DATA ANALYSIS AND INFORMATICS, III
E. Diday et al. (editors)

METHODES NON PARAMETRIQUES EN ANALYSE DISCRIMINANTE ; QUELQUES PROPOSITIONS NOUVELLES

Jean-Michel GAUTIER
ENSAE
3, Av. Pierre Larousse
92240 MALAKOFF

Gilbert SAPORTA
I.U.T. de Paris
143, Av. de Versailles
75016 PARIS

Résumé : Nous proposons une méthode générale non paramétrique s'appliquant aussi bien à des données quantitatives que qualitatives. Elle repose sur trois idées :

- une technique non paramétrique de sélection de variable dite méthode des fenêtres
- l'utilisation d'analyses factorielles par groupe pour réaliser la quantification des variables qualitatives, orthogonaliser les variables et améliorer les estimations de densité.
- l'emploi de noyaux produit gaussiens pour l'estimation de densité.

Summary

A non parametric general method is proposed for numerical data as well as for categorical ones. It is based on three ideas :

- a non parametric seclection of variables called the "window" method
- the use of separate principal components analyses, one for each group, in order to quantify the categorical variables, remove the correlation between numrical variables and improve density estimations
- the use of gaussian product kernels for density estimation.

I - LES METHODES CLASSIQUES ET LEURS LIMITES

1) Méthodes paramétriques

La grande majorité des méthodes de discrimination et des logiciels disponibles utilise ce que nous appellerons par commodité des méthodes paramétriques qu'elles soient d'inspiration géométrique (analyse factorielle discriminante, affectation selon la distance de Mahalanobis minimale aux centres des groupes ...) ou bayesiennes issues du modèle normal multidimensionnel.

a) L'approche géométrique

K groupes étant décrits par p variables numériques, les observations sont considérées comme des points de R^p. On affecte alors un individu $\underline{e}$ au groupe j tel que $d^2(\underline{e} ; \underline{g}_j)$ est minimal pour j = 1,2, ...k. Diverses propositions existent pour le choix de la formule de distance :

- métrique V^{-1} où V est la matrice de covariance totale
$d^2(\underline{e}_i \underline{g}_j) = (\underline{e}-\underline{g}_j)' V^{-1} (\underline{e} - \underline{g}_j)$

- métrique W^{-1} dite de Mahalanobis où W est la moyenne des matrices de covariance de chaque groupe $d^2(\underline{e};\underline{g}_j) = (\underline{e}-\underline{g}_j)' W^{-1} (\underline{e} -\underline{g})$

- métriques locales de Sebestyen $d^2(\underline{e} ; \underline{g}_j) = (\underline{e}-\underline{g}_j)' M_j (\underline{e} -\underline{g}_j)$ où M_j est proportionnel à V_j^{-1}.

Les métriques V^{-1} et W^{-1} ne se justifient en fait que si les V_j ne sont pas significativement différentes et aboutissent à des règles linéaires d'affectation ce qui explique leur usage très fréquent bien qu'en théorie limité aux lois normales. De plus des métriques différentes selon les groupes posent des problèmes de normalisation et se justifient pour des distributions normales de matrices de covariance différentes mais sont peu stables pour de petits échantillons.

La règle géométrique d'affectation procéde par tout ou rien et ne fournit pas d'indication sur les probabilités d'appartenance aux groupes.

b) L'hypothése de multinormalité

Elle est à la base des logiciels classiques (SAS, SPSS, BMDP ..) et permet d'obtenir des probabilités d'appartenance aux différents groupes en appliquant la formule de Bayes. Si f_j est la densité multinormale estimée dans le groupe j de probabilité p_j, on sait que la règle bayesienne revient à affecter une observation $\underline{x}$ au groupe j tel que $p_j f_j(\underline{x})$ est maximal puisque la probabilité a posteriori d'appartenance au groupe j est :

$$p_j f_j(\underline{x}) / \sum_{l=1}^{k} p_l f_l(\underline{x})$$

Lorsque les matrices de variances Σ_j sont égales on retrouve les règles linéaires (fonction de Fisher pour k=2) équivalentes à la minimisation de la distance de Mahalanobis aux centres de gravité sionon on obtient des règles basées sur les métriques locales V_j^{-1}

L'hypothése de multinormalité est ici cruciale et ne peut dont être utilisée sans précautions. De plus le cas des variables qualitatives ne peut être traité.

c) La méthode Disqual

Proposée par G. Saporta pour le traitement des prédicteurs qualitatifs, cette méthode est une extension des méthodes géométriques : par le biais d'une analyse des correspondances multiples sur l'ensemble des prédicteurs on se ramène au cas précédent en utilisant les composantes factorielles des individus comme variables numériques. Disqual utilise comme règle d'affectation la distance au centre de gravité dans un espace factoriel de dimensions réduites, avec la métrique V^{-1}. Ce choix, résultant d'une commodité de calcul puisque V = I pour des composantes normalisées, n'est pas optimal pour le cas de dispersions différentes dans chaque groupe ce qui se produit dans le cas d'interaction entre les variables explicatives et la variable à expliquer.

2) Méthodes non paramétriques

Conçues pour les cas ou l'hypothèse de normalité ne se justifie pas (et ils sont nombreux en pratique ...) elles ont été introduites récemment notamment par Hermans et Habbema (programme "Alloc"). Ces méthodes utilisent l'estimation de la densité dans chaque groupe par la méthode du noyau, dûe à Parzen.

Pour un seul prédicteur :

$$f_j(x) = \frac{1}{n_j h_j} \sum_{x_i \in \text{ groupe } j} K\left(\frac{x-x_j}{h}\right)$$

où K est une densité de probabilité et h_j une constante de lissage propre au groupe J.

Pour p prédicteurs la formule peut se généraliser en utilisant

$$f_j(\underline{x}) = \frac{1}{n_j h_j} \Sigma\, K\left(\frac{\underline{x}-\underline{x}_j}{h_j}\right)$$

où K est une densité p-dimensionnelle. On prend usuellement un noyau K produit de noyaux gaussiens à une dimension et la constante h est estimée par une méthode de pseudo-maximum de vraisemblance.

Indépendemment de la lourdeur des calculs pour de grands échantillons, on peut douter de l'efficacité de l'estimation qui suppose implicitement l'indépendance, au moins locale des variables explicatives : même en substituant non corrélation à indépendance, on sait que des variables globalement non correlées peuvent ne pas l'être dans chaque groupe et vice-versa.

Par ailleurs, la sélection d'un sous-ensemble de critères pose des problèmes car cele-ci suppose que l'on effectue la discrimination et se fait globalement : certaines variables retenues peuvent ne pas être pertinentes pour décrire un groupe particulier.

Aitchison et Aitken ont étendu aux variables binaires la méthode d'estimation par le noyau : si $\underline{x}_i$ est le vecteur des p variables binaires associées à l'individu i la densité en un vecteur $\underline{x}$ est estimée par

$$\frac{1}{h_j} \sum_i \lambda^{p-d(\underline{x}_i, \underline{x})} (1-\lambda)^{d(\underline{x}_i, \underline{x})}$$

où $d(\underline{x}_i ; \underline{x})$ est le nombre de désaccords entre $\underline{x}_i$ et $\underline{x}$ (le nombre de variables dont les valeurs différent pour l'individu i et celui dont on veut calculer la densité) λ est un paramètre de lissage à estimer.

Cette méthode ne s'étend que malaisément aux variables qualitatives à plus de 2 modalités et est criticable à notre avis au moins sur deux points : la distance retenue donne même importance à toutes les différences indépendamment des fréquences marginales des variables (un désaccord nous paraît d'autant plus important qu'il porte sur un critère plus répandu) et ne tient pas compte des dépendances entre variables.

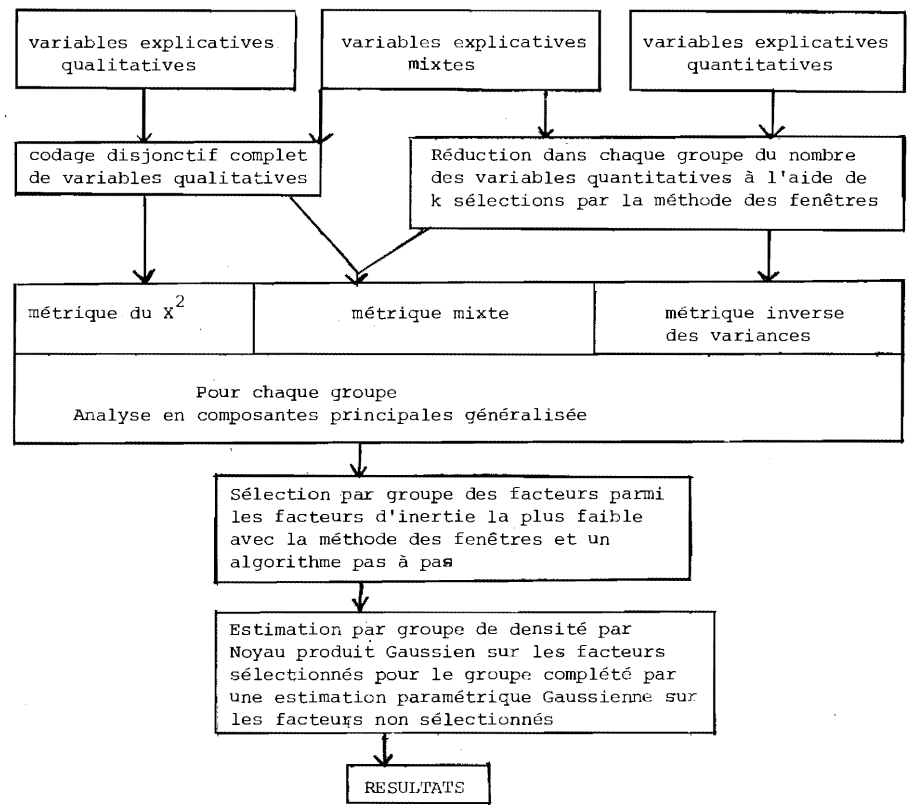
II - PROPOSITIONS NOUVELLES

1) La démarche

Nous sommes partis de la remarque suivante : une estimation de densité pour chaque groupe ne nécessite nullement un calcul non paramétrique (donc lourd) sur toutes les dimensions de l'espace des prédicteurs. On devrait donc notablement améliorer la procédure de discrimination en utilisant une estimation non paramétrique sur les variables décrivant le mieux chaque groupe et en complétant cette estimation par un modèle conditionnel paramétrique sur les autres variables. On pourra de plus choisir des variables orthogonales si l'on a recours au préalable, à une analyse factorielle. Le remplacement des prédicteurs initiaux par les composantes principales d'une analyse factorielle permet en choisissant une métrique adaptée de traiter aisément le cas des prédicteurs mixtes (quantitatifs et qualitatifs) et de justifier partiellement l'emploi de produits d'estimation (pour simplifier les calculs) dans l'estimation non paramétrique de densité. (Bien que évidemment non-corrélation n'implique pas indépendance).

Le schéma suivant précise la démarche proposée :

Méthode non paramétrique d'analyse discriminante à k groupes



Pour l'échantillon analysé

. Détermination des probabilités a posteriori

. Détermination d'un indice d'étrangeté des points

. Affectation bayésienne d'un point à un groupe

→ Sauvegarde des formules de densité pour le calcul des mêmes statistiques sur un échantillon test sur de nouvelles observations

et problèmes à résoudre pour la mise en oeuvre

Fichier de données	Prévoir la possibilité de pondérer les observations
Sélection par groupe de variables quantitatives	Algorithme rapide d'étalonnage d'une distribution Détermination d'une "bonne" taille pour les fenêtres critère d'arrêt pour la sélection
Analyses factorielles par groupe	Recherche de vecteurs propres associés aux axes d'inertie la plus faible choix d'une métrique adaptée
Sélection de facteurs pour chaque groupe	même remarques que pour la sélection locale de variables quantitatives : algorithme d'étalonnage, taille des fenêtres et critère d'arrêt.
Estimation de la densité	Détermination d'un "bon" paramètre de lissage Elaboration de fonctions de densité simplifiées pour le traitement des gros échantillons
Résultats	préciser des probabilités a posteriori et possibilité d'affecter un point à un groupe lorsque cette précision est mauvaise.

2) Méthode de sélection par fenêtre

L'idée directrice est de trouver une technique de sélection non paramétrique reposant sur un critère voisin du pourcentage de bien classés et évitant de faire l'analyse discriminante.

La sélection se faisant localement, les variables sélectionnées devront être adaptés à chaque groupe.

La méthode que nous proposons est la suivante :

Tout d'abord, on appelle fenêtre de taille α sur la variable j l'intervalle limité par les fractiles $\frac{1-\alpha}{2}$ et $\frac{1+\alpha}{2}$

fenêtre de taille α

Etant donné une groupe g et un ensemble de variables explicatives X_1, X_2....X_p (qui sont soit des variables de base, soit des composantes issues d'une analyse factorielle), la variable X_J permettant de caractériser le mieux ce groupe par rapport aux autres est celle qui réalise le maximum de l'expression suivante :

$$C_J(g) = \frac{N_\alpha^J(g)}{N_\alpha^J}$$

où N^J (g) représente le nombre de points du groupe g appartenant à une fenêtre de taille sur la variable J.

N^J représente le nombre total de points de l'echantillon analysé appartenant à la même fenêtre.

On note bien sur que $N_\alpha^j(g) = \alpha . N(g)$

où N(g) est l'effectif du groupe g.

Ce critère mesure la concentration relative des observations du groupe g sur la variable X_j par rapport aux observations des autres groupes.

La valeur maximale du critère est 1 et est atteinte si tous les points des groupes autres que g sont extérieurs à la fenêtre.

Le choix de α est un des paramètres de la méthode, à titre indicatif une valeur α=0,90 semble être raisonnable.

Ceci permet de sélectionner la première variable, les autres variables sont sélectionnées pas à pas selon le principe suivant :

En renumérotant les variables dans l'ordre de leur sélection la $(j+1)^e$ variable est celle qui maximise le critère suivant :

$$C_{1..j,l}(g) = \frac{N_\alpha^1 ..j,l(g)}{N_\alpha^1 ...;j,l} \quad \text{pour l variant de j+1 à p}$$

où N_α^1,..j,l(g) représente le nombre de points du groupe g appartenant à l'intersection des fenêtres de taille α sur les variables 1 , ..j et l.

N_α^1,..j,l (g) représente le nombre total de points de l'échantillon analysé appartenant à la même intersection.

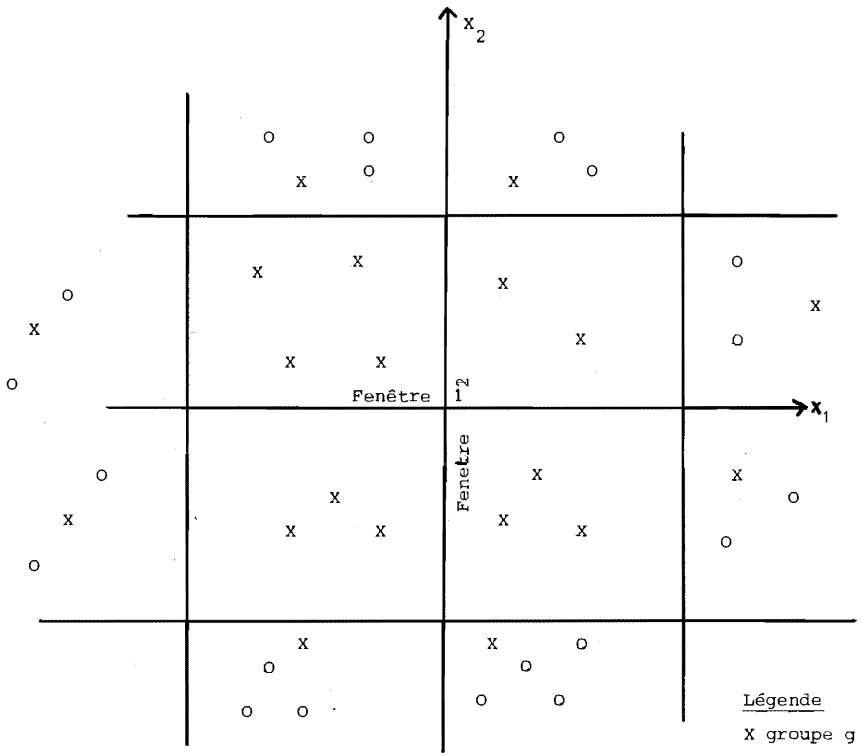
Ce critère revient à chercher à chaque étape la variable qui élimine proportionnellement le plus possible de points des autres groupes et le moins possible de points du groupe g lors de l'intersection de sa fenêtre de taille α avec les fenêtres précédentes.

Ce critère ne varie pas de manière monotone et n'exclue pas totalement les redondances même s'il est globalement plus efficace qu'une sélection indépendante sur la base des $C_j(g)$.

Remarque : Le nombre de points du groupe g présents à l'intersection des fenêtres décroit plus ou moins rapidement selon la valeur de α et le nombre de variables.

Pour un petit nombre de variables à sélectionner (5 ou 6) α=0,90 fournit de bons résultats.

Il ne semble pas possible de définir un test d'arrêt : le nombre de variables sélectionnées sera déterminé a priori soit interactivement en tenant compte de l'évolution des $C_{1,2,..j}(g)$ et des $N_\alpha^{1,2..j}(g)$



Exemple (fictif et idéal)

La fenêtre 1 (à 80%) contient 60 % d'individus des autres groupes, le critère $C_1 = 0,57$

L'intersection avec la fenêtre 2 (toujours à 80 % marginalement pour le groupe g) élimine 40 % des individus du groupe g mais tous les individus des autres groupes et C_{12} vaut 1.

3) Remarques sur l'utilisation d'analyses factorielles par groupes et les métriques utilisées

Des analyses factorielles effectuées séparément pour chaque groupe fournissent des systèmes de variables quantitatives non correlées bien adaptées à chaque groupe qui seront utilisées pour l'estimation de la fonction de densité du groupe par la méthode des noyaux produits.

Dans le cas où certains prédicteurs sont qualitatifs les analyses factorielles en fournissent une quantification. Lorsque tous les prédicteurs sont qualitatifs on utilisera la métrique du X^2, lorsque seuls certains prédicteurs sont qualitatifs on utilisera la méthode suivante :

Soit X le tableau des prédicteurs quantitatifs centré pour le groupe et $A = (A_1 \quad A_2 \quad \ldots \quad A_q)$ le tableau disjonctif des indicatrices des q prédicteurs qualitatifs, on effectuera l'ACP du tableau juxtaposé (X A) avec la métrique M suivante

$$M = \begin{pmatrix} \text{Diag } X'X & 0 \\ 0 & \text{Diag } A'A \end{pmatrix}^{-1}$$ ce qui revient à réduire les variables quantitatives et à utiliser la métrique du X^2 pour les indicatrices associées aux variables qualitatives.

En général le nombre des variables initiales (même après sélection préalable sur les variables quantitatives) sera trop important pour permettre un calcul aisé des fonctions de densité par la méthode des noyaux produits. On ne retiendra donc qu'un nombre limité de composantes principales. Pour les sélectionner on procédera comme au paragraphe II.2 en utilisant une procédure pas à pas basée sur les fenêtres.

En règle générale cette sélection fait apparaître les composantes associées aux plus faibles valeurs propres, car contrairement à l'habitude : le critère utilisé (sélection par fenêtre) tend à minimiser la dispersion des points du groupe, le long de l'axe sous réserve que les autres groupes soient dispersés le long de cet axe ce qui est souvent le cas. Si le nombre de composantes est trop élevé pour permettre une sélection sur l'ensemble on éliminera donc d'abord les composantes principales de plus forte variance.

4) Remarques sur l'estimation de densité

La fonction de densité pour chaque groupe sera estimée à partir des axes factoriels locaux du groupe, sélectionnés selon la méthode indiquée ci-dessus. Les variables étant orthogonales on se contentera pour estimer cette densité q-dimensionnelle du produit des q estimations de densité selon chaque variable.

Bien que le noyau d'Epanechnikov (ou noyau parabolique) ait des propriétés optimales en termes de MISE et soit souvent recommandé, nous préférons utiliser des noyaux gaussiens : en effet l'usage de noyaux bornés pose des problèmes

d'estimation dans les zones de faible densité et conduit à la limite à des estimations indéterminées $\left(\frac{0}{0}\right)$. Par ailleurs les noyaux gaussiens permettent d'obtenir des densités assez lisses.

Pour l'estimation de la constante de lissage on se reportera à la littérature (Habbema, Deheuvels..) il nous semble cependant préférable de procéder à une estimation empirique si l'on dispose de moyens de calculs interactifs avec visualisation des courbes de densité estimées.

On complètera ensuite pour chaque groupe l'estimation de densité obtenue avec la méthode des noyaux sur les facteurs sélectionnés par une estimation paramétrique gaussienne sur les autres facteurs, de façon à rendre comparable les densités estimées dans chaque groupe (même espace de référence).

Ainsi pour le groupe g,

. soient $F_1 \ldots F_e$ les facteurs sélectionnés

$F_{e+1}, \ldots F_p$ les autres facteurs de l'analyse factorielle effectuée sur le groupe $\lambda_1 \ldots \lambda_p$ les valeurs propres associées

. soit $\phi_g(\underline{x})$ la densité estimée au point $\underline{x} \in \mathbb{R}^p$ sur $F_1, \ldots F_l$ par la méthode des noyaux produits.

La densité du groupe g dans $\mathbb{R}^p$ au point x de coordonnées $F_1(x), \ldots F_p(x)$ peut s'écrire

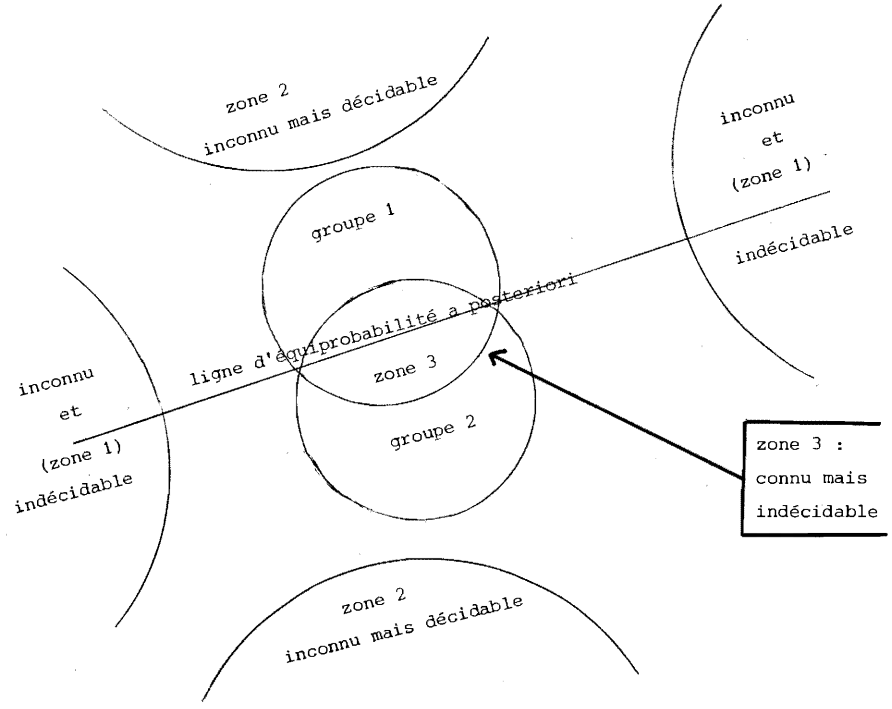
$$f_g(\underline{x}) = \phi_g(\underline{x}) \cdot \frac{e^{-\sum_{j=l+1}^{p} \frac{1}{2}\left(\frac{F_j(x)}{\sqrt{\lambda_j}}\right)}}{(2\pi)^{(p-l)} \prod_{j=l+1}^{p} \sqrt{\lambda_j}}$$

si l'on fait l'hypothèse que la loi conditionnelle de x sachant $(F_1, \ldots F_l)$ est un produit de gaussiennes sur $F_{l+1}, \ldots F_p$ de paramètres $(0, \sqrt{\lambda}_{l+1}) \ldots (0, \sqrt{\lambda_p})$

5) Affectation de nouveaux individus

La formule de Bayes fournit une estimation de la probabilité d'appartenance d'un individu à un groupe en utilisant les densités estimées. A l'évidence il existe des régions de l'espace (peu denses) où ces estimations seront peu fiables. Il n'existe malheureusement pas, à notre connaissance, de résultats concernant la précision des estimations de probabilités permettant de donner des intervalles de confiance.

L'exemple suivant permet de se faire une idée des problèmes :



Il est clair que plus un point s'éloigne des zones de grande densité de l'échantillon, plus ses probabilités d'appartenance aux 2 groupes sont entachées d'imprécision : cependant si une de ces probabilités est beaucoup plus forte que l'autre (c'est le cas de la zone 2) on peut quand même prendre une décision tandis que dans la zone 1 aucune décision raisonnable ne peut être prise.

Plus le nombre de groupes augmente plus les zones indécidables par imprécision sont importantes car le seul cas décidable sera celui d'un point dans une zone peu dense dont la probabilité d'appartenance à l'un des groupes serait très supérieure à toutes les autres.

Nous proposons donc de calculer, en plus des probabilités a posteriori un indice dit d'étrangeté I (x) qui sera d'autant plus grand que le point se situera dans une zone peu dense. Le problème serait élémentaire si on utilisait les mêmes variables pour calculer les densités dans tous les groupes, il suffirait de calculer la densité marginale et de la comparer à une densité moyenne.

Pour tourner la difficulté due au fait que nous utilisons des variables

locales nous proposons l'indice suivant :

$$\frac{1}{l(x)} = \underset{g}{\text{Max}} \left(\text{Max } R_g^j (x)\right)$$

ou $R_g^j (x) = \dfrac{\text{densité marginale en x sur la variable j du groupe g}}{\text{densité moyenne sur la variable j du groupe g}}$

Cet indice est d'autant plus grand que sur le "meilleur" groupe et la "meilleure" variable locale pour ce groupe, la densité relative au point x est faible.

6) Cas de gros échantillons : calculs simplifiés

Il est nécessaire dans ce cas de disposer de formules rapides d'estimation de probabilité qui évitent de faire défiler le fichier pour chaque nouvel individu.

Nous proposons les procédures suivantes :

a) Formule simplifiée d'estimation de densité :

Pour chaque groupe g et sur chaque axe local

- on détermine un étalonnage en centiles de la distribution des points de g sur l'axe j
- on calcule le point moyen (centre de gravité) de chaque intervalle intercentiles
- on estime la densité à partir des points suivants :
  - . points externes du groupe g correspondant aux 1 ou 2 premiers et derniers centiles muni d'un poids 1
  - . centres de gravité des intervalles intercentiles muni d'un poids $\frac{ng}{100}$.

b) Formule simplifiée d'affectation

Valable lorsque l'échantillon analysé n'est pas trop gros, donc dispensé d'une formule simplifiée de densité, mais lorsqu'il y a beaucoup de point à affecter.

Il suffit alors de remplacer la fonction de densité par une fonction en escalier sur la fenêtre à 90 % et de ne faire l'estimation de densité que lorsque le point est extérieur à cette fenêtre.

D'autres améliorations sont rendues indispensables pour le traitement des gros échantillons comme l'adoption d'un algorithme rapide d'étalonnage pour la construction des fenêtres.

REFERENCES

AITCHISON J. AITKEN CGG (1976)
Multivariate binary discrimination by the kernel method Biometrika, 63, 413-420.

DEHEUVELS P. HOMINAL P. (1979)
Estimation non paramétrique de la densité compte tenu d'information sur le support. Rev. Stat. Appli., 27, 47-68.

GAUTIER J.M. SAPORTA G. (1979)
Une méthodologie de discrimination sur variables qualitatives. Actes 2e Congrès AFCET-INRIA. Reconnaissance des Formes et Intelligence Artificielle. Tome I 320-377.

HABBEMA J.D.F. HERMANS J.VAN DEN BROEK K. (1974)
A Stepwise discriminant analysis program using density estimation ; in Compstat 74, Physica Verlag, Wien, 101-110.

KSHIRSAGAR A.M. (1972) Multivariate analysis, Marcel Dekker, New York

NAKACHE J.P. (1980)
Méthodes de discrimination sur variables de nature quelconque, théorie et pratique. Thèse Doctorat es Sciences. Université Paris VI.

SAPORTA G. (1977)
Une méthode et un programme d'analyse discriminante pas à pas sur variables qualitatives. Actes 1ère journées Analyse de Données et Informatique IRIA 201-210.