An Equivalence of the EM and ICE Algorithm for Exponential Family

Jean-Pierre Delmas

An Equivalence of the EM and ICE Algorithm for Exponential Family

Jean Pierre Delmas

***Abstract*—In this correspondence, we compare the expectation maximization (EM) algorithm with another iterative approach, namely, the iterative conditional estimation (ICE) algorithm, which was formally introduced in the field of statistical segmentation of images. We show that in case the probability density function (PDF) belongs to the exponential family, the EM algorithm is one particular case of the ICE algorithm.**

I. Introduction

In many signal processing applications, direct calculations of maximum likelihood (ML) parameter vector estimates are intractable due to the complexity of the likelihood functions. Using the notion of complete data, the EM algorithm and its variations have been used extensively and successfully in many signal processing applications. An alternative iterative estimation method called the called iterative conditional estimation (ICE) was introduced by Pieczynski [1] in the field of statistical segmentation of images [2]–[4]. It is no longer based on the notion of likelihood but on that of conditional expectation. Therefore, this approach is of wider application because it encompasses probability distributions that have both a discrete and a continuous part, which is a case where the notion of likelihood is no longer relevant [5].

After formulating the principle of ICE reshaped in the familiar EM terminology, we compare the EM and ICE algorithms. We show in particular that for the exponential family of PDF's, unlike the EM algorithm, which is invariant to the parameterization, the ICE algorithm yields a specific algorithm for each parameterization. Furthermore, we show that the EM and ICE algorithms are equivalent for the canonical parameter of the structure. Therefore, the EM algorithm appears to be a particular case of the ICE algorithm for these structures. A similar fact has already been pointed out in the context of hidden Markov fields [4]. We conclude by illustrating these parameterizations in some signal processing examples.

II. The ICE Algorithm

Let $\mathbf{y}$ be a realization of the measured random variable (RV) $\mathbf{Y}$, the probability distribution of which depends on a parameter vector θ. In the classical formulation of the EM algorithm [6], [7], [9], one supplements the observed signals $\mathbf{y}$ (which are often called *incomplete data*) to form the *complete data* $\mathbf{x}$ [$\mathbf{y} = h(\mathbf{x})$, where h is a many-to-one mapping]. The EM algorithm iteratively alternates between an E-step, calculating the conditional expectation of the complete data log likelihood, and a M-step, maximizing that expectation with respect to the parameter θ. The mapping h (and, thus, the data $\mathbf{x}$) is chosen in such a way that this M-step is made as simple as possible.

The ICE algorithm, which was introduced in [1] in the context of hidden models (Markov fields), also uses the notion of complete data $\mathbf{x}$, which are now made of the concatenation of the available data $\mathbf{y}$ and of some hidden data $\mathbf{x}'$ that one would like to estimate $\mathbf{x} = (\mathbf{x}', \mathbf{y})$. This estimator is based on the notion of conditional expectation, which is applied to an estimator that is function of the complete data. Reformulated in the EM terminology, the principles of ICE are as follows.

- We suppose that we have at our disposal an estimator $\hat{\theta}$ that is a function of $\mathbf{X}$ ($\hat{\theta}$ *not necessarily* an ML estimator)

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

- Since only $\mathbf{y}$ is observable, we must look for an approximation of $\hat{\theta}$ that is a function of $\mathbf{Y}$. The best approximation in the minimum mean square error sense is a natural choice, i.e., $E_\theta[\hat{\theta}(\mathbf{X})/\mathbf{Y}]$, as this conditional expectation depends on θ, which is unknown by nature; therefore, the following iterative approach was proposed [1].

$$\theta_{k+1} = E_{\theta_k}\Big[\hat{\theta}(\mathbf{X})/\mathbf{Y} = \mathbf{y}\Big]. \tag{2a}$$

- If this conditional expectation cannot be computed analytically, but the conditional law $P_{\mathbf{X}/\mathbf{Y}}$ is known, then one can simulate N realizations $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N$ of $\mathbf{X}$ according to this distribution. θ_{k+1} can be approximated thanks to the law of large numbers by the empirical mean (in practice, one can use only one realization [1])

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

in which $\mathbf{x}_i^k$ denotes a realization of the RV $\mathbf{X}$ according to the law $P_{\mathbf{X}/\mathbf{Y}}$ for the value θ_k of θ. This yields a stochastic approximation of the ICE algorithm.

III. Relations between EM and ICE

At first sight, ICE and EM algorithms are based on completely different principles. Nevertheless, these algorithms can be compared if we use the same complete data $\mathbf{x}$ and if the estimator $\hat{\theta}$ in (1) is chosen to be the ML estimator. Therefore, if we denote by $f_{\mathbf{X}}(\mathbf{x};\theta)$ the PDF of $\mathbf{x}$, we have, according to the EM terminology

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

and in accordance with the ICE principle, we have, thanks to (2a)

$$\theta_{k+1} = E_{\theta_k}\Big[\text{Arg } \operatorname*{Max}_{\theta} \text{Log } f_{\mathbf{X}}(\mathbf{X};\theta)/\mathbf{Y} = \mathbf{y}\Big]. \tag{4}$$

Consequently, if the operations "maximization with respect to θ" and "conditional expectation" commute, the two algorithms become identical. We wish to clarify this condition in the case where the PDF of the complete data belongs the exponential family [8], which is a very frequent case in signal processing applications. Let $f_{\mathbf{X}}(\mathbf{x};\theta)$ be such a PDF with respect to a measure independent of the conventional parameter $\theta \in \mathbb{R}^p$ [with $a(\psi) \neq 0$]

$$f_{\mathbf{X}}(\mathbf{x};\theta) = b(\mathbf{x}) \exp[\langle \phi, s(\mathbf{x})\rangle - a(\phi)]. \tag{5}$$

As expected, the invariance property of the ML estimator implies that the EM algorithm does not depend on the choice of the selected parameter $\mathbf{g}(\theta)$, provided only that $\mathbf{g}$ is a one-to-one mapping. On the other hand, the ICE algorithm depends on this choice. We shall illustrate this point in Section IV. To show that in the exponential family, the EM algorithm is a particular case of the ICE algorithm, we now prove two propositions.

Manuscript received October 1, 1995; revised April 18, 1997. The associate editor coordinating the review of this paper and approving it for publication was Prof. Roger S. Cheng.

The author is with the Département Signal et Image, Institut National des Télécommunications, Evry, France (e-mail: delmas@int-evry.fr).

Publisher Item Identifier S 1053-587X(97)07350-9.

Proposition 1: The ICE and EM algorithms are equivalent for the canonical parameter $\psi = \nabla_\phi a(\phi)$ (which sometimes coincides with the conventional parameter θ).

Proof: Since

$$E_{\psi_k}[\text{Log } f_{\mathbf{X}}(\mathbf{X};\psi)/\mathbf{Y}=\mathbf{y}] = E_{\psi_k}\left[\frac{b(\mathbf{X})}{\mathbf{Y}}=\mathbf{y}\right] + \langle\phi,\, E_{\psi_k}[s(\mathbf{X})/\mathbf{Y}=\mathbf{y}]\rangle - a(\phi)$$

the E-step of EM consists of computing

$$\mathbf{s}_k(\mathbf{y}) \triangleq E_{\psi_k}[\mathbf{s}(\mathbf{X})/\mathbf{Y}=\mathbf{y}].$$

Since $E_{\psi_k}[b(\mathbf{X})/\mathbf{Y}=\mathbf{y}]$ does not depend on ψ, the M-step yields $\psi_{k+1} = \mathbf{s}_k(\mathbf{y})$. As for the ICE algorithm, if in (1) the ML estimator is chosen, then

$$\hat{\psi} = \text{Arg } \{\underset{\psi}{\text{Max}}\ \text{Log } [f_{\mathbf{X}}(\mathbf{X};\psi)]\} = \mathbf{s}(\mathbf{X})$$

and the iteration (2a) of ICE reads

$$\psi_{k+1} = E_{\psi_k}[\mathbf{s}(\mathbf{X})/\mathbf{Y}=\mathbf{y}] = \mathbf{s}_k(\mathbf{y}). \tag{6}$$

Proposition 2: The algorithms EM and ICE are equivalent for the natural parameter ϕ if and only if

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

Proof: The EM algorithm yields

$$\begin{aligned}\phi_{k+1} &= \underset{\phi}{\text{Arg}}\{E_{\phi_k}[\mathbf{s}(\mathbf{X})/\mathbf{Y}=\mathbf{y}] = \nabla_\phi a(\phi)\}\\ &= \mathbf{g}^{-1}\{E_{\phi_k}[\mathbf{s}(\mathbf{X})/\mathbf{Y}=\mathbf{y}]\}\end{aligned}$$

where $\mathbf{g}(\phi) \triangleq \nabla_\phi a(\phi)$. As for the ICE algorithm built on a ML estimator $\hat{\phi}(\mathbf{x})$, it gives

$$\begin{aligned}\phi_{k+1} &= E_{\phi_k}\left\{\underset{\phi}{\text{Arg}}[\mathbf{s}(\mathbf{X}) = \nabla_\phi a(\phi)]/\mathbf{Y}=\mathbf{y}\right\}\\ &= E_{\phi_k}\left[\mathbf{g}^{-1}\mathbf{s}(X)/\mathbf{Y}=\mathbf{y}\right].\end{aligned}$$

Last, the operations E_{ϕ_k} and $\mathbf{g}^{-1}$ commute if and only if $\mathbf{g}^{-1}(.)$ is affine $\Leftrightarrow$ $\mathbf{g}(.)$ is affine $\Leftrightarrow$ (7) holds.

IV. Application Examples

We now illustrate the choice of the different parameterizations by some signal processing examples. First of all, we clarify a case in which the EM and ICE algorithms are different for the parameter θ but equivalent for the parameter ψ.

Example 1—Finite Mixture of Gaussian Distributions: Consider a sequence of n independent RV's $(\mathbb{1}_1^i, \cdots, \mathbb{1}_j^i, \cdots, \mathbb{1}_{q-1}^i, Y_i)_{i=1,\cdots,n}^{j=1,\cdots,q-1}$, where $\mathbb{1}_j^i$ denotes the indicator variable of the distribution j at time i. Each distribution j has a probability α_j, and the conditional distribution of Y_i given $\mathbb{1}_j^i = 1$ is Gaussian $\mathcal{N}(m_j, \sigma_j^2)$ for $j = 1, \cdots, q$. The conventional parameter is

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

and it is easy to show that the PDF of $\mathbf{X} = (\mathbf{1}_1, \mathbf{1}_2, \cdots, \mathbf{1}_{q-1}, \mathbf{Y})$ [with $\mathbf{1}_j \triangleq (\mathbb{1}_j^1, \mathbb{1}_j^2, \cdots, \mathbb{1}_j^n)$ and $\mathbf{Y} \triangleq (Y_1, Y_2, \cdots, Y_n)$] with respect to the product measure of the discrete measure on $\{0,\ 1\}^{n(q-1)}$ by the Lebesgue measure on $\mathbb{R}^n$ belongs to the exponential family (5), where the vectors ϕ and $\mathbf{s}(\mathbf{X})$ with $3q-1$ components and $a(\phi)$ are, respectively, given by

$$\begin{aligned}\phi = n\Bigg[&\text{Log}\left(\frac{a_1}{1-\alpha_1-\cdots-\alpha_{q-1}}\right) - \frac{m_1^2}{2\sigma_1^2} + \frac{m_q^2}{2\sigma_q^2} + \text{Log}\left(\frac{\sigma_q}{\sigma_1}\right)\\ &\cdots, \text{Log}\left(\frac{\alpha_{q-1}}{1-\alpha_1-\cdots-\alpha_{q-1}}\right) + \text{Log}\left(\frac{\sigma_q}{\sigma_{q-1}}\right)\\ &-\frac{m_{q-1}^2}{2\sigma_{q-1}^2} + \frac{m_q^2}{2\sigma_q^2}\frac{m_1}{\sigma_1^2}, \cdots, \frac{m_q}{\sigma_q^2} - \frac{1}{2\sigma_1^2}, \cdots, -\frac{1}{2\sigma_q^2}\Bigg]^T\end{aligned}$$

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

and

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

According to (6), the EM and ICE algorithms are equivalent for the canonical parameter ψ. Since $\mathbf{s}(\mathbf{X})$ is an unbiased estimator of $\psi = \nabla_\phi a(\phi)$, we obtain immediately

$$\begin{aligned}\psi = [&\alpha_1, \cdots, \alpha_{q-1}, \alpha_1 m_1, \cdots, \alpha_{q-1}m_{q-1}\\ &(1-\alpha_1-\cdots-\alpha_{q-1})m_q\\ &\alpha_1(m_1^2+\sigma_1^2), \cdots, \alpha_{q-1}(m_{q-1}^2+\sigma_{q-1}^2)\\ &(1-\alpha_1-\cdots-\alpha_{q-1})(m_q^2+\sigma_q^2)]^T\end{aligned}$$

and the iterations (6), which are common to both algorithms, read

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

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

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

where $\pi_k^{i,j} \triangleq P_k(\mathbb{1}_j^i = 1/\mathbf{Y}_i = \mathbf{y}_i)$ is given by Bayes' rule.

The EM and ICE algorithms, however, are no longer equivalent for the parameter θ, although they keep this equivalence for the components $\alpha_1, \cdots, \alpha_{q-1}$ of θ. If the ML estimator is used as estimator (1), then

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

in which

$$\begin{aligned}U_j &\triangleq \sum_{i=1}^n \mathbb{1}_j^i, \quad V_j \triangleq \sum_{i=1}^n y_i\mathbb{1}_j^i,\\ W_j &\triangleq \sum_{i=1}^n \left(y_i - \frac{V_j}{U_j}\right)^2\mathbb{1}_j^i \quad j = 1, \cdots, q\end{aligned}$$

then, the ICE algorithm yields the iterations

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

As for the parameters θ_j for $j = q, \cdots, 3q-1$, it is proposed in [1] to use the stochastic version of ICE (2b) because the exact expressions of the conditional expectation are too complex to calculate. Therefore, for example

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

Therefore, EM and CE are different algorithms for the parameter θ but are equivalent for the parameter ψ.

Let us now point out a particular case of Gaussian mixture in which the natural parameter ϕ and the canonical parameters ψ are not very pertinent and for which the EM and ICE algorithms are different for the conventional parameter θ.

Example 2: The discrete source separation problem in a noisy mixture is treated in [11] by using the EM algorithm. This case is, in fact, a particular case of Example 1. To show this, consider n observations of $\mathbb{R}^p$: $\mathbf{y}_i = \mathbf{M}\mathbf{x}'_i + b_i$ for $i = 1, \cdots, n$, with $\mathbf{b}_i$ RV's of Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{R})$ and $\mathbf{x}'_i$ RV's, the m components of which are independent and taken from a known alphabet $\mathcal{A}$ of equally likely r values (so that $\mathbf{x}'_i$ takes equally likely $q = r^m$ values $\mathbf{a}_j \in \mathcal{A}$), the RV's $\mathbf{x}'_i$ and $\mathbf{b}_i$ are independent, and $\mathbf{M}$ is some $p \times m$ unknown mixing matrix. The conventional parameter of the model is $\theta = [\mathbf{M}, \mathbf{R}]$. Therefore, we have a mixture of q equally likely Gaussian distributions in which the conditional law of $\mathbf{Y}_i$ given $\mathbf{X}'_i = \mathbf{a}_j$ is $\mathcal{N}(\mathbf{M}\mathbf{a}_j, \mathbf{R})$. If we apply the preceding results (which here degenerate because the *a priori* probabilities α_j are known), we obtain

$$\phi = n\left[\mathbf{R}^{-1}\mathbf{M}\mathbf{a}_1, \mathbf{R}^{-1}\mathbf{M}\mathbf{a}_2, \cdots, \mathbf{R}^{-1}\mathbf{M}\mathbf{a}_q, \tfrac{1}{2}\mathbf{R}^{-1}\right]^T$$

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

and

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

The application of the EM and ICE algorithms to the parameter θ then lead to two different algorithms. If we use as estimator (1) the ML estimator [we denote here $\mathbf{x} = (\mathbf{x}', \mathbf{y})$]

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

with

$$\mathbf{R}_{y,y} \triangleq \frac{1}{n}\sum_{i=1}^{n}\mathbf{Y}_i\mathbf{Y}_i^T$$

$$\mathbf{R}_{y,x'} \triangleq \frac{1}{n}\sum_{i=1}^{n}\mathbf{Y}_i\mathbf{X}_i'^T$$

and

$$\mathbf{R}_{x',x'} \triangleq \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}'_i\mathbf{X}_i'^T$$

the ICE algorithm yields

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

where $P_k(\mathbf{x}'/\mathbf{Y} = \mathbf{y}) \triangleq \Pi_{i\in 1,\cdots,n}P_k(\mathbf{x}'_i/\mathbf{Y}_i = \mathbf{y}_i)$ is given by Bayes' rule. As the exact calculation of θ_{k+1} is too complicated, we use the stochastic version of ICE (2b). As for the EM algorithm, it yields

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

with

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

and

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

Finally, we present a case where the EM and ICE algorithms are equivalent for the parameter ϕ thanks to (7).

Example 3—Linear Gaussian Model Case: In this case, the distribution of $\mathbf{Y}$ is Gaussian $N(\mathbf{H}\theta, \Sigma^{-1})$, where $\mathbf{H} = [\mathbf{h}_1, \cdots, \mathbf{h}_p]$ and Σ are known [10], for which $\mathbf{x}_i = \mathbf{h}_i\theta_i + \mathbf{b}_i$ is chosen with $\mathbf{b}_i$ independent RV's of Gaussian distribution $N(\mathbf{0}, \alpha_i\Sigma)$ with $\sum_{i=1}^{p}\alpha_i = 1$. Here, $f_{\mathbf{X}}(\mathbf{x};\theta)$ is Gaussian, which, therefore, belongs to the exponential family with $\phi = \theta$

$$\mathbf{s}(\mathbf{x}) = \left[\mathbf{h}_1^T\Sigma^{-1}\mathbf{x}_1, \cdots, \mathbf{h}_p^T\Sigma^{-1}\mathbf{x}_p\right]^T$$

and

$$a(\theta) = \tfrac{1}{2}\theta^T\,\mathrm{diag}\left[\mathbf{h}_i^T\Sigma^{-1}\mathbf{h}_i\right]\theta \Rightarrow \mathbf{g}(\theta) = \mathrm{diag}\left[\mathbf{h}_i^T\Sigma^{-1}\mathbf{h}_i\right]\theta$$

in which diag $[a_i]$ denotes the diagonal matrix whose $[\,]_{i,i}$ entry is a_i. Therefore, the common algorithm reads

$$\theta_{k+1} = \mathrm{diag}\left[\mathbf{h}_i^T\Sigma^{-1}\mathbf{h}_i\right]^{-1} \cdot \left[\mathbf{h}_1^T\Sigma^{-1}E_{\phi_k}(\mathbf{X}_1/\mathbf{Y} = \mathbf{y}), \cdots, \mathbf{h}_p^T\Sigma^{-1}E_{\phi_k}(\mathbf{X}_p/\mathbf{Y} = \mathbf{y})\right]^T$$

with $E_{\phi_k}(\mathbf{X}_i/\mathbf{Y} = \mathbf{y}) = \mathbf{h}_i\theta_k^i + \alpha_i[\mathbf{y} - \mathbf{H}\theta_k]$. Then, we recognize the given result in [10].

$$\theta_{k+1} = \theta_k + \mathrm{diag}\left[\alpha_i\left(\mathbf{h}_i^T\Sigma^{-1}\mathbf{h}_i\right)^{-1}\right]\mathbf{H}^T\Sigma^{-1}[y - \mathbf{H}\theta_k].$$

ACKNOWLEDGMENT

The author wishes to thank W. Pieczynski and F. Desbouvries for useful discussions and comments.

REFERENCES

[1] W. Pieczynski, "Statistical image segmentation," *Machine Graphics Vision,* vol. 1, nos. 1/2, 1992.
[2] B. Braathen, W. Pieczynski, and P. Masson, "Global and local methods of unsupervised Bayesian segmentation of images," *Machine Graphics Vision* vol. 2, no. 1, pp. 39–52, 1993.
[3] A. Peng and W. Pieczynski, "Adaptive mixture estimation and unsupervised local Bayesian image segmentation," *Graphical Models Image Process.*, vol. 57, no. 5, 1995.
[4] W. Pieczynski, "Champs de Markov cachés et estimation conditionnelle itérative," *Rev. Traitement Signal*, vol. 11, no. 2, pp. 142–143, 1995.
[5] H. Caillol, W. Pieczynski, and A. Hillion, "Estimation of fuzzy Gaussian mixture and unsupervised statistical image segmentation," *IEEE Trans. Image Processing*, vol. 6, pp. 425–440, 1997.
[6] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood incomplete data via EM algorithm," *J. Roy. Stat. Soc.*, ser. 39, pp. 1–38, 1977.
[7] R. A. Redner and H. F. Walker, "Mixture densities, maximum likelihood and the EM algorithm," *SIAM Rev.*, vol. 26, pp. 195–239, 1984.
[8] R. Sundberg, "Maximum likelihood theory for incomplete data from an exponential family," *Scand. J. Stat.* vol. 1, pp. 49–58, 1974.
[9] N. Laird, "The EM algorithm," *Handbook of Statistics*, C. R. Rao, Ed. New York: Elsevier, 1993, vol. 9.
[10] J. A. Fessler and A. O. Hero, "Complete-data space and generalized EM algorithms," in *Proc. ICASSP*, vol. 4, 1993, pp. 1–4.
[11] A. Belouchrani and J. F. Cardoso, "Maximum likelihood separation for discrete sources," in *Proc. EUSIPCO*, 1994, pp. 768–771.