RELATIONS ENTRE LES ALGORITHMES D'ESTIMATION ITERATIVES EM ET ICE AVEC EXEMPLES D'APPLICATION

Jean Pierre DELMAS

Dépt Signal et Image, Institut National des Télécommunications,
9 rue Charles Fourier, 91011 Evry cedex, France

RESUME

Le problème de l'estimation d'un paramètre multidimensionnel apparaît dans de nombreuses questions de traitement de signal. Dans ce contexte, l'estimateur du maximum de vraisemblance est souvent mis en oeuvre à l'aide de l'algorithme Expectation Maximization (EM). Le but de cet article est de comparer dans le cadre des structures statistiques exponentielles, cet algorithme à une technique alternative d'estimation itérative appelée Itérative Conditional Estimation (ICE), introduite par W.Pieczynski [1] en segmentation statistique d'image [2-3]. Nous montrons en particulier que contrairement à l'algorithme EM, qui est invariant selon le choix du paramétrage, l'algorithme ICE fournit selon ce choix un algorithme spécifique et parmi ceux-ci, il existe un paramétrage pour lequel EM et ICE sont équivalents. Ces choix de paramètres sont illustrés à l'aide de quelques exemples.

ABSTRACT

The problem of estimating a parameter vector appears in many aspects of signal processing. The maximum likelihood estimator is often used in this context with the help of the Expectation Maximization (EM) algorithm. In this paper, we compare for the exponential family of probability density functions, this algorithm to another iterative approach named Iterative Conditional Estimation (ICE), introduced by W.Pieczynski [1] in statistical segmentation of images [2-3]. We show in particular that, unlike the EM algorithm that is invariant to the parametrization, the ICE algorithm yields a specific algorithm for each parametrization. Furthermore, we show the existence of a specific parametrization for which the algorithms EM and ICE are equivalent. The choice of these parametrizations are illustrated by some signal processing examples.

1. INTRODUCTION

Dans une grande variété de problèmes rencontrés en traitement du signal, le calcul direct de l'estimateur du maximum de vraisemblance (MV) d'un paramètre multidimensionnel conduit à un calcul intractable. A l'aide de la notion de données complètes, les algorithmes EM et leurs extensions ont été massivement utilisés avec succès dans ce contexte en traitement du signal et de l'image.

Une technique alternative d'estimation itérative appelée Iterative Conditional Estimation (ICE) a été introduite par W.Pieczynski [1] et fut appliquée en segmentation statistique d'images [2-3]. Basée sur la notion d'espérance mathématique conditionnelle plutôt que sur celle de vraisemblance, cette approche paraît d'application plus large car elle permet de traiter le cas des lois de probabilité comportant une partie continue et une partie discrète pour lesquelles l'estimateur du maximum de vraisemblance perd son sens.

Après avoir reformulé les principes de cette approche dans le contexte de l'algorithme EM, nous comparons les algorithmes ICE et EM. Nous montrons en particulier que dans le cas d'une structure statistique exponentielle, contrairement à l'algorithme EM qui est invariant selon le choix du paramètre estimé, l'algorithme ICE dépend de ce choix et est équivalent à l'algorithme EM pour le paramètre canonique de la structure. L'algorithme EM apparaît ainsi dans le cadre de ces structures, comme un cas particulier de l'algorithme ICE. Nous terminerons par la présentation de quelques exemples puisés en traitement de signal.

2. ALGORITHME ICE

Appelons $\mathbf{y}$ le signal observé, réalisation d'une variable aléatoire $\mathbf{Y}$ dont la loi de probabilité dépend d'un paramètre multidimensionnel $\boldsymbol{\theta}$. Dans le contexte EM [4], ces données observées $\mathbf{y}$ sont complémentées pour former des "données complètes" $\mathbf{x}$ ($\mathbf{y} = h(\mathbf{x})$ avec h fonction non injective) pour obtenir plus aisément un estimateur MV de $\boldsymbol{\theta}$. L'algorithme EM alterne de façon itérative une étape d'espérance mathématique conditionnelle de la log-vraisemblance des données complètes et de maximisation de celle-ci par rapport à $\boldsymbol{\theta}$; le choix des données $\mathbf{x}$ étant guidé par la simplification de cette deuxième étape.

Le principe d'ICE introduit dans [1] dans le cadre de modèles cachés (champs de Markov) utilise aussi la notion de

"données complètes" $\mathbf{x}$, qu'il considère formées par $\mathbf{y}$ et des "données cachées" $\mathbf{x}'$ qu'on désirerait ensuite estimer : $\mathbf{x}=(\mathbf{x}',\mathbf{y})$. Reformulé dans la terminologie EM, le principe d'ICE introduit dans [1] est basé sur la notion d'espérance mathématique conditionnelle appliquée à un estimateur fonction des données complètes. Il repose sur les principes suivants :

- Nous supposons que nous disposons d'un estimateur de $\boldsymbol{\theta}$ fonction de $\mathbf{X}$ (pas nécessairement un estimateur MV) :

$$\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\mathbf{X}) \tag{1}$$

- Or puisque seul $\mathbf{y}$ est observable, nous devons rechercher une approximation de $\hat{\boldsymbol{\theta}}$ comme fonction de $\mathbf{Y}$. Nous proposons de choisir la meilleure approximation de $\hat{\boldsymbol{\theta}}$ comme fonction de $\mathbf{Y}$ au sens de l'erreur quadratique moyenne minimum, c'est à dire $E_{\boldsymbol{\theta}}[\hat{\boldsymbol{\theta}}(\mathbf{X})/\mathbf{Y}]$. Comme cette espérance conditionnelle dépend de $\boldsymbol{\theta}$, qui est par nature inconnue, l'approche itérative suivante a été proposée [1] :

$$\boldsymbol{\theta}_{k+1} = E_{\boldsymbol{\theta}_k}[\hat{\boldsymbol{\theta}}(\mathbf{X})/\mathbf{Y}=\mathbf{y}] \tag{2a}$$

- Si cette espérance conditionnelle ne peut être calculée analytiquement, mais si la loi conditionnelle $P_{\mathbf{X}/\mathbf{Y}}$ est connue, nous pouvons simuler N réalisations $\mathbf{x}_1,\mathbf{x}_2,..,\mathbf{x}_N$ de $\mathbf{X}$ selon la loi $P_{\mathbf{X}/\mathbf{Y}}$. $\boldsymbol{\theta}_{k+1}$ peut alors être approchée selon la loi des grands nombres par la moyenne empirique (en désignant $\mathbf{x}_i^k$ une réalisation de la variable aléatoire $\mathbf{X}$ selon la loi $P_{\mathbf{X}/\mathbf{Y}}$ pour la valeur $\boldsymbol{\theta}_k$ de $\boldsymbol{\theta}$). (En pratique, on peut utiliser une seule réalisation [1]). Nous obtenons :

$$\boldsymbol{\theta}_{k+1} = \frac{1}{N}[\hat{\boldsymbol{\theta}}(\mathbf{x}_1^k)+\hat{\boldsymbol{\theta}}(\mathbf{x}_2^k)+...+\hat{\boldsymbol{\theta}}(\mathbf{x}_N^k)] \tag{2b}$$

qui fournit une approximation stochastique de l'algorithme ICE.

3. CONNECTIONS ENTRE EM ET ICE

A première vue, les principes sur lesquels reposent les algorithmes ICE et EM sont complètement différents. Néanmoins ces algorithmes peuvent être comparés si nous utilisons les mêmes données complètes $\mathbf{x}$ et si l'estimateur (1) choisi dans ICE est l'estimateur MV. En effet si $f_{\mathbf{X}}(\mathbf{x};\boldsymbol{\theta})$ dénote la d.d.p. de $\mathbf{x}$, nous avons selon la méthodologie de EM :

$$\boldsymbol{\theta}_{k+1} = \operatorname{Arg}\operatorname{Max}_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}_k}[\operatorname{Log}(f_{\mathbf{X}}(\mathbf{X};\boldsymbol{\theta}))/\mathbf{Y}=\mathbf{y}] \tag{3}$$

et selon la méthodologie ICE nous avons d'après (2a) :

$$\boldsymbol{\theta}_{k+1} = E_{\boldsymbol{\theta}_k}[\operatorname{Arg}\operatorname{Max}_{\boldsymbol{\theta}} \operatorname{Log}(f_{\mathbf{X}}(\mathbf{X};\boldsymbol{\theta}))/\mathbf{Y}=\mathbf{y}] \tag{4}$$

En conséquence, si les deux opérations maximisation par rapport à $\boldsymbol{\theta}$ et espérance mathématique conditionnelle commutent, les deux algorithmes sont identiques. Nous allons préciser cette condition dans le cas où la d.d.p. des données complètes appartient à la famille exponentielle [5] (cas presque exclusif dans les applications rencontrées en traitement de signal). Soit $f_{\mathbf{X}}(\mathbf{x};\boldsymbol{\theta})$ une telle d.d.p. par rapport à une mesure non dépendant du paramètre conventionnel $\boldsymbol{\theta}\in\mathbb{R}^p$ (avec $a(\boldsymbol{\phi})\neq 0$) :

$$f_{\mathbf{X}}(\mathbf{x};\boldsymbol{\theta}) = b(\mathbf{x})e^{<\boldsymbol{\phi};\mathbf{s}(\mathbf{x})>-a(\boldsymbol{\phi})} \tag{5}$$

Naturellement la propriété d'invariance de l'estimateur du maximum de vraisemblance entraîne que l'algorithme EM est invariant selon le choix du paramètre $\mathbf{g}(\boldsymbol{\theta})$ estimé choisi (pourvu ici que $\mathbf{g}$ soit bijectif). Par contre l'algorithme ICE dépend de ce choix. Nous en donnerons des exemples dans la partie 4. Par suite, les algorithmes ICE et EM sont en général différents pour un choix quelconque de $\mathbf{g}(\boldsymbol{\theta})$. Nous allons montrer qu'ils sont toujours équivalents pour le paramètre canonique $\boldsymbol{\psi} = \nabla_{\boldsymbol{\phi}} a(\boldsymbol{\phi})$ (qui se confond quelquefois avec le paramètre conventionnel $\boldsymbol{\theta}$). En effet puisque :

$$E_{\boldsymbol{\psi}_k}[\operatorname{Log}(f_{\mathbf{X}}(\mathbf{X};\boldsymbol{\psi}))/\mathbf{Y}=\mathbf{y}] = \alpha(\boldsymbol{\psi}_k,\mathbf{y})+\boldsymbol{\phi}^{\mathrm{T}}E_{\boldsymbol{\psi}_k}[\mathbf{s}(\mathbf{X})/\mathbf{Y}=\mathbf{y}]-a(\boldsymbol{\phi})$$

L'étape E de EM consiste à calculer

$$\mathbf{s}_k(\mathbf{y}) \overset{\Delta}{=} E_{\boldsymbol{\psi}_k}[\mathbf{s}(\mathbf{X})/\mathbf{Y}=\mathbf{y}]$$

et l'étape M fournit : $\boldsymbol{\psi}_{k+1} = \mathbf{s}_k(\mathbf{y})$.
Quand à l'algorithme ICE, si nous choisissons comme estimateur (1), l'estimateur du maximum de vraisemblance :

$$\hat{\boldsymbol{\psi}} = \operatorname{Arg}\,[\operatorname{Max}_{\boldsymbol{\psi}} \operatorname{Log}(f_{\mathbf{X}}(\mathbf{X};\boldsymbol{\psi}))] = \mathbf{s}(\mathbf{X}).$$

L'itération d'ICE fournit d'après (2a) :

$$\boldsymbol{\psi}_{k+1} = E_{\boldsymbol{\psi}_k}[\mathbf{s}(\mathbf{X})/\mathbf{Y}=\mathbf{y}] = \mathbf{s}_k(\mathbf{y}) \tag{6}$$

ce qui est la même valeur que celle fournie par EM.

Nous allons maintenant montrer que les algorithmes EM et ICE soient équivalents pour le paramètre naturel $\boldsymbol{\phi}$ si et seulement si $a(\boldsymbol{\phi}) = \boldsymbol{\phi}^{\mathrm{T}}\mathbf{A}\boldsymbol{\phi}+\mathbf{b}^{\mathrm{T}}\boldsymbol{\phi}+c$. En effet, l'algorithme EM fournit :

$$\boldsymbol{\phi}_{k+1} = \operatorname{Arg}_{\boldsymbol{\phi}}\{E_{\boldsymbol{\phi}_k}[\mathbf{s}(\mathbf{X})/\mathbf{Y}=\mathbf{y}] = \nabla_{\boldsymbol{\phi}} a(\boldsymbol{\phi})\} = \mathbf{g}^{-1}\{E_{\boldsymbol{\phi}_k}[\mathbf{s}(\mathbf{X})/\mathbf{Y}=\mathbf{y}]\}$$

en dénotant $\mathbf{g}(\boldsymbol{\phi}) \overset{\Delta}{=} \nabla_{\boldsymbol{\phi}} a(\boldsymbol{\phi})$. Quant à l'algorithme ICE construit à partir de $\boldsymbol{\phi}(\mathbf{x})$ estimateur de vraisemblance, il fournit :

$$\boldsymbol{\phi}_{k+1} = E_{\boldsymbol{\phi}_k}\{\operatorname{Arg}_{\boldsymbol{\phi}}[\mathbf{s}(\mathbf{X}) = \nabla_{\boldsymbol{\phi}} a(\boldsymbol{\phi})]/\mathbf{Y}=\mathbf{y}\} = E_{\boldsymbol{\phi}_k}[\mathbf{g}^{-1}(\mathbf{s}(\mathbf{X}))/\mathbf{Y}=\mathbf{y}]$$

et puisque les opérations $E_{\boldsymbol{\phi}_k}$ et $\mathbf{g}^{-1}$ commutent si et seulement si $\mathbf{g}^{-1}(.)$ est affine $\Leftrightarrow$ $\mathbf{g}(.)$ est affine $\Leftrightarrow$

$$a(\boldsymbol{\phi}) = \boldsymbol{\phi}^{\mathrm{T}}\mathbf{A}\boldsymbol{\phi}+\mathbf{b}^{\mathrm{T}}\boldsymbol{\phi}+c \tag{7}$$

la propriété est démontrée.

4. EXEMPLES D'APPLICATION

Nous allons illustrer ce qui précède par plusieurs

exemples en commençant par deux cas dans lesquels les algorithmes EM et ICE sont équivalents pour le paramètre $\boldsymbol{\psi}$:

- Ex 1 : Mélange de q distributions gaussiennes.

Considérons une suite de n v.a. $(\mathbb{1}_1^i,..,\mathbb{1}_j^i,..,\mathbb{1}_{q-1}^i, Y_i)_{i=1,..,n}^{j=1,..,q-1}$ indépendantes, $\mathbb{1}_j^i$ étant la v.a. indicatrice de la distribution j à l'instant i. Chaque distribution j a pour probabilité α_j et Y_i a pour loi conditionnelle par rapport à $\mathbb{1}_j^i=1$ une loi gaussienne $\mathcal{N}(m_j,\sigma_j^2)$ pour $j=1,..,q$. Ici le paramètre conventionnel est

$$\boldsymbol{\theta} = (\alpha_1,\alpha_2,..,\alpha_{q-1},m_1,m_2,..,m_q,\sigma_1^2,\sigma_2^2,..,\sigma_q^2)$$

et la d.d.p. de $\mathbf{X} = (\mathbf{1}_1,\mathbf{1}_2,..,\mathbf{1}_{q-1},\mathbf{Y})$ (avec $\mathbf{1}_j \triangleq (\mathbb{1}_j^1,\mathbb{1}_j^2,..,\mathbb{1}_j^n)$ et $\mathbf{Y} \triangleq (Y_1,Y_2,..,Y_n)$) par rapport à la mesure produit de la mesure discrète sur $\{0,1\}^{n(q-1)}$ par la mesure de Lebesgue sur $\mathbb{R}^n$ appartient à la famille exponentielle avec pour paramètre naturel le vecteur $\boldsymbol{\phi}$ à $3q-1$ composantes suivant :

$$\begin{aligned}\boldsymbol{\phi} = n[&\mathrm{Log}(\frac{\alpha_1}{1-\alpha_1-...-\alpha_{q-1}})-\frac{m_1^2}{2\sigma_1^2}+\frac{m_q^2}{2\sigma_q^2}+\mathrm{Log}(\frac{\sigma_q}{\sigma_1}),\\ &\mathrm{Log}(\frac{\alpha_2}{1-\alpha_1-...-\alpha_{q-1}})-\frac{m_2^2}{2\sigma_2^2}+\frac{m_q^2}{2\sigma_q^2}+\mathrm{Log}(\frac{\sigma_q}{\sigma_2}),..,\\ &\mathrm{Log}(\frac{\alpha_{q-1}}{1-\alpha_1-...-\alpha_{q-1}})-\frac{m_{q-1}^2}{2\sigma_{q-1}^2}+\frac{m_q^2}{2\sigma_q^2}+\mathrm{Log}(\frac{\sigma_q}{\sigma_{q-1}}),\\ &\frac{m_1}{\sigma_1^2},\frac{m_2}{\sigma_2^2},..,\frac{m_q}{\sigma_q^2},-\frac{1}{2\sigma_1^2},-\frac{1}{2\sigma_2^2},..,-\frac{1}{2\sigma_q^2}]^{\mathrm{T}}\end{aligned}$$

$$\begin{aligned}\mathbf{s}(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^n\ [&\mathbb{1}_1^i,\mathbb{1}_2^i,..,\mathbb{1}_{q-1}^i,\\ &Y_i\mathbb{1}_1^i,Y_i\mathbb{1}_2^i,..,Y_i\mathbb{1}_{q-1}^i,Y_i(1-\mathbb{1}_1^i-\mathbb{1}_2^i-..-\mathbb{1}_{q-1}^i),\\ &Y_i^2\mathbb{1}_1^i,Y_i^2\mathbb{1}_2^i,..,Y_i^2\mathbb{1}_{q-1}^i,Y_i^2(1-\mathbb{1}_1^i-\mathbb{1}_2^i-..-\mathbb{1}_{q-1}^i)]^{\mathrm{T}}\end{aligned}$$

et $$a(\boldsymbol{\phi})=-\mathrm{Log}(1-\alpha_1-..-\alpha_{q-1})+\frac{m_q^2}{2\sigma_q^2}+\mathrm{Log}(\sigma_q)$$

D'après le résultat (6), les algorithmes EM et ICE sont équivalents pour le paramètre canonique $\boldsymbol{\psi}$ (nous rappelons que $\mathbf{s}(\mathbf{X})$ en est un estimateur sans biais).

$$\begin{aligned}\boldsymbol{\psi}=[&\alpha_1,\alpha_2,..,\alpha_{q-1},\\ &\alpha_1m_1,\alpha_2m_2,..,\alpha_{q-1}m_{q-1},(1-\alpha_1-..-\alpha_{q-1})m_q,\\ &\alpha_1(m_1^2+\sigma_1^2),\alpha_2(m_2^2+\sigma_2^2),..,\alpha_{q-1}(m_{q-1}^2+\sigma_{q-1}^2),(1-\alpha_1-..-\alpha_{q-1})(m_q^2+\sigma_q^2)]^{\mathrm{T}}\end{aligned}$$

dont les itérations communes sont d'après (6) :

$$\psi_{k+1}^j = \frac{1}{n}\sum_{i=1}^n \pi_k^{i,j} \quad j=1,..,q-1$$

$$\psi_{k+1}^{q-1+j} = \frac{1}{n}\sum_{i=1}^n y_i\pi_k^{i,j} \quad j=1,..,q$$

$$\psi_{k+1}^{2q-1+j} = \frac{1}{n}\sum_{i=1}^n y_i^2\pi_k^{i,j} \quad j=1,..,q$$

avec $\pi_k^{i,j} \triangleq P_k(\mathbb{1}_j^i=1/\mathbf{Y}_i=\mathbf{y}_i)$ donné par la formule de Bayes.

Par contre pour le paramètre $\boldsymbol{\theta}$, les algorithmes EM et ICE ne sont plus équivalents (en fait, ils le demeurent pour les composantes $\alpha_1,..,\alpha_{q-1}$ de $\boldsymbol{\theta}$). L'algorithme ICE fournit en utilisant pour estimateur (1) l'estimateur MV :

$$\hat{\boldsymbol{\theta}}=(\frac{U_1}{n},\frac{U_2}{n},..,\frac{U_{q-1}}{n},\frac{V_1}{U_1},..,\frac{V_q}{U_q},\frac{W_1}{U_1},..,\frac{W_q}{U_q})$$

avec pour $j=1,..,q$:

$$U_j \triangleq \sum_{i=1}^n \mathbb{1}_j^i,\ V_j \triangleq \sum_{i=1}^n y_i\mathbb{1}_j^i,\ W_j \triangleq \sum_{i=1}^n (y_i-\frac{V_j}{U_j})^2\mathbb{1}_j^i$$

les itérations :

$$\theta_{k+1}^q = \frac{1}{n}\sum_{i=1}^n \pi_k^{i,j} \quad \text{pour } j=1,..,q-1$$

quant aux paramètres θ_j pour $j=q,..,3q-1$, il est proposé dans [1] d'utiliser la version stochastique d'ICE (2b), car les expressions exactes de l'espérance conditionnelle sont trop complexes à calculer. Ainsi par exemple :

$$\theta_{k+1}^{q-1+j} = \sum_{\mathbf{1}_j\in\{0,1\}^n} \frac{V_j(\mathbf{1}_j,\mathbf{y})}{U_j(\mathbf{1}_j)} P_k(\mathbf{1}_j/\mathbf{Y}=\mathbf{y}) \quad \text{pour } j=1,..,q$$

- Ex 2 : Cas de classification incomplète.

En introduction de l'article [6], est cité un problème de classification comprenant 3 classes de probabilités θ_i, $\boldsymbol{\theta} = (\theta_1,\theta_2)$. On dispose de n observations indépendantes comprenant y_1 [resp. y_2] observations dans la classe 1 [resp. 2] et y_3 observations dans la classe 1 ou 2. En rajoutant la donnée inobservée x_1 : nombre total d'observations dans la classe 1, on obtient comme données complètes : $\mathbf{x} = (x_1,y_1,y_2,y_3)$ dont la d.d.p. par rapport à la mesure discrète sur $\{0,1,..,n\}^4$ appartient à la famille exponentielle avec :

$$\boldsymbol{\phi} = n[\mathrm{Log}(\frac{\theta_1}{1-\theta_1-\theta_2}),\mathrm{Log}(\frac{\theta_2}{1-\theta_1-\theta_2})]^{\mathrm{T}}$$

$$\mathbf{s}(\mathbf{X}) = \frac{1}{n}[X_1,\ Y_1+Y_2+Y_3-X_1]^{\mathrm{T}}$$

et $$a(\boldsymbol{\theta}) = -n\mathrm{Log}(1-\theta_1-\theta_2) \Rightarrow \boldsymbol{\psi} = \boldsymbol{\theta}$$

Dans ce cas les algorithmes EM et ICE sont équivalents pour le paramètre $\boldsymbol{\psi} = \boldsymbol{\theta}$.

Signalons un cas particulier de mélange gaussien dans lequel les paramètres naturels $\boldsymbol{\phi}$ et les paramètres canoniques $\boldsymbol{\psi}$ ne sont pas très pertinents et pour lequel les algorithmes EM et ICE sont différents pour le paramètre conventionnel $\boldsymbol{\theta}$.

- Ex 3 : La séparation de sources discrètes

dans un mélange bruité traité dans [8] pour lequel l'algorithme EM a été utilisé, n'est en fait qu'un cas particulier dégénéré de l'exemple 1. En effet considérons n observations de $\mathbb{R}^p$: $\mathbf{y}_i=\mathbf{M}\mathbf{x}'_i+\mathbf{b}_i$ pour $i=1,..,n$ avec $\mathbf{b}_i$ v.a. de loi gaussienne $\mathcal{N}(\mathbf{0},\mathbf{R})$, $\mathbf{x}'_i$ v.a. à m composantes dont chacune appartient à un alphabet connu, de taille r et de valeurs équiprobables (de sorte que $\mathbf{x}'_i$ prend $q=r^m$ valeurs équiprobables $\mathbf{a}_j\in\mathcal{A}$), les v.a. $\mathbf{x}'_i$ et $\mathbf{b}_i$ sont indépendantes et $\mathbf{M}$ est une matrice $p\times m$ de mélange inconnue. Par suite le paramètre conventionnel du modèle est $\boldsymbol{\theta}=[\mathbf{A},\mathbf{R}]$ et nous avons affaire à un mélange de q distributions gaussiennes équiprobables dans lesquelles $\mathbf{Y}_i$ a pour loi conditionnelle à $\mathbf{X}'_i=\mathbf{a}_j$ une loi gaussienne $\mathcal{N}(\mathbf{M}\mathbf{a}_j,\mathbf{R})$.

En appliquant les résultats précédents qui sont dégénérés car ici les probabilités a priori α_j sont connues, on obtient :

$$\boldsymbol{\phi}=n[\mathbf{R}^{-1}\mathbf{M}\mathbf{a}_1,\mathbf{R}^{-1}\mathbf{M}\mathbf{a}_2,..,\mathbf{R}^{-1}\mathbf{M}\mathbf{a}_q,-\frac{1}{2}\mathbf{R}^{-1}]^{\mathrm{T}}$$

$$\boldsymbol{\psi}=\frac{1}{q}[\mathbf{M}\mathbf{a}_1,\mathbf{M}\mathbf{a}_2,..,\mathbf{M}\mathbf{a}_q,\mathbf{M}\mathbf{a}_1\mathbf{a}_1^{\mathrm{T}}\mathbf{M}^{\mathrm{T}}+\mathbf{R},\mathbf{M}\mathbf{a}_2\mathbf{a}_2^{\mathrm{T}}\mathbf{M}^{\mathrm{T}}+\mathbf{R},..,\mathbf{M}\mathbf{a}_q\mathbf{a}_q^{\mathrm{T}}\mathbf{M}^{\mathrm{T}}+\mathbf{R}]^{\mathrm{T}}$$

$$\text{et } \mathbf{s}(\mathbf{X})=\frac{1}{n}\sum_{i=1}^{n}[\mathbf{Y}_i\mathbb{1}_1^i,\mathbf{Y}_i\mathbb{1}_2^i,..,\mathbf{Y}_i\mathbb{1}_q^i,\ \mathbf{Y}_i^2\mathbb{1}_1^i,\mathbf{Y}_i^2\mathbb{1}_2^i,..,\mathbf{Y}_i^2\mathbb{1}_q^i]^{\mathrm{T}}$$

et l'application des algorithmes EM et ICE au paramètre $\boldsymbol{\theta}$ conduit à deux algorithmes différents. En effet en utilisant pour estimateur (1), l'estimateur MV (on note ici $\mathbf{x}=(\mathbf{x}',\mathbf{y})$) :

$$\hat{\boldsymbol{\theta}}(\mathbf{X}) = [\mathbf{R}_{y,x'}\mathbf{R}_{x',x'}^{-1},\ \mathbf{R}_{y,y}-\mathbf{R}_{y,x'}\mathbf{R}_{x',x'}^{-1}\mathbf{R}_{y,x'}^{\mathrm{T}}]$$

avec
$\mathbf{R}_{y,y}\overset{\Delta}{=}\frac{1}{n}\sum_{i=1}^{n}\mathbf{Y}_i\mathbf{Y}_i^{\mathrm{T}}$, $\mathbf{R}_{y,x'}\overset{\Delta}{=}\frac{1}{n}\sum_{i=1}^{n}\mathbf{Y}_i\mathbf{X}_i'^{\mathrm{T}}$ et $\mathbf{R}_{x',x}\overset{\Delta}{=}\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i'\mathbf{X}_i'^{\mathrm{T}}$

l'algorithme ICE donne :

$$\boldsymbol{\theta}_{k+1}=\sum_{\mathbf{x}'\in\mathcal{A}^n}\hat{\boldsymbol{\theta}}(\mathbf{x}',\mathbf{y})\ P_k(\mathbf{x}'/\mathbf{Y}=\mathbf{y})$$

avec $P_k(\mathbf{x}'/\mathbf{Y}=\mathbf{y})\overset{\Delta}{=}\prod_{i\in 1,..,n}P_k(\mathbf{x}_i'/\mathbf{Y}_i=\mathbf{y}_i)$ donné par la formule de Bayes. Comme le calcul exact de $\boldsymbol{\theta}_{k+1}$ est trop complexe, on utilise la version stochastique d'ICE (2b). Quant à l'algorithme EM, il donne :

$$\boldsymbol{\theta}_{k+1} = [\mathbf{R}_{y,x'}^{(k)}\mathbf{R}_{x',x'}^{(k)-1},\ \mathbf{R}_{y,y}-\mathbf{R}_{y,x'}^{(k)}\mathbf{R}_{x',x'}^{(k)-1}\mathbf{R}_{y,x'}^{(k)\mathrm{T}}]$$

avec :

$$\mathbf{R}_{y,x'}^{(k)}\overset{\Delta}{=}\frac{1}{n}\sum_{i=1}^{n}[\sum_{\mathbf{a}\in\mathcal{A}}\mathbf{y}_i\mathbf{a}_j^{\mathrm{T}}\ P_k(\mathbf{x}_i'=\mathbf{a}_j/\mathbf{Y}_i=\mathbf{y}_i)]$$

$$\mathbf{R}_{x',x'}^{(k)}\overset{\Delta}{=}\frac{1}{n}\sum_{i=1}^{n}[\sum_{\mathbf{a}\in\mathcal{A}}\mathbf{a}_j\mathbf{a}_j^{\mathrm{T}}\ P_k(\mathbf{x}_i'=\mathbf{a}_j/\mathbf{Y}_i=\mathbf{y}_i)]$$

Puis nous présentons un cas où les algorithmes EM et ICE sont équivalents pour le paramètre $\boldsymbol{\phi}$ grâce à la relation (7).

- Ex 4 : Cas du modèle linéaire gaussien,

i.e. le cas où la loi de probabilité de $\mathbf{Y}$ est gaussienne $\mathcal{N}(\mathbf{H}\boldsymbol{\theta},\boldsymbol{\Sigma})$ où $\mathbf{H}=[\mathbf{h}_1,..,\mathbf{h}_p]$ et $\boldsymbol{\Sigma}$ sont connus [8], pour lequel on choisit $\mathbf{x}_i=\mathbf{h}_i\theta_i+\mathbf{b}_i$ avec $\mathbf{b}_i$ v.a. indépendantes entre elles de loi de probabilité gaussienne $\mathcal{N}(\mathbf{0},\alpha_i\boldsymbol{\Sigma})$ avec $\sum_{i=1}^{p}\alpha_i=1$. Ici $f_{\mathbf{X}}(\mathbf{x};\boldsymbol{\theta})$ est gaussienne, donc appartient à la famille exponentielle avec ici : $\boldsymbol{\phi}=\boldsymbol{\theta}$,

$$\mathbf{s}(\mathbf{x}) = [\mathbf{h}_1^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{x}_1,..,\mathbf{h}_p^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{x}_p]^{\mathrm{T}}$$

et $\quad a(\boldsymbol{\theta})=\frac{1}{2}\boldsymbol{\theta}^{\mathrm{T}}\mathrm{diag}[\mathbf{h}_i^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{h}_i]\boldsymbol{\theta}\ \Rightarrow \mathbf{g}(\boldsymbol{\theta})=\mathrm{diag}[\mathbf{h}_i^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{h}_i]\boldsymbol{\theta}$

en désignant par $\mathrm{diag}[a_i]$, la matrice diagonale dont les termes $[.]_{i,i}$ sont a_i, pour lequel l'algorithme commun s'écrit :

$$\boldsymbol{\theta}_{k+1}=\mathrm{diag}[\mathbf{h}_i^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{h}_i]^{-1}[\mathbf{h}_1^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathrm{E}_{\boldsymbol{\phi}_k}(\mathbf{X}_1/\mathbf{Y}=\mathbf{y}),..,\mathbf{h}_p^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathrm{E}_{\boldsymbol{\phi}_k}(\mathbf{X}_p/\mathbf{Y}=\mathbf{y})]^{\mathrm{T}}$$

avec $\mathrm{E}_{\boldsymbol{\phi}_k}(\mathbf{X}_i/\mathbf{Y}=\mathbf{y}) = \mathbf{h}_i\theta_k^i + \alpha_i[\mathbf{y}-\mathbf{H}\boldsymbol{\theta}_k]$ et l'on retrouve le résultat donné dans [7] :

$$\boldsymbol{\theta}_{k+1}=\boldsymbol{\theta}_k + \mathrm{diag}[\alpha_i(\mathbf{h}_i^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{h}_i)^{-1}]\mathbf{H}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}[\mathbf{y}-\mathbf{H}\boldsymbol{\theta}_k]$$

Enfin signalons un exemple n'appartenant pas à la famille exponentielle où les algorithmes EM et ICE ne sont pas équivalents.

- Ex 5 : Loi multimodale structurée.

En introduction de l'article [4] est cité un exemple de classification comportant 4 classes de probabilités $[\frac{1}{2}+\frac{\theta}{4}, \frac{1}{4}(1-\theta), \frac{1}{4}(1-\theta), \frac{\theta}{4}]$. On dispose de n observations d'effectifs respectifs (y_1,y_2,y_3,y_4). Il est considéré que la première classe est éclatée en deux sous-classes d'effectifs x_1' et x_2' de probabilités respectives $\frac{1}{2}$ et $\frac{\theta}{4}$. On obtient les données complètes $\mathbf{x}=(x_1',x_2',y_2,y_3,y_4)$ dont la d.d.p. par rapport à la mesure discrète sur $\{0,..,n\}^5$ se met sous la forme :

$$f_{\mathbf{X}}(\mathbf{x};\boldsymbol{\theta}) = b(\mathbf{x})\mathrm{e}^{(x_2'+y_4)\mathrm{Log}\theta+(y_2+y_3)\mathrm{Log}(1-\theta)}$$

donc n'appartient pas à la famille exponentielle. En utilisant l'estimateur MV de $\boldsymbol{\theta}$:

$$\hat{\boldsymbol{\theta}}(\mathbf{X}) = \frac{X_2'+X_4}{X_2'+Y_2+Y_3+Y_4}$$

et en observant que la loi conditionnelle de X_2' par rapport à $\mathbf{Y}=\mathbf{y}$ est une loi binomiale $(y_1,\ \theta/2+\theta)$, l'algorithme ICE fournit l'algorithme différent de EM suivant :

$$\theta_{k+1}=\sum_{i=1}^{y_1}\frac{i+y_4}{i+y_2+y_3+y_4}\,\mathrm{C}_{y_1}^{i}\,(\frac{\theta_k}{2+\theta_k})^{i}\,(\frac{2}{2+\theta_k})^{y_1-i}$$

REFERENCES

[1] W.Pieczynski, *Mixture of distributions, Markov random fields and unsupervised segmentation of images*, Technical Report nb.122, L.S.T.A., Université de Paris VI, November 1990.

[2] B.Braathen, W.Pieczynski & P.Masson, *Global and local methods of unsupervised Bayesian segmentation of images*, Machine Graphics & Vision - Vol. 2, No.1, pp.39-52, 1993.

[3] W.Pieczynski, *Champs de Markov cachés et estimation conditionnelle itérative*, Revue Traitement de signal 1994, Vol.11, No.2.

[4] A.Dempster, N. Laird & D.Rubin, *Maximum Likelihood Incomplete Data via EM Algorithm*, J.Roy.Statit.Soc.Ser.39, 1977, pp.1-38.

[5] R.Sundberg, *Maximum Likelihood Theory for Incomplete Data from an Exponential Family*, Scand Journal of Statistics 1, pp.49-58, 1974.

[6] N.Laird, *The EM Algorithm*, C.R.Rao, ed., Handbook of Statistics, Vol.9 Elsevier Science Publishers 1993.

[7] J.A.Fessler, A.O.Hero, *Complete-Data Space and generalized EM Algorithms*, Proceedings of the ICAASP 1993 IV pp.1-4.

[8] A.Belouchrani, J.F. Cardoso, *Maximum likelihood separation for discrete sources*, Proceedings of EUSIPCO 1994, pp.768-771.