ALGORITHME EM ET CLASSIFICATION NON SUPERVISEE

Catherine Aaron
Université Paris I
SAMOS-MATISSE
90 rue de Tolbiac
75013 Paris
Tél : 01 44 07 89 35
E-mail : catherine_aaron@hotmail.com

Résumé :

Dans ce papier on présente une méthode de classification non supervisée inspirée des méthode supervisées bayesiennes reposant sur l'estimation de densité par l'algorithme EM. Dans un premier temps on présentera la démarche sous hypothèse de densité connue du mélange afin d'examiner comment passer de la densité à une classification en s'abstrayant de la problématique d'estimation de densité. Une fois la problématique dégagée on justifie le choix de l'approximation de densité par mélange de gaussienne (contre une estimation par noyau). Puis on présentera l'algorithme de classification final et quelques résultats.

Mots clefs : ***Classification non supervisée, Algorithme EM, mélange de gaussiennes, estimation de densité, watershede***

Abstract :

In this paper we present an unsupervised clustering algorithm based on supervised Bayesian approach. In a first time we suppose known the mixture density to see how to make cluster with this information. Then we show why, in this purpose, the EM gaussian mixture approach is more adapted to estimate density than a kernel on. Finally we present the clustering algorithm and some results.

Key-Words : Unsupervised clustering, EM algorithm, gaussian mixture; density estimation; watershede.

ALGORITHME EM ET CLASSIFICATION NON SUPERVISEE

Résumé :

Dans ce papier on présente une méthode de classification non supervisée inspirée des méthode supervisées bayesiennes reposant sur l'estimation de densité par l'algorithme EM. Dans un premier temps on présentera la démarche sous hypothèse de densité connue du mélange afin d'examiner comment passer de la densité à une classification en s'abstrayant de la problématique d'estimation de densité. Une fois la problématique dégagée on justifie le choix de l'approximation de densité par mélange de gaussienne (contre une estimation par noyau). Puis on présentera l'algorithme de classification final et quelques résultats.

Mots clefs : ***Classification non supervisée, Algorithme EM, mélange de gaussiennes, estimation de densité, watershede.***

Abstract :

In this paper we present an unsupervised clustering algorithm based on supervised Bayesian approach. In a first time we suppose known the mixture density to see how to make cluster with this information. Then we show why, in this purpose, the EM gaussian mixture approach is more adapted to estimate density than a kernel on. Finally we present the clustering algorithm and some results.

Key-Words : Unsupervised clustering, EM algorithm, gaussian mixture; density estimation; watershede.

I- INTRODUCTION : CLASSIFICATIONS NON SUPERVISEE SOUS HYPOTHESE DE DENSITE CONNUE

Dans tout le papier on recherchera une classification non supervisée d'un ensemble de N points $x_1,....,x_N$ de $\mathbb{R}^P$. Les graphiques illustrant la méthodes sont effectués en dimension 1 dans un seul soucis de lisibilité.

I-1 PRINCIPES

Dans le cas, parfaitement idéal ou l'on connaît le mélange de lois ayant donné lieu au tirage des données, le problème de classification se résout classiquement de manière bayesienne en affectant à chaque point à la classe la plus probable. Tout est alors connu, aussi bien le nombre de classes que la segmentation associée et le pourcentage d'erreur de classification. Malheureusement la connaissance de telles données est, dans la réalité fort improbable.

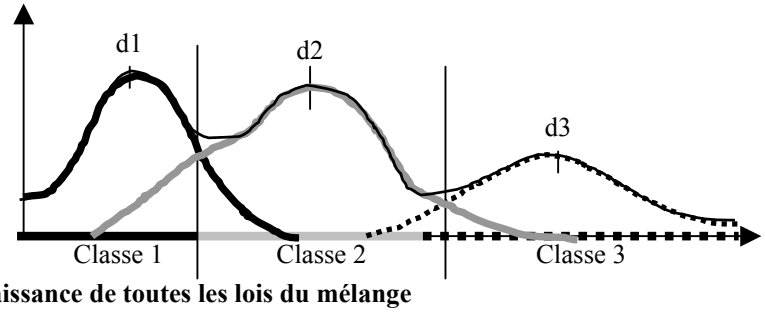

Figure 1 : Cas de la connaissance de toutes les lois du mélange

Dans le cas, toujours idéal, où la densité f, somme pondérée des composantes du mélange, associée au tirage des x_i est connue, plusieurs méthodes de classifications ont été proposées. Une méthode de classification autour des modes a été introduite par Wishart's en 1969 ou l'on considère qu'il y a autant de classes que de modes à la densité.

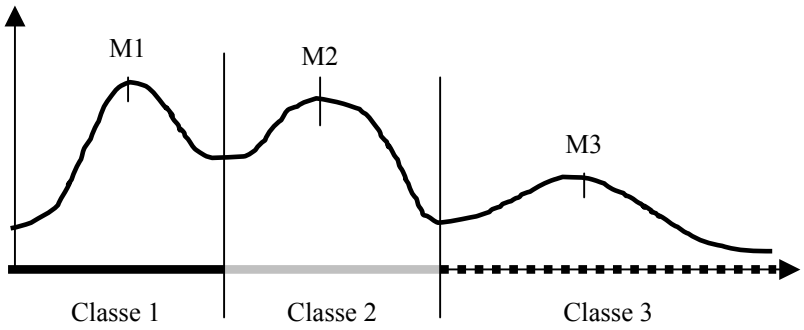

Figure 2 : regroupement autour des modes de la densité

Une autre méthode dite watershed consiste a « immerger » l'espace et a compter le nombre « d'îles », de manière plus formalisées on compte le nombre de composantes connexes aux ensembles $E_\lambda = \{x / f(x) \geq \lambda\}$ dans le cas précédents

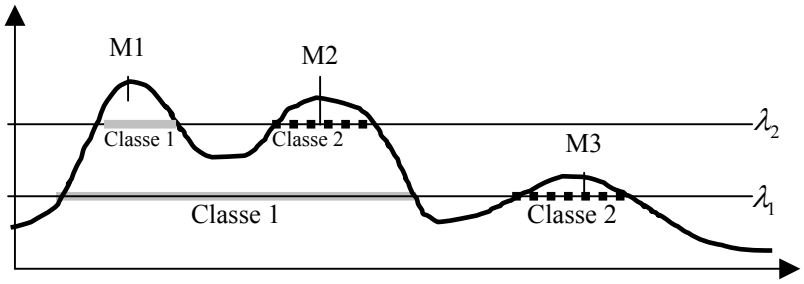

Figure 3 : méthode dite du watershed

On voit alors, sur l'exemple précédent que la méthode de regroupement autour des modes se révèle plus pertinente que la méthode du watershed, d'une part car elle n'impose pas de choix de paramètre, et, d'autre part les trois classes peuvent être obtenues simultanément alors que ce n'est pas le cas pour la seconde méthode.

En outre, d'un point de vue algorithmique la recherche du nombre de composantes connexes a des ensembles type $E_\lambda = \{x / f(x) \geq \lambda\}$, usuellement faites en effectuant un pavage de l'espace, nécessite un temps de calcul énorme pour peu qu'on soit dans des espaces de dimension élevé.

Ces deux points ont motivé notre choix du travail sur les modes de la densité.

I-2 ALGORITHME DE CLASSIFICATION

On travaille ici toujours sous hypothèse de densité f connue à laquelle on rajoute l'hypothèse que les modes sont tous discrets. L'obtention du nombre de classe et des classes associée est alors simultanée et découle d'un algorithme en deux étapes : Dans un premier temps on recherche l'ensemble des maxima locaux a une erreur près par des montées de

gradients. Dans un second temps on regroupe les maxima locaux proches pour obtenir a la fois classification et classes

1- Obtention des modes
Pour chaque point x_i de la base

Montée de gradient sur f, de pas ε fixe, au départ de x_i

Stop quand descente

On obtient un nouveau point m_i situé a une distance d'au plus ε du mode

Pour chaque couple m_i, m_j

Si $d(m_i, m_j) \leq \varepsilon$ alors m_i et m_j correspondent au même mode

2- Classification
Pour chaque point x_i de la base, x_i appartient à la classe mode de m_i

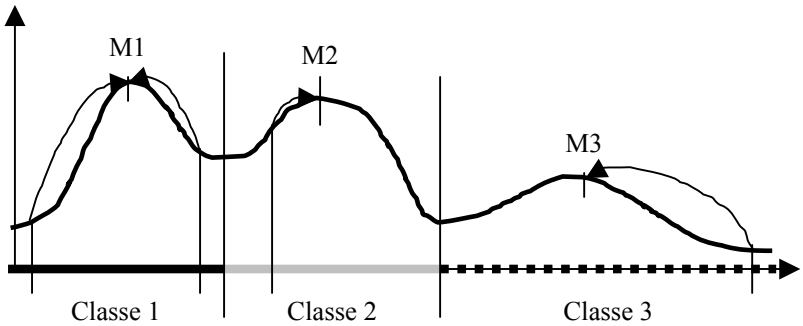

Figure 4 : affectation des points aux classes

Paramétrage de ε :
Le choix de ε est relativement important dans la mesure ou une petite valeur de ce paramètre va accroître le temps de calcul de la montée de gradient. A contrario une trop grande valeur de ce dernier risque de faire « sauter » un mode.
En partant de l'hypothèse qu'il n'y a pas de classes « singleton » on en déduit que chaque point est dans la même classe que son plus proche voisin et on choisit, en pratique :
$\varepsilon = \max_i(d(x_i, x_{\delta(i)}))$ avec $x_{\delta(i)}$ plus proche voisin de x_i.

II- METHODES D'ESTIMATION DE DENSITE

I-1 METHODES A NOYAUX

On recherche une estimation de f comme somme de noyaux centrés sur les observation $\hat{f}(x) = \frac{1}{N}\sum_{i=1}^{N} K(x, x_i, h)$ avec

$\int K(x, m, h)dx = 1$ et

$\forall x, K(x, m, h) \geq 0$

Toute la problématique revient dans le choix de h. Dans la pratique on cherche a minimiser des critère de distance a la densité estimée et la vraie densité. Non seulement pour de telles méthode la problématique est encore ouverte pour le choix d'un h « optimal » mais la

minimisation d'un critère de distance induit, dans la pratique des irrégularités dans la densité estimée qui risque de multiplier le nombre de modes observés

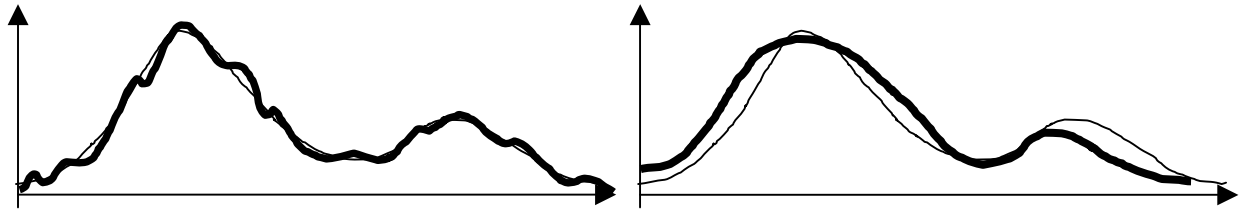

Figure 5 : Dans le premier cas, type estimation à noyau, on aurait 11 classes alors que dans le second cas, pourtant moins bon en estimation de densité, la classification induite est largement meilleure.

En outre, et peut être même plus important encore, les méthodes à noyaux supposent que les données sont tirées suivant une loi (et non suivant un mélange de lois), ainsi dans le cas de mélanges de lois très hétérogènes en terme de dispersion la recherche d'une taille de fenêtre optimale pour chacune des composante est utopique.

II-2 ALGORITHME EM

On suppose que la densité est issue d'un mélange de gaussiennes. C'est a dire que : $f(x)=\sum_{i=1}^{K} p_i \varphi(x,\overrightarrow{m_i},S_i)$. On va alors estimé les paramètres $\left(p_i,\overrightarrow{m_i},S_i\right)$ en maximisant la vraisemblance du tirage. Pour cela on doit partir d'un nombre de gaussienne K fixé a priori puis on recherche un maximum local par les conditions du premier ordre de gradient nul :

$$\begin{cases} p_i = \dfrac{1}{N}\sum_{k=1}^{N} P(C_i|x_k) \\ \overrightarrow{m_i} = E(x|C_i) \\ S_i = V(x|C_i) \end{cases} \quad \text{avec} \quad \begin{cases} P(C_i|x_k) \text{ probabilité d'etre dans la gaussienne } i \text{ sachant } x_k \\ E(x|C_i) \text{ espérance conditionnelle de } x \text{ sachant la gaussienne } i \\ S_i = V(x|C_i) \text{ variance conditionnelle de } x \text{ sachant la gaussienne } i \end{cases}$$

On procède de manière itérative en effectuant :

$$\begin{cases} p_i(t+1) = \dfrac{1}{N}\sum_{k=1}^{N} P(C_i|x_k)(t) \\ \overrightarrow{m_i}(t+1) = E(x|C_i)(t) \\ S_i(t+1) = V(x|C_i)(t) \end{cases}$$

On obtient ainsi des fonctions relativement « lisses » et, en particulier, on sait qu'on obtiendra au plus K modes a la densité estimée. Le choix d'un nombre de gaussiennes optimal est, comme celui de la taille des noyau, un problème ouvert sur lequel un grand nombre de travaux sont publiés et de recherches en cours mais, dans le cas qui nous intéresse, c'est a dire l'application en classification on pourra choisir K comme un nombre de classe maximal a observer.

III ALGORITHME EM ET CLASSIFICATION NON SUPERVISEE

III-1 NOMBRE DE CLASSES ET CLASSIFICATION ASSOCIEE

On se fixe un nombre K de gaussiennes pour l'estimation du mélange de lois. L'algorithme EM nous fournis alors K triplets poids, espérance et variances associées. On sait ainsi qu'il y aura au maximum K modes a la densité totale du mélange qui seront, intuitivement, assez peu éloigné des espérances. Cette considération nous permet d'obtenir le nombre de modes, et donc de classes plus rapidement que dans l'algorithme décrit en I. En effet nous allons nous limiter a observer le résultat d'une montée de gradient au départ de chacune des espérance des gaussiennes.
Au total nous obtenons $k \leq K$ modes définissant les classes.
Ces classes regroupent les gaussiennes pour lesquelles les montées de gradient ont donné le même résultat. Si on considère que la densité estimée $\widehat{f}$ du mélange est la somme pondérée des k densités construites comme sommes pondérée des gaussiennes menant au même mode :

$$\widehat{f}(x) = \sum_{i=1}^{k} \underbrace{\sum_{\mathrm{mod}(m_j)=i} p_j \varphi(x, m_j, S_j)}_{q_i \widehat{f_i}(x)} \text{ avec } q_i = \sum_{\mathrm{mod}(m_j)=i} p_j, \text{ et } \widehat{f_j}(x) = \frac{1}{q_j} \sum_{\mathrm{mod}(m_j)=i} p_j \varphi(x, m_j, S_j)$$

Alors on peut procéder à la segmentation des points de manière bayesienne en les affectant à la classe la plus probable.

Remarque : L'affectation de type « bayesienne » des points aux classes donne des résultats légèrement moins bons que ceux issus d'une montée de gradient mais le gain de temps procuré par cette méthode est tel qu'on la préfère. Ce problème est illustré par la figure suivante :

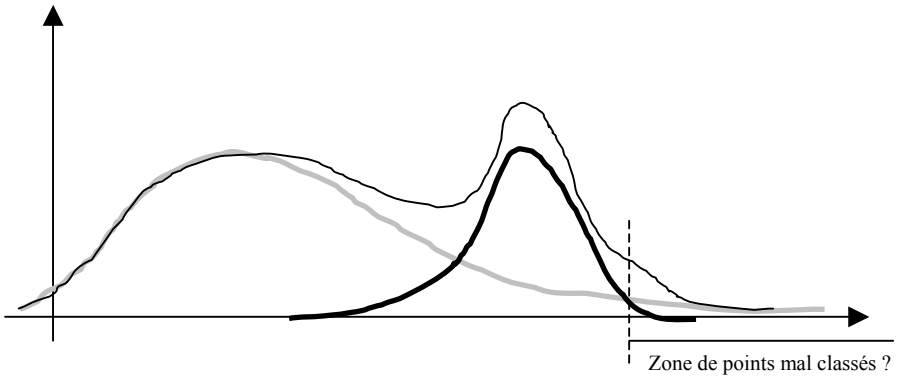

Figure 6 : Cas de défaut de classement de la méthode bayesienne contre un regroupement sur les modes

III-2 RESULTATS

L'algorithme EM est stochastique (en fait ici c'est principalement l'initialisation de l'algorithme qui est sujette a des aléas) et converge vers un maximum local de vraisemblance, les résultats présentés ont été sélectionnés, après plusieurs modélisations en choisissant le modèle de vraisemblance maximale.

Les Iris de Fischer (travail sur les deux composantes principales d'une ACP)

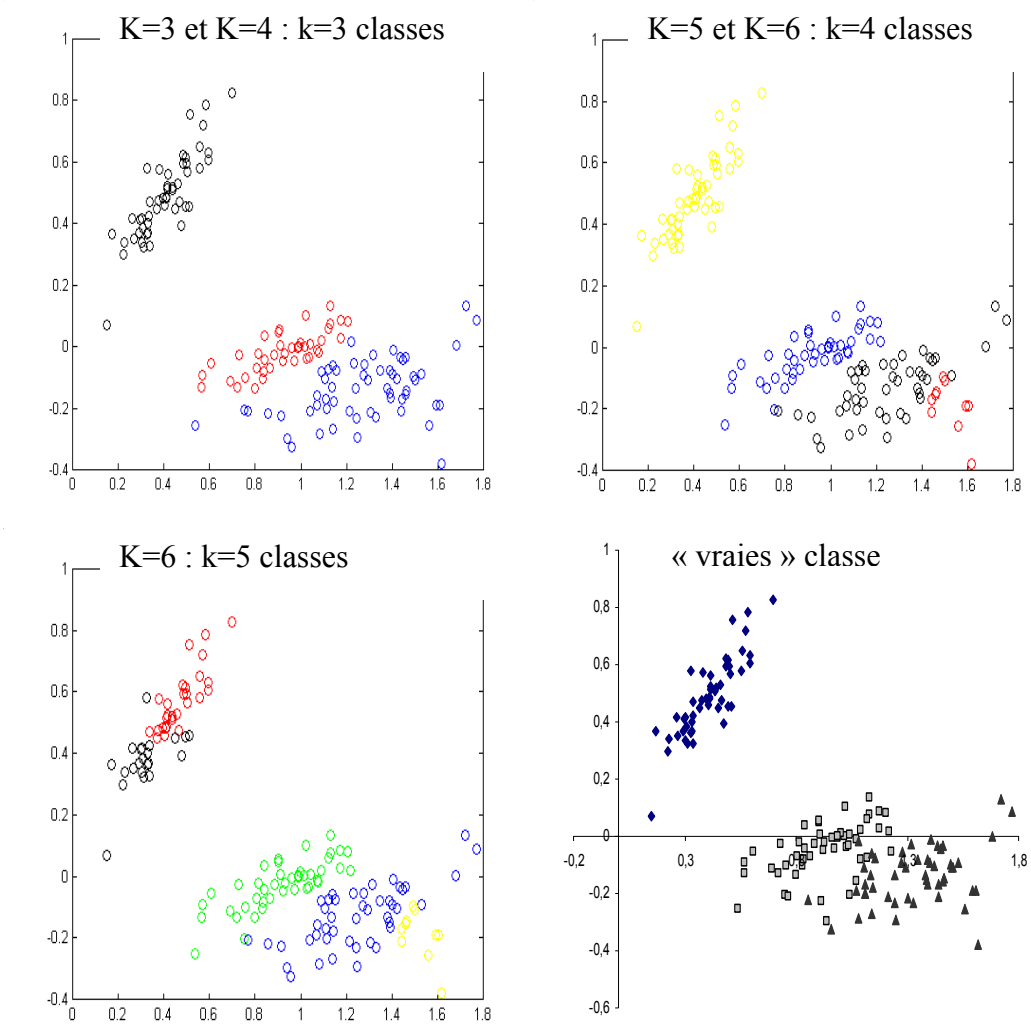

Figure 7 : Résultats sur les Iris de Fisher pour K={3,4,5,6}

Données simulées.

On a simulé des données particulièrement difficile à classer, de manière non supervisée, pour des méthodes « classiques » d'analyse des données. Les données sont le résultat d'un mélange d'une loi gaussienne centrée à l'origine et d'une « couronne » dont le rayon est tiré sur une gaussienne et l'angle résulte d'un tirage uniforme sur $[0, 2\pi]$. Notons en préliminaire qu'un tel tirage ne satisfait pas aux hypothèses effectuées. En effet pour la seconde classe la densité du tirage est de mode continu et non discret.

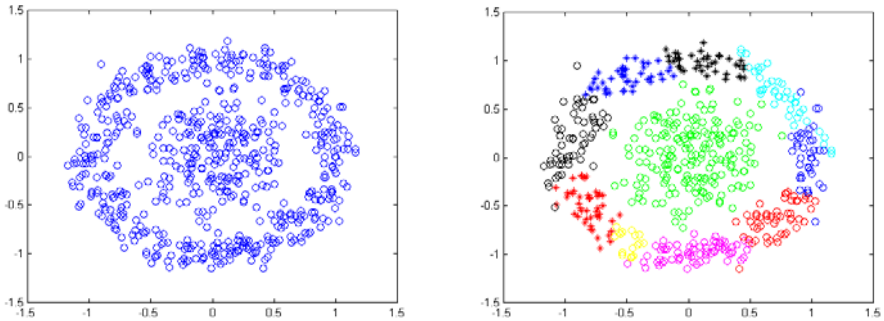

Figure 8 : Résultats pour K=10

III-2 REGROUPEMENT DES CLASSES EN SUPER-CLASSES

Les résultats présentés plus haut sont relativement encourageant mais montrent que la méthode doit être améliorée. En effet sans idée a priori sur le nombre de classes on voit, dans le cas des Iris de Fischer, qu'un choix du paramètre K trop grand, risque d'introduire trop de modes, donc de classes, soit mener aux défauts imputés, a priori, à l'estimation de densité par une méthode à noyau.
L'hypothèse de densité de modes discret induit elle aussi scissions de classes non pertinentes.

Ces deux points nous mène a vouloir regrouper les classes a l'issue de la classification. Pour cela il nous faut définir une matrice de dissimilarité entre densités. Deux pistes ont été poursuivies :

Travailler sur la probabilité d'erreur de classification a posteriori sur les densités deux a deux :

$$d_1(C_i, C_j) = 1 - \frac{1}{q_j + q_i}\left(\int 1_{\{q_i f_i(x) > q_j f_j(x)\}} q_j f_j(x)dx + \int 1_{\{q_i f_i(x) < q_j f_j(x)\}} q_i f_i(x)dx\right)$$

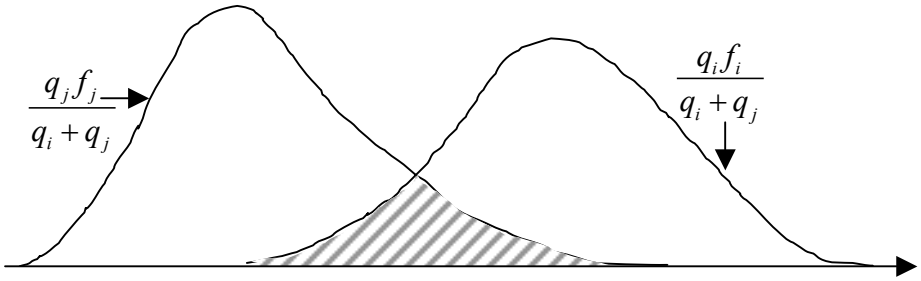

Figure 9 : la distance entre la classe i et la classe j correspond à 1 moins la zone hachurée

Se ramener, pour agréger les classes à une méthode type watershed c'est a dire agréger en fonction de la différence de hauteur entre le plus bas des maxima locaux et le point selle (minimum local en dimension 1) compris entre les deux densités ou encore l'inverse de la hauteur du point selle

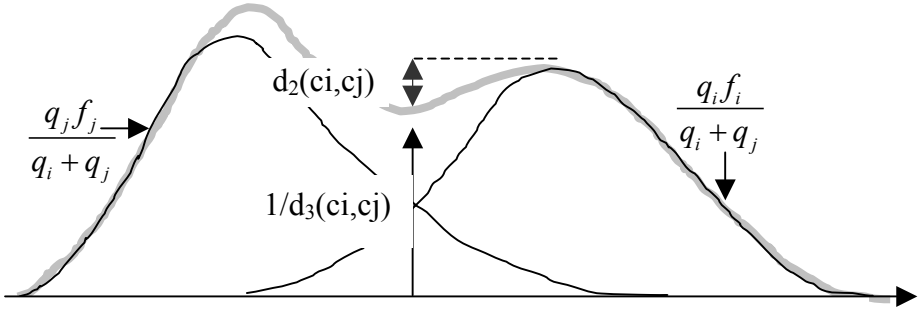

Figure 10 : Distances entre deux classes inspirée de la méthode watershed

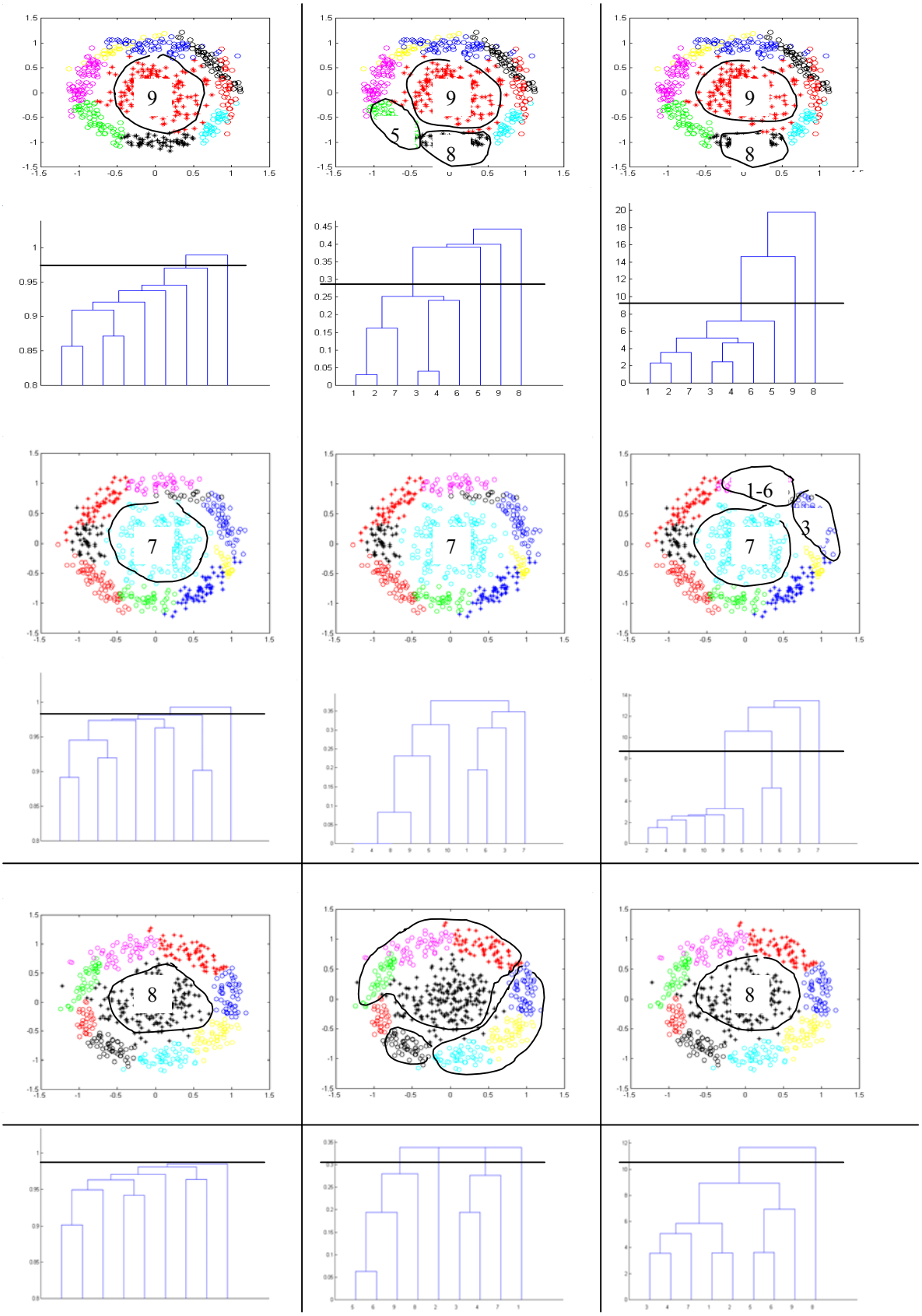

Figure 11 : Regroupement, sur plusieurs simulations suivant les trois méthodes proposées (première colonne distance en probabilité d'être bien classé, seconde différence de hauteur entre les sommets et les points selle et troisième inverse de la hauteur des points selle)

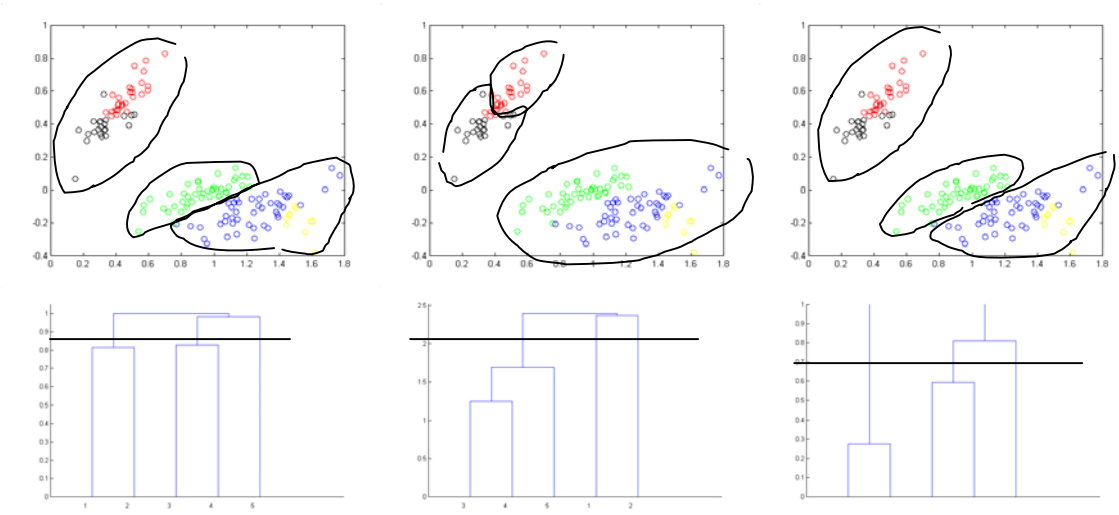

Figure 12 : Regroupement, sur les Iris de Fischer

En conclusion c'est la première méthode qui permet le mieux de retrouver les classes dans tous les cas, c'est aussi la plus intuitive en terme de probabilité et de distance entre des courbes. On retrouve toujours dans l'ordre la hiérarchie intuitive mais il faut l'améliorer car le choix de la scission sur l'arbre ne se voit jamais nettement.

La troisième distance de regroupement fonctionne encore correctement et une telle méthode, est très similaire au watershede mais présente une accélération de l'algorithme notoire.

IV APPLICATION A DES DONNEES REELLES : CLASIFICATION DES PAYS SUR LE PIB

Etant donnée que les méthodes de regroupement ne sont pas encore suffisamment développées pour pouvoir faire des choix de regroupement de classes nous nous sommes limités a la segmentation de données sur une variable ce qui nous permet de visualiser l'estimation de densité et d'en déduire « à l'œil » les éventuels regroupements entre classes.

Dans un premier temps nous avons effectué une segmentation des pays en fonction de leur produit Intérieur Brut par habitant (PIB). Il a été choisit de travailler sur le logarithme du PIB pour réduire la dispersion de se dernier qui empêchait la convergence de l'algorithme. Une estimation de la densité par l'algorithme EM avec 15 centres mène à une densité estimée présentant 11 modes ou classes.

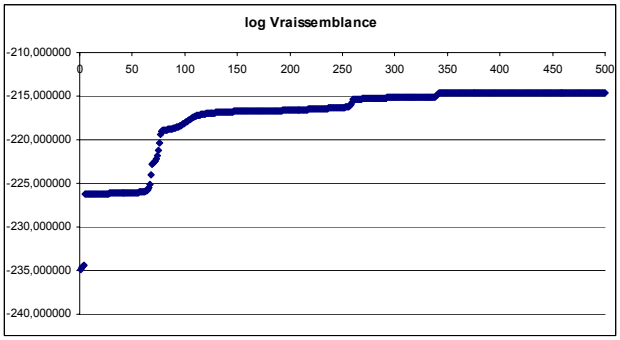

Figure 13 : Convergence de la vraisemblance en fonction des itérations

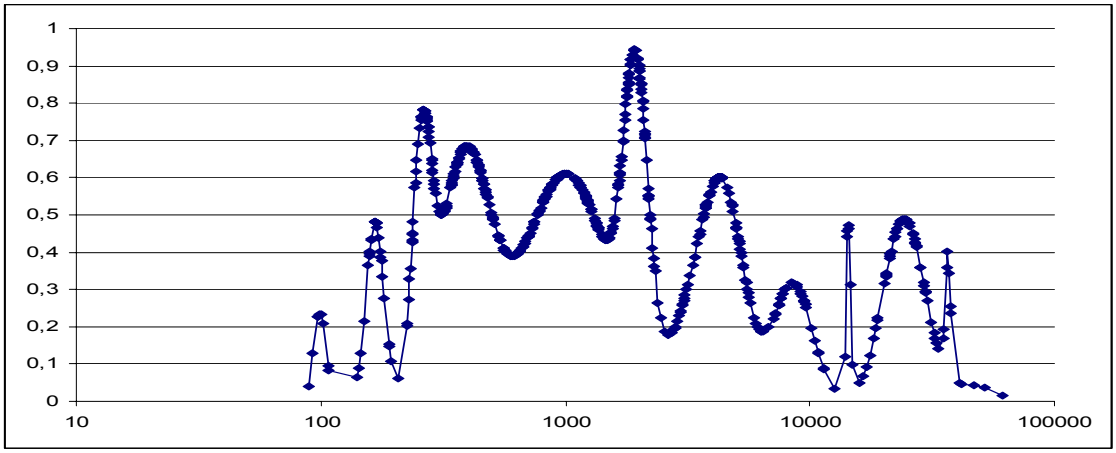

Figure 14 : Densité estimée de log(PIB/hab)

Par rapport à la classification OCDE on observe une segmentation en 4 des pays les moins développés (environ les 4 premiers modes de la densité) et en 3 des pays les plus riches, les autres classes (« lower middle income », « middle income » et « upper middle income ») sont quasiment identiques. La classification pour les pays en 2002 est présentée dans le tableau 1.

V CONCLUSION PERSPECTIVE

La méthode proposée donne des résultats encourageants si ils sont loin d'être parfaits. Pour améliorer ces résultats il faut, a notre sens, se focaliser sur deux points : d'une part sur l'algorithme EM en lui même et notamment se pencher sur la question du nombre de gaussiennes a considérer pour le mélange. Cette question est actuellement largement abordée dans la littérature mais les résultats apportés nécessitent des hypothèses trop restrictive pour notre application.

Un autre point serait de trouver une fonction de la distance d_1 entre gaussienne permettant d'avoir des dendrogrames de super-classes permettant de choisir un regroupement. En effet dans la partie 3 que c'était cette distance qui induisait les regroupement hiérarchiques les plus pertinents mais la lecture des regroupements sur le dendrograme est tout sauf naturelle. Le problème est alors de trouver une fonction de d_1 permettant de choisir un regroupement de la manière « classique ».

Congo, Dem. Rep.	1
Burundi	1
Tajikistan	2
Niger	2
Malawi	2
Liberia	2
Guinea-Bissau	2
Eritrea	2
Uganda	3
Togo	3
Tanzania	3
Nepal	3
Mali	3
Madagascar	3
Least developed countries: UN classification	3
Ghana	3
Gambia, The	3
Central African Republic	3
Cambodia	3
Burkina Faso	3
Zambia	4
Yemen, Rep.	4
Uzbekistan	4
Sudan	4
Sub-Saharan Africa	4
South Asia	4
Senegal	4
Sao Tome and Principe	4
Mongolia	4
Mauritania	4
Kyrgyz Republic	4
Kenya	4
India	4
Guinea	4
Comoros	4
Cameroon	4
Bangladesh	4
Zimbabwe	5
Ukraine	5
Syrian Arab Republic	5
Swaziland	5
Sri Lanka	5
Philippines	5
Paraguay	5
Nicaragua	5
Namibia	5
Morocco	5
Lower middle income	5
Indonesia	5
Honduras	5
Guyana	5
Georgia	5
Egypt, Arab Rep.	5
East Asia & Pacific	5
Cote d'Ivoire	5
Congo, Rep.	5
China	5
Bosnia and Herzegovina	5
Bolivia	5
Belarus	5
Angola	5
Tunisia	6
Suriname	6
Russian Federation	6
Romania	6
Peru	6
Middle income	6
Maldives	6
Macedonia, FYR	6
Kazakhstan	6
Jordan	6
Guatemala	6
El Salvador	6
Ecuador	6
Colombia	6
Bulgaria	6
Albania	6
Venezuela, RB	7
Uruguay	7
Upper middle income	7
Turkey	7
St. Lucia	7
Poland	7
Panama	7
Lebanon	7
Latin America & Caribbean	7
Jamaica	7
Hungary	7
Gabon	7
Estonia	7
Costa Rica	7
Chile	7
Botswana	7
Belize	7
Argentina	7
Trinidad and Tobago	8
St. Kitts and Nevis	8
Seychelles	8
Saudi Arabia	8
Korea, Rep.	8
Czech Republic	8
Barbados	8
Antigua and Barbuda	8
Spain	9
United Kingdom	10
Sweden	10
Singapore	10
Mayotte	10
Italy	10
Hong Kong, China	10
Germany	10
Finland	10
Denmark	10
Belgium	10
Austria	10
Norway	11
Monaco	11

Tableau 1 : Classification des Pays

V BIBLIOGRAPHIE

Archambeau C., Vrins F., Verleysen M. (2001), Flexible and Robust Bayesian Classification by finite Mixture Models., Proceedings of the ESANN'2004 congress.

Bicego M.,Cristani M., Fusiello A., Murino V. (2003), Watershed-based unsupervised clustering Document de travail
http://profs.sci.univr.it/~bicego/bicego_murino_emmcvpr03.pdf.

Everitt B., Landau S.,Leese M. (2001), *Cluster analysis* (1996), Arnold Edition, Londres.

Friedman N, the bayesian structural EM algorithm, Document de travail,
http://www.cs.huji.ac.il/~nir/Papers/Fr2.pdf

Michalis K Titsias., Aristidis C Likas., (2001), Shared Kernel Models for Class Conditional Density Estimation., IEE Transaction on Neural Network Vol 12 N°5