ELSEVIER

Pattern Recognition Letters 19 (1998) 919–927

Pattern Recognition Letters

# Convergence of an EM-type algorithm for spatial clustering

C. Ambroise [a,*], G. Govaert [b]

[a] *LODYC, Université Pierre et Marie Curie, 75252 Paris Cedex, France*
[b] *UMR CNRS 6599, Université de Technologie de Compiègne, F-60200 Compiègne, France*

Received 23 July 1997; received in revised form 25 May 1998

**Abstract**

Ambroise et al. (1996) have proposed a clustering algorithm that is well-suited for dealing with spatial data. This algorithm, derived from the EM algorithm (Dempster et al., 1977), has been designed for penalized likelihood estimation in situations with unobserved class labels. Some very satisfactory empirical results lead us to believe that this algorithm converges (Ambroise et al., 1996). However, this convergence has not been proven theoretically. In this paper, we present sufficient conditions and proof of the convergence. A practical application illustrates the use of this algorithm. 

*Keywords:* EM algorithm; Gaussian mixtures; Spatial data; Penalization

## 1. Introduction

Spatial clustering aims to find classes composed of objects which are both similar according to some measure and geographically close. When classical clustering algorithms (e.g., the EM algorithm for Gaussian mixture estimation) are used for partitioning spatial data, the resulting classes will often be spatially very mixed.

In geology, sociology, image analysis, and in a wide range of other fields, spatial clustering techniques are widely used for finding homogeneous zones. Satellite images are often segmented in order to determine different zones of interest (e.g. forests, cities or rivers). In this particular case, the objects are pixels described by a gray scale or color intensity. Another example consists of statistics describing the number of sick persons per geographic unity (e.g. town or country) which may be used for delimiting different zones of risk.

Several methods exist for taking spatial information into account in a clustering process:

- Modifying existing clustering *algorithms* (Legendre, 1987; Lebart, 1978; Openshaw, 1977). This is done by specifying which objects are neighbors and allowing an object to be assigned to a class if and only if this class already contains a geographical neighbor. This approach has the drawback of producing classes which are necessarily geographically connected. This means that one class is bound to correspond to a single spatial region.
- Integrating the spatial information in the *data set* (Berry, 1966; Jain and Farrokhnia, 1991; Oliver and Webster, 1989). One example consists of considering the geographical

* Corresponding author. E-mail: ca@lodyc.jussieu.fr.

0167-8655/98/$ – see front matter 
PII: S0167-8655(98)00076-2

coordinates as new variables describing the objects; another example is the filtering techniques that extract new features from the original variables which embody the spatial information.

- Choose a *model* which encompasses the spatial aspect of the data. Most of the time, this is equivalent to defining a criterion that includes spatial constraints. This approach comes mainly from image analysis where Markov random fields (Geman and Geman, 1984; Masson and Pieczinsky, 1993) are intensively used.

In a recent paper, the authors have described a clustering algorithm (Ambroise et al., 1996) related to the last approach which is able to deal with objects described by quantitative variables. The spatial distribution of the objects may be regular (e.g., pixels of a image), or irregular (e.g., towns of a given district). The algorithm estimates the parameters of a Gaussian mixture and produces a fuzzy partition made of classes which are spatially homogeneous without being "single spatial region classes".

This paper aims to present a proof of the convergence of this algorithm for spatial clustering. Section 2 introduces the Gaussian mixture model and describes the *Neighborhood EM algorithm* (NEM). Section 3 is dedicated to the convergence proof. In Section 4, an illustrative example based on image segmentation is presented.

## 2. Gaussian mixture and clustering

The probabilistic approach to clustering is mainly based on Gaussian mixture models. In this framework (Celeux and Govaert, 1995), the objects to be classified are considered as a sample $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$ of independent random vectors. The vectors $\boldsymbol{x}_i$ are drawn from a mixture of $K$ Gaussian distributions:

$$f(\boldsymbol{x}_i|\Phi) = \sum_{k=1}^{K} p_k f_k(\boldsymbol{x}_i|\theta_k), \tag{1}$$

where the $p_k$ are the mixing proportions (for $k = 1, \ldots, K,\ 0 < p_k < 1 \quad \text{and} \quad \sum_k p_k = 1$) and $f_k(\boldsymbol{x}|\theta_k)$ denotes the density of a Gaussian distribution with parameter $\theta_k = (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, $\boldsymbol{\mu}_k$ being the mean vector and $\boldsymbol{\Sigma}_k$ the covariance matrix. $\Phi$ is used to denote all the parameters of the mixture: proportions, mean vectors, and covariance matrices. This model also assumes that the unobserved vector of labels, $\boldsymbol{z} = (z_1, \ldots, z_N)$, is an i.i.d. sample of the multinomial distribution:

$$P(Z_i = k) = p_k \quad \text{for } 1 \leqslant i \leqslant N,\ 1 \leqslant k \leqslant K.$$

This kind of model is helpful in a clustering context. One could consider that the sample $\boldsymbol{x}$ is composed of $K$ sub-populations which are all Gaussian distributed. If the parameter $\Phi$ of the mixture is known, then it becomes possible to estimate the unknown labels which describe a partition of the sample into $K$ sub-populations.

### 2.1. The EM algorithm for Gaussian mixtures

The EM algorithm (Dempster et al., 1977) is often used to estimate the unknown parameters of the mixture. It produces a set of parameters that locally maximizes the log-likelihood of the sample, defined as

$$L(\Phi) = \sum_{i=1}^{N} \log f(\boldsymbol{x}_i|\Phi).$$

The principle of the EM algorithm consists of building a sequence of estimates $\Phi^0, \Phi^1, \ldots, \Phi^m$, over which the log-likelihood monotonically increases (for all $m$, $L(\Phi^{m+1}) \geqslant L(\Phi^m)$). At each iteration, $\Phi^{m+1}$ is chosen so that it maximizes the expectation of the likelihood of complete data defined as

$$\begin{aligned} Q(\Phi|\Phi^m) &\stackrel{\Delta}{=} \sum_{\boldsymbol{z}} P(\boldsymbol{z}|\boldsymbol{x}; \Phi^m) \log P(\boldsymbol{x}, \boldsymbol{z}; \Phi) \\ &= \sum_{i=1}^{N} \sum_{k=1}^{K} \log(p_k f_k(\boldsymbol{x}_i)) P(Z_i = k|\boldsymbol{x}_i; \Phi^m). \end{aligned}$$

Thus, starting from an arbitrary value $\Phi^0$, the $(m+1)$th iteration of the EM algorithm can be divided in two steps:

- *E-step* (Expectation): computation of $Q(\Phi|\Phi^m)$;
- *M-step* (Maximization): search for $\Phi^{m+1} = \arg\max_{\Phi} Q(\Phi|\Phi^m)$.

### 2.2. Another interpretation of the EM algorithm

By introducing the variables

$$c_{ik} = \frac{p_k f_k(\boldsymbol{x}_i|\theta_k)}{f(\boldsymbol{x}_i|\boldsymbol{\Phi})} \tag{2}$$

and using the relation $\sum_k c_{ik} = 1$, the likelihood $L(\boldsymbol{\Phi})$ may be written in the following way:

$$\begin{aligned} L(\boldsymbol{\Phi}) &= \sum_{i=1}^{N}\left(\sum_{k=1}^{K} c_{ik}\right) \log f(\boldsymbol{x}_i|\boldsymbol{\Phi}) \\ &= \sum_{i=1}^{N}\sum_{k=1}^{K} c_{ik} \log \frac{p_k f_k(\boldsymbol{x}_i|\theta_k)}{c_{ik}} \\ &= \sum_{i=1}^{N}\sum_{k=1}^{K} c_{ik} \log p_k f_k(\boldsymbol{x}_i|\theta_k) \\ &\quad - \sum_{i=1}^{N}\sum_{k=1}^{K} c_{ik} \log c_{ik} \\ &= D(\boldsymbol{c}, \boldsymbol{\Phi}). \end{aligned}$$

Following Hathaway (1986), this criterion $D$ may be considered as a fuzzy clustering criterion, since the matrix $\boldsymbol{c}$ has all the required properties of a fuzzy classification matrix:

$$\boldsymbol{c} = \left\{ c_{ik} : 0 \leqslant c_{ik} \leqslant 1, \sum_{k=1}^{K} c_{ik} = 1, \sum_{i=1}^{N} c_{ik} > 0 \right. \\ \left. (1 \leqslant i \leqslant N, 1 \leqslant k \leqslant K) \right\}.$$

From this point of view, it is possible to show that the EM algorithm is a grouped coordinate ascent algorithm which optimizes the criterion $D(\boldsymbol{c}, \boldsymbol{\Phi})$ alternatively over the two groups of variables $\boldsymbol{c}$ and $\boldsymbol{\Phi}$:

1. *Initialization* of the mixture parameters.
2. *Iterate:*
   - *E-step*, computation of a new classification matrix $\boldsymbol{c}^{q+1}$:

     $$\boldsymbol{c}^{q+1} = \arg\max_{\boldsymbol{c}} D(\boldsymbol{c}, \boldsymbol{\Phi}^q).$$

   - *M-step*, estimation of the mixture parameters:

     $$\boldsymbol{\Phi}^{q+1} = \arg\max_{\boldsymbol{\Phi}} D(\boldsymbol{c}^{q+1}, \boldsymbol{\Phi}).$$

In Section 2.3, the idea of grouped coordinate ascent will be extended for dealing with objects which have spatial relationships.

### 2.3. An algorithm for fuzzy spatial clustering

In order to take the spatial dependence of objects into account, we suggest considering partitions which are optimal according to a penalized Hathaway criterion. The term of penalization should favour homogeneous classes.

Spatial relationships may be summarized in different ways (e.g. graphs or functions). In the following, to formalize the spatial structure of a given data set we use a matrix $\boldsymbol{V} = (v_{ij})$ defined by

$$v_{ij} = \begin{cases} 1 & \text{if } \boldsymbol{x}_i \text{ and } \boldsymbol{x}_j \text{ are neighbors and } i \neq j, \\ 0 & \text{otherwise.} \end{cases}$$

We propose the following term for regularizing the Hathaway criterion:

$$G(\boldsymbol{c}) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{K} c_{ik} c_{jk} v_{ij}, \tag{3}$$

where $K$ is the number of classes and $c_{ik}$ the probability that $\boldsymbol{x}_i$ belongs to class $k$.

Let us denote by $\boldsymbol{c}_i = [c_{i1} \ldots c_{iK}]^{\mathrm{T}}$ the vector which describes the grade of membership of $\boldsymbol{x}_i$ to the different classes. The penalizing term can be rewritten as

$$G(\boldsymbol{c}) = \sum_{i<j} v_{ij} \boldsymbol{c}_i^{\mathrm{T}} \boldsymbol{c}_j. \tag{4}$$

This term will characterize the level of homogeneity of the partition. The more the classes contain adjacent elements, the greater this term is. Let us consider the simple case of a "hard" partition (hard as opposed to fuzzy), where each object belongs to a single class ($c_{ik} = 1$ if $\boldsymbol{x}_i$ belongs to class $k$, and $c_{ik} = 0$ otherwise). In this context, if all the objects belong to the same class, $G$ will be maximal (it will count the number of neighbors) and if each object has neighbors which belong to different classes, then $G(\boldsymbol{c})$ is equal to its minimum, that is 0.

The new criterion that we consider is the weighted sum of two terms:

$$U(\boldsymbol{c}, \boldsymbol{\Phi}) = D(\boldsymbol{c}, \boldsymbol{\Phi}) + \beta G(\boldsymbol{c}), \tag{5}$$

where $\beta > 0$ is a fixed coefficient.

We propose to use an algorithm having the same structure as the EM algorithm in order to optimize the criterion $U(\boldsymbol{c}, \boldsymbol{\Phi})$, that we call *Neighborhood EM algorithm* (NEM):

1. Initialization: a neighborhood matrix $\boldsymbol{V}$ is computed according to the spatial relationship; arbitrary initial values are chosen for the parameters of the mixtures $\Phi^0$ as well as for the classification matrix $\boldsymbol{c}^0$.
2. At each iteration the following steps are performed until convergence:
   - *Estimation* of a new classification matrix $\boldsymbol{c}$ which maximizes $U(\boldsymbol{c}, \Phi^q)$:
   $$\boldsymbol{c}^{q+1} = \arg\max_{\boldsymbol{c}} U(\boldsymbol{c}, \Phi^q). \tag{6}$$
   - *Maximization* of the criterion over the vector of parameters $\Phi$:
   $$\Phi^{q+1} = \arg\max_{\Phi} U(\boldsymbol{c}^{q+1}, \Phi). \tag{7}$$

Note that this step is identical to the M-step of the EM algorithm since the penalization term does not depend upon $\Phi$.

When each step of the algorithm produces a maximum and if $U$ is limited it can easily be shown that the sequence $\{\boldsymbol{c}^q, \Phi^q\}$ generated by the NEM algorithm converges to a limit.

In Section 3, we detail the two steps of the NEM algorithm and show that the suggested solutions allow us to obtain a unique maximum for each step.

## 3. Convergence of NEM

### 3.1. Estimation step

The method proposed in this section to perform the E-step is inspired from the Hathaway (1986) formulation of the EM algorithm and can be also related to the work of Neal and Hinton (1993). We suggest using the fixed point method to find the classification matrix $\boldsymbol{c}^+$ which maximizes the criterion $U(\boldsymbol{c}, \Phi^q)$.

The necessary optimality Kuhn–Tucker conditions take the following form:

$$\left.\frac{\partial \mathscr{U}}{\partial c_{ik}}\right|_{c=c^+} = \log(p_k f_k(\boldsymbol{x}_i|\theta_k)) - 1 - \log c_{ik}^+ + \lambda_i + \beta\sum_{j=1}^{N} c_{jk}^+ v_{ij} = 0 \quad \forall i,k,$$

$$\sum_{k=1}^{K} c_{ik}^+ = 1 \quad \forall i,$$

where $\mathscr{U}$ is the Lagrangian of $U(\boldsymbol{c}, \Phi)$ that takes into account the constraints on $\boldsymbol{c}$ and $\lambda_i$ are the Lagrangian coefficients. These conditions may be written as the $N \times K + N$ equations:

$$c_{ik}^+ = \exp\left\{\log(p_k f_k(\boldsymbol{x}_i|\theta_k)) - 1 + \lambda_i + \beta\sum_{j=1}^{N} c_{jk}^+ v_{ij}\right\} \quad \forall i,k,$$

$$\sum_{k=1}^{K}\exp\left\{\log(p_k f_k(\boldsymbol{x}_i|\theta_k)) - 1 + \lambda_i + \beta\sum_{j=1}^{N} c_{jk}^+ v_{ij}\right\} = 1 \quad \forall i.$$

Finally, eliminating the $\lambda_i$'s we get the $N \times K$ equations:

$$c_{ik}^+ = \frac{p_k f_k(\boldsymbol{x}_i|\theta_k)\exp\{\beta\sum_{j=1}^{N} c_{jk}^+ v_{ij}\}}{\sum_{l=1}^{K} p_l f_l(\boldsymbol{x}_i|\theta_l)\exp\{\beta\sum_{j=1}^{N} c_{jl}^+ v_{ij}\}} \quad \forall i,k, \tag{8}$$

which can be written in the following closed form:

$$\boldsymbol{c}^+ = F(\boldsymbol{c}^+).$$

In order to solve this system the fixed point method can be used: starting from an initial classification matrix $\boldsymbol{c}^q$, a new matrix $\boldsymbol{c}^{m+1} = F(\boldsymbol{c}^m)$ is computed at each step from the preceding matrix $\boldsymbol{c}^m$. If the operator $F$ is contracting, the sequence generated by the method effectively converges to a fixed point.

**Theorem 1.** *If $\beta < 1/V_{\max}$ where $V_{\max} = \max_i \sum_j v_{ij}$ is the maximum number of neighbors of an object, then the sequence $\{\boldsymbol{c}^m\}$ generated by $\boldsymbol{c}^{m+1} = F(\boldsymbol{c}^m)$ converges to a unique fixed point $c^+$ such that*

$$\boldsymbol{c}^+ = \arg\max_{\boldsymbol{c}} U(\boldsymbol{c}, \Phi)$$

*subject to the constraint that $\boldsymbol{c}$ is a classification matrix.*

**Proof.** The proof adopts the following notations:

- $U_\Phi(\boldsymbol{c})$ is the function $U(\boldsymbol{c},\Phi)$ for a fixed value of $\Phi$.
- $\boldsymbol{c}$ denotes the classification matrix expressed as a vector. It is a column vector composed of $N\times K$ elements:

$$\boldsymbol{c}=(c_{11},c_{12},\ldots,c_{1K},c_{21},\ldots,c_{NK})^{\mathrm{T}}.$$

Each element has two indices and $c_{ik}$ denotes the $i\times k$th element of the vector $\boldsymbol{c}$.

- $F$ is a function from $(0,1)^{N\times K}$ to $(0,1)^{N\times K}$:

$$F(\boldsymbol{c})=(f_{11}(\boldsymbol{c}),f_{12}(\boldsymbol{c}),\ldots,f_{1K}(\boldsymbol{c}),f_{21}(\boldsymbol{c}),\ldots,f_{NK}(\boldsymbol{c}))^{\mathrm{T}}.$$

Let us show that $U_\Phi(\boldsymbol{c})$ is strictly concave. The function $U_\Phi$ is continuous with respect to $\boldsymbol{c}$ and its Hessian matrix $H(\boldsymbol{c})$ is defined for any $\boldsymbol{c}\in(0,1)^{N\times K}$:

$$\frac{\partial^2 U_\Phi}{\partial c_{ik}\partial c_{jl}}=\begin{cases}-1/c_{ik} & \text{if } k=l \text{ and } i=j,\\ \beta & \text{if } k=l \text{ and } v_{ij}=1,\\ 0 & \text{otherwise.}\end{cases}$$

The Hessian matrix is a symmetric matrix whose diagonal terms are $-1/c_{ik}$, and whose other terms are sometimes 0, sometimes $\beta$. From the "Gerschgörin–Hadamard Theorem" (Kreyszig, 1962, p. 821), we can state that for each eigenvalue $\lambda$ of $H(\boldsymbol{c})$, we have

$$|\lambda-H_{ik;ik}|\leqslant\sum_{(j,l)\neq(i,k)}|H_{ik;jl}|$$

for some $i,k$. Using the definition of $H$ and the relation $\beta<1/V_{\max}$, we obtain

$$\left|\lambda+\frac{1}{c_{ik}}\right|<1$$

and hence, as $c_{ik}\in[0,1]$,

$$\lambda<0.$$

This proves that the Hessian matrix is strictly negative definite. It follows that the function $U_\Phi(\boldsymbol{c})$ is strictly concave.

Each linear equality constraint $\sum_{k=1}^{K}c_{ik}^{+}=1$ on $\boldsymbol{c}$ can be replaced by two convex inequality constraints:

$$\sum_{k=1}^{K}c_{ik}^{+}\geqslant 1,\qquad \sum_{k=1}^{K}c_{ik}^{+}\leqslant 1.$$

The problem

$$\begin{aligned}&\max\quad U_\Phi(\boldsymbol{c})\\ &\quad\boldsymbol{c}\ \text{is a classification matrix}\end{aligned}$$

is equivalent to

$$\begin{aligned}&\min\quad -U_\Phi(\boldsymbol{c})\\ &\quad\boldsymbol{c}\ \text{is a classification matrix,}\end{aligned}$$

which is a convex program, since $-U_\Phi(\boldsymbol{c})$ and the constraints are convex. It follows that the required Kuhn–Tucker conditions are satisfied and that $\boldsymbol{c}^{+}$ is a global maximum for the problem (Minoux, 1983, Theorem 6, p. 188).

Let us show that the sequence $\{\boldsymbol{c}^m\}$ generated by $\boldsymbol{c}^{m+1}=F(\boldsymbol{c}^m)$ effectively converges to $\boldsymbol{c}^{+}$.

The partial derivatives of $F$ exist and are continuous. From the fixed point theorem, it follows that the sequence converges if there exists a norm $\|\cdot\|$ of the Jacobian matrix of $F$, denoted $F'$, which is strictly less than 1. Consider $F'(\boldsymbol{c}^m)$, the Jacobian matrix of $F$ computed at $\boldsymbol{c}^m$:

$$F'(\boldsymbol{c}^m)=\begin{bmatrix}\frac{\partial f_{11}}{\partial c_{11}^m} & \frac{\partial f_{11}}{\partial c_{12}^m} & \cdots & \frac{\partial f_{11}}{\partial c_{NK}^m}\\ \frac{\partial f_{12}}{\partial c_{11}^m} & \frac{\partial f_{12}}{\partial c_{12}^m} & \cdots & \frac{\partial f_{12}}{\partial c_{NK}^m}\\ \vdots & \vdots & \cdots & \vdots\\ \frac{\partial f_{NK}}{\partial c_{11}^m} & \frac{\partial f_{NK}}{\partial c_{12}^m} & \cdots & \frac{\partial f_{NK}}{\partial c_{NK}^m}\end{bmatrix}.$$

It is possible to show that this Jacobian matrix can be expressed as

$$\frac{\partial f_{ik}}{\partial c_{jl}^m}=\begin{cases}0 & \text{if } k=l \text{ and } i=j,\\ \beta c_{ik}^{m+1}-\beta(c_{ik}^{m+1})^2 & \text{if } k=l \text{ and } v_{ij}=1,\\ -\beta c_{ik}^{m+1}c_{jl}^{m+1} & \text{if } k\neq l \text{ and } v_{ij}=1,\\ 0 & \text{otherwise.}\end{cases}$$

Let us compute the sum of all the absolute values of any row of the matrix $F'(\boldsymbol{c}^m)$:

$$\begin{aligned}\sum_{jl}\left|\frac{\partial f_{ik}}{\partial c_{jl}^m}\right| &= \sum_{j=1}^{N}v_{ij}|\beta c_{ik}^{m+1}(1-c_{ik}^{m+1})|\\ &\quad+\sum_{j=1}^{N}\sum_{l\neq k}v_{ij}|\beta c_{ik}^{m+1}c_{jl}^{m+1}|\\ &< V_{\max}\beta c_{ik}^{m+1}(1-c_{ik}^{m+1}+1)\\ &< V_{\max}\beta c_{ik}^{m+1}(2-c_{ik}^{m+1}).\end{aligned}$$

If $0 < \beta < 1/V_{\max}$, then the value taken by the polynomial $V_{\max}\beta c_{ik}^{m+1}(2 - c_{ik}^{m+1})$ is less than 1, for all possible $c_{ik}^{m+1}$. It follows that

$$\|F'(\boldsymbol{c}^m)\|_\infty = \max_{ik} \sum_{jl} \left| \frac{\partial f_{ik}}{\partial c_{jl}^m} \right| < 1$$

and hence the sequence $\{\boldsymbol{c}^m\}$ converges to $\boldsymbol{c}^+$ if $0 < \beta < 1/V_{\max}$. □

From a practical point of view, the E-step requires only a few fixed point iterations to compute a reasonable classification matrix. The obtained classification matrix $\boldsymbol{c}^+$ is then used for the next M-step of the $(q+1)$th NEM iteration.

### 3.2. Maximization step

This step is identical to the M-step of the EM algorithm for Gaussian finite mixture. It maximizes the criterion $D(\boldsymbol{c}, \Phi)$ with respect to the parameter vector $\Phi$. The necessary optimality conditions lead to the following estimates:

$$\boldsymbol{\mu}_k^{q+1} = \frac{\sum_{i=1}^N c_{ik}^{q+1} \boldsymbol{x}_i}{n_k^{q+1}}, \tag{9}$$

$$\boldsymbol{\Sigma}_k^{q+1} = \frac{1}{n_k^{q+1}} \sum_{k=1}^K \sum_{i=1}^N c_{ik}^{q+1} (\boldsymbol{x}_i - \boldsymbol{\mu}_k^{q+1})(\boldsymbol{x}_i - \boldsymbol{\mu}_k^{q+1})^{\mathrm{T}}, \tag{10}$$

$$p_k^{q+1} = \frac{n_k^{q+1}}{N}, \tag{11}$$

where $n_k^{q+1} = \sum_{i=1}^N c_{ik}^{q+1}$.

Note that $\Phi^+$ may be a singular solution (Duda and Hart, 1973) (if at least one covariance matrix $\boldsymbol{\Sigma}_k$ is negative definite).

## 4. An application to biological images

Let us illustrate the usefulness of the NEM algorithm [1] with an application to image segmentation.

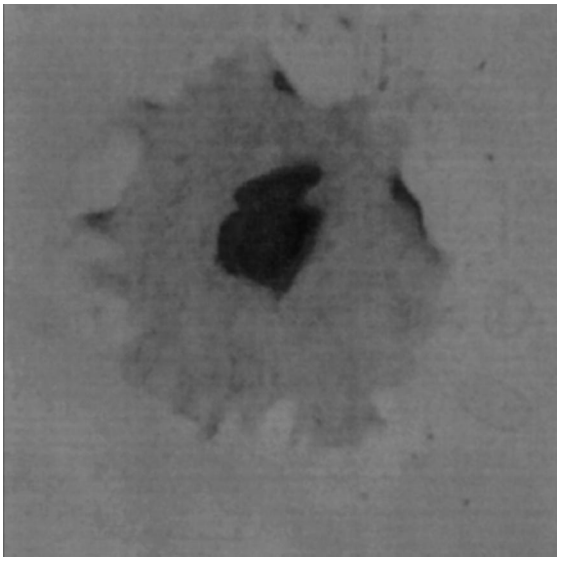

Fig. 1. Image of size 512 × 512 from a cell culture.

Let us consider the following biological experiment: a sample of living cells is laid on a nutritive substance. After a few days new living cells appear and form a thin but visible layer around the original sample. Biologists are interested in determining the surface of the new layer.

We have a collection of 512 × 512 pixel images [2] of such experimental results and present here the analysis of a representative image (Fig. 1).

We aim to distinguish three different kinds of patterns in order to determine automatically the size of the area covered by the new cells. A "good" segmentation from the biologist's point of view should separate the image in three different areas representing:

- the original sample (center of the image);
- the nutritive substance (background of the image);
- the new cells.

We have tested the NEM algorithm with different values of the penalizing coefficient $\beta$. We have considered a very simple neighborhood structure

[1] C code of the NEM algorithm may be found by pointing your World Wide Web browser at http://www.hds.utc.fr/ambroise/Christophe.html.

[2] We are indebted to Eugenio Grapa, from the Université de Technologie de Compiègne, Laboratoire de Biologie Cellulaire Expérimentale, for the biological pictures.

where each pixel has four neighbors, one below, one above, one on the left and one on the right. From Theorem 1 it theoretically follows that this neighborhood structure forces the use of $\beta < 0.25$. As the proof of the theorem uses some approximations, the practical value of $\beta$ may be higher and the NEM algorithm will still converge. Thus we have tested the algorithm with 6 different values between 0 and 2. Notice that when this parameter is 0, the NEM algorithm is equivalent to the EM algorithm for Gaussian mixture.

In these experiments, the algorithms were initialized starting with 20 different random classification matrices. The segmentation which produces the greatest value of the criterion has been selected.

On the original image (Fig. 1), the human vision distinguishes clearly the three classes. In fact, each class is far from having a uniform gray level: some pixels representing the new cells have exactly the same color as the nutritive substance pixels. While the human eye automatically makes the necessary adjustments, the unsupervised segmentation of this kind of image is more difficult than one would think initially.

Fig. 2 shows the result of the obtained segmentation for different values of $\beta$. This result raises the following remarks.

- The EM (NEM with $\beta$ set to zero) algorithm isolates the nutritive substance, but tends to empty the class corresponding to the original sample.
- When $\beta = 0.5$ or $\beta = 1$, the result is very satisfactory and allows the automatic computation of the surface covered by the new cells.
- If we try to run NEM with greater parameter values, the spatial information becomes preponderant and the segmentation does not make a lot of sense. Moreover we have observed that the algorithm did not converge after 100 iterations. Note that this last observation agrees with the sufficient conditions for convergence.

The preceding example showed the practical efficiency of the proposed algorithm on this particular problem. We tested the NEM algorithm on other segmentation tasks and got consistently good results once the $\beta$ parameter was satisfactorily tuned (Ambroise, 1996).

## 5. Concluding remarks

The choice of the penalizing coefficient $\beta$ remains the main difficulty in applying the NEM algorithm. In the preceding example we have used our experience to determine the "optimal" $\beta$ coefficient. When such a procedure is not possible, it would be useful to have an automatic estimation of this parameter. This subject still needs further research.

A particularity of the NEM algorithm consists of providing a fuzzy partition of the data. This may be interesting in some applications where region of doubt should be identified. It should be interesting to compare our approach with other fuzzy segmentation algorithms such as the one proposed by Caillol et al. (1993).

Compared to segmentation algorithms which require Monte Carlo simulations, like Gibbsian EM (Chalmond, 1989), the approach proposed in this paper is deterministic and converges quickly. We have observed empirically that less than 100 NEM iterations are often enough to converge (approximately 5 min on a Sparc Ultra 1).

This algorithm allows us to deal with irregularly distributed spatial data. In that case the computation of a neighborhood graph is the first step of the clustering procedure. This may be done using a Delaunay triangulation for example.

If the class label of some objects is known, the NEM algorithm can take this information into account. Knowing some labels allows us to determine some terms of the classification matrix and during the E-step, only the unknown terms of the classification matrix $\boldsymbol{c}$ are computed. This approach may offer an alternative between supervised and unsupervised learning. It is particularly advantageous when the number of marked objects (pixel whose label is known) is small compared to the number of pixels to be classified:

- In supervised learning the parameter vector $\Phi$ is estimated using a training image (or training set in a more general setting) composed of pixels whose label is known. The other pixels are classified using the estimated values of the parameters. If the size of the learning set is small the estimation of the parameters may be uncertain and the resulting classification quite poor.

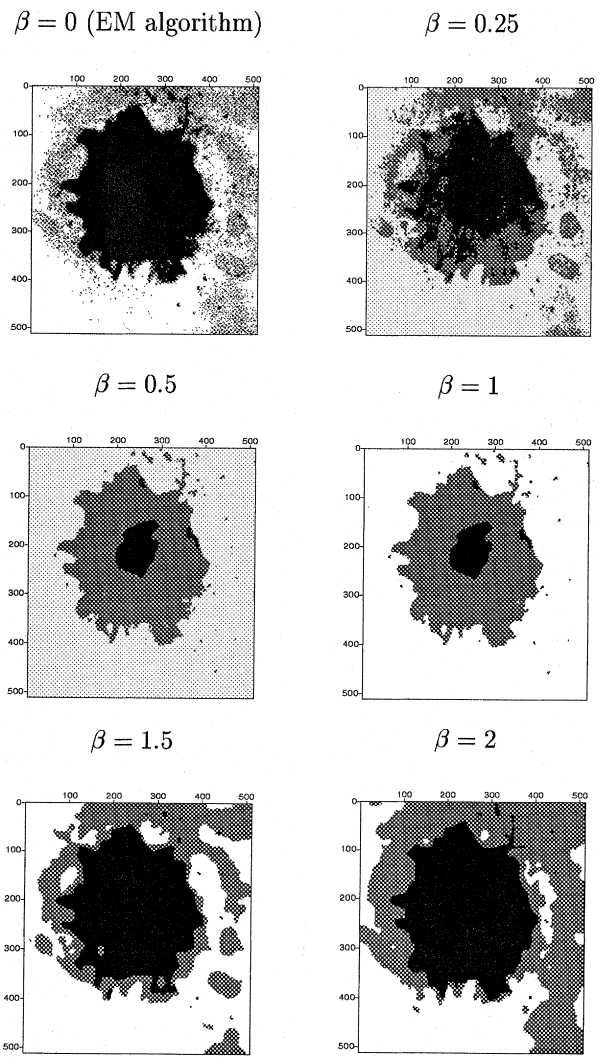


Fig. 2. Segmentation of the image by NEM with different value of the penalization coefficient.

- In unsupervised learning (clustering) there is no learning set and pixels are classified according to some criterion that scales the quality of the partition. All additional information provided by the labeled pixels is not used and hence lost.

In summary, the clustering algorithm proposed in this paper offers a new alternative for spatial clustering. It has been proven that the algorithm locally optimizes a criterion. Further research should thoroughly compare this approach to other spatial clustering approaches.

## References

Ambroise, C., 1996. Approche probabiliste en classification automatique et contraintes de voisinage. Ph.D. Thesis, Université de Technologie de Compiègne, France.

Ambroise, C., Dang, M., Govaert, G., 1996. Clustering of spatial data by the EM algorithm. In: Proceeding of geoENV.

Berry, B.J.L., 1966. Essay on commodity flows and the spatial structure of the Indian economy. Research Paper 111, University of Chicago, Department of Geography.

Caillol, H., Hillion, A., Pieczinsky, W., 1993. Fuzzy random fields and unsupervised image segmentation. IEEE Trans. Geosci. Remote Sensing 31 (4), 801–810.

Celeux, G., Govaert, G., 1995. Gaussian parsimonious clustering models. Pattern Recognition 28, 781–793.

Chalmond, B., 1989. An iterative Gibbsian technique for reconstruction of *m*-ary images. Pattern Recognition 22 (6), 747–761.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Statist. Soc. 39, 1–38.

Duda, R.O., Hart, P.E., 1973. Pattern Classification and Scene Analysis. Wiley, New York.

Geman, S., Geman, D., 1984. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Trans. Pattern Anal. Mach. Intelligence PAMI-6, 721–741.

Hathaway, R.J., 1986. Another interpretation of the EM algorithm for mixture distributions. J. Statist. Probab. Lett. 4, 53–56.

Jain, A.K., Farrokhnia, F., 1991. Unsupervised texture segmentation using Gabor filters. Pattern Recognition 24 (12), 1167–1186.

Kreyszig, E., 1962. Advanced Engineering Mathematics. Wiley, New York.

Lebart, L., 1978. Programme d'agrégation avec contraintes (c.a.h. contiguité). Cahier Analyse Données 3, 275–287.

Legendre, P., 1987. Constrained clustering. In: Developments in Numerical Ecology. NATO ASI Series G 14, pp. 289–307.

Masson, P., Pieczinsky, W., 1993. SEM algorithm and unsupervised statistical segmentation of satellite images. IEEE Trans. Geosci. Remote Sensing 31 (3), 618–633.

Minoux, M., 1983. Programmation Mathématique, Tome 1. Dunod, Paris.

Neal, R.M., Hinton, G.E., 1993. A new view of the EM algorithm that justifies incremental and other variants. Technical Report, Department of Computer Science, University of Toronto, Toronto.

Oliver, M.A., Webster, R., 1989. A geostatistical basis for spatial weighting in multivariate classification. Math. Geol. 21, 15–35.

Openshaw, S., 1977. A geographical solution to scale and aggregation problems in region-building, partitioning and spatial modelling. Trans. Inst. British Geographers 2, 459–472.