

# REVUE DE STATISTIQUE APPLIQUÉE

J. Y. LAFAYE

## **Une méthode de discrétisation de variables continues**

*Revue de statistique appliquée*, tome 27, n° 2 (1979), p. 39-53

[http://www.numdam.org/item?id=RSA\\_1979\\_\\_27\\_2\\_39\\_0](http://www.numdam.org/item?id=RSA_1979__27_2_39_0)

© Société française de statistique, 1979, tous droits réservés.

L'accès aux archives de la revue « Revue de statistique appliquée » (<http://www.sfds.asso.fr/publicat/rsa.htm>) implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme  
Numérisation de documents anciens mathématiques  
<http://www.numdam.org/>

# UNE MÉTHODE DE DISCRÉTISATION DE VARIABLES CONTINUES

J. Y. LAFAYE

I.U.T. Département "STATISTIQUES" Kercado B.P. 1104 — 56008 VANNES

## RESUME

La question soulevée est celle de la transformation d'un tableau de données initialement numériques en un tableau disjonctif complet ou encore en un tableau de notes, relativement à un ensemble de variables préordinales.

Nous envisageons une telle transformation en préalable à des traitements d'analyse de données de type métriques ou combinatoires.

Il s'agit donc de trouver une subdivision "optimale" de l'intervalle de variation de la variable numérique initiale, et d'identifier cette variable avec un caractère dont les différentes modalités correspondraient aux sous-intervalles ainsi mis en évidence.

De façon générale, le problème se pose en termes de séparation de mélanges de lois de probabilité.

Après une revue bibliographique, nous proposons une méthode originale pour obtenir la subdivision cherchée dans une optique non paramétrique.

Nous terminons par l'exposé d'applications pratiques dans le domaine médical et par des résultats concernant des mélanges obtenus par simulation.

## 1. INTRODUCTION

Le caractère peu restrictif des contraintes pesant sur la forme des matrices de similarités dans l'analyse classificatoire des données a permis la définition d'une grande variété d'indice de similarité aussi bien entre variables qu'entre objets. Ces différents indices font référence à divers types de codage.

La liberté de choix d'une forme particulière de similarité requiert donc la capacité d'accéder, à partir des données brutes, à d'autres sortes de représentations.

Nous nous intéressons ici à l'appauvrissement d'une structure riche, variable numérique en l'occurrence. Bien que la richesse de l'information contenue dans les données brutes soit en général une qualité à préserver, il peut être intéressant de procéder néanmoins à une telle opération.

— Ce sera le cas si la précision de la valeur réelle observée ne constitue pas à proprement parler une "information", et si seul un ordre de grandeur est effectivement significatif.

— La discrétisation, conduisant à un codage par des nombres sans dimension, peut être un moyen de réduire l'hétérogénéité d'un ensemble de mesures aux unités trop diverses.

— Enfin, différentes formes d'appréhension d'un même ensemble de données peuvent apporter des éclairages variés et d'un égal intérêt quant à la connaissance des échantillons classifiés. L'exploration de ces différents points de vue nécessite de pouvoir passer au mieux d'un mode de représentation à un autre.

## 2. FORMALISATION DU PROBLEME

Etant donné une variable numérique "v" à valeurs dans l'intervalle I : [a, b [

$$v \in V \quad v : E \quad I \subset \mathbb{R} \\ x \rightarrow v(x)$$

il s'agit de trouver une suite finie strictement croissante

$$\{x_j\}_{j=0,p} \quad \text{où} \quad x_0 = a \quad \text{et} \quad x_p = b$$

Le caractère "c" associé à "v" possède ainsi p modalités

$$\{c_j\}_{j=1,p}$$

et l'on a :

$$C(x) = C_j \Leftrightarrow v(x) \in [x_{j-1} ; x_j[$$

Selon que l'on retient ou non un ordre sur  $\{C_j\}$  on accède à une représentation de variable numérique par une variable préordinaire ou par un caractère à modalités sans structure.

La définition complète de la suite  $\{x_j\}_{j=0,p}$  suppose donc :

- le choix de la valeur p
- le choix des valeurs  $x_j$

pour

$$j = 1 ; (p - 1).$$

Une méthode qui a le mérite de la simplicité est de choisir "p" de façon arbitraire, puis d'opérer une subdivision en p sous-intervalles d'égale amplitude.

Une façon d'agir plus élaborée et qui fait référence à un souci de sauvegarde d'information est d'opérer une subdivision en p intervalles d'effectifs égaux (22).

Cependant, analysant une population hétérogène, si l'on désire que chaque sous-intervalle soit en accord avec la distribution indiquée par l'échantillon et qu'il soit éventuellement possible d'interpréter chacune des modalités des caractères résultants comme correspondant à une sous-population homogène, le problème se pose naturellement en termes de séparation de mélange de lois de probabilité.

Chaque composant unimodal constitue une entité logique que doit respecter un découpage pertinent.

D'une façon formelle, le problème se présente comme suit :

Etant donné une variable aléatoire "X", on considère un échantillon de cette variable comme résultat de tirages successifs dans différentes populations homogènes  $\{\mathcal{P}_i\}_{i=1;k}$

A chaque population  $\mathcal{P}_i$  correspond une distribution unimodale " $\varphi_i$ " de la variable observée "X", ainsi qu'un nombre "p<sub>i</sub>" compris entre 0 et 1, caractéristique de la fréquence relative avec laquelle le tirage effectué s'adresse à la population  $\mathcal{P}_i$ .

C'est la distribution résultante "f" qu'il s'agit de décomposer.

f s'exprime comme une combinaison linéaire convexe de  $\varphi_i, i = 1 ; k$ .

La résolution complète du problème consiste à :

- évaluer  $k$
- préciser le modèle, c'est-à-dire la nature des différents  $\varphi_i$   $i = 1, k$
- estimer les paramètres de chacun des composants.
- préciser le système des poids  $p_i$   $i = 1, k$ .
- déduire de la connaissance des différents composants une subdivision les respectant "au mieux" suivant un critère qu'il s'agit de préciser. Une méthode naturelle peut être de choisir une borne  $x_j$  séparant les réalisations provenant de  $\mathcal{R}_{j-1}$  de celles provenant de  $\mathcal{R}_j$  ( $j = 1, p$ ) la valeur minimisant le risque d'affectation erronée d'une observation à  $\mathcal{R}_{j-1}$  ou à  $\mathcal{R}_j$ . On peut également chercher à optimiser les critères d'inertie par exemple. . .

Signalons que la question de l'existence et de l'unicité d'une solution, au problème de la décomposition des mélanges a été traitée de façon détaillée par Teicher (1963) puis Yakowitz et Spragins (1968).

### 3. DECOMPOSITION DE MELANGES, DISCRETISATION – APERCU DES DIVERSES METHODES

#### a) Méthodes d'estimation classique

Nous dirons peu de mots de ces méthodes désormais classiques qui ont constitué la toute première approche rationnelle de la question. Que ce soit la méthode des moments élaborée par Pearson (1894-1915) limitée à l'origine au cas de distribution gaussiennes puis étendue à un cas plus général par Blischke (1962), ou encore des méthodes utilisant les estimations du maximum de vraisemblance développées par Rao (1948) ou Day (1969).

Si les méthodes sont aussi variées que le sont les méthodes d'estimation elles-mêmes,  $\chi^2$  minimum, estimateurs bayésiens, etc. il reste qu'aucune n'échappe à une complexité rapidement inextricable lorsque se voit augmenté le nombre des composants.

Ainsi les méthodes liées au maximum de vraisemblance se limitent-elles au traitement de deux constituants gaussiens de même variance.

Les éventuelles extensions ne concernent en tout état de cause qu'un nombre très limité de composants et n'abordent jamais le cas multivarié. Elles sont en conséquence de peu d'intérêt pratique.

#### b) Méthodes bayésiennes avec apprentissage

Ce type de méthodes est basé sur des algorithmes d'estimation stochastique des paramètres inconnus du modèle, les observations étant prises en compte une à une. Le critère d'arrêt de l'algorithme étant l'obtention d'une certaine stabilité dans les estimations récursives effectuées à chaque fois.

Une telle optique a été formalisée à l'origine en termes d'apprentissage sans maître par Patrick et Costello (1970) ou encore en termes d'apprentissage avec maître par Abramson et Braverman (1962) et enfin, plus récemment comme "learning with a probabilistic teacher" avec Agralawa (1970).

Ces méthodes, d'un intérêt théorique certain requièrent toutes les hypothèses très restrictives, ce qui réduit considérablement leur domaine d'application. De plus, elles demandent dans leur ensemble des échantillons de taille importante.

Le volume des calculs étant rapidement croissant avec le nombre d'estimations à effectuer, ce genre de technique est mieux adapté aux problèmes où le nombre de paramètres inconnus est faible, soit que l'on ait une bonne connaissance a priori du modèle.

Nous terminerons en signalant un inventaire détaillé de ce type de méthode et de celui traité au précédent paramètre dû à Dorofeyuk (1971).

### c) Algorithmes optimisant un critère

Les deux précédentes classes de méthodes répondent en premier lieu au souci d'estimer les paramètres du modèle, et permettent, à la lumière de ces résultats d'adapter une règle de décision quant au choix des bornes de la subdivision cherchée.

La classe d'algorithmes que nous envisageons ici chemine de façon inverse au niveau conceptuel du moins. On envisage d'abord de procéder à l'allocation de chaque observation à un composant particulier (l'allocation pouvant être remise en cause lors du déroulement de l'algorithme).

Une fois le processus d'allocation terminé, on procède à l'estimation des paramètres.

Dans notre optique propre, le choix de la subdivision peut être envisagée de deux façons, selon que l'on se base comme précédemment sur la connaissance des paramètres du modèle, soit que l'on utilise directement les résultats concernant la séparation des constituants.

Les critères optimisés lors de la procédure de réallocation peuvent être variés ; variés également les algorithmes optimisant un même critère.

Que ce soit un critère de type géométrique optimisant le rapport de l'inertie inter groupes à l'inertie intra groupes telle la méthode de Fischer (1958) basée sur la programmation dynamique.

Cette méthode classique donne lieu à d'intéressantes généralisations de la part de Vinod (1969) en termes de programmation linéaire et Le Chevallier (1977) développant la méthode de Vinod dans une optique non paramétrique.

Dans le même ordre d'idées Young et Corraluppi (1970) proposent une procédure récursive d'optimisation d'un critère d'information pour séparer un mélange gaussien multivarié sans qu'il soit nécessaire de connaître a priori le nombre des constituants.

Enfin, la séparation peut être réalisée par une technique de classification automatique appliquée sur un espace unidimensionnel. Dans ce cadre, A. Schroeder (1976) propose une approche basée sur la méthode des nuées dynamiques optimisant un critère de vraisemblance. Cette technique a l'avantage de permettre la décomposition de mélanges non gaussiens, les " $\varphi_i$ " devant cependant être distribués suivant des lois de probabilité de même types.

Une méthode itérative d'un genre différent est proposée par Benzecri et traitée par Cazes (1976). L'algorithme procède par déconvolutions successives de la densité  $f$  résultante par des densités normales centrées.

### d) Méthodes graphiques

La difficulté à vérifier ou à admettre des hypothèses souvent très restrictives, la lourdeur des calculs à mettre en œuvre, ont poussé à la pratique de méthodes plus ou moins empiriques basées sur des représentations graphiques.

Dans le cas où les composants sont bien séparés et le nombre d'observations suffisant, la seule considération de l'histogramme fournit une estimation correcte du nombre des constituants et de leurs valeurs centrales. L'estimation de la variance est difficile.

Souvent l'histogramme des fréquences est inadéquat. La démarche des méthodes graphiques est de chercher un argument de décision mieux adapté. Dans le cas de mélanges gaussiens, on procède à des transformations sur  $f$  qui aboutissent à l'observation de fonctions quadratiques (24) ou linéaires (3), (5), de la variable.

L'ajustement de paraboles ou de droites permet de conclure quant au nombre de composants et à la valeur des deux premiers moments de chacun.

#### e) Tests de bimodalité

La plupart des méthodes précédemment exposées nécessitent la connaissance a priori du nombre de constituants. Certaines permettent de répondre plus ou moins aisément à la question du dénombrement des modes significatifs correspondant à une sous population homogène, par opposition à ceux erratiques, dus à l'échantillonnage.

Certains procédés spécifiques ont été développés sous la forme de tests de bimodalité.

Engelman et Hartigan (1969) proposent un test basé sur une statistique de rapport de vraisemblance qui a l'inconvénient de stipuler des hypothèses très restrictives. La population mère est sous l'hypothèse nulle gaussienne ( $\mu, \sigma$ ) et, sous l'hypothèse alternative, un mélange de deux sous-populations elles aussi gaussiennes de moyennes  $\mu_1$  et  $\mu_2$  et de même écart type  $= \sigma$ .

Des développements d'applications plus générale sont fournis par Hartigan (1975) qui, étudie les propriétés asymptotiques du test précédent dans le cas où la population mère est arbitraire.

Nous citerons également les travaux de Giacomelli (1971). Et ceux de Hartigan (1977) introduisant la notion de "dip statistic" liée à la mesure de la densité dans les algorithmes de type "single linkage". Hartigan fournit à ce sujet une table de valeurs critiques de la "dip statistic" pour différents seuils et différentes tailles d'échantillons.

## 4 – UNE NOUVELLE METHODE D'AIDE A LA DISCRETISATION DES VARIABLES CONTINUES

### a) But et optique générale

Les méthodes précédentes se sont avérées mal adaptées au traitement préalable d'un fichier médical que nous devons étudier par une série d'analyses classificatoires (15). Ceci en raison du coût qui ne se justifiait pas dans la mesure où la décomposition des mélanges n'était pas le but final, mais uniquement une première étape de l'analyse, et surtout en raison de la taille très réduite du fichier disponible.

Ainsi, avons-nous cherché à mettre au point une technique de discrétisation efficace sur de petits échantillons, présentant en outre des qualités de souplesse de rapidité et nécessitant une place mémoire modeste.

Ces buts peuvent être atteints dans la mesure où l'on accepte de réduire l'ambition de l'estimation de chacun des paramètres du mélange analysé à la simple

séparation des constituants. Ainsi développerons-nous par la suite un algorithme de discrétisation non paramétrique dont I.C. LERMAN nous a fourni l'idée de départ.

### b) Position du problème

Nous nous proposons de mettre en évidence les minima de la densité "f" résultant du mélange. Ceux-ci, une fois reconnus indiqueront simultanément le nombre "k" des composants ainsi que les valeurs  $\{x_j/j = 0, k\}$  des limites de chaque sous intervalle pertinent dans l'intervalle de variation de la variable mesurée.

On peut objecter que les minima de f ne correspondent pas précisément avec le système de bornes optimisant un critère particulier : inertie (10) ou affectation de risque d'erreur minimum.

En effet, cependant, le découpage basé sur les minima de f reste intuitivement justifiée (24). En outre, il constitue dans la majorité des cas une approximation correcte des limites définies par les critères ci-dessus ; en particulier, si l'on fait référence à un modèle où les dispersions sont comparables, les pondérations de même ordre et les dissymétries non excessives. Ces dernières conditions sont peu restrictives vis-à-vis d'un modèle paramétrique.

Notons pour conclure que si l'on considère un critère de risque d'affectation erronée minimum, le choix des minima de f conduit à un nombre de mal classés d'autant plus réduit que le découpage s'effectue dans une zone de faible densité pour chacun des " $\varphi_i$ ". Ceci est d'autant plus vrai que l'on se place dans le cas d'échantillons de petites tailles.

Quoi qu'il en soit, cette approximation est nécessaire si l'on désire éviter la référence à un modèle paramétrique et les lourds calculs que nécessitent les estimations qui en découlent.

### c) Fenêtre mobile

Si dans certains cas, l'histogramme des fréquences permet la localisation des extréma de f, le plus souvent, il s'avère inadapté ; et cela d'autant plus que nous nous intéressons à de petits échantillons.

A l'image des démarches adaptées par BHATTACHARYA ou TANAKA, il s'agit de fournir un argument de décision plus adéquat que le simple histogramme.

A partir de la suite ordonnée des observations, I.C. LERMAN propose, étant donné un réel positif "l" d'imaginer une "fenêtre" de largeur "l" centrée en un point "θ", on calcule la valeur  $F_l(\theta)$  égale au nombre de réalisations dans la fenêtre.

Se donnant une fonction de densité "f", nous définissons de cette manière une famille d'applications de  $\mathbf{R}$  dans  $\mathbf{N}$  :  $\{F_l/l \in \mathbf{R}^+\}$  indicée par le réel positif "l".

$$\begin{aligned} l \in \mathbf{R}^+ \quad F_l : \mathbf{R} &\rightarrow \mathbf{N} \\ \theta &\rightarrow F_l(\theta) \end{aligned}$$

$F_l(\theta)$  est la "l-densité" au point  $\theta$  et correspond à l'intégrale de la densité théorique sur le segment  $\left[ \theta - \frac{l}{2} ; \theta + \frac{l}{2} \right]$ . Autrement dit, si F est une primitive de f :

$$F_l(\theta) = F\left(\theta + \frac{l}{2}\right) - F\left(\theta - \frac{l}{2}\right).$$

De façon analogue, nous pouvons associer à chaque application  $F_\varrho$  un opérateur  $\tilde{F}_\varrho$  sur l'ensemble  $\mathfrak{F}(\mathbf{R}, \mathbf{R})$  des applications réelles d'une variable réelle.

$$\begin{aligned} \tilde{F}_\varrho : \mathfrak{F}(\mathbf{R}, \mathbf{R}) &\rightarrow \mathfrak{F}(\mathbf{R}, \mathbf{R}) \\ f &\rightarrow \tilde{F}_\varrho(f) \end{aligned}$$

où  $F_\varrho(f)$  est défini en tout point  $t$  de  $\mathbf{R}$  par l'égalité

$$\tilde{F}_\varrho(f) \cdot t = F_\varrho(t)$$

Nous noterons par la suite et sans ambiguïté de la même façon  $\tilde{F}_\varrho$  et  $F_\varrho$ , à savoir : " $F_\varrho$ ".

A toute fonction de densité continue  $f$ , nous sommes en mesure d'associer une famille de fonctions de  $\varrho$ -densité  $\{F_\varrho(f)/\varrho \in \mathbf{R}^+\}$ , elles aussi continues.

D'un point de vue expérimental, de tout échantillon issu d'une population distribuée selon la densité  $f$ , il est immédiatement possible de déduire un échantillon d'un quelconque élément de la famille

$$\{F_\varrho(f) ; \varrho \in \mathbf{R}^+\}$$

*Cas extrêmes, exemples :*

Soit  $\{t_j/j = 1, n\}$  une suite d'observations supposée ordonnée et croissante, et  $t_0$  une valeur plus petite que  $t_1$ .

Soit de plus :  $d_m = \min \{t_{j+1} - t_j/j = 0, n - 1 ; t_{j+1} \neq t_j\}$

et

$$d_M = t_n - t_0$$

*1er cas :  $l \leq d_m$*

la représentation de  $F_\varrho(t)$ ,  $t \in \mathbf{R}$  correspond exactement avec l'histogramme le plus fin compatible avec la précision des mesures (fig. 1).

*2e cas :  $l > d_M$*

la représentation de  $F_l(t)$  ;  $t \in \mathbf{R}$  est unimodale et les deux parties monotones correspondent à l'histogramme des fréquences cumulées. L'amplitude des classes  $y$  est définie par la précision des mesures (fig. 2).

*par exemple :*

soit la suite :  $\{1, 3, 3, 3, 4, 5, 7, 7, 8, 10\}$

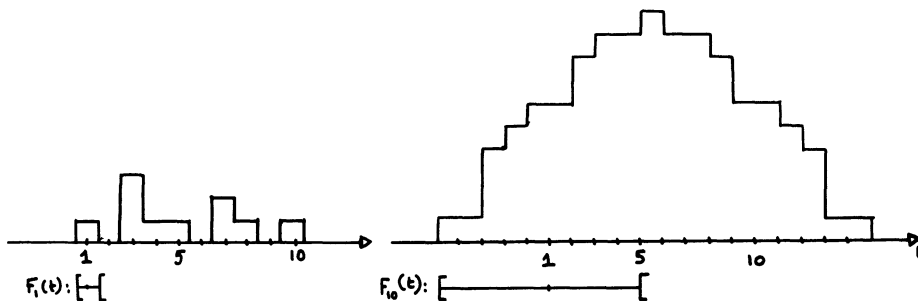


Fig. 1 et 2



Entre ces deux extrêmes, courbe très plurimodale pour  $\ell \leq d_m$  et unimodale pour  $\ell > d_M$ , les éléments de  $\{F_\ell, \ell \in \mathbb{R}^+\}$  passent par des états intermédiaires. C'est précisément l'étude de tels états qui permet la discrétisation par le repérage des extréma significatifs de  $f$ .

d) Caractérisation des extrema de "f"

Les extréma d'un élément particulier  $\{F_\ell(\theta), \theta \in \mathbb{R}\}$  sont sujets à une caractérisation analytique à partir de fonctions se déduisant de  $f$  par translation.

Soit respectivement  $f_{\ell/2}$  et  $\bar{f}_{\ell/2}$  les translatées de  $\bar{f}$  par  $(\ell/2)$  et  $(-\ell/2)$  :

$$\forall x \in \mathbb{R} : f_{\ell/2}(x) = f(x + \ell/2) ; \bar{f}_{\ell/2}(x) = f(x - \ell/2)$$

Soit  $\Delta^\ell$  l'opérateur sur  $\mathcal{F}(\mathbb{R}; \mathbb{R})$  définit comme suit

$$\begin{aligned} \Delta^\ell : \mathcal{F}(\mathbb{R}; \mathbb{R}) &\rightarrow \mathcal{F}(\mathbb{R}; \mathbb{R}) \\ f &\Delta^\ell(f) = f_{\ell/2} - \bar{f}_{\ell/2} \end{aligned}$$

On montre l'identité

$$\frac{dF_\ell}{d\theta}(t^*) = 0 \Leftrightarrow \Delta^\ell(f) \cdot t^* = 0$$

qui caractérise de deux façons équivalents les localisations  $t^*$  des extréma de  $F_\ell$

$\Delta^\ell(f) \cdot t^* = 0$  caractérise également, pour de faibles valeurs de  $\ell$ , les extréma de  $f$ . On déduit donc de ce qui précède que jusqu'à une certaine valeur limite " $\ell_0$ " majorant les valeurs " $\ell$ ", à chaque extremum de  $f$  correspond un extremum de  $F_\ell$  et réciproquement. La valeur  $\ell_0$  peut être précisée analytiquement dès que l'on connaît la densité  $f$  (15).

Les cas intéressants dans notre optique de discrétisation sont ceux où la réciproque est fautive, c'est-à-dire pour des valeurs " $\ell$ " excédant " $\ell_0$ ".

En effet, la considération de  $F_\ell$  au lieu de  $f$  revient à opérer un lissage ou mieux un "filtrage" sur les différents modes de  $f$ , ne laissant subsister que les plus importants, et absorbant les fluctuations dues aux modes mineurs. Ces notions relatives à l'importance relative des extréma sont précisées dans (15). Plutôt que de reprendre ici un formalisme assez lourd, nous nous bornerons à une définition intuitive de "l'amplitude" d'un minimum ou d'un mode telle qu'elle apparaît sur la figure 3.

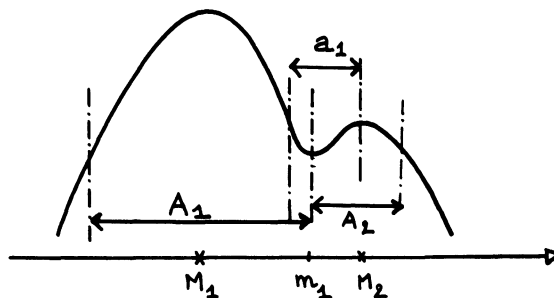


figure 3

Les modes d'abscisses  $M_1, M_2$  et le minimum d'abscisse  $m_1$  ont respectivement pour amplitude les quantités  $A_1, A_2, a_1$ . Elles correspondent schématiquement aux longueurs délimitées par la courbe étudiée sur ses tangentes aux points d'abscisses " $m_1$ " et " $M_2$ ".

Précisément, soit une densité  $f$  présentant une suite de modes  $\{M_i/i = 1, h\}$  d'amplitudes  $\{A_i/i = 1, h\}$  et une suite de minima  $\{m_j/j = 0, h\}$  d'amplitudes  $\{a_j/j = 0, h\}$ . Le mode de rang  $k$  sera absorbé dans la représentation  $F_\ell$  si " $\ell$ " excède une valeur critique  $\delta_k$  égale à :

$$[A_k + \min \{a_{k-1} ; a_{k+1}\}]$$

Pratiquement, l'absorption s'effectue souvent plus tôt pour une valeur  $\ell$  de l'ordre de  $\delta_k = \min \{A_k ; a_{k-1} ; a_{k+1}\}$

Les figures 4 et 5 montrent comment s'opère le filtrage des modes mineurs.

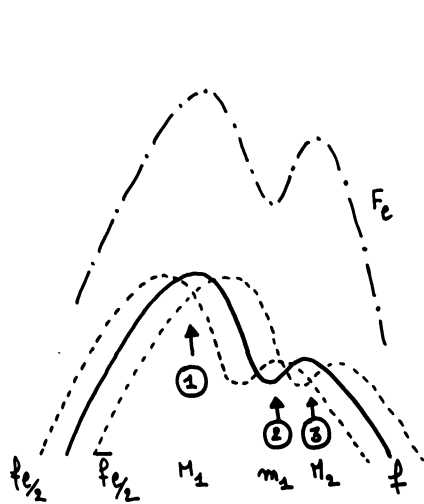


fig. 4. Deux modes sont détectés

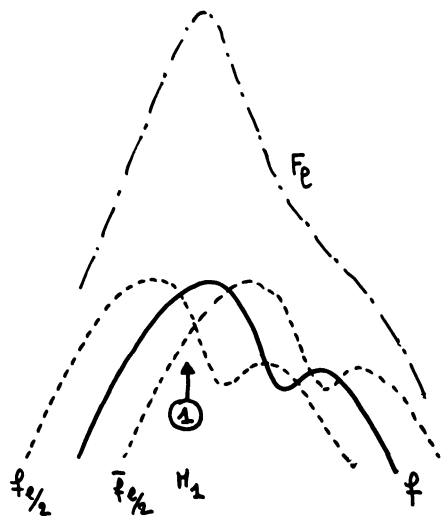


fig. 5. Un seul mode est détecté

### c) Pratique de la discrétisation

Les considérations qui précèdent ont fait référence à des  $\ell$ -densités  $\{F_\ell/\ell \in \mathbb{R}^+\}$ . Pratiquement, on se limite à l'observation d'une suite  $\{F_{\ell_i}/i = 1, N\}$  pour différentes valeurs  $\ell_i$  de la fenêtre (on suppose  $\ell_{i+1} > \ell_i$ ). De même on se limite à l'observation des  $F_{\ell_i}$  en une suite fixe de points :  $\{t_h/h \in \mathbb{N}\}$ . Finalement on observe donc la suite des représentations  $\{\{F_{\ell_i}(t_h)/h \in \mathbb{N}\}/i = 1, N\}$ .

#### Choix de la suite $\{t_h/h \in \mathbb{N}\}$

Le choix le plus naturel est de prendre  $t_h = t_0 + \alpha h$  ou  $t_0$  est une valeur arbitraire inférieure à la plus petite réalisation observée dans l'échantillon. " $\alpha$ " est la distance fixe séparant  $t_h$  de  $t_{h+1}$  et constitue le "pas" d'avancement de la fenêtre. Il se doit d'être assez réduit pour permettre une bonne approximation de  $F_{\ell_i}$  sans excès cependant pour conserver quelques lisibilité au graphique.

A titre d'indication,  $d_m$ ,  $2d_m$ ,  $3d_m$  sont des valeurs expérimentalement raisonnables pour  $\alpha$ .

#### Choix de la suite $\{\ell_i/i = 1, N\}$

L'expérience permet de conseiller  $N \approx 10$  et  $\ell_1 \approx d_M/20$  tandis que  $\ell_N \approx d_M/6$ . On notera que l'on doit veiller à respecter la contrainte  $\alpha < \ell$ . Le

choix  $\alpha = \ell$  conduit à "oublier" des observations et est en conséquence à rejeter ; à fortiori si l'échantillon est réduit.

### *Démarche détaillée*

- I : Obtenir la suite  $\{u_j/j = 1, n\}$  des observations ordonnées par valeurs croissantes.
- II : en déduire les valeurs  $d_m$  et  $d_M$
- III : choisir avec les règles énoncées plus haut une suite croissante  $\{\ell_i/i = 1, N\}$
- IV : choisir une fois pour toute la valeur  $\alpha$  du pas d'avancement de la fenêtre ( $\alpha < \ell_1$ )
- V : choisir pour centre de la première fenêtre ( $t_0$ ) la valeur  $\left(\mu_1 - \frac{\alpha}{2}\right)$
- VI : représenter la suite des courbes  $F_{\ell_i}$  en tout point de  $\{t_0 + h\alpha\}_{h=0, \epsilon}$  où  $\epsilon =$  partie entière de  $\left(\frac{d_M}{\alpha}\right)$
- VII : déduire de l'observation de la suite des extrema des différents  $F_{\ell_i}$  les valeurs de la suite  $\{x_j\}_{j=0, k}$  permettant le découpage cherché.

Pratiquement, l'observation successive des courbes  $F_{\ell_i}$ , montre une évolution du nombre, de la localisation et de l'importance relative des extréma.

– Dans le cas de composants bien séparés, il est aisé de reconnaître la longueur "optimale"  $\ell_i^*$  permettant une nette mise en évidence des constituants.

– Dans les cas plus confus, il est nécessaire de procéder à l'examen simultané de plusieurs  $F_{\ell_i}$

Lorsque  $i$  augmente, on observe une accentuation de certains extrema dont la localisation est stable, une disparition progressive ou une constante migration des autres. Les extrema les plus accentués correspondent naturellement aux extrema significatifs d'une sous-population.

## 5 – SIMULATION ET APPLICATIONS

### a) Décomposition de mélanges obtenus par simulation

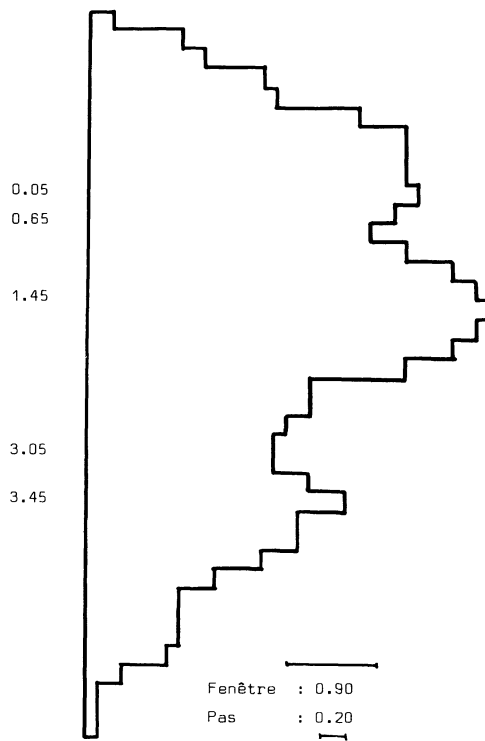
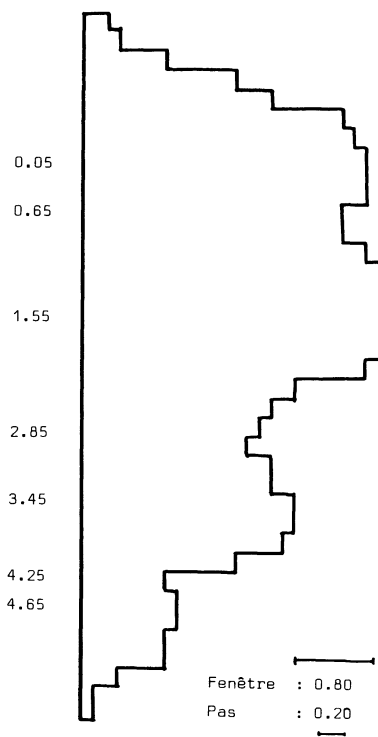
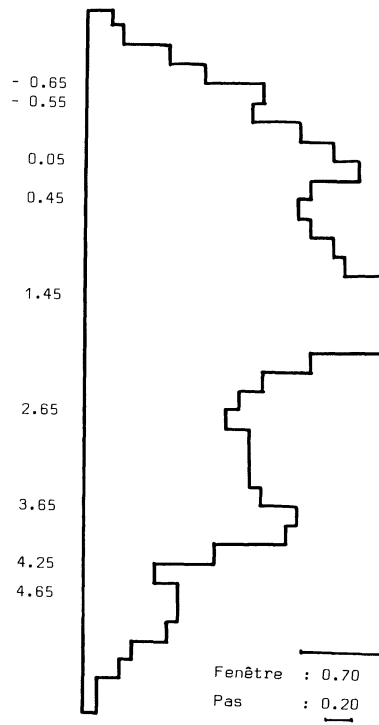
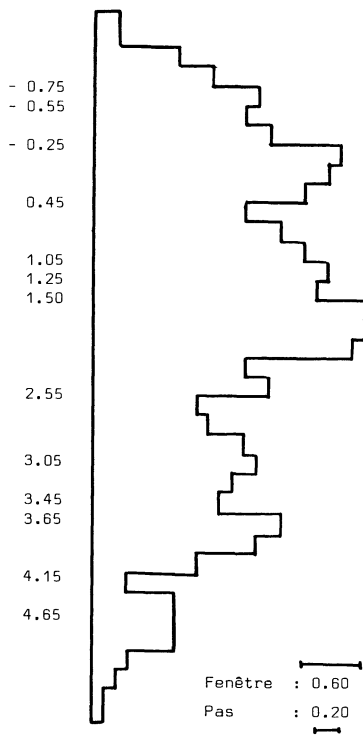
Dans le but de tester la validité de la méthode proposée et d'étudier sa sensibilité dans différents contextes, nous avons procédé à la séparation de nombreux mélanges obtenus par simulation.

Différentes conclusions ressortent de cette analyse qui sont détaillées dans (15). Nous nous limiterons à présenter ici un exemple montrant l'allure générale des courbes  $F_{\ell_i}$ .

L'exemple choisi est un mélange équipondéré de trois constituants gaussiens fortement imbriqués, avec  $n = 150$ .

$$\varphi_1 \sim \mathcal{N}(0 ; 2) \quad \varphi_2 \sim \mathcal{N}(1.5 ; 2) \quad \varphi_3 \sim \mathcal{N}(3.5 ; 3)$$

Si l'on retient comme limites théoriques entre deux sous intervalles consécutifs les points  $\theta_i$  vérifiant  $\varphi_i(\theta_i) = \varphi_{i+1}(\theta_i)$   $i = 1, 2$  on obtient  $\theta_1 = 0.75$  et  $\theta_2 = 3.20$ .



Nous choisissons  $\alpha = 0.2$  et  $t_0 = 1.55$  et pour  $\{\ell_i\}$ , la suite  $\{0.6, 0.7, 0.8, 0.9\}$ .

Les résultats sont les suivants :

$\ell$	minima de $F_\ell$						maxima de $F_\ell$						
0,6	-0.55	0.45	1.25	2.55	3.45	4.15	-0.75	-0.25	1.05	1.50	3.05	3.65	4.65
0.7	-0.55	0.45		2.65		4.25	-0.65	0.05		1.45		3.65	4.65
0.8		0.65		2.85		4.25		0,05		1.55		3.45	4.65
0.9		0.65		3.05				0.05		1.45		3.45	

L'étude du tableau permet de confirmer le résultat qui apparaît clairement d'ailleurs sur les graphiques suivants et en particulier pour  $\ell = 0.9$ . Les valeurs estimées sont  $\hat{\theta}_1 = 0.65$  et  $\hat{\theta}_2 = 3.05$  tandis que les estimations des abscisses des modes sont :  $\hat{M}_1 = 0.05$   $\hat{M}_2 = 1.45$   $\hat{M}_3 = 3.45$ .

### b) Applications à des mélanges réels

Nous avons été amenés à analyser deux fichiers de données médicales

#### • Premier fichier

Dans le but de déterminer des valeurs critiques pour différents paramètres médicaux et différents degrés pathologiques, il s'agit de discrétiser une centaine de variables numériques correspondant à des examens classiques en pathologie générale. Chaque variable est mesurée sur 400 malades décrivant au mieux l'ensemble de la pathologie classique. Une discrétisation en classe de type "très faible", "faible", "normal", "élevé", "très élevé", doit permettre la mise en œuvre ultérieure d'un système de diagnostic automatique "A.D.M." (14).

Les résultats obtenus s'avèrent satisfaisant pour le médecin et permettent de nuancer de façon intéressante le schéma général en 5 classes ci-dessus.

#### • Second fichier

Ce second fichier est limité à la pathologie hépatobiliaire et de taille beaucoup plus modeste puisque croisant 16 paramètres biologiques et 117 malades.

La méthode de la fenêtre mobile permet d'obtenir, à partir des variables numériques initiales des caractères ayant de deux à cinq modalités. Le nombre de modalités semblant d'ailleurs croître, dans une certaine mesure, avec la pertinence et le pouvoir discriminant du paramètre considéré.

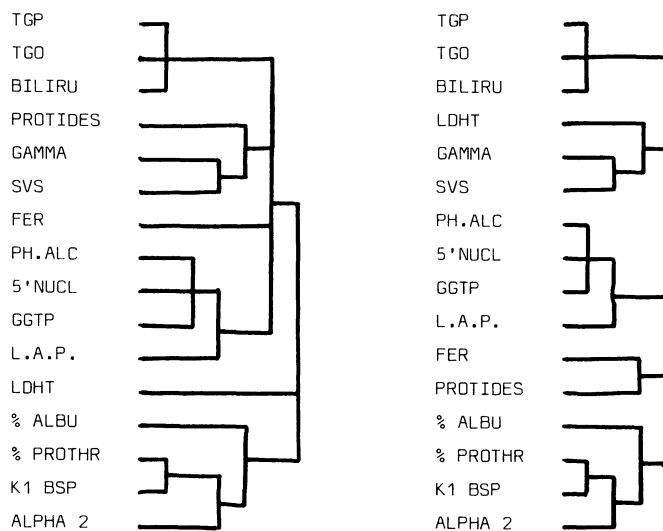
Le point le plus important, dans l'optique de la validation de la méthode de découpage est le suivant :

On observe la presque totale concordance entre les deux classifications obtenues sur l'ensemble des variables :

– à partir des données numériques initiales. L'indice de similarité étant le coefficient de corrélation (fig. 6).

– à partir des données discrétisées les variables étant alors des variables préordinales.

L'indice est celui de la chaîne de programmes de I.C. LERMAN (17) et fait appel à la notion de corrélation entre structures préordinales (fig. 7).



Dans les deux cas, l'algorithme utilisé est celui de la vraisemblance du lien.

Les seules différences observées concernent des variables très neutres. Par contre, les noyaux correspondants aux quatre principaux syndrômes sont conservés.

Cette stabilité doit être interprétée comme une preuve que le codage par une variable préordinaire réduit très faiblement l'information contenue dans les données brutes. Ceci montre que le découpage des variables a bien respecté la structure de la population traitée, et prouve la validité de l'attitude médicale qui retient en fait, pour son diagnostic, seulement un ordre de grandeur.

## 6. CONCLUSION

La méthode de la fenêtre mobile si elle n'est pas en soi une technique de décomposition des mélanges permet toutefois d'opérer de façon efficace la séparation des constituants.

Elle peut éventuellement servir de préalable à l'application d'une méthode d'estimation des paramètres d'un mélange nécessitant la connaissance à priori du nombre des composants.

En outre, elle possède les qualités recherchées d'économie, de souplesse et d'efficacité dans le cas de petits échantillons.

Elle peut être utilisée sur de très petits ordinateurs et prend tout son intérêt en mode conversationnel.

Une voie de poursuite de l'étude serait de tendre à une généralisation au cas multidimensionnel à une quantification de l'erreur entre la subdivision obtenue et une subdivision "optimale" théorique.

## 7. BIBLIOGRAPHIE

- [1] ABRAMSON N., BRAVERMAN D.J. (1962). – “Learning to recognize patterns in random environment”. *I.R.E. Trans. Information Theory – Vol. IT 8* p. 58-63.
- [2] AGRALAWA A.K. (1970). – “Learning with a probabilistic teacher”. *I.E.E.E. Trans. Information Theory – Vol. IT 16*, p. 373-379.
- [3] BHATTACHARYA C.G. (1967). – “A simple method of resolution of a distribution into gaussian components”. *Biometrics – Vol. 23 n° 1*, p. 115-135.
- [4] BLISCHKE W.R. (1962). – “Moment estimators for the parameters of two binomial distributions”. *Ann. Math. Stat. Vol. 33*, p. 444-454.
- [5] CASSIE R.M. (1954). – “The use of probability paper in the analysis of size frequency distributions”. *Austr. J. Marine Freshwater Res. Vol. 5*, p. 253-522.
- [6] CAZES P. (1976). – “Décomposition d’un histogramme en composantes gaussiennes”. *Rev. Stat. Appl. Vol. XXIV*, p. 63-82.
- [7] DAY N.E. (1969). – “Estimating the components of normal distributions”. *Biometrika Vol. 56*, p. 463-474.
- [8] DOROFYUK A.A. (1971). – “Automatic classification algorithms review”. *Automation and control Vol. 32, n° 12*, p. 1928-1958.
- [9] ENGLEMAN L., HARTIGAN (1969). – “Percentage points of a test for clusters”. *J. Amer. Stat. Ass. Vol. 64*, p. 1647-1648.
- [10] FISCHER W.D. (1958). – “On grouping with maximum homogeneity”. *J. Amer. Stat. Ass. Vol. 53*.
- [11] GIACOMELLI F. (1971). – “Subpopulation of blood lymphocytes demonstrated by quantitative cytochemistry”. *J. Histochemistry and cytochemistry, Vol. XIX n° 7*, p. 426-433.
- [12] HARTIGAN J.A. (1975). – “Clustering Algorithms”. WILEY N.Y.
- [13] HARTIGAN J.A. (1977). – “Distribution problems in clustering”. Classification and clustering – J. Van Rizin N.Y.
- [14] KERJEAN A.M. née SALMON (1978). – “Tentative d’établissement de 100 typologies d’examen biologiques. Contribution à l’établissement du système “A.D.M.”. Thèse de Doctorat d’Etat – Univ. de Rennes.
- [15] LAFAYE J.Y. (1978). – “Les différentes formes de l’appréhension des données dans l’exploration fonctionnelle hépatique. Discrétisation de variables numériques. Recherches de profils biologiques par une méthode de classification automatique”. Thèse de Doctorat de 3<sup>e</sup> cycle – Université de Rennes.
- [16] LECHEVALLIER Y. (1974). – “Optimisation de quelques critères en classification automatique et application à l’étude des modifications des protéines sériques en pathologie clinique”. Thèse de Doctorat de 3<sup>e</sup> cycle – Université de Paris VI.
- [17] LERMAN I.C. (1977). – “Reconnaissance et classification des structures finies en analyse des données”. Vol. 1 : Théories et méthodes. Rapport I.R.I.S.A. n° 70 – Rennes.
- [18] MEDGYESSY P. (1961). – “Décomposition of superpositions of distribution function”. Publishing house of the hungarian academy of sciences. Budapest.

- [19] PATRICK E.A., COSTELLO J.P. (1970). – “On unsupervised estimation algorithms”. *I.E.E.E. Trans Information theory* – Vol. IT 16 n° 5 – p. 556-569.
- [20] PEARSON K. (1894). – “Contribution to the mathematical theory of evolution”. *Philos. Trans. Soc. n° 185*.
- [21] RAO C.R. (1948). – “Utilisation of multiple measurement in problems of biological classification”. *J. Roy. Stat. Soc. Série B.* – Vol. 10, n° 2.
- [22] ROUX M., ROUX G. (1975). – “Programme STEAK, pour rendre homogène un tableau de données quelconques et en faire l’A.F.C.”. Laboratoire de statistique – Université de Paris VI.
- [23] SCHROEDER A. (1976). – “Analyse d’un mélange de distributions de même type”. *Rev. Stat. Appl.* – Vol. XXIV n° 1, p. 53-62.
- [24] TANAKA S. (1962). – “A method of analysing a polymodal frequency distribution and its application to the length distribution of the Porgy”. *J. Fish. Res. Bd Canada* – Vol. XIX n° 6, p. 1143-1159.
- [25] TEICHER H. (1963). – “Identifiability of finite mixtures”. *Ann. Math. Stat.* Vol. XXXIV, p. 1265-1269.
- [26] VINOD H.D. (1969). – “Integer programming and the theory of grouping”. *J. Amer. Stat. Ass.* p. 506-519.
- [27] YAKOWITZ S.S., SPRAGINS J.D. (1968). – “On the identifiability of mixtures”. *Ann. Math. Stat.* – Vol. XXXIX n° 1, p. 209-214.
- [28] YOUNG T.Y., CORALLUPPI G. (1970). – “Stochastic estimation of a mixture of normal density functions using an information criterion”. *I.E.E.E. Trans. Information Theory* – Vol. IT n° 3, p. 258-263.