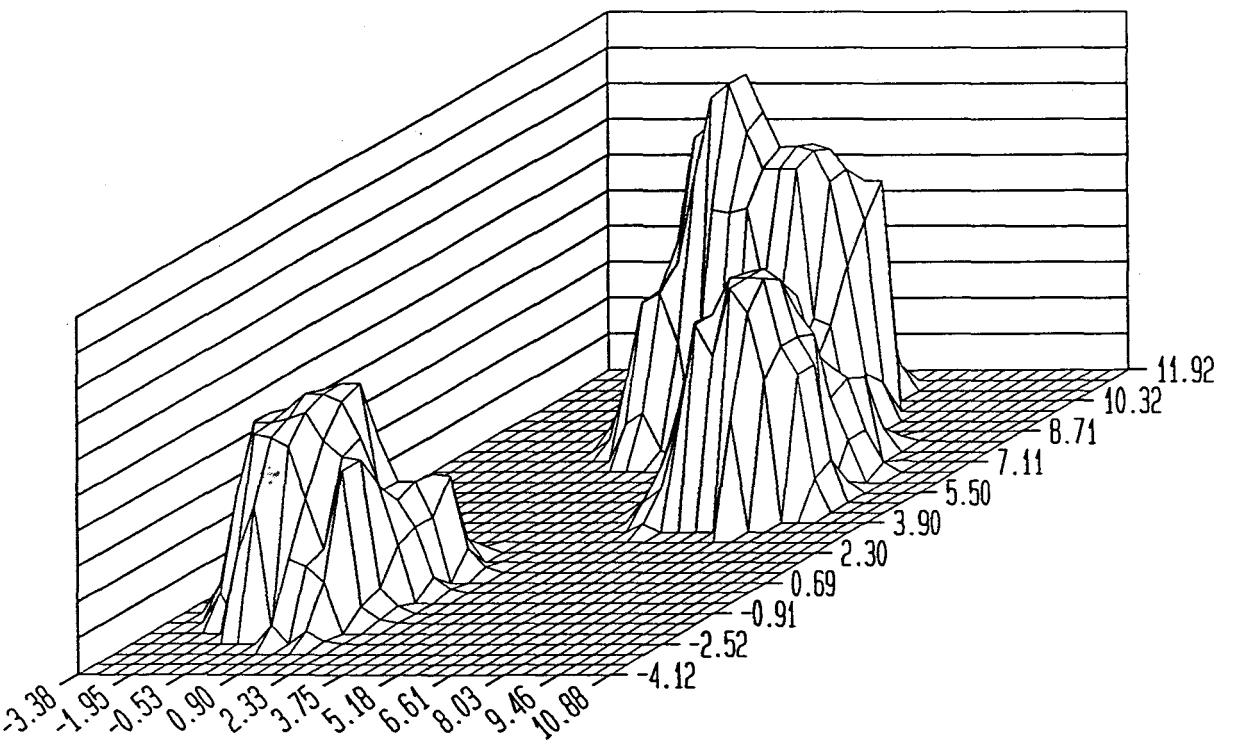


Figure V.c. Fonction de densité de probabilité filtrée (k=20, 3 itérations).

Après une phase de binarisation de la fonction de densité de probabilité filtrée en fixant le seuil à 5% (Cf. fig. V.d.), on applique l'algorithme de détection et d'étiquetage des noyaux des modes. Sur la figure V.e. on remarque que 2 classes, celles correspondant aux vecteurs "moyenne" $(7,5)^T$ et $(3,9)^T$, sont étiquetées identiquement. En fait quelques observations situées entre ces classes ont formées une "passerelle" entre les 2 classes par effet de chaînage lors de la phase d'étiquetage. En effet, l'algorithme de détection et d'étiquetage des noyaux des modes est basé sur la recherche des observations à estimateurs non nuls situées parmi les k plus proches voisins du point d'étude, comme c'est le cas des quelques observations situées entre les 2 classes..

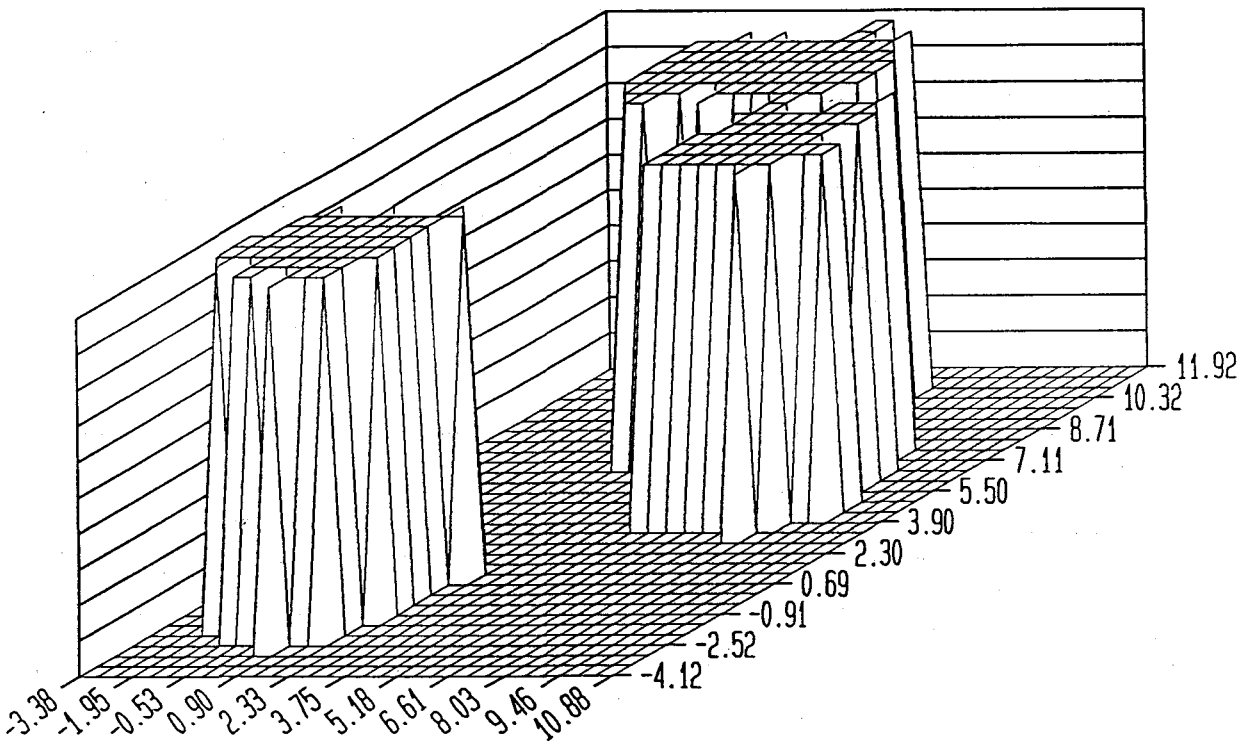


Figure V.d. Binarisation de la fonction de densité de probabilité filtrée (seuil à 5%).

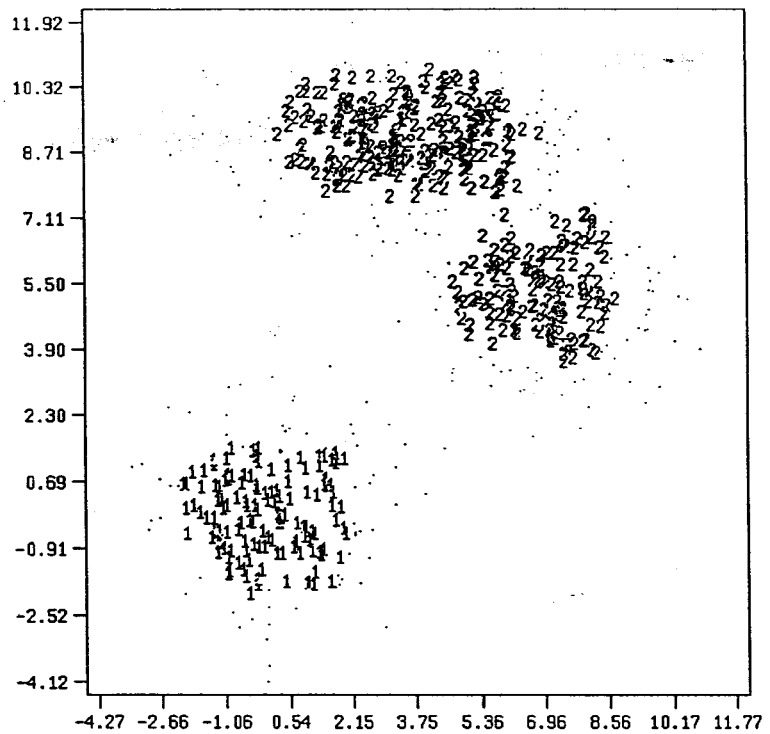


Figure V.e. Noyaux étiquetés des modes.

Sur la figure V.g., les noyaux ont été correctement séparés mais en fixant le seuil de binarisation de la fonction de densité de probabilité à 25% (Cf. fig. V.f.). On voit ici la limite de la méthode de classification proposée. Lorsque les modes ne sont pas bien séparés, la plage de réglage du seuil est relativement plus restreinte que lorsqu'ils sont nettement séparés et où le réglage est alors moins crucial. Toutefois, le dernier exemple constitue une limite au dela de laquelle, si les classes numérotées 2 et 3 etaient légérement plus rapprochées, il serait difficile, même visuellement, de les distinguer.

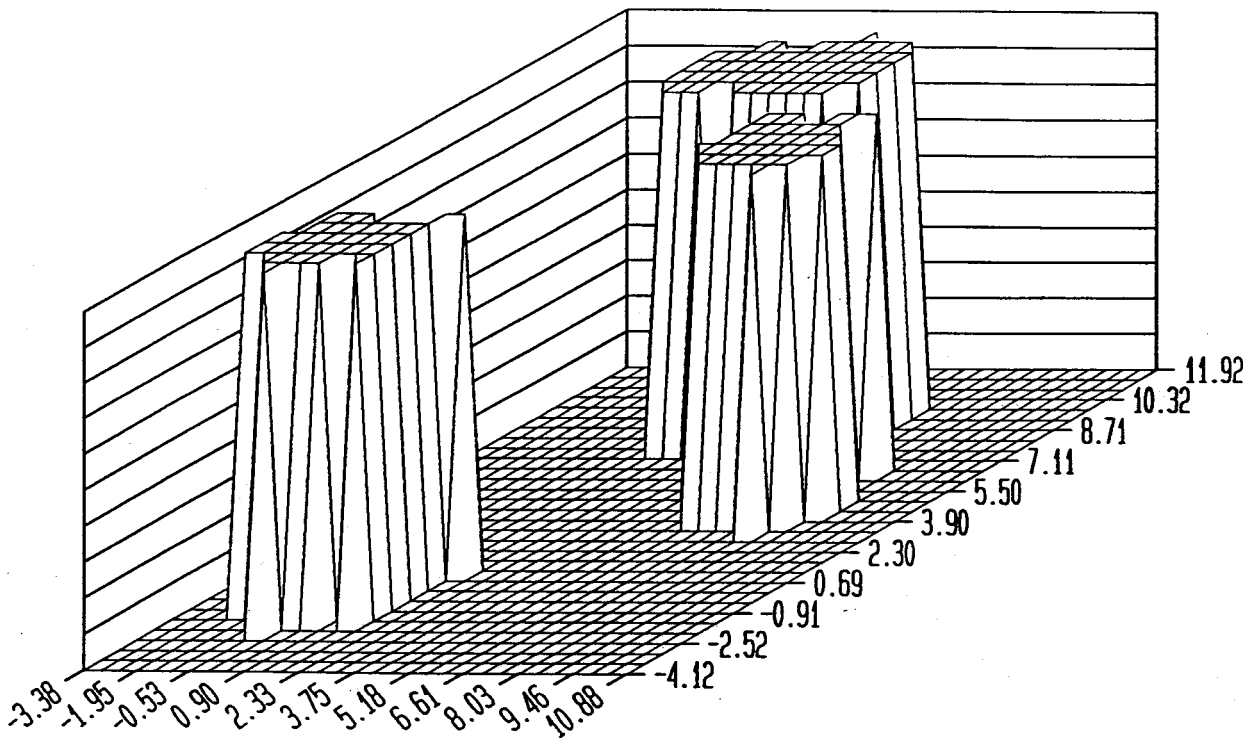


Figure V.f. Binarisation de la fonction de densité de probabilité filtrée (seuil à 25%).

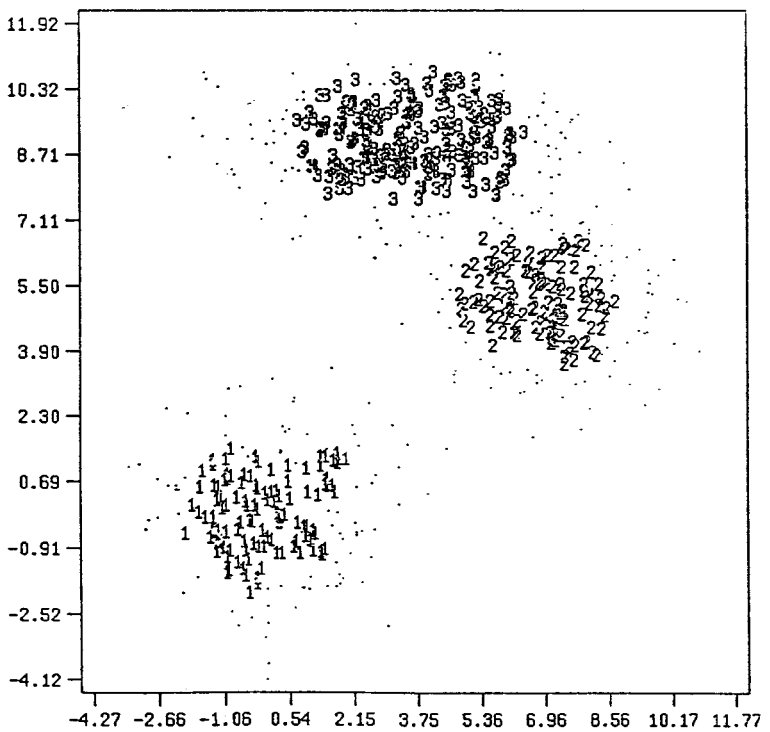


Figure V.g. Noyaux étiquetés des modes.

Sur les figures V.h. et V.i. sont représentés respectivement les classifications obtenues par l'algorithme proposé et par l'algorithme Isodata. Enfin on présente dans les tableaux V.b. et V.c. les caractéristiques statistiques, les matrices de confusion et les taux d'erreurs obtenus après classification.

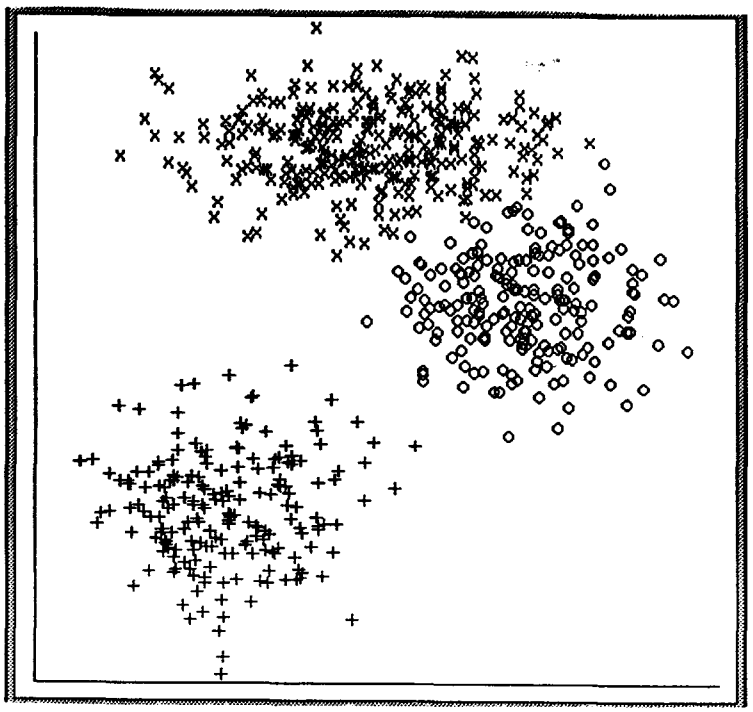


Figure V.h. Classification obtenue par l'algorithme proposé.

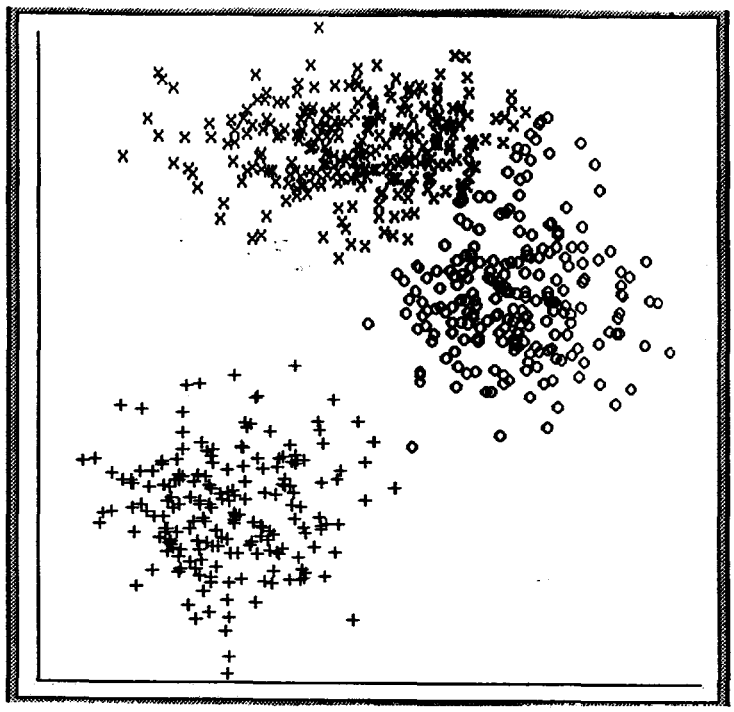


Figure V.j. Classification obtenue par l'algorithme Isodata.

	classe 1	classe 2	classe 3
nombre de points	201	199	300
vecteur moyenne	$\begin{pmatrix} 0.097 \\ -0.070 \end{pmatrix}$	$\begin{pmatrix} 7.055 \\ 5.049 \end{pmatrix}$	$\begin{pmatrix} 3.466 \\ 8.960 \end{pmatrix}$
matrice de covariance	$\begin{bmatrix} 2.214 & 0.090 \\ 0.090 & 1.818 \end{bmatrix}$	$\begin{bmatrix} 2.190 & 0.031 \\ 0.031 & 1.424 \end{bmatrix}$	$\begin{bmatrix} 4.365 & 0.064 \\ 0.064 & 0.926 \end{bmatrix}$
matrice de confusion	$\begin{bmatrix} 200 & 0 & 0 \\ 1 & 193 & 6 \\ 0 & 6 & 294 \end{bmatrix}$		
taux d'erreur	1.86 %		

Tableau V.b. Caractéristiques de l'échantillon après classification.

	classe 1	classe 2	classe 3
nombre de points	200	219	301
vecteur moyenne	$\begin{pmatrix} 0.075 \\ -0.079 \end{pmatrix}$	$\begin{pmatrix} 7.059 \\ 5.343 \end{pmatrix}$	$\begin{pmatrix} 3.211 \\ 8.981 \end{pmatrix}$
matrice de covariance	$\begin{bmatrix} 2.127 & 0.056 \\ 0.056 & 1.814 \end{bmatrix}$	$\begin{bmatrix} 2.041 & 0.144 \\ 0.144 & 2.413 \end{bmatrix}$	$\begin{bmatrix} 3.619 & 0.143 \\ 0.143 & 0.956 \end{bmatrix}$
matrice de confusion	$\begin{bmatrix} 200 & 0 & 0 \\ 0 & 197 & 3 \\ 0 & 22 & 278 \end{bmatrix}$		
taux d'erreur	3.57 %		

Tableau V.c.. Caractéristiques de l'échantillon après classification par l'algorithme Isodata.

VI - CONCLUSION

Les exemples présentés dans ce chapitre montrent l'intérêt de la méthode de classification proposée face à des échantillons de structures très différentes. Même en présence de bruit, cette méthode permet d'obtenir une bonne classification des observations de l'échantillon. La possibilité de discerner des classes de géométrie non gaussienne est intéressante car elle élargit la panoplie d'échantillons qu'il est possible d'analyser. Par rapport aux méthodes existantes, notre approche permet de distinguer des classes présentants des chevauchements relativement importants. Le dernier exemple présenté montre jusqu'à quel degré de chevauchement il est possible d'obtenir une classification correcte, mais l'ajustement du seuil de binarisation de la fonction de densité de probabilité filtrée devient alors très délicat, sinon problèmatique.

CONCLUSION GENERALE

CONCLUSION GENERALE

L'approche utilisée dans ce mémoire pour la classification automatique de données multidimensionnelles est basée sur l'estimation de la fonction de densité de probabilité sous-jacente à la distribution des observations. La méthode des k plus proches voisins a été choisie pour sa capacité d'adaptation à la densité locale de la distribution. Ceci permet d'analyser des échantillons dont les classes sont de densité inégale.

La séparation de l'algorithme d'estimation en deux étapes distinctes, à savoir l'ordonnancement des voisins puis le calcul proprement dit de l'estimateur, permet un gain de temps appréciable surtout si l'analyste veut relancer le processus d'estimation pour différentes valeurs du nombre de voisins.

Mais la mise en mémoire des indices des plus proches voisins n'est pas seulement utile dans la phase d'estimation. En effet, nous avons vu que le nombre k de voisins pris en compte n'a pas une grande influence sur le résultat de l'estimation. L'analyste ne devrait donc pas éprouver le besoin de relancer souvent le processus d'estimation avec différentes valeurs de k. Néanmoins le fait de posséder à tout instant des indices des voisins de chaque observation, permet un gain de temps important dans les phases qui suivent l'estimation de la fonction de densité de probabilité.

L'application d'un filtre non linéaire, à pondération binaire de la contribution

des voisins, permet la recherche des noyaux des modes de la fonction de densité. Les résultats obtenus sur divers exemples montrent la robustesse de ce filtre face à des échantillons dont les classes ont des densités et des formes très différentes, et ceci même en présence de bruit dans la distribution.

Ces deux premières phases de la classification, puis l'étiquetage des noyaux et l'assignation des observations au noyau le plus proches exploitent tous la notion de voisinage à taille variable.

La deuxième approche proposée pour la détection des modes est la recherche des contours des modes. Néanmoins, dans cette approche, le réglage des paramètres est plus délicat. On utilisera donc plutôt cette technique dans le cas d'échantillon où les classes possèdent une densité suffisante, mais dont le degré de chevauchement ne permet d'obtenir de bon résultat par la recherche des noyaux.

Il faut toutefois préciser que dans les exemples traités, le nombre de paramètres à ajuster est très réduit puisque l'on a pris à chaque fois un nombre de voisins égal à 20 quelque soit l'échantillon, mais également pour toutes les phases du processus de classification. Seul la recherche des contours exige un nombre de voisins moindre.

La méthode de classification proposée s'avère efficace même dans le cas où les classes ne sont pas sphériques. Dans l'exemple avec des classes en forme de tore, la classification est correcte avec un taux d'erreur très faible.

Les résultats obtenus conduisent à penser que des approches de classification basées sur la méthode des k plus proche voisins devraient donner toute satisfaction, même si la mise en oeuvre de procédures basées sur la notion de voisinage à taille variable et sans aucune discrétisation de l'espace de représentation des données, est souvent difficile et gourmand en temps de calcul.

Pour essayer d'améliorer encore la capacité d'adaptation à la densité locale de la méthode proposée, il serait intéressant de prendre en compte un nombre de voisins différents en chaque point d'étude.

REFERENCES BIBLIOGRAPHIQUES

REFERENCES BIBLIOGRAPHIQUES

[ASS89] J.P. ASSELIN DE BEAUVILLE
"Panorama sur l'utilisation du mode en classification automatique"
RAIRO-APII, AFCET, n° 2, pp 113-137, 1989.

[BAL65] G.M. BALL
"Data analysis in the social sciences : what about the details ?"
Proc. F.J.C.C. pp 533-560, 1965.

[BAL67] G.H. BALL and D.J. HALL.
"A clustering technique for summarising multivariate data"
Behavioural Science, Vol. 12, pp 153-155, 1967.

[BAY80] C.K. BAYNE, J.J. BEAUCHAMP, C.L. BEGOVITCH AND V.E. KANE
"Monte Carlo comparison of selected clustering procedures"
Pattern Recognition, Vol. 12, pp 51-62, 1980.

[BOT91] C. BOTTE-LECOCQ
"L"analyse de données multidimensionnelles par transformation morphologique binaire"
Thèse de Doctorat, Université des Sciences et Technologies de Lille, 1991.

[CAC62] T. CACOULLOS
"Estimation of multivariate density"
Am. Inst. Stat. Math., Vol. 18, pp 179-189, 1982.

[COV67] J.M. COVER and P.E. HART
"Nearest neighbour pattern classification"
I.E.E.E. Trans. Info. Theory, Vol. IT-13, n°1, pp 21-22, 1967.

[DAL62] R.F. DALY
"The adaptative binary detection problem on the real line"
Technical report 2003-3, Stanford University, Stanford, California, 1962

[DAV 88] E.R. DAVIES
"On the noise suppression and image enhancement characteristics of the median, truncated median, and mode filters."
Pattern Recognition Letters, n°7, pp 87-97, 1988.

[DAY69] N.E. DAY
"Estimating the components of a mixture of normal distributions"
Biometrika, Vol. 56, pp 463-474, 1969

[DID71] E. DIDAY
"Une nouvelle méthode en classification automatique et reconnaissance des formes : la méthode des nuées dynamiques"
Rev. Stat. Appl., Vol. 19, n° 2, pp 20-33, 1971

[DID82] E. DIDAY, J. LEMAN, J. POUGET ET F. TESTU
"Eléments d'analyse de données"
Bordas, Paris, 1982.

[DUD73] R.O. DUDA and P.E. HART
"Pattern Classification and scene analysis"
Wiley, New-York, 1973

[FOR74] F.R. FROM and R.A. NORTHOUSE
"Class : a non parametric clustering algorithm"
Pattern Recognition, Vol. 8, pp 107-114, 1974.

[FRI87] H.P. FRIEMAN and J.RUBI
"On some invariant criteria for grouping data"
J.American Statistical Assn., Vol. 62, pp 1159-1118, 1967.

[FUK70] K. FUKUNAGA and W.L.G. KOONTZ
"A criterion and an algorithm for grouping data"
I.E.E.E. Trans. Comp., Vol. C-19, pp 917-923, 1970.

[FUK75] K. FUKUNAGA and L.D. HOSTETLER
"The estimation of the gradient of a density function with applications in pattern recognition"
I.E.E.E. Trans. Info. Theory, Vol. IT-21, n°1, pp 32-40, 1975.

[FUK84] K. FUKUNAGA and J.M. MANTOCK
"Non parametric Data Reduction"
I.E.E.E. Trans. Pattern Anal. Machine Intell., Vol PAMI-6, n°7, pp115-118, 1984.

[HAM77] R.W. HAMMING
"Digital filters"
Engelwood Cliffs NY, Prentice Hall 1977.

[HAS66] V. HASSELBLAD
"Estimation of parameters for a mixture of normal distributions"
Technometrics, Vol. 8, pp 431-444, 1966.

[HIL68] C.G. HILLBORN and D.G. LAINIOTIS
"Optimal unsupervised learning multicategory dependent hypotheses pattern recognition"
IEEE Trans. on Info. Theory, Vol. IT-14, pp 468-470, 1968.

[IAN79] A. IANNING and S.D. SHAPIRO
"An iterative generalized of the Sobel edge detection operator"
Proc. I.E.E.E. Conf. Pattern Recognition and Image Processing, pp 130-137, 1979.

[JAI88] A.K. JAIN and R.C. DUBES
"Algorithms for clustering data"
Prentice-Hall, Englewood Cliffs, New Jersey, 1988.

[JON68] K.L. JONES
"Problems of grouping individuals and the method of modality"
Behavioral Science, Vol. 13, pp 496-511, 1968.

[JUS78] B.I. JUSTUSSON
"Noise reduction by median filtering"
Proc. I.E.E.E. Conf. Image Processing, pp 502-504, 1978.

[JUS81] B.I. JUSTUSSON
"Median filtering : Statistical Properties"
Topics in Applied Physics, Vol. 43, pp 161-196, 1981.

[KOO76] W.L.G. KOONTZ, P.M. NARENDRA and K. FUKUNAGA
"A graph theoretic approach to non parametric cluster analyser"
I.E.E.E. Trans. Comp., Vol C-25, n°9, pp 936-944, 1976

[LAN67] G.N. LANE and W.T. WILLIAMS
"A general theory of classificatory sorting strategies
1. Hierarchical systems"
Computer J., Vol. 9, pp 973-980, 1967.

[LIP89] R.P. LIPPMAN
"Pattern classification using neural networks"
I.E.E.E. Communications magazine, pp 47-64, 1989.

[LUK79] A. LUKASOVA
"Hierarchical agglomerative clustering procedure"
Pattern Recognition, Vol. 11, pp 365-381, 1979.

[MAC67] J. MACQUEEN
"Some methods for classification and analysis of multivarite observation"
Proc. 5th Berkeley Symp. on Math. Stat. and Prob., Vol 1, pp 281-297, University of California Press, Berkeley and Los Angeles, 1967.

[MAK77] U.E. MAKOV and A.F.M. SMITH
"A quasi-Bayes unsupervised learning procedure for priors"
I.E.E.E. Trans. on Info. Theory, Vol. IT-24, n° 6, pp 761-764, 1977.

[MAT75] G. MATHERSON
"Random sets and integral geometry"
John Wiley , New-York, 1975.

[MOD 77] J.W. MODESTINO et R.W. FRIES.
"Edge detection in noisy images using recursive digital filtering."
Comput. Graphics Image Processing, n°6, pp 409-433, 1977.

[MUR66] V.K. MURTY
"Non parametric estimation of multivariate densities with application"
Multivariate Analysis, Academic Press, New-York, pp 43-56, 1966.

[OLE88] S. OLEJNIK
"Analyse de la convexité d'une fonction de densité de probabilité par étiquetage probabiliste : application à la classification automatique non supervisée"
Thèse de doctorat, Université des Sciences et Technologies de Lille, 1988

[PAR62] E. PARZEN
"On estimation of a probability density function and mode"
Am Math. Stat., Vol. 33, pp 1065-1076, 1962.

[POS82] J.G. POSTAIRE and C.VASSEUR
"A fast algorithm for non parametric probability density function"
I.E.E.E. Trans. on Pattern Analysis and Machine Intell., Vol PAMI-4, n°6, pp 663-666, 1982.

[POS82b] J.-G. POSTAIRE
"Optimisation du processus de classification automatique par analyse de la convexité des fonctions de densités"
Thèse d'ètat, Université de Sciences et Technologies de Lille, 1981.

[PRE 70] J.M.S. PREWIT
"Objects enhancement and extraction, in Picture Processing and Psychopictorics"
B.S. Lipkip et A. Rosenfeld Eds, Academic Press, New-York, pp 75-149, 1970.

[RAB75] L.R. RABINEE, M.R. SAMBUR, and C.E. SCHMIDT.
"Application of non linear smoothing algorithm to speech processing"
I.E.E.E. Trans. on Accoustics, Speech, and Signal processing, Vol. ASSP-23 n°6, pp 552-557, 1975.

[ROB65] L.G. ROBERTS
"Machine perception of three dimensional solids, in optical and electro-optical information processing."
T.J. Tippit et al. Eds, MIT Press, Cambridge Mass., 1965.

[ROS56] A. ROSENBLATT
"Remarks on some non parametric estimates of density function"
Am. Math. Stat., Vol.27, pp 232-237, 1956.

[SCH76] A. SCHROEDER
"Analyse d'un mélange de distributions de probabilité de même type"
Rev. Stat. App., Vol. 24, n°1, pp 39-62, 1976.

[SER82] J. SERRA
"Image analysis and mathematical morphology"
Academic Press, New-York, 1982.

[SOK63] R.R. SOKAL and P.H.A SNEATH
"Principles of numerical taxonomy"
W.H. Freeman Ed., San Francisco, 1963.

[STE86] J.R. STERNBERG
"Grayscale Morphology"
Computer Vision, Graphics and Image Processing, Vol. 35, pp 335-355, 1986.

[TOU87] A. TOUZANI
"Classification automatique par detection des contours des modes des fonctions de densité de probabilité multivariables et étiquetage probabiliste"
Thèse d'état, Université de Sciences et Technologies de Lille, 1987.

[TOU88] A. TOUZANI and J.G. POSTAIRE
"Mode detection by relaxation"
I.E.E.E. Trans. on Pattern Anal. and Machine Intell., Vol PAMI-10, pp 970-978, 1988.

[TOU89] A. TOUZANI and J.G. POSTAIRE
"Clustering by mode boundary detection"
Pattern Recognition letters, Vol. 9, pp 1-12, 1989.

[VAS80] C. VASSEUR and J.G. POSTAIRE
"A convexity testing method for cluster analysis"
I.E.E.E. Trans. Syst. Man. Cybern., Vol. SCM-10, n°3, pp 145-149, 1980.

[WOL70] J.H. WOLF
"Pattern clustering by multivariate mixture analysis"
Multi. Behav. Res., Vol 5, pp 329-350, 1970.

REFERENCES LIEES A CE TRAVAIL

E. CZESNALOWICZ et J.-G. POSTAIRE
"Un algorithme d'estimation rapide par la méthode des k plus proches voisins"
7e Congrès AFCET-INRIA Reconnaissance des Formes et Intelligence Artificielle, pp 633-642, Paris, 29 nov-1er dec, 1989.

E. CZESNALOWICZ et J.-G. POSTAIRE
"Filtrage médian de l'estimateur des k plus proches voisins pour la détection des modes en classification automatique"
8th IASTED International Symposium Applied Informatics, pp 88-91, Innsbruck, Austria, Feb. 20-23; 1990.

E. CZESNALOWICZ et J.-G. POSTAIRE
"Détection des contours des modes sur l'estimateur des k plus proches voisins - Application en classification automatique"
XXIIème Journées de Statistique, pp 82-84, Tours, 28 mai-1er juin, 1990.

E. CZESNALOWICZ and J.-G. POSTAIRE
"Adaptive median filtering for mode detection with application to pattern classification"
9th COMPSTAT Symposium on Computational Statistics, Dubrovnik, Yugoslavia, Sep 9-15, 1990.