A NEW APPROACH IN MIXED DISTRIBUTIONS DETECTION (*)

by E. DIDAY et A. SCHROEDER (†)

Abstract. — Our aim is to detect in a given multivariate sample the possible presence of 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:

$$F(x) = \sum_{1 \leq j \leq k} P_j F_j(x),$$  \hspace{1cm} \text{(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 unknown 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:

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

unknown, but $\varphi$ 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 $\chi^2$, 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

*Revue Française d'Automatique, Informatique et Recherche Opérationnelle*
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 approxi-
mation 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 obser-
vations 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.

juin 1976.
Several practical remarks will then be made, essentially concerning the number of components. Afterwards, the particular case of mixed gaussian distributions will be developed, 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 \( \mathbb{R}^n \);
- \( P_k \): the set of all partitions of \( E \) into \( k \) classes; the elements of \( P_k \) will be called: \( k \)-partitions.

\[ P \in P_k \iff P = (P_1, \ldots, P_k); \]

- \( L \): 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;
- \( L_k \): the set of all \( k \)-tuples of \( L \):

\[ L \in L_k \iff L = (\lambda_1, \lambda_2, \ldots, \lambda_k) \text{ where } \lambda_i \in L, \quad \forall i \in [k] \]

\[ (\text{if } [k] = \{1, 2, \ldots, k\}). \]

The D. C. algorithm solves the following problem: To find a couple \((P, L)\) where \( P \in P_k \) and \( L \in L_k \) that minimizes some criterion function, which will be denoted \( W: L_k \times P_k \rightarrow \mathbb{R}^+ \). 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 \( L_k \) an element of \( P_k \), and from this element of \( P_k \) an element in \( L_k \), so that the values of the criterion function on successive couples \((L, P)\) decrease. To be precise, we shall need some more notation:

\[ D : \quad E \times L \rightarrow \mathbb{R}^+, \]
\[ (x, \lambda) \mapsto D(x, \lambda) \]

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

\[ R : \quad L \times [k] \times P_k \rightarrow \mathbb{R}^+, \]
\[ (\lambda, i, P) \mapsto R(\lambda, i, P), \]

will give a measure of the goodness of fit of kernel \( \lambda \) with the \( i \)th 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:

$$f : L_k \rightarrow P_k,$$

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

where:

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

(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 $\lambda_i$ than to any other kernel of $L$.

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

$$g : P_k \rightarrow L_k,$$

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

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

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

In other words, among all possible kernels, $\lambda_i$ 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 $P \in P_k$, $g(P)$ is a well determined element of $L_k$.]

The criterion to minimize will be defined as:

$$W : L_k \times P_k \rightarrow \mathbb{R}^+, \quad v = (L, P) \Rightarrow W(v) = \sum_{i \in [k]} R(\lambda_i, i, P).$$

The problem is now an optimization problem:

Minimize $W(v)$ for all $v \in L_k \times P_k$. \hfill (2)

2.2. The Algorithm

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

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

juin 1976.
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 $\mathbb{R}^+$, it converges, and we shall see (th. 2) that its limit is attained:

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

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:

$$W(v^*) \leq W(v) \quad \text{for all } v \in L_k \times P_k.$$

For a $v^*$ given by the algorithm, this inequality only holds for $v$ in a part of $L_k \times P_k$, 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. —

**Définition 1**: A function $R : L \times L_k \times P_k \rightarrow \mathbb{R}^+$ is said SEMI-SQUARE, if:

(1) $\forall L \in L_k, \quad W(g \circ f(L), f \circ g \circ f(L)) \leq W(g \circ f(L), f(L))$

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

(1) can be written equivalently:

(1') $\forall P \in f(L_k), \quad W(g(P), f \circ g(P)) \leq W(g(P), P)$.

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

Proof: We shall prove the following inequality:

$$u_{n+1} \leq u_n, \quad \forall n,$$

in two steps:

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

\[ b) \quad W(L^{(n)}, P^{(n+1)}) \leq W(L^{(n)}, P^{(n)}) = u_n, \]
\[ a) \quad u_{n+1} = \sum_{i \in [k]} R(\lambda^{(n+1)}_i, i, P^{(n+1)}) \]

and

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

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

\[ R(\lambda^{(n+1)}_i, i, P^{(n+1)}) = \min_{\lambda \in L} R(\lambda, i, P^{(n+1)}) \]

we have in particular:

\[ \forall i \in [k], \quad R(\lambda^{(n+1)}_i, i, P^{(n+1)}) \leq R(\lambda^{(n)}_i, i, P^{(n+1)}) \]

and therefore, \( a) \) is proved.

\[ b) \quad \text{As soon as } n \geq 1, \quad P^{(n)} = f(L^{(n-1)}) \text{ and then } P^{(n)} \in f(L_k). \text{ Since } R \text{ is assumed to be semi-square, by property (1') of definition 1 applied to } P^{(n)}: \]

\[ W(g(P^{(n)}), f \circ g(P^{(n)})) \leq W(g(P^{(n)}), P^{(n)}) \]

and as:

\[ g(P^{(n)}) = L^{(n)} \quad \text{and} \quad f \circ g(P^{(n)}) = f(L^{(n)}) = P^{(n+1)}, \]

we have:

\[ W(L^{(n)}, P^{(n+1)}) \leq W(L^{(n)}, P^{(n)}), \]

which proves \( b)\).

Then, \( (u_n) \) decreases and as it is a sequence in \( \mathbb{R}^+ \), 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 \( u_{n+1} \leq u_n \) 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 \((3)\).

\[ \forall L \in L_k \quad \forall P \in P_k \quad \left\{ \begin{array}{l} W(L, f(L)) \leq W(L, P). \\
(2) \end{array} \right. \]

\[ \forall L, M \in L_k; \quad W(L, f(M)) \leq W(M, f(M)) \]
\[ \Rightarrow W(L, f(L)) \leq W(L, f(M)). \quad (3) \]

juin 1976.
If (3) is true $R$ is said to be SQUARE.

Proof: One can see easily that (2) $\Rightarrow$ (1) and (2) $\Rightarrow$ (3).

Let us prove that (3) $\Rightarrow$ (1):

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

$$(3) \Rightarrow \forall M \in \mathbb{L}_k, \quad W(g \circ f(M), f \circ g \circ f(M)) \leq W(M, f(M))$$

$$\Rightarrow \quad W(g \circ f(M), f \circ g \circ f(M)) \leq W(g \circ f(M), f(M)).$$

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

$$\forall M \in \mathbb{L}_k, \quad \forall P \in \mathbb{P}_k, \quad W(g(P), P) \leq W(M, P).$$

[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:

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

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

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

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

$$W(L, P) = \sum_{j \in [k]} \sum_{x \in P_j} D(x, \lambda_j)$$

$$= \sum_{x \in E} \sum_{j \in [k]} D(x, \lambda_j) \delta_{P_j}(x)$$

[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:

$$W(L, f(L)) = \sum_{i \in [k]} R(\lambda_i, i, Q) = \sum_{i \in [k]} \sum_{x \in Q_i} D(x, \lambda_i)$$

$$= \sum_{x \in E} \sum_{i \in [k]} D(x, \lambda_i) \delta_{Q_i}(x).$$

*Revue Française d'Automatique, Informatique et Recherche Opérationnelle*
As $Q = f(L)$, and by definition of $f$, we know that $x \in Q$, if

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

for all $j$, which implies, for all $L \in L_k$ and $P \in P_k$: $W(L, f(L)) \leq W(L, P)$.

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).

**Définition:** An element $v = (L, P) \in L_k \times P_k$ is said to be **UNBIASED** for the functions $f$ and $g$, if $P = f \circ g (P)$ and $L = g \circ f (L)$.

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

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

**Theorem 2:** If $R$ has the two following properties:

(i) semi-square,

(ii) $\forall P \in f(L_k), \forall i \in [k], \ R(\lambda, i, P)$ is minimum for a unique $\lambda_0$. Then, $(v_n)$ is convergent and its limit is an unbiased element.

**Proof:** Since $E$ is assumed to be a finite set, $P_k$ is finite too. By definition, $E$ is such that, for any $P \in P_k$, $g (P)$ can only take one value in $L_k$, then $g (P_k) \subset L_k$ 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:

$$\exists M, \ \forall n \geq M, \ u_n = u^*.$$

Thus, $\forall n \geq M, \ W(L^{(n)}, P^{(n)}) = W(L^{(n+1)}, P^{(n+1)})$ 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:

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

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

juin 1976.
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:

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

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:

\[ \forall n \geq M+1, \quad v_n = v_{n+1} = v^*, \]

\[ v^* = (L^*, P^*) = (L^{(n)}, P^{(n)}) = (g(P^{(n)}), P^{(n)}) \]

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

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 \( \mathbb{R}^q \).

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

We want then to find a couple \( (L, P) \), where \( L = (\lambda_1, \ldots, \lambda_k), \lambda_i \in \mathbb{L} \) 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 \( f_{\lambda_i} \).

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

\[ \mathcal{L}(L^*, P^*) = \max_{L \in \mathbb{L}_k} \prod_{i \in [k]} \prod_{x \in P_i} f_{\lambda_i}(x). \]

Revue Française d'Automatique, Informatique et Recherche Opérationnelle
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:
\[ W(L, P) = K - \sum_{i \in [k]} \log V_{\lambda_i}(P_i) \]
(where \(K\) is a constant).

3.2. The Algorithm

Let us take:
- \(E \subseteq \mathbb{R}^t\). The finite set to classify.
- \(L \subseteq \mathbb{R}^s\). Set of the kernels (which will be exactly the set of parameters introduced above).
- \(D : E \times L \to \mathbb{R}^+\),
\[ (x, \lambda) \mapsto D(x, \lambda) = \log(f^*/f_{\lambda}(x)), \]
where \(f^* \geq \max \{ f_{\lambda}(x) | \lambda \in L, x \in E \}\); it is sufficient to know such an \(f\) exists to use the algorithm; for instance, for univariate Gaussian distributions with \(\lambda = (\mu, \sigma)\):
\[ \max_{x \in \mathbb{R}} f_{\lambda}(x) = (2\pi)^{-1/2} \sigma^{-2} \]
and then a possible \(f^*\) is:
\[ f = (2\pi)^{-1/2} d_0^{-1} \quad \text{if} \quad d_0 = \min_{x \in E} |x - y|^2. \]

This definition for \(D\) expresses that the greater \(f_{\lambda}(x)\) is, the nearer to kernel \(\lambda\) the observation \(x\) is. (It can also be said that the likelihood of the sample \(\{x\}\) is large for \(f_{\lambda}\).
- \(R\) is then defined from \(D\):
\[ R : L \times [k] \times P_k \to \mathbb{R}^+, \]
\[ R(\lambda, i, P) = \sum_{x \in P_i} D(x, \lambda), \]
\[ R(\lambda, i, P) = \sum_{x \in P_i} \log(f^*/f_{\lambda}(x)) \]
\[ = \log[\prod_{x \in P_i} f_{\lambda}(x)](\text{where} \ |P_i| = \text{card} P_i) \]
\[ = \log[\prod_{x \in P_i} V_{\lambda}(P_i)]. \]

\(^{(1)}\) The same problem may of course be formalized in other terms.

juin 1976.
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 \(L = (\lambda_1, \ldots, \lambda_k)\)),

\[ W(L, P) = \sum_{i \in [k]} \log \left( \frac{(f^*)^{P_i} | V_{\lambda_i}(P_i)}{V_{\lambda_i}(P_i)} \right), \]

\[ W(L, P) = \log (f^*)^N - \sum_{i \in [k]} \log V_{\lambda_i}(P_i) \] (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

\[ P_i = \{ x \in E \mid D(x, \lambda_i) \leq D(x, \lambda_j), \forall j \neq i \}, \]

\[ P_i = \{ x \in E \mid f_{\lambda_i}(x) \geq f_{\lambda_j}(x), \forall j \neq i \} \]

(\(x\) 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 \(\lambda_i\) is such that:

\[ R(\lambda_i, i, P) = \min_{\lambda \in \mathbf{L}} R(\lambda, i, P) \]
\[ \Leftrightarrow \]
\[ R(\lambda_i, i, P) = \min_{\lambda \in \mathbf{L}} \log \left( \frac{(f^*)^{P_i} | V_{\lambda_i}(P_i)}{V_{\lambda_i}(P_i)} \right) \]
\[ \Leftrightarrow \]
\[ V_{\lambda_i}(P_i) = \max_{\lambda \in \mathbf{L}} V_{\lambda}(P_i) \]
\[ \Rightarrow \lambda_i \text{ is the maximum likelihood estimator of } \lambda, \text{ deduced from sample } P_i. \]

This definition determines uniquely \(g(P)\) since, in usual conditions of regularity for density functions, and at least when \(f_i^*\) 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 \( (f_i)_{i \in L} \).

The sequence \( \mathcal{L} (L^{(n)}, P^{(n)}) \) 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:

\[
W(L^{(n)}, P^{(n)}) = \text{Const.} - \sum_{i \in [k]} \log V_i(n) (P_i^{(n)}) = \text{Const.} - \log \mathcal{L} (L^{(n)}, P^{(n)}),
\]

\( \mathcal{L} (L^{(n)}, P^{(n)}) \) 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 \( f_{L_i} \).

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 \( \mathbb{R}^q \) can be deduced from \( E \) and \( (L, P) \):

\[
\forall z \in \mathbb{R}^q, \quad F(z) = \sum_{i \in [k]} \Pr(z \in P_i)f_{L_i}(z),
\]

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

juin 1976.
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: $\lambda_{1}^{(0)} \neq \lambda_{2}^{(0)} \neq \ldots \neq \lambda_{k}^{(0)}$ 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 choosing $P^{(0)}$: 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](image-url)

**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 $\mathbb{R}^q$ 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 $\mathbb{R}^q$ 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):

\begin{figure}[h]
\centering
\begin{tabular}{cccccc}
1 & 2 & \cdots & j & \cdots & n \\
\vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\
\vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\
\vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\
x & \cdots & \cdots & \cdots & \cdots & \cdots \\
\vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\
\downarrow & \cdots & \cdots & \cdots & \cdots & \cdots \\
\end{tabular}
\caption{Figure 2.}
\end{figure}

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

juin 1976.
DEFINITION: A "stable class" is a subset $A$ of $E$, such that:

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

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

$$x \mathcal{R} y \quad \iff \quad \theta(x, y) = 0.$$ 

The properties of the stable classes are thoroughly developed 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 $(f_\lambda | \lambda \in L)$ be the Gaussian family, i.e.:

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

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

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

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

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

$\bullet$  

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

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

$$\forall x, y \in \mathbb{R}^q; \ d_{V^{-1}}^2(x, y) = (x-y)' V^{-1} (x-y)$$

$$\Rightarrow \quad D(x, \lambda) = \text{Const.} + \frac{1}{2} \left( \log \det V+d_{V^{-1}}^2(x, \mu) \right),$$

Revue Française d'Automatique, Informatique et Recherche Opérationnelle
A NEW APPROACH IN MIXED DISTRIBUTIONS DETECTION

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

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

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

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

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

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

\[ f : L \mapsto P \] with:

\[ P_i = \{ x \in E / D(x, \lambda_i) \leq D(x, \lambda_j), \forall j \neq i, \] with \( x \) assigned to the lower index class in case of equality \};

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

\[ g : P \mapsto L, \] where, for all \( i \in [k] \), \( \mu_i \) 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 \( i \)th kernel \( \lambda_i = (\mu_i, V_i) \) is expressed as the sum of two quantities:

- \( d^2_{\lambda^{-1}}(x, \mu_i) \) = distance from \( x \) to \( \mu_i \) for the metric \( V \)
and

- \( \log \det V_i \), which does not depend on \( x \) but only on \( V_i \) and is a characteristic feature of the dispersion of the \( i \)th distribution.
Therefore \( k \) kernels define on \( \mathbb{R}^q \) \( k \) local metrics.

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

- \( \Delta_i \) is entirely defined by \( \mu_i \) and a positive definite, symmetric matrix \( V_i \), with the distance between \( x \in \mathbb{R}^q \) and \( \mu_i \) in terms of \( \Delta_i \):

\[
\log \det V_i + d_{\Delta_i}^2(x, \mu_i)
\]

and:

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

\[
P_i = \{ x \in E | x \text{ is nearer in terms of } \Delta_i \text{ to } \mu_i \text{ than to any } \mu_i \text{-in terms of } \Delta_j \},
\]

then, \( \mu_i = \text{mean-vector of } P_i \) and \( V_i = \text{covariance matrix of } P_i \).

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

In fact, the contour-lines of the points that are equidistant from \( \mu_i \) in terms of \( \Delta_i \) are the ellipsoids of inertia of the Gaussian distribution of parameters \( (\mu_i, V_i) \); 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 discrimination, Sebestyen has been brought to the following problem:

Knowing a finite population \( E \) of \( N \) elements, in \( \mathbb{R}^q \), to find the distances \( d \) in \( \mathbb{R}^q \) 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 \( Q \)

*Revue Française d'Automatique, Informatique et Recherche Opérationnelle*
to $\mathbb{R}^q$ is equivalent to assign the usual metric to the transformed of $\mathbb{R}^q$ 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 $\mathbb{R}^q$, defined by $Q = W' W$—where $W$ is a linear transformation on $\mathbb{R}^q$ 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

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

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

$$d_Q^2(x, \mu) = \text{Const.}$$

$$\Leftrightarrow (\det V)^{1/q} (x - \mu)' V^{-1} (x - \mu) = \text{Const.}$$

$$\Leftrightarrow (x - \mu)' V^{-1} (x - \mu) = \text{Const.}$$

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 contour 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 the D. C. algorithm, since different clusters have to be compared. It is for the same reason that it needs the additive constants: $\log \det V$, which are associated with the dispersions of the clusters.)

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

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

$$Q(x - y) = [(x - \mu_i) - (y - \mu_i)]' Q [(x - \mu_i) - (y - \mu_i)]$$

$$= (x - \mu_i)' Q(x - \mu_i) + (y - \mu_i)' Q(y - \mu_i) - 2 (x - \mu_i)' Q(y - \mu_i),$$

juin 1976.
\[ D_i^2 = \frac{2}{N_i(N_i-1)} \left[ \sum_{x \in P_i} \sum_{y \in P_i} (x - \mu_i)' Q(x - \mu_i) - \sum_{x \in P_i} \sum_{y \in P_i} (x - \mu_i)' Q(y - \mu_i) \right] \]
\[ = \frac{2}{N_i(N_i-1)} \left[ N_i \sum_{x \in P_i} d_Q^2(x, \mu_i) - \sum_{x \in P_i} (x - \mu_i)' Q(\sum_{y \in P_i} (y - \mu_i)) \right] \]

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
\[ \sum_{i \in [k]} (N_i \text{Log det } V_i + \sum_{x \in P_i} d_Q^2(x, \mu_i)) \]
\[ = \sum_{i \in [k]} \left[ N_i \text{Log det } V_i + \frac{N_i-1}{2} D_i^2 \right]. \]

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:

<table>
<thead>
<tr>
<th>Results obtained with $k = 3$ after 4 iterations</th>
<th>Empirical moments of the 3 drawn samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu_1 = -0.04, \sigma_1 = 0.9$</td>
<td>$\mu_1 = 0.1, \sigma_1 = 1.0$</td>
</tr>
<tr>
<td>$\mu_2 = 3.5, \sigma_2 = 1.2$</td>
<td>$\mu_2 = 3.4, \sigma_2 = 1.8$</td>
</tr>
<tr>
<td>$\mu_3 = -4.9, \sigma_3 = 1.7$</td>
<td>$\mu_3 = -4.9, \sigma_3 = 1.7$</td>
</tr>
</tbody>
</table>

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:

$$P_1 = 1/3, \quad \mu_1 = -2, \quad \sigma_1 = 1, \quad P_2 = 2/3, \quad \mu_2 = +2, \quad \sigma_2 = 1.$$  

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

Input:

$$k = 2,$$

Output:

$$P_1 = 8/25, \quad \mu_1 = -2.2, \quad \sigma_1 = 0.8,$$

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

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:

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

juin 1976.
6.3. 150 points in $\mathbb{R}^2$ have been drawn from three two-dimensional Gaussian populations [see fig. a and b, and appendix 3] of mean-vector $\mu_i$ and covariance matrices $\Sigma_i$ ($i = 1, 2, 3$).

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

and

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

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.

![Ellipsoids of equiprobability (95 %)](image)

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

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

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

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

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

$$\begin{align*}
\mu_3 &= (4; 3), \\
\Sigma_3 &= \begin{pmatrix} 4 & -\frac{2\sqrt{3}}{2} \\ -\frac{2\sqrt{3}}{2} & 2 \end{pmatrix},
\end{align*}$$

(→ a (5/6) $\Pi$ angle).

juin 1976.
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):

\[
\begin{align*}
\mu_1 &= (-.1; -.1), & \Sigma_1 &= \begin{pmatrix} 2.8 & 1.4 \\ 1.4 & 1.0 \end{pmatrix}, \\
\mu_2 &= (-.1; 2.9), & \Sigma_2 &= \begin{pmatrix} .22 & -.03 \\ -.03 & .26 \end{pmatrix}, \\
\mu_3 &= (3.9; 3.0), & \Sigma_3 &= \begin{pmatrix} 3.2 & -1.8 \\ -1.8 & 1.4 \end{pmatrix}.
\end{align*}
\]

(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 preceding 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.

*Revue Française d'Automatique, Informatique et Recherche Opérationnelle*
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 $\mu_1 = 0$ $\sigma_1 = 1$

<table>
<thead>
<tr>
<th>1.7394</th>
<th>6.194</th>
<th>3.708</th>
<th>3.497</th>
<th>1.4361</th>
<th>-5.312</th>
<th>-1.120</th>
<th>10.88</th>
<th>-3.396</th>
<th>-0.0164</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.0423</td>
<td>2.1062</td>
<td>7.442</td>
<td>1.5797</td>
<td>8.246</td>
<td>1.647</td>
<td>5.664</td>
<td>0.0836</td>
<td>-1.4007</td>
<td>-8.435</td>
</tr>
<tr>
<td>-6.672</td>
<td>1.4114</td>
<td>-4.637</td>
<td>-0.0791</td>
<td>3.015</td>
<td>0.975</td>
<td>1.285</td>
<td>3.514</td>
<td>7.351</td>
<td>42.71</td>
</tr>
<tr>
<td>1.785</td>
<td>7.724</td>
<td>-1.7588</td>
<td>4.954</td>
<td>8.018</td>
<td>3.521</td>
<td>-1.036</td>
<td>1.2092</td>
<td>1.1873</td>
<td>1.2409</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>2.9314</th>
<th>4.3458</th>
<th>1.6919</th>
<th>5.0392</th>
<th>5.0085</th>
<th>2.6981</th>
<th>3.1122</th>
<th>2.3901</th>
<th>2.3309</th>
<th>4.4738</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.6453</td>
<td>4.9680</td>
<td>3.9999</td>
<td>1.2877</td>
<td>5.7267</td>
<td>7.706</td>
<td>5.0836</td>
<td>4.341</td>
<td>5.6432</td>
<td>7.7662</td>
</tr>
<tr>
<td>5.8085</td>
<td>2.9557</td>
<td>1.4572</td>
<td>4.1418</td>
<td>3.7362</td>
<td>7.1410</td>
<td>1.2201</td>
<td>5.0516</td>
<td>5.3289</td>
<td>4.5089</td>
</tr>
<tr>
<td>9.066</td>
<td>3.9800</td>
<td>6.0400</td>
<td>2.4195</td>
<td>2.1572</td>
<td>1.1677</td>
<td>1.5911</td>
<td>1.0378</td>
<td>5.9066</td>
<td>2.0994</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>2.8243</th>
<th>-6.9449</th>
<th>6.2504</th>
<th>-6.9985</th>
<th>3.7375</th>
<th>-5.4383</th>
<th>-2.9924</th>
<th>-5.0097</th>
<th>-5.1264</th>
<th>-1.6716</th>
</tr>
</thead>
<tbody>
<tr>
<td>-4.6563</td>
<td>-6.0423</td>
<td>2.3475</td>
<td>3.7037</td>
<td>-5.0953</td>
<td>-1.2384</td>
<td>-5.5724</td>
<td>2.2887</td>
<td>3.5806</td>
<td>-6.8857</td>
</tr>
</tbody>
</table>

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 \( \mu_2 = +2 \).

<table>
<thead>
<tr>
<th>( k )</th>
<th>( x_k )</th>
<th>(Class)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.608</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>-1.590</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>0.235</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>3.949</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>-2.249</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>2.704</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>-2.473</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>0.672</td>
<td>2</td>
</tr>
<tr>
<td>9</td>
<td>0.262</td>
<td>2</td>
</tr>
<tr>
<td>10</td>
<td>1.072</td>
<td>2</td>
</tr>
<tr>
<td>11</td>
<td>-1.773</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td>0.537</td>
<td>2</td>
</tr>
</tbody>
</table>

APPENDIX 3

150 observations from a Gaussian population of parameters:

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

Revue Française d'Automatique, Informatique et Recherche Opérationnelle
A NEW APPROACH IN MIXED DISTRIBUTIONS DETECTION

juin 1976.
APPENDIX 4

150 observations from a Gaussian population or parameters:

\[
\mu_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma_1 = \begin{pmatrix} 4 & 2\sqrt{3} \\ 2\sqrt{3} & 2 \end{pmatrix}, \quad \mu_2 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},
\]

\[
\mu_3 = \begin{pmatrix} 4 \\ 3 \end{pmatrix}, \Sigma_3 = \begin{pmatrix} 4 & -2\sqrt{3} \\ -2\sqrt{3} & 2 \end{pmatrix}
\]

E. DIDAY, A. SCHROEDER

Revue Française d'Automatique, Informatique et Recherche Opérationnelle
REFERENCES


juin 1976.