Estimation de modèles de mélanges finis par un algorithme EM crédibiliste

Finite mixture models estimation with a credal EM algorithm

Patrick Vannoorenberghe

Laboratoire de Télédétection à Haute résolution, ERT43
Université Paul Sabatier, Toulouse 3
Centre de Télédétection Spatiale, 118 route de Narbonne
31062 Toulouse cedex 4 – France

Manuscrit reçu le 11 mai 2006

Résumé et mots clés

Dans cet article, l'estimation d'un modèle de mélange fini est abordée dans le cadre du Modèle des Croyances Transférables (MCT). Ce modèle constitue le socle d'un formalisme non probabiliste pour la représentation d'informations imprécises et incertaines par des fonctions de croyances. Dans ce contexte, un algorithme EM crédibiliste, une extension de l'algorithme EM aux fonctions de croyance, est introduit pour l'apprentissage des paramètres du modèle de mélange fini. Nous montrons comment cet algorithme peut être appliqué dans plusieurs contextes où l'information sur le modèle de génération des données n'est que partiellement disponible. Cette information est représentée, dans le problème d'apprentissage, par des fonctions de croyance qui permettent de modéliser la connaissance disponible sur la composante ayant servie à générer chaque observation de manière la plus fine possible. Plusieurs simulations mettent en évidence des situations où le modèle de génération des données n'est connu que de manière imprécise (apprentissage partiellement supervisé) et où l'on ne posséde auncune information sur la composante d'appartenance de chaque observation (apprentissage non supervisé). Des jeux de données synthétiques permettent de démontrer les bonnes performances de l'approche proposée en terme d'estimation mais également en terme d'apprentissage sur des modèles de mélanges gaussiens.

Modèle de mélange fini, Modèle des croyances transférables, Algorithme EM, Apprentissage.

Abstract and key words

This paper is concerned with finite mixture models estimation in the framework of Transferable Belief Model. This model relies on a non probabilistic formalism for representing and manipulating imprecise and uncertain information with belief functions. Within this framework, a credal EM algorithm, a variant of classical EM algorithm based on belief functions, is introduced for finite mixture parameters learning. This algorithm can be applied in several situations where available information on the data generation model is partially known. In the learning problem, this knowledge is represented with belief functions which allow to represent as better as possible the uncertainty on the component from where each observation has been generated. Several experimentations highlight situations where the algorithm is applied when available information on the learning set is imprecise (partially supervised learning where the actual component of each sample is only known as belonging to a subset of components), and/or uncertain (unsupervised learning where the

knowledge about the actual sample is represented by a belief function). Synthetic data sets allow us to demonstrate the good performance of the proposed approach based on estimated parameters analysis and learning with gaussian finite mixture models.

Finite mixture models, Transferable Belief Model, EM algorithm, Learning.

1. Introduction

Les modèles de mélanges constituent de nos jours un outil d'analyse de données multi-dimensionnelles dont l'utilité dans les domaines qui nécéssitent la modélisation statistique est désormais couramment admise [5, 6]. Dans le domaine de l'apprentissage par exemple, les modèles de mélanges ont permis d'aborder de manière formelle des problèmes de classification non supervisée [7, 4]. Dans ce contexte, ils permettent de modéliser de façon relativement naturelle des observations $\boldsymbol{Y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$ qui auraient été générées par une source k (choisie aléatoirement ou inconnue) d'un ensemble de K sources aléatoires potentielles. Identifier la source (ou ses paramètres associés) qui a permis de générer chacune des observations dans $\boldsymbol{Y}$ constitue un clustering de l'ensemble des observations.

Les modèles de mélanges finis ne se limitent pas simplement au cas non supervisé mais permettent également de représenter des densités de probabilité complexes. Dans ce contexte, on dispose généralement d'un ensemble $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ composé de n observations où chaque exemple $\boldsymbol{x}_i = (\boldsymbol{y}_i, \boldsymbol{z}_i)$ dans $\boldsymbol{X}$ est représenté par son observation $\boldsymbol{y}_i$ à valeur dans Y et son label de classe correspondant $\boldsymbol{z}_i$ à valeur dans $\Omega = \{\omega_1, \omega_2, \ldots, \omega_K\}$, un ensemble fini de K classes possibles. L'apprentissage supervisé consiste à construire une règle de décision qui permet d'assigner à une observation inconnue $\boldsymbol{y}$ une classe dans Ω. Quand le modèle de génération des données est connu, les techniques classiques d'analyse discriminate permettent l'estimation des paramètres du modèle de manière relativement simple. Dans le cas contraire, l'algorithme EM est largement utilisé pour obtenir une convergence vers l'estimation des paramètres par maximum de vraisemblance [8] à partir des données observées.

Dans cet article, nous nous limitons au cas des mélanges gaussiens qui sont les plus couramment utilisés en classification car ils correspondent souvent à la loi de distribution des variables observées. De plus, les mélanges gaussiens sont relativement bien maîtrisés et il existe des procédures efficaces pour les manipuler. Dans [15], l'estimation des paramètres de ce type de mélanges fini est généralisé dans le cadre du Modèle des Croyances Transférables (MCT). L'algorithme proposé, inspiré fortement de l'algorithme EM pour l'estimation de paramètres, peut être appliqué dans plusieurs contextes où l'information sur le modèle de génération des données n'est que partiellement disponible. Cette information est représentée, dans le problème d'apprentissage, par des fonctions de croyance qui permettent de modéliser la connaissance disponible sur la composante ayant servie à générer chaque observation de manière la plus fine possible. La section 2 présente les principales notions liées aux fonctions de croyance qui seront nécessaires à la compréhension de cet article. L'estimation des paramétres d'un modèle de mélange fini par l'algorithme EM est abordée dans la section 3. Nous présenterons l'extension de cet algorithme dans un cadre purement crédibiliste à la section 4. Quelques données synthétiques permettant de mettre en évidence l'algorithme proposé sur des modèles de mélanges gaussiens seront enfin présentées à la section 5.

2. Fonctions de croyance

Soit $\Omega = \{\omega_1, \cdots, \omega_k, \cdots, \omega_K\}$ un ensemble fini, généralement appelé cadre de discernement. Une fonction de croyance peut être définie mathématiquement par une fonction (ou allocation) de masse, notée m^Ω définie de 2^Ω dans $[0, 1]$, qui vérifie :

$$\sum_{\emptyset \neq A \subseteq \Omega} m^\Omega(A) = 1. \tag{1}$$

Chaque sous-ensemble $A \subseteq \Omega$ tel que $m^\Omega(A) > 0$ est appelé élément focal de m^Ω. Ainsi, la masse $m^\Omega(A)$ représente le degré de croyance attribué à la proposition A et qui n'a pas pu, compte tenu de l'état de la connaissance, être affectée à un sous-ensemble plus spécifique que A. Une fonction telle que $m^\Omega(\emptyset) = 0$ est dite normale. Dans le modèle des croyances transférables (MCT), la condition (1) n'est pas supposée et $m^\Omega(\emptyset) > 0$ est acceptée.

Définition 2.1. (Fonction de masse catégorique) Une fonction de masse catégorique focalisé sur le sous-ensemble $B \subseteq \Omega$ est définie telle que :

$$m_B^\Omega(A) = \begin{cases} 1 \text{ si } A = B \\ 0 \text{ sinon.} \end{cases}$$

Definition 2.2. (Fonction de masse certaine) *Une fonction de masse certaine est une fonction de masse catégorique m_B^Ω telle que son élément focal B est un singleton $|B| = 1$.*

Définition 2.3. (Fonction de masse vide) *La fonction de masse vide, notée m_v^Ω, est une fonction de masse catégorique dont l'élément focal est l'ensemble Ω lui même. On a $m_v^\Omega = m_\Omega^\Omega$.*

Définition 2.4 (Fonction de masse bayesienne) *Une fonction de masse bayesienne sur* Ω *est une fonction de masse dont les éléments focaux sont des singletons de* Ω.

Étant donnée une fonction de masse m^Ω, une fonction de croyance bel^Ω et une fonction de plausibilité pl^Ω peuvent être définies respectivement par :

$$bel^\Omega(A) = \sum_{\emptyset \neq B \subseteq \Omega} m^\Omega(B), \ \forall\, A \subseteq \Omega. \tag{2}$$

$$pl^\Omega(A) = \sum_{A \cap B \neq \emptyset} m^\Omega(B), \ \forall\, A \subseteq \Omega. \tag{3}$$

Les fonctions m^Ω, bel^Ω et pl^Ω représentent trois facettes de la même information[1]. De plus, on peut retrouver l'une des trois fonctions à partir de l'une des deux autres en utilisant la transformée de Möbius par simples calculs matriciels [12]. Issues des travaux de G. Shafer [10], les fonctions de croyance sont de nos jours reconnues pour la modélisation d'informations incertaines.

Soient deux jeux de masse m_1^Ω et m_2^Ω définis sur le même référentiel Ω. Ces deux fonctions peuvent être agrégées par un opérateur de combinaison conjonctif noté $\textcircled{\cap}$. Le résultat de cette opération conduit à une fonction de croyance unique à laquelle correspond une fonction de masse, notée $m^\Omega_{\textcircled{\cap}}$, qui peut être définie par :

$$m^\Omega_{\textcircled{\cap}}(A) = (m_1^\Omega \,\textcircled{\cap}\, m_2^\Omega)(A) \triangleq \sum_{B \cap C = A} m_1^\Omega(B)\, m_2^\Omega(C) \ \forall \quad A \subseteq \Omega. \tag{4}$$

Cette règle conjonctive est appelée règle de combinaison de Dempster non normalisée. Si nécessaire, l'hypothèse de normalisation $m^\Omega_{\textcircled{\cap}}(\emptyset) = 0$ peut être retrouvée en divisant chaque masse par un coefficient adéquat. D'autres règles de combinaison ont également été proposées [11], comme la règle de combinaison disjonctive qui s'écrit :

$$m^\Omega_{\textcircled{\cup}}(A) = (m_1^\Omega \,\textcircled{\cup}\, m_2^\Omega)(A) \triangleq \sum_{B \cup C = A} m_1^\Omega(B)\, m_2^\Omega(C) \quad \forall\, A \subseteq \Omega. \tag{5}$$

Dans le cadre du Modèle des Croyances Transférables, on peut distinguer le niveau crédal où les croyances sont manipulées et le niveau pignistique où les décisions sont prises. Au niveau pignistique, une fonction de croyance unique, sorte de résumé exhaustif de l'information disponible au niveau crédal, est utilisée pour la prise de décision. Ph. Smets [14, 13] propose de transformer cette fonction de masse m^Ω en une fonction de probabilité $Bet\,P^\Omega$ définie sur Ω (appelée fonction de probabilité *pignistique*) qui se formalise pour tout $\omega_k \in \Omega$ par :

$$Bet\,P^\Omega(\omega_k) = \sum_{A \ni \omega_k} \frac{m^\Omega(A)}{|A|} \frac{1}{1 - m^\Omega(\emptyset)} \tag{6}$$

où $|A|$ représente la cardinalité du sous-ensemble $A \subseteq \Omega$ et $Bet\,P^\Omega(A) = \sum_{\omega \in A} Bet\,P^\Omega(\omega)$, $\forall A \subseteq \Omega$. Dans cette transformation, la masse de croyance $m^\Omega(A)$ est uniformément distribuée parmi les éléments de A.

Théorème 2.1. (Théorème de Bayes généralisé) *Soit deux espaces finis Y, l'espace des observations et* Ω *un espace non ordonné de paramètres. Le théorème de Bayes généralisé, consiste à définir une fonction de croyance sur* Ω *à partir : d'une observation* $y \subseteq Y$, *d'un ensemble de jeux de masse conditionnels* $m^Y[\omega_k]$ *sur Y, un pour chaque* $\omega_k \in \Omega$ *et d'un a priori vide sur* Ω. *Etant donné l'ensemble de jeux de masse conditonnels (qui peuvent être associés à leur plausibilité* $pl^Y[\omega_k]$*), alors pour* $y \subseteq Y$, *on a la relation :*

$$pl^\Omega[y](A) = 1 - \prod_{\omega_k \in A} (1 - pl^Y[\omega_k](y)) \qquad \forall\, A \subseteq \Omega. \tag{7}$$

Ainsi le théorème de Bayes généralisé est une extension du théorème de Bayes dans le cadre du modèle des croyances transférables [11] et peut être utilisé à bon escient lorsque cela sera nécessaire. Bien d'autres notions et outils ont été définis dans les cadre du MCT mais nous présentons uniquement ici ceux qui sont utiles à la compréhension de la suite de cet article.

3. Apprentissage de modèles de mélanges finis

Soit $\boldsymbol{Y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$ l'ensemble des données i.i.d. d'observation générées à partir d'un modèle de mélange fini dont la densité de probabilité est donnée par :

$$f(\boldsymbol{y}_i; \Psi) = \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{y}_i; \theta_k) \tag{8}$$

où K est le nombre de composantes dans le mélange considéré, π_k sont les proportions du mélange, f_k est la densité de probabilité paramétrée par θ_k, et $\Psi = \{(\pi_k, \theta_k) : k = 1, \ldots, K\}$ sont les paramètres du modèle à estimer. Pour un modèle de mélange gaussien, la fonction $f_k(\boldsymbol{y}_i; \theta_k)$ est une densité de probabilité gaussienne de paramètres $\theta_k = (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ où $\boldsymbol{\mu}_k$ est la moyenne et $\boldsymbol{\Sigma}_k$ la matrice de variance-covariance de la densité f_k.

3.1. Estimation du maximum de vraisemblance

Pour l'estimation de Ψ, on procéde généralement à l'aide d'une méthode qui permet de maximiser la vraisemblance (ou la log-vraisemblance pour des raisons de façilité de calcul) qui peut être exprimée par :

1. On utilisera la notation $g^\Omega[data]$ pour représenter une de ces fonctions définie sur Ω basée sur l'observation des données $[data]$.

$$\log\ L(\Psi;\boldsymbol{Y}) = \log(\prod_{i=1}^{n} f(\boldsymbol{y}_i;\Psi)) \tag{9}$$

$$= \sum_{i=1}^{n} \log(\sum_{k=1}^{K} \pi_k f_k(\boldsymbol{y}_i;\theta_k)). \tag{10}$$

Malheureusement, il n'existe pas de solution analytique à ce problème. L'estimation du maximum de vraisemblance de Ψ implique de résoudre l'équation $\partial \log L(\Psi;\boldsymbol{Y})/\partial\Psi = 0$ qui peut être manipulée de telle sorte que l'estimation du maximum de vraisemblance de Ψ, notée $\widehat{\Psi} = \{\hat{\pi}_k, \hat{\theta}_k\}_{k=1}^{K}$ vérifie :

$$\hat{\pi}_k = \sum_{i=1}^{n} t_k(\boldsymbol{y}_i;\widehat{\Psi})/n \qquad \forall\, k = 1,\ldots,K \tag{11}$$

où $t_k(\boldsymbol{y}_i;\Psi)$ est une quantité à valeurs dans $[0,1]$ qui permet d'indiquer que l'échantillon i lié à l'observation $\boldsymbol{y}_i$ a été généré par la composante k du mélange pouvant être définie par :

$$t_k(\boldsymbol{y}_i;\Psi) = \frac{\pi_k f_k(\boldsymbol{y}_i;\theta_k)}{\sum_{h=1}^{K} \pi_h f_h(\boldsymbol{y}_i;\theta_h)}. \tag{12}$$

De nombreux auteurs ont suggéré de résoudre les équations (11) et (12) par des méthodes de calcul itératif [9] où pour une valeur[2] initiale $\Psi^{(0)}$ de Ψ dans l'équation (11), une nouvelle estimée $\Psi^{(1)}$ peut être calculée pour Ψ dans (12) qui est substituée à nouveau dans (11) de façon à produire une nouvelle mise à jour $\Psi^{(2)}$ et ainsi de suite jusqu'à obtenir la convergence. De nos jours, cette méthode itérative de calcul de la solution de l'équation de vraisemblance peut être menée par application directe de l'algorithme EM inspiré des travaux de Dempster *et al.* [3]. Cet algorithme permet d'assurer que les valeurs prises par la vraisemblance augmentent de façon monotone au fur et à mesure des itérations et ainsi d'assurer la maximisation jusqu'à la convergence. Ce problème d'estimation de modèles de mélanges dans le cadre EM implique de formuler le problème comme un problème de données manquantes.

3.2. Un problème de données manquantes

Pour la maximisation de l'équation (10), l'idée consiste à définir une variable « cachée » qui indique quelle composante du mélange a permis de générer chacune des observations. En formulant le problème de la sorte, le calcul du maximum global de la vraisemblance se décompose en un ensemble de maximisations simples. On suppose que le mode à partir duquel l'échantillon $\boldsymbol{y}_i$ a été tiré est inconnu de telle sorte que les données manquantes sont les labels $\boldsymbol{z}_i$, $i = 1,\ldots,n$. Pour l'échantillon i, le vecteur des labels $\boldsymbol{z}_i$ est un vecteur de dimension K où $z_{ik} = 1$ ou 0 selon que l'observation $\boldsymbol{y}_i$ a été générée par la composante k du mélange ou non. Ainsi, l'échantillon complet sera noté $\boldsymbol{X} = \{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n\}$ avec $\boldsymbol{x}_i = (\boldsymbol{y}_i,\boldsymbol{z}_i)$ et $\boldsymbol{X} = \{\boldsymbol{Y},\boldsymbol{Z}\}$. En utilisant cette variable indicatrice $\boldsymbol{Z}$, l'équation (10) peut être réécrite comme la log-vraisemblance complétée :

$$\log\ L_c(\Psi;\boldsymbol{X}) = \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik} \log(\pi_k f_k(\boldsymbol{y}_i;\theta_k)) \tag{13}$$

où $z_{ik} = 1$ si la densité de probabilité gaussienne qui a permis de générer l'observation $\boldsymbol{y}_i$ est f_k, et 0 sinon. Si la variable $\boldsymbol{Z}$ est connue, le calcul devient trivial. Puisque ce n'est pas le cas, la probabilité *a posteriori* $P_{\Psi}^{\Omega}(z_{ik} = 1|\boldsymbol{Y})$ que $\boldsymbol{y}_i$ ait été générée par la distribution f_k est utilisée. Cette probabilité est définie sur le domaine $\Omega = \{\omega_k\}_{k=1}^{K}$ qui correspond à l'ensemble des composantes du mélange. Nous verrons comment cette probabilité peut être remplacée par la probabilité pignisitique calculée à partir d'un jeu de masse $m^{\Omega}[\boldsymbol{y}_i,\Psi]$ qui quantifie la croyance sur le fait que $\boldsymbol{y}_i$ ait été générée par une des composantes du mélange dans Ω estimée à partir des paramètres Ψ.

3.3. Algorithme EM pour l'estimation des paramètres

L'algorithme EM est appliqué à ce problème en considérant les variables z_{ik} comme des données manquantes et en estimant leurs valeurs. A partir des densités de probabilité $f(\boldsymbol{Y};\Psi)$ et $f(\boldsymbol{X};\Psi)$, on peut écrire $f(\boldsymbol{X};\Psi) = f(\boldsymbol{Y};\Psi)\ f(\boldsymbol{X}|\boldsymbol{Y},\Psi)$ ou encore :

$$\log\ L(\Psi;\boldsymbol{Y}) = \log\ L_c(\Psi;\boldsymbol{X}) - \log\ f(\boldsymbol{X}|\boldsymbol{Y},\Psi). \tag{14}$$

Pour une valeur $\Psi^{(t)}$ des paramètres Ψ estimés à l'itération t, on peut prendre l'espérance conditionelle à $\boldsymbol{Y}$ de l'équation précédente, ce qui permet d'écrire :

$$\log\ L(\Psi;\boldsymbol{Y}) = E_{\Psi^{(t)}}[\log\ L_c(\Psi;\boldsymbol{X})|\boldsymbol{Y}] - E_{\Psi^{(t)}}[\log\ f(\boldsymbol{X}|\boldsymbol{Y},\Psi)] \tag{15}$$

$$= Q(\Psi;\Psi^{(t)}) + H(\Psi;\Psi^{(t)}). \tag{16}$$

Dans cette équation, l'opérateur d'espérance E est associé d'un indice $\Psi^{(t)}$ de façon à insister sur le fait que l'espérance est calculée à partir de $\Psi^{(t)}$, l'estimée de Ψ à l'itération t. À l'itération $(t+1)$, si on prend la valeur de Ψ qui maximise $Q(.;\Psi^{(t)})$ c'est-à-dire $\Psi^{(t+1)} = \arg_{\Psi}\max Q(.;\Psi^{(t)})$, alors on a $Q(\Psi^{(t+1)};\Psi^{(t)}) \geq Q(\Psi^{(t)};\Psi^{(t)})$ et comme $H(\Psi^{(t+1)};\Psi^{(t)}) \leq H(\Psi^{(t)};\Psi^{(t)})$ (Inégalité de Jensen), on obtient la maximisation de la vraisemblance :

$$\log\ L(\Psi^{(t+1)};\boldsymbol{Y}) \geq \log\ L(\Psi^{(t)};\boldsymbol{X}). \tag{17}$$

Pour s'assurer de maximiser la vraisemblance $L(\Psi;\boldsymbol{Y})$, il faut donc s'intéresser à $Q(.;\Psi^{(t)})$ qui peut s'écrire :

$$Q(\Psi;\Psi^{(t)}) = \sum_{i=1}^{n}\sum_{k=1}^{K} P_{\Psi^{(t)}}^{\Omega}(z_{ik} = 1|\boldsymbol{y}_i)\log(\pi_k f_k(\boldsymbol{y}_i;\theta_k)). \tag{18}$$

L'algorithme EM construit une suite d'estimateurs $\widehat{\Psi}^{(t)}$ par itération successive de deux étapes E (pour Espérance) et M (pour Maximisation).

2. Soit $\Psi^{(t)}$ la valeur estimée à l'itération t.

Étape E : L'étape E a pour but de calculer l'espérance conditionelle de la log-vraisemblance complétée $\log L_c(\Psi; \boldsymbol{X})$ sachant les observations $\boldsymbol{Y}$ en utilisant l'estimation courante $\Psi^{(t)}$ à l'itération t. Comme la log-vraisemblance complétée $\log L_c(\Psi; \boldsymbol{X})$ est une fonction linéaire des données non observées z_{ik}, l'étape E à l'itération $(t+1)$ requiert simplement le calcul de l'espérance conditionnelle de Z_{ik} sachant $\boldsymbol{Y}$ où Z_{ik} est la variable aléatoire correspondante à z_{ik} :

$$E_{\Psi^{(t)}}[Z_{ik}|\boldsymbol{Y}] = P^{\Omega}_{\Psi^{(t)}}(z_{ik} = 1|\boldsymbol{Y}) = t_k(\boldsymbol{y}_i; \Psi^{(t)}). \tag{19}$$

En utilisant ce résultat, qui sera généralisé dans le cadre du modèle des croyances transférables à la section 4, le terme à maximiser peut être réécrit :

$$Q(\Psi; \Psi^{(t)}) = \sum_{i=1}^{n}\sum_{k=1}^{K} t_k(\boldsymbol{y}_i; \Psi^{(t)}) \log(\pi_k f_k(\boldsymbol{y}_i; \theta_k)). \tag{20}$$

Étape M : L'étape M à l'itération t impose la maximisation globale de $Q(\Psi; \Psi^{(t)})$ par rapport à Ψ sur l'espace total des paramètres de façon à obtenir l'estimation mise à jour $\Psi^{(t+1)}$. Pour les modèles de mélanges finis, l'estimation mise à jour des proportions $\pi_k^{(t+1)}$ est calculée indépendamment de l'estimation des paramètres $\theta_k^{(t+1)}$ pour $k = 1, \ldots, K$. Si les variables z_{ik} étaient observables, l'estimation au sens de la log-vraisemblance complétée de π_k, notée $\hat{\pi}_k$, serait donnée par :

$$\hat{\pi}_k = \sum_{i=1}^{n} z_{ik}/n \qquad \forall\, k = 1, \ldots, K. \tag{21}$$

Puisque l'étape E requiert simplement de remplacer chaque variable z_{ik} par son espérance conditionnelle $t_k(\boldsymbol{y}_i; \Psi^{(t)})$ calculée à partir des paramètres $\Psi^{(t)}$ dans la log-vraisemblance complétée, l'estimation mise à jour de π_k est calculée en remplaçant z_{ik} par $t_k(\boldsymbol{y}_i; \Psi^{(t)})$ ce qui conduit à :

$$\hat{\pi}_k^{(t+1)} = \sum_{i=1}^{n} t_k(\boldsymbol{y}_i; \Psi^{(t)})/n \qquad \forall\, k = 1, \ldots, K. \tag{22}$$

Ainsi, pour obtenir $\hat{\pi}_k$ à l'itération $(t+1)$, la contribution de chacune des observations $\boldsymbol{y}_i$ est égale à son degré d'appartenance courant à la composante k du modèle de mélange. En ce qui concerne la mise à jour de Θ à l'itération $(t+1)$ de l'étape M, on peut remarquer que Θ^{t+1} peut être obtenue en analysant l'équation :

$$\sum_{i=1}^{n}\sum_{k=1}^{K} t_k(\boldsymbol{y}_i; \Psi^{(t)})\, \partial \log f_k(\boldsymbol{y}_i; \theta_k)/\partial\Theta = 0. \tag{23}$$

Dans le cas d'un mélange gaussien de paramètres $(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, la résolution de cette équation permet d'obtenir :

$$\widehat{\boldsymbol{\mu}}_k^{(t+1)} = \frac{\sum_{i=1}^{n} t_k(\boldsymbol{y}_i, \Psi^{(t)})\, \boldsymbol{y}_i}{\sum_{i=1}^{n} t_k(\boldsymbol{y}_i, \Psi^{(t)})} \tag{24}$$

$$\widehat{\boldsymbol{\Sigma}}_k^{(t+1)} = \frac{\sum_{i=1}^{n} t_k(\boldsymbol{y}_i, \Psi^{(t)})\, (\boldsymbol{y}_i - \widehat{\boldsymbol{\mu}}_k^{(t+1)})(\boldsymbol{y}_i - \widehat{\boldsymbol{\mu}}_k^{(t+1)})}{\sum_{i=1}^{n} t_k(\boldsymbol{y}_i, \Psi^{(t)})}. \tag{25}$$

Convergence : Les étapes E et M sont répétées de façon alternative jusqu'à ce que la différence $L(\Psi^{(t+1)}) - L(\Psi^{(t)})$ soit inférieure à un certain seuil. Dempster *et al.* [3] ont montré que la fonction de vraisemblance ne décroit pas après chaque itération EM de telle sorte que $Q(\Psi; \Psi^{(t+1)}) \geq Q(\Psi; \Psi^{(t)})$ pour $t = 0, 1, 2, \cdots$. Ainsi, la convergence peut être obtenue pour une suite de valeurs bornées de la vraisemblance. Dans la section suivante, nous proposons d'étendre ce formalisme et cet algorithme dans le cadre du Modèle des Croyances Transférables.

4. Algorithme EM crédibiliste

Nous montrons dans la section 4.1, comment peut être étendu le principe du maximum de vraisemblance dans un contexte où les données sont modélisées par des fonctions de croyance. Ce résultat permet d'envisager la généralisation de l'algorithme EM classique pour l'estimation d'un modèle de mélange fini dans le cadre du MCT (*cf.* section 4.2). L'intégration de connaissances *a priori* sur le modèle de génération des données permet la résolution d'une large palette de problèmes d'apprentissage (section 4.3).

4.1. Maximum de vraisemblance et MCT

En théorie des probabilités, beaucoup de procédures d'estimation pour un jeu de paramètres ω sont basées sur la maximisation de la vraisemblance c'est-à-dire $P^Y(\boldsymbol{Y}|\omega)$ considérée comme une fonction de $\omega \in \Omega$. On reconsidère ce problème dans le cadre du MCT. Pour chaque $\omega \in \Omega$, on dispose d'un jeu de masse conditionnel noté $m^Y[\omega]$ défini sur Y, l'espace des observations. L'observation de $\boldsymbol{y} \subseteq Y$ permet le calcul d'un jeu de masse défini sur Ω par application directe du théorème de Bayes généralisé conduisant ainsi à l'obtention d'un jeu de masse noté $m^{\Omega}[\boldsymbol{y}]$.

La question qui se pose concerne l'estimation de ω_0, la valeur actuelle du paramètre Ω ? En maximisant la probabilité pignistique $BetP^{\Omega}[\boldsymbol{y}]$ calculée à partir de $m^{\Omega}[\boldsymbol{y}]$, on peut trouver la valeur la plus probable pour Ω. Cette idée, qui impose de trouver la valeur modale de $BetP^{\Omega}[\boldsymbol{y}]$, nous paraît intuiter au principe général des estimateurs du maximum de vraisemblance dans le cadre du MCT.

Ainsi, on doit trouver le paramètre $\omega_0 \in \Omega$ tel que $BetP^{\Omega}[\boldsymbol{y}](\omega_0) \geq BetP^{\Omega}[\boldsymbol{y}](\omega_i), \forall\, \omega_i \in \Omega$. Dans la pratique, cette maximisation semble impossible à résoudre mais nous utilisons le résultat suivant [2] qui postule que la valeur ω_0 qui

maximise $Bet\,P^{\Omega}[\boldsymbol{y}]$ est la même que la valeur qui maximise la plausibilité $pl^{Y}[\omega_0](\boldsymbol{y})$.

Théorème 4.1. (Maximum de vraisemblance) *Sachant $\boldsymbol{y} \subseteq Y$ et $pl^{Y}[\omega](\boldsymbol{y})$ pour tout $\omega \in \Omega$, soit $pl^{\Omega}[\boldsymbol{y}]$ la plausibilité définie sur Ω et calculée par le théorème de Bayes généralisé, et $Bet\,P^{\Omega}[\boldsymbol{y}]$ sa distribution de probabilité pignistique associée, alors :*

$$Bet\,P^{\Omega}[\boldsymbol{y}](\omega_i) > Bet\,P^{\Omega}[\boldsymbol{y}](\omega_j) \quad \text{si et seulement si} \quad pl^{Y}[\omega_i](\boldsymbol{y}) > pl^{Y}[\omega_j](\boldsymbol{y}). \quad (26)$$

La preuve de ce théorème est donnée à la fin de cet article. Dans notre cas, ce résultat est très utile puisqu'il s'agit de maximiser la vraisemblance avec un *a priori* vide sur Ω. Si on dispose de n i.i.d. observations $\boldsymbol{y}_i, i = 1,\ldots,n$, on a $pl^{Y}[\omega](\boldsymbol{y}_1,...,\boldsymbol{y}_n) = \prod_{i=1}^{n} pl^{Y}[\omega](\boldsymbol{y}_i)$. Ce dernier terme est facile à calculer et permet d'envisager un algorithme réaliste. Maximiser la vraisemblance sur Ω revient ainsi à maximiser sur Ω les plausibilités conditionnelles des données sachant ω. Ce résultat est largement utilisé dans l'algorithme présenté.

4.2. L'algorithme EM crédibiliste pour l'estimation des paramètres

À la section 3.2, nous avons vu qu'en formulant le problème de l'estimation des paramètres d'un mmodèle de mélange fini comme un problème de données manquantes, il suffisait d'expliciter la variable $t_k(\boldsymbol{y}_i; \Psi^{(t)})$ qui représente la probabilité *a posteriori* que $\boldsymbol{y}_i$ ait été générée par la distribution f_k estimée à l'itération t. Dans le cadre du MCT, $t_k(\boldsymbol{y}_i; \Psi^{(t)})$ peut être estimée par la probabilité pignistique associée à un jeu de masses, noté $m^{\Omega}[\boldsymbol{y}_i; \Psi^{(t)}]$, défini sur $\Omega = \{\omega_1,\ldots,\omega_K\}$ l'ensemble fini des hypothèses correspondantes aux différentes composantes du modèle de mélange.

Si l'observation $\boldsymbol{y}_i$ a été générée par la composante k du mélange, $pl^{Y}[\omega_k](\boldsymbol{y}_i)$ est donnée par $f_k(\boldsymbol{y}_i; \theta_k)$. Puisque l'observation $\boldsymbol{y}_i$ est un singleton alors $pl^{\boldsymbol{Y}}[\omega_k](\boldsymbol{y}_i) = f_k(\boldsymbol{y}_i; \theta_k)d\boldsymbol{y}$ où $d\boldsymbol{y}$ permet d'indiquer que la plausibilité est une fonction ensembliste à valeurs dans $[0,1]$ tandis que f_k est une densité à valeurs dans $[0,+\infty[$ qui nécessite d'être normalisée permettant ainsi de faire disparaître le terme $d\boldsymbol{y}$ de façon naturelle. Pour une observation $\boldsymbol{y}_i$ de $\boldsymbol{Y}$, comment peut on dériver la fonction de croyance définie sur Ω à partir des paramètres $\Psi^{(t)}$? La solution est proposée dans la cadre du MCT par l'application du théorème de Bayes généralisé (*cf.* théorème 2.1). Ainsi, on peut définir la plausibilité $pl^{\Omega}[\boldsymbol{y}_i, \Psi^{(t)}]$ à partir des plausibilités conditionnelles $pl^{Y}[\omega_k](\boldsymbol{y}_i)$ par :

$$pl^{\Omega}[\boldsymbol{y}_i, \Psi^{(t)}](A) = 1 - \prod_{\omega_k \in A} (1 - pl^{Y}[\omega_k](\boldsymbol{y}_i)). \quad (27)$$

La fonction de masse $m^{\Omega}[\boldsymbol{y}_i; \Psi^{(t)}]$ qui permet de quantifier le degré de croyance pour que l'observation $\boldsymbol{y}_i$ ait été générée à partir de la distribution f_k dont les paramètres sont calculés à partir de $\Psi^{(t)}$ estimés à l'itération t peut ainsi être retrouvé à partir de $pl^{\Omega}[\boldsymbol{y}_i, \Psi^{(t)}]$.

Dans l'équation (20), on peut donc remplacer la variable $t_k(\boldsymbol{y}_i; \Psi^{(t)})$ par la fonction de masse $m^{\Omega}[\boldsymbol{y}_i, \Psi^{(t)}]$ qui permet d'en quantifier la croyance et la densité $f_k(\boldsymbol{y}_i; \theta_k)$ par la plausibilité $pl^{Y}[.](\boldsymbol{y}_i)$. Avec ces considérations, le terme à maximiser peut être réécrit par la relation suivante :

$$Q(\Psi; \Psi^{(t)}) = \sum_{i=1}^{n} \sum_{A \subseteq \Omega} m^{\Omega}[\boldsymbol{y}_i, \Psi^{(t)}](A)\ \log(pl^{Y}[A](\boldsymbol{y}_i)). \quad (28)$$

Cette équation est l'analogie de l'équation (20) dans le cadre du MCT et nous permet d'étendre l'estimation du modèle de mélange fini aux fonctions de croyance et d'obtenir un algorithme opérationnel. Pour l'estimation des paramètres $(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, la variable $t_k(\boldsymbol{y}_i; \Psi^{(t)})$ sera calculée comme la probabilité pignisitque définie $\forall\ \omega_k \in \Omega$ par la relation suivante :

$$t_k(\boldsymbol{y}_i; \Psi^{(t)}) = \sum_{A \ni \omega_k} \frac{m^{\Omega}[\boldsymbol{y}_i; \Psi^{(t)}](A)}{|A|}. \quad (29)$$

L'analyse de la convergence de l'algorithme proposé est donc identique à l'algorithme EM et n'a rien de spécifique puisqu'elle correspond à remplaçer la probabilité *a posteriori* et les densités de probabilités par des masses et des plausibilités.

4.3. Intégration de connaissances *a priori*

Grâce à sa flexibilité pour modéliser des informations imprécises et incertaines, une fonction de croyance peut représenter différents types de connaissances sur le modèle de génération des données. Ainsi, si l'on dispose d'informations *a priori* sur les labels $\boldsymbol{z}_i$ de certaines des observations $\boldsymbol{y}_i$ dans $\boldsymbol{X}$, l'estimation du modèle de mélange ne peut qu'être facilitée. Dans le cadre du MCT, cette connaissance peut être modélisée par une fonction de masse, notée m_i^{Ω}, définie sur Ω pour chacune des observations $\boldsymbol{y}_i$. Cette fonction quantifie le degré de croyance *a priori* sur le fait que l'observation ait été générée par l'une des K sources du modèle de mélange. Le tableau 1 propose quelques exemples de telles fonctions de masse qui permettent de caractériser la connaissance *a priori* sur le modèle de génération des données dans le cas où $K = 3$.

Ces fonctions de masse m_i^{Ω} définies pour chaque observation $\boldsymbol{y}_i$ sont intégrées dans l'algorithme proposé précédement. Ainsi le terme à maximiser peut être réécrit par l'équation suivante :

$$Q(\Psi; \Psi^{(t)}) = \sum_{i=1}^{n} \sum_{A \subseteq \Omega} (m^{\Omega}[\boldsymbol{y}_i, \Psi^{(t)}]\ \textcircled{\cap} m_i^{\Omega})(A)\ \log(pl^{Y}[A](\boldsymbol{y}_i)). \quad (30)$$

On remarque que cette modélisation va permettre d'intégrer :

- le cas des labels inconnus qui seront pris en compte par une fonction de masse vide $m_i^{\Omega}(\Omega) = 1$. On se place donc ici dans

Tableau 1. Étiquetage imprécis et/ou incertain avec des fonctions de masse dans le cas précis (fonction de masse certaine), imprécis (fonction de masse catégorique), probabiliste (fonction de masse bayesienne), crédal (cas le plus général) et inconnu (fonction de masse vide).

$A \subseteq \Omega$	Précis	Imprécis	Probabiliste	Crédal	Inconnu
$\{\omega_1\}$	0	0	0.2	0.1	0
$\{\omega_2\}$	1	0	0.6	0	0
$\{\omega_1,\omega_2\}$	0	1	0	0.2	0
$\{\omega_3\}$	0	0	0.2	0.3	0
$\{\omega_1,\omega_3\}$	0	0	0	0.3	0
$\{\omega_2,\omega_3\}$	0	0	0	0	0
Ω	0	0	0	0.1	1

le contexte de l'apprentissage non supervisé où la fonction m_i^Ω n'a pas d'influence sur le terme Q (élement neutre de la combinaison conjonctive) ;
- le cas des labels parfaitement connus où la source k ayant générée l'observation $\boldsymbol{y}_i$ est parfaitement déterminé et modélisée par une fonction de masse certaine $m_i^\Omega(\{\omega_k\}) = 1$. C'est fois-ci, c'est le terme $m^\Omega[\boldsymbol{y}_i, \Psi^{(t)}]$ qui n'a pas d'influence sur la combinaison ;
- le cas des labels imprécis où l'on dispose d'une information imprécise sur la (ou les) source(s). On peut citer l'exemple où l'on dispose d'une information qui ne permet pas de différencier entre deux sources (*cf.* tableau 1) pour lequel la fonction sera catégorique du style $m_i^\Omega(B) = 1$. Remarquons qu'après combinaison, la masse affectée aux éléments focaux $A \subseteq \Omega$ de $m^\Omega[\boldsymbol{y}_i, \Psi^{(t)}]$ sera transférée sur $A \cap B$;
- le cas des labels probabilistes où la connaissance sera modélisée par une fonction de masse bayesienne ainsi que
- le cas des labels crédaux qui correspondent au cas le plus général.
Ceci montre que ce type de modélisation où les labels sont manipulés par des fonctions de croyance permet de couvrir toutes les situations (imprécise et/ou incertaine) d'intégration de connaissances *a priori*. Quelques exemples de telles configurations d'apprentissage sont données dans la section suivante.

5. Résultats expérimentaux

Pour illustrer et mettre en évidence les performances de l'algorithme proposé, quelques résultats expérimentaux sur des données simulées sont présentés. Nous présentons tout d'abord un cas où les données générées sont partiellement observées puis nous illustrerons l'efficacité de l'algorithme CrEM sur un problème de clustering. Dans les deux cas, on considére un modèle de mélange gaussien de densité de probabilité $f_k(\boldsymbol{y}_i; \theta_k)$ de paramètres $\theta_k = (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ où $\boldsymbol{\mu}_k$ est la moyenne et $\boldsymbol{\Sigma}_k$ la matrice de variance-covariance qui peut se mettre sous la forme $\boldsymbol{\Sigma}_\mathbf{k} = \sigma_\mathbf{k}\boldsymbol{I}$ où $\boldsymbol{I}$ est la matrice identité.
En ce qui concerne l'implémentation de l'algorithme CrEM, les deux étapes suivantes sont alternées jusqu'à ce que la différence $Q(\Psi^{(t+1)}; \Psi^{(t)}) - Q(\Psi^{(t)}; \Psi^{(t-1)})$ soit inférieure à un certain seuil fixé dans les paramètres de l'algorithme, le terme $Q(\Psi; \Psi^{(t)})$ étant calculé par l'équation (30).
- L'étape E consiste à calculer $t_k(\boldsymbol{y}_i, \Psi^{(t)})$ qui est estimée dans CrEM par l'équation (29) où $m^\Omega[\boldsymbol{y}_i; \Psi^{(t)}]$ est calculé par l'équation (27) à partir de sa plausibilité. L'intégration de connaissances *a priori* est également réalisée dans cette étape à partir de la combinaison conjonctive entre $m^\Omega[\boldsymbol{y}_i; \Psi^{(t)}]$ et m_i^Ω pour chaque observation.
- L'étape M consiste à maximiser la vraisemblance conditionnelle ce qui est fait en deux temps :
  1) en estimant les $\hat{\pi}_k^{(t+1)}$ par l'équation (22) à partir des $t_k(\boldsymbol{y}_i, \Psi^{(t)})$,
  2) en résolvant les équations de vraisemblance ce qui donne pour un mélange gaussien de paramètres $(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, les résultats proposés aux équations (24) et (25).
L'initialisation des paramètres à la première itération de l'algorithme est réalisée de manière aléatoire. L'estimation finale des paramètres est celle qui maximise le critère de vraisemblance sur l'ensemble des tirages réalisés à l'initialisation. Une approche plus efficace pourrait être envisagée [16].

5.1. Cas partiellement supervisé

On suppose un modèle de mélange à $K = 6$ composantes dans un espace d'observation à deux dimensions $Y = \mathbb{R}^2$. Pour chaque composante k du modèle, on génére un échantillon $\boldsymbol{Y}$ dont chaque observation $\boldsymbol{y}$ est tiré à partir d'une distribution Gaussienne $f(\boldsymbol{y}|\omega_k) \sim \mathcal{N}(\mu_\mathbf{k}, \boldsymbol{\Sigma}_\mathbf{k})$ où $\boldsymbol{\Sigma}_\mathbf{k} = \sigma_\mathbf{k}\boldsymbol{I}$. Les paramètres des six densités sont respectivement présentés au tableau 2. Trois composantes du modèle sont paramétrées de telle manière à ce que les observations soient localisées autour des trois sommets d'un triangle isocèle ABC (qui correspondent aux moyennes respectives et pour lesquelles $\sigma_k = 0.5$) tandis que les trois autres composantes sont elles-mêmes localisés sur les sommets d'un autre triangle isocèle DEF (avec $\sigma_k = 2$). À la figure 1, on illustre un exemple d'un tel échantillon (de 50 observations pour chacune des composantes) et les deux triangles isocèles ABC et DEF correspondants (traits fins). Dans cette simulation, on suppose que les labels pour chacune des observations peuvent être de différents types.
- Les labels des composantes ω_1, ω_3 et ω_5 sont supposés connus (observations dont la variance est la plus faible).
- Les labels des composantes ω_2, ω_4 et ω_6 sont tout d'abord supposés imprécis (fonction de masse catégorique) et aléatoirement

Tableau 2. Paramètres des distributions constituant l'ensemble des échantillons, constituon de l'apprentissage et estimations des paramètres (moyenne m_k, variance s_k et proportions π_k pour chacune des composantes du modèle).

	$\omega_1(\diamond)$	$\omega_2(+)$	$\omega_3(\circ)$	$\omega_4(\cdot)$	$\omega_5(*)$	$\omega_6(\times)$
μ_k	$(10,10)^t$	$(17.5,14.3)^t$	$(15,18.6)^t$	$(15,10)^t$	$(20,10)^t$	$(12.5,14.3)^t$
σ_k	2	0.5	2	0.5	2	0.5
Cas Imprécis	50 ω_1	25 ω_2,ω_3 25 ω_2,ω_5	50 ω_3	25 ω_4,ω_1 25 ω_4,ω_5	50 ω_5	25 ω_6,ω_1 25 ω_6,ω_3
m_k	$(9.13,10.3)^t$	$(17.5,14.3)^t$	$(15.6,18.9)^t$	$(14.9,10.1)^t$	$(20.3,9.8)^t$	$(12.4,14.3)^t$
s_k	2.57	0.38	1.85	0.37	3.24	0.35
π_k	0.185	0.152	0.178	0.148	0.179	0.154

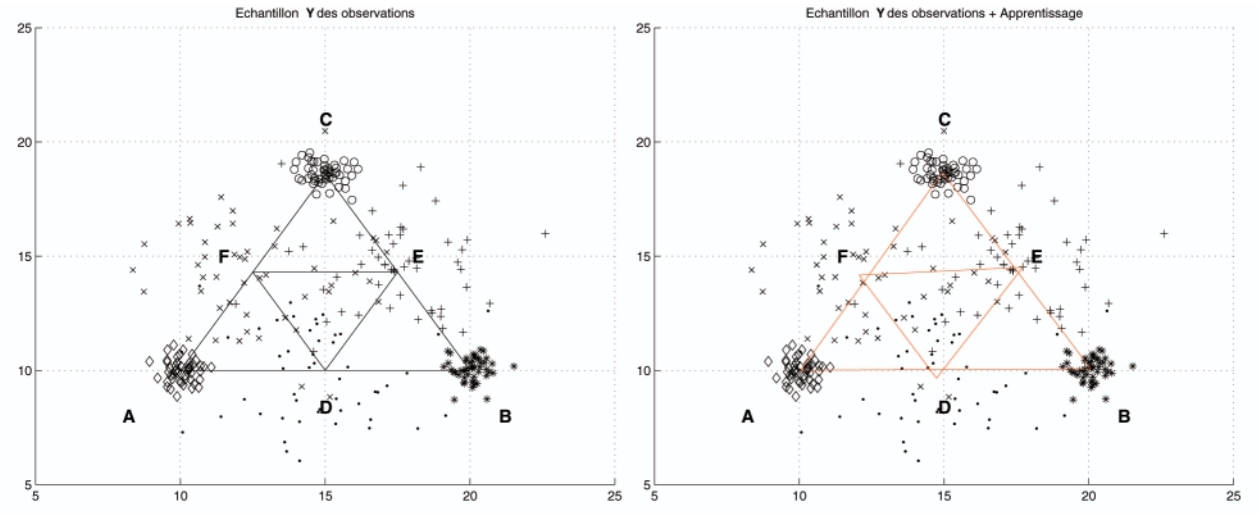

Figure 1. Échantillon dans l'espace des observations (gauche), données d'apprentissage et estimation des paramètres (droite).

réparties en deux groupes. Ainsi, pour les 50 échantillons de la composante ω_2, 25 sont étiquetés $\{\omega_2,\omega_3\}$ c'est-à-dire avec $m_i^{\Omega}(\omega_2,\omega_3) = 1$ et 25 sont étiquetés $\{\omega_2,\omega_5\}$. Il en est de même pour les composantes ω_4 et ω_6 (voir tableau 2).

- Dans un second temps qui correspond à un étiquetage crédal, les labels ont été générés de la façon suivante. L'idée consiste à perturber la certitude sur la composante qui a générée l'observation tout en gardant la vraie composante au sein des éléments focaux de la fonction de masse m_i^{Ω} générée. Ainsi, chaque sous-ensemble de Ω y compris la vraie composante se voit affecté une partie aléatoire de la masse avec une probabilité donnée. On génére ainsi des ensembles d'apprentissage imprécis et incertains qui peuvent être rencontrés dans des applications réelles.

Estimation des paramètres : Nous présentons à la figure 1 (*cf.* figure de droite), les paramètres estimés pour une simulation de l'ensemble des observations. Les triangles illustrent l'application de l'algorithme dans le cas de l'étiquetage imprécis présenté au tableau 2. Comme on peut le remarquer sur cette figure, les moyennes (situées aux sommets des 2 triangles) sont bien estimées par l'algorithme. L'ensemble des estimations est récapitulé dans le tableau 2.

Résultats quantitatifs : De façon à évaluer quantitativement les performances de l'algorithme proposé, nous procédons à 10 tirages de l'ensemble des observations. Pour chacun d'entre eux, on génére les labels comme explicité auparavant. Pour les labels imprécis, nous appliquons l'algoritme EM dans sa version [1] et l'algorithme proposé. Sur les labels crédaux, seul l'algorithme proposé peut être appliqué. Nous présentons au tableau 3, les taux de bonne classification obtenus pour chacun des ensembles d'observation générés. Chaque algorihme produit des résultats très similaires mais l'algorithme proposé est le seul à pouvoir assurer la gestion de labels crédaux, une information bien plus flexible que celle rencontrée dans le cas imprécis.

5.2. Cas non supervisé

De la même manière que précédemment, on considère un ensemble d'observations dans un espace à deux dimensions pour des facilités de représentation. Chaque source génére des données à partir d'une distribution Gaussienne de moyenne et

Tableau 3. Taux de bonne classification pour l'algorihme EM et l'algorithme proposé.

Triangles	1	2	3	4	5	6	7	8	9	10	moyenne	déviation
EM 8	5.3	84.3	86.3	88.0	86.7	87.0	83.3	85.7	90.7	88.0	86.5	2.1
CrEM (Imprécis)	86.3	85.3	88.0	90.3	88.0	87.3	84.0	88.0	91.0	88.0	87.6	2.0
CrEM (Crédal)	87.0	86.6	87.6	90.0	87.6	88.0	85.3	88.3	91.3	86.7	87.8	1.7

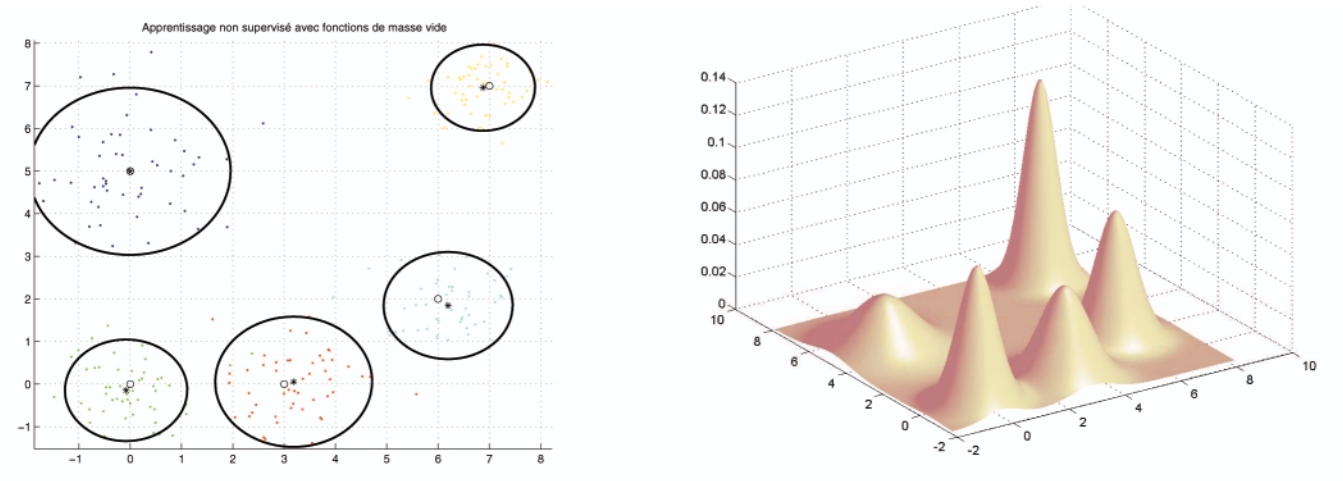

Figure 2. Résultat de l'algorithme CrEM en apprentissage non supervisé, les cercles représentant les courbes de niveau de chacune des composantes du mélange (gauche), densité du mélange estimée (droite).

matrice de variance-covariance différente. Les données d'apprentissage sont représentées à la figure 2.

On réalise un apprentissage non supervisé sur les données générées avec l'algorithme CrEM en initialisant chaque fonction m_i^{Ω} par la fonction de masse vide. L'algorithme est initialisé avec cinq composantes. Les résultats de la phase d'apprentissage sont présentés à la figure 2 où les moyennes estimées sont représentées par des $*$ et où les cercles correspondent aux courbes de niveau de chacune des composantes du mélange. La densité correspondante au mélange fini est présentée sur la partie droite de la figure où l'on peut s'apercevoir que la solution finale donnée par l'algorithme conduit à une bonne estimation des paramètres et donc de la densité. Dans cette situation encore, l'algorithme CrEM s'est accomodé des informations disponibles permettant de concurrencer les méthodes d'estimation de modèles de mélange fini disponibles à l'heure actuelle. Un critère permettant d'estimer le nombre optimal de composantes dans le mélange pourrait également être introduit [4].

6. Conclusions – Perspectives

Dans cet article, l'estimation au sens du maximum de vraisemblance a été généralisée dans le cadre du modèle des croyances transférables permettant ainsi d'enrichir la palette des outils disponibles. La méthodologie présentée est basée sur l'utilisation d'une variante de l'algorithme EM pour estimer les paramètres d'un modèle de mélange fini. Les principales étapes de l'algorithme, baptisé CrEM, ont été détaillées et permettent d'envisager des applications de modélisation statistique réalistes. Des connaissances additionnelles telles que des informations disponibles sur la composante ayant servie à générer chaque observation peuvent être intégrées dans le processus d'estimation par les fonctions de croyance. Plusieurs simulations ont permis de mettre en évidence les bonnes performances de l'algorithme en comparaison de l'algorithme EM classique appliqué à l'estimation de mélanges gaussiens. Ces résultats ont été obtenus pour des problèmes d'apprentissage partiellement ou non supervisé. Cette flexibilité dans la prise en compte d'informations *a priori* confère à l'approche CrEM un avantage indéniable par rapport aux méthodes actuelles d'estimation.

De nombreuses applications de cette méthode peuvent être citées comme par exemple les réseaux bayesiens où l'utilisation des algorithmes EM pour estimer les paramètres de distributions inconnues sont utilisés. L'emploi de l'agorithme CrEM peut ainsi fournir une alternative dans le cadre des réseaux de croyance. Les travaux futurs s'orientent vers le problème difficile de la sélection du modèle en ce qui concerne le nombre de composantes, la forme des composantes, ...

Références

[1] C. AMBROISE, T. DENŒUX, G. GOVAERT, and P. SMETS, Learning from an imprecise teacher : probabilistic and evidential approaches. In *Proccedings of ASMDA'2001, volume 1*, pp. 100-105 Compiègne, France, 2001.

[2] F. DELMOTE and P. SMETS, Target identification based on the Transferable Belief Model interpretation of Dempster-Shafer. *IEEE Transactions on Systems, Man and Cybernetics*, A 34:457-471, 2004.

[3] A.P. DEMPSTER, N.M. LAIRD, and D.B. RUBIN, Maximum likelihood estimation from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, series B*, 39:1-38, 1977.

[4] Mario A.T. FIGUEIREDO and Anil K. JAIN, Unpervised learning of finite mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(3):381-396, 2002.

[5] G. Mc LACHLAN and D. PEEL, *Finite Mixture Models*. Wiley Series in Probability and Statistics. John Wiley & Sons, New-York, 2000.

[6] J.M. MARIN, K. MENGERSEN, and C.P. ROBERT, Bayesian modelling and inference on mixtures of distributions. *Handbook of Statistics*, 25, 2005.

[7] G. McLACHLAN and K. BASFORD, *Mixture Models : Inference and Application to Clustering*. Marcel Dekker, New-York, 1988.

[8] G. McLACHLAN and T. KRISHNAN, *The EM algorithm and extensions*. John Wiley, New-York, 1997.

[9] R. NEAL and G. HINTON, A view of the EM algorithm that justifies incrememtal, sparse, and other variants. In M.I. Jordan, editor, *Learning in Graphical Models*. Kluwer, 1998.

[10] G. SHAFER, *A Mathematical Theory of Evidence*. Princeton univ. Princeton, NJ, 1976.

[11] Ph. SMETS, Belief functions : the disjunctive rule of combination and the genralized Bayesian theorem. *International Journal of Approximate Reasoning*, 9:1-35, 1993.

[12] Ph. SMETS, The application of the matrix calculus to belief functions. *International Journal of Approximate Reasoning*, 31:1-30, 2002.

[13] Ph. SMETS, Decision making in the TBM : the necessity of the pignistic transformation. *International Journal of Approximate Reasoning*, 38:133-147, 2005.

[14] Ph. SMETS and R. KENNES, The transferable belief model. *Artifical Intelligence*, 66:191-234, 1994.

[15] P. VANNOORENBERGHE and Ph. SMETS, Partially supervised learning by a credal EM approach. In Lluis Godo, editor, *Symbolic and Quantitative Approaches to Reasoning with uncertainty, 8th European Conference, ECSQARU 2005*, pp. 956-967, Spain, July 6-8 2005. Springer, 2005.

[16] Z. ZIVKOVIC and F. VAN DER HEIJDEN, Recursive unsupervised learning of finite mixture models. *IEEE Transictions on Pattern Analysis and Machine intelligence*, 26(5):651-656, May 2004.

Preuve du théorème 4.1 Soit $l(\omega_k|\boldsymbol{y}) = pl^Y[\omega_k](\boldsymbol{y})$ la plausibilité conditionnelle que nous supposerons inférieure à 1 pour tout $\omega_k \in \Omega$ c'est-à-dire $l(\omega_k|\boldsymbol{y}) < 1$. Le cas où $pl^Y[\omega_k](\boldsymbol{y}) = 1$ sera étudié ultérieurement. Soit $r_k = l(\omega_k|\boldsymbol{y})/(1 - l(\omega_k|\boldsymbol{y}))$ et $\alpha = \prod_{\omega_k \in \Omega}(1 - l(\omega_k|\boldsymbol{y}))$. L'application du thèoréme de Bayes généralisé permet de calculer la foction de masse $m^\Omega[\boldsymbol{y}]$ définie sur Ω pour tout $\omega \in \Omega$ par :

$$m^\Omega[\boldsymbol{y}](\omega) = \prod_{\omega_i \in \Omega} l(\omega_i|\boldsymbol{y}) \prod_{\omega_i \in \overline{\omega}} (1 - l(\omega_i|\boldsymbol{y})) = \alpha \prod_{\omega_k \in \Omega} r_k.$$

La probabilité pignistique $Bet\,P^\Omega[\boldsymbol{y}]$ déduite de cette fonction de masse pour $\omega_i \in \Omega$ peut s'écrire :

$$\begin{aligned} Bet\,P^\Omega[\boldsymbol{y}](\omega_i) &= \frac{1}{1 - m^\Omega[\boldsymbol{y}](\emptyset)} \sum_{\omega \subseteq \overline{\omega_i}} \frac{m^\Omega[\boldsymbol{y}](\omega_i \cup \omega)}{|\omega| + 1} \\ &= \frac{\alpha}{1 - m^\Omega[\boldsymbol{y}](\emptyset)} \sum_{\omega \subseteq \overline{\omega_i}} \frac{1}{|\omega| + 1} \prod_{\omega_k \in \Omega} r_k \\ &= \frac{\alpha}{1 - m^\Omega[\boldsymbol{y}](\emptyset)} \sum_{\omega \subseteq \overline{\omega_i \cup \omega_j}} \prod_{\omega_k \in \Omega} r_k \left(\frac{r_i}{|\omega| + 1} + \frac{r_i\, r_j}{|\omega| + 2}\right) \\ &\qquad \text{avec } j \neq i. \end{aligned}$$

À partir de ce résultat, on peut écrire la différence entre deux valeurs de la probabilité obtenues pour ω_i et ω_j :

$$\begin{aligned} &Bet\,P^\Omega[\boldsymbol{y}](\omega_i) - Bet\,P^\Omega[\boldsymbol{y}](\omega_j) \\ &= \frac{\alpha}{1 - m^\Omega[\boldsymbol{y}](\emptyset)} \sum_{\omega \subseteq \overline{\omega_i \cup \omega_j}} \prod_{\omega_k \in \Omega} r_k \left(\frac{r_i}{|\omega| + 1} - \frac{r_j}{|\omega| + 1}\right) \\ &= \frac{\alpha\,(r_i - r_j)}{1 - m^\Omega[\boldsymbol{y}](\emptyset)} \sum_{\omega \subseteq \overline{\omega_i \cup \omega_j}} \prod_{\omega_k \in \Omega} \frac{r_k}{|\omega| + 1}. \end{aligned}$$

Puisque $r_k \geq 0$, le produit des termes $r_k/(|\omega| + 1)$ est positif et la somme pour tout $\omega \subseteq \overline{\omega_i \cup \omega_j}$ aussi. Ainsi, le signe de la différence $Bet\,P^\Omega[\boldsymbol{y}](\omega_i) - Bet\,P^\Omega[\boldsymbol{y}](\omega_j)$ est le même que le signe de $r_i - r_j$:

$$Bet\,P^\Omega[\boldsymbol{y}](\omega_i) > Bet\,P^\Omega[\boldsymbol{y}](\omega_j) \qquad \text{ssi} \qquad r_i > r_j.$$

Puisque $r_i > r_j$ si et seulement si $l(\omega_i|\boldsymbol{y}) > l(\omega_j|\boldsymbol{y})$ c'est-à-dire si et seulement si $pl^Y[\omega_i](\boldsymbol{y}) > pl^Y[\omega_j](\boldsymbol{y})$, la valeur maximale de $Bet\,P^\Omega[\boldsymbol{y}](\omega_0)$ est obtenue pour l'hypothèse ω_0 pour laquelle $pl^Y[\omega_0](\boldsymbol{y})$ est maximale.

Si pour $\omega_k \in \Omega_0 \subseteq \Omega$ on a $l(\omega_k|\boldsymbol{y}) = 1$, alors chaque masse positive définie sur Ω est allouée à une partie de Ω_0 et les probabilités pignistiques données aux singletons $\omega_k \in \Omega_0$ sont égales et maximales. Dans le même temps, la plausibilité $pl^Y[\omega_k](\boldsymbol{y})$ est toujours inférieure ou égale à 1, de telle sorte que l'hypothèse $\omega_k \in \Omega_0$ est celle qui maximise la plausibilité, d'où la démonstration du théorème.

Patrick **Vannoorenberghe**

Patrick Vannoorenberghe obtient une thèse de doctorat de l'Université du Littoral-Côte d'Opale en juillet 1997. En 1998, il a été nommé maître de conférences à l'université de Rouen et a exercé ses activités de recherche au sein du laboratoire Perception Systèmes Information. Depuis septembre 2005, il enseigne à l'université Paul Sabatier, Toulouse3 dans le domaine du traitement du signal et des images. Il est chercheur au Laboratoire de Télédétection à Haute Résolution dans le domaine de l'analyse d'images et de la fusion d'information. Il a en charge le développement d'outils d'extraction de connaissances à partir d'images pour des applications spatiales et médicales. Ses domaines de compétence concernent le traitement d'images, la gestion des imprécisions et des incertitudes et la fusion d'informations. Il est l'auteur de nombreux papiers sur les fonctions de croyance, la segmentation d'images et la reconnaissance des formes.