REVUE DE STATISTIQUE APPLIQUÉE

ANNE SCHROEDER

Analyse d'un mélange de distributions de probabilité de même type

Revue de statistique appliquée, tome 24, n° 1 (1976), p. 39-62

<http://www.numdam.org/item?id=RSA_1976__24_1_39_0>

ANALYSE D'UN MÉLANGE DE DISTRIBUTIONS DE PROBABILITÉ DE MÊME TYPE[1]

Anne SCHROEDER
Iria–Laboria (Rocquencourt)

On a présenté dans des articles précédents (DIDAY, SCHROEDER – 1974) un algorithme itératif de reconnaissance des composants d'un mélange de densités de probabilité qui utilise des méthodes d'estimation classiques. Intervient en particulier celle du maximum de vraisemblance qui permet l'optimisation d'un critère de vraisemblance.

Dans ce papier, la méthode est généralisée de façon à pouvoir optimiser ce même critère dans des mélanges de distributions dont les paramètres inconnus ne peuvent être calculés par le maximum de vraisemblance, par exemple les mélanges de lois Gamma. Enfin une application en modélisation de systèmes informatiques est brièvement présentée.

(1) Article remis en Janvier 1975, révisé en Décembre 1975

Mots-clés : **Mélange de densités. Estimation. Classification automatique**

* *
*

1 – INTRODUCTION

1.1. – Le problème

Une des tâches courantes de la statistique consiste à déduire d'un échantillon observé une distribution de probabilité sur la population-mère dont cet échantillon est tiré. Cette estimation peut être un but en soi, mais n'est, la plupart du temps, qu'un intermédiaire indispensable pour pousser plus avant l'analyse statistique par des tests, l'établissement d'un modèle, etc.

Compte tenu, soit de la nature physique du phénomène observé, soit de l'échantillon lui-même, on peut être conduit au choix d'une loi de distribution de type connu (binomiale, Γ, gaussienne à une ou plusieurs dimensions, Wishart, etc.).

Mais bien souvent, l'existence de plusieurs sources aléatoires à l'origine du phénomène, ou bien l'observation d'un histogramme manifestement multimodal indiquent que la distribution de la population-mère peut être un mélange de lois de probabilité simples et régulières. Se pose alors le problème d'estimer à partir d'un échantillon fini, les paramètres des différentes distributions intervenant dans le mélange.

Si on suppose connus le nombre k de composants du mélange et la famille de lois de probabilité $(f_\lambda/\lambda \in \mathbf{L})$ à laquelle appartiennent les distribution des différents composants, la densité du mélange peut s'écrire :

$$\forall\, x \in \mathbf{R}^q \qquad \varphi(x) = \sum_{1 \leqslant j \leqslant k} P_j f_{\lambda_j}(x) \tag{1}$$

où
- λ_j = valeur du paramètre λ dont dépend la famille (f_λ), pour le j^e composant.
- P_j = probabilité a priori d'apparition du j^e composant.

et le paramètres inconnus à estimer sont les $(P_j/1 \leqslant j \leqslant k)$ et les $(\lambda_j/1 \leqslant j \leqslant k)$.

1.2. – Différentes approches

Il existe un grand nombre de techniques pour estimer les paramètres d'un mélange.

Les plus classiques sont des techniques d'estimation, qui, posant a priori le modèle (1), estiment directement les paramètres à l'aide d'estimateurs

calculés sur les observations ; plus récemment, ont été mises au point des techniques de type bayesien, d'apprentissage, etc. qui procèdent par approximations successives, liées à l'introduction des observations pour estimer les paramètres.

Sauf mention expresse, ces méthodes requièrent l'entrée du nombre de composants.

Les techniques du premier type ne diffèrent que par la méthode d'estimation qu'elles utilisent : méthode des moments (PEARSON (1894)) avec estimateurs du maximum de vraisemblance (RAO (1948) DAY (1969)) du χ^2 minimum, etc. La plupart d'entre elles s'appliquent aux mélanges gaussiens et sont souvent restreintes aux distributions unidimensionnelles. Citons par exemple : RAO (1948), qui résoud le cas de mélanges à 2 composants, BATTACHARYA (1967) qui donne une méthode graphique d'évaluation du nombre de composants mais, d'une part, nécessite un très grand nombre d'observations (# 10 000) et d'autre part, requiert que les distributions-mères soient relativement séparées. Une présentation générale de ce genre de méthode est donnée dans DOROFEYUK (1971).

Une approche sensiblement différente du problème de l'estimation du modèle (1) est celle de COOPER et COOPER (1964) : les paramètres inconnus du modèle sont calculés à partir des moments de la distribution globale observée.

Chez tous les auteurs, l'extension des méthodes au cas multidimensionnel nécessite de nombreuses hypothèses supplémentaires : par exemple, DAY (1969) et COOPER et COOPER (1964) n'abordent sur le plan pratique que le cas de deux distributions ne différant que par leurs moyennes.

Les approches du second type sont diverses : la plupart d'entre elles se rattachent aux techniques d'estimation bayesienne : PATRICK et HANCOCK (1966), PATRICK et COSTELLO (1970), AGRAWALA (1970). Les hypothèses nécessaires diffèrent de l'une à l'autre mais sont en général très contraignantes. YOUNG et CORRALUPI (1970) présentent un algorithme d'approximation stochastique basé sur un critère d'information ; bien qu'exclusivement applicable aux mélanges gaussiens unidimensionnels, cet algorithme semble très performant et a l'avantage de ne pas demander que soit fixé le nombre de composants du mélange contrairement à l'ensemble des autres méthodes.

L'ensemble de ces méthodes d'approximation permet de formaliser le problème de la résolution des mélanges en termes d'apprentissage avec ou sans maître (AGRAWALA (1970), PATRICK (1972), DUDA et HART (1973).

Pour le cas particulier des mélanges gaussiens unidimensionnels J.P. BENZECRI propose une méthode basée sur une série de déconvolutions sucessives (cf. (2)) realisée et testée par P. CAZES.

1.3. Approche proposée

Dans la pratique, choisir la famille $(f_\lambda / \lambda \in \mathbf{L})$ est assez simple : lois gaussiennes lorsqu'on a peu d'information a priori, mais beaucoup d'observations ; lois Γ, χ^2 ou log-normale lorsque le phénomène est essentiellement positif ou l'échantillon nettement asymétrique, etc.

Au contraire, choisir le nombre de composants peut être beaucoup plus délicat. La connaissance que l'on a de la nature du phénomène observé indique parfois ce nombre, mais la plupart du temps, il n'y a pas de moyen de la connaître a priori. Même dans le cas unidimensionnel, l'observation d'un histogramme est évidemment insuffisante.

Par rapport à ces deux choix, l'algorithme de reconnaissance des composants d'un mélange que nous proposons présente une certaine généralité car, d'une part la famille de densités cherchée est laissée au choix de l'utilisateur et d'autre part la rapidité des calculs permet de l'appliquer plusieurs fois en se fixant des nombres de composants différents et en comparant ensuite les estimations obtenues pour chaque essai.

Il s'agit d'un algorithme itératif détectant parallèlement une partition en classes de l'échantillon observé et des distributions associées à ces classes.

Cette idée de la recherche simultanée d'une partition et de "noyaux" caractéristiques des classes de cette partition a été initialement utilisée en classification automatique non hiérarchique : il s'agit de la méthode des Nuées Dynamiques dûe à E. DIDAY ; les noyaux sont alors des éléments d'un échantillon à classer. DIDAY expose la méthode en (8) et propose l'utilisation du même schéma avec des noyaux de divers types en vue de résoudre des problèmes spécifiques : par exemple, en prenant comme noyaux les éléments principaux d'inertie des classes, la méthode fournira des analyses factorielles locales à forte inertie (Analyse Factorielle Typologique (11)) ; si les noyaux sont des polynômes d'interpolation d'un point moyen des classes, l'algorithme permet de reconstituer les données manquantes d'un tableau en tenant compte des données présentes pour regrouper les observations en classes et réduire ainsi le nombre d'interpolations à effectuer, (lissage typologique, développé par Y. OK).

Pour aborder le problème des mélanges de distributions de probabilité, le même schéma sera à nouveau utilisé en prenant cette fois comme noyaux des distributions de probabilité.

Dans un premier temps nous avions étudié une forme particulière de l'algorithme utilisant la méthode d'estimation du maximum de vraisemblance et optimisant un critère de vraisemblance ((10), (11)). Dans (22), cette variante est replacée dans un contexte plus général qui permet d'étendre la méthode à d'autres techniques d'estimation liées à l'optimisation d'autres critères ; l'aspect pratique y est également plus largement développé.

Dans ce papier, nous nous placerons à un nouveau niveau de généralisation : une étape du schéma introduit par E. DIDAY et repris en (22) consiste à chaque itération à passer d'une partition à un ensemble de noyaux en optimisant la fonction critère choisie qui est fonction du couple (partition, ensemble de noyaux) ; ici, cette étape n'est plus une optimisation mais seulement une amélioration du critère et les propriétés de convergence sont démontrées sous cette nouvelle hypothèse. Comme on le verra dans la suite, cette modification élargit considérablement les possibilités d'application de l'algorithme ; une transformation "améliorante" est beaucoup plus facile à définir qu'une transformation optimale. (Cette extension à également été utilisée en classification avec distances adaptatives par GOVAERT (9)).

L'application immédiate qui en sera faite sera celle de l'analyse des mélanges de distributions Gamma avec optimisation d'un critère de vraisemblance. Trouver, pour un échantillon donné, les paramètres d'une distribution Gamma qui améliorent sa vraisemblance par rapport à des paramètres antérieurs est en effet facile alors que ne l'est pas l'estimation de ces paramètres par le maximum de vraisemblance.

En entrée, on demande :

- l'échantillon à analyser (taille et dimension quelconques)
- la forme de la famille des lois de probabilité cherchées (lois gaussiennes, lois multinomiales, lois gamma, éventuellement famille exponentielle, etc.).
- le nombre de composants recherchés dans le mélange.

En sortie, on obtient une partition de l'échantillon en k classes et k lois de probabilité qui leur sont attribuées.

Dans un premier paragraphe, on indiquera le schéma général de l'algorithme de reconnaissance des mélanges proposé. Ses propriétés seront ensuite données précisément. Puis deux cas particuliers d'estimation seront étudiés. Enfin, on examinera le comportement de l'algorithme pour deux familles de distributions usuelles. et on présentera une application en informatique.

2 – L'ALGORITHME PROPOSE

2.1. – Schéma général

Soient E, sous ensemble fini de $\mathbf{R}^q$, l'échantillon dont on dispose, et $(f_\lambda/\lambda \in \mathbf{L})$ la famille des densités de probabilité à laquelle on suppose que les distributions des différents composants appartiennent ; λ est un paramètre réel ou vectoriel et $\mathbf{L}$ son espace de définition – $\mathbf{L} \subset \mathbf{R}^s$ – (par exemple si $(f_\lambda/\lambda \in \mathbf{L})$ est la famille des distributions gaussiennes sur $\mathbf{R}^q$, $\lambda = (\mu, \Sigma)$ où $\mu \in \mathbf{R}^q$ est le vecteur moyen de la distribution et Σ est sa matrice de covariance. $\Sigma \in \mathcal{S}$ = espace des matrices réelles symétriques définies positives de rang q – alors $\mathbf{L} = \mathbf{R}^q \times \mathcal{S} \subset \mathbf{R}^{q(q+3)/2}$). On note $\mathbf{P}_k$ l'ensemble des partitions de E en k classes, et $\mathbf{L}_k$ l'ensemble des k-uples d'éléments de $\mathbf{L}$.

L'algorithme se déroule alors de la façon suivante : Partant d'une partition initiale quelconque $(P_1^\circ, P_2^\circ, \ldots, P_k^\circ) \in \mathbf{P}_k$ de E, ou bien d'un quelconque k-uple de valeurs du paramètre inconnu $(\lambda_1^\circ, \lambda_2^\circ, \ldots, \lambda_k^\circ)$, on applique successivement les deux fonctions suivantes jusqu'à l'obtention d'éléments stables de $\mathbf{L}_k$ et $\mathbf{P}_k$:

$$\begin{array}{ccc} \mathbf{L}_k & \xrightarrow{f} & \mathbf{P}_k \\ L^\circ & \longmapsto & P^\circ \end{array}$$

puis

$$\begin{array}{ccccc} \mathbf{L}_k \times \mathbf{P}_k & \xrightarrow{g} & \mathbf{L}_k & \xrightarrow{f} & \mathbf{P}_k \\ (L^\circ, P^\circ) & \longmapsto & L^1 & \longmapsto & P^2 \ldots \end{array} \qquad (2)$$

$$L_k \times P_k \xrightarrow{g} L_k \xrightarrow{f} P_k$$

$$(L^{n-1}, P^{n-1}) \longmapsto L^n \longmapsto P^n \ldots$$

Plus précisément f et g sont définies comme suit :

$$- \; f : L_k \longrightarrow P_k$$

$$L \longmapsto P ,$$

avec $L = (\lambda_1, \ldots, \lambda_k)$ et $P = (P_1, \ldots, P_k)$ tels que

$\forall$ i, $1 \leqslant i \leqslant k$: P_i est l'ensemble des observations qui sont plus "proches" de la loi f_{λ_i} que de toutes les autres (avec choix du plus petit indice en cas d'égalité éventuelle).

Il faut pour cela, se donner une fonction D mesurant la "distance" d'une observation $x \in E$ à une distribution f_λ :

$$D : E \times L \longrightarrow R^+$$

$$(x, \lambda) \longmapsto D(x, \lambda)$$

Le choix de cette fonction peut se faire de diverses façons selon l'optique statistique dans laquelle on se place :

Une observation x peut être considérée comme "proche" de la valeur du paramètre inconnu, si la valeur $f_\lambda(x)$ est grande, ou bien si x est proche de l'espérance mathématique de f_λ, etc.

(Différents choix de D seront vus en détail au paragraphe 2.3).

Alors :

$$P_i = \{ x \in E / D(x, \lambda_i) \leqslant D(x, \lambda_j), \forall_j \neq i,$$

avec $$i < j \quad \text{si} \quad D(x, \lambda_i) = D(x, \lambda_j) \}$$

Notons R la fonction suivante :

$$R : L \times \mathcal{P}(E) \longrightarrow R^+$$

($\mathcal{P}(E)$ est l'ensemble des parties de E)

$$(\lambda, A) \longmapsto R(\lambda, A) = \sum_{x \in A} D(x, \lambda)$$

alors,

$$g : L_k \times P_k \longrightarrow L_k$$

$$((\lambda_1 \ldots \lambda_k), (P_1 \ldots P_k)) \longmapsto (\mu_1 \ldots \mu_k)$$

tel que, pour tout i ($1 \leqslant i \leqslant k$), μ_i "améliore" la "ressemblance" de P_i à sa distribution f_{μ_i}, au sens suivant :

$$R(\mu_i, P_i) < R(\lambda_i, P_i)$$

$$\sum_{x \in P_i} D(x, \mu_i) < \sum_{x \in P_i} D(x, \lambda_i)$$

(s'il n'existe pas de $\mu_i \neq \lambda_i$ vérifiant cette inégalité stricte, on gardera $\mu_i = \lambda_i$)

Le critère alors optimisé sera fonction de la partition P^* et du k-uple L^* de $\mathbf{L}$, obtenus à la convergence, et aura la forme :

$$W(L^*, P^*) = \sum_{1 \leqslant i \leqslant k} R(\lambda_i^*, P_i^*) = \sum_{1 \leqslant i \leqslant k} \sum_{x \in P_i^*} D(x, \lambda_i^*)$$

on appellera $(v_n) = (L^n, P^n)$ la suite des couples de $\mathbf{L}_k \times \mathbf{P}_k$ fournis par l'algorithme et (u_n) celle des valeurs prises par le critère sur les (v_n) :

$$\forall\, n \in N \qquad v_{n+1} = (L^{n+1}, P^{n+1})$$

tels que :

$$L^{n+1} = g(L^n, P^n) \text{ et } P^{n+1} = f(L^{n+1})$$

$$\forall\, n \in N \qquad u_n = W(v_n) = W(L^n, P^n) = \sum_{1 \leqslant i \leqslant k} \sum_{x \in P_i^n} D(x, \lambda_i^n)$$

2.2. – Schéma simplifié

Il arrivera que l'on soit placé dans le cas particulier où D est telle que :

pour tout $A \subset E$, il existe un minimum unique en λ à la fonction

$$R(\lambda, A) = \sum_{x \in A} D(x, \lambda),$$

alors la fonction g peut faire mieux qu'ameliorer, elle minimise, et ne depend alors que de la seule partition, c'est-à-dire :

$$g : \mathbf{P}_k \rightarrow \mathbf{L}_k$$

$$P \mapsto L, \qquad \text{tel que :}$$

$$\forall\, i\ (1 \leqslant i \leqslant k)\ \lambda_i \qquad \text{est tel que :}$$

$$R(\lambda_i, P_i) = \sum_{x \in P_i} D(x, \lambda_i) = \inf_{\lambda \in L} \sum_{x \in P_i} D(x, \lambda)$$

Le schéma (2) de l'algorithme prend alors la forme simplifiée suivante :

$$\mathbf{L}_k \underset{g}{\overset{f}{\rightleftarrows}} \mathbf{P}_k \tag{3}$$

Dans ce cas, les paramètres λ_i sont donc estimés à partir des k classes P_i considérées comme échantillons ; selon le choix qui aura été fait pour la fonction D, la fonction g définira une certaine procédure d'estimation.

On verra dans le paragraphe 2.4. quelques techniques usuelles d'estimation pouvant ainsi être utilisées, chacune de ces méthodes étant associée à l'optimisation d'un critère statistique différent.

2.3. – Propriétés de l'algorithme

2.3.1. – Convergence

Définition 1 :

Une suite (L^n, P^n) de $L_k \times P_k$ sera dite *convergente* si elle est stationnaire, c'est-à-dire si il existe un entier M, tel que :

$$\forall\ n \geqslant M \qquad (L^n, P^n) = (L^{n+1}, P^{n+1})$$

Définition 2 :

Dans le cas particulier où $R(\lambda, A) = \sum_{x \in A} D(x, \lambda)$ admet pour tout

$A \subset E$, un minimum unique en λ, (cf. schéma simplifié 2.2) on dira qu'un (L, P) est *non-biaisé* pour les fonctions f et g, s'il est tel que :

$$P = f \circ g\,(P) \qquad \text{et} \qquad L = g \circ f\,(L)$$

C'est-à-dire que chaque P_i est constituée des éléments de E les plus proches de λ_i au sens de D, tandis que les λ_i sont tel que :

$$R(\lambda_i, P_i) = \inf_{\lambda \in L} R(\lambda, P_i)$$

Les deux théorèmes suivants assurent la convergence de l'algorithme général :

Théorème 1 :

La suite réelle $(u_n) = (W(v_n))$ décroît et converge vers un minimum local u*.

Démonstration :

Pour démontrer la décroissance de (u_n), $W(L^n, P^n) \leqslant W(L^{n-1}, P^{n-1})$, on vérifiera successivement les deux inégalités :

a) $$W(L^n, P^n) \leqslant W(L^n, P^{n-1})$$

et

b) $$W(L^n, P^{n-1}) \leqslant W(L^{n-1}, P^{n-1})$$

a) $$W(L^n, P^n) = \sum_{1 \leqslant i \leqslant k} \sum_{x \in P_i^n} D(x, \lambda_i^n)$$

et

$$W(L^n, P^{n-1}) = \sum_{1 \leqslant i \leqslant k} \sum_{x \in P_i^{n-1}} D(x, \lambda_i^n)$$

Comme $P^n = f(L^n)$, chaque partie P_i^n est précisément composée des éléments x de E tels que :

$$D(x, \lambda_i^n) \leqslant (x, \lambda_j^n) \qquad \text{pour tout } j \neq i,$$

avec $j < i$ en cas d'égalité, d'où immédiatement l'inégalité a).

b) $$W(L^n, P^{n-1}) = \sum_{1 \leqslant i \leqslant k} R(\lambda_i^n, P_i^{n-1})$$

et

$$W(L^{n-1}, P^{n-1}) = \sum_{1 \leqslant i \leqslant k} R(\lambda_i^{n-1}, P_i^{n-1})$$

Comme $L^n = g(L^{n-1}, P^{n-1})$, l'inégalité b) résulte directement de la définition de g.

(u_n) est donc décroissante, comme elle est de plus réelle et positive, donc bornée inférieurement, sa convergence est assurée.

De plus, E étant fini, les espaces $\mathbf{P}_k$ et $g(\mathbf{L}_k \times \mathbf{P}_k)$ le sont aussi (pour un L° fixé) et donc les suites (u_n) et (v_n) ne peuvent prendre qu'un nombre fini de valeurs.

Comme (u_n) converge, elle atteint donc sa limite :

$$\exists\, u^* \quad \exists M \quad \forall\, n \geqslant M \quad u_n = u^* \qquad \text{(CQFD)}$$

Théorème 2 :

La suite (v_n) converge dans $\mathbf{L}_k \times \mathbf{P}_k$ et atteint sa limite (L^*, P^*).

Démonstration :

Il a été vu au théorème 1 que la suite $(u_n) = (W(L^n, P^n))$ est stationnaire, on a donc :

$$\forall\, n \geqslant M \quad W(L^{n-1}, P^{n-1}) = W(L^n, P^n)$$

et encore, d'après la double inégalité a) et b) du théorème 1 :

$$W(L^{n-1}, P^{n-1}) = W(L^n, P^{n-1}) = W(L^n, P^n)$$

la première égalité s'écrit aussi :

$$\sum_{1 \leqslant i \leqslant k} R(\lambda_i^{n-1}, P^{n-1}) = \sum_{1 \leqslant i \leqslant k} R(\lambda_i^n, P_i^{n-1})$$

Or, comme $L^n = g(L^{n-1}, P^{n-1})$ on sait que les $(\lambda_i^n / 1 \leqslant i \leqslant k)$ sont tels que :

$$R(\lambda_i^n, P_i^n) \leqslant R(\lambda_i^{n-1}, P_i^{n-1})$$

avec égalité si et seulement si $\lambda_i^n = \lambda_i^{n-1}$ ce qui est donc ici précisément le cas puisque les sommes de ces quantités sont égales. En conséquence :

$$\forall\, n \geqslant M \quad \forall\, i\ (1 \leqslant i \leqslant k) \quad \lambda_i^n = \lambda_i^{n-1} \Leftrightarrow \forall\, n \geqslant M \quad L^n = L^{n-1}$$

$$\Rightarrow P^n = f(L^n) = P^{n-1}$$

La suite (v_n) est donc bien stationnaire : elle converge en atteignant sa limite (L^*, P^*).

(CQFD)

De la stationarité de (v_n), on déduit immédiatement le résultat suivant :

Corollaire :

Dans le cas particulier du schéma simplifié 2.2, la limite (L^*, P^*) de (v_n) est un élément non-braisé.

2.3.2. – Estimation de la distribution globale

Notons d'abord que le critère W (L, P) que nous cherchons à minimiser passe par un minimum local au couple (L^*, P^*) non-biaisé obtenu à la convergence de l'algorithme. Cet optimum local dépend évidemment de la partition initiale quelconque P°.

En pratique, il sera utile de procéder à plusieurs tirages aléatoires pour P° et d'interpréter les couples (L^*, P^*) obtenus en fonction des valeurs corres- W (L^*, P^*) du critère. (On trouvera dans (22) des précisions sur de tels tirages et sur leurs fréquences d'apparition)

Supposons que E soit un échantillon représentatif d'une distribution de mélange sur $\mathbf{R}^q$, de densité :

$$x \in \mathbf{R}^q \qquad \varphi(x) = \sum_{1 \leqslant i \leqslant k} p_i f_{\lambda_i}(x) - \text{où } p_i = \text{Prob}\,(x \in P_i) -$$

où les P_i sont des parties disjointes de $\mathbf{R}^q$ et les f_{λ_i} des densités de probabilité dépendant de paramètres $\lambda_i \in \mathbf{R}^s$. Ceci revient à supposer que les observations sont issues de k phénomènes aléatoires de probabilités a priori P_i ayant chacun une loi de probabilité différente f_{λ_i}.

Sous ces hypothèses, l'algorithme proposé donne la solution suivante : les f_{λ_i} sont données par L^* et les Prob $(x \in P_i)$ estimées par les fréquences $\dfrac{|P_i^*|}{N}$ où $|P_i^*| = \text{card}\,(P_i^*)$.

2.4. – Choix particuliers de méthodes d'estimation et critères associés

Finalement, toute forme de l'algorithme est entièrement définie par la donnée d'une application D – on remarque d'ailleurs facilement que le fait de prendre D à valeurs dans R^+ n'est pas strictement nécessaire ; il suffit en réalité pour démontrer les propriétés de convergence, de s'assurer que l'espace des valeurs est un sous-ensemble borné inférieurement de **R**.

2.4.1. – Critère de vraisemblance

Soit $D : E \times L \to \mathbf{R}$

$$(x, \lambda) \to D(x, \lambda) = \text{Log}\,[f^* / f_\lambda(x)]$$

Cette définition exprime qu'une observation x sera d'autant plus proche du noyau λ que la densité f_λ sera grande en x.

Pour que cette définition conduise à un ensemble de valeurs pour D qui soit borné inférieurement, il faut choisir la constante f^* de façon à ce que :

$$f^* \geqslant \max\{f_\lambda(x) / \lambda \in L' \subset L, x \in E\}$$

où $\mathbf{L}'$ est tel que : $g(\mathbf{P}_k) = (\mathbf{L}')^k$, c'est-à-dire où $\mathbf{L}'$ est la partie de $\mathbf{L}$ composée des valeurs de λ qui peuvent être obtenues à partir de g comme élément d'un k-uple de $g(\mathbf{P}_k)$.

Par conséquent, l'existence de f* est équivalente au fait que la fonction g qui sera déduite de D, est définie partout sur $\mathbf{P}_k$ et ne peut conduire qu'à des valeurs bornées pour les densités f_λ.

Ce choix pour la fonction D entraîne pour f, la définition suivante :

$$f : \mathbf{L}_k \rightarrow \mathbf{P}_k$$
$$L \mapsto f(L) = (P_1, \ldots, P_k)$$

où

$$P_i = \{ x \in E / f_{\lambda_i}(x) \geqslant f_{\lambda_j}(x), \forall\, j \neq i,$$

x étant affecté à la classe de plus petit indice en cas d'égalité}

D'autre part :

$$\forall \lambda \in L \quad \forall A \subset E \quad R(\lambda, A) = \sum_{x \in A} D(x, \lambda)$$

prend la forme : $R(\lambda, A) = \text{cste} - \text{Log}\, V_\lambda(A)$ si $V_\lambda(A)$ note la vraisemblance de l'échantillon, A pour la densité f_λ.

Pour les distributions f_λ qui admettent pour λ un estimateur du maximum de vraisemblance au vu de l'échantillon A, pour tout $A \subset E$, on se trouve donc dans le cas d'application du schéma simplifié (3) et la fonction g prend la forme :

$$g : \mathbf{P}_k \rightarrow \mathbf{L}_k$$

$P \rightarrow g(P) = (\lambda_1, \ldots, \lambda_k)$ où $\forall\, i / 1 \leqslant i \leqslant k, \lambda_i$ est l'estimé de λ par le maximum de vraisemblance au vu de l'échantillon P_i.

Le critère à minimiser est alors :

$$W(L, P) = \text{Cste} - \text{Log} \prod_{1 \leqslant i \leqslant k} V_{\lambda_i}(P_i)$$

Se trouvant dans les conditions d'application des théorèmes 1 et 2 et du corollaire de 2.3.1., on peut énoncer :

Théorème 3

Pour les distributions admettant des estimateurs du maximum de vraisemblance pour leurs paramètres inconnus, et pour tout sous-échantillon de E, l'algorithme défini ci-dessus par les fonctions f et g converge en faisant croître le produit des vraisemblances des classes P_i pour les distributions f_{λ_i} qui leur sont associées.

Le couple (L^*, P^*) obtenu à la convergence est un élément non-biaisé ; c'est-à-dire que les λ_i^* sont estimés par le maximum de vraisemblance à partir des échantillons P_i^* et que ceux-ci sont eux-mêmes constitués des observations tirées plus vraisemblablement de la distribution $f_{\lambda_i}^*$ que des (k-1) autres.

2.4.2. – Une variante liée à la méthode des moindres carrés

Soit maintenant

$$D : (x, \lambda) \mapsto D(x, \lambda)$$

avec

$$D(x, \lambda) = d_M^2 (x, m_\lambda)$$

où :

M est une métrique sur $\mathbf{R}^q$ et m_λ l'espérance mathématique de la loi de probabilité f_λ ; m_λ est supposée elle-même fonction du paramètre inconnu λ.

Avec ce choix pour D, on considère que la distance d'une observation x à un noyau λ ne dépend que de la distance de x à la moyenne théorique de f_λ, au sens d'une certaine métrique M à déterminer. Il s'agit d'une hypothèse statistique classique tres restrictive, mais qui peut être justifiée dans certains cas.

Alors, la fonction f prend la forme suivante :

$$f : L = (\lambda_1 \dots \lambda_k) \rightarrow f(L) = (P_1, \dots, P_k)$$

où

$$P_i = \{ x \in E / d_M^2 (x, m_{\lambda_i}) \leqslant d_M^2 (x, m_{\lambda_j}), \forall j \neq i \quad \text{avec} \quad i < j$$

en cas d'égalité $\}$

ce qui signifie que chaque x est affecté à la classe dont il approche le plus la moyenne théorique au sens de M.

$$\forall \lambda \in L \quad \text{et} \quad \forall A \subset E \qquad R(\lambda, A) = \sum_{x \in A} d_M^2 (x, m_\lambda)$$

n'est autre que la dispersion quadratique de A autour de m_λ au sens de M. Pour A fixé, R (λ, A) passe par un minimum lorsque m_λ est le centre de gravité de A et λ est défini ainsi sans équivoque dès que m_λ est fonction injective de λ. On est alors placé dans le cadre du schéma 2.2 et g est définie par : $g : P \rightarrow g(P) = (\lambda_1 \dots \lambda_k)$ tel que, pour tout i, $1 \leqslant i \leqslant k$, λ_i vérifie :

$$\sum_{x \in P_i} d_M^2 (x, m_{\lambda_i}) = \operatorname*{Inf}_{\lambda \in L} \sum_{x \in P_i} d_M^2 (x, m_\lambda)$$

$\Leftrightarrow$ λ_i est l'estimateur de λ au sens des moindres carrés pour l'échantillon P_i et la métrique M.

Et le critère à minimiser n'est autre que :

$$W(L, P) = \sum_{1 \leqslant i \leqslant k} \sum_{x \in P_i} d_M^2 (x, m_{\lambda_i}) =$$

somme des M-dispersions de chaque classe P_i autour de la moyenne théorique de la distribution f_{λ_i} correspondante.

Bien que placée dans une optique très différente, cette variante est équivalente – d'un point de vue technique – aux méthodes de classification dite "du centre de gravité" (cf. (8))

On pourrait également prendre :

$$D(x, \lambda) = d^2_{M_\lambda}(x, m_\lambda)$$

où la métrique M_λ serait elle-même fonction du paramètre λ. L'algorithme, dans ce cas, regrouperait les éléments de E au vu de leur distance aux moyennes m_λ pour les métriques M_λ, différentes selon les classes et selon le n° de l'itération du processus. Un tel procédé de classification avec distance adaptive est présenté dans (9).

3 – CAS PARTICULIERS DE DEUX FAMILLES DE DISTRIBUTIONS USUELLES.

3.1. – Lois gaussiennes multidimensionnelles.

La famille analysée est alors $(f_\lambda / \lambda \in L)$ avec :

$$\forall x \in R^q \qquad f_\lambda(x) = (2\pi)^{-q/2} (\det V)^{-1/2} \exp\left[-(x-\mu)' V^{-1} (x-\mu)/2\right]$$

avec

$$\lambda = (\mu, V)$$

où

$\mu \in R^q$ est le vecteur moyenne de la distribution,
V, sa matrice de variance-covariance $(q \times q)$

Et les fonctions D, f deviennent :

$$D(x, \lambda) = \text{Cste} + \frac{1}{2}\left[\text{Log det } V + d^2_{V^{-1}}(x, \mu)\right]$$

– $f : L \mapsto P$ est telle que, pour tout $i \leqslant k$, on ait :

$$P_i = \{x \in E / \text{Log det} V_i + d^2_{V_i^{-1}}(x, \mu_i) \leqslant \text{Log det} V_j + d^2_{V_j^{-1}}(x, \mu_j), \forall j \neq i\}$$

La famille des distributions gaussiennes admet pour ses paramètres inconnus des estimateurs du maximum de vraisemblance conduisant à des densités bornées sauf sur les échantillons engendrant dans R^q un sous-espace de dimension (dim (P_i)) inférieure à la dimension de leurs matrices de covariance plus 1 (en particulier si $|P_i| < q + 1$), car alors, $\det V = 0$.

a) Si on fait l'hypothèse que l'algorithme ne conduira pas à de telles parties, on est placé dans le cas du schéma simplifié décrit en 2.4.1., et la fonction g est donnée par :

– $g : P \mapsto L$ est telle que, pour tout $i \leqslant k$, μ_i et V_i sont les estimations par le maximum de vraisemblance de la moyenne et de la matrice de variance-covariance de l'échantillon P_i.

On sait que ces estimations sont uniques et valent :

$$\mu_i = \frac{1}{|P_i|} \sum_{x \in P_i} x \quad \text{et} \quad V_i = \frac{1}{|P_i|} \sum_{x \in P_i} (x - \mu_i)(x - \mu_i)' \quad (|P_i| = \text{card}(P_i))$$

Le théorème 3 assure la convergence de l'algorithme vers un minimum local de la fonction critère :

$$W(L,P) = \text{Cste} + \frac{1}{2} \sum_{1 \leqslant i \leqslant k} \left(|P_i| \operatorname{Log} \det V_i + \sum_{x \in P_i} d^2_{v_i^{-1}}(x, \mu_i) \right)$$

Faire l'hypothèse qu'on ne rencontrera au cours du déroulement de l'algorithme, aucune partie de dimension inférieure à une valeur donnée n'est absolument pas justifiable théoriquement ; on vérifie toutefois qu'elle est vérifiée dans la plupart des expériences.

b) Si on ne fait pas cette hypothèse, la forme générale de l'algorithme peut s'appliquer pour permettre d'optimiser le même critère de vraisemblance : les fonctions D et f restent inchangées, la fonction d'estimation g devient :

$$g : L_k \times P_k \to L_k$$

$$(L,P) \mapsto M ,$$

où

$$L = ((\mu_1, V_1), \ldots, (\mu_k, V_k)),\ P = (P_1, \ldots, P_k)$$

et

$$M = ((\mu'_1, V'_1), \ldots, (\mu'_k, V'_k))$$

tels que, pour tout i ($1 \leqslant i \leqslant k$) :

$$\mu'_i = \frac{1}{|P_i|} \sum_{x \in P_i} x \qquad \text{(maximum de vraisemblance)}$$

Si

$$\dim(P_i) \geqslant q + 1$$

et si :

$$\mathcal{V}_i = \frac{1}{|P_i|} \sum_{x \in P_i} (x - \mu'_i)(x - \mu'_i)'$$

alors

$$V'_i = \mathcal{V}_i \qquad \text{(maximum de vraisemblance)}$$

Si

$$\dim(P_i) < q + 1 \qquad \text{alors} \qquad V'_i = V_i .$$

Remarque :

Les lignes de niveau des points équidistants d'un μ_i au sens de D sont les ellipsoïdes d'équiprobabilité de la distribution gaussienne de paramètres (μ_i , V_i) qui a pour équation :

$$(x - \mu_i)' V_i^{-1} (x - \mu_i) = \text{Cste}$$

Les propriétés géométriques des solutions fournies dans ce cas sont détaillées dans (11) et (22).

En ce sens, l'algorithme peut être utilisé en classification automatique pour reconnaître les formes ellipsoïdales.

Exemples d'application à des données construites par simulation :

a) 150 points de $\mathbf{R}^2$ simulés de trois distributions gaussiennes bidimensionnelles équiprobables (cf. fig (a)) de paramètres :

$$\mu_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \qquad \mu_2 = \begin{pmatrix} 5 \\ 0 \end{pmatrix} \qquad \mu_3 = \begin{pmatrix} 0 \\ 5 \end{pmatrix}$$

$$V_1 = \begin{pmatrix} 1/4 & 0 \\ 0 & 1/4 \end{pmatrix} \qquad V_2 = \begin{pmatrix} 4 & 0 \\ 0 & 4 \end{pmatrix} = V_3$$

Après 5 tirages initiaux, le tirage ayant donné la meilleure valeur du critère (W = 850.) a apporté les résultats suivants en 10 itérations (cf.fig. (b)) :

$P_1 = 52/150 \qquad P_2 = 46/150 \qquad P_3 = 52/150$

$$\mu_1 = \begin{pmatrix} -.09 \\ -.01 \end{pmatrix} \qquad \mu_2 = \begin{pmatrix} 5.09 \\ .31 \end{pmatrix} \qquad \mu_3 = \begin{pmatrix} .16 \\ 4.89 \end{pmatrix}$$

$$V_1 = \begin{pmatrix} .21 & -.02 \\ -.02 & .29 \end{pmatrix} \qquad V_2 = \begin{pmatrix} 3.12 & .44 \\ .44 & 3.96 \end{pmatrix} \qquad V_3 = \begin{pmatrix} 4.06 & -.96 \\ -.96 & 3.25 \end{pmatrix}$$

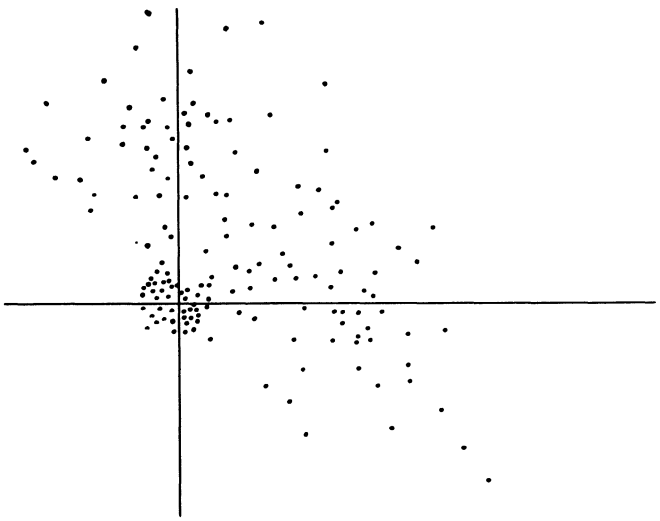

Fig (a) – Données initiales

b) Un échantillon artificiel unidimensionnel a été construit de la façon suivante : (cf. fig (c))

50 points tirés de la loi normale $\mu_1 = 0, \quad \sigma_1 = 1$

50 points tirés de la loi normale $\mu_2 = 3, \quad \sigma_2 = 2$

50 points tirés de la loi normale $\mu_3 = -5, \quad \sigma_3 = 2$

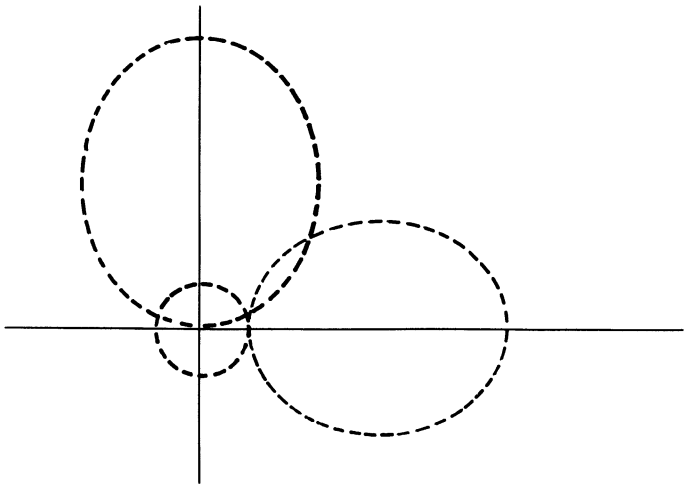

Fig (b) – Ellipsoïdes d'équiprobabilité (95 %) obtenues.

Moments empiriques des 3 échantillons ainsi obtenus :

$$\mu_1 = .1 \qquad \mu_2 = 3.4 \qquad \mu_3 = -4.9$$

$$\sigma_1 = 1. \qquad \sigma_2 = 1.8 \qquad \sigma_3 = 1.7$$

Résultats fournis par l'algorithme avec k = 3 et après 4 itérations pour atteindre la convergence :

$$\mu_1 = -.0 \qquad \mu_2 = 3.5 \qquad \mu_3 = -4.9$$

$$\sigma_1 = .9 \qquad \sigma_2 = 1.2 \qquad \sigma_3 = 1.7$$

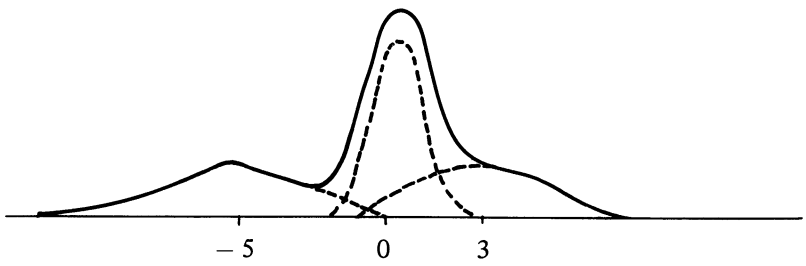

Fig (c)

3.2. – Lois Gamma

Considérons maintenant la famille $(f_\lambda / \lambda \in L)$ suivante :

$$\forall x \in R \qquad f_\lambda(x) \begin{cases} = \dfrac{\alpha^\beta}{\Gamma(\beta)} (x-\gamma)^{\beta-1} \exp(-\alpha(x-\gamma)) & \text{si } x \geqslant \gamma \\ = 0 & \text{si } x < \gamma \end{cases}$$

avec

$$\lambda = (\alpha, \beta, \gamma) \in \mathbf{R}^+ \times [1, +\infty] \times \mathbf{R} \subset \mathbf{R}^3.$$

et où Γ est la fonction eulérienne usuelle :

$$\Gamma(\beta) = \int_0^{\infty} t^{\beta-1} \exp(-t)\, dt$$

– Si β est entier, la distribution f_λ s'appelle loi d'Erlang (alors

$$\Gamma(\beta) = (\beta - 1)\ !\ )$$

et on remarque que c'est la somme de β exponentielles négatives.

Remarque :

La famille des distributions Gamma peut être définie pour tout β positif; toutefois selon que β est inférieur ou supérieur à 1, les densités considérées sont d'allures bien différentes et nous nous restreindrons donc à une seule de ces deux sous-familles. En outre, le problème concret qui est à l'origine de ce travail sur les mélanges de Gamma (cf. 4.) est en réalité une recherche de distributions d'Erlang, physiquement interprétées comme sommes d'exponentielles ; il est donc naturel pour cette application de se limiter au cas où $\beta \geqslant 1$.

On verra dans la suite l'utilité technique de cette hypothèse.

En choisissant à nouveau d'optimiser un critère de vraisemblance, on prendra pour D et f les mêmes fonctions qu'au paragraphe précédent :

$$D(x, \lambda) = \text{Cste} + \text{Log}\,\Gamma(\beta) - \beta\,\text{Log}\,\alpha - (\beta - 1)\,\text{Log}\,(x - \gamma) + \alpha\,(x - \gamma)$$

$$f(\lambda_1, \ldots, \lambda_k) = (P_1, \ldots, P_k) \qquad \text{où pour tout i,}$$

$$P_i = \{\, x \in E / \forall\, j \neq i, f_{\lambda_i}(x) \geqslant f_{\lambda_j}(x) \quad \text{avec} \quad j > i \text{ en cas d'égalité} \}$$

Le problème reste de trouver une fonction $g : \mathbf{L}_k \times \mathbf{P}_k \to \mathbf{L}_k$ qui améliore la vraisemblance et qui donne aux paramètres des valeurs ne conduisant qu'à des densités bornées ($\gamma < +\infty, \alpha < +\infty,\ 1 \leqslant \beta < +\infty$)

Notations :

$L = (\lambda_1 \ldots \lambda_k) \in \mathbf{L}_k$ s'écrira :

$$L = (A, B, C) \quad \text{où} \quad A = (\alpha_1 \ldots \alpha_k)$$

$$B = (\beta_1 \ldots \beta_k)$$

$$C = (\gamma_1 \ldots \gamma_k)$$

Nous proposons la fonction g suivante qui consiste à améliorer la vraisemblance en deux étapes et dont on vérifiera aisément qu'elle remplit les conditions requises :

$$g = g_2 \circ g_1 \quad \text{où :}$$

– $g_1 : \mathbf{L}_k \times \mathbf{P}_k \to \mathbf{L}_k \times \mathbf{P}_k$

$$((A, B, C), P) \mapsto ((A, B, C'), P)$$

tel que :

$$\forall\ i/1 \leqslant i \leqslant k : V_{(\alpha_i\beta_i)\gamma_i'}(P_i) = \max_{\gamma \in R} V_{(\alpha_i\beta_i)\gamma}(P_i)$$

Plus précisément, on a :

$$V_{(\alpha_i\beta_i)\gamma}(P_i) = \begin{cases} \prod_{x \in P_i} \frac{\alpha_i^{\beta_i}}{\Gamma(\beta_i)} (x-\gamma)^{\beta_i - 1} \exp[-\alpha(x-\gamma)] \\ \qquad \text{si } x > \gamma \text{ pour tout } x \in P_i \\ 0 \qquad \text{si il existe un } x \in P_i \text{ tel que } x \leqslant \gamma \end{cases}$$

d'où, si $x > \gamma$ pour tout $x \in P_i$:

$$\text{Log } V_{(\alpha_i\beta_i)\gamma}(P_i) = \alpha_i \mid P_i \mid \gamma + (\beta_i - 1) \sum_{x \in P_i} \text{Log}(x-\gamma) + \text{Cste}$$

Quantité dont on cherche le maximum pour

$$\gamma < \inf\{\ x/x \in P_i\}\ .$$

– Si $\beta_1 = 1$, $\text{Log } V_{(\alpha_i\beta_i)\gamma}(P_i) = \alpha_i \mid P_i \mid \gamma + \text{Cste}$

Il n'existe pas de maximum exact en γ sur

$$\{\gamma/\gamma < \inf\{\ x/x \in P_i\}\},$$

on prendra donc :

$$\gamma_i' = \inf\{\ x/x \in P_i\} - \epsilon$$

où ϵ (> 0) est une constante à determiner (qui soit par exemple de l'ordre de la précision de la machine utilisée ; il ne faut pas perdre de vue que, de toutes façons, le fait d'obtenir exactement $\beta = 1$ n'est vrai que moyennant la discrétisation introduite par l'utilisation du calculateur).

– Si $\beta_i > 1$, alors :

$$\frac{d\,\text{Log } V_{(\alpha_i\beta_i)\gamma}(P_i)}{d\gamma} = \alpha \mid P_i \mid - (\beta_i - 1) \sum_{x \in P_i} \frac{1}{x - \gamma}$$

de plus :

$$\frac{d^2\,\text{Log} V_{(\alpha_i\beta_i)\gamma}(P_i)}{d\gamma^2} = -(\beta_i - 1) \sum_{x \in P_i} \frac{1}{(x-\gamma)^2} < 0$$

et :

$$\frac{d\,\text{Log} V_{(\alpha_i\beta_i)\gamma}(P_i)}{d\gamma} \xrightarrow[\gamma \to -\infty]{} \alpha \mid P_i \mid > 0$$

$$\frac{d\,\text{Log} V_{(\alpha_i\beta_i)\gamma}(P_i)}{d\gamma} \xrightarrow[\gamma \to \inf\{x/x \in P_i\}]{} -\infty$$

La dérivée de la fonction de vraisemblance de γ est donc strictement décroissante de $\alpha \mid P_i \mid$ à $-\infty$ et admet, par suite, un zéro unique qui correspond à un maximum de la vraisemblance.

Dans ce cas, γ_i' est l'estimé par le maximum de vraisemblance de γ au vu de l'échantillon P_i et à $(\alpha_i \beta_i)$ fixés.

– $g_2 : L_k \times P_k \rightarrow L_k$

$$((A, B, C), P) \rightarrow (A', B', C') \quad \text{tel que :}$$

$$\forall\, i/1 \leqslant i \leqslant k : V_{(\alpha_i' \beta_i') \gamma_i}(P_i) = \max_{\substack{\alpha \in R^+ \\ \beta \in R^+}} V_{(\alpha \beta) \gamma_i}(P_i)$$

C'est-à-dire que (α_i, β_i) est le couple estimé de (α, β) par le maximum de vraisemblance au vu de l'échantillon P_i et à γ_i fixé.

Ce couple est obtenu ainsi :

$$\text{Log } V_{(\alpha\beta)\ \gamma_i}(P_i) = \sum_{x \in P_i} \text{Log } f_{(\alpha\beta)\ \gamma_i}(x)$$

$$\equiv \mid P_i \mid (-\text{Log } \Gamma(\beta) + \beta \text{ Log } \alpha) + (\beta - 1) \sum_{x \in P_i} \text{Log } (x - \gamma_i) - \alpha \sum_{x \in P_i} (x - \gamma_i)$$

d'où :

$$\frac{d\,[\text{Log } V_{(\alpha\beta)\ \gamma_i}(P_i)]}{d\,\alpha} = \mid P_i \mid \frac{\beta}{\alpha} - \sum_{x \in P_i} (x - \gamma_i)$$

et

$$\frac{d\,[\text{Log } V_{(\alpha\beta)\ \gamma_i}(P_i)]}{d\,\beta} = \mid P_i \mid \left(-\frac{\Gamma'(\beta)}{\Gamma(\beta)} + \text{Log } \alpha\right) + \sum_{x \in P_i} \text{Log } (x - \gamma_i)$$

– Ces deux dérivations sont bien valides puisque $\beta \geqslant 1$ et

$$\gamma_i < \inf\{\, x/x \in P_i\} -$$

Le couple $(\hat{\alpha}, \hat{\beta})$ maximisant la vraisemblance devra donc vérifier :

$$\begin{cases} \hat{\beta}/\hat{\alpha} = \dfrac{1}{\mid P_i \mid} \displaystyle\sum_{x \in P_i} \text{Log } (x - \gamma_i) \\[2ex] \dfrac{\Gamma'(\hat{\beta})}{\Gamma(\hat{\beta})} - \text{Log } \hat{\alpha} = \dfrac{1}{\mid P_i \mid} \displaystyle\sum_{x \in P_i} \text{Log } (x - \gamma_i) \end{cases}$$

Les conditions d'application des théorèmes de convergence étant remplies, l'algorithme décrit tend à maximiser le produit des vraisemblances des classes P_i pour les distributions correspondantes f_{λ_i}.

Exemple d'application :

On considère un échantillon de 200 valeurs simulées de la façon suivante :

100 points tirés d'une loi Gamma de paramètres :

$\alpha_1 = 1/3$ $\beta_1 = 1/$ $\gamma_1 = 0$

100 points tirés d'une loi Gamma de paramètres :

$\alpha_1 = 1$ $\beta_1 = 6$ $\gamma_1 = 3$

La figure (d) suivante montre les deux histogrammes observé (en trait plein) et estimé (en pointillé) pour le meilleur de 10 tirages aléatoires initiaux.

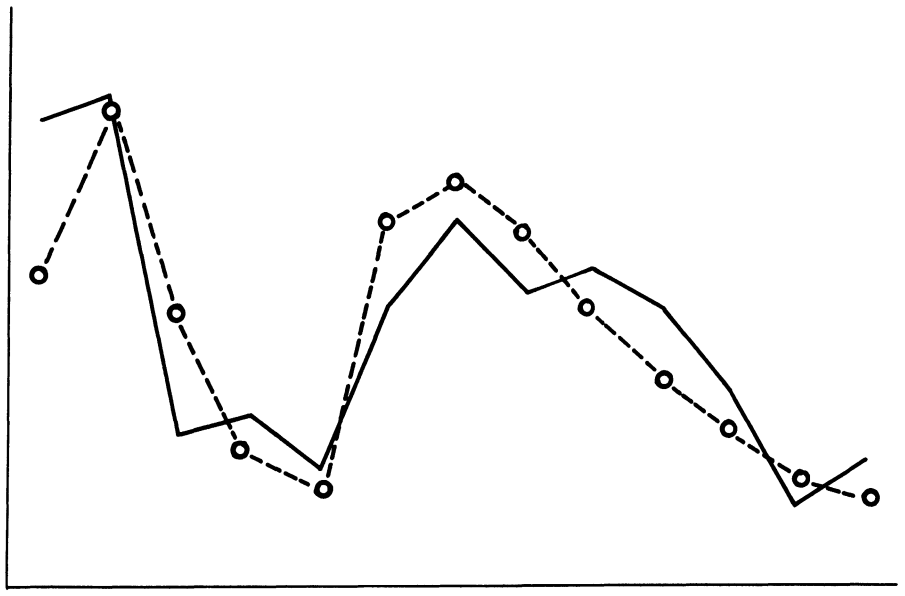

Fig (d)

Les estimations individuelles des paramètres ne sont pas bonnes ce qui n'est pas très étonnant étant donnée la méthode d'estimation utilisée à chaque pas qui n'a aucune des propriétés classiques. Cependant, l'algorithme a fourni une combinaison de Gamma qui ajuste globalement l'histogramme observé de façon admissible.

Résultats :

$|P_1| = 78$ $\alpha_1 = .8$ $\beta_1 = 1.3$ $\gamma_1 = .4$

$|P_2| = 122$ $\alpha_2 = .4$ $\beta_2 = 1.6$ $\gamma_2 = 4.9$

Deuxième exemple : (Fig. (e)) :

Echantillon initial : 100 observations $\rightarrow$ $\begin{cases} \alpha = 4. \\ \beta = 1. \\ \gamma = 0. \end{cases}$

et 100 observations → $\begin{cases} \alpha = 1. \\ \beta = 6. \\ \gamma = -1. \end{cases}$

Meilleur de 5 tirages : 105 observations → $\begin{cases} \alpha = 2.10 \\ \beta = 1.07 \\ \gamma = .002 \end{cases}$

95 observations → $\begin{cases} \alpha = .98 \\ \beta = 5.8 \\ \gamma = -.80 \end{cases}$

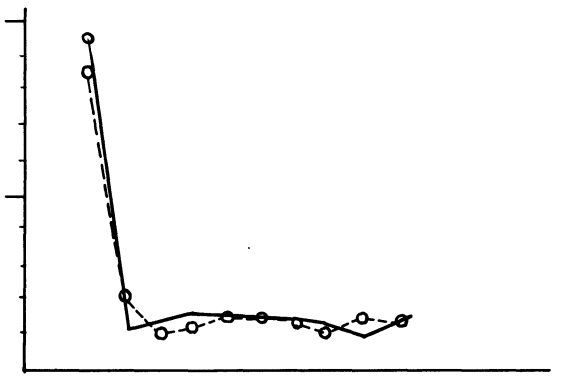

Fig (e)

Remarque :

Dans la mesure où le maximum de vraisemblance pose des problèmes pour l'estimation des paramètres des distributions Gamma, on pourrait choisir la fonction D et par suite, les fonctions f, R, g de façon à ce que la méthode d'estimation définie par g, soit, par exemple celle des moments. La méthode des moments présente en effet l'avantage de donner des estimateurs des trois paramètres α, β, γ très faciles à calculer à partir des trois premiers moments empiriques d'un échantillon A de n observations :

$$\alpha = 2m_2/m_3 \qquad \beta = 4m_2^3/m_3^2 \qquad \gamma = m_1 - 2m_2^2/m_3$$

où :

$$\forall\, i/1 \leqslant i \leqslant 3 : m_i = \frac{1}{n} \sum_{x \in A} x^i$$

Toutefois nous n'avons pas retenu cette possibilité à cause de l'imprécision de l'estimation par les moments ; en effet, les estimés fournis sont fluctuants tant que l'échantillon n'est pas de très grande taille, ce qui est extrêmement gênant dans le cas d'un algorithme itératif où les classes peuvent varier considérablement de taille d'une itération à l'autre.

L'optimisation en deux étapes de la vraisemblance, telle que nous l'avons

présentée plus haut, est beaucoup plus satisfaisante sur ce point et s'est montrée plus efficace sur des échantillons simulés.

4 – UNE APPLICATION : AIDE A LA MODELISATION DE SYSTEMES INFORMATIQUES. (cf. (15))

De façon générale, l'objet de la modélisation est de fournir une description mathématique du fonctionnement d'un système physique quelconque (stock de marchandises, système informatique, système biologique,...).

Le modèle permet, connaissant les valeurs prises par un certain nombre de paramètres d'entrée, de calculer des valeurs probables pour d'autres paramètres (de sortie).

Dès que le système analysé est un peu complexe, les modèles utilisés sont de type stochastique, c'est-à-dire que les entrées comme les sorties ne sont plus les valeurs numériques certaines des paramètres, mais leurs distributions de probabilité.

En modélisation de systèmes informatiques, un certain nombre de travaux théoriques permettent d'utiliser des modèles de réseaux à files d'attente où les temps de service peuvent être supposés distribués comme des mélanges de distributions d'Erlang ((4), (6)) ou même de façon quelconque dans le cas d'approximation par un processus de diffusion (14).

Notre application a consisté à utiliser l'algorithme présenté dans cet article sur un échantillon de mesures prélevées sur un système réel pour calculer des distributions effectives pour les temps de service ; ces formules peuvent servir ensuite dans des modèles mathématiques et aussi dans des simulations pour générer des échantillons.

Le problème informatique et les résultats sont présentés en détail dans (15).

5 – CONCLUSION

Le problème de la reconnaissance des composants d'un mélange s'il est constamment posé dans la pratique, est encore loin d'être résolu de façon générale. L'algorithme que nous proposons présente vis-à-vis des techniques existantes une certaine souplesse dans le choix du nombre de composants, du type de lois recherché dans le mélange et dans la dimension de la population observée. Il resterait à étudier des variantes permettant d'optimiser des critères statistiques habituels tels que le χ^2 ou la distance de Kolmogorov-Smirnov entre distribution observée et distribution estimée.

Pour analyser des mélanges de distributions de types différents (par exemple, normales et Log-normales...) on peut utiliser le même schéma ; il suffit de se donner en plus une mesure de "ressemblance" entre un échantillon et une forme de distribution. Il conviendrait d'approfondir le cas où cette ressemblance serait précisément liée à la vraisemblance (cf. (16)).

BIBLIOGRAPHIE

(1) AGRAWALA A.K. (1970) – "Learning with a probabilistic teacher" – *IEEE Trans. on Information theory* – vol. IT 16, n°4 –

(2) BENZECRI J.P. (1972) – "La Régression" Laboratoire de Statistique Mathématique – Université Paris VI.

(3) BHATTACHARYA C.G. (1967) – "A simple method of résolution of a distribution into gaussian components" *Biometrics* – March.

(4) BUZEN J.P. (1971) – "Queuing network models of multiprogramming" Ph. D. thesis – Harvard.

(5) COOPER D.B. et COOPER P.W. (1964) – "Non supervised adaptive signal detection and pattern recognition" *Information and Control* – n° 7 – p. 416 à 444.

(6) COX D.R. (1955) – "A use of complex probabilities in the theory of Stochastic Processes" *Proc. Camb. Phil. Soc.* – 51.

(7) DAY N.E. (1969) – "Estimating the components of a mixture of normal distributions" *Biometrika* – vol. 56 – n° 3 – p. 463 à 467

(8) DIDAY E. (1972) – "Nouvelles méthodes et nouveaux concepts en classification automatique et reconnaissance des formes" Thèse d'Etat – Université Paris VI.

(9) DIDAY E. et GOVAERT G. (1974) – "Classification avec Distance adaptative" *Compte rendu à l'Académie des Sciences – Paris* – Série A – t. 278 – p. 993 à 995 –

(10) DIDAY E., SCHROEDER A. (1976) – "A new approach in mixed distributions detection" RAIRO – Recherche Operationnelle – vol. 10, n° 6.

(11) DIDAY E., SCHROEDER A. et OK Y. (1974) – "The Dynamic Clusters Method in Pattern Recognition" *Proceedings of IFIP Congress* – Stockholm.

(12) DOROFEYUK A.A. (1971) – "Automatic Classification algorithms" *Automation and remote control* (review) vol. 32 – n° 12 – Part 1 – p. 1928 à 1958 – Dec.

(13) DUDA R.O. et HART R.E. (1973) – "Pattern Classification and Scene Analysis" Wiley N.Y.

(14) GELENBE E. (1974) – "On approximate computer systems models" International Workshop on Computer architectures and networks modelling and evaluation – North Holland Pub. C°

(15) LEROUDIER J., SCHROEDER A. (1975) – "A statistical approach to the estimation of service times distributions for operating systems modelling" – E. Gelenbe, D. Potier (Eds.) *International Computing Symposium* 1975 – (North – Holland Publ. C°)

(16) LINDSEY J.K. (1974) – "Comparison of Probability distributions" *J.R. Statist. Soc. B* n° 1 – p. 38 à 47.

(17) – PATRICK E.A. (1972) – "Fundamentals of Pattern Recognition" Prentice Hall inc. – N.J.

(18) PATRICK E.A. et COSTELLO J.P. (1970) – "On unsupervised estimation algorithms" *IEEE Trans. on Information theory* vol. IT 16 – n° 5 – p. 556 à 569 –

(19) PATRICK E.A. et HANCOCK J.C. (1966) – "Nonsupervised sequential classification and recognition of patterns" *IEEE Trans. on Information Theory* – vot. IT 12 – n° 3 p. 362 à 372 -

(20) PEARSON K. (1894) – "Contributions to the mathematic theory of evolution" *Philos. Trans. Soc.* – n° 185

(21) RAO C.R. (1948) – "Utilization of multiple measurement in problems of biological classification" *J.R. Statist. Soc. B.* – vol X – n° 2 – p. 159 à 193.

(22) SCHROEDER A. (1974) – "Reconnaissance des Composants d'un mélange". Thèse de 3e cycle – Université Paris VI

(23) YOUG T.Y. et CORALUPPI G. (1970) – "Stochastic estimation of a mixture of normal density functions using an information criterion". *IEEE Trans. on Information Theory* – vol. IT 16 – n° 3 p. 258 à 263 –