REVUE FRANÇAISE D'AUTOMATIQUE, INFORMATIQUE, RECHERCHE OPÉRATIONNELLE. RECHERCHE OPÉRATIONNELLE

E. DIDAY

A. SCHROEDER

A new approach in mixed distributions detection

Revue française d'automatique, informatique, recherche opérationnelle. Recherche opérationnelle, tome 10, n° V2 (1976), p. 75-106

<http://www.numdam.org/item?id=RO_1976__10_2_75_0>

R.A.I.R.O. Recherche opérationnelle
(vol. 10, n° 6, juin 1976, p. 75 à 106)

A NEW APPROACH IN MIXED DISTRIBUTIONS DETECTION (*)

by E. Diday et A. Schroeder [1]

Abstract. — *Our aim is to detect in a given multivariate sample the possible presence oj sub-samples drawn from probability distributions of some known type.*

In the first paragraph, the Dynamic Clusters method is described in the general case of unspecified kernels. Afterwards, it is applied to the above quoted problem of mixed distributions detection, taking probability distribution functions as kernels.

As presented here, the algorithm maximizes a likelihood criterion function.

1. INTRODUCTION

1.1. The Problem

From a multivariate sample, all multidimensional techniques (principal components, factor analysis, multidimensional scaling, clustering techniques, hierarchical or not) give a description of the population, that have afterwards to be interpreted either by further statistical methods (e. g. discrimination, ...) or by common sense. These techniques need no probabilistic assumption, neither necessarily refer to any probabilistic model.

On the other hand, problems and techniques of modelization exist which try to adapt a stochastic model to a real phenomenon.

Our own purpose consists to detect if a given sample contains sub-samples which could have been drawn from populations distributed according to a known family of probability distribution functions.

More precisely, given a sample, we shall try to find if it results from the "effects" of several stochastic phenomena arising from different distributions.

This problem is the classical Resolution of Mixtures when one can assume that the overall distribution function—which will be written $F(x)$—actually follows the model:

$$F(x) = \sum_{1 \leqq j \leqq k} P_j F_j(x), \tag{1}$$

(*) Reçu janvier 1975, version révisée décembre 1975.

[1] I.R.I.A. Rocquencourt.

- P_j: *a priori* probability of the *j*th distribution;
- $F_j(x)$: *j*th distribution function belonging to a known family.

In terms of signal theory, this same problem can be expressed (in the case: $k = 2$) as "a model of observation signal consisting of a mixture of two unknown pulse waveforms of some duration T, which occur independently successively at random with probabilities P_1 and P_2, imbedded in additive zero-mean stationary Gaussian noise with u.known power" [from W. D. Gregg and J. C. Hancock (1969)].

1.2. Various Approaches

A number of techniques have been proposed to this end. We shall classify them in two categories, according to the problem they solve.

Those, that use model (1) in its analytical form, and that estimate the P_j and the unknown parameters on which the F_j depend:

$$F_j(x) = \varphi(x, \theta_j) \qquad \text{with} \quad \theta_j \in \mathbf{R}^s$$

unknown, but φ known.

Those that are first looking for components from a mixture of a given type in the observed sample, and estimate afterwards the unknown parameters.

The first category techniques are estimation techniques and they only differ in the type of estimators they use: method of moments [Pearson (1894)] with maximum likelihood estimates [Rao (1948), Day (1969)], minimum χ^2, etc. Most of them are adapted to Gaussian distributions and often to the univariate case; let us note two of them: Rao (1948), specific for two classes mixtures; Battacharya (1967), who gives a graphical method to determine the number of classes, but needs a large number of observations to be collected and the different distributions to be adequately separated; for a general review of these techniques, *see* Dorofeyuk (1971).

Day (1969) also deals with two components, that may be multivariate.

Another approach to the estimation problem of model (1) is Cooper and Cooper's (1964): the unknown parameters are deduced from the moments of the overall observed distribution.

In order to study the multivariate case, particularly when the number of components is larger than two, many assumptions have to be made (e. g. Day assumes that the covariance matrices are equal; from a practical point of view, Cooper and Cooper study the case of two distributions only differing by their means).

The second category techniques include bayesian methods, stochastic approximation, supervised or non supervised learning,...

The algorithms of approximation type are very different. Most of them attempt a bayesian approach [Patrick and Hancock (1966), Patrick and

Costello (1970), Patrick (1972), Agrawala (1970)] though giving quite various techniques. The hypotheses differ from one method to another, but are usually very restrictive.

This kind of approach enables to formalize the mixed distributions detection problem in terms of unsupervised learning [Agrawala (1970), Patrick (1972), Duda and Hart (1973)].

An information theory criterion can also be used for a stochastic approximation algorithm [Young and Coraluppi (1970)] which is very interesting, because it is not needed to know *a priori* the actual number of components in the mixture; it is however restricted to one-dimensional Gaussian distributions.

1.3. The Dynamic Clusters Approach

Our own approach may be roughly classified in the second category techniques. We only want to detect in the population the possible presence of samples of some known distribution, but we do not make any assumptions about the global distribution.

However, if the user can make hypotheses on the representativeness of the population as a sample of a global population and can admit model (1) for the overall distribution, then our algorithm will give him a solution of the mixture type, as soon as goodness of fit may be proved.

We shall use an algorithm of the "Dynamic clusters" type (*cf.* Diday, [6] to [9]) i. e. an algorithm that detects parallely clusters among the observations and some typical features for these clusters: the "kernels"; for us the feature to find will be fitness with a probability density of some known type.

Our algorithm will be presented in a general form for any density function, any dimension for the sample space and any number of components.

The only input we need is the form of the probability distribution function and the number of components. However, we can see that this last hypothesis is not really restrictive, because the actual number can be found by several means even if the algorithm has run with another *a priori* number.

D. C. algorithm has been thoroughly studied in E. Diday ([6] to [9]) in two particular cases: when the kernels are linear manifolds (in factor analysis), and when they are subsets of the population, in non-hierarchical clustering; in that case, the algorithm belongs to the class of methods such as Isodata (Hall and Ball, 1965), *k*-means (MacQueen, 1967), iterative relocation (Wishart, 1971).

In the first part of the paper, we shall recall the D. C. (Dynamic Clusters) algorithm in the general case—for any set of kernels—and its convergence properties. Afterwards, the problem will be formalized in precise terms and D. C. applied to it.

Several practical remarks will then be made, essentially concerning the number of components. Afterwards, the particular case of mixed gaussian distributions will be developped, as it brings interesting results from a geometrical viewpoint. We shall then conclude with several examples.

2. THE DYNAMIC CLUSTERS ALGORITHM

2.1. Notation

– Let E be any finite subset of $\mathbf{R}^q$;

– $\mathbf{P}_k$: the set of all partitions of E into k classes; the elements of $\mathbf{P}_k$ will be called: k-partitions.

$$P \in \mathbf{P}_k \quad \Leftrightarrow \quad P = (P_1, \ldots, P_k);$$

– $\mathbf{L}$: a set that will be called the space of "kernels". These kernels will be associated with subsets of E, as a characterization of these subsets depending on the application of the algorithm;

– $\mathbf{L}_k$: the set of all k-tuples of $\mathbf{L}$:

$$L \in \mathbf{L}_k \quad \Leftrightarrow \quad L = (\lambda_1, \lambda_2, \ldots, \lambda_k) \text{ where } \lambda_i \in \mathbf{L}, \qquad \forall i \in]k]$$
$$(\text{if}]k] = \{1, 2, \ldots, k\}).$$

The D. C. algorithm solves the following problem: To find a couple (P, L) where $P \in \mathbf{P}_k$ and $L \in \mathbf{L}_k$ that minimizes some criterion function, which will be denoted W: $\mathbf{L}_k \times \mathbf{P}_k \longrightarrow \mathbf{R}^+$. Many problems can be written in these terms as long as any choice is possible for the kernels and the criterion function.

The general idea of the algorithm is then quite simple: It consists in deducing from any element of $\mathbf{L}_k$ an element of $\mathbf{P}_k$, and from this element of $\mathbf{P}_k$ an element in $\mathbf{L}_k$, so that the values of the criterion function on successive couples (L, P) decrease. To be precise, we shall need some more notation:

$$.D: \quad E \times \mathbf{L} \rightarrow \mathbf{R}^+, \quad (x, \lambda) \mapsto D(x, \lambda)$$

will express a "distance" between an observation and a kernel.

$$.R: \quad \mathbf{L} \times]k] \times \mathbf{P}_k \rightarrow \mathbf{R}^+, \quad (\lambda, i, P) \mapsto R(\lambda, i, P),$$

will give a measure of the goodness of fit of kernel λ with the ith element P_i of the k-partition P.

The way to associate a k-partition with a k-tuple of kernels will then be given by the following function:

$$.f: \quad \mathbf{L}_k \rightarrow \mathbf{P}_k,$$

$$L = (\lambda_1, \ldots, \lambda_k) \quad \Rightarrow \quad f(L) = P = (P_1, \ldots, P_k),$$

where:

$$P_i = \{ x \in E / D(x, \lambda_i) \leqq D(x, \lambda_j), \forall j \in]k] \}.$$

(In case of equality, x will be assigned to the lower index class.)

P_i is therefore built with all elements of E that are "nearer" (in D sense) to λ_i than to any other kernel of L.

Reciprocally, to associate a k-tuple of kernels to a k-partition, we shall introduce the function g:

$$.g: \quad \mathbf{P}_k \rightarrow \mathbf{L}_k,$$

$$P = (P_1, \ldots, P_k) \quad \Rightarrow \quad g(P) = L = (\lambda_1, \ldots, \lambda_k),$$

where, $\forall\, i \in]k]$, λ_i is given by:

$$R(\lambda_i, i, P) = \min_{\lambda \in \mathbf{L}} R(\lambda, i, P).$$

In other words, among all possible kernels, λ_i is chosen as the nearest (in terms of R) to P_i. [If this definition leads to several L, one has to define a unique choice, so that, for any $P \in \mathbf{P}_k$, $g(P)$ is a well determined element of $\mathbf{L}_k$].

The criterion to minimize will be defined as:

$$.\, W: \quad \mathbf{L}_k \times \mathbf{P}_k \rightarrow \mathbf{R}^+,$$

$$v = (L, P) \quad \Rightarrow \quad W(v) = \sum_{i \in]k]} R(\lambda_i, i, P).$$

The problem is now an optimization problem:

$$\text{Minimize } W(v) \quad \text{for all } v \in \mathbf{L}_k \times \mathbf{P}_k. \qquad (2)$$

2.2. The Algorithm

The D. C. algorithm is based on two sequences:

- (v_n), in $\mathbf{L}_k \times \mathbf{P}_k$, i. e. $v_n = (L^{(n)}, P^{(n)})$;
- and $u_n = W(v_n) \in \mathbf{R}^+$.

Let $P^{(0)}$, be any initial k-partition (it can either be drawn at random, or chosen) and $L^{(0)} = g(P^{(0)}) \rightarrow v_0 = (L^{(0)}, P^{(0)})$. The sequence (v_n) is then defined recursively:

$v_{n+1} = (L^{(n+1)}, P^{(n+1)})$ is deduced from v_n by: $P^{(n+1)} = f(L^{(n)})$ and $L^{(n+1)} = g(P^{(n+1)})$.

We shall show that — under certain constraints — the sequence $u_n = W(v_n)$ decreases. As it is a sequence in $\mathbf{R}^+$, it converges, and we shall see (th. 2) that its limit is attained:

$$\exists M, \quad \forall n \geqq M, \qquad u_n = u^*.$$

A couple $v^* = (L^*, P^*)$ such as $W(v^*) = u^*$ will be called a LOCAL OPTIMUM for the problem.

A couple (L^*, P^*) will be called a GLOBAL OPTIMUM if:

$$W(v^*) \leqq W(v) \quad \text{for all } v \in \mathbf{L}_k \times \mathbf{P}_k.$$

For a v^* given by the algorithm, this inequality only holds for v in a part of $\mathbf{L}_k \times \mathbf{P}_k$, this is why it is called a "local" optimum. For further information on this optimality, *see* Diday [9].

We shall see now how (u_n) decreases.

— proofs of the results are given for self-consistency but are not necessary to understand the following sections. —

Definition 1: A function $R : \mathrm{L} \times]k] \times \mathbf{P}_k \rightarrow \mathbf{R}^+$ is said *SEMI-SQUARE*, if:

$$\text{(1)} \qquad \forall L \in \mathbf{L}_k, \quad W(g \circ f(L), f \circ g \circ f(L)) \leqq W(g \circ f(L), f(L))$$

(The sign $\circ$ is used to denote the composition of functions).

(1) can be written equivalently:

$$\text{(1')} \qquad \forall P \in f(\mathbf{L}_k), \quad W(g(P), f \circ g(P)) \leqq W(g(P), P).$$

Theorem 1: *If R is semi-square, the sequence (u_n) decreases and therefore converges.*

Proof: We shall prove the following inequality:

$$u_{n+1} \leqq u_n, \qquad \forall n,$$

in two steps:

$$a) \qquad u_{n+1} = W(L^{(n+1)}, P^{(n+1)}) \leqq W(L^{(n)}, P^{(n+1)})$$

and

$$b) \qquad W(L^{(n)}, P^{(n+1)}) \leqq W(L^{(n)}, P^{(n)}) = u_n,$$

$$a) \qquad u_{n+1} = \sum_{i \in]k]} R(\lambda_i^{(n+1)}, i, P^{(n+1)})$$

and

$$W(L^{(n)}, P^{(n+1)}) = \sum_{i \in]k]} R(\lambda_i^{(n)}, i, P^{(n+1)}).$$

Since, for each $i \in]k]$, $\lambda_i^{(n+1)}$ is deduced from $P^{(n+1)}$ by function g, so that:

$$R(\lambda_i^{(n+1)}, i, P^{(n+1)}) = \min_{\lambda \in \mathbf{L}} R(\lambda, i, P^{(n+1)})$$

we have in particular:

$$\forall i \in]k], \quad R(\lambda_i^{(n+1)}, i, P^{(n+1)}) \leqq R(\lambda_i^{(n)}, i, P^{(n+1)})$$

and therefore, *a*) is proved.

b) As soon as $n \geqq 1$, $P^{(n)} = f(L^{(n-1)})$ and then $P^{(n)} \in f(\mathbf{L}_k)$. Since R is assumed to be semi-square, by property (1′) of definition 1 applied to $P^{(n)}$:

$$W(g(P^{(n)}), f \circ g(P^{(n)})) \leqq W(g(P^{(n)}), P^{(n)})$$

and as:

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

we have:

$$W(L^{(n)}, P^{(n+1)}) \leqq W(L^{(n)}, P^{(n)}),$$

which proves *b*).

Then, (u_n) decreases and as it is a sequence in $\mathbf{R}^+$, it converges.

Q. E. D.

Remark: The property that R be semi-square is necessary and sufficient to prove *b*), but it is only sufficient to have (u_n) decreasing since we could have $u_{n+1} \leqq u_n$ without the intermediate inequalities we have used.

Proposition: *The two following conditions* (2) *and* (3) *on* R *are sufficient to have* R *semi-square. Moreover* (2) *implies* (3).

$$(2) \qquad \left.\begin{array}{l} \forall L \in \mathbf{L}_k \\ \forall P \in \mathbf{P}_k \end{array}\right\} W(L, f(L)) \leqq W(L, P),$$

$$(3) \qquad \forall L, M \in \mathbf{L}_k; \quad W(L, f(M)) \leqq W(M, f(M)) \Rightarrow W(L, f(L)) \leqq W(L, f(M)).$$

If (3) is true R is said to be *SQUARE.*

Proof: One can see easily that (2) ⇒ (1) and (2) ⇒ (3).

Let us prove that (3) ⇒ (1):

As (3) stands for all L and M, we can write it for $L = g \circ f(M)$, then:

$$(3) \quad \Rightarrow \quad \forall M \in \mathbf{L}_k,$$

$$\begin{aligned} & W(g \circ f(M), f \circ g \circ f(M)) \leqq W(M, f(M)) \\ \Rightarrow \quad & W(g \circ f(M), f \circ g \circ f(M)) \leqq W(g \circ f(M), f(M)). \end{aligned}$$

By definition of g and W from R, we have:

$$\forall M \in \mathbf{L}_k, \quad \forall P \in \mathbf{P}_k, \qquad W(g(P), P) \leqq W(M, P).$$

[This had in fact already been proved in the inequality *a*) of theorem 1.]

Therefore, the left-hand side of the implication is always true and the implication is reduced to its right-hand side, for all M, which is exactly the condition (1).

Q. E. D.

COROLLARY 1: *If R is defined as:*

$$\forall \lambda \in \mathbf{L}, \quad \forall P \in \mathbf{P}_k, \qquad R(\lambda, i, P) = \sum_{x \in \mathbf{P}_i} D(x, \lambda),$$

then the sequence (u_n) decreases and converges.

Proof: We shall see that the condition (2) is true for such an R: Let us take

$$L = (\lambda_1, \ldots, \lambda_k) \qquad \text{and} \qquad f(L) = Q = (Q_1, \ldots, Q_k).$$

Then, for all $P \in \mathbf{P}_k$:

$$\begin{aligned} W(L, P) &= \sum_{j \in]k]} \sum_{x \in P_j} D(x, \lambda_j) \\ &= \sum_{x \in E} \sum_{j \in]k]} D(x, \lambda_j)\, \delta_{P_j}(x) \end{aligned}$$

[where $\delta_{P_j}(.)$ is the characteristic function of P_j, i. e.: $\delta_{P_j}(x) = 1$ if $x \in P_j$ and $= 0$ otherwise.]

On the other hand:

$$\begin{aligned} W(L, f(L)) &= \sum_{i \in]k]} R(\lambda_i, i, Q) = \sum_{i \in]k]} \sum_{x \in Q_i} D(x, \lambda_i) \\ &= \sum_{x \in E} \sum_{i \in]k]} D(x, \lambda_i)\, \delta_{Q_i}(x). \end{aligned}$$

As $Q = f(L)$, and by definition of f, we know that $x \in Q_i$ if

$$D(x, \lambda_i) \leqq D(x, \lambda_j)$$

for all j, which implies, for all $L \in \mathbf{L}_k$ and $P \in \mathbf{P}_k$: $W(L, f(L)) \leqq W(L, P)$.

Q. E. D.

As we have studied the convergence of (u_n), we shall now see how (v_n) can give a solution to the problem (2).

DEFINITION: An element $v = (L, P) \in \mathbf{L}_k \times \mathbf{P}_k$ is said to be *UNBIASED* for the functions f and g, if $P = f \circ g(P)$ and $L = g \circ f(L)$.

DEFINITION: A sequence $(L^{(n)}, P^{(n)})$ in $\mathbf{L}_k \times \mathbf{P}_k$ is said to be convergent if there exists a M such that:

$$\forall n \geqq M, \quad (L^{(n)}, P^{(n)}) = (L^*, P^*).$$

THEOREM 2: *If R has the two following properties:*

(i) *semi-square,*

(ii) $\forall P \in f(\mathbf{L}_k)$, $\forall i \in]k]$, *$R(\lambda, i, P)$ is minimum for a unique λ_0. Then, (v_n) is convergent and its limit is an unbiased element.*

Proof: Since E is assumed to be a finite set, $\mathbf{P}_k$ is finite too. By definition, E is such that, for any $P \in \mathbf{P}_k$, $g(P)$ can only take one value in $\mathbf{L}_k$, then $g(\mathbf{P}_k) \subset \mathbf{L}_k$ is finite too and (v_n) and (u_n) can only take a finite number of values. As (u_n) converges, its limits u^* is reached:

$$\exists M, \quad \forall n \geqq M, \qquad u_n = u^*.$$

Thus, $\forall n \geqq M$, $W(L^{(n)}, P^{(n)}) = W(L^{(n+1)}, P^{(n+1)})$ which implies that the two inequalities *a*) and *b*) of theorem 1 are equalities:

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

The second equality may be written:

$$\sum_{i \in]k]} R(\lambda_i^{(n)}, i, P^{(n+1)}) = \sum_{i \in]k]} R(\lambda_i^{(n+1)}, i, P^{(n+1)}).$$

The hypothesis (ii) implies:

$$\forall i \in]k], \quad R(\lambda_i^{(n+1)}, i, P^{(n+1)}) \leqq R(\lambda_i^{(n)}, i, P^{(n+1)}),$$

with equality if and only if $\lambda_i^{(n+1)} = \lambda^{(n)}$.

We then have two sums of positive which are equal while every term of one of them is less than or equal to the corresponding term of the other. This is only possible if all corresponding terms are equal and therefore:

$$\lambda_i^{(n+1)} = \lambda_i^{(n)}, \qquad \forall i \in]k] \quad \Leftrightarrow \quad L^{(n+1)} = L^{(n)}.$$

Then $P^{(n+1)} = f(L^{(n)})$ and $P^{(n+2)} = f(L^{(n)})$ imply that $P^{(n+1)} = P^{(n+2)}$ and the convergence of (v_n) is proved:

$$\forall n \geqq M+1, \quad v_n = v_{n+1} = v^*,$$

$$v^* = (L^*, P^*) = (L^{(n)}, P^{(n)}) = (g(P^{(n)}), P^{(n)})$$

$$\Rightarrow \begin{cases} P^* = P^{(n)}, \\ L^* = g(P^{(n)}) = g(P^*), \end{cases}$$

and

$$v^* = (L^*, P^*) = (L^{(n+1)}, P^{(n+1)}) = (g \circ f(L^{(n)}), f(L^{(n)}))$$

$$\Rightarrow \begin{cases} P^* = f(L^*) = f \circ g(P^*), \\ L^* = g \circ f(L^*). \end{cases}$$

Therefore v^* is an unbiased element for f and g.

Q. E. D.

3. MIXED DISTRIBUTIONS DETECTION

3.1. The Problem

We shall now write in mathematical terms the problem we have informally described in the introduction.

Let E be a set of N observations on which q measures have been taken: then E is a finite subset of $\mathbf{R}^q$.

Suppose we are given a family of probability density functions: $(f_\lambda)_{\lambda \in \mathbf{L}}$, which depends on the parameter λ, with $\lambda \in \mathbf{L} \subset \mathbf{R}^s$. [For instance, if $q = 1$, this family could be that of Gaussian univariate distribution with $\lambda = (\mu, \sigma)$, $s = 2$, $\mathbf{L} = \mathbf{R} \times \mathbf{R}^+$.]

We want then to find a couple (L, P), where $L = (\lambda_1, \ldots, \lambda_k)$, $\lambda_i \in \mathbf{L}$ and $P = (P_1, \ldots, P_k)$ is a k-partition of E, such that for all $i \in]k]$, P_i may be considered as a "likely" sample of the distribution f_{λ_i}.

To this end, we shall try to maximize the product of the likelihoods of the k "samples" P_i for the densities f_{λ_i}, or, in other words, to find L^* and P^* such that:

$$\boxed{\mathscr{L}(L^*, P^*) = \max_{\substack{L \in \mathbf{L}_k \\ P \in \mathbf{P}_k}} \prod_{i \in]k]} \prod_{x \in P_i} f_{\lambda_i}(x).}$$

Let us now note

$$V_{\lambda_i}(P_i) = \prod_{x \in P_i} f_{\lambda_i}(x).$$

We shall show that, in fact, the D. C. algorithm make the following criterion [1] on (L, P) decrease:

$$W(L, P) = K - \sum_{i \in]k]} \operatorname{Log} V_{\lambda_i}(P_i)$$

(where K is a constant).

3.2. The Algorithm

Let us take:

- $E \in \mathbf{R}^q$. The finite set to classify.
- $\mathbf{L} \in \mathbf{R}^s$. Set of the kernels (which will be exactly the set of parameters introduced above).
- $$\begin{aligned} D:\ & E \times \mathbf{L} \to \mathbf{R}^+, \\ & (x, \lambda) \mapsto D(x, \lambda) = \operatorname{Log}(f^*/f_\lambda(x)), \end{aligned}$$

where $f^* \geqq \max \{ f_\lambda(x)/\lambda \in \mathbf{L},\ x \in E \}$; it is sufficient to know such an f exists to use the algorithm; for instance, for univariate Gaussian distributions with $\lambda = (\mu, \sigma)$:

$$\max_{x \in \mathbf{R}} f_\lambda(x) = (2\Pi)^{-1/2} \sigma^{-2}$$

and then a possible f^* is:

$$f = (2\Pi)^{-1/2} d_0^{-1} \qquad \text{if} \quad d_0 = \min_{\substack{x \in E \\ y \in E}} |x - y|^2.$$

This definition for D expresses that the greater $f_\lambda(x)$ is, the nearer to kernel λ the observation x is. (It can also be said that the likelihood of the sample $\{ x \}$ is large for f_λ).

- R is then defined from D:

$$\begin{aligned} R:\ & \mathbf{L} \times]k] \times \mathbf{P}_k \to \mathbf{R}^+, \\ & (\lambda, i, P) \mapsto R(\lambda, i, P) = \sum_{x \in P_i} D(x, \lambda), \end{aligned}$$

$$\begin{aligned} R(\lambda, i, P) &= \sum_{[x \in P_i]} \operatorname{Log}(f^*/f_\lambda(x)) \\ &= \operatorname{Log}\Big[(f^*)^{|P_i|} / \prod_{x \in P_i} f_\lambda(x)\Big] (\text{where } |P_i| = \operatorname{card} P_i) \\ &= \operatorname{Log}\big[(f^*)^{|P_i|} / V_\lambda(P_i)\big]. \end{aligned}$$

[1] The same problem may of course be formalized in other terms.

- Finally, the criterion is:

$$W: \ \mathbf{L}_k \times \mathbf{P}_k \to \mathbf{R}^+,$$
$$(L, P) \mapsto W(L, P) = \sum_{i \in]k]} R(\lambda_i, i, P)$$

[if $L = (\lambda_1, \ldots, \lambda_k)$],

$$W(L, P) = \sum_{i \in]k]} \operatorname{Log}[(f^*)^{|P_i|}/V_{\lambda_i}(P_i)],$$

$$W(L, P) = \operatorname{Log}(f^*)^N - \sum_{i \in]k]} \operatorname{Log} V_{\lambda_i}(P_i) \ (\text{since } \sum_{i \in]k]} |P_i| = N).$$

So the above formalization leads us to a criterion which expresses that we shall maximize the product of the likelihoods of the k samples P_i.

The two fundamental functions of the algorithm, f and g, become:

- $$f: \ \mathbf{L}_k \to \mathbf{P}_k,$$
$$L = (\lambda_1, \ldots, \lambda_k) \ \Rightarrow \ P = (P_1, \ldots, P_k),$$

where

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

(x being assigned to the lower index class in case of quality).

The elements of E are therefore assigned to the class to which they more likely belong.

- $$g: \ \mathbf{P}_k \to \mathbf{L}_k,$$
$$P = (P_1, \ldots, P_k) \ \Rightarrow \ L = (\lambda_1, \ldots, \lambda_k)$$

where λ_i is such that:

$$R(\lambda_i, i, P) = \min_{\lambda \in \mathbf{L}} R(\lambda, i, P)$$

$$\Leftrightarrow \quad R(\lambda_i, i, P) = \min_{\lambda \in \mathbf{L}} \operatorname{Log}[(f^*)^{|P_i|}/V_{\lambda_i}(P_i)]$$

$$\Leftrightarrow \quad V_{\lambda_i}(P_i) = \max_{\lambda \in \mathbf{L}} V_\lambda(P_i)$$

$\Leftrightarrow$ λ_i is the maximum likelihood estimator of λ, deduced from sample P_i.

This definition determines uniquely $g(P)$ since, in usual conditions of regularity for density functions, and at least when f_λ is the general family of the exponential type distributions, there exists one and only one maximum likelihood estimator for. (*See* for instance Fourgeaud and Fuchs.)

3.3. Convergence of the Algorithm

THEOREM 3: *Given:*

- *a number of classes, k;*
- *a family of distribution functions* $(f_\lambda)_{\lambda \in \mathbf{L}}$.

The sequence $\mathscr{L}(L^{(n)}, P^{(n)})$ *of the products of the k likelihoods increases and converges.*

The corresponding sequence $v_n = (L^{(n)}, P^{(n)})$ *converges towards an unbiased element.*

Proof: Since R is defined as in the hypotheses of corollary 1 (*see* 2.2), the sequence $W(L^{(n)}, P^{(n)})$ decreases, and since:

$$W(L^{(n)}, P^{(n)}) = \text{Const.} - \sum_{i \in]k]} \text{Log}\, V_{\lambda_i}(n)(P_i^{(n)}) = \text{Const.} - \text{Log}\, \mathscr{L}(L^{(n)}, P^{(n)}),$$

$\mathscr{L}(L^{(n)}, P^{(n)})$ increases and converges.

The convergence of (v_n) is ensured by theorem 2.

Q. E. D.

4. PRACTICAL ASPECTS AND INTERPRETATION

4.1. Meaning of the k Classes Obtained

When the D. C. algorithm is applied to mixed distributions detection, the k classes obtained are attached to the two following constraints:

- the notion of likelihood which has been taken as a quality criterion;
- the family of probability densities which is initially chosen.

Before anything else, goodness of fit tests must be made for each class i, between the sample P_i and its computed probability density function f_{λ_i}.

4.2. The Overall Density and the Resolution of Mixtures

In the specific case where E can be supposed a representative sample arising from an underlying distribution and when model (1) is assumed, then an overall distribution on $\mathbf{R}^q$ can be deduced from E and (L, P):

$$\forall z \in \mathbf{R}^q, \quad F(z) = \sum_{i \in]k]} \Pr(z \in P_i) f_{\lambda_i}(z),$$

where the probabilities $\Pr(z \in P_i)$ are estimated by the frequencies: card $(P_i)/N$.

The D. C. algorithm then gives a solution to the mixtures resolution problems this solution must of course be checked afterwards with goodness of fit test; (for such an application of the D. C. algorithm, *see* example 4.2).

4.3. The Initial Partition

It has been seen that the algorithm needs a partition of E as a starting point. The first idea which comes to mind to build this initial partition is of course a random classification. Several numerical tests have proved that it was not convenient to do so: In fact, if the k classes $P_1^{(0)}$, $P_2^{(0)}$ and $P_k^{(0)}$ are really uniformly distributed on E, the maximum-likelihood estimators of the unknown parameters will take almost equal values; this means that

$$L^{(0)} = g(P^{(0)}) = (\lambda_1^{(0)}, \ldots, \lambda_k^{(0)})$$

will be such that: $\lambda_1^{(0)} \# \lambda_2^{(0)} \# \ldots \# \lambda_k^{(0)}$ and the new partition $P^{(1)} = f(L^{(0)})$ will be very loose; the algorithm has then difficulties to converge and takes anyway a large number of iterations.

Consequently, we have adopted a particular way of chosing $P^{(0)}$: We cut up the ranges of E in all q dimensions according to the given number k.

— If $q = 2$, $k = 3$ and given an $\alpha \in]0, 1]$, the cutting up is shown in figure 1.

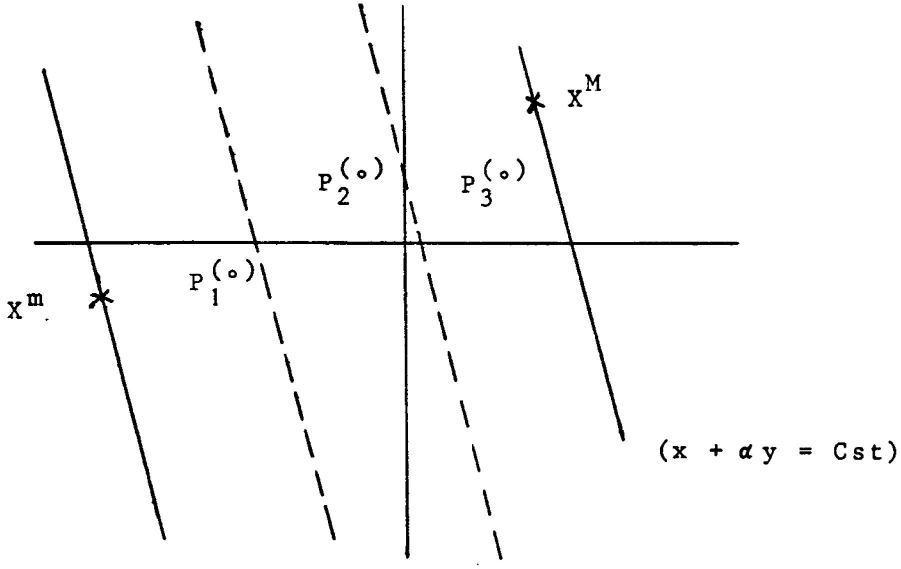

Figure 1.

$X^M \in E$ such as: $X_1^M + \alpha X_2^M = \max_{\{x \in E\}} (x_1 + \alpha x_2)$.

$X^m \in E$ such as: $X_1^m + \alpha X_2^m = \min_{\{x \in E\}} (x_1 + \alpha x_2)$.

By this mean, the space $\mathbf{R}^q$ is cut up by hyperplanes, but one can imagine many other ways to get the initial partition: cutting up for instance, by curves, or around spheres, etc.

Moreover, it is not necessary to start from a partition, the user may have rather chosen initial kernels $L^{(0)}$; $P^{(1)}$ is then taken as $f(L^{(0)})$ and the algorithm goes on as above. A way to get an initial $L^{(0)}$ is for instance to draw at random k elements of E, take them as mean vectors for the k-initial distributions and calculate the other parameters of the distributions so that the range of E in $\mathbf{R}^q$ is entirely covered.

4.4. Notion of Stable Class. The Number k

Knowing that different initial partitions lead to different unbiased elements, it seems necessary to compare the elements obtained from a given set of data with several trials with different initial partitions.

To this end, let us consider the following table (*Fig.* 2):

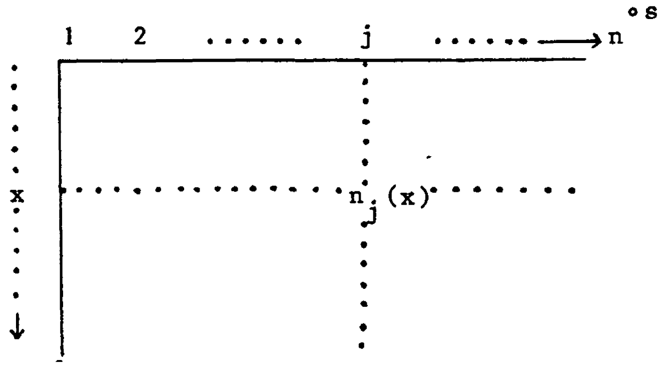

Figure 2.

$n_j(x)$ = number of the class to which x belongs in the partition obtained at the jth trial.

The elements of E are then compared with one another with the help of the following measure of dissimilarity:

$\forall x, y \in E$, $\theta(x, y)$ = number of trials in which x and y have not been classified in the same class.

In this way, x and y are near if they often belong to the same class of the obtained partitions. $\theta(x, y) = 0$ means that they always belong to the same group.

DEFINITION: A "stable class" is a subset A of E, such that:

$$\forall x, \quad y \in A \;\Leftrightarrow\; \theta(x, y) = 0.$$

The set $\mathcal{F}$ of all stable classes is nothing else than the quotient space E/R where R is the relation of equivalence:

$$x R y \;\Leftrightarrow\; \theta(x, y) = 0.$$

The properties of the stable classes are thoroughly developped in Diday ([6] to [9]) in the case of non-hierarchical clustering where they are called "strong patterns".

The example 4.4 shows how they are of use as a complement of information about the true number of classes existing in the population.

5. THE PARTICULAR CASE OF GAUSSIAN DISTRIBUTIONS

5.1. Definition of D. R, W, f and g in the Gaussian Case

Let now the family of probability densities $(f_\lambda / \lambda \in \mathbf{L})$ be the Gaussian family, i. e.:

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

where $\lambda = (\mu, V)$ with:

$$\begin{cases} \mu \in \mathbf{R}^q: & \text{mean-vector of the distribution,} \\ V: & \text{its covariance matrix } (q \times q). \end{cases}$$

— Here, $\mathbf{L} = \mathbf{R}^q \times \mathcal{E}$ [if $\mathcal{E}$ is the space of all $(q \times q)$ symmetric positive definite matrices] and we have $\mathbf{L} \subset \mathbf{R}^s$ with $s = q(q+1)$.

Now, we can define D, R, W, f and g:

- $$D(x, \lambda) = \text{Const.} + \frac{1}{2}[\text{Log} \det V + (x-\mu)' V^{-1} (x-\mu)].$$

— As V^{-1} is a positive definite and symmetric matrix, it defines on $\mathbf{R}^q$ a quadratic distance which we shall denote d_{V-1}.

$$\forall x, y \in \mathbf{R}^q; \quad d^2_{V^{-1}}(x, y) = (x-y)' V^{-1} (x-y)$$

$$\Rightarrow \quad D(x, \lambda) = \text{Const.} + \frac{1}{2}[\text{Log} \det V + d^2_{V^{-1}}(x, \mu)],$$

- $$R(\lambda, i, P) = \sum_{\{x \in P_i\}} D(x, \lambda),$$

$$R(\lambda, i, P) = \text{Const.} + \frac{1}{2}\Big[\sum_{x \in P_i} [\text{Log}\det V + d^2_{V^{-1}}(x, \mu)]\Big],$$

$\sum_{x \in P_i} d^2_{V^{-1}}(x, \mu)$ is the quadratic dispersion of the set P_i around the point μ for the metric V^{-1},

- $$W(L, P) = \text{Const.} + \sum_{i \in]k]} R(\lambda_i, i, P) \qquad \text{where} \quad \lambda_i = (\mu_i, V_i),$$

$$W(L, P) = \text{Const.} + \frac{1}{2} \sum_{i \in]k]} \sum_{x \in P_i} (\text{Log}\det V_i + d^2_{V_i^{-1}}(x, \mu_i))$$

$$= \text{Const.} + \frac{1}{2}\Big[\sum_{i \in]k]} |P_i| \,\text{Log}\det V_i + \sum_{i \in]k]} \sum_{x \in P_i} d^2_{V_i^{-1}}(x, \mu_i)\Big],$$

- $f : L \mapsto P$ with:

$$P_i = \{ x \in E / D(x, \lambda_i) \leqq D(x, \lambda_j), \forall j \neq i,$$

with x assigned to the lower index class in case of equality $\}$:

$$\forall\, i \in]k]; \quad P_i = \{ x \in E / \text{Log}\det V_i + d^2_{V_i^{-1}}(x, \mu_i) \leqq \text{Log}\det V_j + d^2_{V_j^{-1}}(x, \mu_j), \forall j \}.$$

- $g : P \mapsto L$, where, for all $i \in]k]$, μ_i and V_i are the maximum-likelihood estimates of the mean-vector and of the covariance matrix of the sample P_i.

We know these estimates are given by:

$$\mu_i = \frac{1}{|P_i|} \sum_{x \in P_i} x$$

and

$$V_i = \frac{1}{|P_i|} \sum_{x \in P_i} (x - \mu_i)(x - \mu_i)'.$$

5.2. Geometrical Interpretation

In this particular case we see that the function f reclassifies the elements of E in the following way:

The "distance" between an $x \in E$ and the ith kernel $\lambda_i = (\mu_i, V_i)$ is expressed as the sum of two quantities:

– $d^2_{V_i^{-1}}(x, \mu_i)$ = distance from x to μ_i for the metric V

and

— Log det V_i, which does not depend on x but onlyon V_i and is a characteristic feature of the dispersion of the ith distribution.

Therefore k kernels define on $\mathbf{R}^q$ k local metrics.

The unbiased element obtained at the point of convergence gives a system of k local metrics Δ_i around k different points μ_i (the mean-vectors) such that:

– Δ_i is entirely defined by μ_i and a positive definite, symmetric matrix V_i, with the distance between $x \in \mathbf{R}^q$ and μ_i in terms of Δ_i:

$$\text{Log}\det V_i + d^2_{V_i^{-1}}(x, \mu_i)$$

and:

– if $P = (P_1, \ldots, P_k)$ is the k-partition of E which is determined by the λ_i, $(i = 1, \ldots, k)$, i. e.:

$P_i = \{ x \in E/x$ is nearer in terms of Δ_i to μ_i than to any μ_j-in terms of $\Delta_j \}$, then, μ_i = mean-vector of P_i and V_i = covariance matrix of P_i.

We can then consider that our algorithm has given a solution to the following problem: To find local metrics in $\mathbf{R}^q$ that express in some way the features of E.

In fact, the countour-lines of the points that are equidistant from μ_i in terms of Δ_i are the ellipsoids of inertia of the Gaussian distribution of parameters (μ_i, V_i): our algorithm, in this case, is therefore able to detect ellipsoidal clusters.

5.3. Sebestyen's Problem (*see* Sebestyen and Romeder)

In his works on clustering and descrimination, Sebestyen has been brought to the following problem:

Knowing a finite population E of N elements, in $\mathbf{R}^q$, to find the distances d in $\mathbf{R}^q$ that minimizes the mean of the squares of the d-distances between all N points, two by two. If this mean is denoted by D^2:

$$D^2 = \frac{1}{N(N-1)} \sum_{x \in E} \sum_{y \in E} d^2(x, y).$$

In fact, he searches d in the class of the euclidean metrics that are defined by a positive definite symmetric matrix, and therefore looks for such a matrix Q which minimizes D^2:

$$D^2 = \frac{1}{N(N-1)} \sum_{x \in E} \sum_{y \in E} (x-y)' \, Q(x-y).$$

N.B.: One remembers that any positive definite and symmetric matrix Q can be written: $Q = W' \, W$, where W is triangular. Then, to assign the metric Q

to $\mathbf{R}^q$ is equivalent to assign the usual metric to the transformed of $\mathbf{R}^q$ by the application: $x \mapsto Wx$:

$$d_Q^2(x, y) = (x-y)'\,Q(x-y) = (x-y)'\,W'\,W(x-y) = (Wx-Wy)'(Wx-Wy).$$

The problem is then solved by the following *theorem:*

The metric on $\mathbf{R}^q$, *defined by* $Q = W'\,W$ *— where* W *is a linear transformation on* $\mathbf{R}^q$ *that keeps volumes constant, i. e.* $\det W = 1$ *— which minimizes* D^2 *on the finite set* E, *is:*

$$Q = (\det V)^{1/q}\,V^{-1},$$

where V *is the covariance matrix of* E.

Sebestyen's metric therefore is $d_{V^{-1}}$, with a multiplicative constant which comes from the constraint

$$Q = W'\,W \qquad \text{and} \qquad \det W = 1.$$

The deformation on E is then given by V^{-1}: the countour-lines of equidistant points around the mean vector μ of E are ellipsoids:

$$\begin{aligned} & d_Q^2(x, \mu) = \text{Const.} \\ \Leftrightarrow\quad & (\det V)^{1/q}(x-\mu)'\,V^{-1}(x-\mu) = \text{Const.} \\ \Leftrightarrow\quad & (x-\mu)'\,V^{-1}(x-\mu) = \text{Const.} \end{aligned}$$

which is the equation of an ellipsoid, the axes of which are given by V^{-1}.

Suppose now *given* a k-partition P of E, Sebestyen finds local metrics associated with the classes P_i that minimize the mean of the dispersion D_i^2 of each class.

As for the D. C. algorithm, no k-partition is initially given. Classes and local metrics are simultaneously researched and one can remark that it leads to the same ellipsoidal countour lines as Sebestyen's: the metrics differ only by constants, but not in direction. (It is natural, for instance, that there is no constraint on volumes in D. C. algorithm, since different clusters have to be compared. It is for the same reason that it needs the additive constants: Log det V_i which are associated with the dispersions of the clusters.)

On the other hand, the Sebestyen's criterion D_i^2 may be written, if

$$\mu_i = \frac{1}{N_i}\sum_{x \in P_i} x = \text{mean-vector of } P_i (N_i = \text{Card}(P_i)),$$

$$\begin{aligned}(x-y)'\,Q(x-y) &= [(x-\mu_i)-(y-\mu_i)]'\,Q[(x-\mu_i)-(y-\mu_i)] \\ &= (x-\mu_i)'\,Q(x-\mu_i)+(y-\mu_i)'\,Q(y-\mu_i)-2(x-\mu_i)'\,Q(y-\mu_i),\end{aligned}$$

$$D_i^2 = \frac{2}{N_i(N_i-1)}\Big[\sum_{x \in P_i} \sum_{y \in P_i} (x-\mu_i)' Q (x-\mu_i) - \sum_{x \in P_i} \sum_{y \in P_i} (x-\mu_i)' Q (y-\mu_i)\Big]$$

$$= \frac{2}{N_i(N_i-1)}\Big[N_i \sum_{x \in P_i} d_Q^2(x, \mu_i) - \sum_{x \in P_i} (x-\mu_i)' Q \Big(\sum_{y \in P_i} (y-\mu_i)\Big)\Big]$$

and since

$$\sum_{y \in P_i} (y-\mu_i) = 0,$$

we have:

$$D_i^2 = \frac{2}{N_i-1} \sum_{x \in P_i} d_Q^2(x, \mu_i).$$

Then, given a $P = (P_1, \ldots, P_k)$, Sebestyen minimizes the D_i^2 by choosing the optimal Q, while the D. C. algorithm finds simultaneously the k-partition P and the local metrics that tend to minimize

$$\sum_{i \in]k]} (N_i \operatorname{Log} \det V_i + \sum_{x \in P_i} d_Q^2(x, \mu_i))$$

$$= \sum_{i \in]k]} \left[N_i \operatorname{Log} \det V_i + \frac{N_i-1}{2} D_i^2 \right].$$

This paragraph leads us to remark that the local transformations we have found in the aim of maximizing local likelihoods, while searching Gaussian distributions, belong to the family of those that minimize the mean of the square of the distances within the clusters.

6. NUMERICAL EXPERIMENTS

6.1. In an univariate population drawn from three Gaussian populations, we drew:

- 50 observations from a population of parameters (*see* appendix 1) $\mu_1 = 0$ and $\sigma_1 = 1$,
- $50 \rightarrow \mu_2 = 3$ and $\sigma_2 = 2$,
- $50 \rightarrow \mu_3 = -5$ and $\sigma_3 = 2$.

Five trials have been performed; execution time for these trials on CII IRIS 80: .12 minutes.

The best trial, with the 115th, 50th, 9th observations as starting points (*cf.* 4.3), has given the following results:

Results obtained with $k = 3$ after 4 iterations		Empirical moments of the 3 drawn samples	
$\mu_1 = -\ .04$	$\sigma_1 = \ .9$	$\mu_1 = \ .1$	$\sigma_1 = 1.0$
$\mu_2 = \ 3.5$	$\sigma_2 = 1.2$	$\mu_2 = \ 3.4$	$\sigma_2 = 1.8$
$\mu_3 = -4.9$	$\sigma_3 = 1.7$	$\mu_3 = -4.9$	$\sigma_3 = 1.7$

6.2. The data: an artificial sample proposed by Duda and Hart (1973) 25 observations drawn (*see* appendix 2) from the one-dimensional two components Gaussian mixture:

$$P_1 = 1/3, \quad \mu_1 = -2, \quad \sigma_1 = 1,$$

$$P_2 = 2/3, \quad \mu_2 = +2, \quad \sigma_2 = 1.$$

– Using the D. C. algorithm, in the particular case of gaussian distributions:

Input:

$$k = 2,$$

Output:

$$P_1 = \ 8/25, \quad \mu_1 = -2.2, \quad \sigma_1 = \ .8,$$

$$P_2 = 17/25, \quad \mu_2 = \ 1.8, \quad \sigma_2 = 1.2.$$

The convergence is achieved in 2 iterations, we obtained exactly the two drawn samples, associated with their maximum-likelihood parameters with any initial drawing. (Execution time for 5 trials: .8 minutes on CII IRIS 80.)

– Using Duda and Hart method (which necessarily requires Gaussian distributions):

Input:

$$k=2, \quad \sigma_1 = \sigma_2 = 1, \quad P_1 = 1/3, \quad P_2 = 2/3;$$

Output:

$$\mu_1 = -2.1, \ \mu_2 = 1.7.$$

6.3. 150 points in $\mathbf{R}^2$ have been drawn from three two-dimensional Gaussian populations [*see fig. a* and *b*, and appendix 3] of mean-vector μ_i and covariance matrices Σ_i ($i = 1, 2, 3$).

$$\mu_1 = (0;\ 3), \qquad \mu_2 = (3;\ 0), \qquad \mu_3 = (3;\ 3)$$

and

$$\Sigma_1 = \Sigma_2 = \Sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

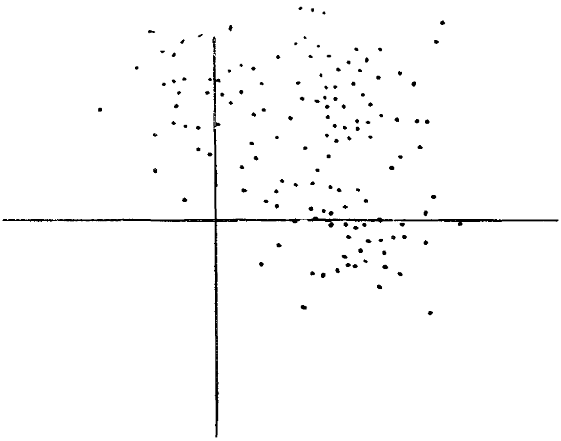

Fig. *a*. — Original data.

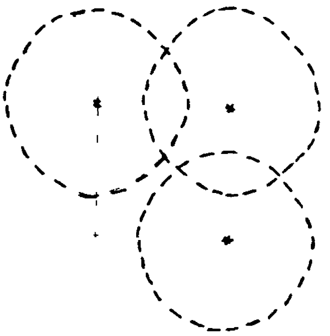

Fig. *b*. — Ellipsoids of equiprobability (95 %) of the distributions from which the original data arise.

At the point of convergence, of the best of 5 trials three classes are found after 6 iterations (*see fig. c*). The parameters of which are:

$$\mu_1 = (-.3;\ \ 2.9), \quad \Sigma_1 = \begin{pmatrix} .6 & -.2 \\ -.2 & 1.1 \end{pmatrix}, \quad |P_1| = 49,$$

$$\mu_2 = (\ 2.9;\ -.2), \quad \Sigma_2 = \begin{pmatrix} 1.0 & -.1 \\ -.1 & .8 \end{pmatrix}, \quad |P_2| = 46,$$

$$\mu_3 = (\ 2.9;\ \ 3.0), \quad \Sigma_3 = \begin{pmatrix} 1.0 & -.2 \\ -.2 & .8 \end{pmatrix}, \quad |P_3| = 55.$$

(This trial had the 53th, 87th and 41st observations as initial mean-vectors, *cf.* § 4.3.)

This example shows the efficiency of the algorithm even if the classes are not clearly separated. Here the execution time for the five trials is 48 minutes on CII IRIS 80.

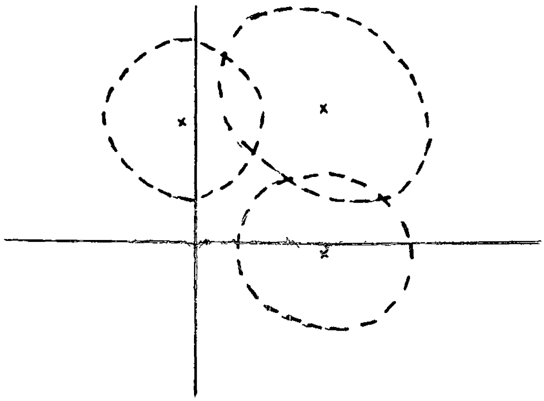

Fig. *c.* **— Ellipsoids od equiprobability (95 %) of the distributions given by the algorithm after 9 iterations.**

6.4. 150 points in $\mathbf{R}^2$ have been drawn from three two-dimensional Gaussian populations (see *fig. d* and *e*, and appendix 4):

$$\mu_1 = (0; 0), \qquad \Sigma_1 = \begin{pmatrix} 4 & \frac{2\sqrt{3}}{2} \\ \frac{2\sqrt{3}}{2} & 1 \end{pmatrix}.$$

(The principal axis of the equiprobable ellipsoids of this distribution make a $\pi/6$ angle with the 1st coordinate axis.)

$$\mu_2 = (0; 3), \qquad \Sigma_2 = \begin{pmatrix} 1/4 & 0 \\ 0 & 1/4 \end{pmatrix},$$

$$\mu_3 = (4; 3), \qquad \Sigma_3 = \begin{pmatrix} 4 & -\frac{2\sqrt{3}}{2} \\ -\frac{2\sqrt{3}}{2} & 1 \end{pmatrix} \qquad (\rightarrow \text{a } (5/6)\ \Pi \text{ angle}).$$

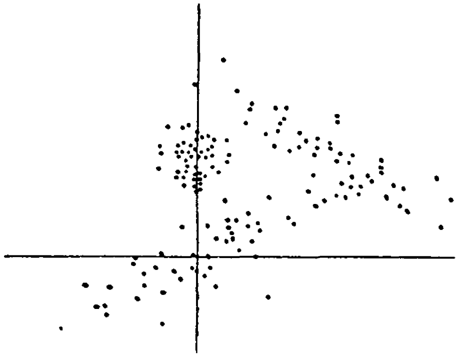

Fig. *d*. — Original data.

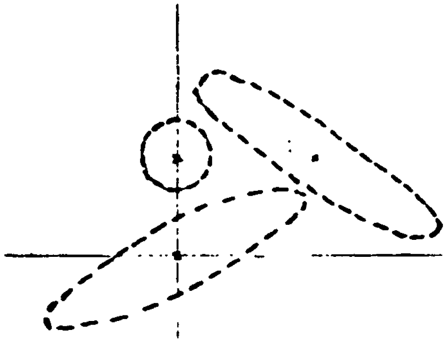

Fig. *e*. — Ellipsoids od equiprobability (95 %) of the distributions from which the original data arise.

Two trials have been done:

1) Asking for 3 classes, the algorithm has been used starting from different initial partitions and the results achieved with the best criterion value are : (see *fig. f*):

$$\mu_1 = (-.1;\ -.1), \qquad \Sigma_1 = \begin{pmatrix} 2.8 & 1.4 \\ 1.4 & 1.0 \end{pmatrix},$$

$$\mu_2 = (-.1;\ 2.9), \qquad \Sigma_2 = \begin{pmatrix} .22 & -.03 \\ -.03 & .26 \end{pmatrix},$$

$$\mu_3 = (\ 3.9;\ 3.0), \qquad \Sigma_3 = \begin{pmatrix} 3.2 & -1.8 \\ -1.8 & 1.4 \end{pmatrix}.$$

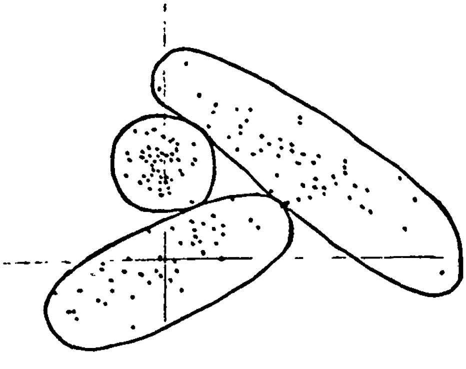

Fig. *f*. — Clusters that have been detected by the algorithm after 8 iterations, asking for three classes.

(Execution time for 5 trials: 0,58 minutes on CII IRIS 80.)

2) Asking for 4 classes, the algorithm has been applied starting from four different initial partitions. The classification obtained at the trial that gives the best criterion value is shown on figure *g*. (Two of the initial classes are recognized, the third one is cut up in two parts.)

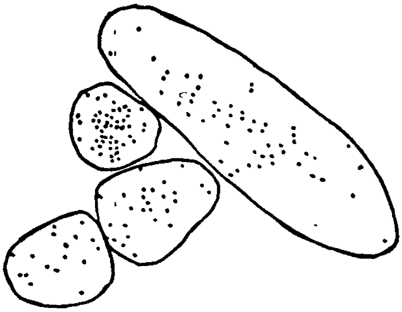

Fig. *g*. — **Clusters that have been detected by the algorithm after 8 iterations, asking for four classes.**

Comparing the results obtained from the four different initial drawing there are 23 stable classes (see *fig. h*) — sets of points that have been classified together in all the 4 trials, see 4.4 — but, by gathering these stable classes as soon as they are classed together in three of the four trials, there remains only three patterns (see *fig. i*) which are the three that have been given.

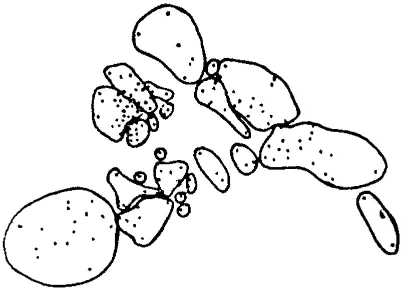

Fig. *h* — **Stable classes got after four different drawings asking for four classes.**

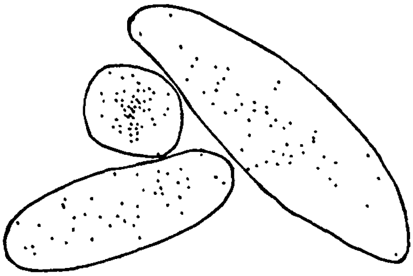

Fig. *i*. — **Sets of points that have been classified together in three of the four drawings asking for four classes, using the array of stable classes.**

This example shows how the use of several initial partitions and of stable classes can be of help when the actual number of components of the population is not known.

juin 1976.

6.5. An Application in Operating Systems Modeling

Modeling is an attempt to describe in mathematical terms a physical system (operating system, biological system...). Knowing some input parameters, the model permits to compute likely values for other parameters (output).

As soon as the considered systems becomes complex, stochastic models are not anymore the numerical values of some parameters, but their probability densities.

In operating systems modeling, theoretical results allow to use queuing networks models where service times may be assumed distributed as mixtures of gamma densities or even anyhow, in the case of approximation by a diffusion process (in this case, gaussian mixtures have been estimated).

Our application consisted in using the algorithm presented here on a sample of measures that have been picked up on a real operating system to estimate the service times distributions; these formulas may now be used in mathematical models or in simulation to generate artificial samples.

The computing aspect of the problem and all results are thoroughly described in [17].

7. EXTENSIONS

Many extensions in various directions may be considered to enlarge the algorithm field of applications.

Let us introduce those that have been recently studied.

Though this likelihood based method has proved its efficiency, we have tried to replace it in a more general context to that it could be extended other estimation methods' and to the optimization of other criteria [26].

Another step in generalization is the following: the scheme presented here consists in optimizing the criterion function at each iteration by computing a new set of kernels for the preceeding partition; the proposed generalization replaces the optimization by a plain "improving" of the chosen criterion: convergence properties may be proved under this new assumption [27] and this extension widely enlarges the possibilities of the algorithm.

For instance, it allows to optimize a likelihood criterion when dealing with distributions that do not admit maximum-likelihood estimates for their unknown parameters, such as Gamma distributions (*see* application 6.5).

8. CONCLUSION

This paper introduces the general Dynamic Clusters algorithm as a useful tool in mixed distributions detection, and presents one way among many to apply it.

The interest of the proposed method can be seen in the described experiments. Anyway, this work is only a first step in that sort of application and further research has to be made on the following points:

– the initial choice of a partition or of kernels;

– the interpretation of the stable classes; have they a probabilistic significance in the case of distributions detection?

– the problem of the number of classes—which is close to the precedent;

– the choice of other criterion functions associated with other estimation techniques for the function g;

– the use of labelled samples if they are some, which would lead to a supervised learning approach.

APPENDIX 1

50 observations from a Gaussian population of parameters $\mu_1 = 0 \quad \sigma_1 = 1$

1 7394	6194	3708	3497	1 4361	- 5312	- 1120	1088	- 3396	- 0164
- 0423	2 1062	7442	1 5797	8246	1647	5664	- 0836	-1 4007	- 8435
- 6672	1 4114	- 4637	- 0791	3015	- 0975	1285	3514	7351	4271
1785	7724	-1 7588	4954	8018	3521	- 1036	1 2092	1 1873	1 2409
- 2401	1 3913	5089	-1 4681	6116	8819	- 2125	2124	- 3618	- 2588

50 observations from a population of parameters $\mu_2 = 3 \quad \sigma_2 = 2$

2 9314	4 3458	1 6919	5 0392	5 0085	2 6981	3 1122	2 3901	2 3309	4 4738
1 8650	4 9257	7697	2863	2 7909	4 1684	3 8922	3 8375	1 9952	1 4338
2 6453	4 9680	3 9999	1 2877	5 7267	7706	5 0836	4341	5 6432	7 7662
5 8085	2 9557	1 4572	4 1418	3 7362	7 1410	1 2201	5 0516	5 3289	4 5089
9066	3 9800	6 0400	2 4195	2 1572	1 1677	1 5911	1 0378	5 9066	2 0994

50 observations from a population of parameters $\mu_3 = -5 \quad \sigma_3 = 2$

2 8243	-6 9449	6 2504	-6 9985	3 7375	-5 4383	-2 9924	-5 0097	-5 1264	-1 6716
3 8914	-4 3042	6 8031	6 0803	-3 2544	-4 8036	5 5317	-3 9576	-1 9603	4 1438
3 2204	-6 0275	-7 1818	-2 8435	4 4249	6 9575	-5 9212	2 9093	- 1652	-4 8075
-3 3577	-2 8789	-3 0543	-6 4160	-3 0076	6 3012	-6 7394	-3 7253	-5 6971	-4 6550
-4 6563	-6 0423	2 3475	3 7037	-5 0953	-1 2384	-5 5724	2 2887	3 5806	-6 8857

juin 1976.

APPENDIX 2

25 observations drawn from the one-dimensional two components Gaussian mixture [*see* Duda and Hart (1973)].

$$p(x/\mu_1, \mu_2) = \frac{1}{3\sqrt{2\pi}}\exp\left[-\frac{1}{2}(x-\mu_1)^2\right] + \frac{2}{3\sqrt{2\pi}}\exp\left[-\frac{1}{2}(x-\mu_2)^2\right],$$

with $\mu_1 = -2$ and $\mu_2 = +2$.

k	x_k	(Class)
1	0.608	2
2	−1.590	1
3	0.235	2
4	3.949	2
5	−2.249	1
6	2.704	2
7	−2.473	1
8	0.672	2
9	0.262	2
10	1.072	2
11	−1.773	1
12	0.537	2
13	3.240	2
14	2.400	2
15	−2.499	1
16	2.608	2
17	−3.458	1
18	0.257	2
19	2.569	2
20	1.415	2
21	1.410	2
22	−2.653	1
23	1.396	2
24	3.286	2
25	−0.712	1

APPENDIX 3

150 observations from a Gaussian population of parameters:

$$\mu_1 = \begin{pmatrix} 0 \\ 3 \end{pmatrix} \Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \mu_2 = \begin{pmatrix} 3 \\ 0 \end{pmatrix} \quad \Sigma_2 = \Sigma_1 \quad \mu_3 = \begin{pmatrix} 3 \\ 3 \end{pmatrix} \quad \Sigma_3 = \Sigma_1$$

1	1.7675	2.8858	51	2.5946	- .8165	101	3.2078	3.6286
2	- .5153	4.6630	52	3.1067	.1518	102	2.4902	2.8357
3	-1.9992	2.0056	53	3.2886	.2590	103	2.0712	3.3565
4	- .3574	4.0662	54	2.7717	- .8121	104	2.0150	2.6180
5	- .1515	3.3470	55	4.0327	1.7961	105	1.4491	2.4994
6	- .6922	3.4861	56	2.2511	- .9148	106	3.5598	1.4344
7	- .7895	2.7930	57	3.2123	- .6538	107	3.3170	4.1116
8	- .5074	3.3836	58	4.0135	- .6892	108	1.8639	2.7601
9	- .9394	3.1642	59	1.1704	.7486	109	2.3297	3.5567
10	.9305	1.5333	60	1.9004	1.9004	110	5.2037	2.4988
11	.0375	1.7220	61	1.8690	1.4305	111	2.2544	3.9827
12	- .1495	2.5318	62	2.1106	- .6687	112	4.6936	2.4067
13	-1.2349	2.6928	63	2.8425	.1129	113	2.8717	2.5960
14	1.9367	3.3711	64	1.0593	-1.3042	114	3.9879	4.9157
15	-2.2663	3.9912	65	4.7734	2.1586	115	2.0817	3.1297
16	- .0277	3.6071	66	4.1071	- .3107	116	1.5994	4.5373
17	- .7697	3.7215	67	.6815	.7086	117	3.8605	3.0562
18	- .3691	2.8654	68	4.1252	.0472	118	2.7686	2.5011
19	.7129	2.6985	69	4.6177	- .0938	119	2.8671	2.5013
20	.5994	1.4029	70	4.5792	- .9878	120	2.2855	3.4976
21	.1799	2.1303	71	2.9162	- .0828	121	1.9090	2.4740
22	.6849	3.1563	72	4.2844	.3939	122	3.8850	2.3661
23	.4901	4.7646	73	2.4606	1.1082	123	1.1284	3.9302
24	-1.2234	2.1811	74	2.0768	- .3467	124	3.8058	3.2864
25	- .7513	3.2052	75	2.0768	- .3467	125	2.6790	2.3463
26	- .4974	3.6017	76	3.6275	- .0208	126	1.8218	2.5004
27	.7771	1.7630	77	3.0737	- .0049	127	1.8199	2.8859
28	.1396	2.1626	78	3.7948	.1579	128	1.5227	.8116
29	- .1566	4.1088	79	3.1061	-1.0087	129	4.7574	1.8967
30	- .1958	3.1893	80	2.4829	-2.4738	130	3.8401	1.0753
31	.2346	2.1566	81	3.9428	- .7644	131	3.2242	2.3822
32	- .8296	3.2361	82	2.3104	1.6781	132	2.7846	2.6149
33	-1.0896	1.0071	83	3.3773	.9482	133	3.6897	2.2499
34	- .0711	2.9173	84	3.4703	.5865	134	4.0765	1.9652
35	.3805	1.4403	85	2.4265	.9851	135	2.2519	4.5419
36	- .0777	5.3859	86	4.3260	-1.3684	136	3.8229	3.5649
37	1.1096	2.3529	87	3.1177	1.9240	137	3.6697	2.5119
38	-1.6433	3.6442	88	3.7716	-1.1405	138	1.6122	1.9875
39	.1539	4.6894	89	1.5706	.8411	139	4.6461	3.3177
40	-1.2868	2.3383	90	3.4790	-1.6889	140	2.3666	4.0192
41	-2.1055	4.8253	91	2.7388	.2964	141	3.3848	3.2558
42	-1.0522	2.9071	92	3.1213	.9787	142	3.0094	4.3619
43	- .3636	4.0151	93	3.0786	- .7954	143	3.5936	2.8688
44	.2877	2.9269	94	.3801	- .5808	144	4.4768	4.9558
45	- .0014	2.4252	95	1.5727	- .3265	145	3.5181	2.9150
46	.4027	1.2087	96	4.0155	1.2681	146	2.8177	1.8883
47	- .5714	1.4259	97	2.9389	.5474	147	3.2431	3.0944
48	-1.0525	3.6767	98	1.4941	- .1286	148	2.0992	4.5718
49	.8277	2.2273	99	3.5144	- .6985	149	2.4069	2.5815
50	.4388	3.2734	100	3.6396	-2.0335	150	4.5484	2.3426

APPENDIX 4

150 observations from a Gaussian population or parameters:

$$\mu_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \Sigma_1 = \begin{pmatrix} 4 & \frac{2\sqrt{3}}{2} \\ \frac{2\sqrt{3}}{2} & 1 \end{pmatrix} \qquad \mu_2 = \begin{pmatrix} 0 \\ 3 \end{pmatrix} \Sigma_2 = \begin{pmatrix} \frac{1}{4} & 0 \\ 0 & \frac{1}{4} \end{pmatrix}$$

$$\mu_3 = \begin{pmatrix} 4 \\ 3 \end{pmatrix} \Sigma_3 = \begin{pmatrix} 4 & -\frac{2\sqrt{3}}{2} \\ -\frac{2\sqrt{3}}{2} & 1 \end{pmatrix}$$

1	88	2 85	51	1 61	46	101	4 74	2 15
2	26	2 62	52	1 02	68	102	5 69	2 20
3	-1 00	3 05	53	55	47	103	3 26	2 55
4	18	2 72	54	1 25	1 25	104	2 90	3 68
5	- 18	3 34	55	-3 18	80	105	4 07	3 09
6	- 35	2 56	56	44	- 27	106	68	5 86
7	- 39	3 84	57	04	- 35	107	3 97	2 91
8	25	2 94	58	2 71	1 17	108	4 34	1 77
9	- 47	2 62	59	1 20	26	109	3 38	3 31
10	46	2 71	60	- 86	-1 23	110	3 81	3 45
11	02	3 21	61	1 83	- 23	111	2 04	3 72
12	- 07	2 39	62	3 69	1 75	112	1 61	4 58
13	- 62	2 42	63	1 50	93	113	3 35	3 14
14	97	3 02	64	24	- 61	114	7 23	78
15	1 13	3 31	65	82	1 72	115	5 02	2 40
16	01	3 70	66	32	01	116	5 46	1 79
17	38	2 44	67	1 19	1 10	117	4 35	2 42
18	19	3 29	68	77	47	118	4 38	2 11
19	36	3 29	69	66	- 44	119	5 07	2 30
20	30	3 63	70	-2 22	-1 85	120	4 55	2 42
21	09	3 15	71	1 80	99	121	82	3 44
22	34	3 22	72	- 15	14	122	5 30	2 61
23	25	2 96	73	- 99	07	123	6 15	1 44
24	- 61	3 31	74	-2 12	1 16	124	2 51	4 48
25	- 38	2 24	75	2 59	1 70	125	5 90	2 09
26	- 25	3 12	76	1 02	58	126	2 82	3 28
27	39	3 05	77	-2 68	-1 55	127	2 33	3 93
28	07	2 23	78	-2 76	-1 50	128	3 14	3 43
29	08	2 38	79	2 82	1 05	129	3 98	4 12
30	10	2 74	80	1 04	-2 03	130	3 40	3 63
31	12	2 51	81	3 53	1 60	131	2 32	3 78
32	- 42	3 44	82	-1 84	- 09	132	1 14	5 02
33	55	3 02	83	-2 48	- 88	133	2 16	4 50
34	04	2 18	84	92	87	134	5 31	2 71
35	19	2 43	85	2 08	1 77	135	2 73	3 19
36	- 04	3 49	86	4 71	1 93	136	5 35	2 77
37	56	2 56	87	- 41	87	137	5 93	1 47
38	- 82	3 92	88	-2 22	1 94	138	1 51	4 47
39	08	2 00	89	1 46	1 33	139	6 86	2 44
40	- 54	3 18	90	1 74	03	140	4 16	2 22
41	-1 05	2 61	91	3 25	2 05	141	6 02	1 77
42	- 53	3 01	92	88	1 08	142	8 22	59
43	- 18	3 08	93	43	- 71	143	4 01	4 18
44	14	2 37	94	-1 67	-1 30	144	7 33	1 84
45	00	2 01	95	29	90	145	2 41	4 14
46	20	3 24	96	29	90	146	- 33	5 23
47	- 29	3 81	97	-1 50	- 55	147	2 22	3 28
48	53	3 22	98	-1 48	93	148	3 83	3 35
49	41	3 51	99	- 06	44	149	1 42	4 04
50	22	3 58	100	50	88	150	4 45	3 00

REFERENCES

1. A. K. Agrawala, *Learning with a Probabilistic Teacher* I.E.E.E. *Transactions* on Information Theory, IT-16, No. 4, May 1970.
2. G. H. Ball, D. J. Hall, Isodata : *A Novel Method of Data Analysis and Pattern Classification*, Technical Report, SRI Project 5533, Stanford Research Institute, Menlo Park, California U.S.A., 1965.
3. C. G. Bhattacharya, *A Simple Method of Resolution of a Distribution into Gaussian Components*, Biometrics, March 1967.
4. D. B. Cooper and P. W. Cooper, *Non Supervised Adaptive Signal Detection and Pattern Recognition*, Information and Control, 7, 1964, p. 416.
5. P. W. Cooper, *Some Topics on Non Supervised Adaptive Detection for Multivariate Normal Distributions*, Computer and Information Sciences, 11, 1967.
6. N. E. Day, *Estimating the Components of a Mixture of Normal Distributions*, Biometrika, 56, 3, 1969, p. 463.
7. E. Diday, *Optimisation en classification automatique et Reconnaissance des formes*, R.A.I.R.O., V-3, November 1972, p. 61.
8. E. Diday, *The Dynamic Clusters Method in Non-Hierarchical Clustering*, International Journal of Computer and Information Sciences, 2, No. 1, 1973.
9. E. Diday, *Une nouvelle méthode de classification automatique et reconnaissance des formes: La méthode des Nuées Dynamiques*, Revue de Statistique Appliquée, XIX, 2, 1970.
10. E. Diday, *Nouvelles méthodes et nouveaux concepts en classification automatique et reconnaissance des formes*, Thèse de Doctorat d'État ès-sciences Mathématiques, 1972.
11. Dorofeyuk, *Automatic Classification Algorithms*, (Review) Automation and Remote Control, 32, No. 12 part 1, December 1971, p. 1928.
12. R. O. Duda and R. E. Hart, *Pattern Classification and Scene Analysis*, Wiley, New York, 1973.
13. C. Fourgeaud and A. Fuchs, *Statistique*, Dunod, Paris, 1971.
14. E. Gelenbe, J. C. A. Boekhorst and J. L. W. Kessels, *Minimizing Wasted Space in Partitioned Segmentation*, Communications of the A.C.M., 16, No. 6, June 1973.
15. W. D. Gregg and J. C. Hancock, *An Optimum Decision, Directed Scheme for Gaussian Mixtures*, I.E.E.E. Transactions on Information Theory, IT-14, No. 3, May 1968.
16. Y. C. Ho and A. K. Agrawala, *On Pattern Classification Algorithms;* Introduction and Survey, Proceedings of the I.E.E.E., 56, No. 12, December 1972.
17. J. Leroudier and A. Schroeder, *A Statistical Approach to the Estimation of Service Times Distributions for Operating Systems Modeling*, E. Gelenbe, D. Potier (eds), International Computing Symposium 1975, North-Holland publ. C°, June 1975.
18. J. MacQueen, *Some Methods for Classification and Analysis of Multivariate Observations*, The 5th Berkley Symposium on Mathematics, Statistics and Probability, 1, No. 1, 1967.
19. E. A. Patrick and F. P. Fischer, *Cluster Mapping with Experimental Computer Graphics*, I.E.E.E. Trans. On Computers, c 18, No. 11, november 1969.
20. E. A. Patrick and J. C. Hancock, *Non Supervised Sequential Classification and Recognition of Patterns*, I.E.E.E. Transactions on Information Theory, IT-12, No. 3, July 1966.

juin 1976.

21. E. A. Patrick and J. P. Costello, *On Unsupervised Estimation Algorithms*, I.E.E.E. Transactions on Information Theory, IT-16, No. 5; September 1970.

22. E. A. Patrick, *Fundamentals of Pattern Recognition*, Prentice Hall Inc. NJ, 1972.

23. K. Pearson, *Contributions to the Mathematic Theory of Evolution*, Philos Trans. Soc., No. 185, 1894.

24. C. R. Rao, *Utilization of Multiple Measurements in Problems of Biological Classification*, Journal of the Royal Statistical Society, Series B, X, No. 2, 1948.

25. J. M. Romeder, *Méthodes de discrimination*, Thèse de 3^{e} cycle, Laboratoire de Statistique Mathématique, Université Paris VI, 1969.

26. A. Schroeder, *Reconnaissance des composants d'un mélange*, Thèse 3^{e} cycle, Université Paris VI, 1974.

27. A. Schroeder, *Analyse d'un mélange de distributions de probabilité de même type*, Rapport de Recherche, n^{o} 104, I.R.I.A., Laboria, 1975.

28. G. S. Sebestyen, *Decision-Making Process in Pattern Recognition*, A.C.M. Monograph series, MacMillan, New York, 1962.

29. J. S. Spragins, *Learning Without a Teacher*, I.E.E.E. Transactions on Information Theory, IT-12, No. 2, April 1966.

30. D. Wishart, *Some Problems in the Theory and Application of the Methods of Numerical Taxonomy*, Ph. D. Thesis University of St-Andrews, 1971.

31. T. Y. Young and G. Coraluppi, *Stochastic Estimation of a Mixture of Normal Density Functions Using an Information Criterion*, I.E.E.E. Transactions on Information Theory, IT-16, No. 3.