REVUE DE STATISTIQUE APPLIQUÉE

J. G. POSTAIRE

Une approche statistique unique pour l'analyse des mélanges et la détection des modes en classification automatique

Revue de statistique appliquée, tome 31, n° 4 (1983), p. 17-36

<http://www.numdam.org/item?id=RSA_1983__31_4_17_0>

UNE APPROCHE STATISTIQUE UNIQUE POUR L'ANALYSE DES MELANGES ET LA DETECTION DES MODES EN CLASSIFICATION AUTOMATIQUE

J.G. POSTAIRE

Centre d'Automatique
Université de Lille 1
59655 Villeneuve d'Ascq Cedex

RESUME

Cet article présente une approche statistique globale et cohérente de différents problèmes de classification automatique. On montre comment l'analyse locale de la convexité des fonctions de densité de probabilité permet d'aborder, avec les mêmes outils mathématiques, des aspects aussi différents de la classification non supervisée que la recherche des groupements par détection des modes et l'optimisation de la classification par identification des mélanges.

Deux algorithmes sont proposés pour chacun de ces problèmes. L'un est destiné aux grands échantillons, l'autre aux échantillons de taille réduite. L'accent est porté sur la comparaison des performances et des domaines d'application de ces quatre procédures dont les comportements sont analysés sur des données artificielles.

Dans tous les cas, la classification est obtenue uniquement à partir de l'information apportée par les observations non étiquetées de l'échantillon disponible. La variété des conditions d'utilisation et des hypothèses d'application de l'approche proposée permet une réponse souple et adaptée aux nombreuses exigences des problèmes de classification automatique non supervisée.

1. INTRODUCTION

La classification automatique apparaît encore trop souvent comme un arsenal de méthodes qui n'ont en commun que leur finalité et qui font appel à autant de notions mathématiques et de concepts scientifiques qu'il existe d'algorithmes. L'utilisateur le plus averti reste souvent perplexe devant le foisonnement des techniques qui permettent d'identifier les classes en présence dans un ensemble de données multidimensionnelles. Selon le nombre d'observations à classer et la dimension des données, selon la possibilité d'utiliser un modèle paramètrique et compte tenu des informations dont il dispose a priori, l'analyste doit choisir une stratégie parmi un ensemble de méthodes très diverses et souvent peu comparables, chacune ayant ses points forts et ses faiblesses.

Ce manque de cohérence et d'unité apparaît surtout en classification non supervisée, c'est-à-dire lorsqu'il s'agit d'identifier les classes en présence dans un échantillon à partir de la seule information qui peut être extraite des observations à classer. Ce type de situation correspond à des démarches exploratoires où on ne dispose d'aucune information a priori sur les données à classer, ne serait-ce que sous la forme de quelques prototypes. Très souvent, au début de l'analyse, on ne connait même pas le nombre de classes en présence.

Mots-Clés : **Classification automatique ; Convexité ; Analyse des mélanges ; Recherche des groupements.**

Certaines méthodes de classification non supervisée sont basées sur des concepts de distance [1-6] tandis que d'autres se réfèrent à des notions statistiques [7-11]. Si certains algorithmes sont totalement heuristiques [12-14], d'autres se veulent optimaux en ce sens qu'ils minimisent un critère [1-2, 5, 15-16]. Il faut même noter que certaines méthodes, qui semblent au premier abord apparentées aux approches statistiques, ne font en fait référence à des notions de corrélation et de variance que pour définir un critère algébrique à extrèmiser. Il ne s'agit alors que d'une généralisation de la notion de distance [5, 17-18].

Malgré cette grande diversité d'approches dont les références citées ne donnent qu'un aperçu incomplet et limité, il n'existe aucune méthode permettant d'optimiser une classification non supervisée au sens de la minimisation du taux d'erreur, c'est-à-dire au sens de la théorie de la décision.

Face à cette profusion de méthodes qui masque certaines lacunes, nous avons tenté de résoudre les problèmes classiques de classification non supervisée par une approche unique et cohérente. Le principal but de cet article est de montrer comment l'utilisation du concept de convexité permet de regrouper différents aspects de la classification automatique sous une même théorie unificatrice.

L'approche proposée apporte une solution au problème classique de la recherche des groupements (clustering) lorsqu'on n'émet aucune hypothèse particulière sur la nature statistique de la distribution des données à classer. Mais elle permet également d'optimiser une classification par minimisation du taux d'erreur. Cette possibilité était jusqu'à présent réservée à des problèmes particuliers, avec un nombre réduit de classes [19-20], ou une connaissance a priori de certains paramètres [21-24] ou encore avec des hypothèses restrictives [25-28]. L'optimisation dans le cas très général de la classification non supervisée sans hypothèses restrictives a été obtenue grâce à une méthode d'identification des mélanges basée sur l'analyse de la convexité des fonctions de densité de probabilité. Nous verrons également comment analyser, par la même approche, des données de faible dimension comme des données de dimension élevée, que l'on dispose de petits échantillons ou d'échantillons de taille importante.

Cet article est un essai de synthèse de différents travaux publiés séparément. L'accent est donc mis sur les grandes lignes de l'approche et les aspects fondamentaux de la démarche. Le lecteur intéressé trouvera dans la bibliographie des références d'articles détaillés portant aussi bien sur les fondements théoriques de la méthodologie que sur les aspects pratiques de son implantation sur calculateur numérique.

Les résultats présentés reposent sur la mise au point d'un test qui permet de connaître localement le sens de la convexité de la fonction de densité sous-jacente à une distribution d'observations multidimensionnelles (section 2). On montre comment la description des fonction de densité en termes de convexité permet de proposer un schéma pour l'identification des mélanges gaussiens (section 3). Mais l'utilisation du test de convexité ne pouvant être envisagée que pour des échantillons relativement importants, une variante de cette méthode d'identification des mélanges est proposée pour les petits échantillons. Le schéma directeur est conservé, mais l'analyse de la convexité de la fonction de densité multivariable est remplacée par l'analyse de la convexité de ses densités marginales (section 4).

Quelle que soit la taille de l'échantillon disponible et la dimension du problème, on peut donc, en acceptant l'hypothèse normale, optimiser une classification sur la base de l'identification du mélange sous-jacent (section 5). Mais dans le cas où l'on estime devoir s'affanchir de cette hypothèse, il est encore possible

de regrouper les observations après avoir détecté les modes de la fonction de densité sous-jacente grâce à l'analyse de sa convexité ou de celle de ses densité marginales (section 6).

De nombreux exemples permettent de juger les performances de la procédure d'optimisation ainsi que celles de la méthode de recherche des groupements pour des échantillons de taille et de dimension variées. Ils montrent l'intérêt de cette nouvelle approche qui apporte une réponse souple et adaptée aux exigences des problèmes de classification non supervisée.

2. CONVEXITE DES FONCTIONS DE DENSITE

Nous verrons dans cet article que la détermination de la convexité des fonctions de densité de probabilité apparaît comme une approche intéressante des problèmes de classification automatique. Usuellement, on dit qu'une fonction f(X) est convexe sur un domaine convexe D de R^n si :

$$f[\lambda X_1 + (1 - \lambda) X_2] \leqslant \lambda f(X_1) + (1 - \lambda) f(X_2) \qquad (1)$$

quels que soient X_1, $X_2 \in D$ et $\lambda \in [0, 1]$. Or les fonctions de densité ne sont généralement connues dans les problèmes de classification que sous la forme d'une estimation explicite, obtenue à partir des observations disponibles. La définition classique est donc mal adaptée à la détermination de la convexité de ces fonctions dont on ne connait pas l'expression analytique. En effet, cette définition fait appel à une notion globale, en ce sens que l'inégalité (1) doit être vérifiée pour tout couple de points X_1, X_2 appartenant au domaine sur lequel la fonction est convexe. Lorsqu'on ne dispose pas de la forme analytique de la fonction étudiée, il est difficile d'utiliser cette définition globale pour analyser sa convexité.

Pour pallier cette difficulté, nous proposons une méthode qui consiste à tester localement (ou point par point) le sens de la convexité des fonctions multivariables. Cette analyse fait appel à une notion de convexité en un point que nous définissons comme suit :

Une fonction f(X) est dite "localement convexe" au point X s'il existe un voisinage de X convexe sur lequel la fonction f(X) est convexe. La détermination de la convexité d'une fonction en un point X peut alors être envisagée à partir de l'analyse des variations de sa valeur moyenne :

$$\rho[f(X), D] = \frac{\int_D f(X)\, dX}{V(D)}$$

calculée sur des domaines D appropriés, centrés en X et de volume V(D) variable. Ces domaines, appelés "domaines d'observation", sont obtenus à partir d'un domaine de référence, centré en X et symétrique par rapport à son centre, par une homothétie de centre X (cf. Fig. 1). On montre alors que si f(X) est localement convexe en un point, la valeur moyenne de f(X) calculée sur les domaines d'observation ainsi définis est une fonction croissante de la taille de ces domaines. Le sens de variation est opposé dans le cas concave [29].

Cette relation simple entre le sens de variation de la valeur moyenne de la fonction calculée sur des domaines d'observation de taille croissante et le sens de la convexité de la fonction au point où sont centrés les domaines est aisément exploitable dans le contexte de la classification automatique. Partant d'un ensemble d'observations multidimensionnelles, il est en effet maintenant possible de tester le sens de la convexité de la fonction de densité sous-jacente.

En tout point X où on désire effectuer ce test, on estime d'abord f(X) par la méthode de ROSENBLATT-PARZEN sous la forme :

$$\hat{f}(X) = \frac{k/q}{V(D)}$$

où k est le nombre d'observations de l'échantillon disponible de taille q situées dans le domaine D centré en X et de volume V(D). La taille du domaine D est ajustée en fonction du nombre q d'observations afin d'assurer la convergence en moyenne quadratique de l'estimateur.

$\hat{f}(X)$ est considérée comme la valeur limite de la valeur moyenne de f(X) calculée sur des domaines d'observation centrés en X lorsque leur taille tend vers zéro. Pour connaître le sens de variation de cette valeur moyenne lorsque la taille des domaines d'observation augmente, il suffit d'estimer la valeur moyenne $\hat{\rho}[f(X), D']$ de f(X) sur un domaine D' de taille légèrement supérieure à celle de D sous la forme :

$$\hat{\rho}[f(X), D'] = \frac{k'/q}{V(D')}$$

k' étant le nombre d'observations de l'échantillon situées dans le domaine D'.

Si :

$$\hat{\rho}[f(X), D'] > \hat{f}(X)$$

on conclut que la valeur moyenne de f(X) est une fonction croissante de la taille des domaines d'observation sur lesquels elle est calculée. La fonction f(X) est alors localement convexe en X. Dans le cas contraire, on conclut qu'elle est localement concave en X.

Une des principales critiques faites à la méthode d'estimation des fonctions de densité de ROSENBLATT-PARZEN utilisée dans le test de convexité décrit ci-dessus est que son utilisation demeure limitée à des problèmes de faible dimension. Or le test de convexité proposé nécessite deux foix plus de calculs que l'estimation des fonctions de densité, puisqu'en chaque point il faut d'abord estimer la fonction elle-même, puis sa valeur moyenne sur un domaine d'observation.

Il serait donc irréaliste d'envisager l'utilisation de ce test sans disposer d'une procédure permettant d'accélérer la mise en œuvre de la méthode d'estimation de ROSENBLATT-PARZEN. C'est pourquoi nous avons mis au point un algorithme rapide d'estimation utilisant le noyau cubique [30]. A titre d'exemple, pour un problème à 6 dimensions avec un échantillon de 1 000 observations, cet algorithme permet d'estimer la fonction de densité sous-jacente environ 10 000 fois plus vite que l'algorithme conventionnel. Utilisé pour tester le sens de la convexité locale des fonctions de densité à partir des observations, cet algorithme permet de traiter des problèmes de forte dimension en des temps tout à fait raisonnables.

Il faut toutefois garder à l'esprit le fait que l'application de cette procédure ne permet de déterminer le sens de la convexité locale de la fonction analysée

que si celle-ci est définie au voisinage des points de test. C'est pourquoi nous allons maintenant examiner les propriétés de convexité des fonctions de densité normales, souvent proposées comme modèles paramètriques dans les problèmes de classification automatique.

3. IDENTIFICATION DES MELANGES GAUSSIENS PAR ANALYSE MULTIVARIABLE

3.1. Détermination des paramètres d'une distribution normale par analyse de la convexité de sa fonction de densité

En étudiant le signe du HESSIEN(*) d'une fonction de densité normale d'expression :

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

on peut montrer que celle-ci est concave à l'intérieur d'un hyperellipsoïde d'équation :

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

et que sa concavité n'est pas définie à l'extérieur de ce domaine appelé "domaine de concavité" de p(X) [31].

L'analyse des propriétés géométriques de cet hyperellipsoïde montre que ses axes principaux ont la direction des vecteurs propres de l'inverse de la matrice de covariance Σ. De plus, les demi-longueurs de ces axes ne sont autres que les valeurs propres de la matrice Σ^{-1}. (cf. Fig. 2).

Le test de convexité présenté au paragraphe précédent permet donc de proposer une nouvelle méthode d'identification des paramètres d'une distribution normale à partir d'un échantillon. Le schéma directeur est le suivant :

a) Déterminer, point par point et à partir des observations disponibles, le domaine où la fonction de densité sous-jacente est concave.

b) Modéliser ce domaine par un hyperellipsoïde dont on détermine les directions principales et la longueur des axes principaux.

c) Le centre de l'hyperellipsoïde indique le vecteur moyenne $\bar{X}$ de la distribution. La direction et la longueur des axes principaux indiquent les vecteurs propres et valeurs propres de la matrice Σ^{-1}. Le calcul de la matrice de covariance est donc immédiat.

La procédure a) est réalisée en appliquant le test de convexité à tous les points d'un réseau hypercubique. Lorsque le test indique que la fonction p(x)

(*) Rappelons que les éléments $p_{i,j}(X)$ du HESSIEN de p(X) sont de la forme :

$$p_{i,j}(X) = \frac{\partial^2 p(X)}{\partial x_i\ \partial x_j}\ ;\ i = 1, \ldots, n\ ;\ j = 1, \ldots, n$$

avec : $$X = [x_1, x_2, \ldots, x_i, \ldots, x_n]^T$$

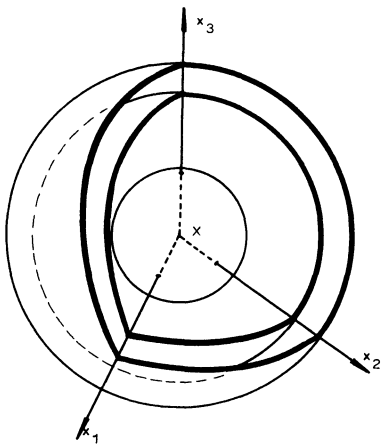


Figure 1. – Domaines d'observation sphériques dans un espace à trois dimensions

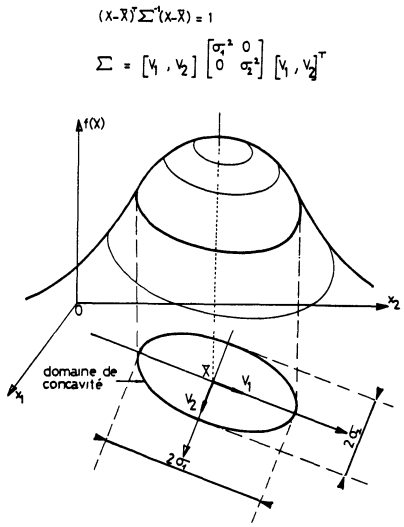
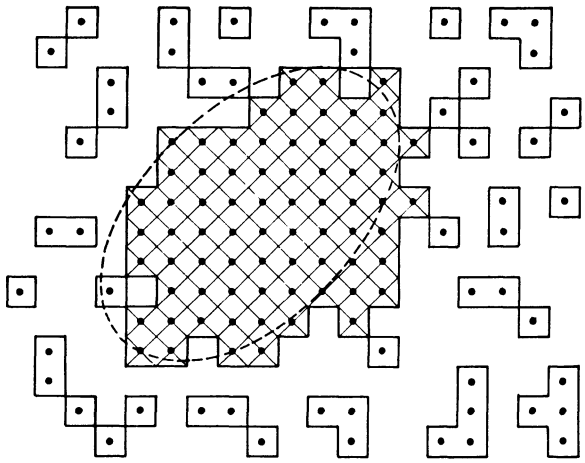


Figure 2. – Propriétés géométriques du "domaine de concavité" de la fonction de densité d'une distribution normale à deux dimensions.



CARRES A CONVEXITE NEGATIVE

CARRES A CONVEXITE NEGATIVE MATERIALISANT LE DOMAINE DE CONCAVITE

DOMAINE DE CONCAVITE

Figure 3. – Résultat du test de convexité appliqué à une distribution normale d'observations bidimensionnelles.

est convexe en un point, on dit que ce point a une convexité positive, par opposition aux points à convexité négative où le test indique que la fonction est concave.

Tous les points du réseau situés à l'intérieur du domaine de concavité de p(X) seront des points à convexité négative. Par contre à l'extérieur de ce domaine, la convexité de la fonction n'étant pas définie, le test de convexité fera apparaître indifféremment des points à convexité négative et positive (cf. Fig. 3). Une simple procédure d'agrégation, portant sur les points à convexité négative, permet donc d'isoler par chaînage les points adjacents matérialisant le domaine de concavité de la fonction.

Une technique de calcul apparentée à la recherche de l'ellipsoïde d'inertie d'un solide permet alors de déterminer les axes principaux du domaine matérialisé par les points isolés par la procédure d'agrégation. La détermination des paramètres de la distribution est alors immédiate.

La détermination, à partir des observations, du domaine à l'intérieur duquel une fonction de densité est concave permet donc de calculer des valeurs approchées du vecteur moyenne et de la matrice de covariance d'une distribution normale. Si l'intérêt de cette nouvelle approche pour l'identification d'une distribution unique reste très limité, il n'en est pas de même pour ce qui est de l'analyse des mélanges gaussiens qui a motivé cette étude.

3.2. Identification des mélanges gaussiens par analyse de la convexité des fonctions de densité

On s'intéresse maintenant à des observations dont la distribution peut être décrite par un mélange gaussien de fonction de densité :

$$f(X) = \sum_{k=1} p(X \mid C_k)\, P(C_k)$$

où $p(X \mid C_k), k = 1, 2, \ldots, k$ sont les fonctions de densité de chacune des classes en présence et où $P(C_k), k = 1, 2, \ldots, k$ sont les probabilités a priori de ces composantes.

Il est d'usage, en classification automatique, de supposer qu'il existe une correspondance bijective entre les modes de f(X) et les classes en présence [7]. Qui plus est, nous verrons sur des exemples que les degrés de chevauchement entre composantes couramment rencontrés en pratique permettent d'assimiler, avec une très bonne approximation, les domaines de concavité des mélanges à ceux de leurs composantes. Le test de convexité permet de déterminer les domaines de concavité d'un mélange à partir des observations. En suivant le schéma directeur présenté au paragraphe précédent pour chacune des composantes ainsi mise en évidence, il devient alors possible d'obtenir des valeurs approchées du vecteur moyenne et de la matrice de covariance de chaque classe. Les seuls paramètres restant à déterminer pour identifier complètement les mélanges sont les probabilités a priori des différentes classes. Il est aisé de montrer qu'elles sont sensiblement proportionnelles au nombre d'observations situées à l'intérieur des domaines de concavité associés aux classes mises en évidence.

L'identification d'un mélange d'un nombre inconnu de composantes peut donc être entreprise selon le schéma directeur suivant [31] :

a) Déterminer point par point et à partir des observations disponibles les domaines de concavité de la fonction de densité sous-jacente.

b) Modéliser ces domaines par des hyperellipsoïdes.

c) A partir des caractéristiques géométriques de ces hyperellipsoïdes, déterminer des valeurs approchées des vecteurs moyenne et matrices de covariance de chacune des classes mises en évidence.

d) Dénombrer le nombre d'observations situées à l'intérieur de chaque domaine de concavité pour calculer les probabilités a priori des classes constituant le mélange.

4. IDENTIFICATION DES MELANGES GAUSSIENS PAR ANALYSE MONOVARIABLE

Le test de convexité qui permet de connaître le sens de la convexité des fonctions de densité multivariables repose sur une technique d'estimation non paramètrique qui nécessite un nombre d'observations disponibles d'autant plus important que la dimension des données est élevée. Pour les échantillons trop petits pour permettre l'utilisation de ce test, on utilise une variante de la méthode précédente qui repose sur l'analyse de la convexité des densités marginales des fonctions de densité multivariables dont l'estimation ne pose pas de problème, même avec des échantillons de taille très réduite [32].

Sous l'hypothèse d'indépendance des variables, il est aisé de montrer que les densités marginales d'une fonction normale présentent des segments concaves qui sont les projections du domaine de concavité sur les axes. L'analyse de la convexité des densités marginales d'un mélange gaussien estimées à partir des observations indiquent donc les projections sur les axes des domaines de concavité des composantes. Le produit euclidien de ces segments permet ensuite de reconstituer des domaines hyperparallélépipédiques dont certains, appelés "domaines caractéristiques", définissent les composantes du mélange. D'autres, artificiellement introduits par le produit euclidien, sont facilement rejetés du fait qu'ils ne contiennent que très peu d'observations (cf. Fig. 4).

Les centres des domaines caractéristiques retenus indiquent les vecteurs moyenne des composantes mises en évidence ; leurs dimensions dans les différentes directions indiquent les éléments diagonaux des matrices de covariance. Comme

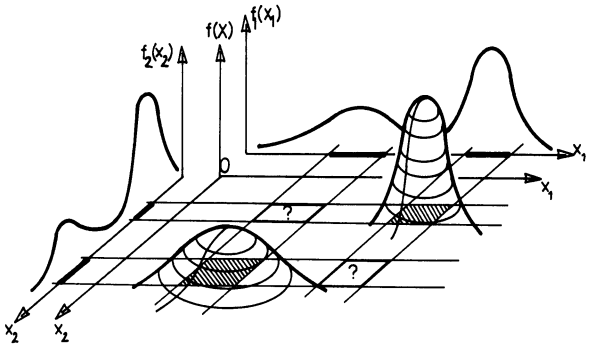


Figure 4. – Détermination des "domaines caractéristiques" d'un mélange gaussien à partir des segments concaves de ses fonctions de densité marginales.

précédemment, le dénombrement des observations situées à l'intérieur de chacun de ces domaines permet de calculer les probabilités a priori des différentes classes. Le schéma directeur de cette méthode d'identification des mélanges est alors le suivant [33] :

a) Estimer les densités marginales de la distribution des observations disponibles et en déterminer les segments concaves.

b) Déterminer les domaines caractéristiques des composantes par le produit euclidien de ces segments concaves.

c) A partir des caractéristiques géométriques de ces domaines, déterminer les valeurs des vecteurs moyenne et des matrices de covariance des classes mises en évidence.

d) Dénombrer le nombre d'observations situées à l'intérieur de chaque domaine caractéristique pour calculer les probabilités a priori des classes constituant le mélange.

5. OPTIMISATION DU PROCESSUS DE CLASSIFICATION

A ce stade, nous disposons de deux méthodes d'analyse des mélanges gaussiens, l'une destinée à de grands échantillons de taille relativement réduite, l'autre mieux adaptée aux petits échantillons, même de dimension élevée.

L'identification du mélange dont sont supposées extraites les observations à classer permet d'envisager une classification optimale puisque l'on dispose de valeurs approchées de tous les paramètres nécessaires au calcul des fonctions de décision classiques :

$$g_k(X) = -\frac{1}{2}\,\mathrm{Log}\,|\Sigma_k| - \frac{1}{2}(X-\bar{X}_k)^T\,\Sigma_k^{-1}(X-\bar{X}_k) + \mathrm{Log}\,P(C_k)\,,\ k=1,2,\ldots,k$$

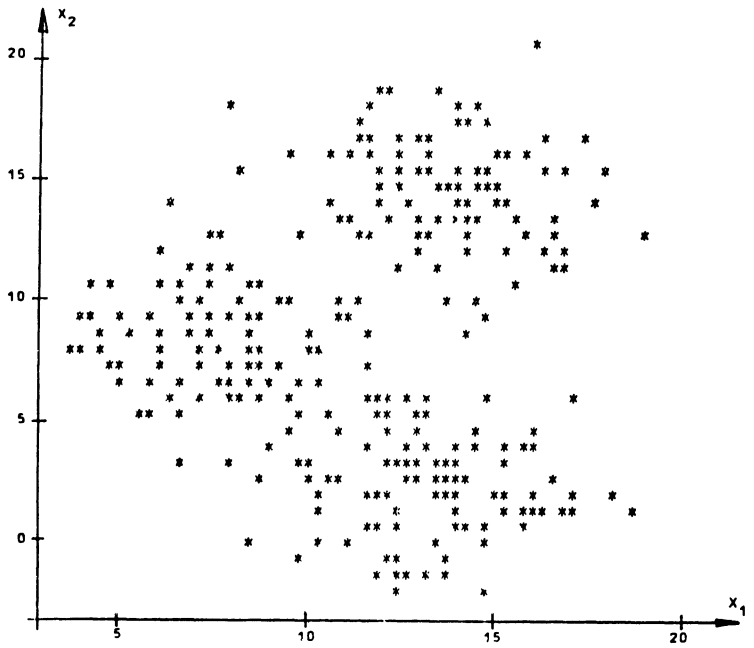
En affectant une observation X à la classe C_k telle que :

$$g_k(X) \geqslant g_i(X)\ ,\ i=1,2,\ldots,k\ ,\ i\neq k$$

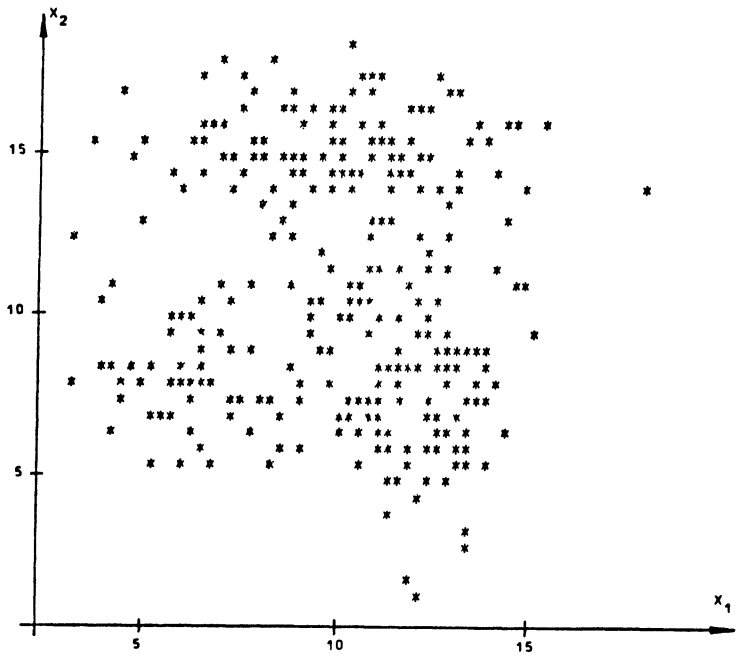
on est assuré de minimiser le risque d'erreur en ce sens qu'aucune autre procédure ne peut conduire à un taux d'erreur plus petit. La différence entre le taux d'erreur minimum théorique et le taux d'erreur effectif dépend du degré d'approximation des valeurs des paramètres des mélanges obtenues par les méthodes proposées [34].

Nous présentons maintenant quelques exemples de classification sur des données générées artificiellement qui permettent de juger de la qualité de ces approximations. Nous commencerons par trois exemples bi-dimensionnels qui peuvent être aisément visualisés.

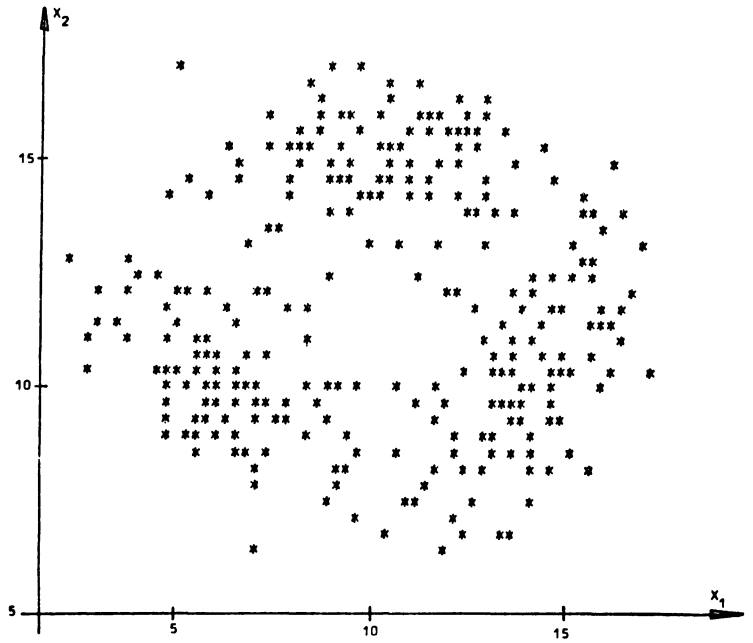
Exemple 1 : Les données du premier exemple, représentées figure 5a, sont issues d'un mélange de trois classes sphériques équiprobables dont les paramètres sont consignés dans le tableau 1a. Les deux méthodes d'analyse des mélanges ont été appliquées à ces données et les valeurs approchées des paramètres des classes ainsi obtenues figurent également dans ce tableau. Pour les deux méthodes exposées ci-dessus, le taux d'erreur de la classification basée sur ces valeurs approchées reste très proche du taux d'erreur théorique optimal.



a) exemple 1,



b) exemple 2,



c) exemple 3.

Figure 5. – Représentation graphique des échantillons bidimensionnels utilisés pour illustrer la procédure d'optimisation.

Exemple 2 : Le second exemple, représenté figure 5b et dont les caractéristiques sont indiquées au tableau 1b est destiné à mettre en évidence les possibilités d'analyse de mélanges de classes non sphériques, avec des probabilités a priori différentes. Ici encore, les valeurs approchées des paramètres obtenues par les deux méthodes d'analyse des mélanges permettent de classer les données avec des taux d'erreur très proches du taux théorique minimum. La différence entre le taux d'erreur effectif et le taux optimal correspond à seulement 3 observations mal classées sur un total de 300 pour la méthode d'analyse multivariable et à 5 observations pour la méthode d'analyse monovariable.

Exemple 3 : Le troisième exemple, représenté figure 5c, est destiné à montrer les possibilités d'analyse des mélanges de composantes ayant des matrices de covariance non diagonales. La méthode d'analyse monovariable étant réservée aux cas où les variables sont indépendantes, il ne sera traité que par la méthode d'analyse multivariable. Les paramètres du mélange et leurs valeurs approchées ainsi obtenues sont donnés dans le tableau 1c. Encore une fois, le taux d'erreur de la classification reste très proche de la valeur optimale.

Pour des données de dimension plus élevée, le comportement des deux procédures reste très semblable à celui décrit dans ces exemples bi-dimensionnels. Le tableau 2 indique quelques résultats pour des données multidimensionnelles avec des degrés de chevauchement entre classes variables. Ces résultats sont encore très satisfaisants, mais ce qu'il convient surtout de remarquer, c'est la progression quasi linéaire du temps d'exécution des algorithmes avec la dimension des données. Cette propriété est bien entendu le fruit du nouvel algorithme d'estimation non paramètrique mentionné section 2.

TABLEAU 1

Comparaison des performances des deux méthodes d'analyse des mélanges pour des échantillons bidimensionnels.

	VALEURS EXACTES DES PARAMETRES DU MELANGE	VALEURS APPROCHEES DES PARAMETRES DU MELANGE	
		ANALYSE MULTIVARIABLE	ANALYSE MONOVARIABLES
CLASSE 1	$\bar{X}_1=\begin{bmatrix}8\\8\end{bmatrix}$ $\Sigma_1=\begin{bmatrix}4 & 0\\0 & 4\end{bmatrix}$ $P(C_1)=0,33$	$\hat{\bar{X}}_1=\begin{bmatrix}8,09\\8,52\end{bmatrix}$ $\hat{\Sigma}_1=\begin{bmatrix}4,41 & 0,32\\0,32 & 5,56\end{bmatrix}$ $\hat{P}(C_1)=0,34$	$\hat{\bar{X}}_1=\begin{bmatrix}8,48\\8,37\end{bmatrix}$ $\hat{\Sigma}_1=\begin{bmatrix}3,56 & 0\\0 & 4,75\end{bmatrix}$ $\hat{P}(C_1)=0,36$
CLASSE 2	$\bar{X}_2=\begin{bmatrix}14\\14\end{bmatrix}$ $\Sigma_2=\begin{bmatrix}4 & 0\\0 & 4\end{bmatrix}$ $P(C_2)=0,33$	$\hat{\bar{X}}_2=\begin{bmatrix}13,74\\14,29\end{bmatrix}$ $\hat{\Sigma}_2=\begin{bmatrix}3,83 & -0,09\\-0,09 & 3,20\end{bmatrix}$ $\hat{P}(C_2)=0,31$	$\hat{\bar{X}}_2=\begin{bmatrix}11,86\\14,18\end{bmatrix}$ $\hat{\Sigma}_2=\begin{bmatrix}3,16 & 0\\0 & 2,99\end{bmatrix}$ $\hat{P}(C_2)=0,30$
CLASSE 3	$\bar{X}_3=\begin{bmatrix}14\\2\end{bmatrix}$ $\Sigma_3=\begin{bmatrix}4 & 0\\0 & 4\end{bmatrix}$ $P(C_3)=0,33$	$\hat{\bar{X}}_3=\begin{bmatrix}14,21\\2,46\end{bmatrix}$ $\hat{\Sigma}_3=\begin{bmatrix}5,67 & 0,67\\0,67 & 6,67\end{bmatrix}$ $\hat{P}(C_3)=0,35$	$\hat{\bar{X}}_3=\begin{bmatrix}13,05\\2,18\end{bmatrix}$ $\hat{\Sigma}_3=\begin{bmatrix}3,16 & 0\\0 & 5,98\end{bmatrix}$ $\hat{P}(C_3)=0,34$
	TAUX D'ERREUR THEORIQUE MINIMUM: 3,27%	TAUX D'EPREUR DE LA CLASSIFICATION: 3,33%	TAUX D'ERREUR DE LA CLASSIFICATION: 5,0%

a) exemple 1,

	VALEURE EXACTES DES PARAMETRES DU MELANGE	VALEURS APPROCHEES DES PARAMETRES DU MELANGE	
		ANALYSE MULTIVARIABLE	ANALYSE MONOVARIABLE
CLASSE 1	$\bar{X}_1=\begin{bmatrix}7\\8\end{bmatrix}$ $\Sigma_1=\begin{bmatrix}2 & 0\\0 & 2\end{bmatrix}$ $P(C_1)=0,20$	$\hat{\bar{X}}_1=\begin{bmatrix}7,38\\7,60\end{bmatrix}$ $\hat{\Sigma}_1=\begin{bmatrix}1,76 & -0,20\\-0,20 & 1,77\end{bmatrix}$ $\hat{P}(C_1)=0,21$	$\hat{\bar{X}}_1=\begin{bmatrix}7,41\\7,72\end{bmatrix}$ $\hat{\Sigma}_1=\begin{bmatrix}1,69 & 0\\0 & 1,88\end{bmatrix}$ $\hat{P}(C_1)=0,22$
CLASSE 2	$\bar{X}_2=\begin{bmatrix}12\\8\end{bmatrix}$ $\Sigma_2=\begin{bmatrix}2 & 0\\0 & 6\end{bmatrix}$ $P(C_2)=0,40$	$\hat{\bar{X}}_2=\begin{bmatrix}12,19\\7,37\end{bmatrix}$ $\hat{\Sigma}_2=\begin{bmatrix}2,61 & -0,08\\-0,08 & 4,58\end{bmatrix}$ $\hat{P}(C_2)=0,40$	$\hat{\bar{X}}_2=\begin{bmatrix}12,42\\7,29\end{bmatrix}$ $\hat{\Sigma}_2=\begin{bmatrix}2,54 & 0\\0 & 4,51\end{bmatrix}$ $\hat{P}(C_2)=0,36$
CLASSE 3	$\bar{X}_3=\begin{bmatrix}10\\15\end{bmatrix}$ $\Sigma_3=\begin{bmatrix}7 & 0\\0 & 2\end{bmatrix}$ $P(C_3)=0,40$	$\hat{\bar{X}}_3=\begin{bmatrix}10,14\\14,49\end{bmatrix}$ $\hat{\Sigma}_3=\begin{bmatrix}10,55 & 0,61\\0,61 & 1,73\end{bmatrix}$ $\hat{P}(C_3)=0,39$	$\hat{\bar{X}}_3=\begin{bmatrix}9,38\\14,80\end{bmatrix}$ $\hat{\Sigma}_3=\begin{bmatrix}7,60 & 0\\0 & 2,14\end{bmatrix}$ $\hat{P}(C_3)=0,42$
	TAUX D'ERREUR THEORIQUE MINIMUM: 4,3%	TAUX D'ERREUR DE LA CLASSIFICATION: 5,3%	TAUX D'ERREUR DE LA CLASSIFICATION: 6,0%

b) exemple 2,

	VALEURS EXACTES DES PARAMETRES DU MELANGE	VALEURS APPROCHEES DES PARAMETRES DU MELANGE ANALYSE MULTIVARIABLE
CLASSE 1	$\bar{X}_1 = \begin{bmatrix} 10 \\ 15 \end{bmatrix}$ $\Sigma_1 = \begin{bmatrix} 7 & 0 \\ 0 & 2 \end{bmatrix}$ $P(C_1)=0,33$	$\hat{\bar{X}}_1 = \begin{bmatrix} 9,8 \\ 14,8 \end{bmatrix}$ $\hat{\Sigma}_1 = \begin{bmatrix} 8,25 & -0,69 \\ -0,69 & 2,17 \end{bmatrix}$ $\hat{P}(C_1)=0,37$
CLASSE 2	$\bar{X}_2 = \begin{bmatrix} 6 \\ 10 \end{bmatrix}$ $\Sigma_2 = \begin{bmatrix} 4,5 & 2,5 \\ 2,5 & 4,5 \end{bmatrix}$ $P(C_2)=0,33$	$\hat{\bar{X}}_2 = \begin{bmatrix} 6,6 \\ 9,5 \end{bmatrix}$ $\hat{\Sigma}_2 = \begin{bmatrix} 4,49 & 2,74 \\ 2,74 & 5,60 \end{bmatrix}$ $\hat{P}(C_2)=0,31$
CLASSE 3	$\bar{X}_3 = \begin{bmatrix} 14 \\ 10 \end{bmatrix}$ $\Sigma_3 = \begin{bmatrix} 4,5 & -2,5 \\ -2,5 & 4,5 \end{bmatrix}$ $P(C_3)=0,33$	$\hat{\bar{X}}_3 = \begin{bmatrix} 13,9 \\ 9,8 \end{bmatrix}$ $\hat{\Sigma}_3 = \begin{bmatrix} 4,85 & -3,46 \\ -3,46 & 6,25 \end{bmatrix}$ $\hat{P}(C_3)=0,33$
	TAUX D'ERREUR THEORIQUE MINIMUM: 3,9%	TAUX D'ERREUR DE LA CLASSIFICATION: 5,3%

c) exemple 3.

TABLEAU 2

Comparaison des performances des deux méthodes d'analyse des mélanges pour des données multidimensionnelles. (Le degré de chevauchement des classes de l'échantillon n° 2 n'a pas permis de retrouver la structure du mélange par la méthode d'analyse monovariable).

	PARAMETRES DES MELANGES ANALYSES			ANALYSE MULTIVARIABLE		ANALYSE MONOVARIABLE	
	VECTEURS MOYENNE	MATRICES DE COVARIANCE	TAUX D'ERREUR THEORIQUE MINIMUM	TAUX D'ERREUR DE LA CLASSIFICATION	TEMPS DE CALCUL	TAUX D'ERREUR DE LA CLASSIFICATION	TEMPS DE CALCUL
ECHANT.N°1	$\bar{X}_1=\begin{bmatrix}8\\8\end{bmatrix}$ $\bar{X}_2=\begin{bmatrix}14\\14\end{bmatrix}$ $\bar{X}_3=\begin{bmatrix}14\\?\end{bmatrix}$	$\Sigma_k=4\ I_2, k=1,2,3$	3,3%	3,4%	27"	5,0%	21"
ECHANT.N°2		$\Sigma_k=8\ I_2, k=1,2,3$	10,1%	11,3%	26"	-	-
ECHANT.N°3	$\bar{X}_1=\begin{bmatrix}8\\8\\8\end{bmatrix}$ $\bar{X}_2=\begin{bmatrix}14\\14\\14\end{bmatrix}$ $\bar{X}_3=\begin{bmatrix}14\\14\\2\end{bmatrix}$	$\Sigma_k=4\ I_3, k=1,2,3$	1,7%	2,3%	35"	3,3%	25"
ECHANT.N°4		$\Sigma_k=8\ I_3, k=1,2,3$	6,1%	7,7%	34"	8,3%	29"
ECHANT.N°5	$\bar{X}_1=\begin{bmatrix}8\\8\\8\\8\end{bmatrix}$ $\bar{X}_2=\begin{bmatrix}14\\14\\14\\14\end{bmatrix}$ $\bar{X}_3=\begin{bmatrix}14\\14\\14\\2\end{bmatrix}$	$\Sigma_k=4\ I_4, k=1,2,3$	1,4%	1,7%	43"	2,3%	31"
ECHANT.N°6		$\Sigma_k=8\ I_4, k=1,2,3$	4,5%	6,3%	45"	7,0%	30"

Il importe de noter que la seule intervention de l'analyste consiste à ajuster le pas de discrétisation qui définit la densité du réseau de points où est testée la convexité des fonctions de densité mono ou multivariables. L'analyse des variations du nombre de modes mis en évidence en fonction du pas de discrétisation permet d'ajuster ce dernier de manière heuristique. On détermine pour cela la plus grande plage de variation de ce pas pour laquelle le nombre de modes détectés reste constant. En se réferant au concept de "stabilité du nombre de classes détectées" [35], le pas est finalement ajusté à la valeur correspondant au milieu de cette plage de variation.

Une seconde précision doit également être apportée, concernant la mise en œuvre de la méthode d'analyse monovariable. Le partage de l'espace en régions de décision relatives à chacune des classes mises en évidence par cette méthode peut être amélioré par une procédure d'analyse locale itérative faisant suite à la procédure d'analyse globale décrite dans la section 4. Cette analyse locale est une transposition directe de l'analyse globale aux observations situées dans chaque région de décision obtenue à l'itération précédente. En général, cette procédure permet de mettre en évidence des classes qui étaient passées inaperçues lors de l'analyse globale et améliore sensiblement la détermination des vecteurs moyenne et matrices de covariance des différentes classes mises en évidence [36].

6. RECHERCHE DE GROUPEMENTS

Par le jeu des nombreux paramètres qui caractérisent les fonctions de densité normales, le modèle gaussien permet de modéliser des distributions très variées. Mais il se présente des cas où, malgré sa souplesse, l'utilisation de ce modèle risque d'imposer une structure aux données plutôt que d'aider à découvrir leur organisation véritable.

Une approche classique du problème de recherche de groupements dans un échantillon d'observations consiste à mettre en évidence les modes de la fonction de densité sous-jacente, une classe étant associée à chaque mode [7]. Traditionnellement, la détection des modes d'une distribution fait appel à des techniques de type "gradient", avec tous les inconvénients que cela comporte. En assimilant les modes aux domaines concaves de la fonction de densité sous-jacente plutôt qu'à ses sommets locaux, on peut les détecter grâce au test de convexité, mono- ou multivariable. Les observations situées à l'intérieur de chaque domaine où la fonction de densité est concave constituent le noyau de la classe ainsi détectée. Le problème de la recherche des groupements se résoud alors en affectant les observations restantes à leur plus proche noyau [37].

Pour illustrer les possiblités de détection des modes et de recherche des groupements par analyse de la convexité des densités de probabilité, nous présentons deux exemples bi-dimensionnels constitués de classes aisément identifiables par examen visuel mais difficilement modélisables (cf. Fig. 6). Dans les deux cas, les groupements obtenus concordent, à quelques exceptions près, avec les conclusions d'une inspection visuelle des données.

Notons que l'analyse de la convexité des densités marginales permet également de détecter les modes lorsque le nombre d'observations disponibles n'est pas suffisant pour appliquer le test de convexité multivariable. La mise en œuvre d'une analyse locale itérative à la suite de l'analyse globale améliore sensiblement les performances de la méthode [38].

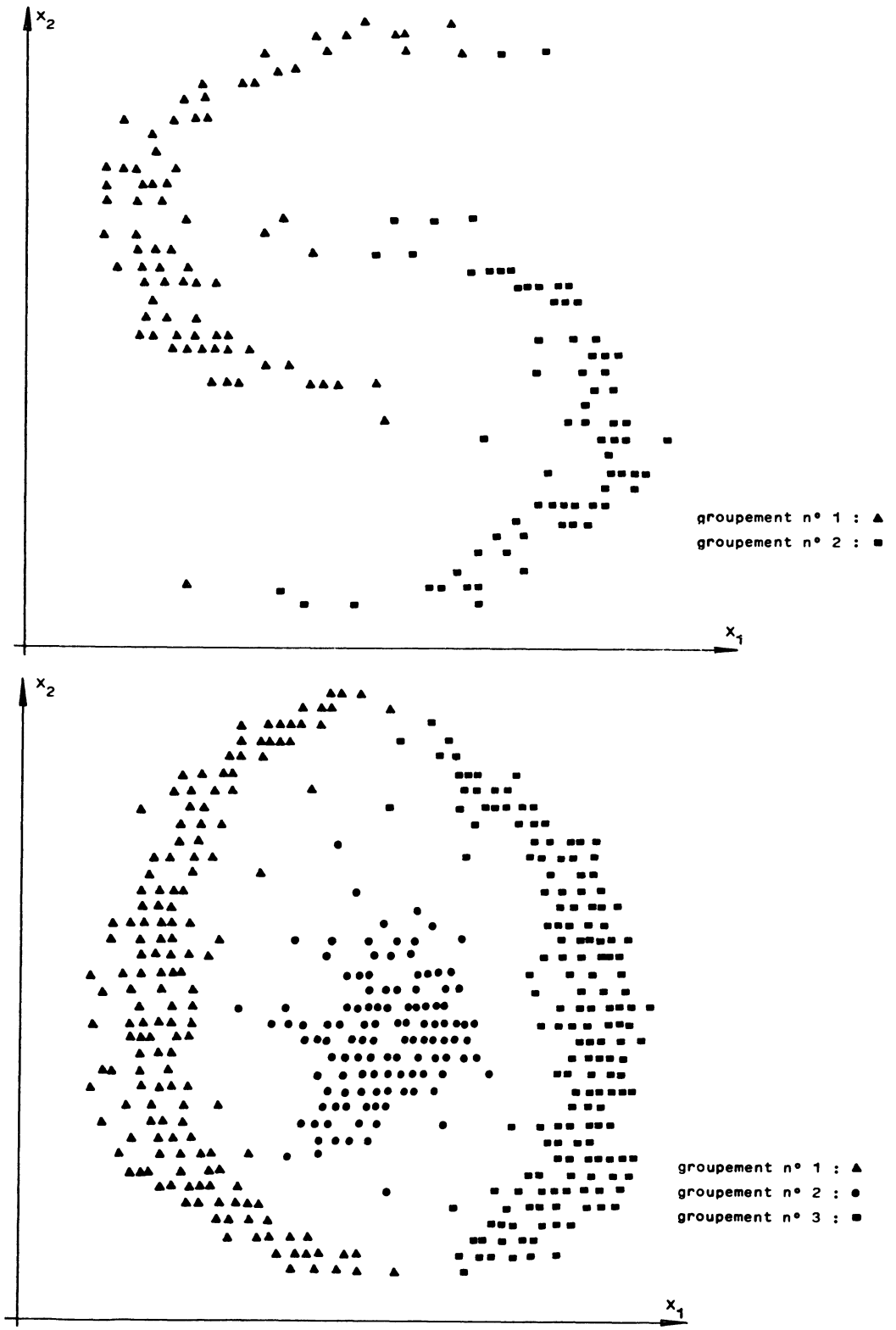


Figure 6. – Représentation graphique des échantillons utilisés pour illustrer la procédure de recherche des groupements.

7. CONCLUSION

Dans cet article de synthèse, nous avons montré comment l'utilisation du concept de convexité fournit un cadre général pour aborder l'aspect statistique de la classification automatique. La même approche mathématique permet de résoudre efficacement des problèmes d'analyse de mélanges aussi bien que des problèmes de détection des modes des fonctions de densité de probabilité.

Servant de base aux procédures de classification présentées, nous avons proposé une technique nonparamétrique qui permet, à partir des observations, de déterminer localement le sens de la convexité de la fonction de densité sous-jacente. L'opération d'intégration de cette fonction sur des domaines d'observation qui est à la base de cette approche, assure une grande robustesse aux opérateurs mis en jeu. Dans le cas très général où l'analyste ne dispose d'aucune autre information que celle qui peut être extraite de l'échantillon qui lui est soumis, cette technique fournit une description succinte et concise de la distribution des observations en termes de convexité.

Sans aucune hypothèse restrictive, l'analyse de la convexité des fonctions de densité multivariables permet de mettre en évidence les composantes d'un mélange gaussien et de déterminer, pour chacune d'elles, des valeurs approchées du vecteur moyenne, de la matrice de covariance et de la probabilité a priori. La construction de fonctions de décision sur la base de ces valeurs approchées permet ensuite d'optimiser le processus de classification.

L'utilisation de techniques non paramétriques pour l'analyse de la convexité des fonctions de densité a posé deux problèmes majeurs, d'ailleurs bien connus des utilisateurs. Le premier concerne les temps de calculs prohibitifs nécessaires à la mise en œuvre de ces techniques dès que la dimension des données dépasse quelques unités. Nous avons été conduits à proposer un algorithme d'estimation rapide qui, par delà son application immédiate dans le cadre de cette étude, peut contribuer à redonner un nouvel intérêt à ces techniques d'estimation nonparamétriques.

Le second problème rencontré est lié à la dégradation des performances de l'estimateur lorsque la taille de l'échantillon disponible vient à diminuer. Nous avons ainsi été amenés à proposer des variantes de la méthode générale, en substituant l'analyse des densités marginales de probabilité à celle des fonctions de densité multivariables elles-mêmes.

Afin de guider l'utilisateur potentiel, la figure 7 indique la méthode la plus appropriée pour traiter un problème en fonction des hypothèses acceptables, du nombre d'observations disponibles q et de la taille n de l'échantillon.

Le concept de convexité apporte donc un nouvel outil bien adapté aux problèmes de classification automatique, particulièrement lorsque l'on dispose de très peu d'information a priori sur la structure des données à analyser, la seule information disponible étant celle qui peut être extraite de l'ensemble des observations. Dans le cadre unificateur de l'analyse de la convexité des fonctions de densité, l'analyste dispose d'une panoplie de méthodes qui répondent aux exigences les plus strictes des problèmes de classification non supervisée. Il importait que ces diverses méthodes soient présentées simultanément en un seul article de synthèse.

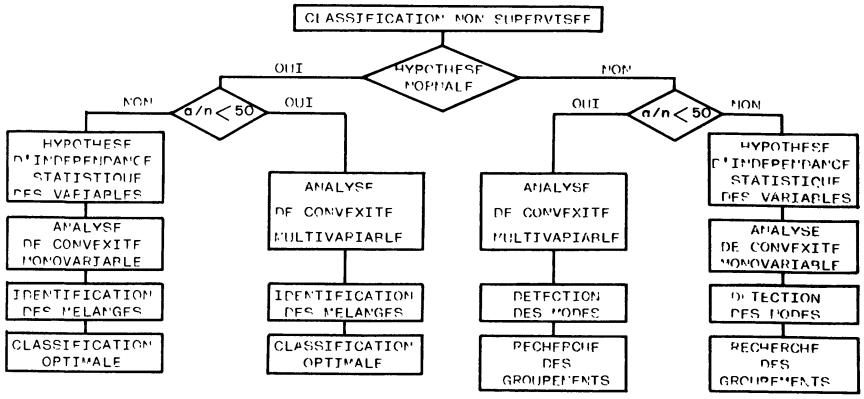


Figure 7. – Choix d'une méthode de classification en fonction des hypothèses d'application et des conditions d'utilisation. (La valeur limite du rapport q/n qui conditionne l'utilisation de l'analyse multivariable n'est donnée qu'à titre indicatif car elle dépend essentiellement de la structure des données).

REMERCIEMENTS

Les travaux présentés dans cet article constituent une partie d'une thèse d'état soutenue à l'Université de Lille 1 sous le titre "Optimisation du processus de classification automatique par analyse de la convexité des fonctions de densité de probabilité". Que le Professeur P. VIDAL, qui a dirigé cette thèse, trouve ici l'expression de nos plus sincères remerciements pour son aide attentive et efficace.

Nous sommes également reconnaissants envers le Professeur M. NAJIM qui a généreusement mis les moyens matériels de son Laboratoire à notre disposition.

REFERENCES

[1] G.H. BALL et D.J. HALL. (1967). – A clustering technique for summarizing multivariate data, *Behavorial Sc.*, Vol. 12, p. 153-155.

[2] E. DIDAY (1971). – Une nouvelle méthode en classification automatique et reconnaissance des formes : la méthode des nuées dynamiques, *Revue de Stat. Appl.* Vol. 19, n° 2, p. 20-33.

[3] K. FUKUNAGA et W.L.G. KOONTZ (1970). – A criterion and an algorithm for grouping data, *IEEE Trans. on Computers*, Vol.C-19, p. 917-923.

[4] G.N. LANCE et W.T. WILLIAMS (1967). – A general theory of classificatory sorting strategies 1 – Hierarchical systems, *Computer Jour.*, Vol. 9, p. 973-980.

[5] J. MAC QUEEN (1967). – Some methods for classification and analysis of multivariate observations, *Proc. 5th Berkeley Symp. on Math. Stat. and Prob.*, p. 281-297.

[6] C.T. ZAHN (1971). – Graph theoretical methods for detecting and describing gestalt clusters. *IEEE Trans. on Computers*, Vol.C-20, p. 68-86.

[7] J.L. FLEISS et J. ZUBIN (1969). – On the method and theory of clustering, *Multivariate Behavioral Research*, p. 235-250.

[8] I. GITMAN (1973). – An algorithm for nonsupervised pattern classification, *IEEE Trans.on Syst., Man & Cyb.*, Vol. SMC-3, p. 66-74.

[9] I. GITMAN et M.D LEVINE (1970). – An algorithm for detecting unimodal fuzy sets and its application as a clustering technique, *IEEE Trans. on Computers*, Vol.C-19, p. 583-593.

[10] R. MIZOGUCHI et M. SHIMURA (1976). – Non parametric learning without a teacher based on mode estimation, *IEEE Trans. on Computers*, Vol. C-25, p. 1109-1117.

[11] D. WISHART (1978). – Estimating the modes of a multivariate sample density, *Proc. Joint Meeting of the Classif. Soc. & British Pattern Recognition Ass. London.*

[12] B.G. BATCHELOR et B.R. WILKINS (1969). – Methods fo location of clusters of patterns to initialize a learning machine, *Electronics letters*, Vol. 5, n° 20, p. 481-483.

[13] K. FUKUNAGA et L.D. HOSTETLER (1975). – The estimation of the gradient of a density fonction with applications in pattern recognition, *IEEE Trans. on Inf. Theory*, Vol. II-21, n° 1, p. 32-40.

[14] J. KITTLER (1976). – A locally sensitive method for cluster analysis, *Pattern Recognition*, Vol. 8 p. 23-33.

[15] G.S. SEBESTYEN (1962). – Pattern recognition by an adaptive process of sample set construction, *IRE Trans. on Info. Theory*, Vol. II-8, p. 82-91.

[16] R.L. THORNDIKE (1953). – Who belong in the family? *Psychometrica*, Vol.18, p. 267-276.

[17] H.P. FRIEDMAN et J. ROBIN (1967) – On some invariant criteria for grouping data, *J. Amer. Stat. Ass.*, Vol. 62, p. 1159-1178.

[18] M.A. VOGEL et A.K.C. WONG (1979). – PFS clustering method, *IEEE Trans. on Pattern Anal. & Machine Intelligence*, Vol.PAMI-1, n° 3, p. 237-245.

[19] U.E. MAKOV et A.F.M. SMITH. – A quasi-Bayes Unsupervised learning procedure for priors. *IEEE Trans. Info. Theory*, Vol. II-23, n° 6, p. 761-764.

[20] R. MIZOGUCHI et M. SHIMURA (1975). – An approach to unsupervised learning classification, *IEEE Trans. on Computers*, Vol. C-24, n° 10, p. 979-983.

[21] R.F. DALY (1962). – *The adaptive binary detection problem on the real line*, Techn. Report 2003-3, Stanford Univ., Calif.

[22] C.G. HILLBORN et D.G. LAINIOTIS (1968). – Optimal unsupervised learning multicategory dependant hypotheses pattern recognition, *IEEE Trans. Info. Theory*, Vol. II-14, p. 468-470.

[23] A. SCHROEDER (1976). – Analyse d'un mélange de distributions de probabilité de même type, *Rev. Statist. App.*, Vol. 24, n° 1, p. 32-62.

[24] D. KAZAKOS (1977). – Recursive estimation of prior probabilities using a mixture, *IEEE Trans. Info. Theory*, Vol. II-23, n° 2, p. 203-211.

[25] N.E. DAY (1969). – Estimating the components of a mixture of normal distributions, *Biometrika*, Vol. 56, p. 463-474.

[26] J.H. WOLFE (1970). – Pattern clustering by multivariate mixture analysis, *Multiv. Behav. Res.*, Vol. 5, p. 329-350.

[27] D.B. COOPER et P.W. COOPER (1964). – Non supervised adaptive signal detection and pattern recognition, *Info. & Control*, Vol. 7, p. 416-444.

[28] P.W. COOPER (1967). – Some topics on nonsupervised adaptive detection for multivariate normal distributions, *Comp. & Info. Sc.*, Vol. II, p. 123-146, Academic Press, N.Y.

[29] C. VASSEUR et J.G. POSTAIRE (1980). – A convexity testing method for cluster analysis, *IEEE Trans. on Syst., Man & Cyb.*, Vol. SMC-10, n° 3, p. 145-149.

[30] J.G. POSTAIRE et C. VASSEUR (1982). – A fast algorithm for nonparametric probability density estimation, *IEEE Trans. on Pattern Anal. & Machine Intelligence*, Vol. PAMI-4, n° 6, p. 663-666.

[31] J.G. POSTAIRE et C. VASSEUR (1981) – An approximate solution to normal mixture identification with application to unsupervised pattern classification, *IEEE Trans. on Pattern Anal. & Machine Intelligence*, Vol. PAMI-3, n° 2, p. 163-179 (1981).

[32] J. G. POSTAIRE et M. LIMOURI (1978). – The convexity concept in cluster analysis, *4th Int. Joint conf. on Pattern Recog.*, Kyoto, Japon.

[33] J.G. POSTAIRE (1982). – An unsupervised Bayes classifier for normal patterns based on marginal densities analysis, *Pattern Recog.*, Vol. 15, n° 2, p. 103-111.

[34] J.G. POSTAIRE (1982). – Fonctions convexes et optimisation du processus de classification automatique : I. Optimisation par analyse de la convexité des fonctions de densité multivariables, *RAIRO/Automatique*, Vol. 16, n° 4, p. 357-379.

[35] D.J. EIGEN, F.R. FROMM & R.A. NORTHOUSE. – Cluster analysis based on dimensional information with application to feature selection and classification, *IEEE Trans. Syst., Man & Cyb.*, Vol. SMC-4, n° 3, p. 284-294.

[36] J.G. POSTAIRE (1983). – Fonctions convexes et optimisation du processus de classification automatique : II. Optimisation par analyse de la convexité des fonctions de densité marginales, *RAIRO/Automatique*, Vol. 17, n° 1, p. 39-59.

[37] C. VASSEUR et J.G. POSTAIRE (1979). – Convexité des fonctions de densité : application à la détection des modes en reconnaissance des formes, *RAIRO/Automatique*, Vol. 13, n° 2, p. 171-188.

[38] M. LIMOURI (1979). – Classification automatique non supervisée par analyse des densités marginales de probabilité, *Diplôme d'Etudes Sup.*, Univ. de Rabat, Maroc.