Algorithms for Model-based Block Gaussian Clustering

M. Nadif\(^1\) and G. Govaert\(^2\)

\(^1\) CRIP5, Université Paris Descartes, Paris, France
\(^2\) Heudiasyc, UMR CNRS 6599, Université de Technologie de Compiègne, Compiègne, France

Abstract - When the data consists of a set of objects described by a set of continuous variables, the clustering can concern the sets of objects (rows), variables (columns) or the both sets simultaneously. Considering the last type of clustering, we propose a new mixture model and develop an adapted Generalized EM (GEM) algorithm as part of the maximum likelihood, and a Classification GEM (CGEM) version as part of the classification maximum likelihood approach. The different steps of these new algorithms are presented showing the interest in data mining context. Some illustrative synthetic data allow us to evaluate their performances in comparison with EM and Classification EM (CEM) applied on the sets of objects given a partition of variables, and EM and CEM applied only on the set of objects.

Keywords: Block mixture model, latent block model, EM

1. Introduction

Cluster analysis is an important tool in a variety of scientific areas including pattern recognition, information retrieval, micro-arrays and data mining. Although many clustering procedures such as hierarchical clustering and \(k\)-means, aim to construct an optimal partition of objects or, sometimes, variables, there are other methods, known as block clustering methods or latent block models, which consider the two sets simultaneously and organize the data into homogeneous blocks. Here, we restrict to block clustering methods defined by a partition of objects and a partition of variables. If \(x\) denotes an \(n \times d\) data matrix defined by \(x = \{(x_{ij}); i \in I \text{ and } j \in J\}\), where \(I\) is a set of \(n\) objects (rows, observations, cases) and \(J\) is a set of \(d\) numeric variables (columns, attributes), the basic principle of these methods is to make permutations of objects and variables in order to construct a correspondence structure on \(I \times J\).

Another advantage of block clustering methods is that they reduce the initial data matrix \(x\) into a simpler data matrix having the same structure. Moreover, these methods are fast and can process large data sets. Far less computation is required than for processing the two sets separately and consequently these methods are of interest in data mining.

These last years, block clustering has become an important challenge in data mining context. In the text mining field, Dhillon [8] has proposed a spectral block clustering method by exploiting the duality between rows (documents) and columns (words). In the analysis of microarray data where data are often presented as matrices of expression levels of genes under different conditions, block clustering of genes and conditions has permitted to overcome the problem of the choice of similarity on the two sets found in conventional clustering methods [5]. Also, these kinds of methods have practical importance in a wide variety of applications such as text, web-log and market basket data analysis and bioinformatics.

For numeric variables, there are some manners to attempt the block clustering objective. The simple one consists to clustering the variables and from the clusters obtained, we apply a clustering algorithm on the set of objects. To cluster variables, we can use partitioning methods such as \(k\)-means, or hierarchical methods such as the well-known SAS Varclus procedure which is a type of oblique component analysis related to multiple group factor analysis [13]. The \(k\)-means applied on variables centered and Varclus are used as variable-reduction methods where the variables of the same cluster are correlated as much as possible, and two variables belonging to different clusters are as uncorrelated as possible. In order to obtain the homogeneous blocks, there are other methods more adapted which consist to clustering the both sets simultaneously, we can cite for examples the works of Hartigan [14], [1] and Bock [3]. Latter, we will interest particularly in the \textit{Croeuc} algorithm proposed by Govaert [10].

In this paper, we will focus on these kinds of methods. We will study the block clustering problem in embedding it in the mixture approach. We will propose a new mixture model taking into account the block clustering situation and perform soft and hard innovative and suitable co-clustering algorithms. To propose the both algorithm, we set this problem in the maximum likelihood (ML) approach [6], [20] and in the classification maximum likelihood (CML) approach [19]. Results on simulated data are given to illustrate the behaviors of these algorithms.

This paper is organized as follows. Section 2 begins with a review of the ML and CML approaches, the EM and Classification EM algorithms. As we are interested in the modeling of our block clustering problem, we describe our mixture model in Section 3. In Section 4, we propose the adapted soft and hard algorithms. Section 5 is devoted to describe some parsimonious models. To achieve our aim we study the behaviors of our algorithms and compare it with two methodologies using separately on the clustering on the set of objects and the set of variables in Section 6. Finally, the last section summarizes the main points of this paper.
2. Mixture Model and EM algorithm

2.1. Notation

We now define the notation that is used consistently throughout this paper. The partition of a sample \( I \) into \( g \) clusters is noted \( z \) and will be represented by the classification matrix \( (z_{ik}; i = 1, \ldots, n; k = 1, \ldots, g) \) where \( z_{ik} = 1 \) if \( i \) belongs to the \( k \)th cluster and 0 otherwise. A similar notation will be used for the partition \( w \) of a set \( J \) into \( m \) clusters represented also by the classification matrix \( (w_{ij}; j = 1, \ldots, m; \ell = 1, \ldots, m) \). Moreover, to simplify the notation, the sums and the products relating to rows, columns or clusters will be subscripted respectively by letters \( i, j, k \) or \( \ell \) without indicating the limits of variation, which will be implicit. So, for example, the sum \( \sum_{i} \) stands for \( \sum_{i=1}^{n} \), and \( \sum_{i,j,k,\ell} \) stands for \( \sum_{i=1}^{n} \sum_{j=1}^{d} \sum_{k=1}^{g} \sum_{\ell=1}^{m} \).

2.2. Finite Mixture model

Finite mixture models underpin a variety of techniques in major areas of statistics including cluster analysis; see for instance [17]. With a mixture model-based approach clustering, it is assumed that the data to be clustered are generated by a mixture of underlying probability distributions in which each component represents a different cluster. Given observations \( x = (x_1, \ldots, x_n) \), let \( \varphi_k(x_i; \alpha_k) \) be the density of an observation \( x_i \) from the \( k \)th component, where the \( \alpha_k \)'s are the corresponding parameters and let \( g \) be the number of components in the mixture. The probability density function is

\[
f(x_i; \theta) = \sum_k \pi_k \varphi_k(x_i; \alpha_k),
\]

where \( \pi_k \) is the probability that an observation belongs to the \( k \)th component and \( \theta \) is the vector of the unknown parameters \( (\pi_1, \ldots, \pi_g; \alpha_1, \ldots, \alpha_g) \).

The problem of clustering can be studied in the mixture model using two different approaches. The ML approach and the CML approach. The ML approach has been by far the most commonly used approach to the fitting of mixture distribution and is appropriate to tackle this problem. It estimates the parameters of the mixture, and the partition of \( I \) is derived from these parameters using the maximum a posteriori principle (MAP). Classical optimization techniques such as Newton-Raphson or gradient methods can be used but, in mixture context, the EM algorithm [7] has been successfully applied and is one of the most widely used procedures.

2.3. EM algorithm

Let us remind that this algorithm is a method for maximizing the log-likelihood \( L(\theta) \) iteratively, using the maximization of the conditional expectation of the complete-data log-likelihood given a previous current estimate \( \theta^{(c)} \) and the observed data \( x \).

In mixture model, we take the complete-data to be the vector \( (x, z) \) where the unobservable vector \( z \) is the label data; the complete-data log-likelihood \( L_c(\theta; x, z) \) noted also \( L_c(z; \theta) \) is written as

\[
L_c(z; \theta) = \sum_{i,k} z_{ik} \log \pi_k \varphi_k(x_i; \alpha_k)
\]

and its conditional expectation can we written

\[
Q(\theta, \theta^{(c)}) = \sum_{i,k} z_{ik} \{ \log(\pi_k) + \log \varphi_k(x_i; \alpha_k) \},
\]

where

\[
s_{ik}^{(c)} = P(z_{ik} = 1|x, \theta^{(c)}) = \frac{\pi_k \varphi_k(x_i; \alpha_k)}{\sum_{k'} \pi_{k'} \varphi_{k'}(x_i; \alpha_k)}
\]

denotes the conditional probability, given \( x \) and \( \theta^{(c)} \), that \( x_i \) arises from the mixture component with density \( \varphi_k(x; \alpha_k) \). Each iteration of EM has two steps: an E-step and a M-step. The \( (c + 1) \)st M-step finds the conditional expectation of the complete-data log-likelihood. Note that in the mixture case this step reduces to the computation of the conditional density of the \( s_{ik}^{(c)} \). The \( (c + 1) \)st M-step finds \( \theta^{(c+1)} \) maximizing \( Q(\theta, \theta^{(c)}) \).

The characteristics of the EM algorithm are well documented; see for instance [18]. It leads in general to simple equations, has the nice property of increasing the log-likelihood at each iteration until stationarity, and in many circumstances, it derives sensible parameter estimates and consequently it is a popular tool to obtain maximum likelihood estimation. The EM algorithm can be viewed as the soft algorithm, and the partition can be derived from the parameters by using the MAP.

Note that, as pointed out by Hathaway [15], in the classical mixture model context the EM algorithm can be viewed as an alternated maximization of the following fuzzy clustering criterion

\[
F_c(s; \theta) = L_c(s; \theta) + H(s)
\]

where \( s = (s_{ik}) \) denotes a \( n \times g \) data matrix which expresses a fuzzy partition, \( L_c(s; \theta) = \sum_{i,k} s_{ik} \log(\pi_k f(x_i; \alpha)) \) is the fuzzy complete data log-likelihood and \( H(s) = -\sum_{i,k} s_{ik} \log s_{ik} \) is the entropy of the clustering. In fact, maximizing \( F_c \) w.r. to \( s \) yields the E-step and maximizing \( F_c \) w.r. to \( \theta \) yields the M-step. At the convergence, the log-likelihood \( L_c(\theta) \) is therefore equal to \( F_c(s; \theta) \) where \( s \) is computed from \( \theta \).

2.4. CEM algorithm

To tackle the problem of clustering, we can consider the CML approach, the partition is added to the parameters to be estimated. The maximum likelihood estimation of these new parameters leads to optimize in \( \theta \) and \( z \) the complete data log-likelihood (2). This optimization can be done by the Classification EM (CEM) algorithm [4], a variant of EM, which converts the posterior probabilities \( s_{ik} \)'s to a discrete classification in a C-step before performing the M-step. We obtain a hard version of EM.

Some of the most popular heuristic clustering methods can be viewed as approximate estimations of probability models. For instance, the sum-of-squares criterion optimized by the k-means algorithm corresponds to the CML.
approach associated with a Gaussian mixture model with strong constraints. Hence, CEM is a generalization of k-means.

3. Model-based block clustering

To tackle the problem of block clustering, recently we have proposed a block mixture model. We have studied the cases of binary data and contingency tables [11],[12]. The model proposed is symmetric and appears less adapted to our situation where the objects are described by a set of continuous variables. Hereafter, we propose another model based on the classical formulation (1) with the added parameter \( \alpha \).

The pdf is written as

\[
f(x_i; \theta) = \sum_k \pi_k \varphi_k(x_i; \omega, \alpha)
\]

where

\[
\varphi_k(x_i; \omega, \alpha) = \prod_{j} \prod_{\ell} \frac{1}{\sqrt{2\pi \sigma_{\ell}^2}} e^{-\frac{1}{2\sigma_{\ell}^2}(x_{ij} - \mu_{\ell k})^2}
\]

\[
= \prod_{j} \left( \frac{1}{\sqrt{2\pi \sigma_{\ell}^2}} e^{-\frac{1}{2\sigma_{\ell}^2}(x_{ij} - \mu_{\ell k})^2} \right)^{w_{j\ell}}.
\]

The unknown parameter \( \theta \) is formed by \( \pi, \omega \) and \( \alpha \) with \( \pi = (\pi_1, \ldots, \pi_\gamma) \), \( \omega \) is a partition of the variables and \( \alpha \) represents the means and variances for each block \( (\mu_{11}, \ldots, \mu_{\gamma m}, \sigma_{1j}^2, \ldots, \sigma_{\gamma j}^2) \).

The log-likelihood can be written

\[
L(\theta) = \log f(x; \theta) = \sum_i \log \sum_k \pi_k \varphi_k(x_i; \omega, \alpha)
\]

and, if we denote \( z_k = \sum_i z_{ik} \) and \( w_\ell = \sum_j w_{j\ell} \) (the cardinality of each cluster), the corresponding complete data log-likelihood \( L_c(z, \omega; \theta) \) is written as

\[
\sum_{i,k} z_{ik} \log \left( \pi_k \varphi_k(x_i; \omega, \alpha) \right),
\]

and takes therefore, up to the constant \(-\frac{nd}{2} \log 2\pi\), the following form:

\[
\sum_k z_{ik} \log \pi_k - \frac{1}{2} \sum_{i,j,k,\ell} z_{ik} w_{j\ell} \left( \log \sigma_{k\ell}^2 + \frac{1}{\sigma_{k\ell}^2}(x_{ij} - \mu_{k \ell})^2 \right).
\]

In the following, we will also use the following extending definition of this complete data log-likelihood \( L_c \) to a fuzzy partition \( s \) represented by the classification matrix \((s_{ik}; i = 1, \ldots, n; k = 1, \ldots, g)\) where \( s_{ik} \in [0, 1] \) with \( \sum_k s_{ik} = 1 \).

\[
L_c(s, \omega; \theta) = \sum_{i,k} s_{ik} \log \left( \pi_k \varphi_k(x_i; \omega, \alpha) \right)
\]

and the relation (3) becomes

\[
\sum_k s_k \log \pi_k - \frac{1}{2} \sum_{i,j,k,\ell} s_{ik} w_{j\ell} \left( \log \sigma_{k\ell}^2 + \frac{1}{\sigma_{k\ell}^2}(x_{ij} - \mu_{k \ell})^2 \right),
\]

where \( s_k = \sum_i s_{ik} \).

4. Soft and hard algorithms

Setting our model under the ML approach, we will propose an associated EM algorithm. Before describing The E and M steps, we choose to initialize EM by \((w^{(0)}, \theta^{(0)})\) obtained with the classical \( k \)-means applied on the set of variables.

4.1. E-step

In the mixture case, this step reduces to the computation of the conditional density \( s_{ik}^{(c)} \) since the conditional expectation is written

\[
Q(\theta, \theta^{(c)}) = \sum_{i,k} s_{ik}^{(c)} \log(\pi_k \varphi_k(x_i; \omega, \alpha)),
\]

where

\[
s_{ik}^{(c)} = P(z_{ik} = 1|\theta, \theta^{(c)}) = \frac{\pi_k \varphi_k(x_i; \omega^{(c)}, \alpha^{(c)}) \sum_k \pi_k \varphi_k(x_i; \omega^{(c)}, \alpha^{(c)})}{\sum_k \pi_k \varphi_k(x_i; \omega^{(c)}, \alpha^{(c)})}.
\]

This denotes the conditional probability, given \( x \) and \( \theta^{(c)} \), that \( x_i \) arises from the mixture component with density \( \varphi_k(x_i; \omega, \alpha) \).

a) Compute of \( s_{ik} \): For convenience, if we denote \( S_{ik} = \log(\pi_k \varphi_k(x_i; \omega, \alpha)) \) the conditional probabilities can be written

\[
s_{ik} = \frac{e^{s_{ik}}}{\sum_k e^{s_{ik}}}.
\]

After some calculations, the term \( S_{ik} \) takes the following form

\[
\log \pi_k - \frac{1}{2} \sum_\ell \left( w_{\ell} \log \sigma_{k\ell}^2 + \frac{1}{\sigma_{k\ell}^2}(e_{\ell} + w_{\ell}(u_{\ell} - \mu_{k \ell})^2) \right)
\]

where \( u_{\ell} = \sum_j w_{j\ell} x_{ij} \) and \( e_{\ell} = \sum_j w_{j\ell}(x_{ij} - u_{\ell})^2 \).

4.2. M-Step

The maximization of \( Q(\theta, \theta^{(c)}) \) is not straightforward and, using the Generalized EM algorithm [7] for which the M-step requires \( \theta^{(c+1)} \) to be chosen such that \( Q(\theta^{(c+1)}, \theta^{(c)}) \geq Q(\theta^{(c)}, \theta^{(c)}) \); when one chooses \( \theta^{(c+1)} \) to increase the function \( Q(\theta, \theta^{(c)}) \) rather than maximize it over \( \theta \). Note that \( Q(\theta, \theta^{(c)}) \) is the fuzzy complete data log-likelihood \( L_c(s\omega, \omega, \theta) \), then from (3), the expression (4) takes this form

\[
\sum_k s_k^{(c)} \log \pi_k - \frac{1}{2} \sum_{i,j,k,\ell} s_{ik} w_{j\ell} \left( \log \sigma_{k\ell}^2 + \frac{1}{\sigma_{k\ell}^2}(x_{ij} - \mu_{k \ell})^2 \right).
\]

To reach our objective we will alternate the optimization of \( Q(\theta, \theta^{(c)}) \), we maximize \( Q(\omega, \theta^{(c)}) \) w.r. to \( \omega \) given \( s \) and \( \theta^{(c)} \) then maximize \( Q(\theta, \theta^{(c)}) \) w.r. to \( \theta \) given \( s \) and \( \omega \).

1) Maximization of \( Q(\omega, \theta^{(c)}) \) w.r. to \( \omega \): The previous expression of \( L_c(s^{(c)}\omega, \omega, \theta) \) can be written as

\[
\sum_k s_k^{(c)} \log \pi_k + \sum_\ell w_{j\ell} \beta_j^{(c)} \theta \]

where

\[
\beta_j^{(c)} = \frac{1}{2} \sum_{i,k} s_{ik} \left( \log \sigma_{k\ell}^2 + \frac{1}{\sigma_{k\ell}^2}(x_{ij} - \mu_{k \ell})^2 \right).
\]
The $\ell$th cluster of the optimal partition corresponds to the highest $S_{j\ell}^{(c)}$ for the variable $j$. $w_{j\ell}^{(c)} = 1$ if $\ell = \arg\max_{k=1,...,g} S_{j\ell}^{(c)}$ and $w_{j\ell}^{(c)} = 0$ otherwise.

Note that, this step is easily performed by using sufficient statistics

$$v_{jk} = \sum_i s_{ik} x_{ij}$$

and

$$f_{jk} = \sum_i s_{ik} (x_{ij} - v_{jk})^2,$$

allowing to compute easily $S_{j\ell}$ which becomes

$$-\frac{1}{2} \sum_k \left( s_k^{(c)} \log \sigma_{k\ell}^2 + \frac{1}{\sigma_{k\ell}^2} (f_{jk} + s_k (v_{jk} - \mu_{k\ell})^2) \right).$$

2) Maximization of $Q(\theta|\theta^{(c)})$ w.r. to $\theta$ given $w$ and $s$:

This step consists in computing the maximum likelihood $\pi$ et $\alpha = (\mu_{11}, \ldots, \mu_{gm}, \sigma_{11}, \ldots, \sigma_{gm})$. Using the $k$th component and the $\ell$th cluster, the expression of $L_c(s,w;\theta)$ is

$$\sum_k s_k \log \pi_k$$

$$- \frac{1}{2} \sum_{k,\ell} \left( s_k w_{\ell} \log \sigma_{k\ell}^2 + \frac{1}{\sigma_{k\ell}^2} \sum_{i,j} s_{ik} w_{\ell j} (x_{ij} - \mu_{k\ell})^2 \right).$$

It leads to

$$\pi_k^{(c+1)} = \frac{s_k^{(c)}}{n},$$

and

$$\mu_{k\ell}^{(c+1)} = \frac{\sum_{i,j} s_{ik}^{(c)} w_{\ell j}^{(c)} x_{ij}}{s_k^{(c)} w_{\ell}^{(c)}}.$$ 

To accelerate the compute of $\mu_{k\ell}$, we use the sufficient statistic $v_{jk}$ previously computed, then

$$\mu_{k\ell}^{(c+1)} = \frac{\sum_{j} w_{j\ell}^{(c)} v_{jk}}{s_k^{(c)} w_{\ell}^{(c)}}.$$ 

In the same manner by using $f_{jk}$, we compute the variance for the block $k\ell$ by

$$\sigma_{k\ell}^{(c+1)} = \frac{\sum_{j} w_{j\ell}^{(c)} (f_{jk} + s_{ik}^{(c)} (v_{jk} - \mu_{k\ell})^2)}{s_k^{(c)} w_{\ell}^{(c)}}.$$ 

Concerning the context of clustering with the ML approach, after we estimate parameter $\theta$, we can give a probabilistic clustering of the $n$ objects in term of their fitted posterior probabilities of component membership $s_{ik}$ obtained at the end of EM. Then, we can obtain a partition by using C-step which assigns each object to the component of the mixture to which it has the highest posterior of probability of belonging. With the optimal $w$ partition, we obtain therefore a block clustering where a partition of objects is characterized by a partition of variables. Note that with our algorithm, the different computations in the M-step are performed on a reduced matrix using sufficient statistics $v_{jk}$ and therefore it is suitable for large data sets. In the other hands, the GEM algorithm proposed can be viewed as a soft algorithm to cluster simultaneously the set of objects and the set of variables.

A hard version Classification EM [4] can be performed by substituting $Q(\theta|\theta^{(c)})$ by $L_c(z,w;\theta')$. The main modifications concern therefore the conditional maximization of complete data log-likelihoods w.r. to $w$ given $z$ and $\theta$ and w.r. to $\theta$ given $z$ and $w$. Then the different steps are:

1) Start from an initial position $(w^{(0)}, \theta^{(0)})$ obtained for example by the classical $k$-means.

2) E-step Computation of $s_{ik}$

3) C-step Computation of $z$: the $k$th cluster of $z^{(c)}$ is defined with $z_{ik}^{(c)} = 1$ if $k = \arg\max_{k=1,...,g} s_{ik}^{(c)}$ and $z_{ik}^{(c)} = 0$ otherwise.

4) M-step Maximization of $L_c$

(a) Compute $w$ by maximizing $L_c$ w.r. to $w$ given $z$: the $\ell$th cluster of $w^{(c+1)}$ is defined with $w_{j\ell}^{(c+1)} = 1$ if $\ell = \arg\max_{k=1,...,g} S_{j\ell}^{(c)}$ and $w_{j\ell}^{(c+1)} = 0$ otherwise.

(b) Compute $\theta$ by maximizing $L_c$ w.r. to $\theta$ given $w$: by standard calculations, one arrives at the following re-estimations parameters.

$$\pi_k^{(c+1)} = \frac{z_{ik}^{(c)}}{n},$$

$$\mu_{k\ell}^{(c+1)} = \frac{\sum_{j} w_{j\ell}^{(c+1)} v_{jk}}{z_{ik}^{(c)} w_{\ell}^{(c)}},$$

and

$$\sigma_{k\ell}^{(c+1)} = \frac{\sum_{j} w_{j\ell}^{(c)} (f_{jk} + z_{ik} (v_{jk} - \mu_{k\ell})^2)}{z_{ik}^{(c)} w_{\ell}^{(c)}}.$$ 

5) Iterate the steps 2), 3) and 4) until the convergence.

5. Parsimonious models

For the classical Gaussian mixture model, Banfield and Raftery [2] have proposed different variants of this model. They have considered a parametrization of the covariance matrix $\Sigma_k$ in terms of its eigenvalue decomposition, $\Sigma_k = \lambda_k D_k A_k D_k^T$ (the superscript $T$ denotes matrix transposition), where $\lambda_k$ defines the volume of the $k$th cluster, $D_k$ is an orthogonal matrix, which defines its orientation and $A_k$ is a diagonal matrix with determinant 1, which defines its shape. This parametrization allows one to propose many general criteria and the simplest one corresponds to spherical clusters and equal volumes that lead to the famous $k$-means criterion. From our model, we can apply a similar parametrization on the parameters $\pi_k$ and $\sigma_{k\ell}^2$. Some parsimonious models allow one to take into account particular constraints on the proportions and the degree of homogeneity in the blocks. Then we can propose the following models where the cluster proportions are supposed different

- $[\pi_k, \sigma_{k\ell}^2]$: variances different in the blocks.
- $[\pi_k, \bar{\sigma}_{k\ell}^2]$: variances depending only on the clusters of variables.
- $[\pi_k, \bar{\sigma}_{\ell}^2]$: variances depending only on the clusters of objects.
- $[\pi_k, \bar{\sigma}^2]$: a same variance for all blocks.


From these models, we can impose that the proportions are equal and the simplest model is $[\pi, \sigma^2]$. In this case, the complete data log-likelihood can be written:

$$L_c(z, w, \pi, \alpha) = -n \log g - \frac{nd}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i,j,k,\ell} z_{ik} w_{j\ell} (x_{ij} - \mu_{k\ell})^2$$

and it is easy to see that the maximization of $L_c$ and the minimization of $W(z, w, \mu)$ where

$$W(z, w, \mu) = \sum_{i,j,k,\ell} z_{ik} w_{j\ell} (x_{ij} - \mu_{k\ell})^2$$

are equivalent. Then we have a signification of the block clustering algorithm called Crobec minimizing $W$: the proportions are supposed equal and the variances for all the blocks are the same. Moreover, the algorithm used, minimizing alternatively the two criteria $W(z, \mu|w)$ and $W(w, \mu|z)$, appears as a particular version of the CGEM algorithm for the parsimonious model $[\pi, \sigma^2]$.

6. Numerical experiments

6.1. Experiment conditions

To evaluate the interest of our model and the algorithms proposed GEM and CGEM, we compared their performances to other algorithms EM, CEM, EM-w and CEM-w described hereafter:

- **GEM**: Generalized algorithm from our model.
- **CGEM**: Classification version of GEM.
- **EM**: EM applied only on the set of objects, we ignore the clustering on the variables.
- **CEM**: Classification version of EM applied on the set of variables and the set of objects but separately. The partition obtained is used by EM-w and CEM-w.
- **EM-w**: Classical EM applied with optimal partition $w$ obtained by CEM.
- **CEM-w**: Classification version of EM-w.

In our experiments, we selected several kinds of data arising from $3 \times 3$-component mixture model corresponding to three degrees of overlap of the clusters: well separated, moderately separated and poorly separated. The concept of cluster separation is difficult to visualize easily for our model, but the degree of overlap can be measured by the true error rate approximated by comparing the partitions simulated with those we obtained by applying a classification step. From our numerical experiments, we present only 3 situations corresponding to 3 levels of overlap degrees: M1 for clusters well separated, M2 for moderately separated and M3 for poorly separated. Furthermore, to have an idea of the degree of overlap, we used the principal components analysis (PCA) on data and transposed data. The figures 1, 2 and 3 represent the factorial planes for M1, M2 and M3.

To compare two partitions $z$ and $z'$ having the same number of clusters, the error rate or the proportions of misclassified objects can be defined as follows: If $C$ is the confusion matrix between the two partitions, relabel the components of the partition $z'$ such that the trace of
matrix $C$ is maximal (to obtain this maximum value in our experiments, we enumerate all possible relabeling), then compute
\[
\delta(z, z') = 1 - \frac{1}{n} \sum_{i,k} z_{ik} z'_{ik}
\]
We extend this formula to compute the proportions of misclassified variables $\delta(w, w')$.

6.2. Results and discussion

The simulations results for the six algorithms are summarized in Table I which displays the overall apparent error for each model. Note that, for the clustering of variables $\delta(w, w')$ is the same for CEM, EM-w and CEM-w because the partition obtained by CEM is used by EM-w and CEM-w.

From these experiments, the main point arising are the following.

- The EM algorithm working only on the set of objects and CEM on the both sets separately, are suitably effective only when the clusters are well separated. This shows the risk of the use of such methods in order to detect homogeneous blocks.
- The CGEM algorithm, even if it is faster than GEM and better than CEM and EM, does not give encouraging results when the clusters are not well separated.
- Incontestably the GEM algorithm outperforms the other algorithms. It gives encouraging results for the three models. Note that to accelerate the convergence of EM we can develop an efficient lazy EM noted eLEM [16] or initialize it with CGEM.

7. Summary and future works in progress

When the data consists of a large number of variables taking values over a large number of cases, such as in data mining context, block clustering algorithms can be an interesting approach. By clustering cases and variables simultaneously, they allow us to find patterns into homogeneous blocks, which can be viewed as a summary of the data.

In this work, we have proposed a mixture model taking into account algorithms such as block clustering algorithms. Setting the problem of block clustering under the ML and CML approaches, we have proposed the associated GEM and CGEM algorithms. Under the CML approach, our model gives an interpretation of the Crovatic criterion and allowed us to propose a general criterion, which depends on unknown proportions and standard deviation for each block. Moreover, we proved that the Crovatic algorithm is just a particular version of the CGEM algorithm.

From our experiments illustrated by three situations, the GEM algorithm appears the better but CGEM which is faster, can be used to initialize GEM to avoid to run GEM several times. The results are encouraging results on simulated data but it would be now necessary to apply this algorithm to real data.

8. References


| Table I: Comparison results between the six algorithms |
|----------------------------------|---|---|---|---|---|---|
| Model | GEM | CGEM | CEM | EM | EM-w | CEM-w |
| M1 | 1 | 1 | 0 | 0 | 1 | 1 |
| M2 | 11 | 12 | 0 | 19 | 15 | 15 |
| M3 | 29 | 41 | 41 | 39 | 44 | 42 |
| M1 | 0 | 0 | 0 | 0 | 0 |
| M2 | 5 | 5 | 30 | 30 |
| M3 | 20 | 35 | 50 | 50 | 50 |


