N° d'ordre: 406

50376
1987
5

THÈSE

présentée à

L'UNIVERSITE DES SCIENCES ET TECHNIQUES DE LILLE FLANDRES ARTOIS

pour obtenir le titre de

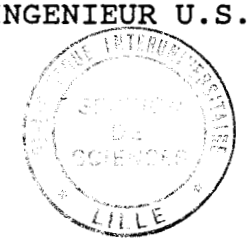
DOCTEUR INGENIEUR

Spécialité: AUTOMATIQUE

par

FIZAZI Hadria

INGENIEUR U.S.T.O



CLASSIFICATION AUTOMATIQUE DE PETITS ECHANTILLONS DE GRANDE DIMENSION. APPLICATION A LA BIOMETRIE DE L'ABEILLE.

Soutenue le 2 février 1987 devant la Commission d'Examen:

MM.	P. VIDAL	Président
	J.G. POSTAIRE	Rapporteur
	C. LANGRAND	Examinateur
	J. LOUVEAUX	Examinateur
	M. STAROSWIECKI	Examinateur

A mes Parents

Je sais ce que je fus, je sais ce que je suis,
je veux ce que je dois, je fais ce que je puis.
(BOUSCAL).

Il ne faut point juger des hommes par ce
qu'ils ignorent, mais par ce qu'ils savent,
et par la manière dont ils le savent.
(VAUVENARGUES).

AVANT PROPOS

Le travail présenté dans ce mémoire a été effectué au Centre d'automatique de l'Université des Sciences et Techniques de Lille Flandres Artois.

Je remercie vivement Monsieur le Professeur Pierre VIDAL pour son acceuil au sein de son laboratoire. Qu'il trouve ici l'expression de ma gratitude pour l'honneur qu'il me fait en assurant la présidence du jury.

J'adresse ma plus vive reconnaissance à Monsieur Jack Gérard POSTAIRE, Professeur à l'USTL Flandres Artois, grâce à qui, de par son dynamisme et son soutien constant, ce présent travail a pu aboutir. Ne m'ayant jamais épargné de son temps et ses connaissances, qu'il trouve ici mes plus vifs remerciements.

Que Monsieur Claude LANGRAND, Professeur à l'USTL Flandres Artois, veuille bien trouver ici l'expression de ma sincère reconnaissance pour l'honneur qu'il me fait en acceptant de participer à ce jury.

Je suis également très honorée par la présence de Monsieur Jean LOUVEAUX, Directeur de Recherche à l'INRA-CNRS. J'exprime ma profonde gratitude pour l'intérêt qu'il a bien voulu porter à ces travaux en acceptant de participer à ce jury.

Mes remerciements vont également à Monsieur Marcel STAROSVIECKI, Professeur à l'EUDIL. Sa présence dans ce jury m'honore grandement.

Je tiens à remercier le Gouvernement Algérien pour m'avoir aidé financièrement durant tout mon séjour en France, en m'accordant une bourse d'étude.

Ma reconnaissance va de même à tous ceux qui, de loin ou de près, par leur compétence ou leur amitié m'ont soutenu dans l'élaboration de ce travail.

SOMMAIRE

CHAPITRE I: IDENTIFICATION DES MELANGES ET CLASSIFICATION.

CHAPITRE II: IDENTIFICATION DES MELANGES PAR ANALYSE DE CONVEXITE DES FONCTIONS DE DENSITE.

CHAPITRE III: RESTAURATION DES PROPRIETES DE CONVEXITE PAR FILTRAGE NON LINEAIRE.

CHAPITRE IV: CLASSIFICATION OPTIMALE DES PETITS ECHANTILLONS.

CHAPITRE V: FUSIONNEMENT DES GROUPEMENTS D'OBSERVATIONS.

CHAPITRE VI: CLASSIFICATION DES ABEILLES

NOTATIONS UTILISEES

n	: Dimension de l'espace de représentation.
Q	: Taille de l'échantillon.
X_q	: Observation.
$p(x_i)$	: Fonction de densité marginale de probabilité (f.d.m.p.).
$\hat{p}(x_i)$	: Estimation de $p(x_i)$.
h_Q	: Largeur de la fenêtre de Parzen.
$\varphi(u)$	: Fonction réelle définie sur R appelée noyau.
k_Q	: Nombre de plus proches voisins (p.p.v.).
x_0	: Point où on estime la f.d.m.p.
$h(x_0)$	: Largeur de la fenêtre centrée en x_0 et qui englobe les k_Q p.p.v.
a_n	: Coefficient de normalisation.
x_{imax}	: La plus grande valeur sur l'axe i.
x_{imin}	: La plus petite valeur sur l'axe i.
M	: Nombre d'intervalles égaux et adjacents.
$\bar{x}_i$	: Moyenne de la distribution normale.
v_i	: Variance de la distribution normale.
σ_i	: Ecart type de la distribution normale.
x'_i	: Coordonnées normalisées des observations.
$\mathcal{X}=\{X_q\}$	: Ensemble des observations disponibles distribuées normalement.
$\bar{X}_i$	: Vecteur moyenne.
$\sum$	: Matrice de covariance.
$\hat{p}^*(x_i)$	: Forme discrète de $\hat{p}(x_i)$.
$p^*(x_i)$	: Forme discrète de $p(x_i)$.
$\{\hat{p}_m(x_i)\}$	: Ensemble des valeurs qui représentent $\hat{p}^*(x_i)$.
$\{p_m(x_i)\}$	: Ensemble des valeurs qui représentent $p^*(x_i)$.
E_i	: Erreur quadratique moyenne.
h^*_Q	: Valeur optimale de la largeur de fenêtre h_Q.
k^*_Q	: Valeur optimale du nombre de p.p.v.
E^*_i	: Erreur quadratique minimale suivant l'axe i.
d_i	: Segment de convexité de la fonction $p(x_i)$
$\hat{\bar{X}}$	: Valeur approchée du vecteur moyenne X.
$\hat{p}_m(x_i)$	: Version discrète de l'estimation $p(x_i)$.
l	: Largeur de la fenêtre de l'opérateur de filtrage non linéaire.

$\hat{p}'_m(x_i)$: Fonction de densité marginale de probabilité estimée-filtrée.

$E^{*'}_i$: Minimum des erreurs quadratiques après filtrage.

$g_k(X)$: Fonction de décision.

$P(C_k)$: Probabilité de la classe C_k.

$\bar{X}_k$: Vecteur moyenne de la classe C_k.

$\sum_k$: Matrice de covariance de la classe C_k.

R_k : Région de décision.

h^*_i : Valeur optimale de h_Q suivant l'axe i.

k^*_i : Valeur optimale de k_Q suivant l'axe i.

G_m : Groupement d'observations.

$g_{i,m}$: Numéro d'ordre du groupement d'observations G_m suivant l'axe i.

$E_{i,m}$: Erreur quadratique moyenne suivant l'axe i entre la f.d.m.p et le modèle analytique associée au groupement d'observations G_m.

$G^*_m(i)$: Groupement d'observations qui se rapproche le plus d'une distribution normale.

$E^*_{i,m}$: Le minimumn minimorum des $E_{i,m}$.

$G^*_{m*}(i)$: Groupement d'observations possédant un minimum minimorum $E^*_{i,m}$.

$h^*_{m*}(i)$: Largeur de la fenêtre de Parzen associée au groupement $G^*_{m*}(i)$ pour laquelle on obtient $E^*_{i,m}$.

$k^*_{Q,m*}(i)$: Nombre de p.p.v. associés à $G^*_{m*}(i)$ pour lequel on obtient $E^*_{i,m}$.

$Q(G_m)$: Taille du groupement d'observations G_m.

$h_m(i)$: Largeur de la fenêtre de Parzen associée au groupement G_m suivant l'axe i .

$k_{Q,m}(i)$: Nombre de p.p.v. associés au groupement G_m suivant l'axe i.

$E'_{i,m}$: Erreur quadratique calculée à partir du paramètre $h_m(i)$ ou bien $k_{Q,m}(i)$ associée au groupement G_m.

$E^g_{i,m}$: Erreur quadratique située sur la partie gauche du maximum de la f.d.m.p. associée au groupement G_m.

$E^d_{i,m}$: Erreur quadratique située sur la partie droite du maximum de la f.d.m.p. associée au groupement G_m.

----------//----------

CHAPITRE I

IDENTIFICATION DES MELANGES ET CLASSIFICATION

CHAPITRE I

IDENTIFICATION DES MELANGES ET CLASSIFICATION

I - 1 INTRODUCTION.

L'histoire montre comment la classification a longtemps été au service des sciences de la nature. La classification est considérée comme un processus de connaissance objectif. Elle doit permettre, sur la seule base des faits observables, de découvrir l'ordre du réel caché derrière la prolifération des détails.

L'objet principal de toute classification est de définir des groupes (ou classes) à partir d'un ensemble d'individus dont la structure est inconnue a priori. Le but de la classification est de condenser l'essentiel des informations multiples observées sur un lot d'individus. La description complexe et détaillée de chaque individu est remplacée par son appartenance à une classe bien définie. Le cerveau humain étant incapable d'accomplir une synthèse multidimensionnelle, l'automatisation de la classification constitue un pas décisif dans le processus d'analyse et de compréhension des données.

La classification automatique permet, dans un domaine inconnu, de découvrir dans les phénomènes étudiés des structures qui n'étaient pas directement visibles sur les données. Elle apparait ainsi comme une méthode exploratoire, créatrice d'hypothèses. D'autre part, elle permet de retrouver, en les précisant, des structures que l'on soupçonnait déjà.

La classification automatique apparait souvent comme un arsenal de méthodes qui n'ont en commun que leur finalité et qui font appel à des notions mathématiques et des concepts scientifiques très divers. Souvent, l'utilisateur reste perplexe devant le foisonnement des techniques utilisables pour identifier les classes en présence dans un ensemble de données multidimensionnelles.

Selon le nombre d'observations et la dimension des données, selon la possibilité d'utiliser un modèle et compte tenu des informations dont il dispose, l'analyste choisit une stratégie parmi un ensemble de méthodes très diverses et souvent peu comparables, chacune ayant ses avantages et ses inconvénients.

Ce manque de cohérence 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 pas le nombre de classes en présence, ni la fonction de densité, ni la probabilité a priori attachées à chacune d'elles.

Il est toutefois possible d'envisager une optimisation du classement en compensant le manque de connaissances sur le mélange par les informations apportées par les individus à classer eux-mêmes.

En supposant que les fonctions de densité des différentes classes appartiennent à un ensemble de fonctions représentables par quelques paramètres (fonctions normales, de Bernoulli, ..., etc), le problème de l'optimisation du processus de classification se trouve posé en termes d'analyse du mélanges.

I- 2 LIMITATIONS DES METHODES D'ANALYSE DES MELANGES.

Le problème de l'optimisation du processus de classification peut se ramener simplement au problème de l'estimation des paramètres d'un mélange.

L'analyse des mélanges a été abordé par de nombreuses méthodes. Historiquement, le premier travail sur la question remonte à 1894, lorsque K. Pearson utilise les moments de la distribution des observations pour déterminer les paramètres d'un mélange de deux densités normales et monovariables [PEA.94].

Buchanan-Wollaston et Hodgesson [BUC.29] proposent une méthode graphique consistant à ajuster une loi normale sur chaque pic de l'histogramme expérimental représentant la distribution des données. On retrouve la même approche beaucoup plus tard chez Bhattacharya [BAT.67].

Doetsch [DOE.36] utilise la transformée de Fourier pour décomposer les mélanges gaussiens monovariables en supposant connue la valeur de la fonction de densité en tout point.

Dans le cadre de l'analyse des mélanges unidimensionnels, il convient de citer encore Rao [RAO.48] qui traite le cas des mélanges à deux composantes et, plus récemment, Benzecri [BEN.72] et Cazes [CAZ.76] qui utilisent une série de déconvolutions successives. Si l'on peut aborder avec succès l'analyse des mélanges monovariables, il semble que peu de solutions entièrement satisfaisantes n'aient été apportées à l'analyse des mélanges multivariables.

En effet, certaines techniques ne sont applicables que sous des hypothèses restrictives [DAY.69], [WOL.70]. D'autres nécessitent des informations a priori impossibles à fournir lors de l'analyse d'un échantillon inconnu [HIL.68], [SCH.76], [KAZ.77].

De plus, les rares procédures qui ne sont pas limitées dans ce sens nécessitent une telle somme de calculs que leur champ d'application se trouve considérablement restreint.

Même si on se limite au cas d'observations distribuées normalement, une seule technique permet de déterminer toutes les caractéristiques d'un mélange gaussien à partir des seules informations apportées par l'ensemble des individus à classer. Seule l'approche par analyse de la convexité des fonctions de densité de probabilité permet, en ayant pour entrée les observations, de fournir en sortie le nombre de classes présentes et, pour chacune d'elles, le vecteur moyenne, la matrice de covariance et la probabilité a priori [POS.83a].

I - 3 CONVEXITE ET OPTIMISATION DU PROCESSUS DE CLASSIFICATION.

La classification peut être optimisée en n'utilisant aucune autre information que celle apportée par les observations disponibles avec pour unique hypothèse de travail, la normalité de la distribution des individus de chaque classe. Cette hypothèse est généralement admise lorsqu'on dispose de peu d'information a priori, mais de beaucoup d'observations.

Pour obtenir une description complète des mélanges gaussiens, on utilise une procédure permettant de déterminer localement la convexité de toute fonction multivariable continue. En faisant appel à des techniques d'estimation non paramétrique, il est possible de déterminer la convexité locale d'une fonction de densité multivariable quelconque à partir des observations. Cette procédure est exploitée pour analyser les mélanges gaussiens en déterminant, à partir des observations, les domaines où la fonction de densité est convexe [POS.82b] [POS.83a].

L'optimisation de la classification revient à déterminer, en analysant ces domaines, les paramètres du mélange à savoir:

1 / Le nombre de composantes présentes dans le mélange.

2 / Le vecteur moyenne de chaque composante.

3 / La matrice de covariance de chaque composante.

4 / La probabilité a priori de chaque composante.

La méthode repose sur un test qui s'apparente aux techniques d'estimations non paramétriques. Pour toutes ces techniques, le nombre d'observations nécessaires pour obtenir des résultats satisfaisants est d'autant plus important que la dimension du problème est élevée. Dans la pratique, nous pouvons être confrontés à des problèmes de forte dimension pour lesquels la taille réduite des échantillons disponibles ne permet d'obtenir ni une estimation correcte de la fonction de densité sous-jacente ni, a fortiori, une information précise sur sa convexité.

I- 4 L'OPTIMISATION DE LA CLASSIFICATION DES PETITS ECHANTILLONS.

En émettant l'hypothèse d'indépendance statistique des caractères mesurés sur les individus, il est possible de déterminer les paramètres du mélange à partir de l'analyse de la convexité des fonctions de densité marginale monovariables. Pratiquemment, ces fonctions peuvent être estimées avec un nombre restreint d'observations, soit par la méthode du noyau [ROS.56], [PAR.62], soit par la méthode des plus proches voisins [COV.67], [LOF.65], et ce, quelque soit la dimension des données.

Toutefois, l'expérience montre qu'il est extrèmement délicat d'ajuster la taille du noyau si on utilise l'estimateur de Parzen-Rosenblatt, ou le nombre de voisins à prendre en compte si on utilise la méthode des plus proches voisins. Les résultats obtenus sont d'autant plus sensibles à l'ajustement des paramètres des estimateurs que le nombre d'observations à classer est réduit. Une mesure de l'adéquation entre une fonction de densité de probabilité estimée et son modèle analytique est proposée afin de quantifier l'effet des différents paramètres de réglage des estimateurs (chapitre II).

Généralement les fonctions de densité marginale sont bruitées, c'est à dire que les contributions individuelles de chaque observation masquent les véritables modes de ces fonctions. Pour pallier cet inconvénient, on propose dans le chapitre III une méthode de filtrage non-linéaire qui permet de restaurer l'estimateur dans le cas où celui-ci est trop bruité pour être directement exploité.

Cette procédure, qui élimine les petites variations locales non significatives tout en préservant les modes effectifs des fonctions de densité, a été testée sur des données multidimensionnelles normales. En comparant l'adéquation des estimations brutes et des estimations filtrées au modèle analytique gaussien sous-jacent aux distributions analysées, on met en évidence l'amélioration de la qualité des fonctions de densité marginale estimées après filtrage ainsi que la robustesse de la procédure par rapport aux réglages des paramètres.

D'autre part, le filtre proposé permet de restaurer les propriétés de convexité des fonctions de densité marginale estimées. On peut ainsi aborder, avec succès, l'identification des mélanges gaussiens par analyse de la convexité de leurs fonctions de densité marginale, même pour des petits échantillons et ce, quelle que soit la dimension des observations. La procédure débouche alors sur des techniques de classification automatique optimales au sens de la théorie de décision.

Le fait de travailler axe par axe engendre certaines difficultés, les contributions des classes en présence dans le mélange pouvant ne pas s'individualiser au niveau des fonctions de densité marginale.

Pour remédier à de telles difficultés, une analyse locale permet d'affiner la classification en séparant en plusieurs classes, le cas échéant, des groupements mis en évidence par une analyse globale (chapitre IV).

Il peut également arriver que l'analyse globale scinde en plusieurs groupements les observations provenant d'une seule classe. En calculant l'adéquation de la fonction de densité marginale de chaque groupement mis en évidence à un modèle gaussien, il est possible de valider les composantes détectées sur les densités marginales lorsqu'elles correspondent effectivement à des classes réelles. Grâce à cette analyse, on peut fusionner des groupements tronqués pour reconstituer les classes véritables (chapitre V).

Dans le chapitre VI, on applique ces techniques d'analyse de convexité, de filtrage non-linéaire et de séparation/fusionnement des groupements à une population d'abeilles provenant de deux îles de la GUADELOUPE. Cet exemple, qui s'inscrit dans le cadre d'études apidologiques, met en évidence l'intérêt de l'approche proposée pour la classification de petits échantillons.

CHAPITRE II

IDENTIFICATION D'UNE DISTRIBUTION NORMALE PAR ANALYSE DE LA CONVEXITE DE SES FONCTIONS DE DENSITE.

CHAPITRE II

IDENTIFICATION D'UNE DISTRIBUTION NORMALE PAR ANALYSE DE LA CONVEXITE DE SES FONCTIONS DE DENSITE MARGINALE.

II-1 INTRODUCTION

L'identification d'un mélange gaussien par analyse de la convexité des fonctions de densité marginale repose sur des techniques d'estimation non paramétrique de ces fonctions à partir des observations disponibles.

Nous abordons, dans ce chapitre, l'étude critique de ces méthodes en les illustrant, pour la clarté de l'exposé, par les problèmes liés à l'identification d'une distribution normale multidimensionnelle unique.

II-2 ESTIMATION DES FONCTIONS DE DENSITE MARGINALE PAR LA METHODE DU NOYAU

II- 2- 1 Principe d'estimation.

Lorsque les individus à classer peuvent être caractérisés par n attributs $x_1, x_2, \ldots, x_n$, on peut leur associer une représentation vectorielle:

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

dans un espace euclidien à n dimensions.

Supposons que l'on dispose de Q observations $X_q = [x_{1,q}, \ldots, x_{i,q}, \ldots, x_{n,q}]^T$ pour estimer les n fonctions de densité marginale $p(x_i)$, i=1,2,...,n de la distribution de ces observations. La méthode du noyau donne une estimation $\hat{p}(x_i)$ de $p(x_i)$ en tout point x_0 sous la forme:

$$\hat{p}(x_i)|x_0 = \frac{1}{Q} \sum_{q=1}^{Q} \frac{1}{h_Q} \varphi\left[\frac{x_{i,q}-x_0}{h_Q}\right]$$

où le paramètre h_Q est appelé largeur de la fenêtre d'estimation et $\varphi(u)$ est une fonction réelle définie sur R, appelée "noyau" de la fenêtre qui doit satisfaire les relations suivantes:

1 / $\varphi(u) > 0$

2 / $\int_{-\infty}^{+\infty} \varphi(u).\, du = 1$

3 / limite $\|u\|\varphi(u) = 0$
$\|u\| \to \infty$

afin d'assurer, en dehors des contraintes sur la largeur h_Q de la fenêtre, la convergence de l'estimateur $\hat{p}(x_i)$ vers $p(x_i)$.

Il existe différents types de fonctions $\varphi(u)$ répondant à ces exigences. On peut citer les plus utilisées [RYZ.69], [BAN.77].

1 / Le noyau gaussien:

$$\varphi(u) = (1/\sqrt{2\pi}) \exp\{ - (1/2)\, u^2\}$$

2 / Le noyau triangulaire:

$\varphi(u) = 1 - |u|$ si $|u| \leq 1$

$\varphi(u) = 0$ autrement.

3 / Le noyau cubique:

$$\begin{cases} \varphi(u) = 1 & \text{si} \quad |u| \leq 1/2 \\ \varphi(u) = 0 & \text{autrement.} \end{cases}$$

Le noyau cubique est l'un des plus utilisés dans la pratique. Ce choix est généralement motivé par des considérations de programmation sur calculateur numérique car l'estimateur de $p(x_i)$ est alors obtenu d'une façon très simple.

La qualité de cet estimateur dépend de la largeur h_Q de la fenêtre qui doit être ajustée en fonction du nombre Q d'observations disponibles, afin d'assurer la convergence de l'estimateur.

Plus précisément, on doit choisir h_Q comme fonction de Q telle que [ROS.56], [WEG.72], [RYZ.69]:

1 / $\lim_{Q \to \infty} h_Q = 0$

2 / $\lim_{Q \to \infty} Q.h_Q \to \infty$

La détermination du paramètre h_Q est décrite dans le paragraphe suivant.

II- 2- 2 Choix de la largeur h_Q de la fenêtre.

La largeur h_Q de la fenêtre est un paramètre important qui influence la qualité de l'estimation. Ce paramètre doit être ajusté en fonction du nombre Q d'observations disponibles pour assurer la convergence de l'estimateur. En prenant le paramètre h_Q de la forme:

$$h_Q = h / \sqrt{Q}$$

$$h_Q = h / \text{Log } Q$$

l'estimateur $\hat{p}(x_i)$ (i=1,2, ..., n) tend asymptotiquement vers $p(x_i)$ lorsque la taille Q de l'échantillon disponible augmente indéfiniment.

Il est cependant bien connu des praticiens que si h est trop petit, l'estimation présente trop de variabilité et que s'il est trop grand, l'estimation trop lissée, manque de résolution [DUD.73].

La seule certitude dont dispose l'analyste et que, quelque soit le choix de h, l'estimateur convergera asymptotiquement vers la valeur de la fonction de densité si Q augmente indéfiniment. Pour un échantillon de taille finie, il faut chercher un compromis entre ces deux extrêmes et trouver une valeur permettant une estimation satisfaisante de $p(x_i)$.

II- 3 ESTIMATION DES FONCTIONS DE DENSITE MARGINALE PAR LA METHODE DES K_Q PLUS PROCHES VOISINS.

II- 3- 1 Principe d'estimation.

Une alternative à la méthode du noyau est la méthode des "k_Q plus proches voisins". Au lieu de fixer la taille de la fenêtre d'estimation en fonction du nombre Q d'observations disponibles, on cherche la fenêtre qui englobe un nombre prédéterminé d'observations dans le voisinage du point d'estimation [MAG.71].

Dans la littérature, on peut trouver plusieurs algorithmes de recherche des k_Q plus proches voisins [FUK.73], [FRE.75], [YUN.76], [BON.85], qui ont des performances comparables. La méthode des k_Q plus proches voisins consiste:

1 / à se fixer le paramètre k_Q.

2 / à faire croître un domaine D centré en x_0 (point où l'on estime la fonction de densité de probabilité) jusqu'à ce que celui-ci contienne les k_Q plus proches voisins de x_0 [COV.67].

Plus précisément, si on désire estimer $p(x_i)$ au point x_0, on calcule la quantité:

$$\hat{p}(x_i)|x_0 = \frac{k_Q / Q}{h\,[x_0]}$$

où $h(x_0)$ est la largeur de la plus petite fenêtre centrée en x_0, qui englobe les k_Q plus proches voisins du point x_0.

Le paramètre k_Q est ajusté en fonction du nombre Q d'observations disponibles.

II- 3- 2 Choix du paramètre k_Q.

Le choix du paramètre k_Q est le problème le plus délicat pour l'utilisation de l'algorithme des k_Q plus proches voisins. Ce paramètre dépend de plusieurs facteurs dont l'influence est difficile à apprécier: taille, caractéristiques de l'échantillon à analyser, nombre de classes, ..., etc.

Pour assurer la convergence de cet estimateur, il faut ajuster le paramètre k_Q en fonction du nombre Q d'observations disponibles. On peut prendre:

$$k_Q = k / \sqrt{Q}$$

$$k_Q = k / \text{Log } Q$$

Ces fonctions assurent la convergence de l'estimateur lorsque Q tend vers l'infini, quelque soit le choix de k [HIL.68]. Mais pour des échantillons de taille Q finie, l'ajustement de k est délicat. En effet, si k est trop petit, l'estimateur est très sensible aux fluctuations de la distribution des observations; par contre, si k est trop grand, l'estimateur ne reflète pas assez finement la fonction de densité de probabilité.

Un compromis entre ces deux extrêmes doit, comme dans la méthode précédente, être recherché.

II- 4 NORMALISATION ET DISCRETISATION DE L'ESPACE DE REPRESENTATION.

Il est généralement conseillé, avant d'envisager tout traitement des données, de normaliser leur espace de représentation [MEI.72]. L'espace normalisé est ensuite discrétisé en divisant chaque axe en intervalles égaux et adjacents.

Pour normaliser l'espace de représentation des données, on définit un coefficient de normalisation a_i pour chaque attribut i tel que:

$$a_i = x_{imax} - x_{imin} \qquad i=1,2,\ldots,n.$$

x_{imax} est la plus grande valeur du i^{eme} caractère.
x_{imin} est la plus petite valeur de $i^{ème}$ caractère.

Les nouvelles coordonnées des observations deviennent alors:

$$x_i' = \frac{x_i - x_{imin}}{a_i}$$

Après cette transformation, les coordonnées des observations vérifient la condition:

$$0 \leq x_i' \leq 1 \qquad i=1,2,\ldots,n.$$

Dans l'espace normalisé nous discrétisons les fonctions de densité marginale de probabilité en divisant chaque plage de variation d'un attribut de 0 à 1 en M intervalles égaux et adjacents.

II- 5 DETERMINATION DE LA QUALITE DE L'ESTIMATION.

Pour apprécier l'effet des différents paramètres de réglage des estimateurs, il importe de pouvoir déterminer la qualité de l'estimation. On propose de quantifier cette qualité en déterminant l'écart quadratique entre la fonction de densité marginale estimée et le modèle analytique correspondant à la distribution des observations.

Le modèle analytique d'une distribution normale est caractérisé par sa moyenne $\bar{x}_i$ et sa variance V_i.

Soit $\mathcal{X} = \{X_q\}, \quad q=1, 2, 3, \ldots, Q$

avec $X_q = [x_{1,q}, \; x_{2,q}, \; \ldots, \; x_{n,q}]^T$, l'ensemble des observations disponibles distribuées normalement à partir desquelles on désire estimer la fonction de densité marginale de probabilité $p(x_i)$.

Si on note $\bar{X} = [\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_n]^T$ le vecteur moyenne et

$$\sum = \begin{vmatrix} \sigma_1 & & 0 \\ & \sigma_2 & \\ 0 & & \sigma_n \end{vmatrix}$$

la matrice de covariance de la distribution des observations X_q, la fonction de densité marginale $p(x_i)$ a la forme analytique:

$$p(x_i) = \frac{1}{\sigma_i \sqrt{2\pi}} \exp\{ - \frac{1}{2} \left[\frac{(x - \bar{x}_i)}{\sigma_i} \right]^2 \}$$

Supposons que l'on estime, à partir de l'ensemble $\mathcal{H}$ la fonction $p(x_i)$ en M points équidistants le long de l'axe associé à l'attribut x_i de l'espace de représentation normalisé. Soit

$$\{\hat{p}_m(x_i)\} \qquad m= 1, 2,..., M$$

l'ensemble des valeurs qui représente ainsi la forme discrète $\hat{p}^*(x_i)$ de $\hat{p}(x_i)$.

On peut, de la même manière, définir la forme discrète $p^*(x_i)$ de $p(x_i)$: $\{p_m(x_i)\}$ m=1, 2,...,M.

Dans ces conditions, l'erreur quadratique moyenne entre la fonction estimée et son modèle analytique prend la forme:

$$E_i = \frac{1}{M} \sum_{m=1}^{M} \{p_m(x_i) - \hat{p}_m(x_i)\}^2$$

II- 6 MISE EN OEUVRE DES METHODES D'ESTIMATION ETUDIEES

Afin de tester les méthodes d'estimation étudiées aux paragraphes II-2 et II-3, nous générons artificiellement un échantillon de taille réduite comprenant trente observations bidimensionnelles distribuées suivant une loi normale. On se place sous l'hypothèse d'indépendance statistique des composantes des observations pour traiter cet exemple.

L'échantillon considéré est caractérisé par un vecteur moyenne $\bar{X}$ de valeur:

$$\bar{X} = [6, 10]^T$$

et par une matrice de covariance diagonale:

$$\sum = \begin{vmatrix} 1.2 & 0 \\ 0 & 0.5 \end{vmatrix}$$

La représentation graphique des observations de cet échantillon est donnée par la fig.1.

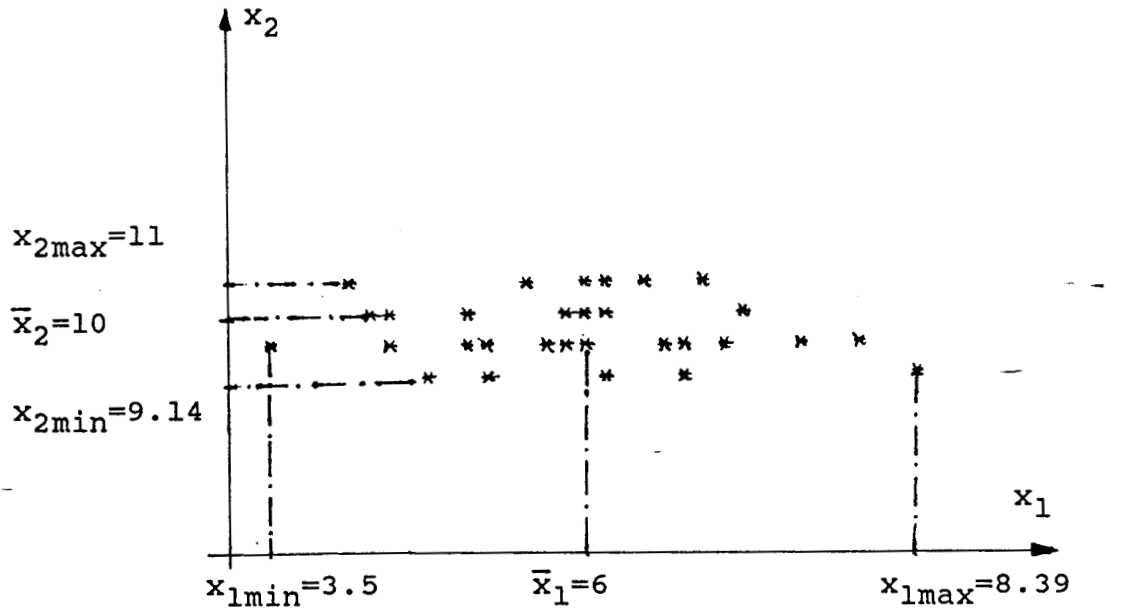


Figure 1: Représentation graphique d'un échantillon à 30 observations.

II- 6- 1 Résultats obtenus par la méthode d'estimation du noyau.

Nous estimons, par la méthode du noyau les fonctions de densité marginale de probabilité obtenues à partir de l'échantillon défini précédemment et ceci, suivant chaque attribut.

Pour chacune des fonctions de densité marginale estimées, on détermine les variations de l'erreur quadratique moyenne E_i i=1, 2 entre la fonction de densité marginale de probabilité estimée $\hat{p}(x_i)$ et son modèle analytique $p(x_i)$ en fonction de l'ajustement de la largeur h_Q de la fenêtre de Parzen. Les résultats obtenus sont consignés dans le tableau 1.

h_Q	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.50	0.60	0.80
E_1	0.143	0.030	0.025	0.016	0.008	0.012	0.017	0.030	0.054	0.125
E_2	0.460	0.451	0.434	0.222	0.157	0.186	0.235	0.238	0.292	0.495

Tableau 1

Ces résultats sont illustrés par la figure 2 qui représente les variations de l'erreur quadratique en fonction de la largeur h_Q.

Les courbes en trait continu et en trait discontinu indiquent respectivement les variations de E_1 et de E_2.

On constate ainsi que la variation de l'erreur quadratique moyenne E_i i=1, 2, en fonction de la largeur h_Q de la fenêtre, fait apparaître les valeurs optimales h^*_Q de h_Q pour lesquelles on obtient la meilleure adéquation entre la courbe estimée et son modèle analytique. Dans cet exemple particulier on a:

pour i=1 $\quad h^*_Q = 0.3$

pour i=2 $\quad h^*_Q = 0.3$

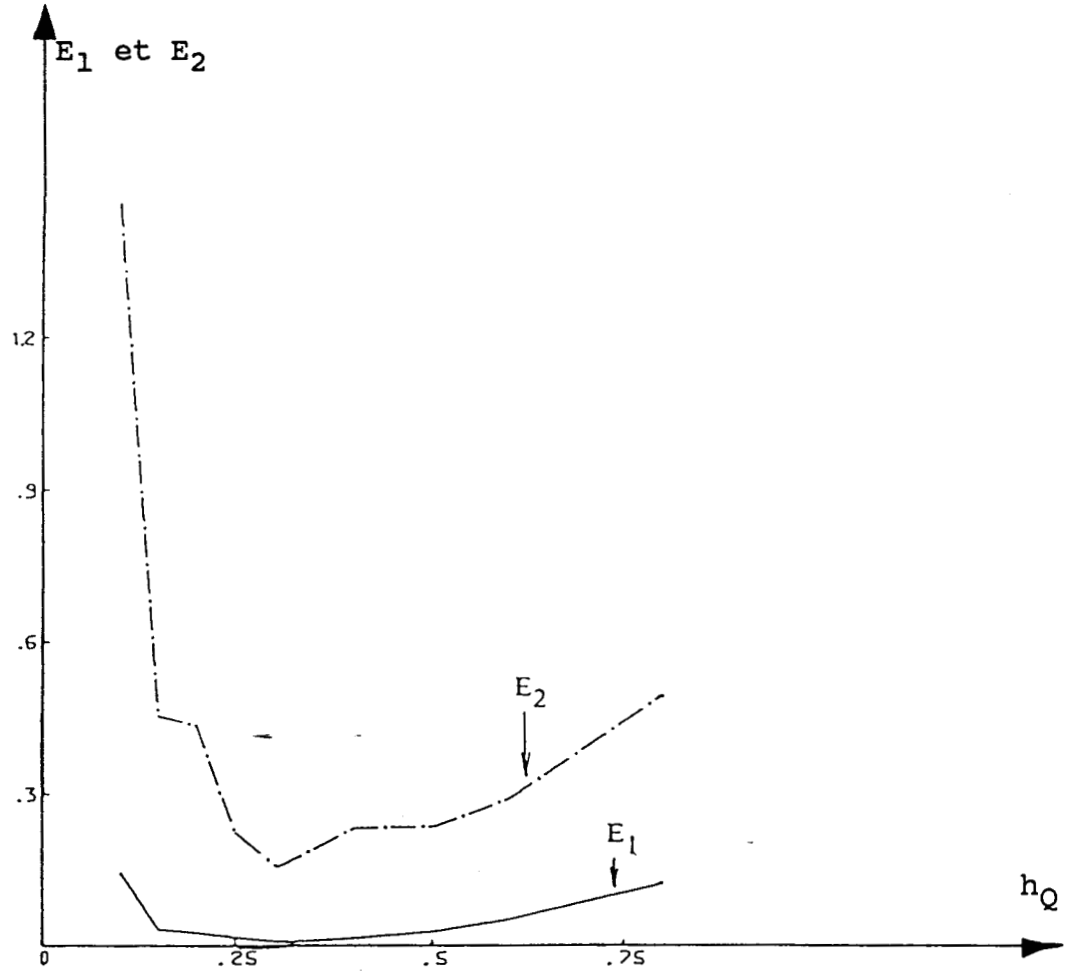


Figure 2: Variations des erreurs quadratiques en fonction de la largeur h_Q de la fenêtre de Parzen.

En prenant comme largeur h_Q de la fenêtre, la valeur obtenue par cet ajustement optimal des paramètres, on obtient des erreurs quadratiques E_i, telles que:

$$E^*_1 = 0.008$$
$$E^*_2 = 0.157$$

La figure 3 illustre les résultats de l'estimation de ces fonctions de densité marginale de probabilité pour cet ajustement optimal.

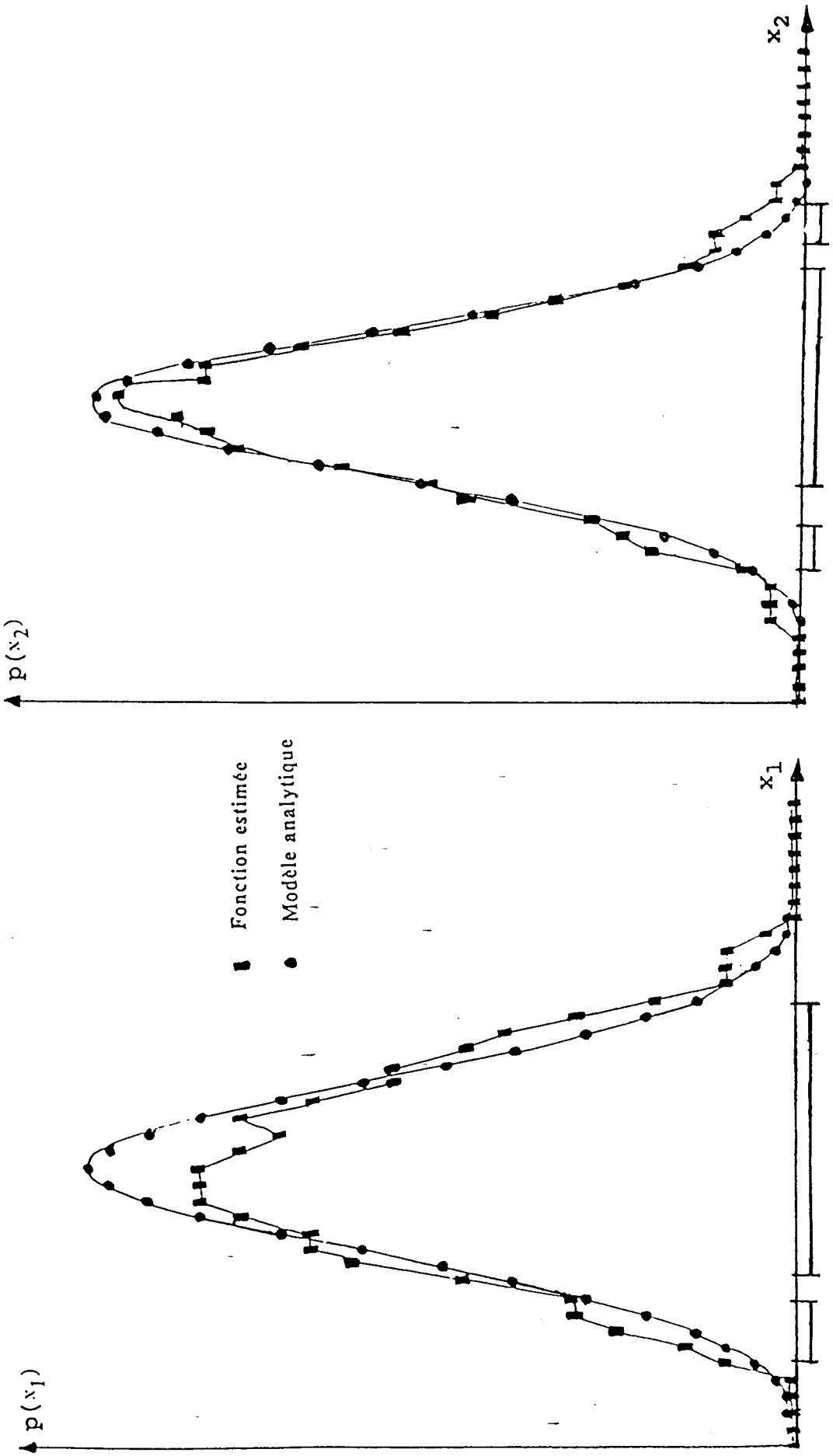


Figure 3 : Fonctions de densité estimées par la méthode du noyau et leurs modèles analytiques

II- 6- 2 Résultats obtenus par la méthode d'estimation des plus proches voisins.

On applique la méthode d'estimation des plus proches voisins sur l'échantillon représenté par la figure 1 afin d'estimer les fonctions de densité marginale de probabilité.

On étudie ensuite l'influence du paramètre k_Q, nombre des plus proches voisins, sur l'erreur quadratique moyenne entre la fonction de densité marginale estimée et le modèle analytique correspondant. Les résultats obtenus sont consignés dans le tableau 2.

k_Q	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
E_1	0.21	0.17	0.14	0.13	0.12	0.11	0.10	0.08	0.09	0.10	0.10	0.11	0.12	0.13	0.14
E_2	1.24	0.85	0.57	0.53	0.51	0.49	0.48	0.50	0.54	0.60	0.66	0.67	0.69	0.72	0.76

Tableau 2

Ces résultats sont illustrés par la figure 4 qui représente les variations des erreurs quadratiques en fonction du nombre k_Q des plus proches voisins.

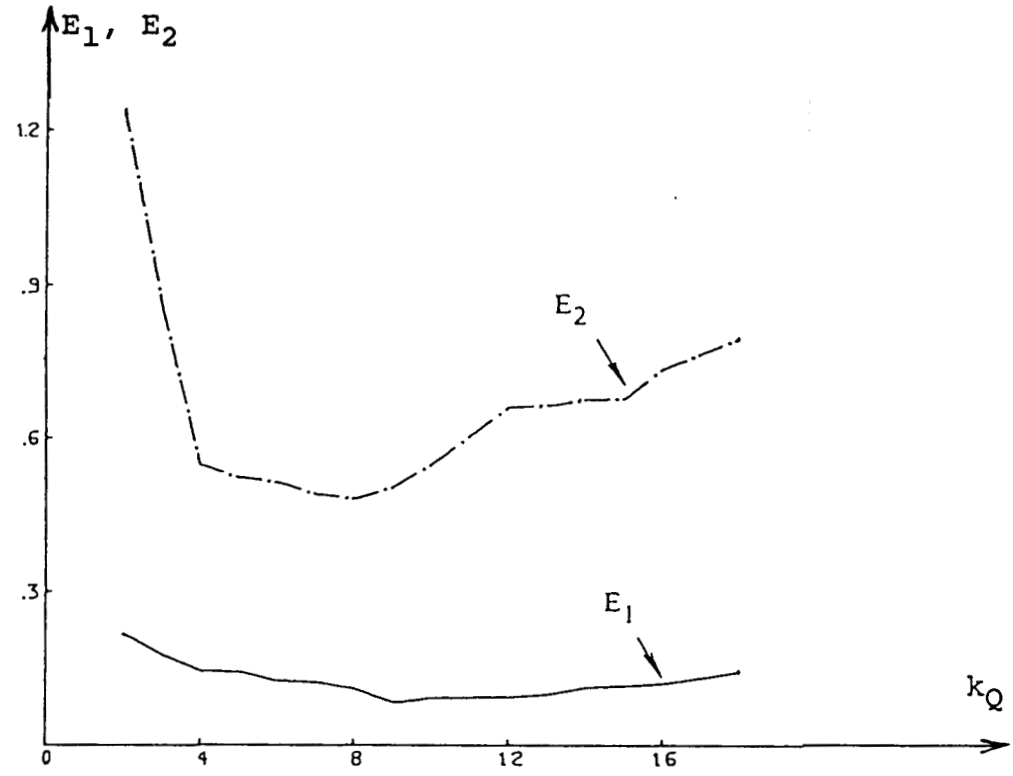


Figure 4: Variation des erreurs quadratiques en fonction du nombre de plus proches voisins.

La courbe en trait continu indique les variations de E_1 et celle en trait pointillé les variations de E_2.

Les variations de ces erreurs quadratiques mettent en évidence les valeurs optimales k^*_1 et k^*_2 des paramètres k_1 et k_2 respectivement, pour lesquels les erreurs quadratiques passent par un minimum E^*_1 et E^*_2. Dans l'exemple considéré on a:

pour $k^*_1 = 9$ $\qquad E^*_1 = 0.08$
pour $k^*_2 = 8$ $\qquad E^*_2 = 0.48$

A partir des valeurs optimales k^*_1 et k^*_2, on estime les fonctions de densité marginale de probabilité $p(x_1)$ et $p(x_2)$. Les estimations obtenues sont présentées sur la figure 5.

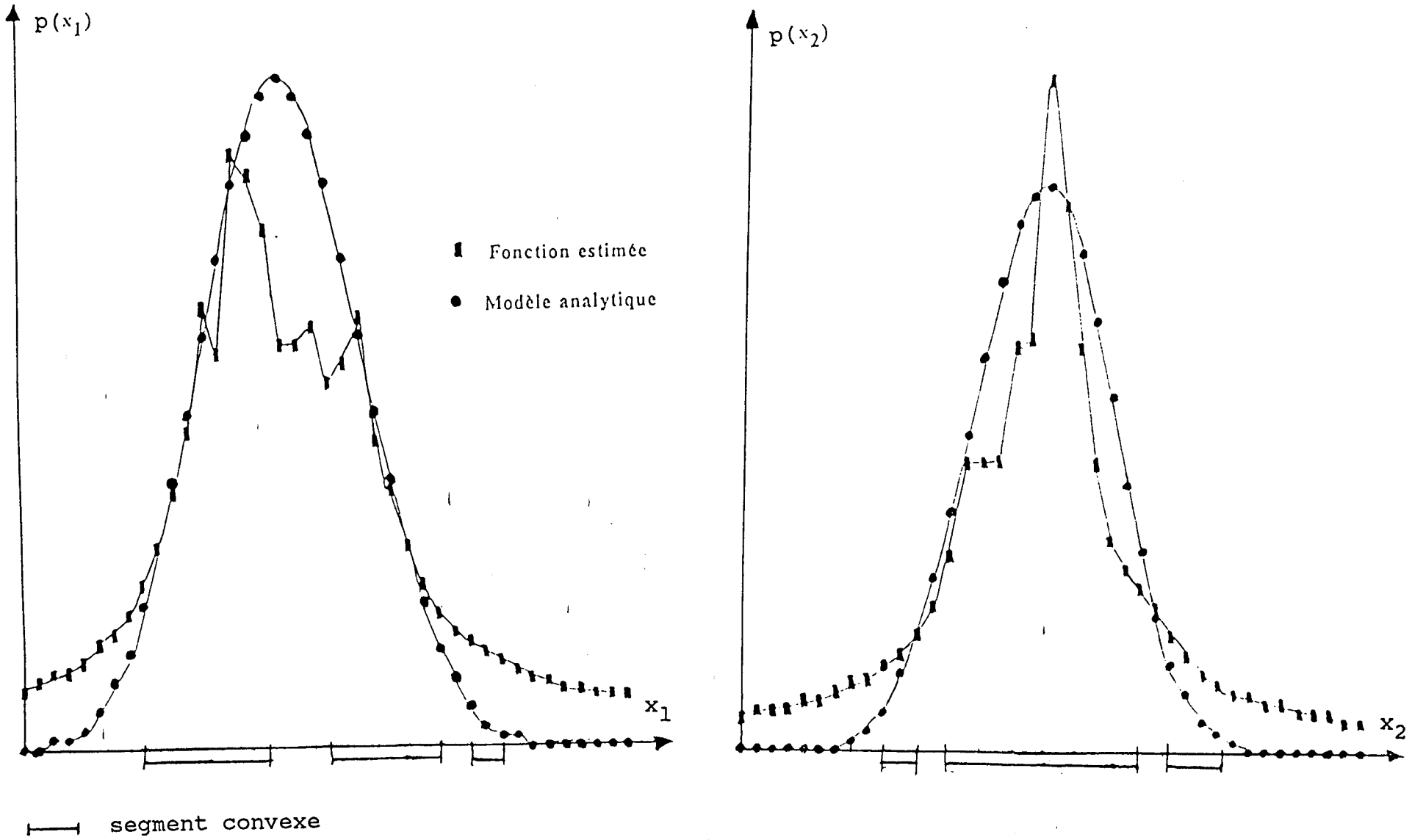


Figure 5 : Fonctions de densité estimées par la méthode des K_Q plus proches voisins et leurs modèles analytiques

II- 7 ANALYSE DE LA CONVEXITE DES FONCTIONS DE DENSITE MARGINALE.

La forme particulièrement simple de ces fonctions de densité marginale estimées par la méthode du noyau ou par celle des plus proches voisins est exploitée pour identifier les paramètres de ces fonctions de densité monovariables.

On sait que, sous l'hypothèse d'indépendance statistique des caractères, chaque fonction de densité marginale monovariable $p(x_i)$ $i=1,n$ peut être décomposée en un segment convexe compris entre deux segments concaves. Le segment convexe est centré sur la valeur moyenne $\bar{x}_i$ et sa longueur est égale au double de l'écart type σ_i . Ce segment, noté d_i, tel que:

$$d_i = [\bar{x}_i - \sigma_i, \quad \bar{x}_i + \sigma_i]$$

est appelé "segment de convexité" de la fonction $p(x_i)$, $i= 1, 2,...,n$.

La convexité de la fonction de densité marginale de probabilité $\hat{p}(x_i)$ $i =1,n$ est testée en tous les points où elle a été estimée. La procédure consiste à tester la position de ces points par rapport aux segments de droite joignant les points adjacents (cf figure 6).

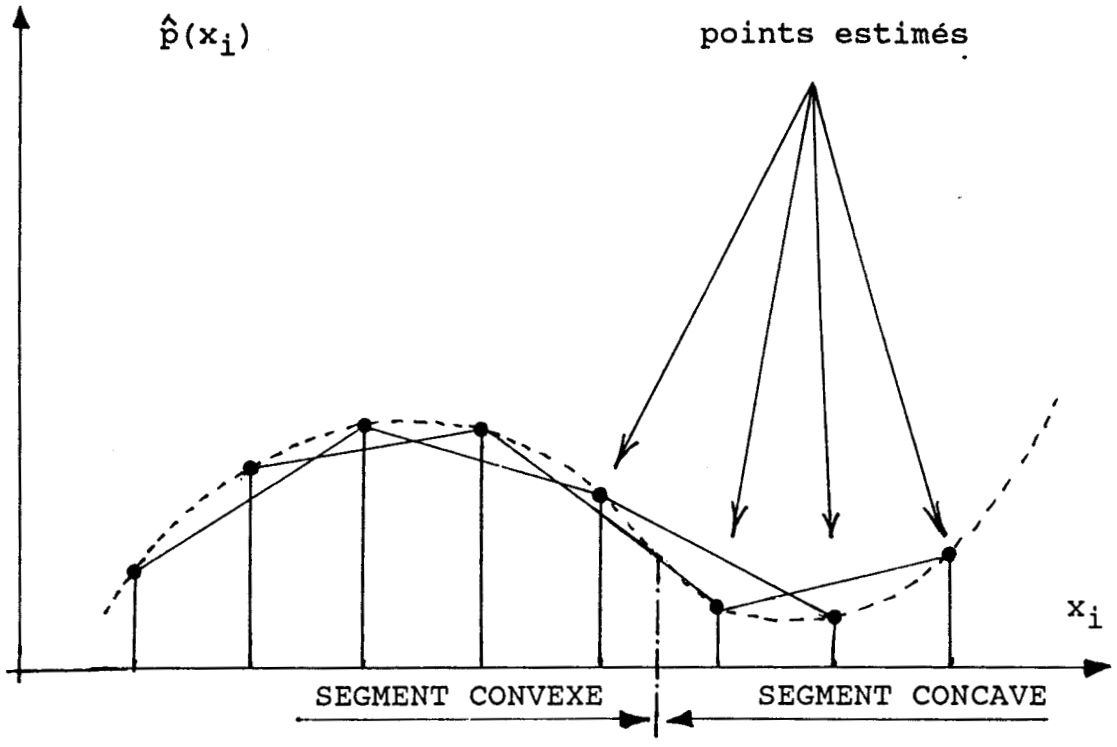


Figure 6: Estimation discrète des fonctions de densité marginale et test de convexité.

Mais, dans la pratique, nous pouvons être confrontés, après analyse directe de la convexité des fonctions de densité marginale de probabilité à une série alternée de points à convexité opposée, surtout dans les zones d'inflexion.

Pour pallier la dégradation des performances du test de convexité, lorsque nous nous trouvons dans de tels cas, et afin d'assurer la cohérence des résultats obtenus en chaque point, les résultats locaux sont comparés tout au long de la courbe représentant la fonction de densité marginale de probabilité.

Lorsqu'on obtient deux résultats successifs différents, on ajourne la prise de décision en mémorisant ces résultats, tout en attendant une confirmation ou infirmation par l'analyse de convexité des points suivants. On peut schématiser d'une manière globale cette partie d'algorithme par la description suivante:

a/ Si la convexité au point x_m est la même que celle de x_{m-1} alors le problème ne se pose pas.

b/ Si la convexité au point x_m est différente de celle de x_{m-1} alors on mémorise ce résultat et on teste la convexité au point x_{m+1}.

b1/ Si la convexité au point x_{m+1} est la même que celle de x_m alors il y a eu changement de convexité à partir du point x_m.

b2/ Si la convexité au point x_{m+1} est différente de celle au point x_m alors la convexité assignée au point x_m est celle de ces deux voisins.

Cette étude de la convexité ainsi établie nous permet d'éliminer l'effet de quelques variations résiduelles de convexité non significatives.

Les coordonnées des milieux des segments de convexité ainsi déterminés pour les n densités marginales d'une distribution fournissent une valeur approchée $\hat{\bar{X}}$ de son vecteur moyenne. Les carrés des demi-longueurs de ces segments donnent des valeurs approchées des éléments diagonaux de sa matrice de covariance.

Nous appliquons le test de convexité, aux meilleures estimations des fonctions de densité marginale obtenues pour l'échantillon représenté par la figure 1, par la méthode du noyau, autrement dit, aux estimations pour lesquelles on a obtenus une erreur quadratique minimale.

L'analyse de la convexité fait apparaître deux ou trois segments convexes sur chacune des deux fonctions de densité marginale de probabilité (cf. figure 3). Ces résultats ne nous satisfont pas sachant qu'initialement nous avons considéré un échantillon constitué d'une classe unique. Cet exemple montre donc que même en utilisant la meilleure estimation possible des fonctions de densité marginale, la dégradation des propriétés de convexité ne permet pas, compte tenu de la taille réduite de l'échantillon disponible, de retrouver la structure des données.

De manière à pouvoir, malgré tout, utiliser le test de convexité qui a déjà fait ses preuves pour des échantillons de grande taille, nous avons donc été amenés à modifier l'organisation du traitement des fonctions de densité marginale de probabilité: celles-ci sont soumises avant ce test à un filtrage non linéaire exposé en détail au chapitre suivant.

De manière analogue, on applique le test de convexité aux fonctions de densité marginale de probabilité estimées par la méthode des plus proches voisins. On rencontre le même problème que dans le cas des fonctions de densité marginale de probabilité estimées par la méthode du noyau. En effet, nous obtenons deux segments convexes au lieu d'un seul attendu sur chacune des fonctions de densité marginale estimée (cf. figure 5). Pour pallier cette dégradation de la méthode pour les petits échantillons, nous proposerons au chapitre suivant la même solution que celle proposée par la méthode du noyau.

II - 8 CONCLUSION.

L'utilisation des techniques d'estimation non paramétrique des fonctions de densité des probabilité et d'analyse directe de la convexité de ces fonctions, ne nous a pas permis d'identifier une distribution normale à partir des observations. En effet, on ne peut exploiter les résultats obtenus par application directe du test de convexité sur les fonctions de densité marginale de probabilité estimées comme, nous l'avons montré à travers ce chapitre.

Pour pallier cet effet, nous appliquons un opérateur de filtrage non linéaire.

CHAPITRE III

RESTAURATION DES PROPRIETES DE CONVEXITE PAR FILTRAGE NON LINEAIRE.

CHAPITRE III

RESTAURATION DES PROPRIETES DE CONVEXITE PAR FILTRAGE NON LINEAIRE

III - 1 INTRODUCTION

L'analyse des performances des méthodes d'estimation non paramétriques classiques a montré que, pour de petits échantillons, il est difficile d'estimer correctement les fonctions de densité marginale de probabilité.

L'analyse directe de la convexité des fonctions de densité marginale ne permet pas de détecter les modes correspondants aux classes en présence dans l'échantillon considéré. En effet, les fonctions de densité marginale sont très "bruitées" et présentent de nombreux modes parasites dont la présence ne peut s'expliquer que par le nombre insuffisant d'observations disponibles.

Pour remédier à cette dégradation des résultats obtenus par estimation non paramétrique classique, on propose dans ce chapitre une technique de filtrage non linéaire qui permet de reconstituer la forme des fonctions de densité marginale de probabilité estimées à partir d'un nombre très restreint d'observations.

III - 2 PRINCIPE DU FILTRAGE NON LINEAIRE.

Soit $\hat{p}_m(x_i)$, $m = 1, 2, \ldots, M$; $i = 1, 2, \ldots, n$ la version discrète de l'estimation de $\hat{p}(x_i)$. Pour chaque point d'échantillonnage d'indice m, où on désire estimer la fonction de densité marginale de probabilité, on calcule la valeur moyenne de l'estimation sur une fenêtre de largeur l située à gauche du point d'indice m, c'est à dire la quantité:

$$\frac{\sum_{j=m-l}^{m} \hat{p}_j(x_i)}{l}$$

De même, on calcule la valeur moyenne de l'estimation sur une fenêtre de même largeur l, mais cette fois-ci, située à droite du point où l'on estime la fonction de densité marginale, soit la quantité:

$$\frac{\sum_{j=m}^{m+l} \hat{p}_j(x_i)}{l}$$

La version filtrée de l'estimation de $\hat{p}(x_i)$ au point d'échantillonnage d'indice m est le produit de la moyenne sur une fenêtre de largeur l à gauche du point d'indice m, par la moyenne sur une fenêtre de même largeur, à droite de ce point (cf. figure 1) soit :

$$\hat{p}'_m(x_i) = (1/l)^2 \sum_{j=m-l}^{m} \hat{p}_j(x_i) \sum_{j=m}^{m+l} \hat{p}_j(x_i)$$

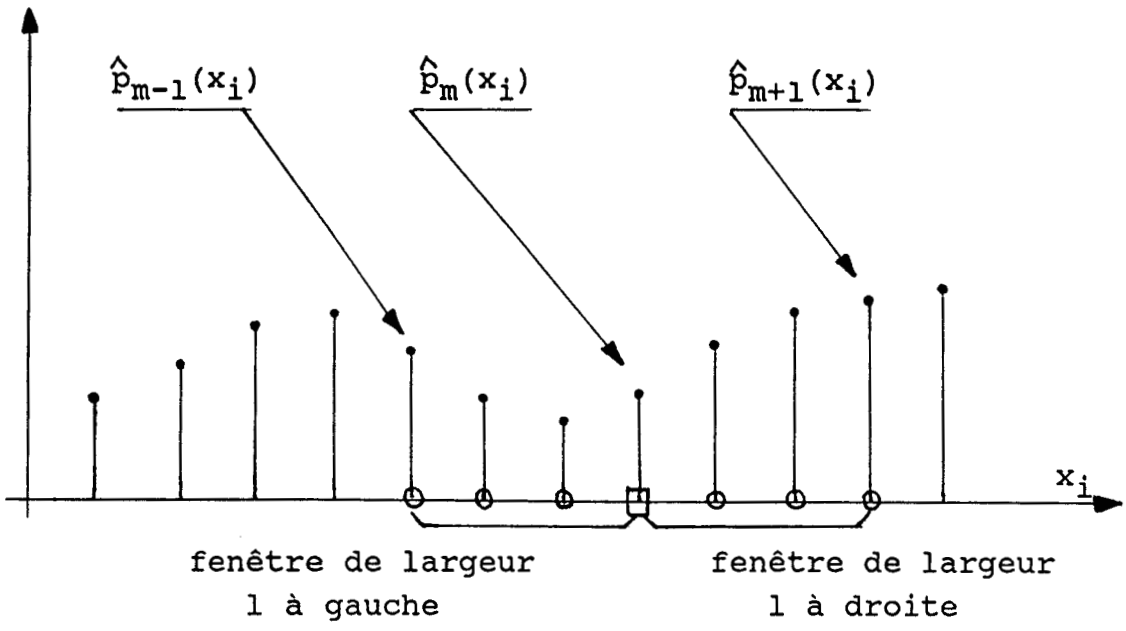


Figure 1: Fenêtre de calcul de l'opérateur de filtrage non linéaire.

Pour les points d'échantillonnage situés aux extrémités de la fonction de densité marginale de probabilité, on distingue deux cas:

a/ $1 \leq m \leq l$.

La moyenne sur la fenêtre à gauche est calculée sous la forme:

$$\frac{\sum_{j=1}^{m} \hat{p}_j(x_i)}{m}$$

b/ $M-l \leq m \leq M$.

La moyenne sur la fenêtre à droite est calculée sous la forme:

$$\frac{\sum_{j=m}^{M} \hat{p}_j(x_i)}{m}$$

III - 3 APPLICATION DE L'OPERATEUR DE FILTRAGE NON LINEAIRE A L'IDENTIFICATION D'UNE DISTRIBUTION NORMALE

Pour illustrer les performances de la méthode proposée, on applique cet opérateur de filtrage non linéaire aux meilleures estimations des fonctions de densité marginale, associées à l'échantillon de trente observations présenté sur la figure 1 du chapitre II, et ceci pour les deux méthodes d'estimation non paramétrique étudiées.

En faisant varier la largeur l de la fenêtre de l'opérateur de filtrage, nous pouvons juger l'effet de l'ajustement de ce paramètre de réglage sur les performances de la méthode.

Les tableaux 1 et 2 indiquent l'évolution de l'erreur quadratique entre la fonction de densité estimée-filtrée $\hat{p}'_m(x_i)$ et le modèle analytique $p_m(x_i)$, en fonction du paramètre l.

Le tableau 1 représente les résultats des variations de l'erreur quadratique dans le cas de l'estimation par la méthode du noyau et le tableau 2, ceux obtenus dans le cas de l'estimation par la méthode des k_Q plus proches voisins.

Largeur l	Méthode du noyau E_1	Méthode du noyau E_2	Méthode des p.p.v E_1	Méthode des p.p.v E_2
2	0.065	0.128	0.127	0.324
4	0.045	0.121	0.320	0.160
6	0.025	0.116	0.094	0.037
7	0.018	0.110	0.010	0.038
8	0.012	0.106	0.013	0.061
9	0.007	0.086	0.035	0.109
10	0.004	0.064	0.055	0.186
11	0.002	0.038	0.075	0.280
12	0.001	0.015	0.108	0.413
13	0.001	0.019	0.140	0.550
14	0.003	0.030	0.153	0.647
15	0.006	0.062	0.175	0.695
16	0.012	0.106	0.184	0.753
17	0.018	0.153	0.192	0.792
18	0.023	0.194	0.201	0.814

Tableau 1

Tableau 2

Les résultats obtenus montrent la robustesse de la procédure par rapport au réglage de la largeur l de la fenêtre de l'opérateur de filtrage non linéaire.

En effet, on constate que pour de grandes plages de variations de l, l'erreur quadratique entre la fonction de densité marginale estimée-filtrée et son modèle analytique est inférieure à l'erreur entre la meilleure fonction de densité marginale de probabilité estimée brute et ce modèle.

Ces plages de variation apparaissent encadrées de traits gras sur les tableaux 1 et 2. On note d'autre part, que pour les meilleurs ajustements du paramètre l, on obtient des valeurs de l'erreur quadratique nettement inférieures aux valeurs minimales correspondant aux réglages optimaux des paramètres des estimateurs. On obtient ainsi, après filtrage, des fonctions de densité marginale $\hat{p}'_m(x_i)$ beaucoup plus proches de leurs formes analytiques réelles $p_m(x_i)$ que les fonctions de densité marginale de probabilité $\hat{p}_m(x_i)$ puisque les erreurs quadratiques sont environ dix fois plus faibles (cf. tableau 3, algorithme I).

Méthodes d'estimation utilisées	Erreurs quadra--tiques	Estimations bruitées	estimations filtrées Algorithme 1	2	3
Méthode du noyau	E_1	0.008	0.001	0.012	0.025
	E_2	0.157	0.015	0.024	0.085
Méthode des P.P.V.	E_1	0.085	0.009	0.015	0.350
	E_2	0.480	0.037	0.088	1.120

Tableau 3

En effet, la somme des carrés des écarts quadratiques entre la fonction de densité marginale estimée-filtrée $\hat{p}'_m(x_i)$ et le modèle analytique $p_m(x_i)$ peut être réduite à:

$$E^{*\prime}_1 = 0.001 \qquad \text{et} \qquad E^{*\prime}_2 = 0.015$$

alors qu'elle ne pouvait descendre en dessous de:

$$E^{*}_1 = 0.008 \qquad \text{et} \qquad E^{*}_2 = 0.157$$

par la méthode du noyau. De même, cette somme peut être réduite à:

$$E^{*\prime}_1 = 0.009 \qquad \text{et} \qquad E^{*\prime}_2 = 0.037$$

alors qu'elle ne pouvait descendre en dessous de:

$$E^{*}_1 = 0.085 \qquad \text{et} \qquad E^{*}_2 = 0.480$$

par la méthode des k_Q plus proches voisins.

Les figures 2(a) et 2(b) représentent sur les mêmes diagrammes les fonctions de densité marginale estimées-filtrées $\hat{p}'_m(x_i)$ et leurs modèles réels $p_m(x_i)$ i=1,2 pour les deux méthodes d'estimations envisagées.

L'analyse directe de la convexité de ces fonctions de densité estimées ne permettait pas d'identifier les segments convexes ou concaves de ces fonctions. Par contre, après filtrage, on les détecte correctement, comme en témoignent les figures 2(a) et 2(b).

Ces segments nous permettent de calculer les paramètres statistiques de la distribution analysée. Pour l'exemple considéré, on trouve alors:

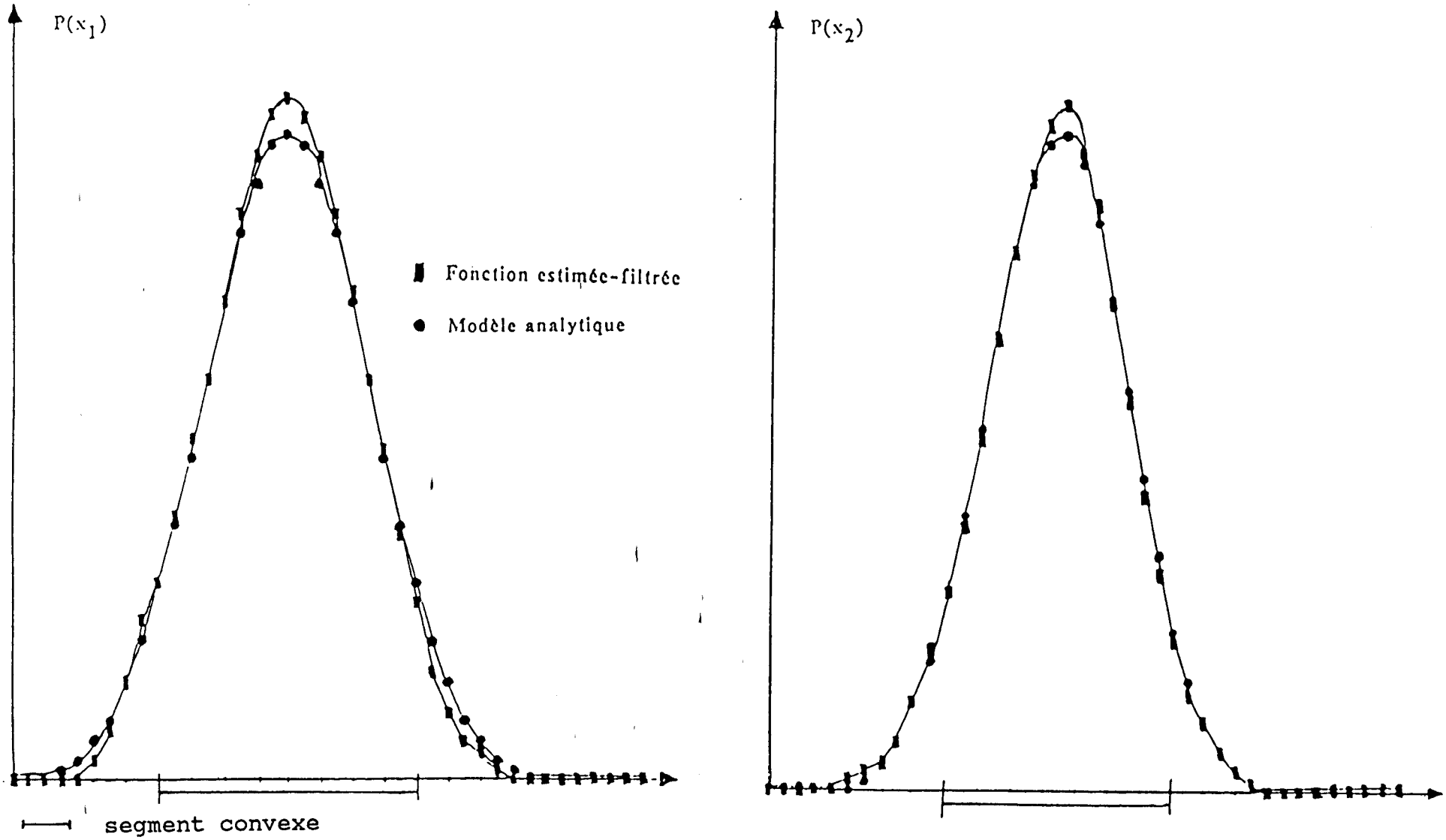


Figure 2 (a): Estimation des fonctions de densité marginale de la distribution normale estimées par la méthode du noyau après filtrage

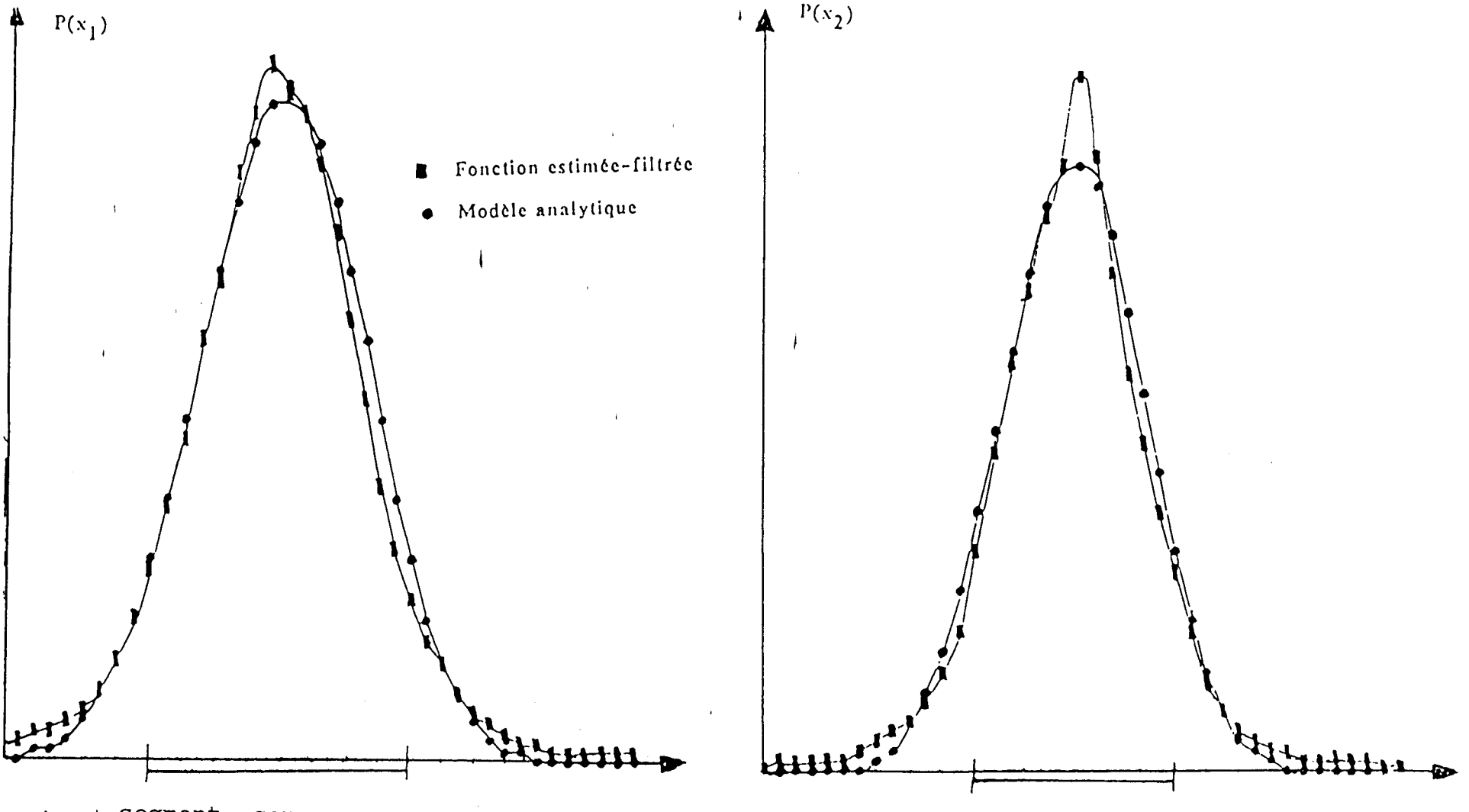


Figure 2(b): Estimation des fonctions de densité marginale de la distribution normale estimées par la méthode des k_Q plus proches voisins après filtrage.

a/ Dans le cas de la méthode du noyau.

$$\hat{\bar{X}} = [5.89,\ 9.97]^T \quad ; \quad \hat{\sum} = \begin{vmatrix} 1.12 & 0 \\ 0 & 0.48 \end{vmatrix}$$

b/ Dans le cas de la méthode des k_Q plus proches voisins.

$$\hat{\bar{X}} = [5.80,\ 9.82]^T \quad ; \quad \hat{\sum} = \begin{vmatrix} 1.12 & 0 \\ 0 & 0.52 \end{vmatrix}$$

III - 4 COMPARAISON AVEC D'AUTRES TECHNIQUES DE FILTRAGE NON LINEAIRE.

III-4-1 Opérateur de filtrage à moyenne glissante.

Une alternative à l'opérateur de filtrage non linéaire proposé est le filtre utilisant une moyenne glissante classique.

La densité de probabilité en chaque point où on estime la fonction sera égale à la moyenne des densités $\hat{p}_m(x_i)$, calculée sur une fenêtre de largeur 1 centrée au point x_i . Cela revient à calculer:

$$\hat{p}'_m(x_i) = (1/L) \sum_{m-1/2}^{m+1/2} \hat{p}_j(x_i)$$

où L est la largeur de la fenêtre de l'opérateur de filtrage à moyenne glissante.

Nous appliquons alors cet opérateur de filtrage aux meilleures estimations obtenues pour l'exemple de la figure 1 du chapitre II. Les résultats sont donnés par les figures 3(a) et 3(b) qui indiquent les fonctions de densité marginale estimées et filtrées par l'opérateur de filtrage décrit dans ce paragraphe. En évaluant l'écart quadratique entre ces fonctions estimées-filtrées et leurs modèles analytiques, nous obtenons les résultats consignés dans le tableau 3, algorithme II.

III - 4 - 2 Opérateur de filtrage non linéaire de A. Rosenfeld et P. Torre.

Une autre technique de filtrage comparable à celle exposée dans ce chapitre a été proposée par A. Rosenfeld et P. Torre [ROS.83]. Cette méthode de filtrage consiste à calculer en chaque point d'échantillonnage d'indice m, la quantité:

$$W_i = \sum_{j=1}^{m-1} \hat{p}_j(x_i) \qquad \sum_{j=m}^{M} \hat{p}_j(x_i) \qquad i=1,\ 2,\ldots,n$$

Toutefois, sans contrôle de la largeur des fenêtres servant à calculer les sommes intervenant dans l'opérateur de filtrage, cette technique conduit à des résultats moins satisfaisants que ceux obtenus avec l'opérateur de filtrage non linéaire proposé au paragraphe III-2 .

Ceci a été mis en évidence après utilisation de l'opérateur de filtrage de A. Rosenfeld et P. Torre, sur les fonctions de densité marginale $\hat{p}_m(x_i)$ des observations de l'échantillon de la fig.1 chapitre 2.

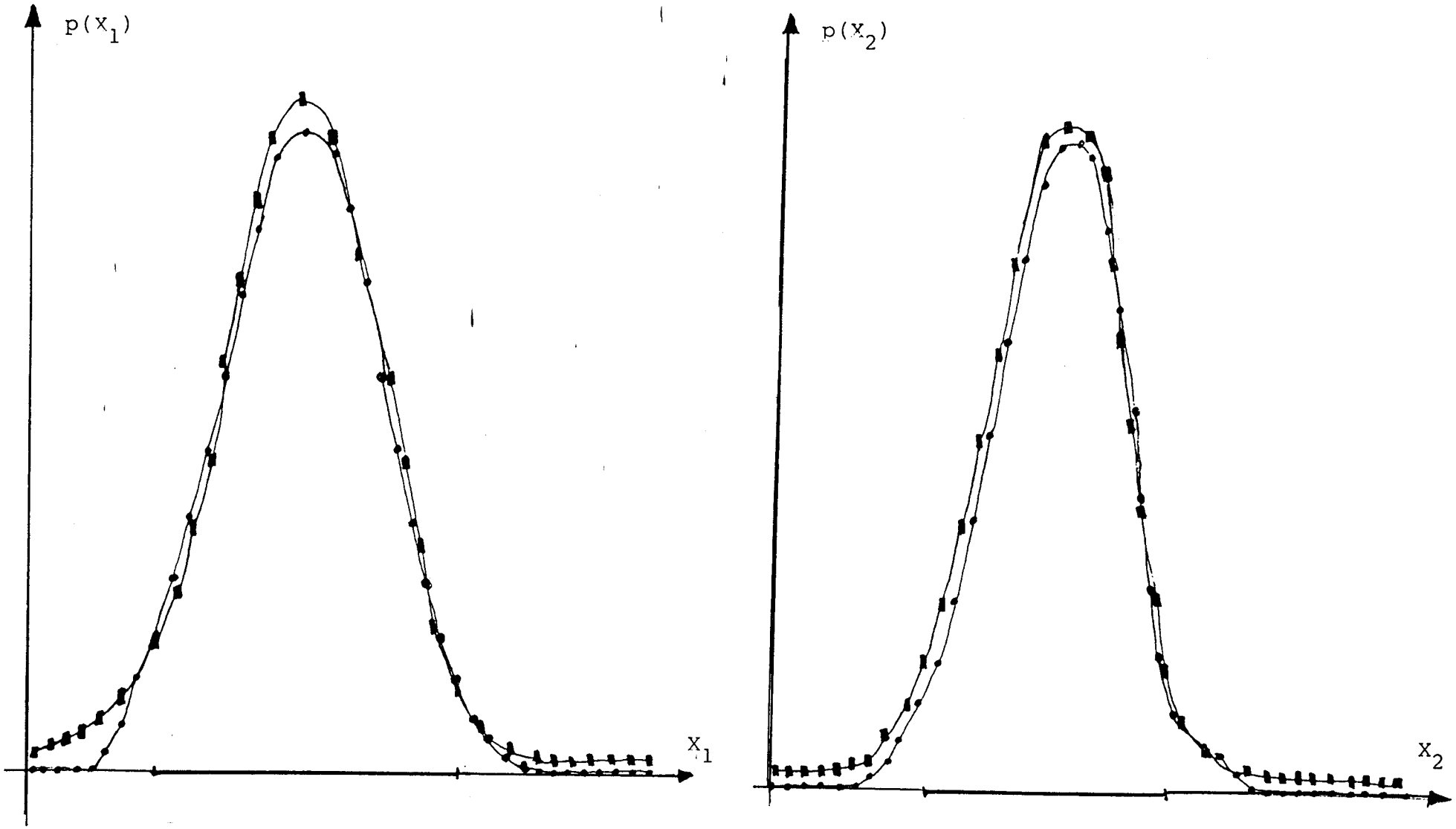
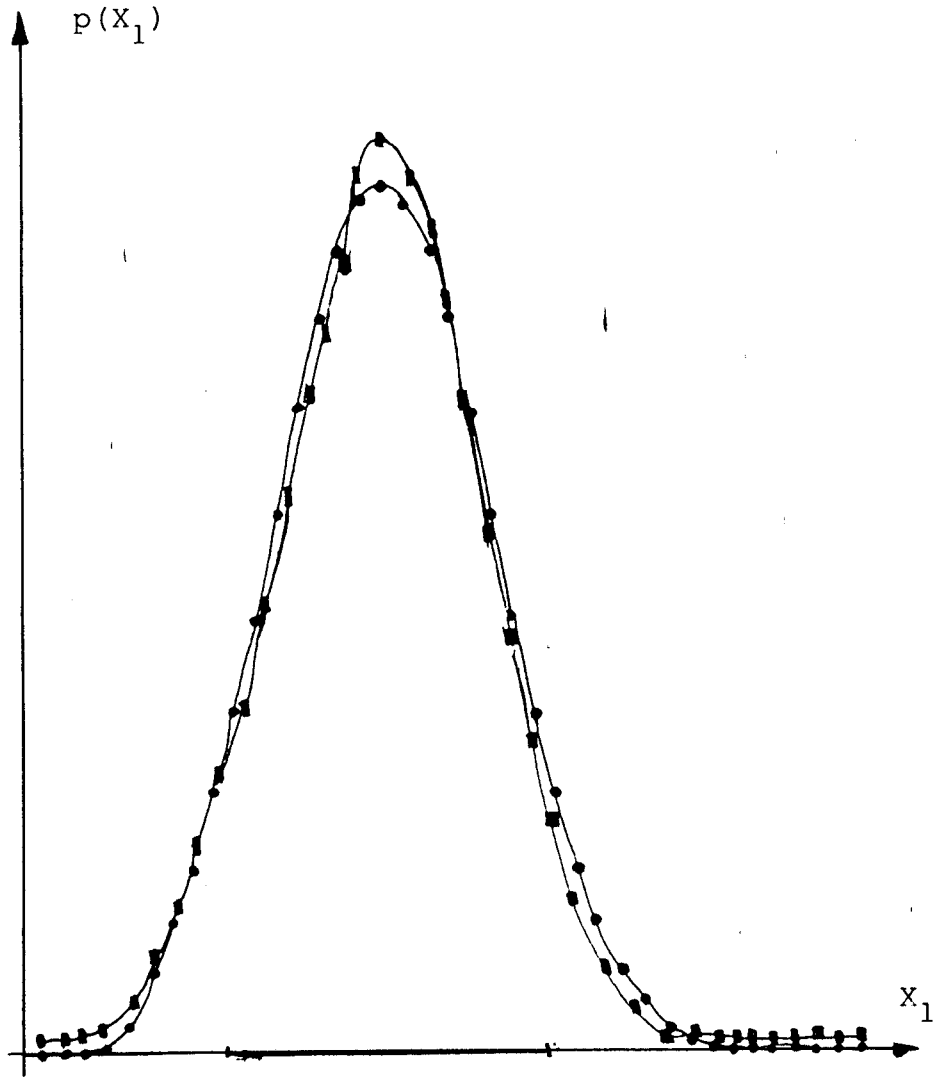


Figure 3(a) : Fonction de densité estimée par la méthode du noyau après application de l'opérateur de filtrage à moyenne glissante



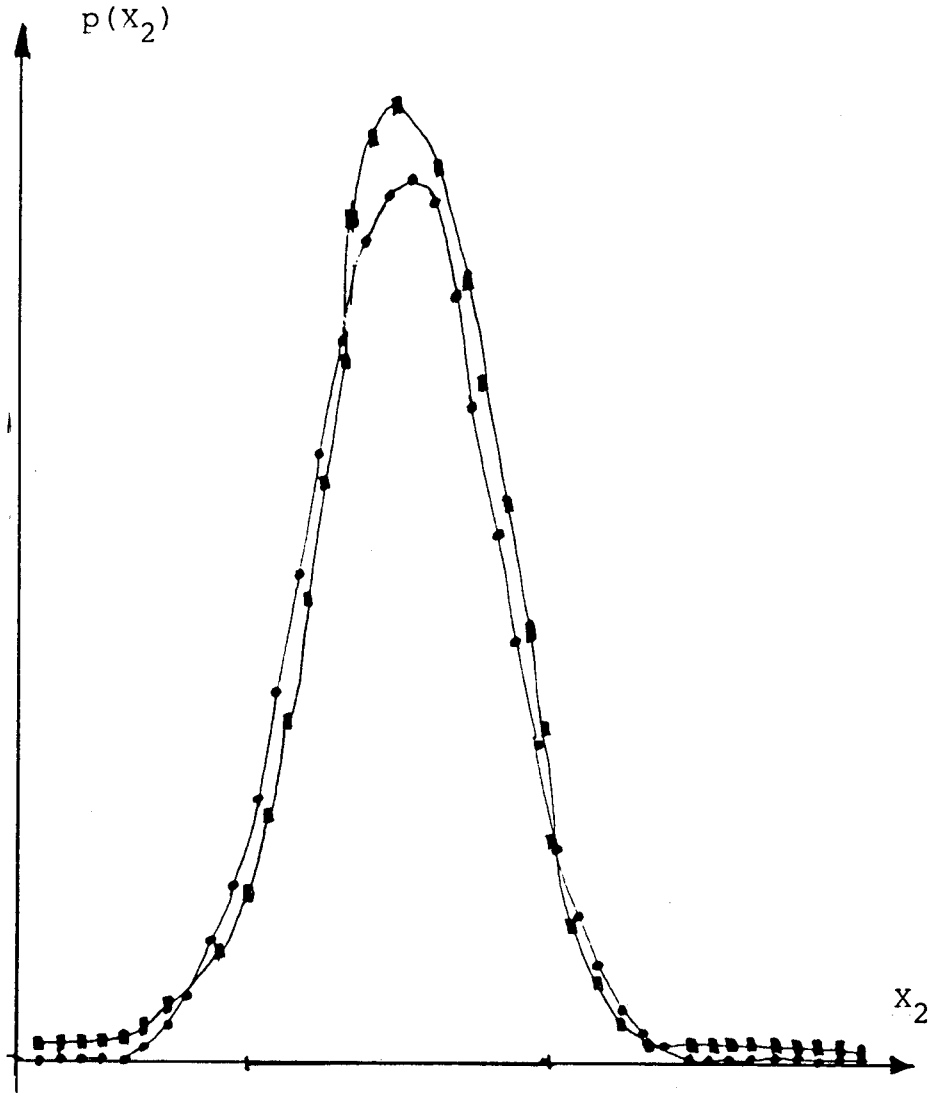


Figure 3(b) : Fonction de densité estimée par la méthode des plus proches voisins après application de l'opérateur de filtrage à moyenne glissante

Les résultats obtenus après application de ce filtre sont illustrés par la figure 4. Plus précisément, la figure 4(a) et 4(b) représentent les fonctions de densité marginale estimées respectivement par la méthode du noyau et par la méthode des k_Q plus proches voisins. Ces deux estimations sont filtrées par l'opérateur de filtrage de A. Rosenfeld et P. Torre.

En évaluant, l'écart quadratique entre la fonction estimée-filtrée par l'opérateur de filtrage décrit dans ce paragraphe $\hat{p}'_m(x_i)$ et son modèle réel $p_m(x_i)$, on obtient les résultats du tableau 3 (cf.algorithme III).

En comparaison de l'utilisation de l'opérateur de filtrage de l'algorithme III, ou bien d'une moyenne glissante classique, l'expérience montre que l'opérateur de filtrage non linéaire proposé présente l'avantage de supprimer les variations du signal non significatives sans toutefois en altérer le contenu. Le tableau 3 met en évidence la supériorité de l'algorithme proposé par rapport aux opérateurs de filtrage décrits dans le paragraphe III-4-1 et III-4-2 .

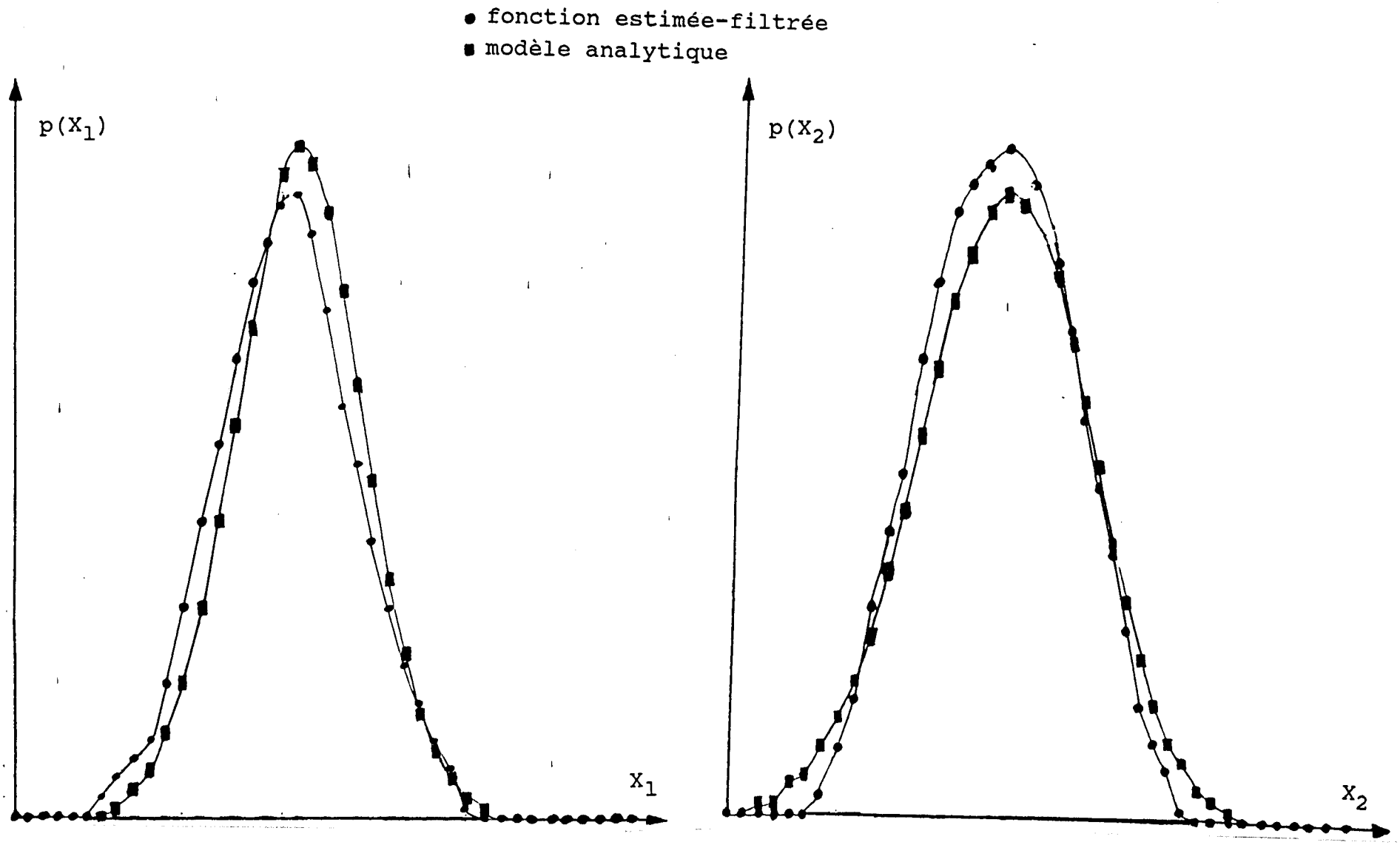


Figure 4(a): Fonctions de densité marginale estimées par la méthode du noyau après application de l'opérateur de filtrage non linéaire de Rosenfeld.

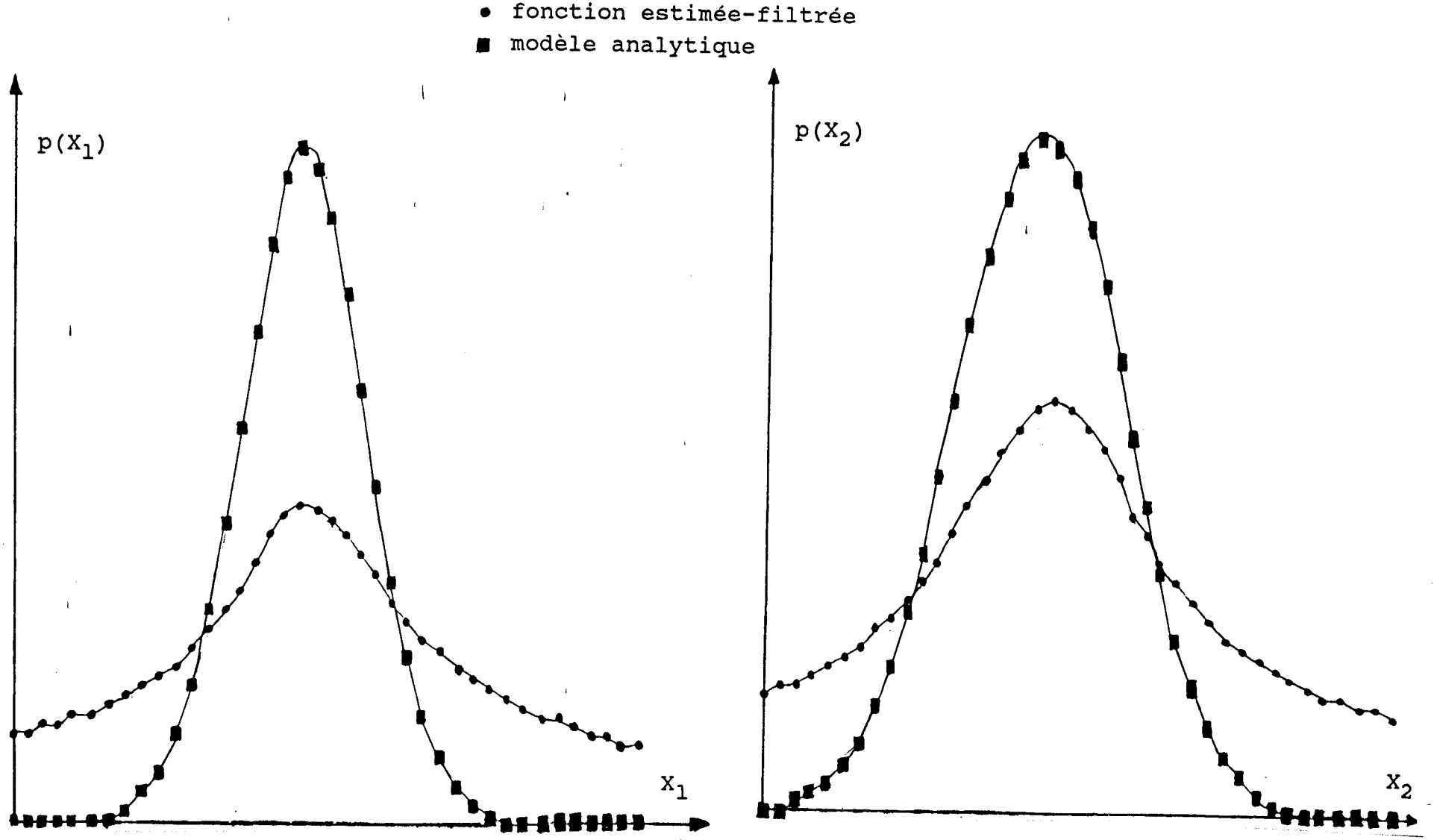


Figure 4(b): Fonctions de densité marginale estimées par la méthode des k_Q plus proches voisins après application de l'opérateur de filtrage non linéaire de Rosenfeld.

III - 5 CONCLUSION.

La technique d'estimation-filtrage que nous avons présentée dans ce chapitre permet de restaurer les propriétés de convexité des fonctions de densité marginale de probabilité estimées à partir d'échantillons de taille réduite.

L'utilisation de cette procédure permet d'améliorer les performances d'une méthode classique de classification automatique basée sur l'analyse de la convexité des fonctions de densité marginale de probabilité. Cette étude fait l'objet du chapitre IV suivant.

CHAPITRE IV

CLASSIFICATION OPTIMALE DES PETITS ECHANTILLONS.

CHAPITRE IV

CLASSIFICATION OPTIMALE DES PETITS ECHANTILLONS.

IV - 1 OPTIMISATION DU PROCESSUS DE CLASSIFICATION.

Il existe de nombreux problèmes d'analyse de données pour lesquels on ne dispose pas de prototypes. Nous n'avons alors qu'un ensemble d'observations multidimensionnelles qu'il s'agit de regrouper en différentes classes.

Dans ces conditions, on se trouve confronté à un problème de classification automatique, qui se démarque du problème de classement par le manque total d'informations a priori sur les données à analyser.

On rappelle que, dans le cadre de cette étude, les composantes constituant le mélange à analyser suivent des lois de distribution normales, comme il a été précisé au chapitre II. La classification optimale de ces observations se ramène donc à l'estimation des paramètres du mélange de ces lois. La solution que nous envisageons consiste à estimer et à modéliser les fonctions de densité de probabilité marginale de la distribution des observations disponibles. Cette analyse permet de calculer les valeurs approchées des vecteurs moyenne, matrice de covariance, et probabilités à priori des différentes classes en présence [POS.82a], [POS.83a], [POS.82b].

Une fois que le mélange est identifié, le problème du classement optimal revient à trouver un ensemble de fonctions $g_k(X)$, appelées fonctions de décision, telles qu'une observation X est attribuée à la classe C_k si et seulement si:

$$g_k(X) \geq g_i(X) \qquad k = 1,2,3,\ldots, K \quad \text{et} \quad k \neq i$$

où K est le nombre de classes mises en évidence.

Les valeurs approchées des paramètres statistiques des mélanges analysés permettent de calculer les fonctions de décision. Ces dernières peuvent être choisies de la forme [MAR.60].

$$g_k(X) = p(X|C_k) \cdot P(C_k) \qquad k=1,2,\ldots,K$$

où $P(C_k)$ est la probabilité a priori de la classe C_k avec:

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

où n est la dimension de l'espace de représentation des données. Σ_k et $\bar{X}_k$, sont respectivement la matrice de covariance et le vecteur moyenne de la classe C_k.

Comme les fonctions de densité sont de type exponentiel, il est préférable, pour les mélanges gaussiens, d'utiliser le logarithme des fonctions de décision $g_k(X)$ [DUD.73]. Dans ces conditions, on peut prendre:

$$g'_k(X) = \text{Log } g_k(X) \cdot (2\pi)^{n/2}$$

$$= \text{Log}\left(p(X|C_k) \cdot P(C_k)\right) \cdot (2\pi)^{n/2}$$

$$= -(1/2)\text{Log}\,\Sigma_k - (1/2)(X-\bar{X}_k)^T \Sigma_k^{-1} (X-\bar{X}_k)$$

$$+ \text{Log}P(C_k)$$

Nous remplaçons dans l'expréssion $g'_k(X)$, les valeurs exactes du vecteur moyenne $\bar{X}_k$, de la matrice de covariance, et de la probabilité a priori de la classe $C_k, P(C_k)$, par les valeurs approchées $\hat{\bar{X}}_k$, $\hat{\Sigma}_k$ et $\hat{P}(C_k)$ obtenues par analyse des fonctions de densité marginale.

En effet, il a été montré que la détermination des segments convexes des fonctions de densité marginale contient toute l'information nécessaire pour identifier des mélanges gaussiens sous l'hypothèse d'indépendance statistique des caractères [POS.82a], [POS.83b].

Selon une procédure désormais classique, ces segments permettent d'identifer des domaines modaux à partir desquels on peut calculer les vecteurs moyenne et la matrice de covariance des différentes classes en présence, après élimination des classes introduites artificiellement par le produit euclidien de ces segments convexes (cf.figure 1).

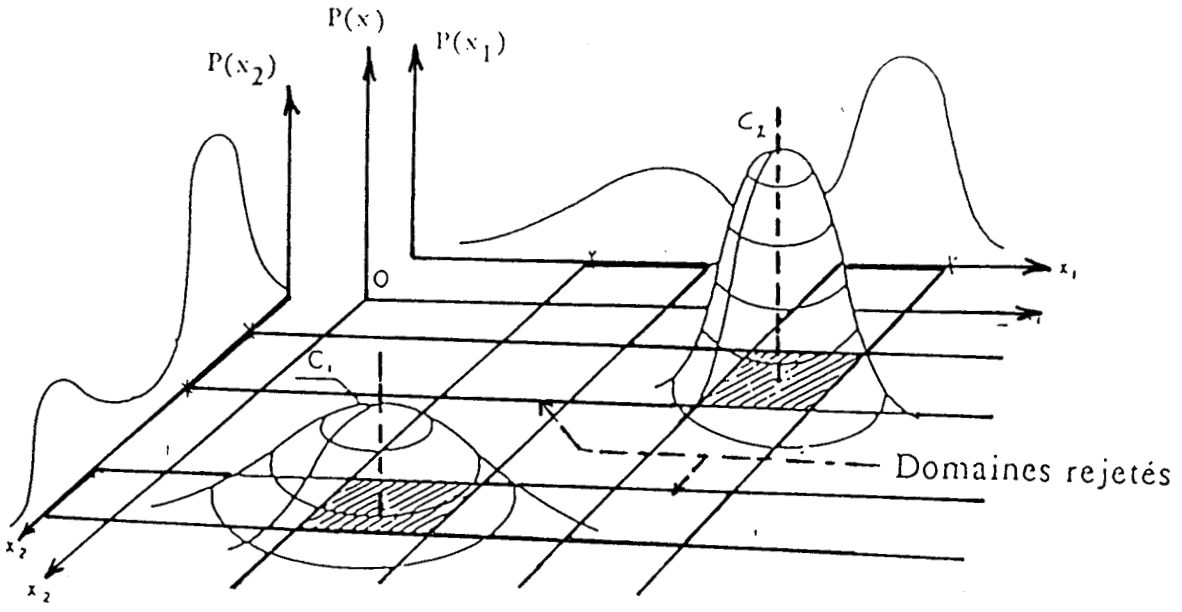


Figure 1: Détermination des domaines caractéristiques du mélange à partir des segments convexes de ses fonctions de densité marginale.

Les probabilités a priori sont obtenues en comptant le nombre d'observations situées dans les domaines caractéristiques de chaque classe.

Les fonctions de décisions deviennent alors:

$$\hat{g}'_k(X) = -(1/2)\,\mathrm{Log}\,\hat{\Sigma}_k - (1/2)(X-\hat{\bar{X}}_k)^T \hat{\Sigma}_k^{-1} (X-\hat{\bar{X}}_k) + \mathrm{Log}\hat{P}(C_k)$$

Ces fonctions définissent des surfaces de décision d'équations:

$$\hat{g}'_k(X) - \hat{g}'_i(X) = 0 \qquad k \neq i$$

Ces surfaces partitionnent l'espace en régions de décision R_k, k=1,2,...,K chacune d'elles ne contenant que les observations assignées à la classe correspondante.

Le problème du classement optimal revient donc à calculer un ensemble de fonction de décision $g_k(X)$, telles que l'observation X est attribuée à une classe C_k si et seulement si:

$$\hat{g}'_k(X) > \hat{g}'_i(X) \qquad k \neq i$$

L'exemple présenté à la section IV-2 illustre cette procédure d'optimisation de la classification.

IV - 2 APPLICATION DE LA TECHNIQUE D'ESTIMATION-FILTRAGE A LA CLASSIFICATION DES DONNEES MULTIVARIABLES.

La technique d'estimation-filtrage proposée au chapitre III, paragraphe III-2, permet de restaurer les propriétés de convexité des fonctions de densité marginale de probabilité estimées à partir de petits échantillons. Nous proposons d'utiliser cette procédure pour la classification automatique d'échantillon de taille réduite.

IV - 2 - 1 - Analyse d'un échantillon bidimensionnel.

EXEMPLE:

En utilisant un exemple à deux classes d'observations bidimensionnelles généré artificiellement pour lequel nous connaissons les modèles analytiques des fonctions de densité en présence, nous allons quantifier l'effet de l'opérateur de filtrage non linéaire sur des données de taille réduite.

L'échantillon est constitué de deux composantes contenant trente (30) observations chacune (cf. figure 2). Ces composantes suivent une loi normale. Les paramètres statistiques de cet échantillon sont indiqués sur le tableau 1.

	Vecteur moyenne	Matrice de covariance	Probabilité a priori
Classe C_1	$\overline{X}_1=[7,\ 7.7]^T$	$\Sigma_1=\begin{vmatrix} 1.2 & 0 \\ 0 & 0.5 \end{vmatrix}$	$P(C_1)=0.5$
Classe C_2	$\overline{X}_2=[9,\ 5]^T$	$\Sigma_2=\begin{vmatrix} 0.5 & 0 \\ 0 & 0.8 \end{vmatrix}$	$P(C_2)=0.5$

Tableau 1.

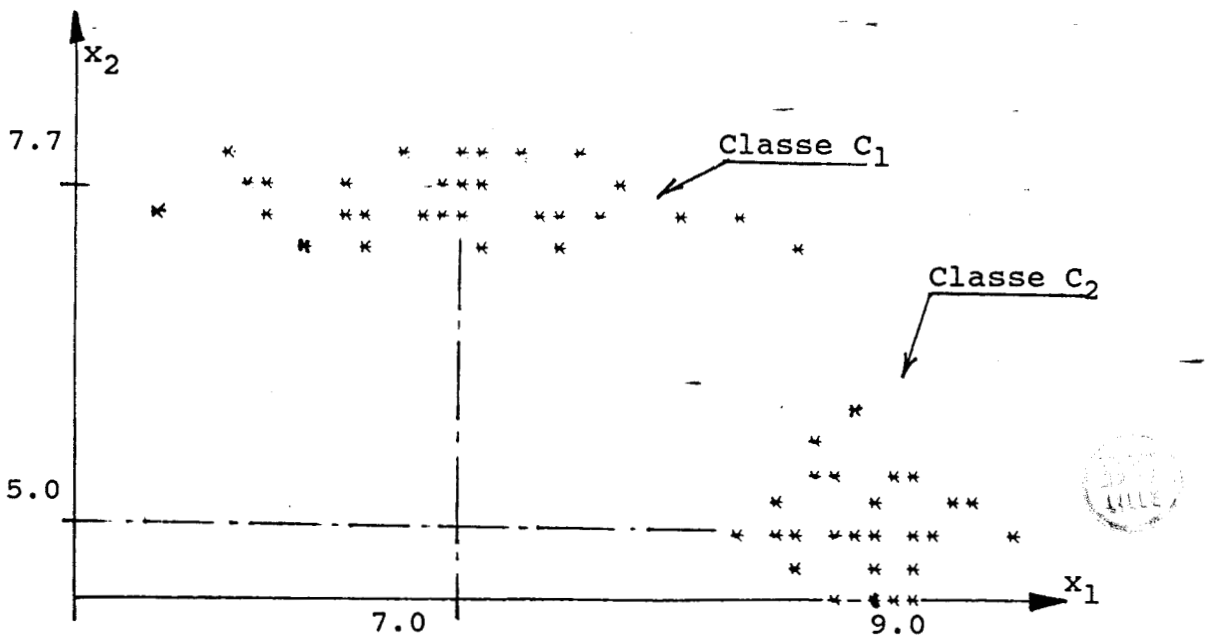


Figure 2: Représentation graphique d'un échantillon bidimentionnel.

Sur cet exemple, on applique les deux méthodes d'estimation non paramétrique décrites au chapitre II et nous étudions les erreurs quadratiques moyennes E_i entre les fonctions estimées $\hat{p}(x_i)$ et leurs modèles analytiques $p(x_i)$ $i=1,2$:

$$E_i = (1/M) \sum_{m=1}^{M} [\hat{p}_m(x_i) - p_m(x_i)]^2$$

où $p_m(x_i)$ et $\hat{p}_m(x_i)$, $m = 1,2,\ldots,M$ représentent les formes discrètes des fonctions continues $p(x_i)$ et $\hat{p}(x_i)$ $i=1,2$. (cf. chapitre II).

Les variations de E_i en fonction de la largeur h_Q de la fenêtre utilisée en employant la méthode du noyau mettent en évidence, pour chaque fonction de densité marginale de probabilité, une valeur optimale h^*_i $i=1,2$ de h_Q, pour laquelle l'erreur quadratique entre la fonction estimée et son modèle analytique correspondant passe par un minimum E^*_i, $i=1,2$.

$E^*_1 = 0.035$ $\qquad$ $E^*_2 = 0.042$

De manière analogue, dans le cas de la méthode d'estimation des plus proches voisins, on constate qu'il existe pour chaque fonction de densité marginale de probabilité, une valeur optimale k^*_i, $i=1,2$ de k_Q, pour laquelle l'erreur quadratique passe par un minimum E^*_i $i=1,2$.

$E^*_1 = 0.081$ $\qquad$ $E^*_2 = 0.140$

Le tableau 2 indique les valeurs optimales correspondantes des paramètres h_Q et k_Q, pour chaque axe et chaque méthode d'estimation.

	Valeur optimale correspondant à l'erreur quadratique minimale
Largeur de la fenêtre de Parzen	h(1) = 0.14 h(2) = 0.13
Nombre des plus proches voisins	k(1) = 9 k(2) = 8

Tableau 2

Les fonctions de densité marginale de probabilité estimées par les deux méthodes et à l'aide des valeurs optimales h^*_i et k^*_i, i=1,2 sont présentées par les figures 3(a) et 3(b). La figure 3(a) indique les fonctions de densité marginale estimées par la méthode du noyau et la figure 3(b) indique celles estimées par la méthode des k_Q plus proches voisins. Ces fonctions de densité marginale sont présentées avec leurs modèles analytiques gaussiens dont les caractéristiques statistiques sont déterminées à partir de celles du mélange représenté par la figure 2 .

Afin de retrouver les paramètres statistiques du mélange, nous appliquons sur les fonctions de densité marginale estimées, le test de convexité (cf. chapitre II paragraphe II-7). Ce dernier nous permet de déterminer les segments concaves et convexes des fonctions de densité marginale. Pour les deux méthodes d'estimation utilisées et pour cet exemple, l'analyse directe de la convexité ne permet pas de mettre en évidence les segments concaves et convexes attendus (cf. figure 3). Nous allons montrer que, malgré ce manque de fiabilité des estimateurs, même pour le réglage optimal des paramètres, l'opérateur de filtrage non linéaire proposé au chapitre III permet de retrouver la structure du mélange.

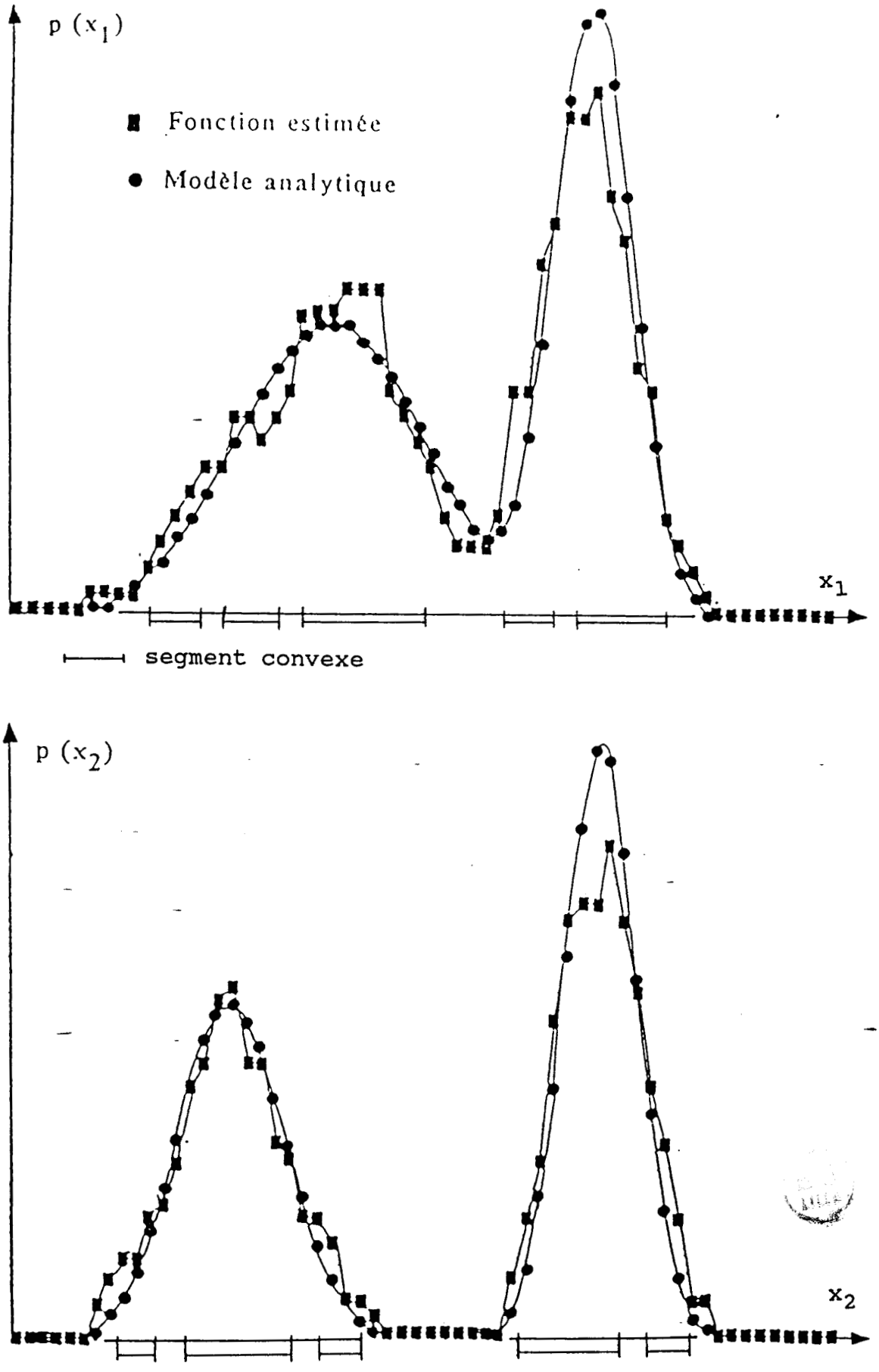


Figure 3 (a) : Estimation du mélange par la méthode du noyau

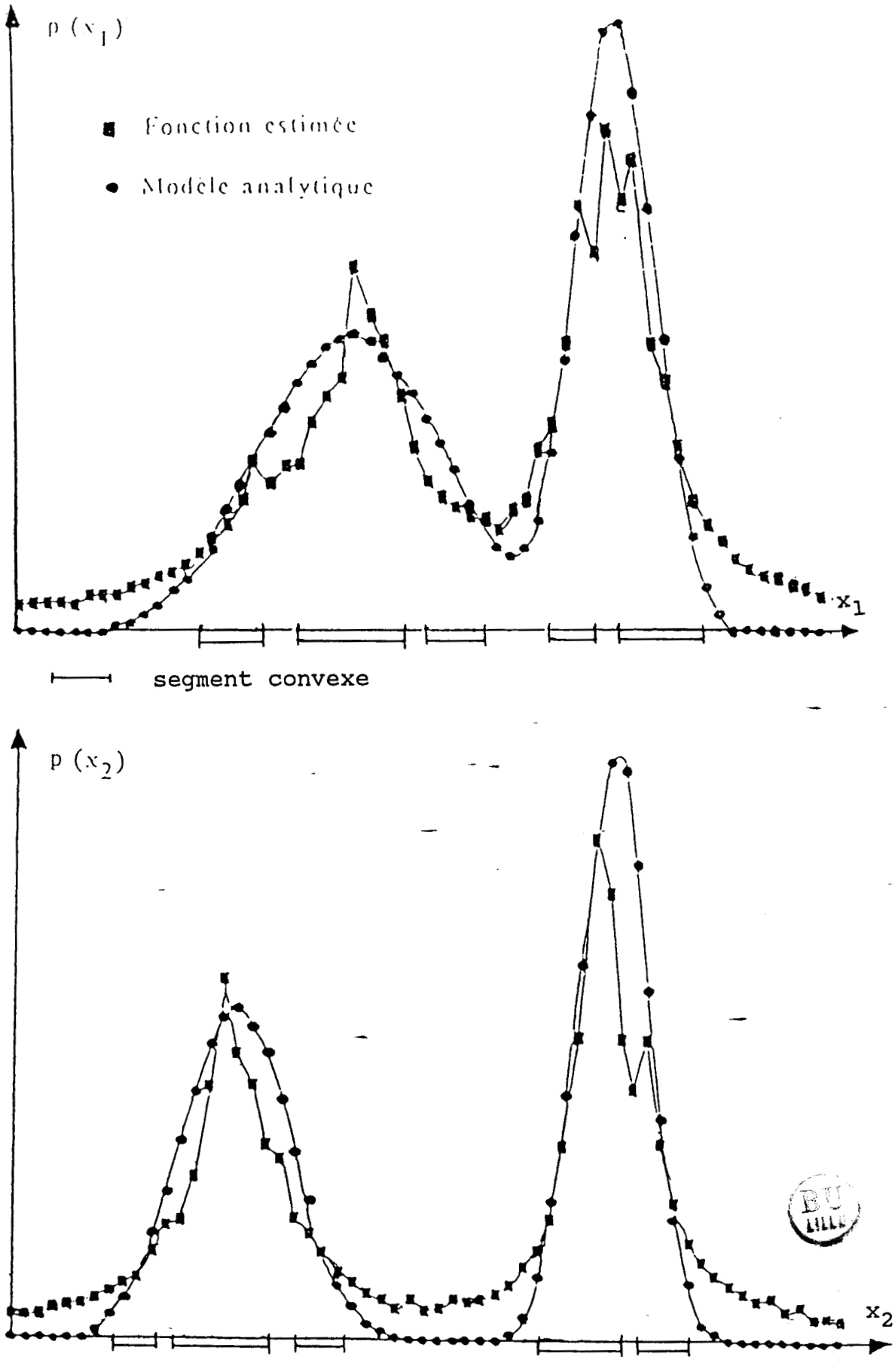


Figure 3 (b): Estimation du mélange par la méthode des plus proches voisins

IV -2 -2 Application de l'opérateur de filtrage non linéaire.

Comme l'analyse directe de la convexité des fonctions de densité marginale estimées ne permet pas toujours d'identifier les segments concaves et convexes, on utilise le filtre non linéaire décrit dans le chapitre III paragraphe II-1 pour restaurer les propriétés de convexité de ces fonctions.

Après filtrage, on obtient des erreurs quadratiques entre les fonctions de densité estimées-filtrées et leurs modèles analytiques nettement inférieures à celles obtenues avec les fonctions estimées non filtrées.

Les valeurs de ces erreurs quadratiques optimales sont indiquées dans le tableau 3 pour les deux méthodes d'estimation utilisées ainsi que les valeurs $\bar{l}$ de la largeur de la fenêtre de l'opérateur de filtrage non linéaire correspondant, qui ont été obtenue de la même manière que dans le chapitre III paragraphe III - 3.

	Erreur quadratique optimale après filtrage	Largeur de la fenêtre de l'opérateur de filtrage non linéaire
Méthode d'estimation du noyau	$E^*_1 = 0.0100$ $E^*_2 = 0.0063$	$\bar{l} = 5$ $\bar{l} = 8$
Méthode des plus proches voisins	$E^*_1 = 0.0507$ $E^*_2 = 0.0270$	$\bar{l} = 5$ $\bar{l} = 4$

Tableau 3

En appliquant le test de convexité aux fonctions de densité marginale estimées-filtrées, on détecte des segments convexes représentatifs de la structure du mélange (cf. figure 4).

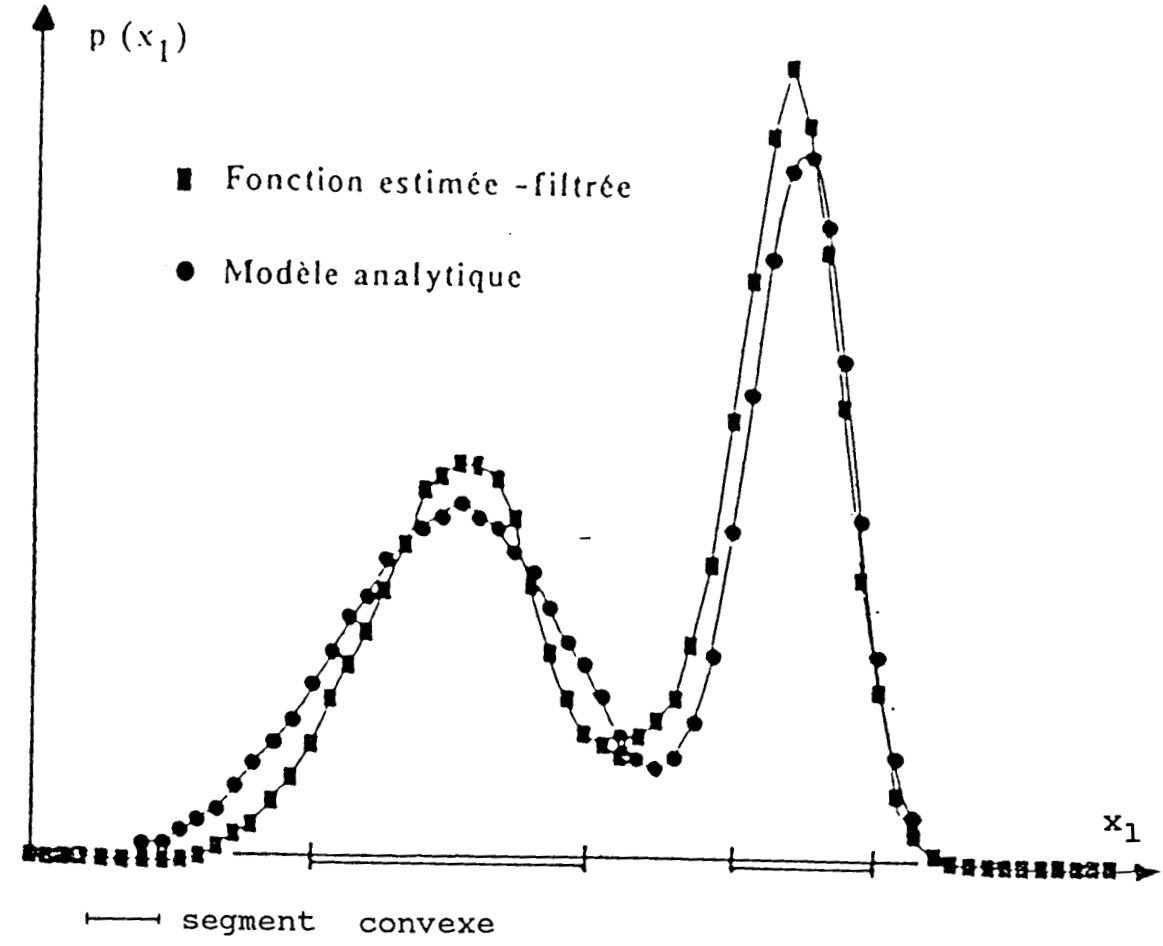
Les figures 4(a) et 4(b) représentent respectivement la fonction de densité estimée par la méthode du noyau et celle estimées par la méthode des k_Q plus proches voisins après filtrage non-linéaire. Ces fonctions de densité marginale estimées-filtrées sont présentées avec leurs modèles analytiques.

La détection des segments convexes sur les fonctions de densité marginale estimée-filtrée nous permet de déterminer les valeurs approchées des paramètres statistiques de la distribution considérée (cf. tableau 4).

	Vecteur moyenne	Matrice de covariance	Probabilité a priori
Classe C_1	$\hat{\underline{X}}_1=[7.5,\ 8.01]^T$	$\hat{\Sigma}_1=\begin{vmatrix} 1.5 & 0 \\ 0 & 0.48 \end{vmatrix}$	$\hat{P}(C_1)=0.48$
Classe C_2	$\hat{\underline{X}}_2=[8.7,\ 5.41]^T$	$\hat{\Sigma}_2=\begin{vmatrix} 1.5 & 0 \\ 0 & 0.9 \end{vmatrix}$	$\hat{P}(C_2)=0.52$

Tableau 4

On constate que, quelque soit la méthode d'estimation utilisée, les valeurs des paramètres statistiques obtenues après filtrage sont proches des valeurs exactes caractérisant le mélange.



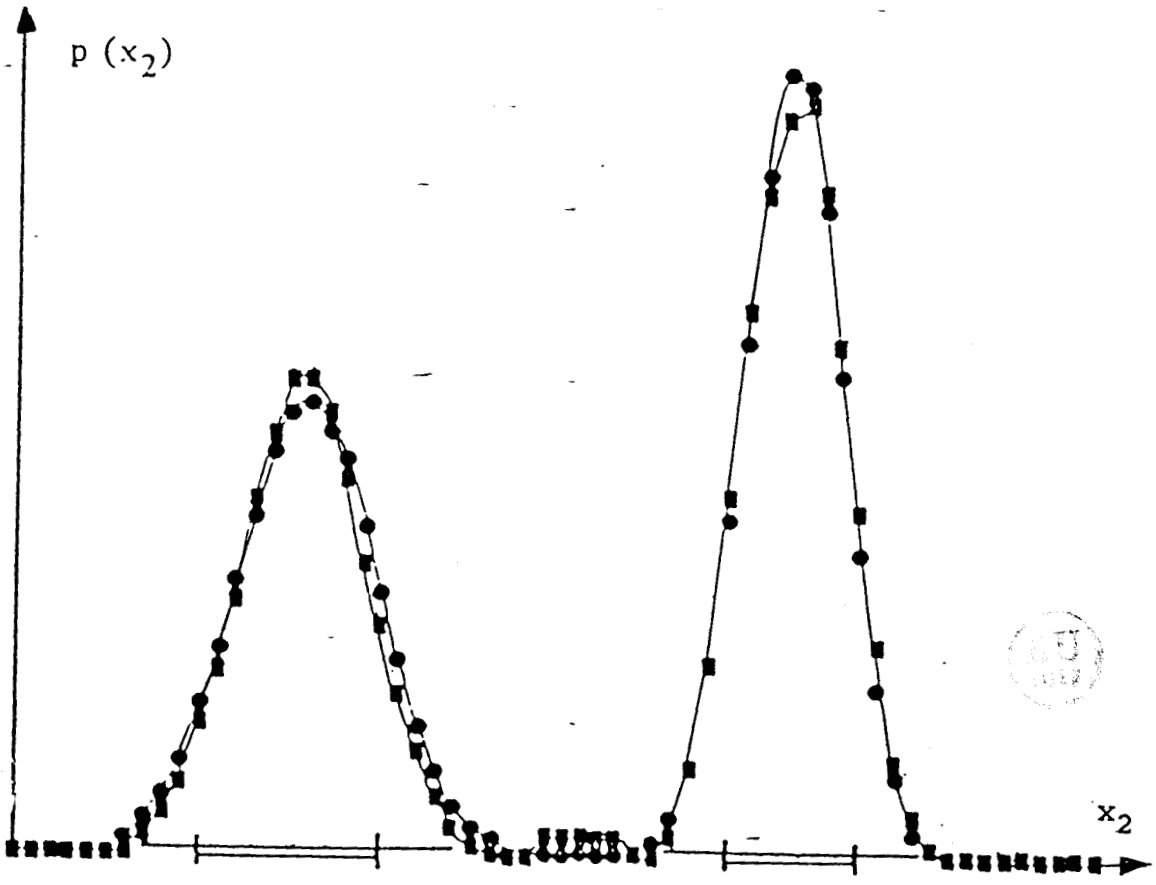
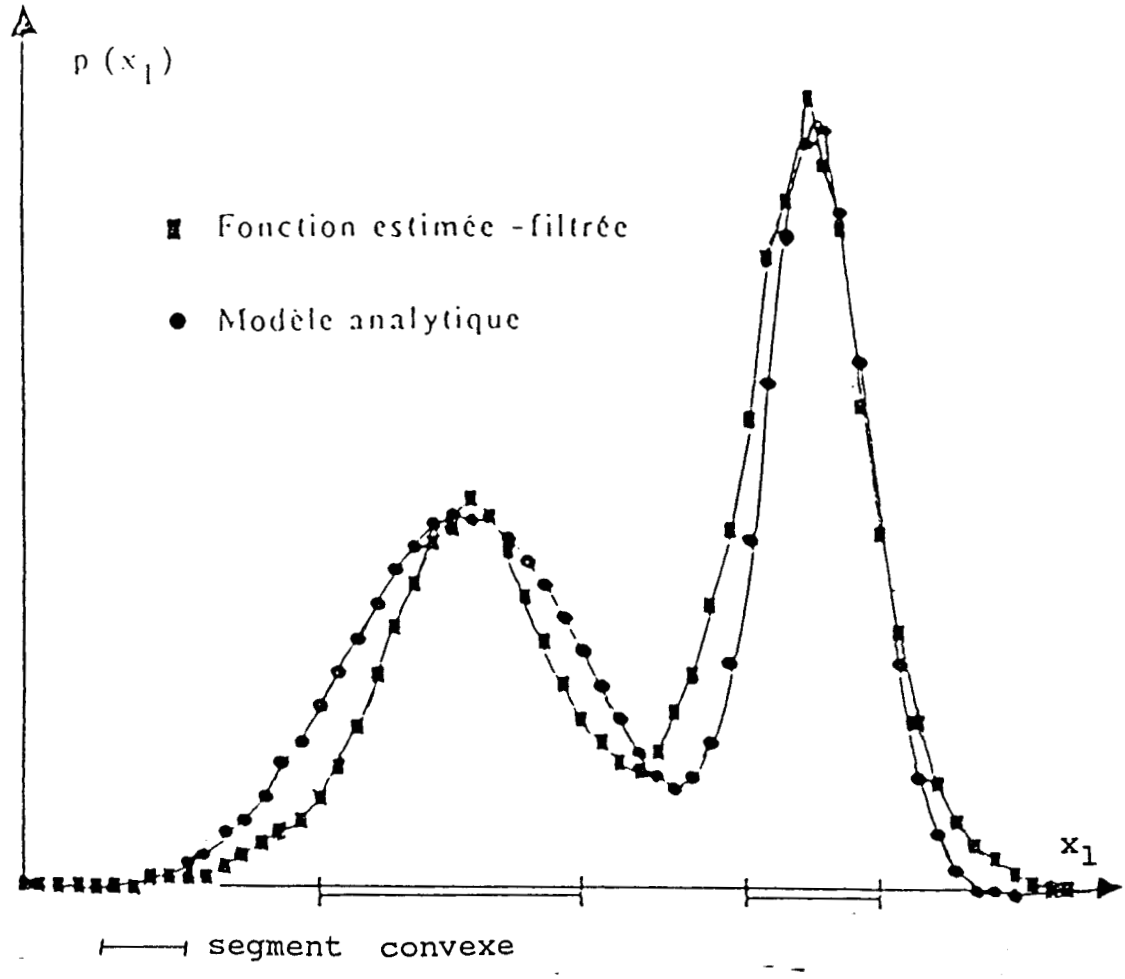


Figure 4(a) : Estimation du mélange par la méthode du noyau après filtrage.



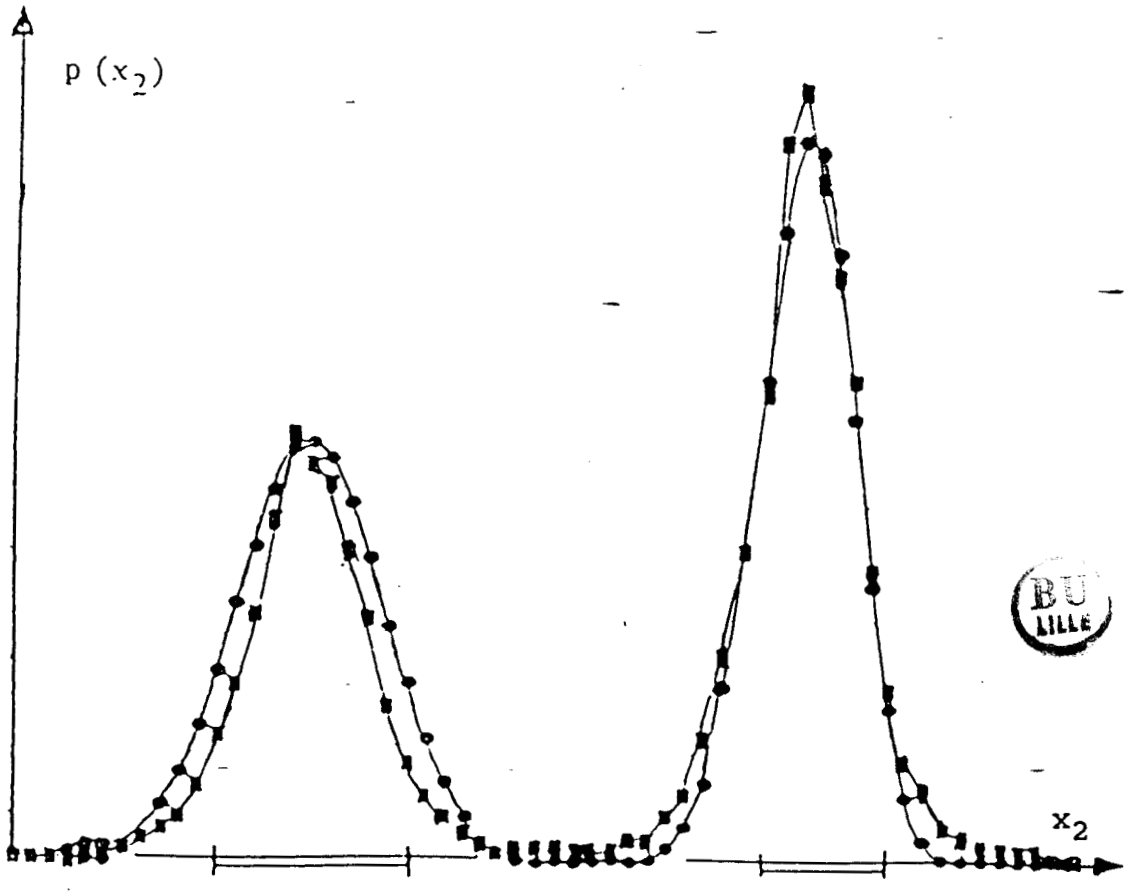


Figure 4(b) : Estimation du mélange par la méthode des plus proches voisins après filtrage.

Dans la pratique, on ne dispose pas des modèles analytiques des fonctions de densité marginale. Dans le cadre de l'exploration de la structure d'un échantillon totalement inconnu de taille Q et de dimension n, le choix de la largeur de la fenêtre de Parzen ou du nombre de plus proches voisins, ainsi que l'ajustement de la largeur de la fenêtre de l'opérateur de filtrage non linéaire doivent être laissés à l'initiative de l'analyste.

Nous avons montré, au chapitre III que les résultats obtenus de l'estimation-filtrage sont très robustes par rapport à l'ajustement de ces différents paramètres. L'analyste, guidé par la visualisation des fonctions estimées, puis estimées-filtrées saura donc effectuer des choix judicieux.

Cette possibilité d'ajustement interactif des paramètres constitue l'un des grands avantages de la méthode d'analyse des fonctions de densité marginale. Une telle approche est en effet impossible lorsqu'on utilise la fonction de densité de probabilité multidimensionnelle sous-jacente à la distribution des observations.

Il importe cependant de remarquer qu'il existe des cas où les composantes multivariables sont nettement séparées dans l'espace de représentation des données alors que les fonctions de densité marginale se chevauchent au point d'être indiscernables par analyse de convexité. Pour pallier cet inconvénient, on procède à une analyse globale de l'échantillon suivie d'une analyse locale.

IV - 3 - ANALYSE GLOBALE - ANALYSE LOCALE.

En général, lorsque les composantes qui constituent le mélange sont nettement séparées, ou bien présentent un faible degré de chevauchement, l'analyse de convexité des fonctions de densité marginale permet d'identifier correctement le mélange .

Mais, si les composantes se chevauchent au niveau des fonctions de densité marginale, nous pouvons obtenir par cette analyse des segments convexes qui peuvent être le résultat de la superposition des contributions de deux ou plusieurs composantes (cf. figure 5).

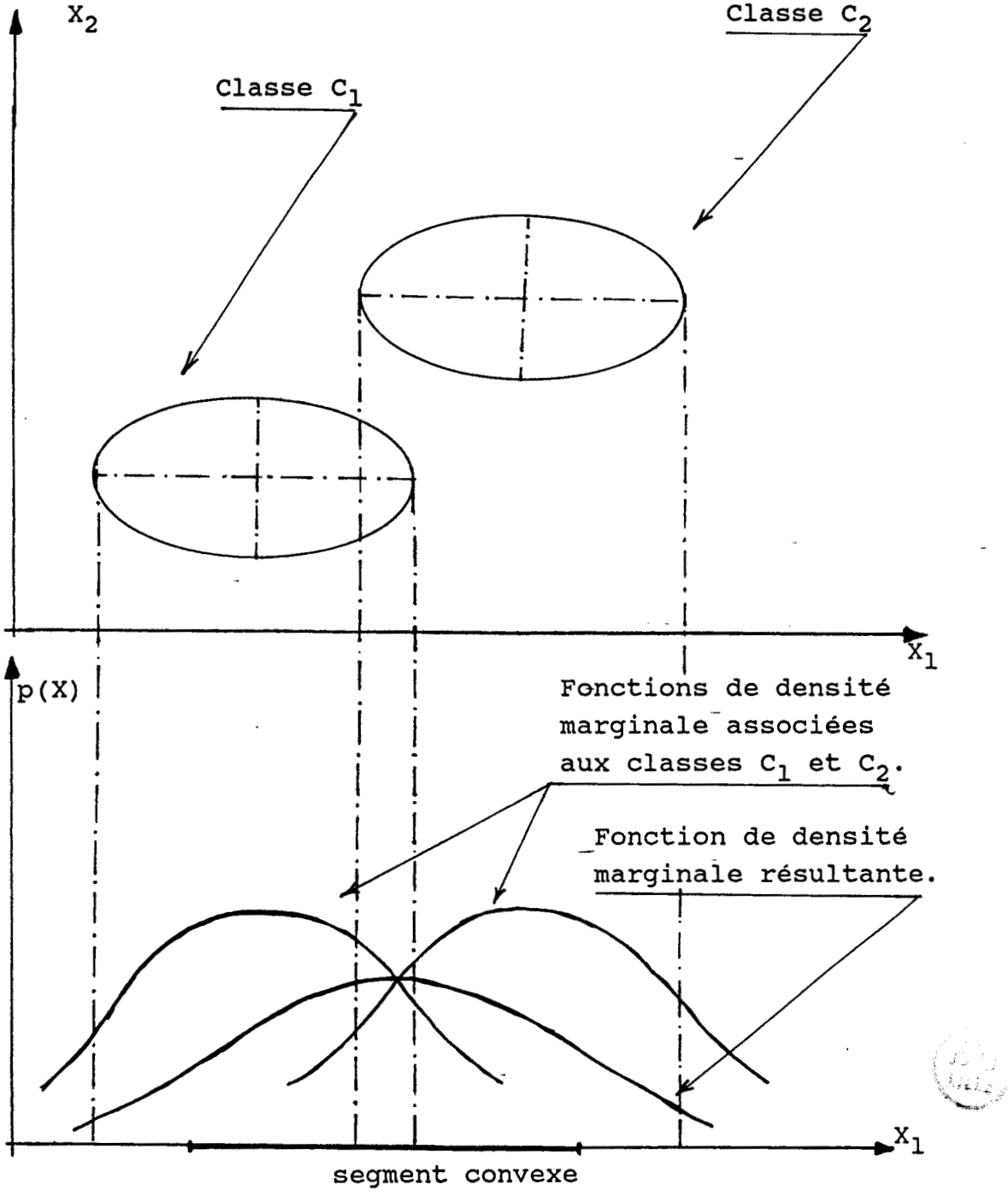


Figure 5: L'unique segment convexe de la fonction de densité marginale résultante ne permet pas de détecter la présence des deux composantes.

Pour éviter une dégradation des performances dans de telles situations, on fait appel à une analyse locale de la structure des données dans chacune des régions de décision R_k k=1,2,...,K mises en évidence par l'analyse globale de l'échantillon.

L'analyse locale a été utilisée en classification automatique afin d'améliorer la qualité des résultats [MEI.72], [NOR.73]. Il a été montré que ces techniques locales qui peuvent faire disparaître des classes ou en faire apparaître de nouvelles, sont en général itératives [EIG.74].

Pour ce faire, on applique l'analyse de convexité des fonctions de densité marginale aux observations situées dans chaque région R_k, k=1,2,...,K. Cette analyse peut faire apparaître des classes dans une région où l'itération précédente ne permettait pas de les détecter. A partir des segments convexes des densités marginales, on obtient de nouveaux domaines caractéristiques.

A chaque itération, ces domaines nous permettent de calculer de nouvelles fonctions de décision qui définissent de nouvelles régions de décision et ainsi de suite. Cette procédure itérative est arrêtée s'il y a stabilisation des résultats de la classification.

Dans le cas contraire, lorsque des surfaces de décision oscillent de part et d'autres de certaines observations l'opérateur peut arrêter la procédure itérative, lorsqu'il constate que les variations du nombre d'observations à reclasser à chaque itération tendent à se stabiliser [POS.83a,b].

Les performances de cette procédure de classification par analyse itérative ont été déjà étudiées sur des données artificielles et sur des problèmes de reconnaissance des formes en dendrométrie [MHI.84]; [MHI.83]; [POS.83a,b]. mais uniquement pour des échantillons de taille importante.

Cette méthode, qui a fait ses preuves sur des échantillons de forte dimension en utilisant les estimateurs brutes des fonctions de densité marginale de probabilité est directement utilisable pour les échantillons de petite dimension, à condition de s'appuyer sur l'analyse de la convexité des fonctions estimées-filtrées.

Il faut cependant remarquer que dans la pratique, il existe des cas où, malgré la mise en oeuvre de la procédure d'analyse locale, la méthode d'identification des mélanges gaussiens présentée conduit à des résultats erronés.

Après avoir présenté, le type de situation où la méthode nécessite une séparation de groupements mis en évidence par une analyse globale, nous exposons, au chapitre suivant, une procédure de fusionnement destinée à pallier un autre type de dégradation des performances de la méthode.

CHAPITRE V

FUSIONNEMENT DES GROUPEMENTS D'OBSERVATIONS.

CHAPITRE V

FUSIONNEMENT DES GROUPEMENTS D'OBSERVATIONS

V - 1 - INTRODUCTION

Grâce au test de convexité des fonctions de densité marginale estimées-filtrées et aux techniques d'analyse globale suivies des techniques d'analyse locale, nous avons vu qu'on pouvait identifier un mélange et déterminer des valeurs approchées de ses paramètres statistiques.

Il peut toutefois arriver que la technique d'analyse du mélange proposée assigne des observations à deux groupements qui ne constituent en fait qu'une seule classe.

Dans de telles situations la procédure itérative locale exposée au chapitre précédent ne peut corriger efficacement les résultats dégradés de l'analyse globale.

L'exemple d'échantillon bidimensionnel de la figure 1 montre comment les segments convexes des fonctions de densité marginale, obtenus à la suite de l'analyse globale, définissent quatre domaines caractéristiques alors qu'il n'y a que trois classes dans l'exemple étudié.

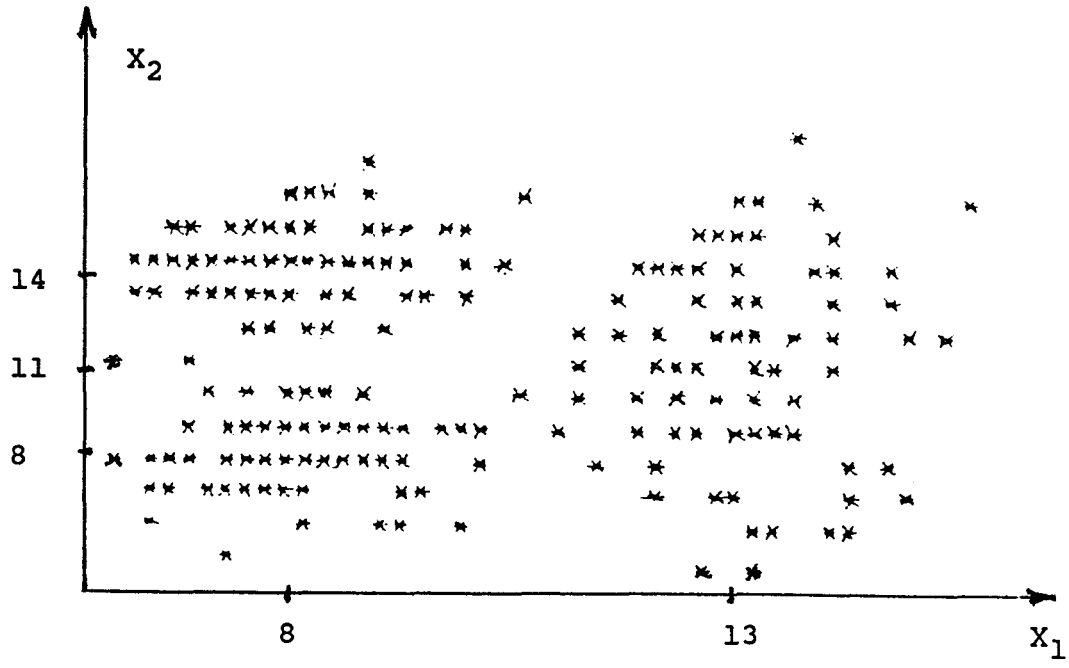


Figure 1: Représentation graphique d'un échantillon bidimensionnel à 3 composantes.

L'analyse de convexité des fonctions de densité marginale de probabilité estimées-filtrées par l'opérateur de filtrage non linéaire proposé met en évidence, dans le cas de la figure 2(a), deux segments convexes suivant chaque axe. La procédure classique du produit euclidien détermine quatre domaines caractéristiques D_k $k = 1,2,..,4$ qui définissent quatre groupements d'observations (cf. figure 2(b)) alors qu'initialement l'échantillon ne présentait que trois composantes.

La procédure d'analyse globale a aboutit à séparer une composante unique en deux groupements distincts.

Afin d'assurer le succès de l'approche étudiée dans tous ces cas de figure, il importe de pouvoir déceler de telles situations qui doivent conduire à un fusionnement de groupements non significatifs individuellement.

Nous proposons une méthode pour fusionner ces classes non significatives qui peuvent être obtenues par analyse globale des fonctions de densité marginale de probabilité.

Figure 2: Résultat de l'analyse globale pour l'exemple d'échantillon bidimentionnel à 3 composantes.

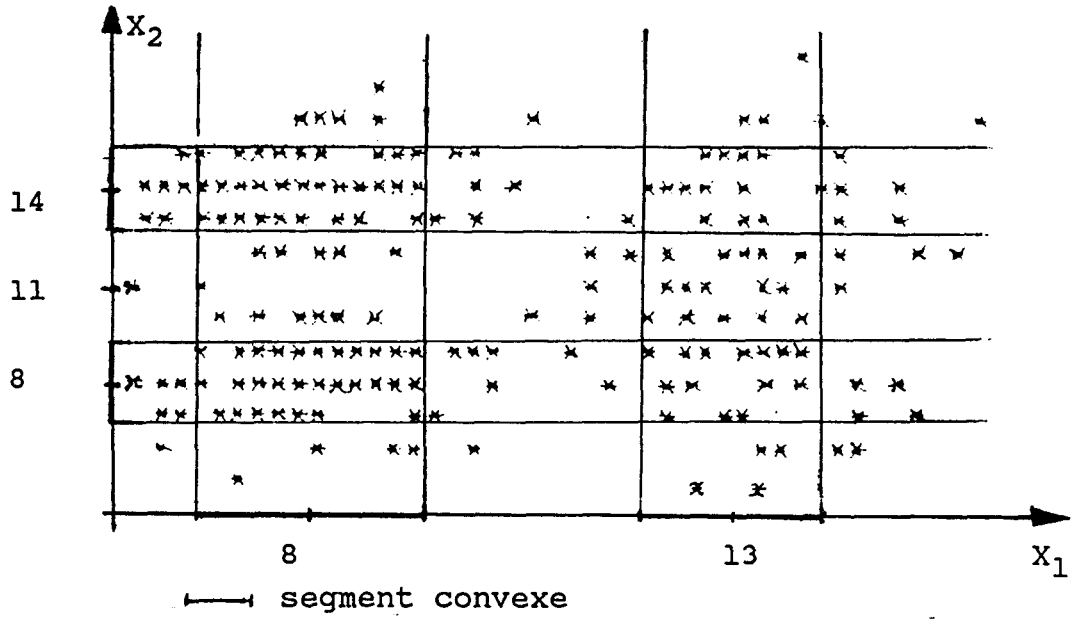


Figure 2(a): Détermination des segments convexes des fonctions de densité marginale.

(▼,✱) : Classes correspondant effectivement à un mode.

(■,●) : Classes devant être fusionnées.

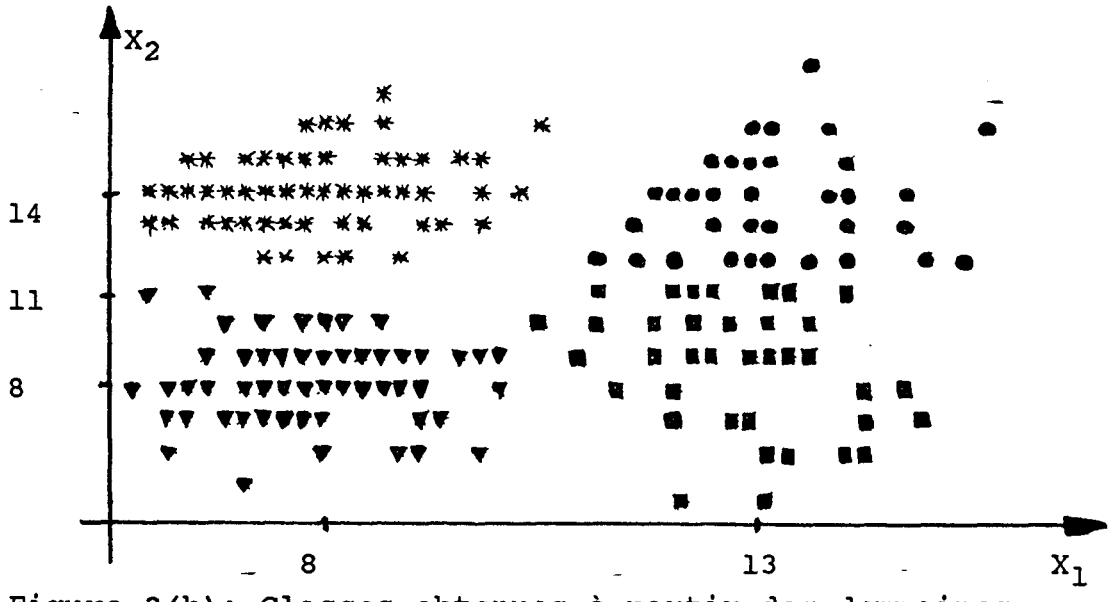


Figure 2(b): Classes obtenues à partir des dommaines caractéristiques.

V - 2 PRINCIPE DE FUSIONNEMENT.

La procédure de fusionnement est basée sur une quantification de l'adéquation des groupements obtenus aux modèles gaussiens.

Plus précisément, pour chaque groupement résultant de l'analyse globale, on estime le vecteur moyenne et la matrice de covariance par la méthode du maximum de vraissemblance [KAZ.77]. On estime ensuite les fonctions de densité marginale de probabilité associées à ce groupement.

On calcule alors l'erreur quadratique entre chaque fonction de densité marginale de probabilité estimée et son modèle gaussien obtenu à partir des valeurs des paramètres du vecteur moyenne et de la matrice de covariance.

Les valeurs des erreurs quadratiques ainsi obtenues permettent de confirmer ou infirmer les résultats de l'analyse globale. Cette décision est basée sur une comparaison de l'ensemble des erreurs quadratiques associées à toutes les classes. Toute erreur dépassant un seuil adapté au problème traité entraine une procédure de fusionnement du groupement concerné avec son ou ses voisins.

V - 3 TEST DE NORMALITE

Le problème consiste maintenant à tester si le groupement associé à chacun des domaines caractéristiques D_K obtenus constitue effectivement une classe gaussienne.

Pour ce faire, on suppose temporairement que chaque groupement étudié est le résultat d'une distribution normale d'observations. Il s'agit alors de vérifier si les fonctions de densité marginale de probabilité estimées-filtrées à partir des observations constituant chaque groupement suivent effectivement le modèle normal correspondant.

Si tel est le cas, le groupement est réellement une classe gaussienne. Dans le cas contraire, le groupement est constitué d'une partie de composante du mélange et doit être fusionné avec le ou les groupements manquants.

Sous l'hypothèse normale, les fusionnements seront déclenchés en fonction de l'étude des erreurs quadratiques qui existent entre les fonctions estimées-filtrées et leurs modèles analytiques.

V - 4 CODAGE DES GROUPEMENTS D'OBSERVATIONS.

On attribut, à chacun des modes détectés sur chaque axe, un numéro d'ordre. Cette numérotation est donnée par ordre croissant de l'origine vers l'infini, le mode le plus proche de l'origine portant le numéro 1, le suivant 2, ..., etc.

On pourra ainsi repérer tout groupement par les numéros d'ordre des modes des n fonctions de densité marginale de probabilité qui ont permis de le détecter. Plus précisément si $g_{i,m}$ est le numéro d'ordre, sur l'axe i, du mode associé au groupement G_m, on peut représenter chacun des m (m= 1,2,...,M) groupements d'observations (cf. figure 3) par un vecteur sous la forme:

$$G_m = [g_{1,m}, g_{2,m}, \cdots, g_{n,m}]^T$$

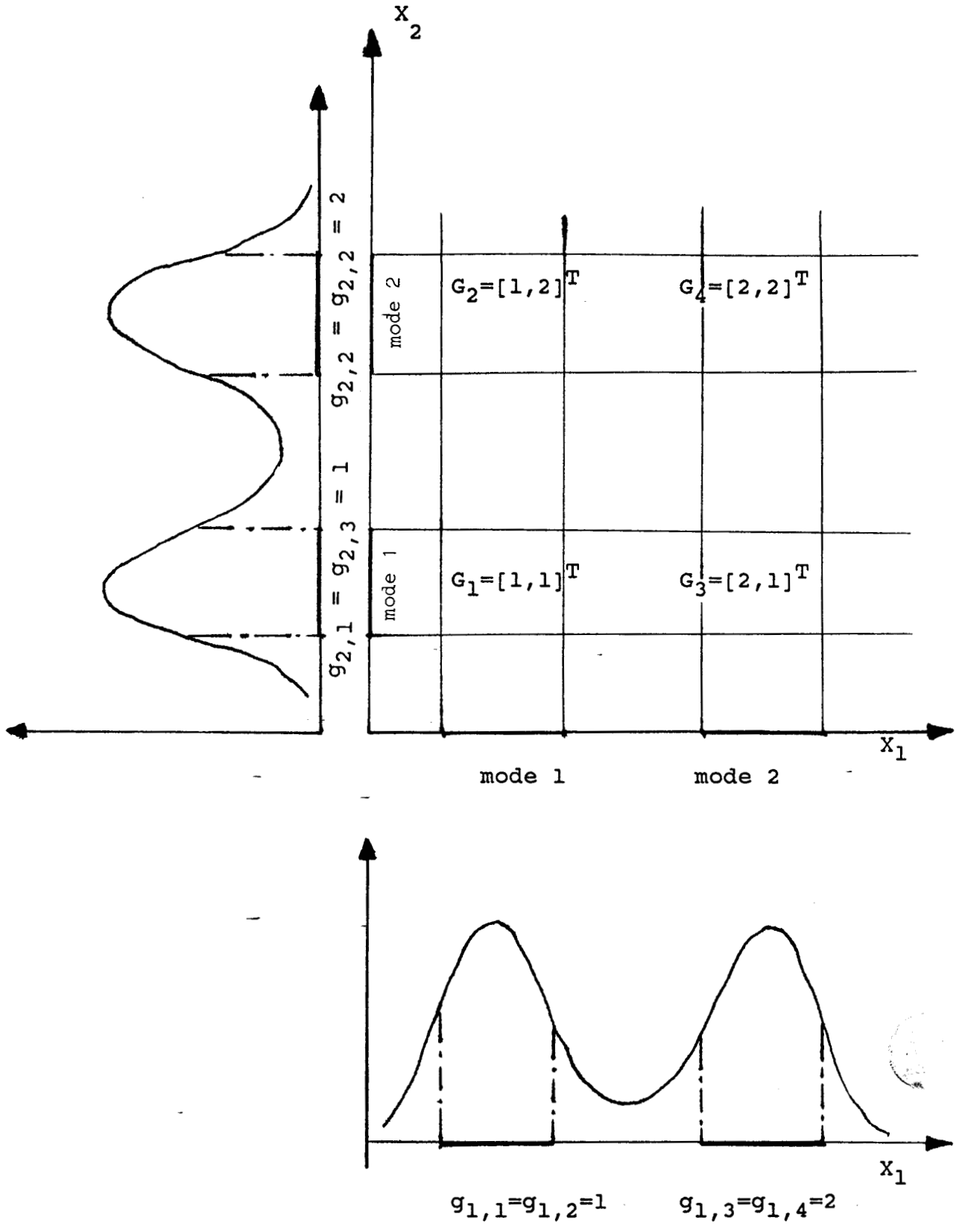


Figure 3 : Représentation des groupements d'observations.

V - 5 DETERMINATION DES GROUPEMENTS DE REFERENCE.

L'étude qui suit doit être conduite sur chacun des axes de l'espace de représentation. Pour simplifier l'exposé, on ne considère dans ce qui suit, qu'un seul axe, l'axe d'indice i.

Pour chaque groupement d'observations G_m, m=1, 2,...,M, on compare sa fonction de densité marginale de probabilité estimée $\hat{p}_m(x_i)$ et son modèle analytique $p_m(x_i)$ obtenu à partir de la moyenne et de la variance des observations contenues dans G_m.

Soit $E_{i,m}$ l'erreur quadratique entre la fonction de densité marginale et le modèle analytique correspondant. On compare alors, suivant chaque axe i, les valeurs des erreurs quadratiques $E_{i,m}$ m= 1,2,...,M, afin de détecter le groupement G_m dont la fonction de densité marginale de probabilité sur l'axe i se rapproche le plus d'une distribution normale. Soit $G^*_m(i)$ ce groupement.

Pour ce faire, on étudie, pour chacun des groupements d'observations G_m, les variations de l'erreur quadratique $E_{i,m}$ en fonction de la largeur de la fenêtre de réglage de l'estimateur dans le cas de la méthode du noyau ou bien en fonction du nombre des plus proches voisins dans le cas de la méthode d'estimation des plus proches voisins. Pour chaque groupement G_m on retient, par la suite, la valeur pour laquelle l'erreur quadratique $E_{i,m}$ est minimale.

On compare ensuite, toutes les valeurs minimales associées aux M groupements et on détecte leur minimum $E^*_{i,m}$. Ce minimum minimorum de ces erreurs calculées, sert de référence pour l'ajustement. Soit $G^*_{m*}(i)$ le groupement correspondant et soit $h^*_{m*}(i)$ ou $k^*_{Q,m*}(i)$ les paramètres optimaux à ajuster des estimateurs utilisés.

A partir de la largeur de la fenêtre de Parzen ou bien de la valeur du nombre de plus proches voisins, retenue pour estimer la fonction de densité marginale de probabilité du groupement de référence $G^*_{m*}(i)$, nous ajustons les largeurs des fenêtres de réglage de l'estimateur ou du nombre de plus proches voisins, en respectant les regles de convergence présentées au chapitre II paragraphe (II-2), afin de comparer sous des conditions identiques, les erreurs quadratiques associées aux fonctions de densité marginale de tous les groupements restants G_m, $m=1,2,...,M$, $m \neq m^*$.

V - 6 CALCUL DES ERREURS QUADRATIQUES.

Soit $Q(G_m)$, la taille du groupement $G_m=[g_{1,m}, g_{2,m}, \ldots, g_{i,m}]^T$. Nous ajustons la largeur des fenêtres de Parzen pour les groupements d'observations G_m, $m \neq m^*$, à partir de la relation (1) ou bien le nombre de plus proches voisins à partir de la relation (2):

$$h_m(i) = h_0(i) \,/\, \sqrt{Q(G_m)} \qquad (1) \quad \text{(cf.chapitre II)}$$

$$k_{Q,m}(i) = k(i) \,/\, \sqrt{Q(G_m)} \qquad (2) \quad " \quad "$$

la valeur des paramètres $h_0(i)$ et $k(i)$ est déterminée en considérant les relations-précédentes pour le groupement de référence G^*_{m*} avec:

$$h_0(i) = h^*_{m*}(i).\sqrt{Q(G^*_{m*}(i))} \qquad (3)$$

et

$$k(i) = k^*_{Q,m*}(i).\sqrt{Q(G^*_{m*}(i))} \qquad (4)$$

La largeur de la fenêtre de Parzen pour estimer la fonction de densité marginale de probabilité pour chacun des groupements d'observations G_m et suivant l'axe i $(i=1,2,...n)$ est calculée à partir de la relation:

$$h_m(i) = h^*_{m*}(i).\sqrt{Q(G^*_{m*}(i))} \,/\, \sqrt{Q(G_m)} \qquad (5)$$

De même pour la valeur du nombre de plus proches voisins. Celle-ci est calculée à partir de la relation suivante:

$$k_{Q,m}(i) = k^*_{Q,m*}(i) \sqrt{Q(G^*_{m*}(i))} / \sqrt{Q(G_m)} \qquad (6)$$

C'est à partir de la valeur des paramètres $h_m(i)$ et $k_{Q,m}(i)$ ainsi calculées, qu'on pourra déterminer les erreurs quadratiques $E'_{i,m}$ entre les fonctions de densité marginale associées aux groupements G_m et leurs modèles analytiques. On considérera ces erreurs pour déclencher éventuellement le fusionnement des groupements d'observations.

V - 7 SEUIL DE DETECTION DE FUSIONNEMENT.

Pour déterminer les groupements d'observations à fusionner, nous comparons les erreurs quadratiques qui viennent d'être calculées à celles associées au groupement de référence $G^*_{m*}(i)$. Les groupements à fusionner sont ceux dont l'erreur quadratique $E'_{i,m}$ pour au moins un attribut vérifie la relation:

$$E'_{i,m} \geq \alpha E^*_{i,m*} \qquad (7)$$

où α est un seuil à ajuster.

V - 8 ALGORITHME D'EXPLOITATION

Si une erreur $E'_{i,m}$ correspondant au groupement d'observations $G_m = [g_{1,m}\ g_{2,m}\ \cdots\ g_{n,m}]^T$ m= 1,2,...,M vérifie la relation (7), nous passons à la deuxième étape qui consiste à comparer les erreurs quadratiques à gauche et à droite du maximum de la fonction de densité marginale de probabilité estimée $\hat{p}_m(x_i)$. Il s'agit de :

a/ Déterminer le maximum de la fonction de densité marginale de probabilité $\hat{p}_m(x_i)$ pour les groupements d'observations à fusionner.

b/ Déterminer l'erreur quadratique entre la fonction de densité marginale de probabilité $\hat{p}_m(x_i)$ et le modèle analytique $p_m(x_i)$ correspondant, sur la partie située à gauche de ce maximum. Soit $E^g_{i,m}$ cette erreur (cf. figure 4).

c/ Déterminer l'erreur quadratique entre la fonction de densité marginale de probabilité $\hat{p}_m(x_i)$ et le modèle - analytique $p_m(x_i)$ correspondant, sur la partie située à droite de ce maximum. Soit $E^d_{i,m}$ cette erreur (cf. figure 4)

d/ Comparer les deux erreurs quadratiques $E^g_{i,m}$ et $E^d_{i,m}$.

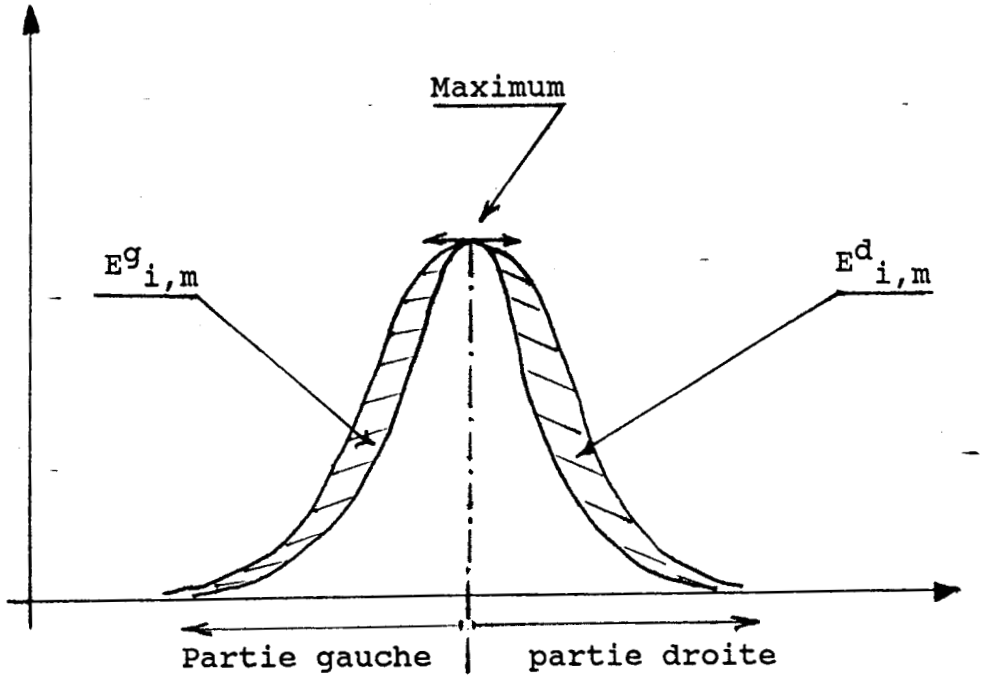


Figure 4: Erreur quadratique à gauche et à droite.

Si pour un groupement incomplet on a:

$$E^g_{i,m} \geq \beta \; E^d_{i,m} \qquad (8)$$

où ß est un deuxième seuil à ajuster, alors on peut dire que la partie manquante au groupement d'observations est située à gauche de ce groupement. Le numéro d'ordre de ce dernier, suivant l'axe où il est incomplet, est affecté d'une étiquette "-" comme sur l'exemple ci-après.

$$G_m = [g_{1,m}, g^-_{2,m}, \cdots, g_{n,m}]^T$$

Afin de retrouver la partie manquante au groupement d'observations incomplet G_m, il faut trouver l'autre groupement incomplet $G_{m'}$ tel que ses erreurs quadratiques de part et d'autre du maximum vérifient la relation:

$$E^d_{i,m'} \geq ß \; E^g_{i,m'} \qquad (9)$$

Cette recherche est réalisée en testant dans l'ordre croissant les groupements qui se succèdent sur l'axe considéré.

De même, si pour un groupement G_m on trouve que:

$$E^d_{i,m-} \geq ß \; E^g_{i,m} \qquad (10)$$

ces groupements auront un numéro d'ordre suivant l'axe i étiqueté d'un signe +. Par exemple:

$$G_m = [g_{1,m}, g^+_{2,m}, \cdots, g_{n,m}]^T$$

Pour retrouver la partie manquante à ce groupement d'observations G_m, on teste les groupements incomplets $G_{m'}$ dont les erreurs quadratiques $E^g_{i,m'}$ et $E^d_{i,m'}$ vérifient la relation suivante:

$$E^g_{i,m'} \geq ß \; E^d_{i,m'} \qquad (11)$$

V - 9 ALGORITHME DE FUSIONNEMENT DE GROUPEMENTS D'OBSERVATIONS.

Généralement, la partie manquante à un groupement d'observations G_m dont le numéro d'ordre $g_{i,m}$ est affecté d'un signe -, est le groupement incomplet $G_{m'}$, dont le numéro d'ordre, sur le même axe est l'entier immédiatement inférieur:

$$g_{i,m'} = (g^{-}_{i,m} - 1),$$

autrement dit, au groupement incomplet:

$$G_m = [g_{1,m}, g^{-}_{2,m}, \cdots, g_{n,m}]^T$$

on fusionne le groupement incomplet:

$$G_{m'} = [g_{1,m'}, g_{2,m'}, \cdots, g_{n,m'}]^T$$

on aura: $g_{2,m'} = g^{-}_{2,m} - 1.$

De manière analogue, la partie manquante à un groupement d'observations G_m dont le numéro d'ordre $g_{i,m}$ est affecté d'une étiquette " + ", est le groupement incomplet $G_{m'}$ dont le numéro d'ordre $g_{i,m'}$ sur le même axe est l'entier immédiatement supérieur:

$$g_{i,m'} = g^{+}_{i,m} + 1$$

autrement dit, le groupement incomplet:

$$G_m = [g_{1,m}, g^{+}_{2,m}, \cdots, g_{n,m}]^T$$

on fusionne le groupement incomplet:

$$G_{m'} = [g_{1,m'}, g_{2,m'}, \cdots, g_{n,m'}]^T$$

on aura: $g_{2,m'} = g^{+}_{2,m} + 1$.

Le fusionnement des groupements incomplets affectés des étiquettes + ou - se réalise de la manière décrite par l'algorithme suivant. Au départ du traitement, tous les groupements sont présentés de la manière suivante:

$$G_m = [g_{1,m}, g_{2,m}, \cdots, g_{n,m}]^T$$

sauf pour les groupements incomplets. Ces derniers sont affectés d'un signe + ou - selon la partie manquante et l'axe considéré. Par exemple:

$$G_m = [g_{1,m}, g^{\pm}_{2,m}, \cdots, g_{n,m}]^T$$

pour l'axe i=2.

Les fusionnements des groupements d'observations incomplets G_m sont réalisés comme suit:

a/ On ordonne, suivant l'axe i, les numéros d'ordre $g_{i,m}$ des groupements d'observations G_m.

b/ Tous les groupements d'observations incomplets G_m de numéro d'ordre affecté d'un signe +, soit $g^+_{i,m}$, sont fusionnés avec les groupements d'observations $G_{m'}$ de numéro d'ordre $g^-_{i,m'}$ tel que $g^-_{i,m'} = g^+_{i,m} + 1$. Par exemple soit:

$$G_m = [g_{1,m}, g^+_{2,m}, \cdots, g_{n,m}]^T$$

un groupement incomplet. Celui-ci va être fusionné avec le groupement d'observations $G_{m'}$:

$$G_{m'} = [g_{1,m'}, g^-_{2,m'}, \cdots, g_{n,m'}]^T$$

pour lequel $g^-_{2,m'} = g^+_{2,m} + 1$

Il faut cependant signaler qu'il existe des cas où le numéro d'ordre d'un groupement d'observations G_m, suivant un axe i, est affecté de deux signes + et - c'est à dire d'un numéro d'ordre $g^{\pm}_{i,m}$.

Dans ces cas, le groupement incomplet G_m sera fusionné avec le groupement d'observations $G_{m'}$ qui le précède, de numéro d'ordre $g^+_{i,m'} = (g^-_{i,m} - 1)$ et en même temps, il sera fusionné avec le groupement qui le succède $G_{m''}$, de numéro d'ordre $g^-_{i,m''} = (g^+_{i,m} + 1)$.

On fait ainsi l'exploration des groupements d'observations tout le long de l'axe considéré et ceci pour chacun des axes de représentation.

Nous appliquons ensuite les techniques d'analyse étudiée dans le chapitre II, III et IV aux fonctions de densité marginale de probabilité du groupement d'observations obtenu par fusionnement.

V - 10 VERIFICATION

Connaissant la valeur $h^*_{m*}(i)$ du paramètre de réglage de l'estimateur de Parzen ou bien celle du nombre $k^*_{Q,m*}(i)$ plus proches voisins, nous pouvons calculer les valeurs $h_Q(i)$ ou bien $k(i)$. Sachant aussi que le groupement d'observation obtenu par fusionnement a pour taille, la somme des tailles des groupements qui le constituent, il est possible de déterminer la largeur $h_m(i)$ de la fenêtre de Parzen ou bien $k_Q(i)$ associé à ce groupement.

A partir de ces valeurs, nous estimons par la méthode du noyau ou bien celle des plus proches voisins, les fonctions de densité marginale de probabilité du groupement d'observations fusionné. A ces fonctions on applique l'opérateur de filtrage non linéaire décrit au chapitre III paragraphe III-2, suivi du test de convexité des fonctions de densité afin de déterminer les paramètres statistiques des groupements obtenus.

V - 11 APPLICATION DE LA TECHNIQUE DE FUSIONNEMENT DE GROUPEMENTS D'OBSERVATIONS SUR UN ECHANTILLON BIDIMENSIONNEL

Pour illustrer le comportement de l'algorithme de fusionnement décrit ci-dessus, on considère un exemple bidimensionnel généré artificiellement qui permet de visualiser les résultats obtenus. Cet exemple est constitué de trois composantes équiprobables, contenant chacune 100 observations. Les paramètres statistiques sont résumés dans le tableau 1 suivant:

	Vecteur moyenne	Matrice de covariance	Probabilité a priori
Classe C_1	$\bar{X}_1=[7,8]^T$	$\Sigma_1=\begin{vmatrix}1 & 0\\ 0 & 1\end{vmatrix}$	$P(C_1)=0.33$
Classe C_2	$\bar{X}_2=[8,14]^T$	$\Sigma_2=\begin{vmatrix}1 & 0\\ 0 & 1\end{vmatrix}$	$P(C_2)=0.33$
Classe C_3	$\bar{X}_3=[13,11]^T$	$\Sigma_3=\begin{vmatrix}1 & 0\\ 0 & 3\end{vmatrix}$	$P(C_3)=0.33$

Tableau 1

L'analyse directe de la convexité des fonctions de densité marginale de probabilité a mis en évidence, pour l'exemple bidimensionnel représenté par la figure 1, deux segments convexes sur chacun des axes. La procédure classique du produit euclidien des segments convexes détermine quatre groupements d'observations G_m m=1,2,3,4, tels que:

$$G_1 = [1, 1]^T$$
$$G_2 = [1, 2]^T$$
$$G_3 = [2, 1]^T$$
$$G_4 = [2, 2]T$$

Pour chacun des groupements d'observations et suivant chaque caractère, on étudie l'influence de la largeur $h_m(i)$ de la fenêtre de Parzen sur l'erreur quadratique $E_{i,m}$, m=1, 2,...,M. Les valeurs des erreurs quadratiques obtenues sont résumées dans le tableau 2. A partir de ces valeurs nous pouvons tracer les courbes représentatives (cf. figure 5).

La figure 5 indique les variations de l'erreur quadratique $E_{i,m}$ en fonction des paramètres $h_m(i)$. Ces variations mettent en évidence des valeurs optimales des erreurs quadratiques. Ces dernières sont résumées dans le tableau 3.

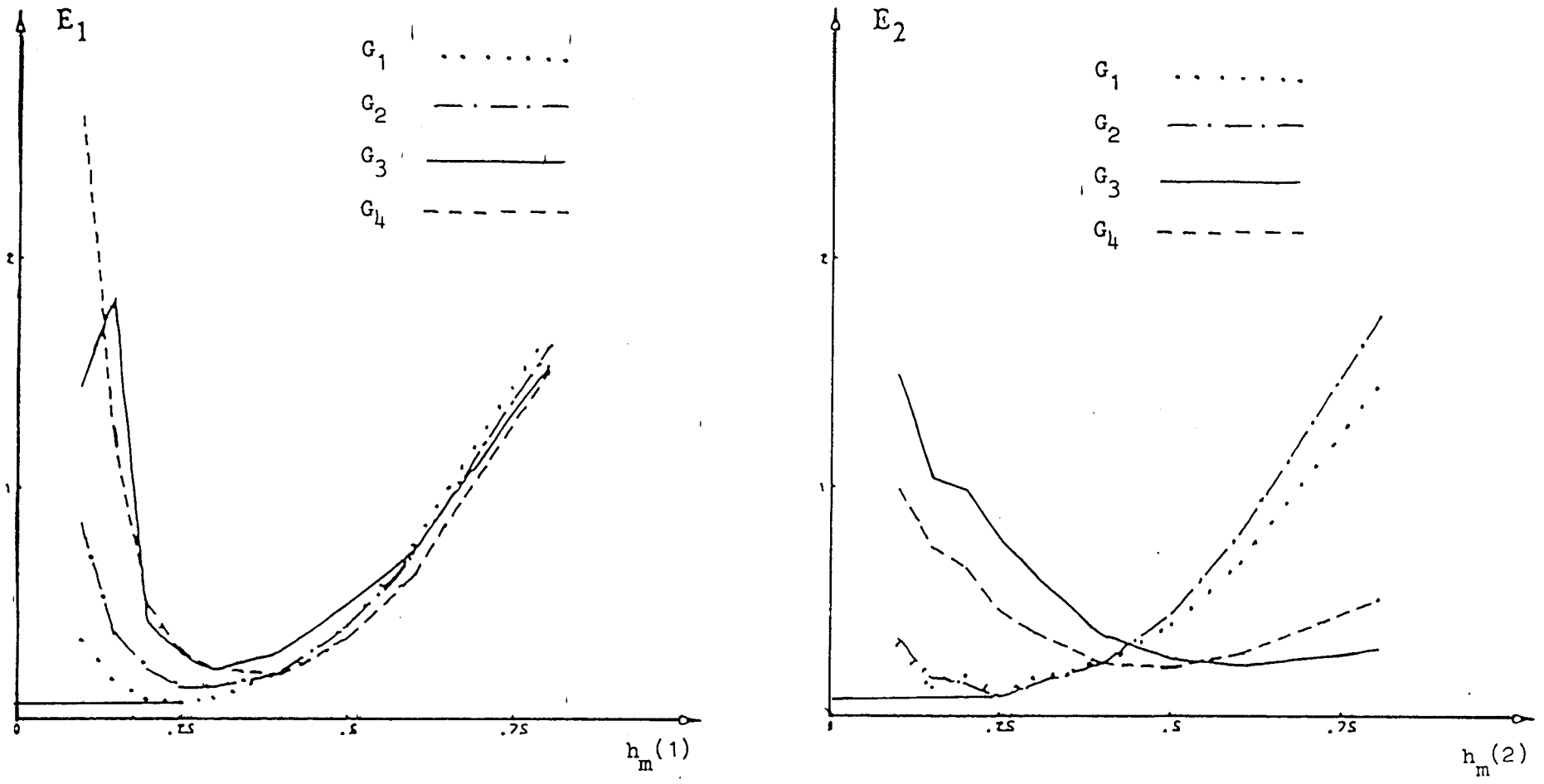


Figure 5 : Variations des erreurs quadratiques en fonction de la largeur de la fenêtre de Parzen $h_m(i)$.

	$h_m(i)$	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.50	0.60	0.80
i=1	G_1	0.0343	0.0171	0.0087	0.0073	0.0096	0.0157	0.0206	0.0424	0.0754	0.1681
	G_2	0.0841	0.0383	0.0219	0.0137	0.0142	0.0174	0.0204	0.0411	0.0715	0.1610
	G_3	0.1442	0.1832	0.0428	0.0281	0.0213	0.0250	0.0298	0.0562	0.0726	0.1532
	G_4	0.2623	0.1223	0.0504	0.0321	0.0222	0.0215	0.0192	0.0363	0.0628	0.1502
i=2	G_1	0.0317	0.0126	0.0173	0.0091	0.0167	0.0198	0.0237	0.0416	0.0682	0.1463
	G_2	0.0346	0.0164	0.0141	0.0085	0.0142	0.0189	0.0234	0.0463	0.0805	0.1751
	G_3	0.1481	0.1032	0.0976	0.0765	0.0616	0.0472	0.0354	0.0254	0.0222	0.0301
	G_4	0.0989	0.0732	0.0643	0.0462	0.0368	0.0341	0.0230	0.0257	0.0281	0.0529

tableau 2.

	$E^*_{1,m}$	$h^*_m(1)$	$E^*_{2,m}$	$h^*_m(2)$
$G_1=[1,1]^T$	0.0073	0.25	0.0091	0.25
$G_2=[1,2]^T$	0.0137	0.25	0.0085	0.25
$G_3=[2,1]^T$	0.0213	0.30	0.0222	0.60
$G_4=[2,2]^T$	0.0192	0.40	0.0230	0.40

Tableau 3.

D'après le tableau 3, nous pouvons déduire le minimum minimorum $E^*_{i,m*}$. Ce dernier est donné par le groupement d'observations G_1 suivant le premier caractère (i=1) et par le groupement G_2 suivant le deuxième caractère (i=2). Autrement dit on a:

pour i=1: $E^*_{1,1} = 0.0073$ et $G^*_{m*}(i) = G_1(1) = [1, 1]T$

pour i=2: $E^*_{2,2} = 0.0085$ et $G^*_{m*}(i) = G_2(2) = [1, 2]T$

Connaissant $Q(G^*_{m*})$, la taille du groupement de référence G^*_{m*}, ainsi que la valeur de $h^*_{m*}(i)$ de la fenêtre de Parzen correspondant au groupement $G^*_{m*}(i)$, suivant l'axe i, nous pouvons ajuster la valeur $h_m(i)$ du paramètre de réglage de l'estimateur pour les groupements d'observations G_m, m=1, 2, 3, 4 à partir de la relation 5 soit:

$$h_m(i) = h^*_{m*}(i) \sqrt{Q(G^*_{m*}(i))} \; / \sqrt{Q(G_m)}$$

C'est à partir de la valeur du paramètre $h_m(i)$, ainsi calculée, qu'on détermine les erreurs quadratiques $E'_{i,m}$ entre la fonction de densité marginale associée à chacun des groupements G_m et leurs modèles analytiques. Les résultats obtenus sont consignés dans le tableau 4.

	$Q(G_m)$	$h_m(1)$	$E'_{1,m}$	$h_m(2)$	$E'_{2,m}$
G_1	99	0.25	0.0073	0.25	0.0091
G_2	101	0.25	0.0137	0.25	0.0085
G_3	45	0.37	0.0298	0.38	0.0354
G_4	55	0.33	0.0222	0.35	0.0341

Tableau 4

Pour chacun des groupements d'observations $G_m(i)$ m= 1,2,3,4. obtenus, nous vérifions la relation (7), soit:

$$E'_{i,m} \geq \alpha E^*_{i,m*} \qquad (7)$$

Nous avons déterminé le seuil α de détection des groupements à fusionner à partir de plusieurs essais réalisés sur differents échantillons générés artificiellement. L'expérience montre qu'avec α=4, on peut obtenir en général de bons résultats.

Pour l'exemple considéré, nous utilisons les informations données par le tableau 4 afin de déduire les groupements d'observations qui vérifient la relation 7. Suivant l'axe i = 1, le problème de fusionnement ne se pose pas, par contre on détecte des groupements d'observations incomplets suivant l'axe i = 2. Ce sont les groupements G_3 et G_4 avec:

$$G_3 = [2, 1]^T \qquad \text{et} \qquad G_4 = [2, 2]^T$$

Pour tout groupement, dont l'erreur quadratique $E'_{i,m}$ répond à la relation (1), on vérifie si les erreurs quadratiques $E^g_{i,m}$ et $E^d_{i,m}$ de ce groupement d'observations sont telles que:

$$E^g_{i,m} \geq \beta \; E^d_{i,m} \qquad (8)$$

L'ajustement de β est plus délicat que celui du seuil α. Les exemples ont montré que les valeurs $\beta \geq 2$ donnent des résultats satisfaisants dans des situations variées.

Pour l'exemple considéré, les valeurs des erreurs quadratiques, situées à gauche et à droite du maximum de la fonction de densité marginale du groupement d'observations incomplet (G_3, G_4) sont:

Pour G_3 on a $E^g_{2,3} = 0.0113$ et $E^d_{2,3} = 0.0241$
Pour G_4 on a $E^g_{2,4} = 0.0227$ et $E^d_{2,4} = 0.0114$

En se basant sur l'algorithme d'exploitation ainsi que celui de fusionnement, nous pouvons représenter les groupements d'observations incomplets de la manière suivante:

$$G_3 = [2, 1^+]^T \quad \text{et} \quad G_4 = [2, 2^-]^T$$

D'après le principe de fusionnement, le groupement manquant à G_2 est le groupement incomplet G_3. On fusionne alors ces deux groupements d'observations et on vérifie si le résultat du fusionnement donne une classe.

Connaissant la largeur $h^*_{m*}(i)$ de la fenêtre de Parzen correspondant à chacun des groupements de référence déduit à partir du tableau 3, on peut calculer la valeur $h_0(i)$. De cela, et sachant aussi que le groupement d'observations obtenu par fusionnement a pour taille, la somme des tailles des groupements qui le constituent, il est possible de déterminer la largeur $h_m(i)$ de la fenêtre de Parzen associée à ce groupement. Ceci est réalisé à partir de la relation suivante:

$$h_m(2) = h_2(2) \sqrt{Q(G_2(2))} \; / \sqrt{Q(G_3(2)) + Q(G_4(2))}$$

$$h_m(2) = 0.25 \sqrt{101} \; / \sqrt{45 + 55}$$

$$h_m(2) = 0.25$$

A partir de cette valeur de la fenêtre de Parzen, on estime par la méthode du noyau, les fonctions de densité marginale de probabilité du groupement fusionné. Les résultats obtenus sont indiqués par la figure 6.

A ces fonctions, on applique l'opérateur de filtrage non linéaire décrit dans le chapitre III. En prenant comme largeur de fenêtre de cet opérateur, la valeur $l = 8$, nous obtenons la figure 7. Cette dernière représente l'estimation obtenue après filtrage.

L'application du test de convexité sur les fonctions de densité estimées-filtrées, met en évidence un segment convexe. La longueur de ce segment est égale au double de l'écart type et son milieu correspond à la valeur moyenne. Les paramètres statistiques de ces composantes obtenues sont résumés dans le tableau 4. On retrouve les composantes initiales (cf. figure 8).

V - 12 CONCLUSION.

Comme l'analyse globale peut dans certains cas séparer une composante unique en deux groupements distincts. L'exemple présenté ci-dessus montre comment la procédure de fusionnement de groupements non significatifs individuellement, permet de déceler de telles situations.

Donc, Grâce à la technique d'analyse globale et la procédure d'estimation-filtrage nous pouvons séparer ou fusionner des groupements d'observations, ce qui nous permet de reconstituer les différentes composantes d'un échantillon de taille réduite.

Nous utilisons ces techniques sur une population d'abeilles afin de faire une classification.

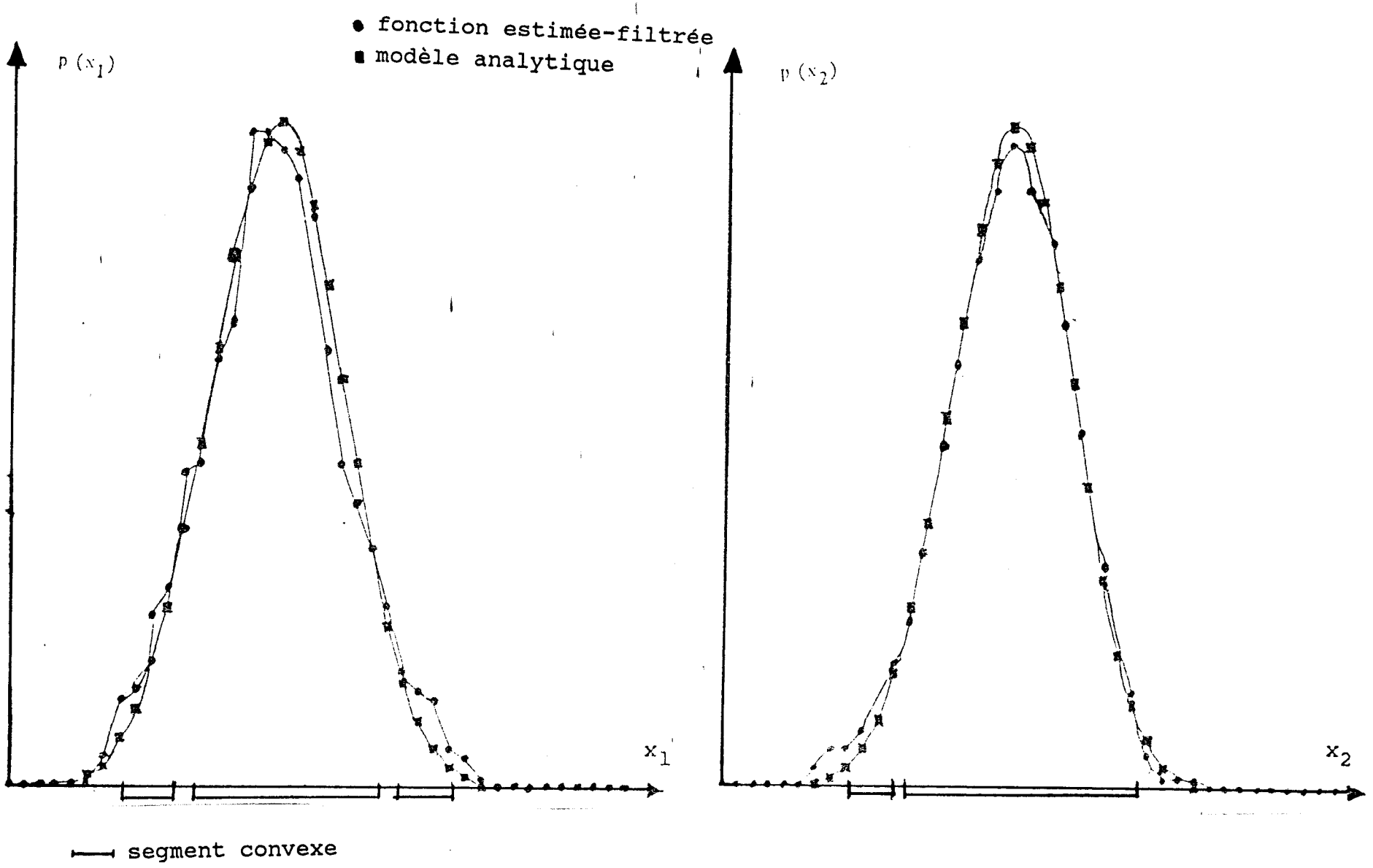


Figure 6: Comparaison des fonctions de densité marginale estimées-filtrées et de leur modèle analytique du groupement d'observations obtenu par fusionnement.

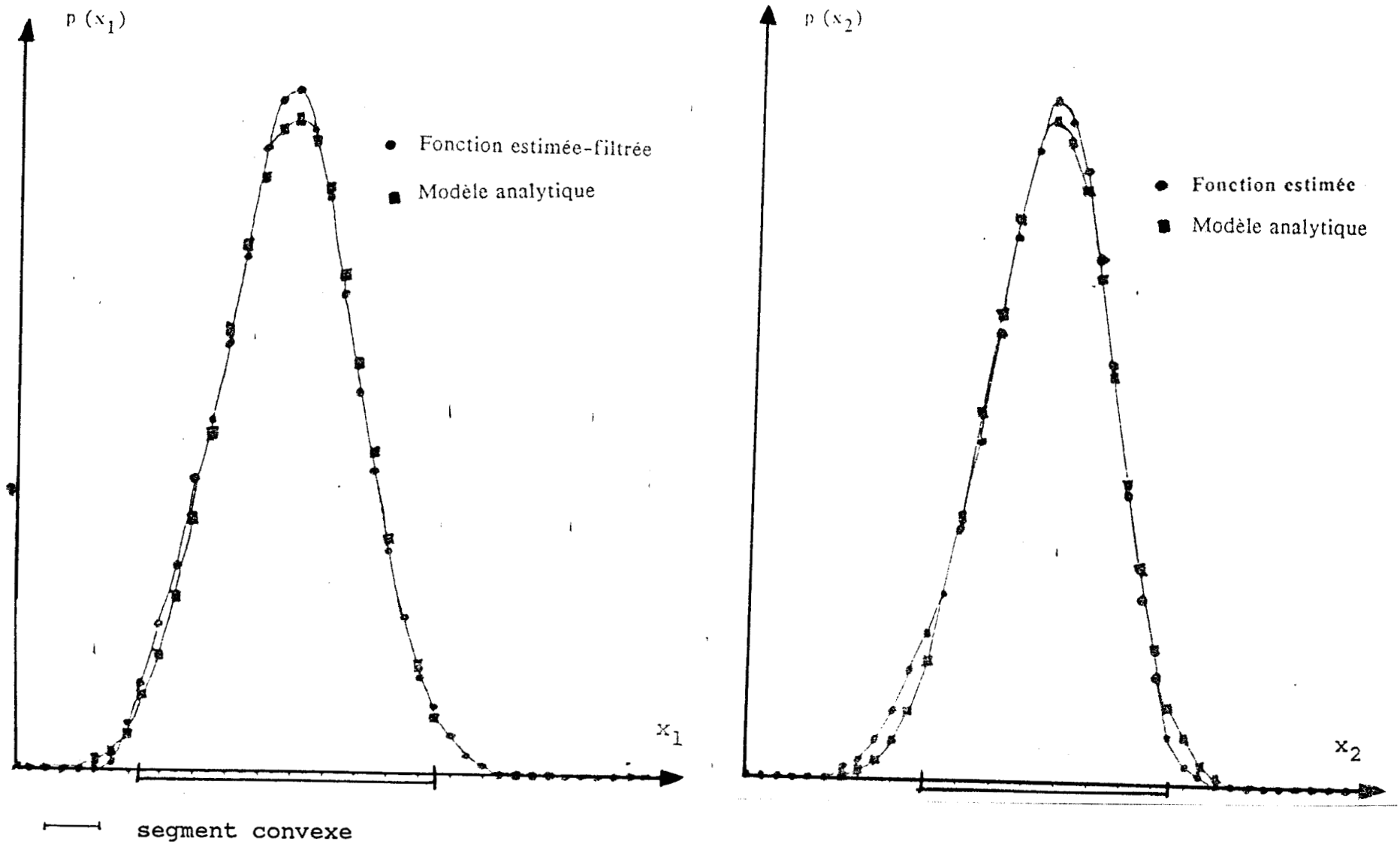


Figure 7: Comparaison des fonctions de densité marginale estimées-filtrées et de leur modèle analytique du groupement d'observations obtenu par fusionnement.

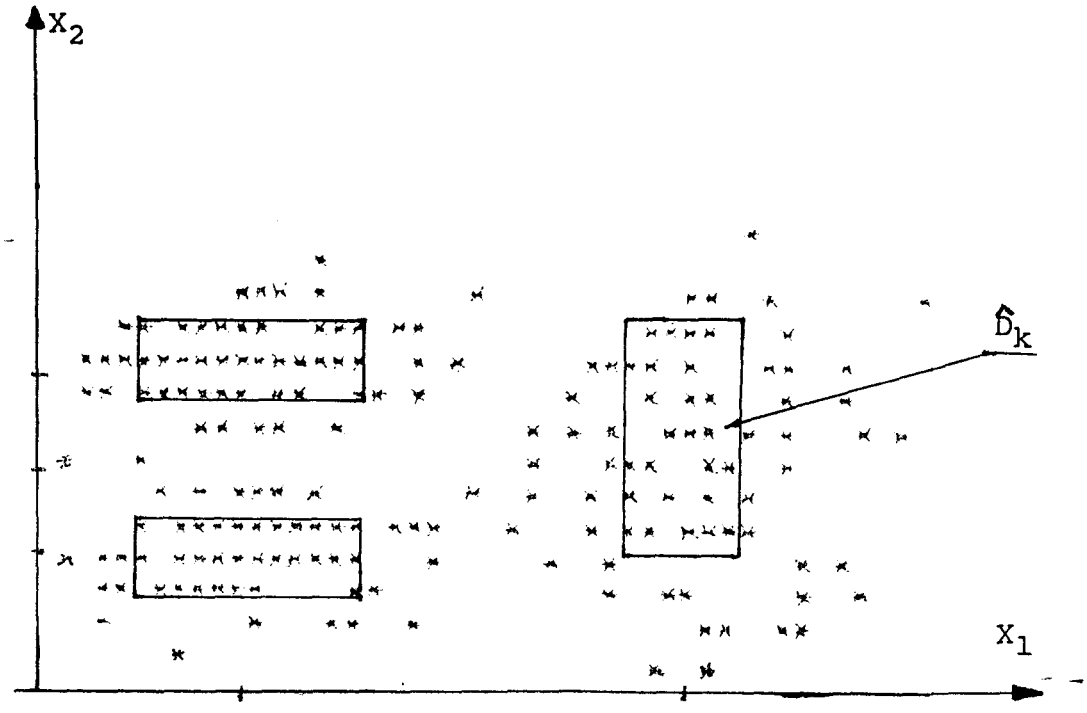


Figure 8: Représentation graphique des observations de l'échantillon et domaines approchés des domaines caractéristiques obtenus.

	Vecteur moyenne	Matrice de covariance	probabilité a priori
Classe C_1	$\bar{X}_1 = \begin{bmatrix} 8 \\ 8 \end{bmatrix}$	$\Sigma_1 = \begin{vmatrix} 1 & 0 \\ 0 & 1 \end{vmatrix}$	$P(C_1)=0.33$
	$\hat{\bar{X}}_1 = \begin{bmatrix} 7.97 \\ 7.99 \end{bmatrix}$	$\hat{\Sigma}_1 = \begin{vmatrix} 1.3 & 0 \\ 0 & 1.08 \end{vmatrix}$	$\hat{P}(C_1)=0.34$
Classe C_2	$\bar{X}_2 = \begin{bmatrix} 8 \\ 14 \end{bmatrix}$	$\Sigma_2 = \begin{vmatrix} 1 & 0 \\ 0 & 1 \end{vmatrix}$	$P(C_2)=0.33$
	$\hat{\bar{X}}_2 = \begin{bmatrix} 7.90 \\ 13.78 \end{bmatrix}$	$\hat{\Sigma}_1 = \begin{vmatrix} 1.05 & 0 \\ 0 & 1.1 \end{vmatrix}$	$\hat{P}(C_2)=0.34$
Classe C_3	$\bar{X}_3 = \begin{bmatrix} 13 \\ 11 \end{bmatrix}$	$\Sigma_3 = \begin{vmatrix} 1 & 0 \\ 0 & 3 \end{vmatrix}$	$P(C_3)=0.33$
	$\hat{\bar{X}}_3 = \begin{bmatrix} 12.96 \\ 11.03 \end{bmatrix}$	$\hat{\Sigma}_3 = \begin{vmatrix} 0.95 & 0 \\ 0 & 2.96 \end{vmatrix}$	$\hat{P}(C_1)=0.32$

Tableau5: Valeurs exactes et approchées des paramètres statistiques de l'exemple d'échantillon à 3 composantes.

CHAPITRE VI

CLASSIFICATION DES ABEILLES

CHAPITRE VI

CLASSIFICATION DES ABEILLES

VI - 1 - INTRODUCTION

Avec l'évolution des techniques apicoles, l'homme est intervenu de plus en plus profondément dans la vie des abeilles. Les apiculteurs se sont aperçus depuis longtemps que les populations d'abeilles des différents pays ou régions du monde ne se ressemblaient pas. La première différence morphologique remarquée a été la couleur des abeilles, car elle est la plus visible à l'oeil nu. Mais peu à peu avec l'évolution des connaissances scientifiques, les tentatives de sélection, l'augmentation des importations d'abeilles, la nécessité de mieux différencier les populations par des mesures biométriques s'est précisée [FRE.81].

La biométrie de l'abeille fait actuellement appel à un ensemble de mesures ou d'évaluations de caractères morphologiques qui permettent de définir avec plus ou moins de certitude et de précision la race des abeilles [FRE.81]. Le plus souvent, aucun de ces critères biométriques n'est suffisant à lui seul pour la discrimination escomptée. L'utilisation simultanée d'un certain nombre d'entre eux s'avère indispensable.

Des études précédentes ont montré la possibilité de distinguer morphologiquement des populations d'abeilles domestiques au sein d'une même race [TAS.75], [RUT.78], [COR.78].

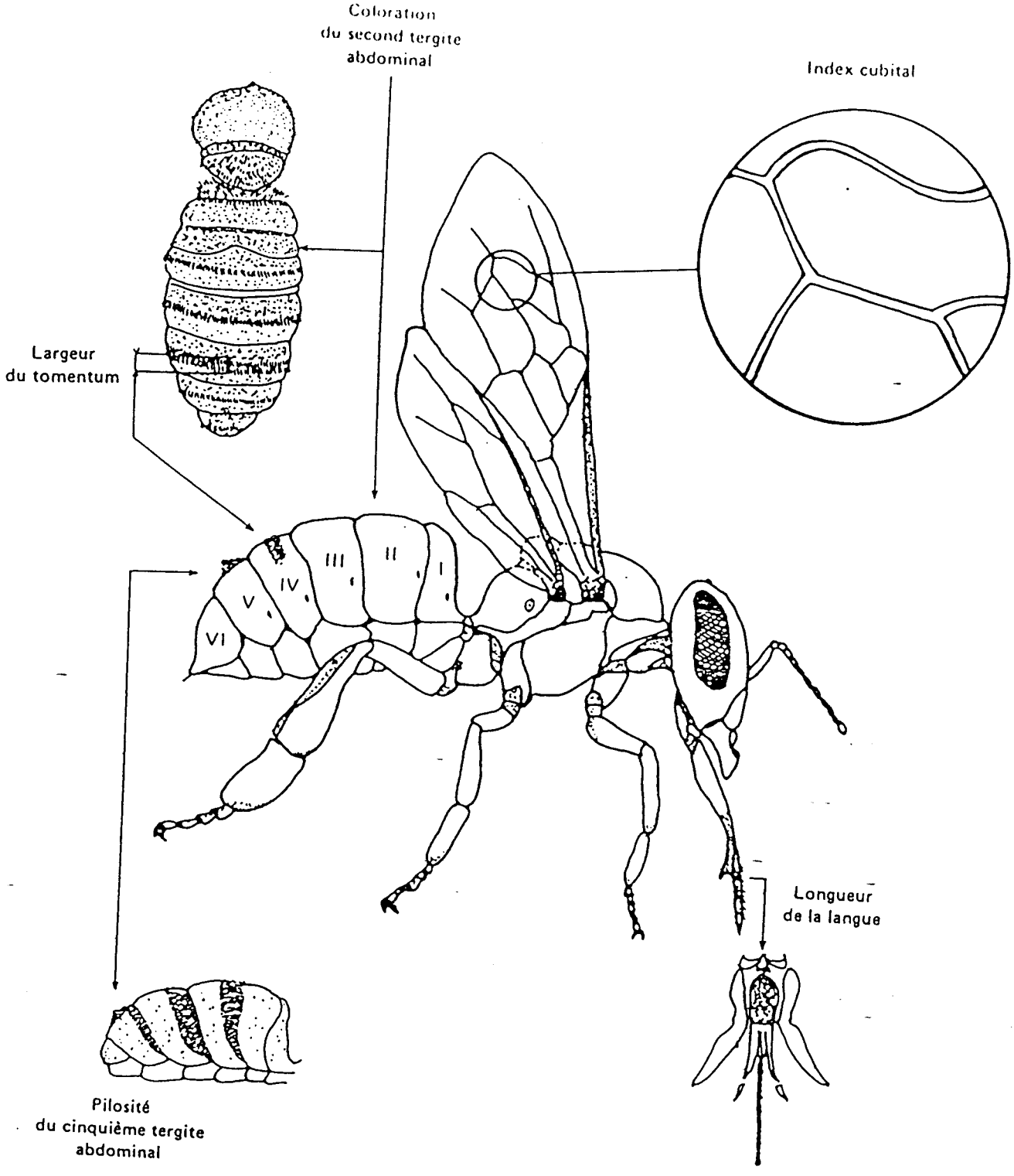


Figure 1: Principales mesures effectuées en biométrie.

En effet, les mesures de différents paramètres sur les abeilles permettent de représenter celles-ci dans des espaces multidimensionnels. L'existence d'écotypes différents se traduira par des nuages compacts de points représentant les individus. L'analyse morphométrique des abeilles pose donc un problème de classification automatique classique que nous aborderons dans un contexte non supervisé afin de ne pas imposer une structure a priori aux données.

VI - 2 CARACTERES BIOMETRIQUES ET METHODES DE MESURES.

Il existe une cinquantaine de caractères morphologiques utilisables en biométrie de l'abeille, mais la plupart des analyses biométriques n'utilisent que cinq caractères, retenus pour leur pouvoir discriminateur par plusieurs spécialistes, (GOETZE 1963), (RUTTNER 1968) et (MACKENSEN 1951) mondialement reconnus (cf figure 1). Il s'agit de:

VI -2 -1 Coloration.

La détermination de la présence et de l'importance de la couleur sur le deuxième tergite de l'abdomen a constitué, à l'origine, la base de la taxinomie des races d'abeilles. Elle est actuellement controversée en raison de la difficulté et du caractère parfois subjectif des mesures et aussi pour sa grande variabilité naturelle.

GOETZE (directeur de la station d'apiculture de BONN en ALLEMAGNE) a établi une classification en neuf catégories, du noir au jaune. L'estimation est faite par comparaison avec des croquis de référence (cf. figure 2).

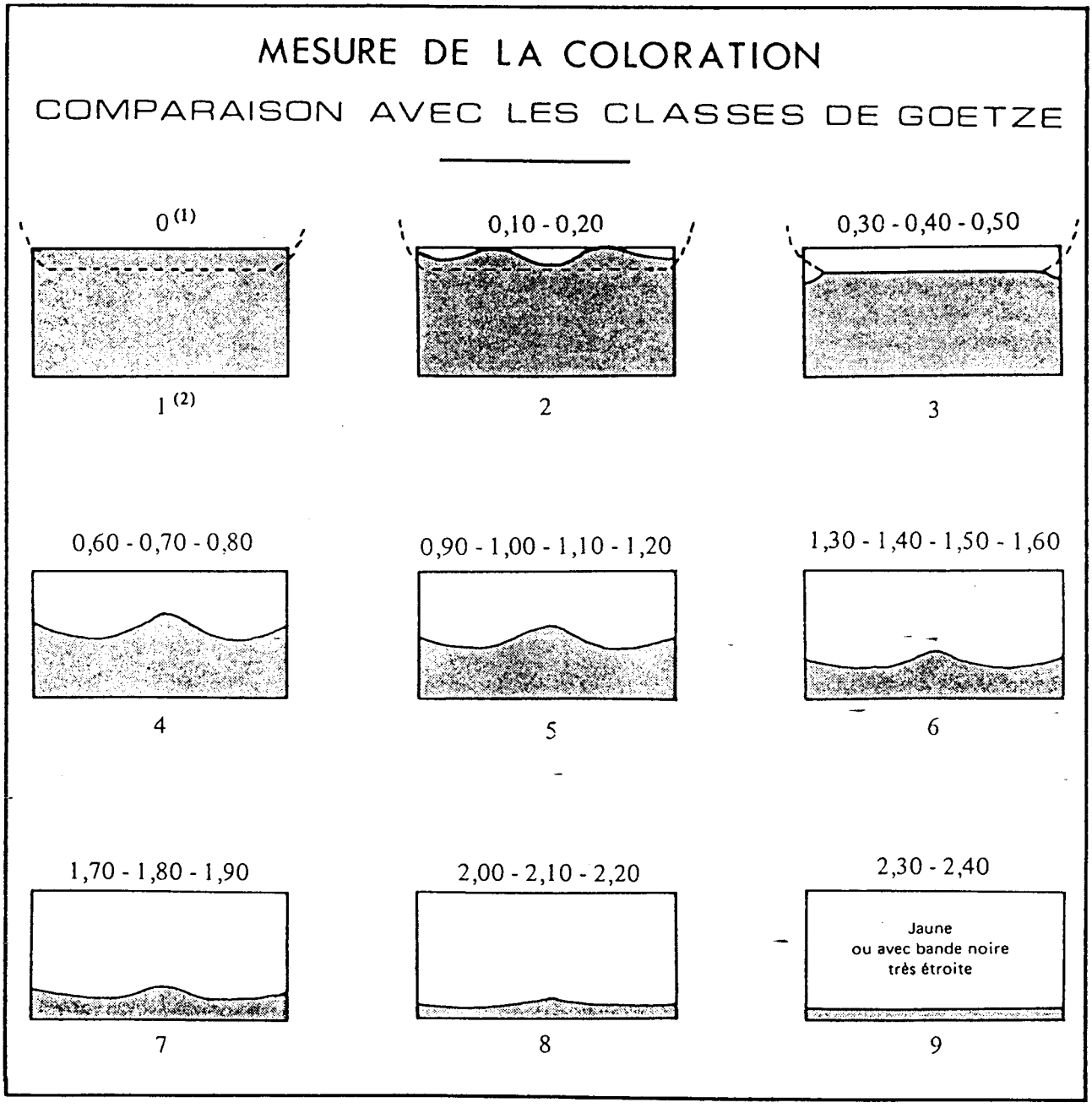

MESURE DE LA COLORATION
COMPARAISON AVEC LES CLASSES DE GOETZE
0 (1)
1 (2)
0,10 - 0,20
2
0,30 - 0,40 - 0,50
3
0,60 - 0,70 - 0,80
4
0,90 - 1,00 - 1,10 - 1,20
5
1,30 - 1,40 - 1,50 - 1,60
6
1,70 - 1,80 - 1,90
7
2,00 - 2,10 - 2,20
8
2,30 - 2,40
Jaune
ou avec bande noire
très étroite
9

Figure 2.

VI -2 -2 Pilosité

C'est la longueur des poils mesurée sur le cinquième tergite abdominal avec un réticule oculaire. Souvent, la limite de la bande de pilosité est mal définie, ce qui peut fausser les résultats des mesures.

VI -2 -3 Tomentum

C'est la longueur de la bande pileuse sur le quatrième tergite abdominal. La limite de la bande du tomentum n'est pas droite, elle est plus ou moins sinueuse. Il est préférable d'exprimer le tomentum sous forme d'index par le rapport suivant [FRE.81], [COR.78].

$$I_t = \frac{\text{Partie pileuse (mm)}}{\text{Partie glabre du tergite (mm)}}$$

VI -2 -4 Langue

La longueur est mesurée après avoir coupé et retourné la tête de l'abeille. Cette mesure interdit le travail en série, puisqu'elle exige quelques manipulations avant les mesures.

VI -2 -5 Index cubital

C'est le rapport de A sur B (A/B) où A et B sont les longueurs des deux nervures formant un angle obtus à la base de la troisième cellule cubital de l'aile droite de l'abeille. Il est mesuré à l'aide du dispositif de RUTTNER. [FRE.81], [TAS.75].

VI - 3 MESURE ET ANALYSE DES PARAMETRES MORPHOLOGIQUES.

VI -3 -1 Méthodes de mesure par colonie.

Un rucher est un ensemble de ruches. Dans chaque ruche loge une colonie, autrement dit, un ensemble d'abeilles. L'ensemble des colonies d'abeilles vivant sur une aire géographique donnée peut être considéré comme une population. A l'intérieur d'une population locale il existe entre les colonies des relations qui ont pour conséquence la formation d'un pool génétique plus ou moins riche. Le biométricien travaille, en général, au niveau de la colonie. Il prélève dans des conditions bien déterminées un ou plusieurs échantillons de trente abeilles. [TAS.75], [COR.78].

Les méthodes d'analyse actuellement employée, notamment par l'I.N.R.A, (Institut Nationnal de Recherche Agronomique), sont relativement simples mais exigent beaucoup de soin; mal réalisées elles peuvent être à l'origine d'erreurs importantes. Le travail de l'analyste est organisé de telle sorte que les mesures puissent être faites en série.

Pour présenter une colonie, les analystes prennent, pour chaque caractère morphologique, un échantillon de trente abeilles à partir duquel ils mesurent la valeur moyenne du caractère considéré [FRE.81], [COR.78]. Par exemple, pour l'index cubital, ils prennent l'échantillon de trente abeilles, donc trente index cubitaux et calculent la valeur moyenne I_{cmoy} de cet index. Ce dernier représentera tous les index individuels, il correspond à l'une des classes C_i établies par DREHER (cf. tableau 1).

GOETZE, RUTHNER et DREHER ont établi en 1963 des classes de référence. Ces classes sont définies par:

$$C_i = [I_{cmax}, \quad I_{cmin}]$$

et par leur index cubital moyen tel que:

$$I_{cmoy} = \frac{I_{cmax} + I_{cmin}}{2}$$

L'index cubital mesuré prendra la valeur moyenne I_{cmoy} de la classe à laquelle il correspond.

Limites de classe	Classes	Moyenne de Classe arrondie à 2 décimales
0.923		
	5	0.96
1.000		
	6	1.04
1.083		
	7	1.13
1.173		
	8	1.22
1.272		
	9	1.32
1.380		
	10	1.44
1.500		
	11	1.56
1.631		
	12	1.70
1.777		
	13	1.86
1.941		
	14	2.03
2.125		
	15	2.33
2.333		
	16	2.45
2.571		
	17	2.70
2.846		
	18	3.00
3.166		
	19	3.35
3.545		
	20	3.77
4.000		
	21	4.28
4.555		
	22	4.90
5.250		
	23	5.69
6.142		
	24	6.73
7.333		
	25	8.16
9.000		

Tableau 1: Classes de DREHER.
(Limites et moyennes des classes).

Notons que ces classes peuvent être trop importantes dans certaines zones et trop faibles dans une autre. Ces méthodes de mesure de l'index cubital ne peuvent donc pas être parfaites. Elles peuvent être la cause d'une fausse décision prise vis à vis des abeilles.

D'une manière analogue à celle de la mesure de l'index cubital, les analystes prennent un échantillon de trente abeilles pour la mesure de pilosité. Ils calculent la valeur moyenne P_{moy} de cette dernière. Ils font de même pour les autres caractères, chaque échantillon étant différent de l'autre. En fin de mesure, on obtient la valeur moyenne de chaque caractère morphologique. Ces derniers représenteront le prototype de la colonie pour laquelle chaque caractère a été calculé à partir de trente abeilles différentes (cf. figure 3).

Les analystes prennent, en général, de grandes précautions pour harmoniser leurs méthodes de mesure afin d'assurer leur reproductibilité d'un opérateur à l'autre, et pour le même opérateur, au cours du temps.

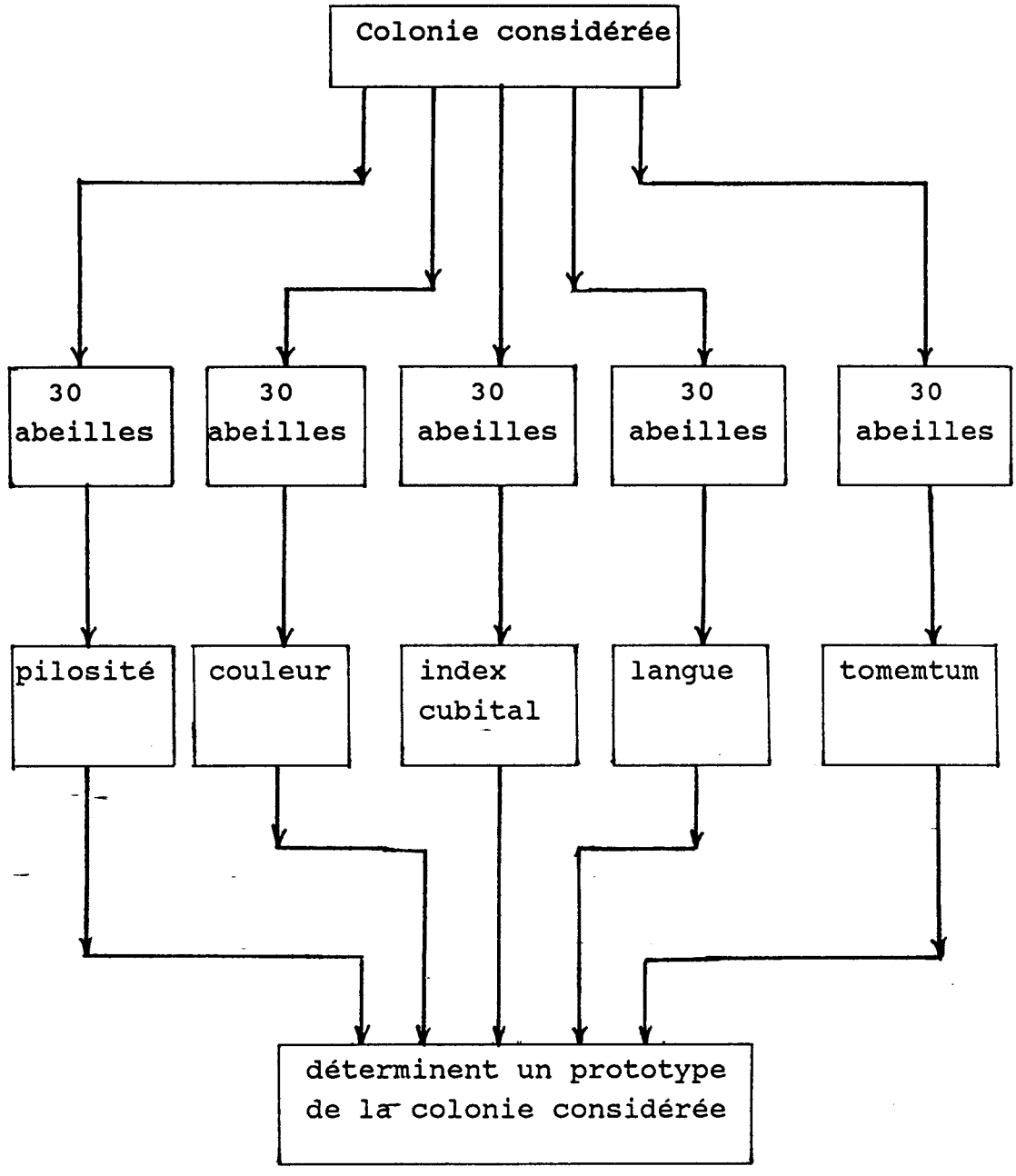


Figure 3: Méthode de la moyenne par colonie.

VI -3 -2 Méthode de mesure abeille par abeille.

Etant donné que chaque colonie est représentée par un prototype qui, en réalité, n'existe pas, il est indispensable d'analyser les caractères morphologiques par la méthode abeille par abeille. Cette méthode fournira des renseignements sur une colonie plus exacts que ceux fournis par la première méthode.

Cette procédure élimine le travail en série. Tout d'abord elle consiste à numéroter les abeilles destinées aux mesures au début de chaque analyse, et à établir sur une même abeille les mesures des cinq caractères. Le nombre d'abeilles par échantillon n'est pas fixé d'une façon impérative, et peut varier sans inconvénients [FRE.81], [COR.75].

Cette méthode nécéssite en fait un nombre réduit d'échantillon. Dans notre exemple, chaque colonie est représentée par un seul échantillon de trente abeilles dont on mesure cinq caractères. Il est ainsi possible de représenter une colonie par un prototype. Ce dernier est caractérisé par les valeurs moyennes des cinq caractères biométriques de la colonie (cf. figure 4).

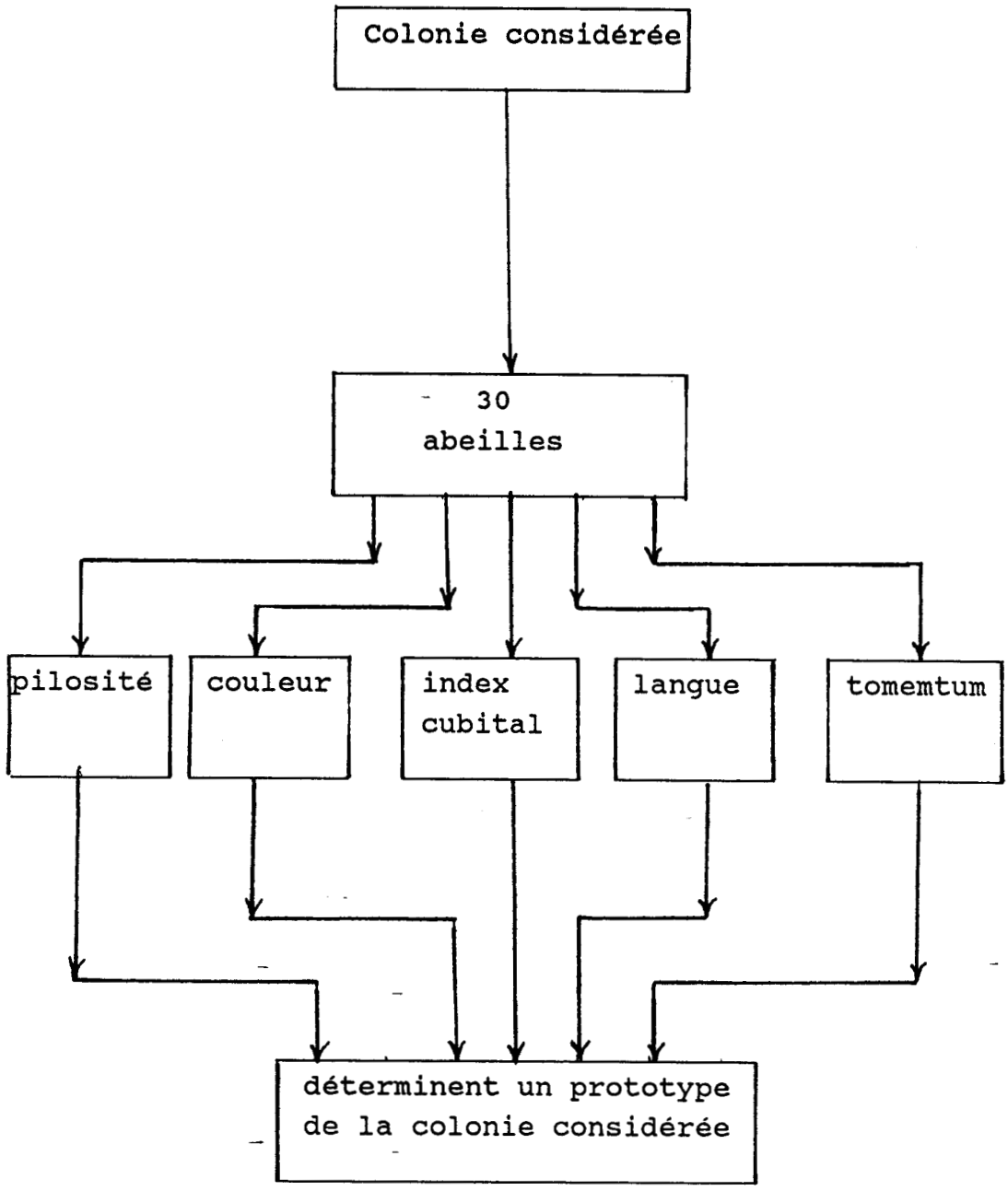


Figure 4: Méthode abeille par abeille.

VI - 3 -3 La biométrie de l'abeille: un problème de classification

On constate que, quelle que soit la méthode de mesure utilisée, les analystes se contentent en général de classer des colonies plutôt que les abeilles elles mêmes.

Les biométriciens de l'abeille ont donc l'habitude de condenser l'information existante dans leurs observations en déterminant un prototype par colonie. Cette approche est critiquable sur le plan méthodologique car elle suppose que les colonies sont homogènes, ce qui n'est pas assuré. On peut également se poser la question de savoir si un échantillon de trente abeilles suffit pour déterminer les valeurs moyennes des paramètres.

Les travaux de SNEDECOR [SNE.71], nous montrent que la taille de l'échantillon est-fonction de la dispersion des caractères. Il faut donc analyser ce problème avec précision. Mais comme la dispersion doit être calculée à l'intérieur de classes homogènes que nous devons justement identifier, le problème ne semble pas avoir de solution. Nous devons donc aborder l'étude morphologique des abeilles comme un problème de classification automatique dans son sens le plus large.

VI - 4 APPLICATION DES METHODES PROPOSEES A LA CLASSIFICATION DES ABEILLES

Chaque abeille étant caractérisée par six paramètres: la couleur "c", le tomentum "t", la langue "L", la pilosité "p", et par les deux longueurs A et B des nervures constituant la cellule cubital, peut être représentée par un point X tel que:

$$X = [c,\ t,\ L,\ p,\ A,\ B]^T$$

dans un espace de représentation à six dimensions.

L'application des méthodes d'estimation utilisées ainsi que les techniques d'analyse d'échantillon, (à savoir: la technique estimation-filtrage et la technique séparation/fusionnement) doivent permettre d'établir, si elles existent, la présence des différentes classes au sein de l'échantillon étudié, chaque classe correspondant à un écotype spécifique.

L'échantillon servant à notre application est un échantillon constitué de mille trois cents trente six (1336) abeilles sur lesquelles on a mesuré les six paramètres précédemment définis. Ces abeilles proviennent de ruchers de l'archipel de la GUADELOUPE.

Compte tenu de la dimension de l'espace de représentation, le nombre d'abeilles disponibles dans l'échantillon est relativement faible et justifie l'utilisation des méthodes d'analyse étudiées dans les chapitres précédents. Nous supposerons que les observations suivent des lois normales et que les paramètres caractérisant les abeilles sont statistiquement indépendants.

Nous appliquons, sur les observations constituants l'échantillon disponible, les méthodes d'estimation non paramétrique étudiées dans le chapitre II pour chacun des six paramètres caractérisant les abeilles à analyser.

VI - 4 - 1 Estimation par la méthode des plus proches voisins.

Nous estimons par la méthode des plus proches voisins les fonctions de densité marginale de probabilité obtenues à partir de l'échantillon d'abeilles disponible et ceci pour chacun des six paramètres.

Les résultats sont représentés par la figure 5. Plus précisément les figures 5(a), 5(b), 5(c), 5(d), 5(e) et 5(f) représentent respectivement les fonctions de densité marginale de probabilité estimées pour les caractères: couleur, tomentum, pilosité, langue, longueur A et longueur B. Ces résultats ont été obtenus pour $k_Q = 20$.

Le fait de travailler axe par axe, autrement dit le fait d'étudier chaque paramètre individuellement, permet la visualisation, sur écran d'un calculateur, des fonctions de densité marginale de probabilité. La visualisation des résultats facilite l'exploration de l'échantillon à étudier et l'ajustement du paramètre k_Q de l'estimateur.

L'analyse des fonctions de densité marginale de probabilité relatives à chacun des caractères mesurés sur les abeilles permet de mettre en évidence les caractères les plus discriminants. Dans notre cas les fonctions de densité marginale de probabilité obtenues pour chacun des caractères sont très bruitées. Pour pallier ces variations résiduelles et peut être non significatives, on applique l'opérateur de filtrage non linéaire décrit au chapitre III et qui a fait ses preuves sur des petits échantillons générés artificiellement.

VI - 4 - 2 Application de l'opérateur de filtrage non linéaire.

Nous avons vu précédemment que l'opérateur de filtrage non linéaire permet de restaurer les propriétés de convexité des fonctions de densité marginale de probabilité. La largeur l de la fenêtre de cet opérateur est ajustée de manière interactive, au vu des résultats obtenus axe par axe. En effet, nous avons vu que si la valeur de l est trop importante, on peut voir disparaître des modes significatifs. Par contre, si la valeur de l est faible, les fonctions de densité marginale de probabilité restent entachées de bruits.

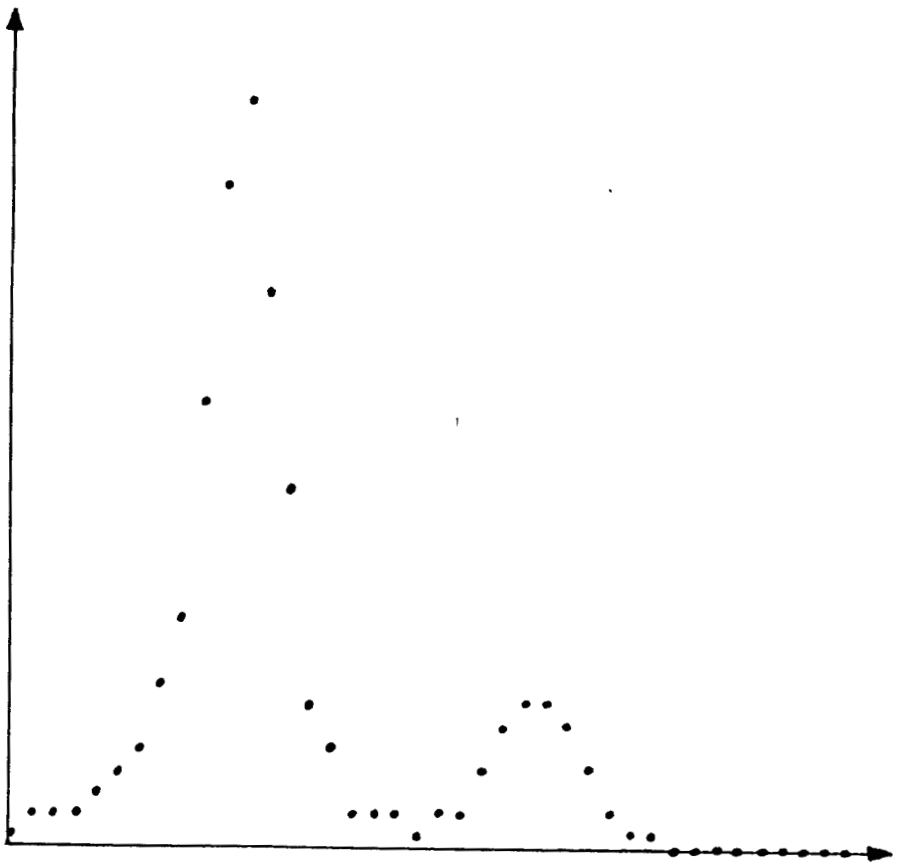


Fig. 5 (a)
avant filtrage

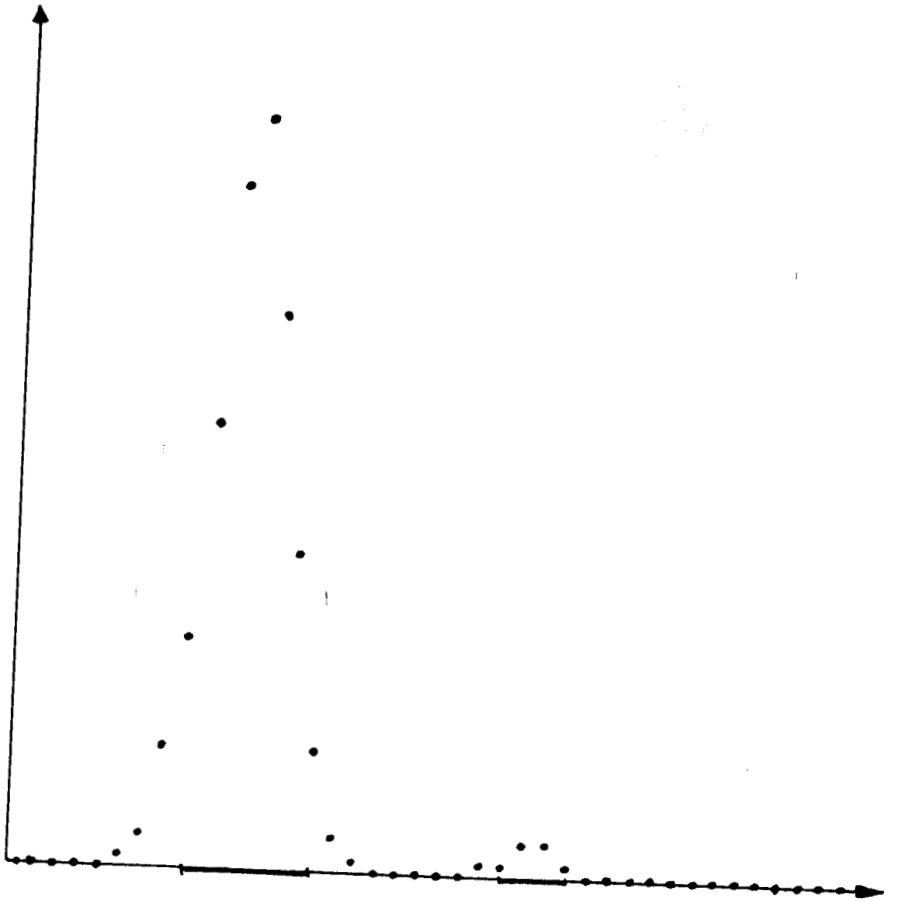


Fig. 8(a)
après filtrage

Fonctions de densité marginale de probabilité
associées au caractère couleur

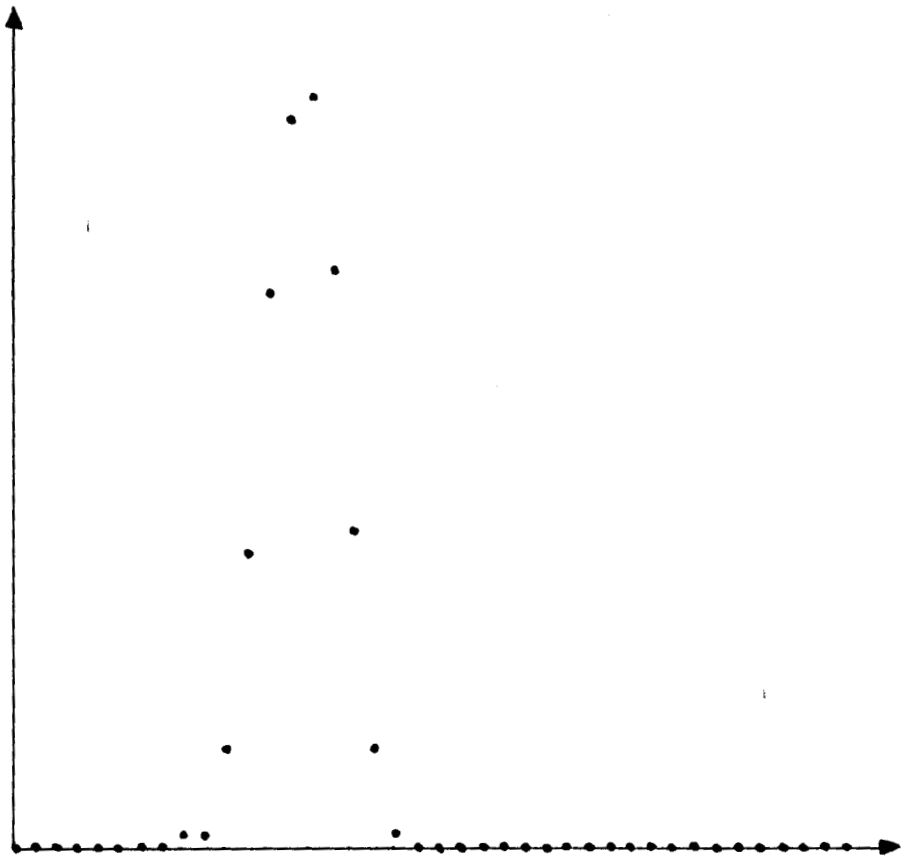


Fig. 5 (b)
avant filtrage

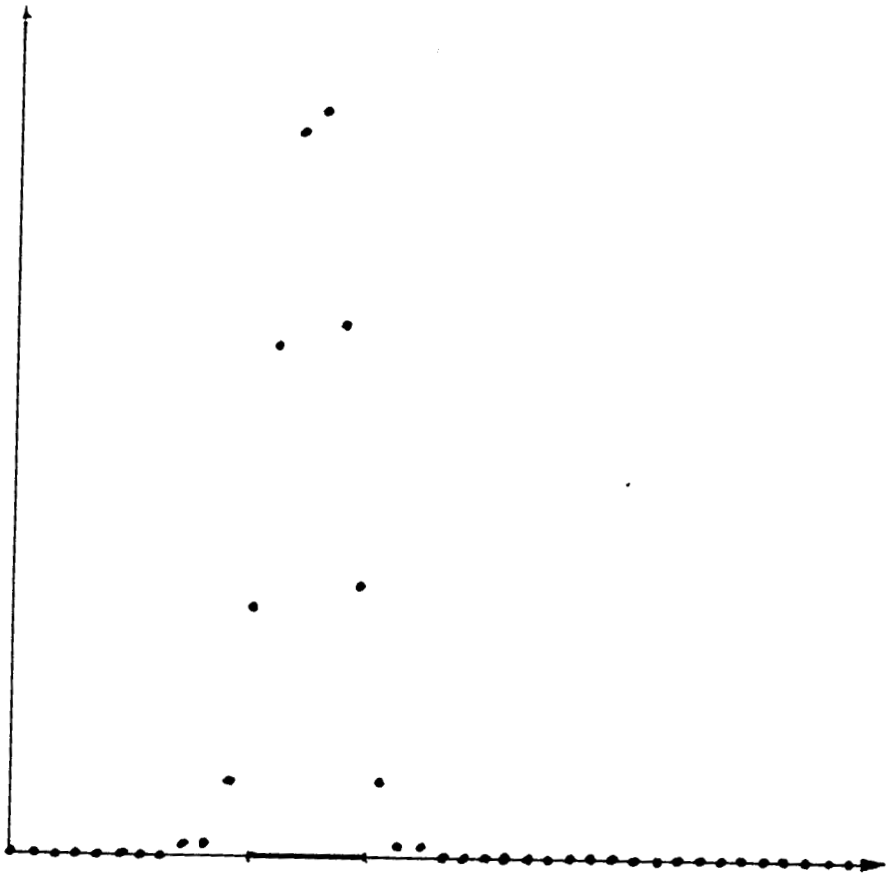


Fig. 8(b)
après filtrage

Fonctions de densité marginale de probabilité
associées au caractère pilosité

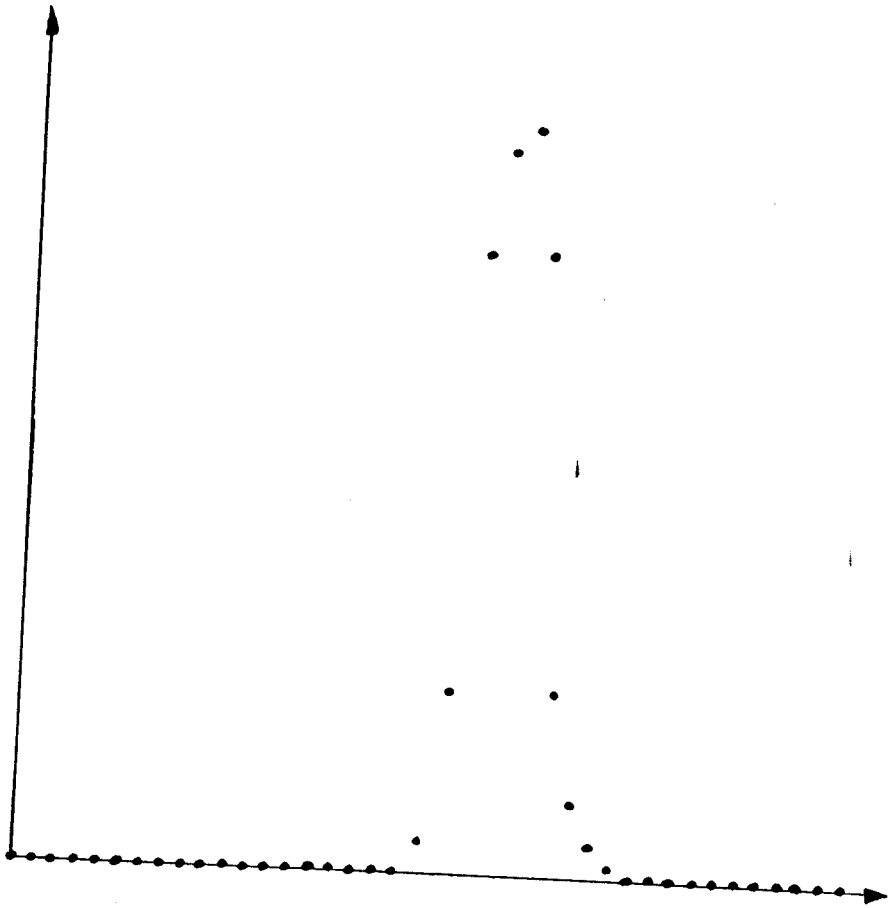


Fig. 5 (c)

avant filtrage

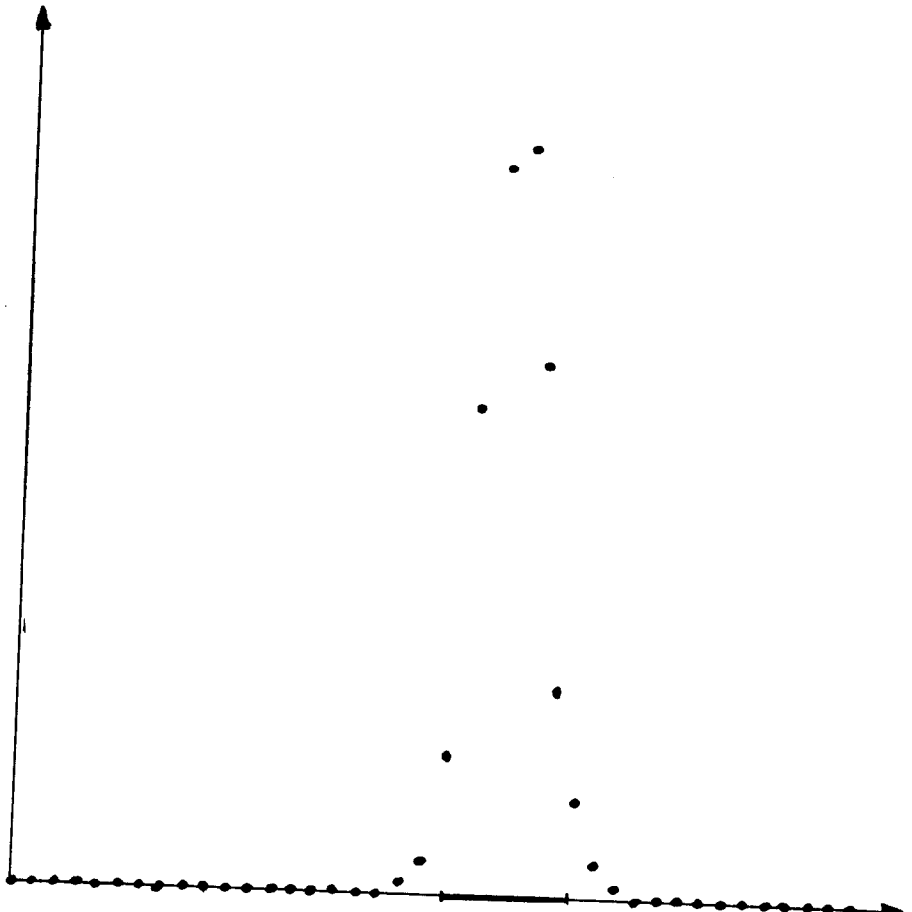


Fig. 8 (c)

après filtrage

Fonctions de densité marginale de probabilité associées au caractère Tomentum

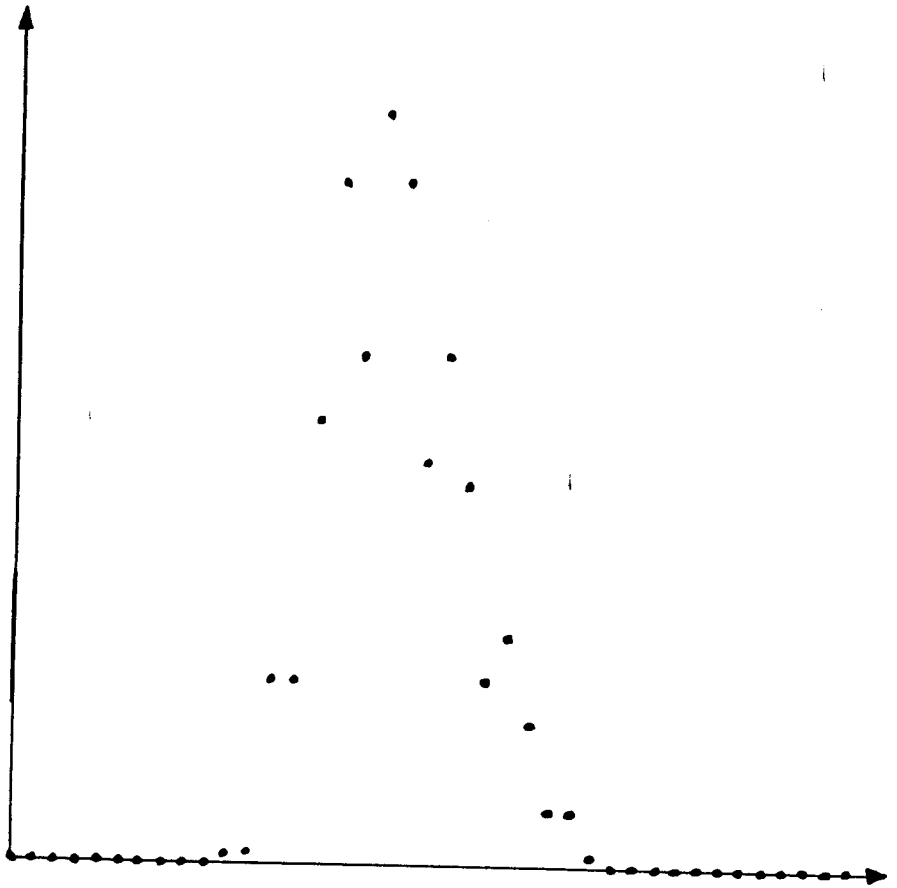


Fig. 5(d)

avant filtrage

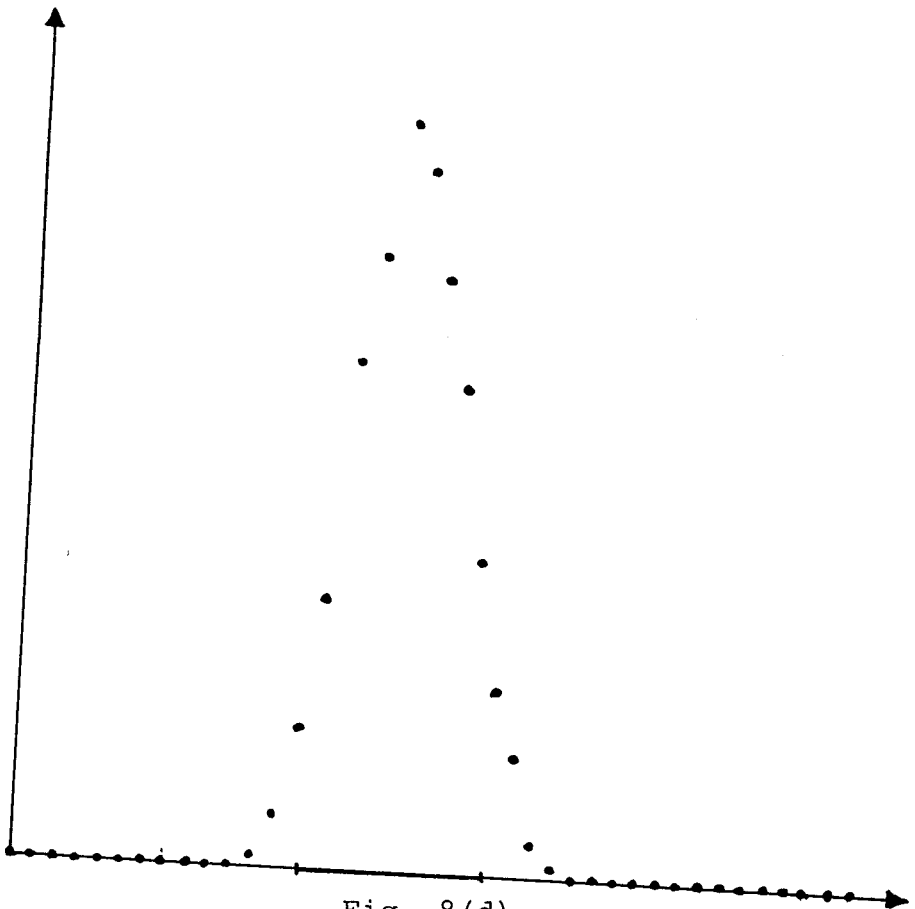


Fig. 8(d)

après filtrage

Fonctions de densité marginale de probabilité
associées au caractère Langue

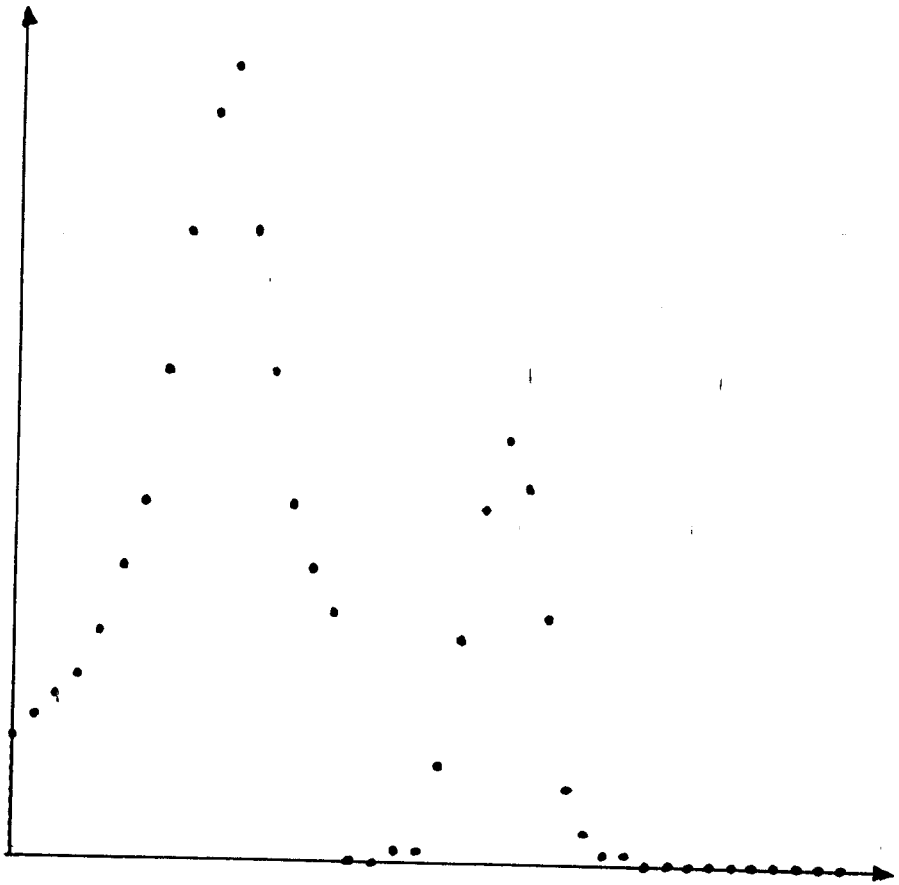


Fig. 5(e)

avant filtrage

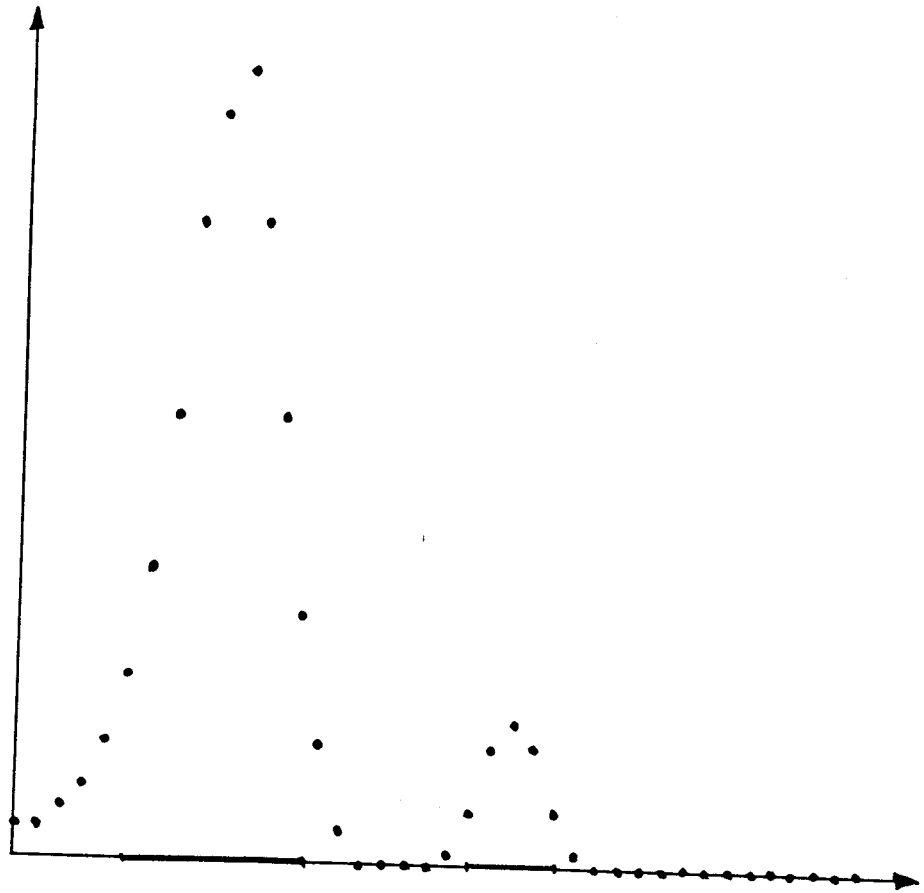


Fig. 8(e)

après filtrage

Fonctions de densité marginale de probabilité associées au caractère longueur A des nervures de la cellule cubitale

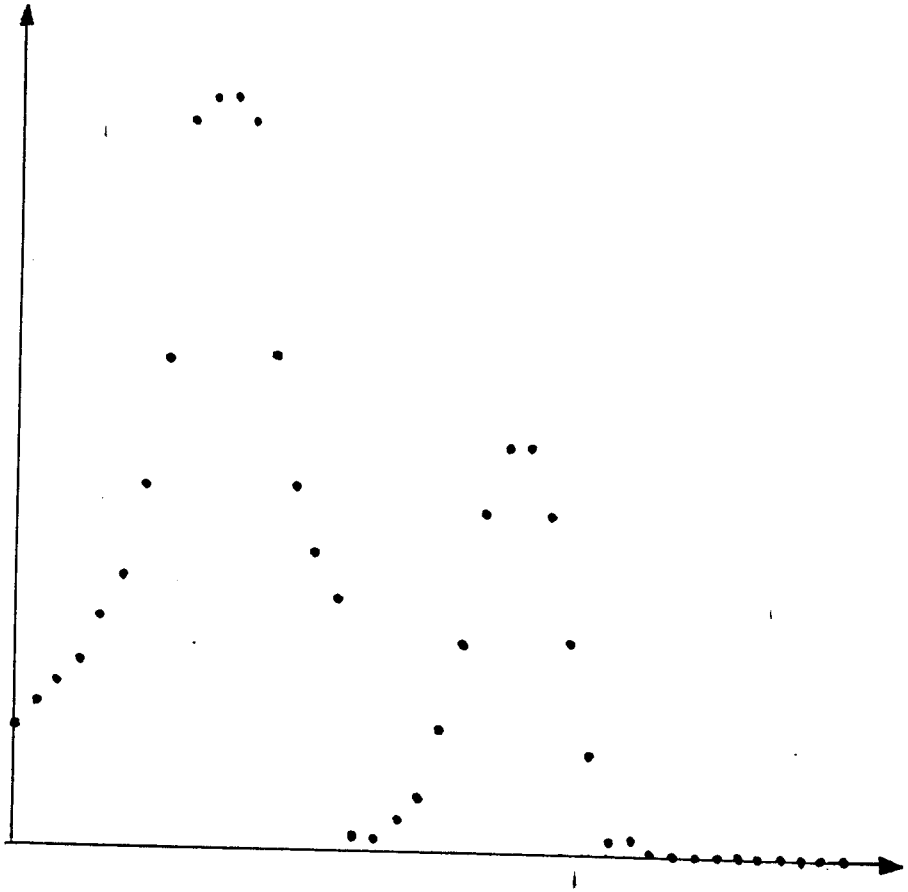


Fig. 5(f)

avant filtrage

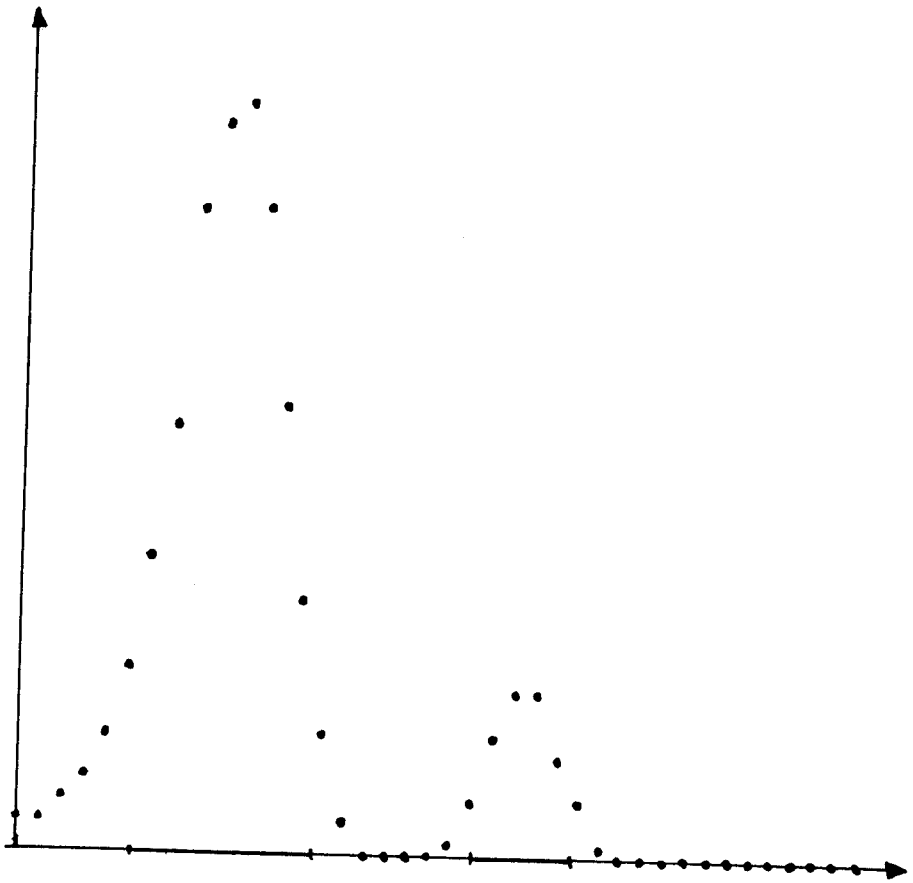


Fig. 8(f)

après filtrage

Fonctions de densité marginale de probabilité associées au caractère longueur B des nervures de la cellule cubitale

Les figures 6(a,b,c,d,e et f) indiquent les estimations filtrées pour une petite largeur de la fenêtre de l'opérateur de filtrage, soit l = 2. Ces résultats montrent bien la subsistance de petites variations locales non significatives.

Par contre, les figures 7(a,b,c,d,e et f) sont obtenues en choisissant une fenêtre de l'opérateur de filtrage trop large. Les courbes représentant les estimations filtrées avec l = 25 masquent les véritables modes de ces fonctions. Ce phénomène de lissage excessif apparait surtout dans le cas des caractères qui ont une fonction de densité marginale de probabilité multimodale. Avec l = 25, elles deviennent unimodales.

Après visualisation de plusieurs essais, nous avons trouvé que la valeur l = 5 donne les résultats les plus satisfaisants pour chacun des paramètres caractérisant l'abeille.

En appliquant l'opérateur de filtrage non linéaire avec une cette largeur de fenêtre à chacune des fonctions de densité marginale de probabilité représentées par la figure 5, nous obtenons les estimations filtrées indiquées par les figures 8(a,b,c,d,e et f). Ces dernières correspondent respectivement aux caractères: couleur, tomentum, pilosité langue, longueur A et longueur B.

VI - 5 - 3 Classification

Aux estimations filtrées, on applique le test de convexité décrit au chapitre II. L'analyse des fonctions de densité marginale de probabilité relatives à chacun des paramètres mesurés sur les abeilles nous a permis de mettre en évidence les caractères les plus discriminants.

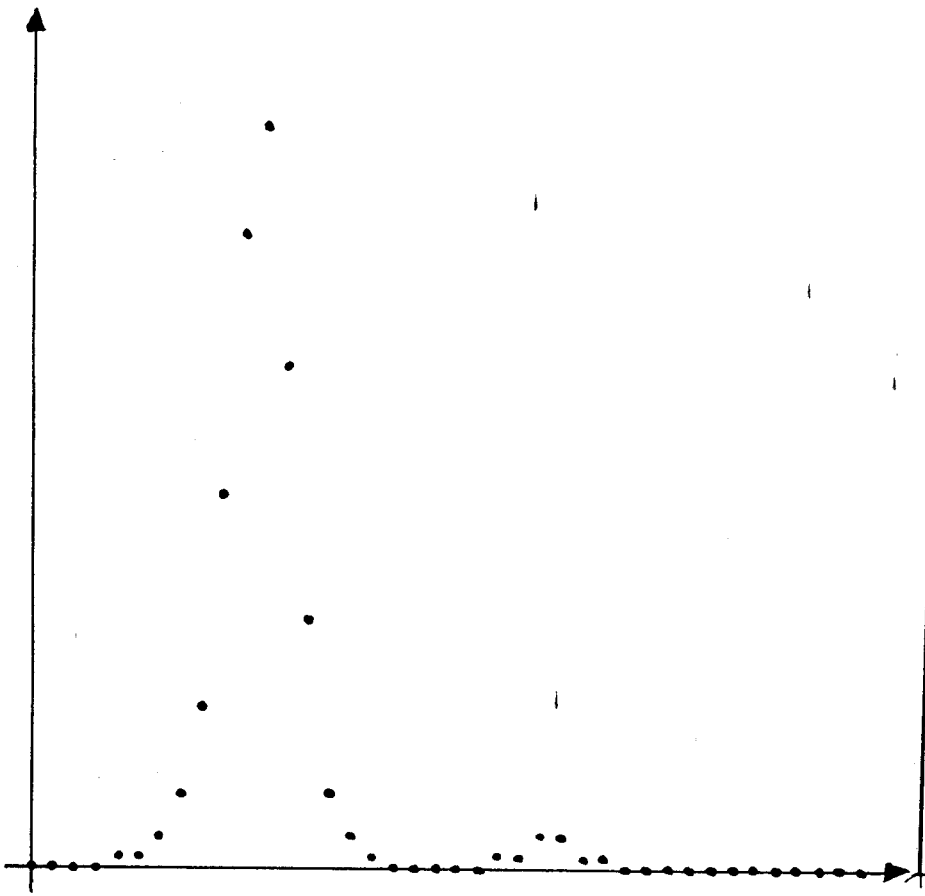


Fig. 6(a)

avant filtrage

l = 2

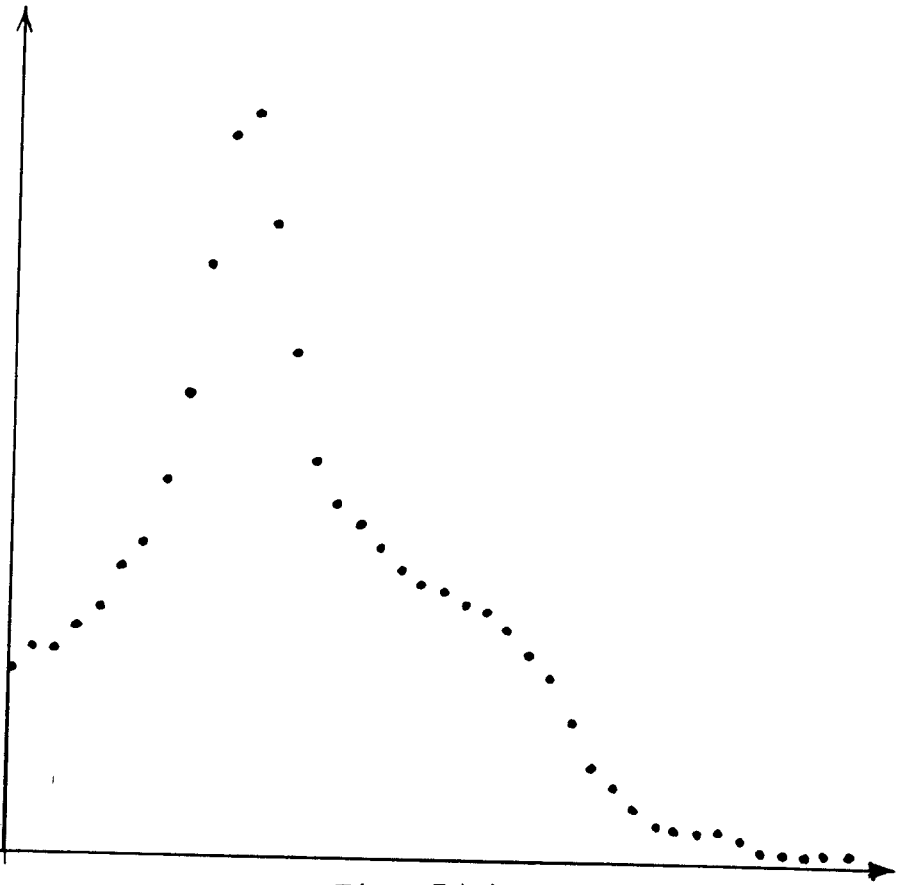


Fig. 7(a)

après filtrage

l = 25

Fonctions de densité marginales de probabilité estimées filtrées, représentant le caractère couleur

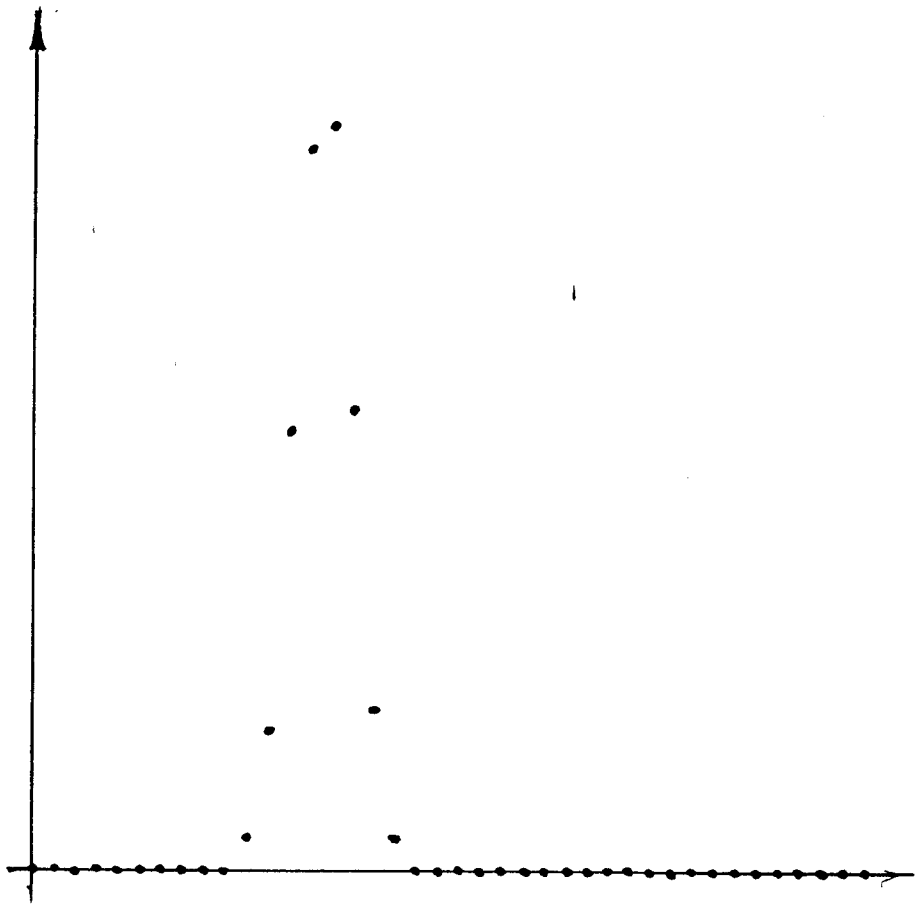


Fig. 6(b)

avant filtrage

l = 2

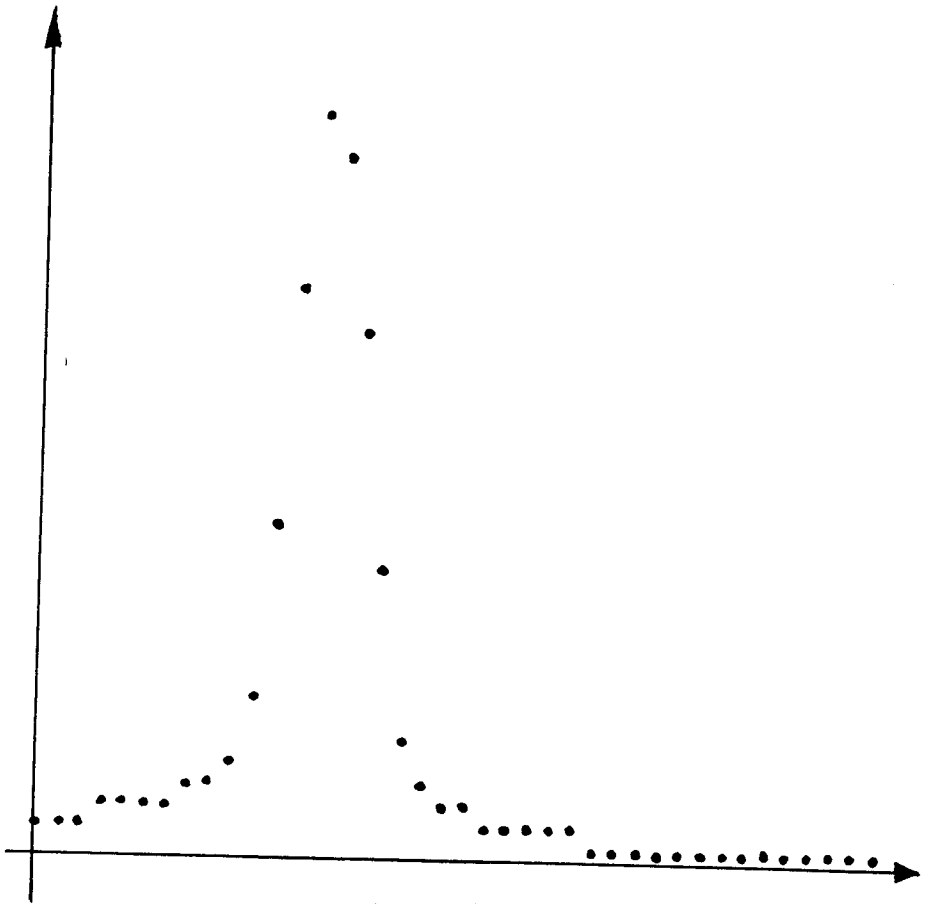


Fig. 7(b)

après filtrage

l = 25

Fonctions de densité marginales de probabilité estimées filtrées, représentant le caractère pilosité

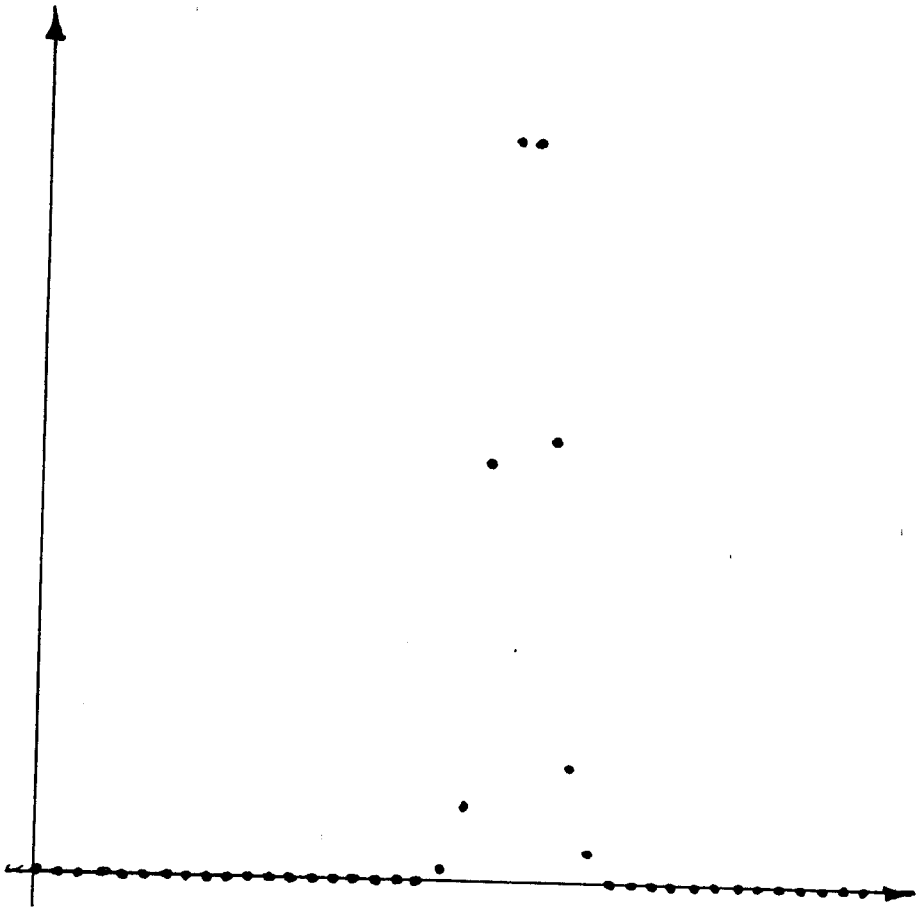


Fig. 6(c)

avant filtrage

1 = 2

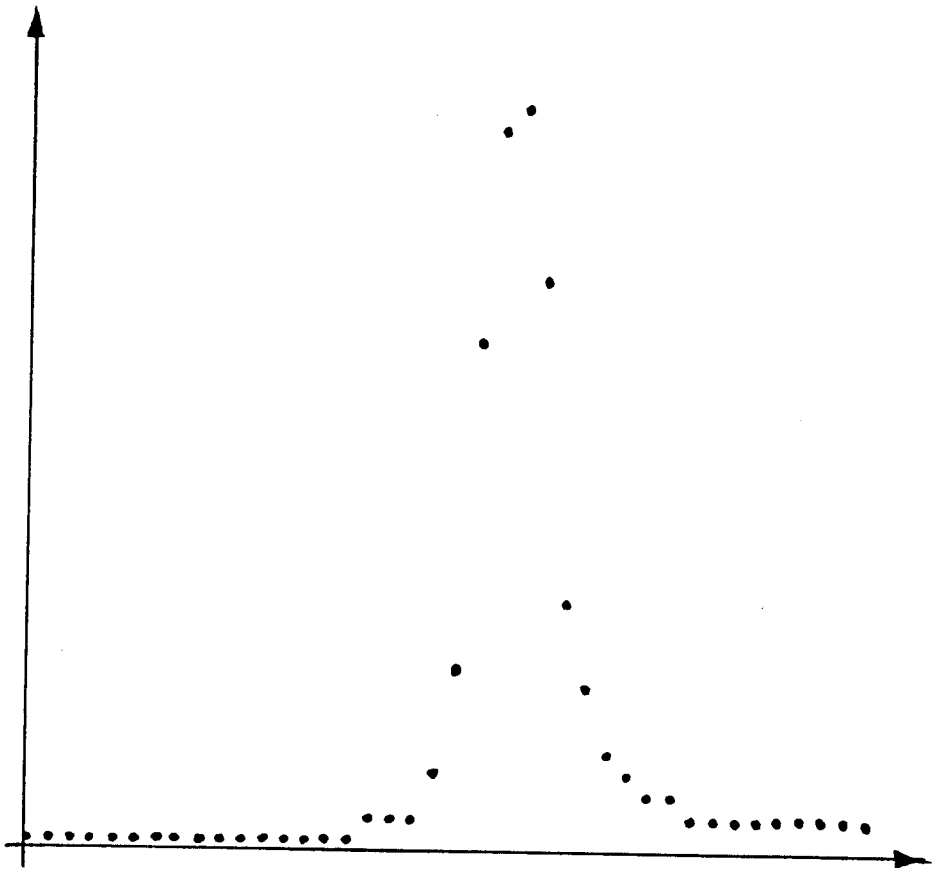


Fig. 7(c)

après filtrage

1 = 25

Fonctions de densité marginales de probabilité estimées filtrées, représentant le caractère tomentum

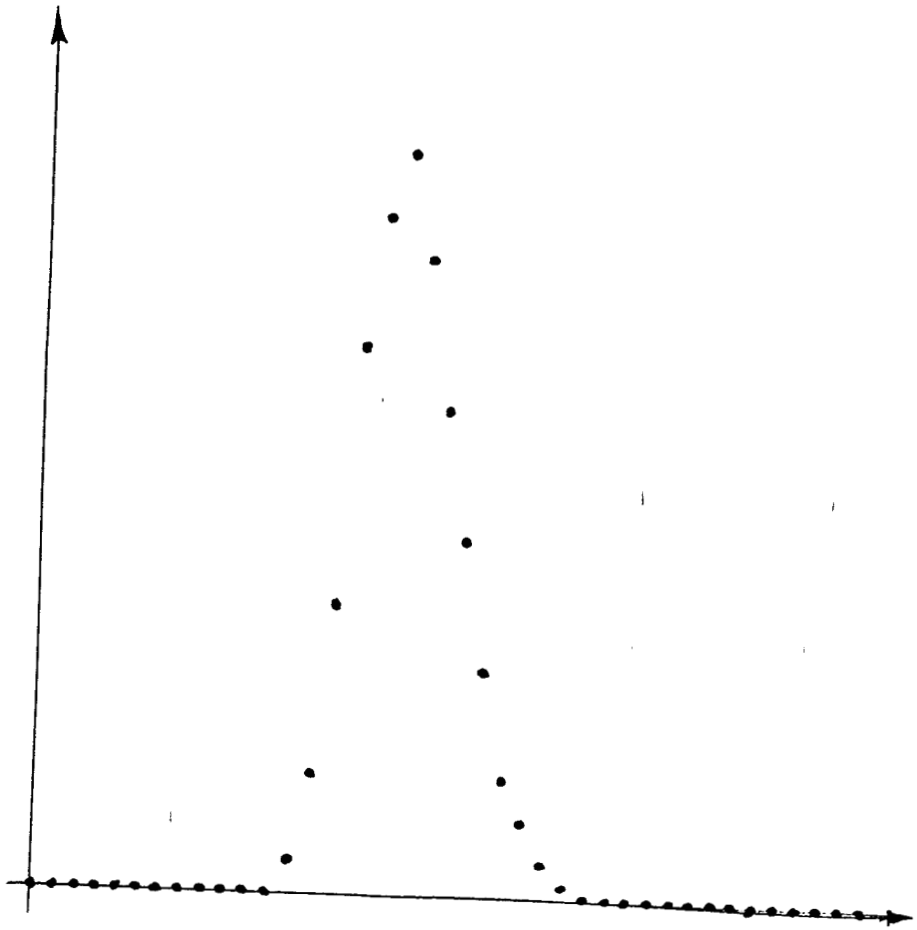


Fig. 6(d)

avant filtrage

l = 2

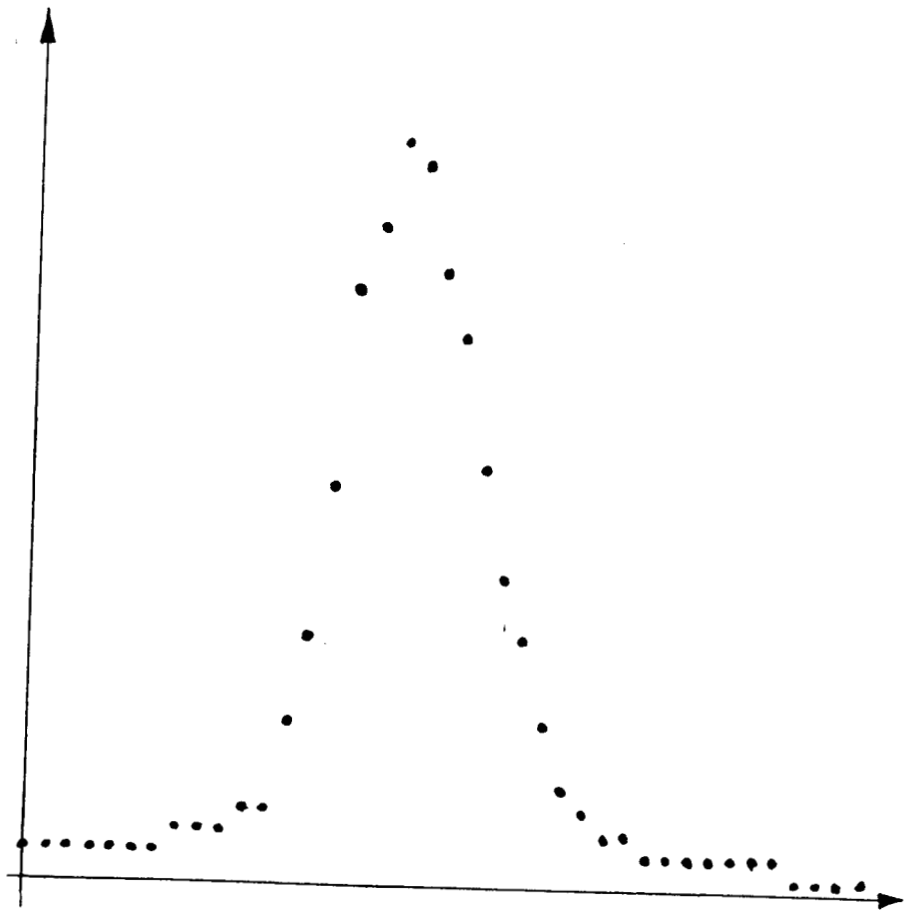


Fig. 7(d)

après filtrage

l = 25

Fonctions de densité marginales de probabilité estimées filtrées, représentant le caractère langue

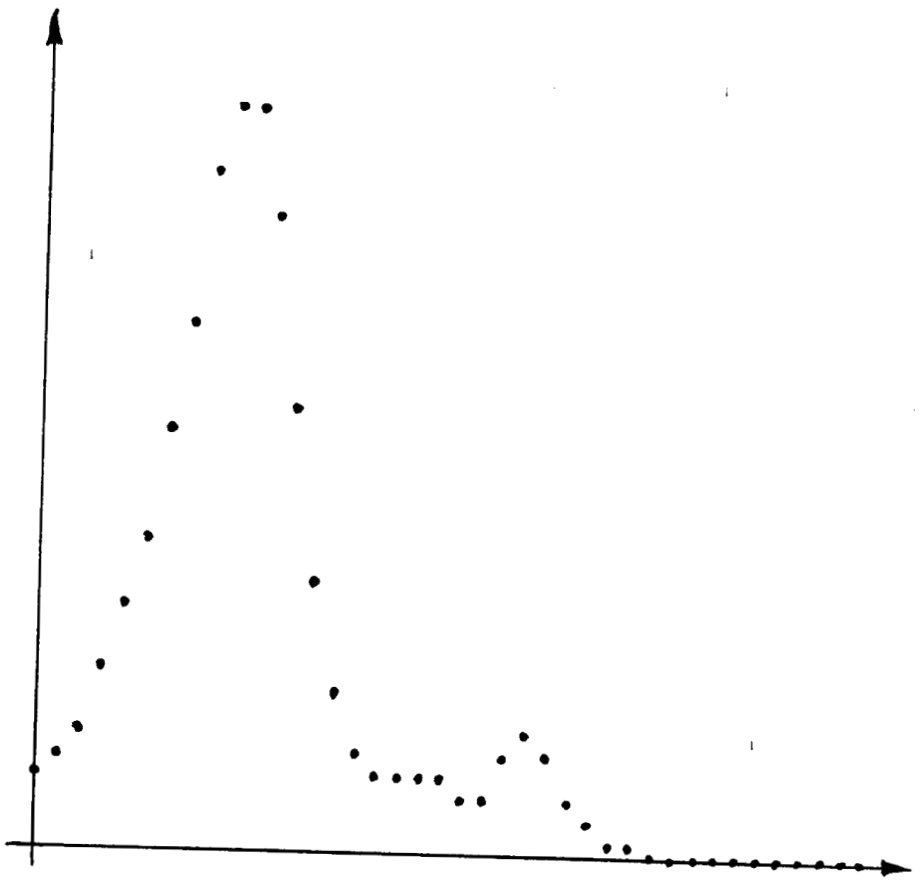


Fig. 6(e)

avant filtrage

l = 2

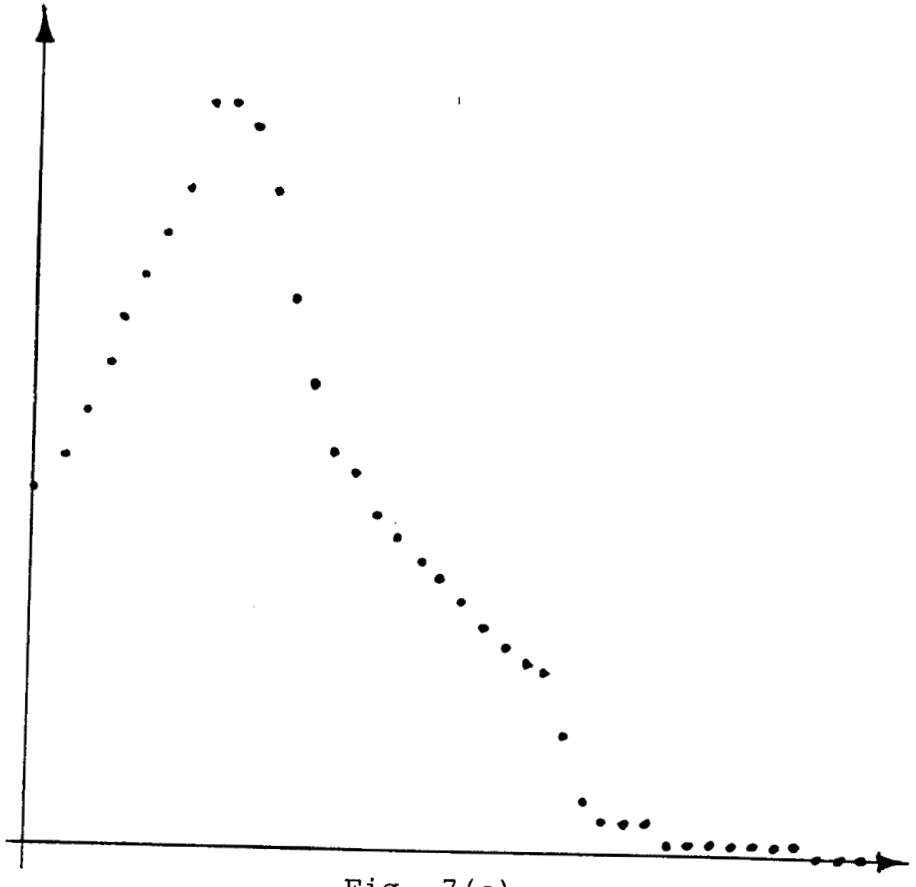


Fig. 7(e)

après filtrage

l = 25

Fonctions de densité marginales de probabilité estimées filtrées, représentant le caractère longueur A des nervures de la cellule cubitale

Seules les fonctions de densité marginale de probabilité multimodales permettront de distinguer plusieurs classes dans la distribution des observations. Nous n'utilisons donc par la suite que les caractères couleur, longueur A et longueur B des nervures de la cellule cubital auquels sont associées les fonctions de densité marginale de probabilité estimées filtrées multimodales des figures 8(a), 8(e), 8(f).

L'analyse de convexité appliquée aux fonctions de densité marginale de probabilité correspondant à ces caractères discriminants, met en évidence deux segments convexes sur chacun d'eux (cf. figures 8(a), 8(e), 8(f)). Par la méthode classique du produit euclidien de ces segments on obtient huit domaines caractéristiques.

L'étude de ces domaines caractéristiques permet de déterminer, pour chaque composante mise en évidence, les valeurs approchées du vecteur moyenne, de la matrice de covariance et de la probabilité a priori.

Cette étude nous a permis de rejeter les domaines artificiellement introduits par la procédure et qui ne contenaient que très peu d'observations. Nous n'avons retenu que deux domaines caractéristiques. Ces derniers possèdent une grande probabilité a priori.

Plus précisément les six autres domaines avaient une probabilité a priori inférieure à 0.073. C'est à dire que le domaine caratéristique le plus important parmi les six domaines rejetés ne contenait que 98 abeilles sur 1336.

Les deux domaines retenus ont des probabilités égales à 0.164 (correspondant à 220 abeilles) et 0.470 (629 abeilles).

Ainsi nous utiliserons les fonctions de décision $g_k(X)$ sur la base des valeurs approchées du vecteur moyenne, de la matrice de covariance et de la probabilité a priori de ces deux domaines retenus pour optimiser le processus de la classification.

Les paramètres statistiques des deux classes mises en évidence sont présentés sur le tableau 2 suivant.

	Vecteur moyenne	Matrice de covariance	Probabilité a priori
Classe C_1	$\begin{bmatrix} 31.80 \\ 300.00 \\ 153.00 \end{bmatrix}$	$\begin{bmatrix} 1.80 & & 0 \\ & 11 & \\ 0 & & 8.40 \end{bmatrix}$	0.74
Classe C_2	$\begin{bmatrix} 15.60 \\ 161.50 \\ 81.00 \end{bmatrix}$	$\begin{bmatrix} 1.20 & & 0 \\ & 17.50 & \\ 0 & & 14.00 \end{bmatrix}$	0.26

Tableau 2: Valeurs des paramètres statistiques des deux classes d'abeilles mises en évidence.

Ainsi l'étude du mélange d'abeilles a mis en évidence deux classes. Nous allons vérifier si ces classes sont de véritables classes normales. Ceci est réalisé en quantifiant l'écart quadratique entre la fonction de densité marginale relative à chacune des classes et le modèle analytique correspondant. Ce dernier est déterminé à partir des paramètres statistiques estimés sur la base des abeilles constituant le groupement correspondant.

Nous pouvons donc conclure que les deux classes se rapprochent de manière satisfaisante du modèle gaussien. Aucune fusion ni séparation des groupements n'est donc nécessaire.

VI - 4 - 5 Détermination des paramètres statistiques des classes mises en évidence.

L'analyse de convexité des fonctions de densité marginale de probabilité pour chaque composante, a mis en évidence un segment convexe sur chacun des caractères. La longueur de ce dernier est égale au double de l'écart type de la classe associée. Son milieu coincide avec la moyenne de cette même classe.

A partir de cette analyse nous avons déterminé les valeurs approchées des paramètres statistiques pour chacune des composantes.

Les résultats obtenus, en considérant les six paramètres qui caractérisent chaque abeille, sont consignés dans le tableau 4.

	Vecteur moyenne	Matrice de covariance	Probabilité a priori
Classe C1	$\begin{bmatrix} 31.80 \\ 300.00 \\ 153.00 \\ 21.84 \\ 18.03 \\ 73.99 \end{bmatrix}$	$\begin{bmatrix} 1.80 & & & & & 0 \\ & 11.00 & & & & \\ & & 8.40 & & & \\ & & & 1.12 & & \\ & & & & 0.83 & \\ 0 & & & & & 0.37 \end{bmatrix}$	0.74
Classe C2	$\begin{bmatrix} 15.60 \\ 161.50 \\ 81.00 \\ 22.34 \\ 17.93 \\ 74.66 \end{bmatrix}$	$\begin{bmatrix} 1.20 & & & & & 0 \\ & 17.50 & & & & \\ & & 14.00 & & & \\ & & & 1.39 & & \\ & & & & 0.97 & \\ 0 & & & & & 0.49 \end{bmatrix}$	0.26

Tableau 4

VI - 4 - 4 Vérification

Les valeurs des erreurs quadratiques minimales, entre la fonction de densité marginale de probabilité et le modèle analytique, obtenues pour chacune des classes et suivant chaque caractère sont consignées dans le tableau 3 .

	Erreurs quadratiques minimales					
	E_1	E_2	E_3	E_4	E_5	E_6
Classe C_1	0.032	0.022	0.051	0.031	0.034	0.045
Classe C_2	0.027	0.019	0.043	0.022	0.031	0.029

Tableau 3

A partir du tableau 3 nous pouvons déduire, et cela pour chaque caractère étudié, le groupement de référence. Ce dernier servira de base pour l'ajustement des paramètres de réglage de l'estimateur.

Pour notre exemple, c'est la classe C_2 qui possède les valeurs minimales des erreurs quadratiques et cela quelque soit le caractère considéré. Nous pouvons expliquer ceci par l'importance du nombre d'abeilles qui constituent la classe C_2

La suite du traitement consiste à vérifier si les valeurs des erreurs quadratiques associées à la classe C_1 vérifient la relation 7 du chapitre V, c'est à dire la relation:

$$E'_{i,m} > \alpha\, E^*_{i,m} \qquad \text{(7) du chap. V}$$

On constate que quelque soit le caractère considéré, les erreurs quadratiques associées à la classe C_1 ne vérifient pas cette relation tant que α est supérieur à 1.5. En effet il n'y a pas une grande différence entre les erreurs associées à la classe C_1 et celles associées à la classe C_2.

VI - 5 CONCLUSION

Le tableau des valeurs approchées des paramètres statistiques explique bien le fait que les caractères: tomentum, pilosité, et langue ont des fonctions de densité marginale de probabilité qui n'ont qu'un seul mode.

Nous constatons que les deux composantes obtenues ont les moyennes des caractères tomentum, pilosité et langue qui sont égales. Ce qui veut dire que pour ces caractères les deux classes se chevauchent et le degré de chevauchement est très important.

Par contre les trois autres caractères, à savoir la couleur et les deux longueurs A et B des nervures de la cellule cubital, ont des moyennes différentes. Ceci explique la présence des deux modes sur la fonction de densité marginale de probabilité de chaque composante.

Grâce aux différentes techniques présentées à travers les chapitres de ce mémoire, nous avons analysé un échantillon d'abeilles. Nous avons mis en évidence la présence de deux classes.

L'interprétation de ces résultats est maintenant du ressort de l'apidologue qui pourra interpréter ces résultats en fonction de l'écosystème dans lequel évoluent les populations d'abeilles analysées et des races en présence dans les ruchers.

Pour pallier cette limitation, nous avons proposé des techniques de fusionnement qui viennent compléter celles de séparations déjà existantes pour affiner les résultats.

L'approche proposée a finalement été appliquée à une population réduite d'abeilles en vue d'analyser leur morphologie dans le cadre d'études biométriques. Cette application a permis de mettre en évidence l'intérêt de travailler axe par axe. En effet, la visualisation de tous les résultats intermédiaires correspond à chacun des attributs mesurés sur les objets à classer (ici les abeilles) permet d'ajuster les paramètres de réglage des algorithmes de manière interactive. L'analyste garde ainsi un contrôle effectif sur le déroulement de la procédure, ce qui assure une meilleure fiabilité des résultats

Une critique importante qui peut être formulée est la nécessité d'accepter l'hypothèse d'indépendance statistique des attributs mesurés sur les objets. Pour lever cette hypothèse, il est indispensble de travailler directement dans l'espace multidimensionnel. C'est la raison pour laquelle nous envisageons d'adapter la technique de filtrage non linéaire à une dimension proposée à des fonctions multivariables. Dans ce cas, la méthode d'identification des mélanges gaussiens par analyse de fonction de densité sous-jacente pourrait être appliquée à des petits échantillons.

Il serait également souhaitable de pouvoir adapter les procédures de séparation/fusionnement destinées à affiner la classification à ces méthodes multidimensionnelles.

Ces deux aspects de la généralisation des outils d'analyse des données présentés dans ce mémoire constitueront l'essentiel, de nos prochains travaux de recherche.

BIBLIOGRAPHIE

[BAN.77] **GERALD BANON** (1977)
"Estimation non paramétrique de fonction de densité de probabilité par le processus de markov".
Thèse d'état Univ. P. Sabatier Toulouse.

[BAT.67] **C.G. BATTACHARYA** (1967).
" A simple method of resolution of a distribution into gaussian components". Biometrics.

[BEN.72] **J.P. BENZECRI** (1972)
"La regression".
Publication labo. statist. Math. Univ. PARIS VI.

[BON.85] **F. BONNEAU & J.M. PROTH** (1985)
"Analyse discriminante: Méthode du type plus proches voisins utilisant un prétraitement des données".
INRIA, n 440, pp2-6.

[BUC.29] **H.G. BUCHANON-WOLLASTON & W.G HODGESSON** (1929)
"A new method of treating frequency curves in fischery statistics, with some results".
Jour. cons. Vol.4, pp 207-225.

[CAZ. 76] **P. CAZES** (1976)
"Décomposition d'un histogramme en composantes gaussiennes". Rev. Stat. App. Vol.24, n 1, pp63-82.

[COO.64] **D.B. COOPER & P.W. COOPER** (1964)
"Nonsupervised adaptive signal detection and pattern recognition ". Info. & Control, Vol.7, pp416-444.

[COO.67] **P.W. COOPER** (1967)
"Some topics on nonsupervised adaptive signal detection for multivariate normal distributions".
Comp & Info. Sc. Vol.II, ppl23-146,
Academic Press, N. Y.

[COR.78] **J.M. CORNUET & J. FRESNAYE & P. LAVIE** (1978).
"Etude biométrique de deux populations d'abeilles cévénoles". Apidologie, Vol. 9(1) , pp 41-55.

[COV.67] **T.M. COVER & P.E. HART** (1967)
"Nearest neighbor pattern classification".
IEEE Trans. Info. Theory, Vol.IT-13, pp21-27.

[DAY.69] **N.E. DAY** (1970)
"Estimating the components of a mixture of normal distributions". Biometrika, Vol.56, pp463-474.

[DOE.36] **G. DOETSCH** (1936).
"Zerlegung einer funktion in gausche fehlerkuven und zeitliche zuruckverfolgung eines temperaturzustandes"
Math. Zeitschrift Vol.41, pp283-318;

[DUD.73] **R.O. DUDA & P.E. HART** (1973)
"Pattern Classification and scene analysis".
J. Willey, New-York, pp88,95.

[EIG.74] **D.J. EIGEN & F.R. FROMM & R.A. NORTHOUSE** (1974)
"Cluster analysis based on dimensional information with applications to feature selection and classification".
IEEE Tr on SMC, vol.4, n 3, pp284,294.

[FLE.69] **J.L. FLEISS & J. ZUBIN** (1969)
"On the method and theory of clustering".
Multivariate Behavioral Research pp235,250.

[FRE.65] **J. FRESNAYE** (1965).
"Etude biométrique de quelques caractères morphologiques de l'abeille noire française".
Ann. Abeille, Vol.8(4), pp271-283.

[FRE.75] **J.H. FREDMAN** (1975)
"An algorithm for finding nearest neighbors".
IEEE (octobre 1975).

[FRE.81] **J. FRESNAYE** (1981).
"Biométrie de l'abeille". I.N.R.A.

[FUK.73] **K. FUKUNAGA & P. NARENDRA** (1973)
"A branch and bound algorithm for computing k nearest neighbors". IEEE (mai 1973).

[HIL.68] **C.G. HILLBORN & D.G. LAINOTIS** (1968)
"Optimal unsupervised learning multicategory dependant hypotheses pattern recognition".
IEEE. Trans. Info. Theory, Vol.IT 14, pp468-470.

[KAZ.77] **D. KAZAKOS** (1977)
"Recursive estimation of prior probabilities using a mixture".
IEEE Trans. Info. Theory, Vol.IT-23, n 2, pp203-211.

[LOF.65] **D.O. LOFSTGAARDEN & C.P. QUESENBERY** (1965)
"A non parametric estimate of multivariate density function".
Ann. Math. Stat. Vol.36, pp1049-1051.

[MAC.51] **O. MACKENSON** (1951).
"Breeding improved Honey bees".
Amer. bee J. 91(7) ,p 292-294. (8) pp328-330.

[MAG.71] **T.J. MAGNER** (1971)
"Convergence of the nearest neighbor rule".
IEEE Trans. on info Theory Vol.IT-17, pp566-577.

[MAR.60] **T. MARILL & D.M. GREEN** (1960)
"Statistical recognition functions and the design of pattern recognizers".
IRE Trans. Elec. Comp. vol.EC-9, pp472-477.

[MEI.72] **W.S. MEISEL** (1972)
"Computer oriented approach to pattern recognition".
Academic press, New-York.

[MHI.83] **O. M'HIRIT & J.G. POSTAIRE** (1983)
" Analyse de la forme des tiges du cèdre du Maroc. Application à la détermination des courbes de profil". Ann. Sci. For. vol.40(4), pp353,372.

[MHI.84] **O. M'HIRIT & J.G. POSTAIRE** (1984)
"Analyse de la forme des tiges pour la construction des tarifs de cubage. Application au cèdre du Maroc". (Cedsus atlantica, Manetti).
Ann. Sci. For. vol.41(3), pp303,322.

[NOR.73] **R.A. NORTHOUSE & F.R. FROMM & D.J. EIGEN** (1973)
"A global-local approach to cluster analysis".
IEEE conf. on Decision and control, pp550,555.

[PAR.62] **E. PARZEN** (1962)
"On estimation of a probability density function and mode". Ann. Math. Stat. Vol.33, pp1065-1076.

[PEA.94] **K. PEARSON** (1894)
"Contribution to the mathematical theory of evolution".
Philo. Trans. Royal Soc. of London Vol.185, pp71-110.

[POS.81] **J. G. POSTAIRE** (1981)
"Optimisation du processus de classification Automatique par analyse de la convexité des fonctions de densité de probabilité".
Thèse Doc.es-Sciences Lille I.

[POS.82a] **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, pp357,379.

[POS.82b] **J.G. POSTAIRE** (1982).
"Unsupervised Bayes classifier for normal patterns based on marginal densities analysis".
Pattern Recognition, Vol.15, n 2, pp103-111.

[POS.83a] **J.G. POSTAIRE** (1983)
"Fonctions convexes et optimisation du processus de classification automatique:
II. Optimisation par analyse de la convexité des fonctions de densité marginale".
Rairo Automatique, Vol.17, n 1, pp39-59.

[POS.83b] **J.G. POSTAIRE & O. M'HIRIT** (1983)
"Application des techniques de classification automatique et de reconnaissance des formes à l'estimation du volume des arbres forestiers".
RAIRO Automatique vol.17, n 2, pp131,147.

[RAO.48] **C.R. RAO** (1948)
"Utilization of multiple measurements in problems of biological classification".
Jour. Rev. Statist. Soc. B. VolL.10, n 2, pp159-193.

[ROS.56] **A. ROSENBLATT** (1956)
"Remarks on some non parametric estimates of density function".
Ann. Math. Stat. Vol.27, pp832-837.

[ROS.83] **A. ROSENFELD & P. TORRE** (1983).
"Histogramme concavity analysis as an aid in treshold selection". IEEE. Trans. on Systems, Man and Cybernetics, Vol.SMC-13, n 3.

[RUT.78] **F. RUTTNER & L. TASSENCOURT & J. LOUVEAUX** (1978)
"biometrical-statistical analysis of the geographie variability of Apis mellifica".
Apidologie, Vol.9(4), pp363-381.

[RYZ.69] **J. VAN RYZIN** (1969)
"On strong consistency of density estimates".
Ann. Math. Vol.40, pp1765-1772.

[SCH.76] **A. SCHRODER** (1976)
"Analyse d'un mélange de distribution de probabilité de même type". Rev. Stat. App., Vol.24, n 1, pp32-62.

[SNE.71] **G.B. SNEDECOR & W.G. COCHRAN** (1971)
"Méthodes statistiques".
Ass. de Coord. Tech. Agr. (Ch.6).

[TAS.75] **L. TASSENCOURT & J.M CORNUET & J. FRESNAYE** (1975)
"Discrimination et classification de populations à partir de caractères biométriques ".
Apidologie, Vol.6(2), pp145-187

[TOM.75] **I. TOMEK** (1975)
"A modification of a clustering method".
IEEE Tr on SMC vol.5, n 3, pp394,396.

[WEG.72] **E.S. WEGMAN** (1972)
"Non parametric probability density estimation
I. A. Summary of available méthods".
Technometrics Vol.14, pp533-546.

[WOL.70] **J.H. WOLFE** (1970)
"Pattern clustering by multivariate mixture analysis".
Multiv. Behav. Res., Vol.5, pp329-350.

[YUN.76] **T. YUNEK** (1975)
"A technique to identify nearest neighbors ".
IEEE (octobre 1976).