Unsupervised image segmentation using a telegraph parameterization of Pickard random fields

Jérôme Idier, Yves Goussard and Andrea Ridolfi

Abstract

This communication presents a non-supervised three-dimensional segmentation method based upon a discrete-level unilateral Markov field model for the labels and conditionaly Gaussian densities for the observed voxels. Such models have been shown to yield numerically efficient algorithms, for segmentation and for estimation of the model parameters as well. Our contribution is twofold. First, we deal with the degeneracy of the likelihood function with respect to the parameters of the Gaussian densities, which is a well-known problem for such mixture models. We introduce a bounded *penalized* likelihood function that has been recently shown to provide a consistent estimator in the simpler cases of independent Gaussian mixtures. On the other hand, implementation with EM reestimation formulas remains possible with only limited changes with respect to the standard case. Second, we propose a *telegraphic* parameterization of the unilateral Markov field. On a theoretical level, this parameterization ensures that some important properties of the field (e.g., stationarity) do hold. On a practical level, it reduces the computational complexity of the algorithm used in the segmentation and parameter estimation stages of the procedure. In addition, it decreases the number of model parameters that must be estimated, thereby improving convergence speed and accuracy of the corresponding estimation method.

I. Introduction

In this paper, we present a method for segmenting images modeled as N-ary Markov random fields (MRFs). Such image representations have proved useful for segmentation because they can explicitly model important features of actual images, such as the presence of homogeneous regions separated by sharp discontinuities. However, Markov-based segmentation methods are often computationally intensive and therefore difficult to apply in a three-dimensional (3D) context. In addition, specification of the MRF parameter values is often difficult to perform. This can be done in a heuristic manner, but such an approach is strongly application-dependent and becomes very burdensome for complex models (i.e., large neighborhoods and large number of levels). Deriving *unsupervised* methods in which the MRF parameters are estimated from the observed data is more satisfactory, but such a task generally requires approximations in order to be mathematically feasible [1], and the corresponding amount of computations is generally much higher than for a segmentation operation.

In order to overcome such difficulties, Devijver and Dekesel [2] proposed an unsupervised segmentation approach based on a hidden markov model (HMM) that belongs to a special class of *unilateral* MRFs: Pickard random fields (PRFs). The PRF is observed through an independent Gaussian process and the labels as well as the model parameters are estimated using maximum likelihood techniques. Because of the specific properties of PRF models, a significant reduction of the computational burden is achieved, and application of such methods to 3D problems can be envisioned. However, three kinds of difficulties remain: firstly, from a theoretical standpoint, the estimated MRF parameters are not necessarily consistent with the assumed stationarity of the model; secondly, the likelihood function of the observation parameters presents attractive singular points, as it is well-known in Gaussian mixture identification problems [3], and this hinders the convergence of the estimation procedure; thirdly, the convergence of the estimation procedure is made even more difficult by the fairly large number of parameters that need to be estimated.

Yves Goussard is with École Polytechnique, Biomedical Engineering Institute, C.P. 6079, Station Centre-Ville, Montréal (Québec) H3C 3A7, Canada. Jérôme Idier and Andrea Ridolfi are with Laboratoire des Signaux et Systèmes, École Supérieure d'Électricité, Plateau de Moulon, 91192 Gif-sur-Yvette Cedex, France.

Here, we present a segmentation method that extends the technique introduced by Devijver and Dekesel and corrects some of its deficiencies. First, the method is based upon a parsimonious parameterization of PRFs, referred to as a *telegraph model*, which simplifies the parameter estimation procedure, speeds up its convergence and ensures that some necessary conditions (such as marginal stationarity of the rows and columns) are fulfilled. Second, the singularity of the likelihood function of the parameters is dealt with by using a well-behaved *penalized* likelihood function that lends itself to the derivation of an efficient maximization procedure. Therefore, the resulting unsupervised segmentation method presents a safe overall convergence and exhibits a moderate amount of computations, which makes it suitable to process large 3D images as illustrated in the sequel.

II. Approach

Throughout the paper, random variables and realizations of thereof are respectively denoted by uppercase and corresponding lowercase symbols; in addition, notations such as $f(\boldsymbol{y}\,|\,\boldsymbol{x})$ and $\Pr(\boldsymbol{x}\,|\,\boldsymbol{y})$ are employed as shorthands for $f_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}\,|\,\boldsymbol{x})$ and $\Pr(\boldsymbol{X}=\boldsymbol{x}\,|\,\boldsymbol{Y}=\boldsymbol{y})$, whenever unambiguous.

As stated in the introduction, the image to be segmented is modeled as a hidden N-ary PRF $\boldsymbol{X}$. N-ary PRFs were studied by Pickard in a two-dimensional (2D) framework [4], [5]; these fields are stationary and their joint probability is determined by a measure τ on a four-pixel elementary cell $\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)$ that must fulfill several symmetry and independence conditions [5]. Conversely, it is shown in [6] that stationary MRFs on a finite rectangular lattice can be characterized by their marginal distribution on a four-pixel elementary cell, and that in some important cases (Gaussian fields, symmetric fields), the only stationary fields are PRFs. As a consequence of the stationarity of $\boldsymbol{X}$, the marginal probability of each row and column presents the structure of a stationary and reversible Markov chain whose initial and transition probabilities can be easily deduced from τ. According to [7], most of the latter results have three-dimensional (3D) counterparts that apply to 3D PRFs.

Here, we assume that the observations $\boldsymbol{y}$ of PRF $\boldsymbol{X}$ fulfill the following properties:

$$f(\boldsymbol{y}\,|\,\boldsymbol{x}) \;=\; \prod_{i,j} f(y_{\{i,j\}}\,|\,x_{\{i,j\}})\,, \tag{1}$$

$$f(y_{\{i,j\}}\,|\,X_{\{i,j\}}=n) \;=\; \mathcal{G}_n\,, \tag{2}$$

where i, j and $n \in \{1\,,\ldots\,,N\}$ respectively denote the row, column and state indices, and where $\mathcal{G}_n$ represents the Gaussian density with mean u_n and variance v_n.

These assumptions correspond to the common situation of an image degraded by independent Gaussian noise and, as underlined by Devijver and Dekesel [2] in a 2D context, they are well suited to marginal maximum *a posteriori* (MMAP) segmentation of $\boldsymbol{X}$ as well as to maximum likelihood (ML) estimation of the PRF and noise parameters. The key to the derivation of numerically efficient segmentation algorithms is the approximation

$$\Pr(x_{\{i,j\}}\,|\,\boldsymbol{y}) \approx \Pr(x_{\{i,j\}}\,|\,\boldsymbol{y}_{\{i,\cdot\}},\boldsymbol{y}_{\{\cdot,j\}})\,, \tag{3}$$

where $\boldsymbol{y}_{\{i,\cdot\}}$ and $\boldsymbol{y}_{\{\cdot,j\}}$ respectively denote i-th row and j-th column of $\boldsymbol{y}$. The above approximation amounts to neglecting interactions in the diagonal directions and to rely only on interactions in the horizontal and vertical directions. This may cause a lower accuracy for segmentation of objects with diagonally-oriented boundaries. However, this effect is not severe, as shown in [2] and in Section VII of this article, and with this approximation, the marginal posterior likelihood only involves 1D restrictions of $\boldsymbol{y}$ which present Markov chain structures. In order to take advantage of this property, Bayes rule is applied to (3) and the orthogonality properties of measure τ yield the following simplified expression

$$\Pr(x_{\{i,j\}}\,|\,\boldsymbol{y}_{\{i,\cdot\}},\boldsymbol{y}_{\{\cdot,j\}}) \propto f(\boldsymbol{y}_{\{i,\cdot\}}\,|\,x_{\{i,j\}})\,f(\boldsymbol{y}_{\{\cdot,j\}}\,|\,x_{\{i,j\}})\,\Pr(x_{\{i,j\}})\,. \tag{4}$$

The above expression only involves 1D quantities; this has two important consequences. First, due to the Markov chain structures of $\boldsymbol{X}_{\{i,\cdot\}}$ and $\boldsymbol{X}_{\{\cdot,j\}}$, the first two terms of the right hand side of (4) can be evaluated in an efficient manner by means of 1D forward-backward algorithms. Second, the only parameters of interest in the *a priori* PRF model are those which control the distribution of rows and columns $\boldsymbol{X}_{\{i,\cdot\}}$ and $\boldsymbol{X}_{\{\cdot,j\}}$ thereby simplifying the parameter estimation stage outlined below.

The PRF representation associated with assumptions (1)-(2) is also well suited to ML estimation of the model parameter vector $\boldsymbol{\theta}$. The ML estimate $\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} f(\boldsymbol{y};\boldsymbol{\theta})$ cannot be expressed in closed form. Devijver and Dekesel [2] proposed to evaluate $\boldsymbol{\theta}$ through maximization of the following criterion:

$$J(\boldsymbol{y};\boldsymbol{\theta}) \propto \prod_i f(\boldsymbol{y}_{\{i,\cdot\}};\boldsymbol{\theta}) \prod_j f(\boldsymbol{y}_{\{\cdot,j\}};\boldsymbol{\theta})\,, \tag{5}$$

They showed that iterative maximization of J can be carried out by an expectation-maximization (EM) algorithm and that the quantities required for the EM iterations can be evaluated by the same forward-backward procedures as the ones used for segmentation of $\boldsymbol{X}$. Even though Devijver and Dekesel presented J as a mere approximation of the exact likelihood function, it is clear by inspection that J can be interpreted as a generalization of the pseudo-likelihood function proposed in [8]. More generally, we conjecture that the above estimator can be cast within the framework of minimum contrast estimation and that its convergence and consistency properties can be investigated with techniques similar to those presented in [9, pp.157–162].

This method, in the form proposed by Devijver and Dekesel [2], proved to provide interesting segmentation results in a non supervised framework at a reasonable computational cost. It nonetheless presents several limitations and deficiencies. First, it is limited to 2D problems; second, the distributions of $\boldsymbol{X}_{\{i,\cdot\}}$ and $\boldsymbol{X}_{\{\cdot,j\}}$ are parameterized in a standard manner by the initial and transition probabilities. Consequently, the stationarity and reversibility of each row and column of PRF $\boldsymbol{X}$ is not guaranteed. In addition, $O(N^2)$ parameters must be estimated, which requires a significant amount of computations and induces convergence difficulties, even for moderate numbers of states; third, the likelihood function used for estimation of $\boldsymbol{\theta}$ presents singular points. Intuitively, this is caused by the normal densities $f(\boldsymbol{y}_{\{i,\cdot\}} \,|\, \boldsymbol{x}_{\{i,\cdot\}};\boldsymbol{\theta})$ which enter the right hand side of (5) through decompositions of the form:

$$f(\boldsymbol{y}_{\{i,\cdot\}};\boldsymbol{\theta}) = \sum_{\boldsymbol{x}_{\{i,\cdot\}}} \prod_j f(y_{\{i,j\}} \,|\, x_{\{i,j\}};\boldsymbol{\theta})\ \Pr(\boldsymbol{x}_{\{i,\cdot\}};\boldsymbol{\theta})\,, \tag{6}$$

and which degenerate when $x_{\{i,j\}} = n$, $u_n = y_{\{i,j\}}$ and $v_n \searrow 0$ for some n, j. For estimation of parameters of mixtures of Gaussian densities, this behavior is well known and well documented [3], [10]. The consequence of this degeneracy is the divergence of the EM procedure if a reestimated value of $\boldsymbol{\theta}$ reaches a neighborhood of any singular point.

The main purpose of this article is to propose several extensions and refinements of the segmentation method introduced by Devijver and Dekesel [2] in order to alleviate its main limitations. The major improvements are

1. extension of the technique to a three dimensional (3D) framework;
2. correction of the degeneracy of the likelihood function through adjunction of an appropriate penalization function, while retaining the possibility of estimating the model parameters with an EM procedure in a slightly modified form;
3. parameterization of the 1D restrictions of $\boldsymbol{X}$ with a *telegraph model* (TM) which guarantees their stationarity and reversibility while remaining compatible with the EM procedure used for model estimation. In addition, convergence of the EM procedure is improved by the reduced dimension ($O(N)$) of the TM parameter vector with respect to standard parameterization of Markov chains.

Before addressing these three points, we briefly recall the equations of the EM algorithm and derive two properties that will simplify the subsequent derivations.

III. EM Reestimation formulas for parameter estimation

A. General EM procedure

The EM algorithm is an iterative procedure which increases the likelihood $f(\boldsymbol{y}\,;\boldsymbol{\theta})$ of a parameter vector $\boldsymbol{\theta}$ given observations $\boldsymbol{y}$ at each iteration. Starting from an initial value $\boldsymbol{\theta}^0$, a series of successive estimates $\boldsymbol{\theta}^k$ is generated by alternating the following two steps:

$$\text{Expectation (E):} \quad \text{Evaluate } Q(\boldsymbol{\theta},\boldsymbol{\theta}^k\,;\boldsymbol{y})\,, \tag{7}$$

$$\text{Maximization (M):} \quad \boldsymbol{\theta}^{k+1}=\arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta},\boldsymbol{\theta}^k\,;\boldsymbol{y})\,, \tag{8}$$

where the function Q is defined as

$$Q(\boldsymbol{\theta},\boldsymbol{\theta}^0\,;\boldsymbol{y}) \;\triangleq\; \sum_{\boldsymbol{x}} \Pr(\boldsymbol{x}\,|\,\boldsymbol{y}\,;\boldsymbol{\theta}^0)\log(f(\boldsymbol{y}\,|\,\boldsymbol{x}\,;\boldsymbol{\theta})\Pr(\boldsymbol{x}\,;\boldsymbol{\theta}))\,, \tag{9}$$

$$= \; E[\log(f(\boldsymbol{y}\,|\,\boldsymbol{X}\,;\boldsymbol{\theta})\Pr(\boldsymbol{X}\,;\boldsymbol{\theta}))\,|\,\boldsymbol{y}\,;\boldsymbol{\theta}^0]\,, \tag{10}$$

$\boldsymbol{X}$ being an auxiliary variable whose practical role is to make the *extended likelihood* $f(\boldsymbol{y}\,|\,\boldsymbol{x}\,;\boldsymbol{\theta})\Pr(\boldsymbol{x}\,;\boldsymbol{\theta})$ easier to compute than the original likelihood $f(\boldsymbol{y}\,;\boldsymbol{\theta})$. The above equations are given for a continuous-valued variable $\boldsymbol{y}$ and a discrete-valued auxiliary variable $\boldsymbol{x}$, as this corresponds to our application. Transposition to a discrete-valued $\boldsymbol{y}$ and/or a continuous-valued $\boldsymbol{x}$ is straightforward. In all cases, the EM algorithm can be shown to increase the likelihood at each iteration and to converge to a critical point of the likelihood function $f(\boldsymbol{y}\,;\boldsymbol{\theta})$. A detailed analysis of the properties of the EM algorithm can be found in [11] in the context of hidden Markov chains and in [12], [10] in a more general framework. Here, we provide the equations of a complete EM algorithm for estimation of the parameters of a 1D HMM, as this will be the base for the derivations in Sections V and VI. The hidden Markov chains $X_{\{i,\cdot\}}$ and $X_{\{\cdot,j\}}$ are discrete-valued and in accordance with assumptions (1)-(2) the observations $Y_{\{i,\cdot\}}$ and $Y_{\{\cdot,j\}}$ are conditionally independent and Gaussian. For a generic discrete-valued hidden Markov chain $X_t \in \{1,\ldots,N\}$, $1\leq t\leq T$, with conditionally independent Gaussian observations y_t, $1\leq t\leq T$, the equations of the complete EM algorithm are given in Table II, in compliance with the compact notations defined in Table I. It should be underlined that quantity $p^0_{t,n}$ computed by the forward-backward algorithm is precisely the marginal likelihood $\Pr(X_t=n\,|\,\boldsymbol{y})$ used for estimation of $\boldsymbol{X}$. This illustrates the point made in Section II that the forward-backward algorithm is the basic tool for both the segmentation step and the parameter estimation step.

B. Decoupling of the M step

Assume that parameter vector $\boldsymbol{\theta}$ can be partitioned into two subvectors $\boldsymbol{\theta}_{Y|X}$ and $\boldsymbol{\theta}_X$ which respectively control the conditional probability function $f(\boldsymbol{y}|\boldsymbol{x})$ and the probability distribution $\Pr(\boldsymbol{x})$. Such a situation is commonly encountered and can be taken advantage of in order to decouple the M step of the EM algorithm into two — hopefully simpler — independent maximization problems.

Under these assumptions, the probability product which enters the definition of Q in (9) can be expressed as

$$f(\boldsymbol{y}\,|\,\boldsymbol{x}\,;\boldsymbol{\theta})\;\Pr(\boldsymbol{x}\,;\boldsymbol{\theta}) = f(\boldsymbol{y}\,|\,\boldsymbol{x}\,;\boldsymbol{\theta}_{Y|X})\;\Pr(\boldsymbol{x}\,;\boldsymbol{\theta}_X)\,. \tag{11}$$

For any set value of parameter vector $\boldsymbol{\theta}^0$, define functions $Q_{Y|X}$ and Q_X as

$$Q_{Y|X}(\boldsymbol{\theta}_{Y|X},\boldsymbol{\theta}^0\,;\boldsymbol{y}) \;\triangleq\; \sum_{\boldsymbol{x}} Pr(\boldsymbol{x}\,|\,\boldsymbol{y}\,;\boldsymbol{\theta}^0)\log f(\boldsymbol{y}\,|\,\boldsymbol{x}\,;\boldsymbol{\theta}_{Y|X})\,, \tag{12}$$

$$Q_X(\boldsymbol{\theta}_X,\boldsymbol{\theta}^0\,;\boldsymbol{y}) \;\triangleq\; \sum_{\boldsymbol{x}} \Pr(\boldsymbol{x}\,|\,\boldsymbol{y}\,;\boldsymbol{\theta}^0)\log\Pr(\boldsymbol{x}\,;\boldsymbol{\theta}_X)d\boldsymbol{x}\,. \tag{13}$$

$$
\begin{array}{lll}
\boldsymbol{y} = [y_1, \ldots, y_T]^{\mathrm{t}}, & & \boldsymbol{y}_s^t = [y_s, \ldots, y_t]^{\mathrm{t}}, \\
\multicolumn{3}{c}{\mathcal{G}_n = f(y_t \,|\, X_t = n) = (2\pi v_n)^{-1/2} \exp\left[-(y-u_n)^2/2v_n\right],} \\
p_n = \Pr(X_1 = n \,;\, \boldsymbol{\theta}), & & P_{mn} = \Pr(X_t = n \,|\, X_{t-1} = m \,;\, \boldsymbol{\theta}), \\
p_n^0 = \Pr(X_1 = n \,;\, \boldsymbol{\theta}^0), & & P_{mn}^0 = \Pr(X_t = n \,|\, X_{t-1} = m \,;\, \boldsymbol{\theta}^0), \\
p_{t,n}^0 = \Pr(X_t = n \,|\, \boldsymbol{y} \,;\, \boldsymbol{\theta}^0), & & p_{t,mn}^0 = \Pr(X_{t-1} = m, X_t = n \,|\, \boldsymbol{y} \,;\, \boldsymbol{\theta}^0), \\
\alpha_n^0 = \sum_{t=1}^T p_{t,n}^0, & \beta_n^0 = \sum_{t=2}^{T-1} p_{t,n}^0, & s_n^0 = \sum_{t=2}^T p_{t,nn}^0, \\
\eta_n^0 = (\alpha_n^0 + \beta_n^0)/2, & & \gamma_n^0 = \eta_n^0 - s_n^0, \\
\multicolumn{3}{l}{\mathcal{F}_{t,n} = P(X_t = n \,|\, \boldsymbol{y}_1^t \,;\, \boldsymbol{\theta}^0),\ \mathcal{N}_{t,n} = f(y_t \,|\, \boldsymbol{y}_{1,T-1}),\ \mathcal{B}_{t,n} = \dfrac{f(\boldsymbol{y}_{t+1}^T \,|\, X_t = n \,;\, \boldsymbol{\theta}^0)}{f(\boldsymbol{y}_{t+1}^T \,|\, \boldsymbol{y}_1^t \,;\, \boldsymbol{\theta}^0)}.}
\end{array}
$$

TABLE I
NOTATIONS.

It can be immediately deduced from (9) and (11) that function Q can be expressed as

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^0 \,;\, \boldsymbol{y}) = Q_{Y|X}(\boldsymbol{\theta}_{Y|X}, \boldsymbol{\theta}^0 \,;\, \boldsymbol{y}) + Q_X(\boldsymbol{\theta}_X, \boldsymbol{\theta}^0 \,;\, \boldsymbol{y}), \tag{14}$$

which shows that the M step of the EM algorithm can be decoupled into two operations: maximization of $Q_{Y|X}$ with respect to $\boldsymbol{\theta}_{Y|X}$ and maximization of Q_X with respect to $\boldsymbol{\theta}_X$.

C. Independent realizations

Another special case of interest occurs when $\boldsymbol{y}$ is made up of independent realizations $\boldsymbol{y}_i \,;\, 1 \leq i \leq I$. For instance, this corresponds to the case of the pseudo-likelihood defined in (5). As a consequence, the corresponding auxiliary processes $\boldsymbol{X}_i$ are also independent and it is not difficult to obtain

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^0 \,;\, \boldsymbol{y}) = \sum_{i=1}^I Q^i(\boldsymbol{\theta}, \boldsymbol{\theta}^0 \,;\, \boldsymbol{y}_i), \tag{15}$$

where functions Q^i are defined by

$$\begin{aligned}
Q^i(\boldsymbol{\theta}, \boldsymbol{\theta}^0 \,;\, \boldsymbol{y}_i) &\triangleq \sum_{\boldsymbol{x}_i} \Pr(\boldsymbol{x}_i \,|\, \boldsymbol{y}_i \,;\, \boldsymbol{\theta}^0) \log(f(\boldsymbol{y}_i \,|\, \boldsymbol{x}_i \,;\, \boldsymbol{\theta}) \Pr(\boldsymbol{x}_i \,;\, \boldsymbol{\theta})), && (16) \\
&= E[\log(f(\boldsymbol{y}_i \,|\, \boldsymbol{X}_i \,;\, \boldsymbol{\theta}) \Pr(\boldsymbol{X}_i \,;\, \boldsymbol{\theta})) \,|\, \boldsymbol{y}_i \,;\, \boldsymbol{\theta}^0]. && (17)
\end{aligned}$$

In addition, if parameter vector $\boldsymbol{\theta}$ can be partitioned into two subvectors $\boldsymbol{\theta}_{Y|X}$ and $\boldsymbol{\theta}_X$, it is straightforward to check in the same manner as in Paragraph III-B that each function Q_i can be decomposed as

$$Q^i(\boldsymbol{\theta}, \boldsymbol{\theta}^0 \,;\, \boldsymbol{y}_i) = Q_{Y|X}^i(\boldsymbol{\theta}_{Y|X}, \boldsymbol{\theta}^0 \,;\, \boldsymbol{y}_i) + Q_X^i(\boldsymbol{\theta}_X, \boldsymbol{\theta}^0 \,;\, \boldsymbol{y}_i), \tag{18}$$

where the expressions of $Q_{Y|X}^i$ and Q_X^i can be deduced from (12) and (13) by substituting $\boldsymbol{y}_i$ and $\boldsymbol{x}_i$ for $\boldsymbol{y}$ and $\boldsymbol{x}$, respectively.

- Forward step :

$$\begin{aligned} & & \mathcal{N}_1 &= \textstyle\sum_{n=1}^N p_n^0\, \mathcal{G}_n, \\ \text{for } n=1,\dots,N: & & \mathcal{F}_{1,n} &= p_n^0\, \mathcal{G}_n / \mathcal{N}_1, \\ \text{for } t=2,\dots,T: & & & \\ & & \mathcal{N}_t &= \textstyle\sum_{n=1}^N \big(\sum_{m=1}^N \mathcal{F}_{t-1,m}\, P_{mn}^0\big)\, \mathcal{G}_n, \\ \text{for } n=1,\dots,N: & & \mathcal{F}_{t,n} &= \big(\textstyle\sum_{m=1}^N \mathcal{F}_{t-1,m}\, P_{mn}^0\big)\, \mathcal{G}_n / \mathcal{N}_t. \end{aligned}$$

- Backward step :

$$\begin{aligned} \text{for } n=1,\dots,N: & & \mathcal{B}_{T,n} &= 1, \\ \text{for } t=T-1,\dots,1: & & & \\ \text{for } n=1,\dots,N: & & \mathcal{B}_{t,n} &= \textstyle\sum_{m=1}^N \mathcal{B}_{t+1,m}\, P_{nm}^0\, \mathcal{G}_m / \mathcal{N}_{t+1}. \end{aligned}$$

- For $t = T-1,\dots,1$:

$$\begin{aligned} \text{for } n=1,\dots,N: & & p_{t,n}^0 &= \mathcal{F}_{t,n}\mathcal{B}_{t,n}, \\ \text{for } m,n=1,\dots,N: & & p_{t,mn}^0 &= \mathcal{F}_{t-1,m} P_{nm}^0 \mathcal{B}_{t,n} \mathcal{G}_n / \mathcal{N}_t. \end{aligned}$$

- Reestimation step :

$$\begin{aligned} \text{for } n=1,\dots,N: & & p_n &= p_{1,n}^0, \\ \text{for } m,n=1,\dots,N: & & P_{mn} &= \textstyle\sum_{t=2}^T p_{t,mn}^0 / \sum_{t=1}^{T-1} p_{t,m}^0, \\ \text{for } n=1,\dots,N: & & u_n &= \textstyle\sum_{t=1}^T p_{t,n}^0\, y_t / \alpha_n^0, \quad v_n = \sum_{t=1}^T p_{t,n}^0\, (y_t - u_n)^2 / \alpha_n^0. \end{aligned}$$

TABLE II

STANDARD REESTIMATION EM FORMULAS THAT YIELD $\boldsymbol{\theta} = \big(\{p_n\}, \{P_{mn}\}, \{u_n\}, \{v_n\}\big)$ AS THE MAXIMIZER OF $Q(., \boldsymbol{\theta}^0, \boldsymbol{y})$ FOR A FINITE STATE HOMOGENEOUS HMM WITH GAUSSIAN OBSERVATIONS. THE FORWARD-BACKWARD ALGORITHM PROVIDED HERE TAKES THE NORMALIZED FORM GIVEN IN [2].

IV. 3D EXTENSION

A. Segmentation of 3D PRFs

This paragraphs relies on an extension of the construction of stationnary MRFs and PRFs presented in [4], [5] to the 3D case. The results are available in [7] and will not be derived here. We model $\boldsymbol{X}$ as a 3D Pickard random field and we consider MMAP estimation of a voxel $X_{\{i,j,k\}}$ of the 3D array under approximations similar to those outlined in Section II. More specifically, the marginal likelihood $\Pr(x_{\{i,j,k\}} \,|\, \boldsymbol{y})$ is approximated as

$$\Pr(x_{\{i,j,k\}} \,|\, \boldsymbol{y}) \approx \Pr(x_{\{i,j,k\}} \,|\, \boldsymbol{y}_{\{i,\cdot,k\}}, \boldsymbol{y}_{\{\cdot,j,k\}}, \boldsymbol{y}_{\{i,j,\cdot\}}), \tag{19}$$

where $\boldsymbol{y}_{\{i,\cdot,k\}}$, $\boldsymbol{y}_{\{\cdot,j,k\}}$ and $\boldsymbol{y}_{\{i,j,\cdot\}}$ denote the three 1D restrictions of $\boldsymbol{y}$ which contain voxel $y_{\{i,j,k\}}$. Here again, this approximation amounts to neglect interactions in the diagonal directions. It can be shown

that (see [7]):

$$\Pr(x_{\{i,j,k\}} \,|\, \boldsymbol{y}_{\{i,\cdot,k\}}, \boldsymbol{y}_{\{\cdot,j,k\}}, \boldsymbol{y}_{\{i,j,\cdot\}}) \propto \Pr(x_{\{i,j,k\}}) f(\boldsymbol{y}_{\{i,\cdot,k\}} \,|\, x_{\{i,j,k\}}) \\ f(\boldsymbol{y}_{\{\cdot,j,k\}} \,|\, x_{\{i,j,k\}}) f(\boldsymbol{y}_{\{i,j,\cdot\}} \,|\, x_{\{i,j,k\}}) \,. \quad (20)$$

As in the 2D case, the terms in the right hand side of (20) only involve 1D quantities. More specifically, due to the hidden Markov chain structures of the 1D restrictions of $\boldsymbol{y}$, the conditional probabilities in the right hand side of (20) can be evaluated using the same 1D forward-backward algorithms as in the 2D case, and the only parameters of interest of the PRF prior model are those which control the behavior of 1D Markov chains $\boldsymbol{X}_{\{i,\cdot,k\}}$, $\boldsymbol{X}_{\{\cdot,j,k\}}$ and $\boldsymbol{X}_{\{i,j,\cdot\}}$.

B. Parameter estimation

Here again, the ML estimator of $\boldsymbol{\theta}$ cannot be expressed in closed form and an EM procedure is applied to the pseudo-likelihood obtained by taking the product of marginal likelihoods of all 1D restrictions of $\boldsymbol{Y}$. Therefore, we have :

$$f(\boldsymbol{y} \,;\, \boldsymbol{\theta}) \underset{\sim}{\propto} \prod_{r \in \mathcal{R}_1} f(\boldsymbol{y}_{(r)} \,;\, \boldsymbol{\theta}) \,, \quad (21)$$

where $\{\boldsymbol{y}_{(r)}; r \in \mathcal{R}_1\}$ is a shorthand notation for $\{\boldsymbol{y}_{\{i,\cdot,k\}}; i,k\} \cup \{\boldsymbol{y}_{\{\cdot,j,k\}}; j,k\} \cup \{\boldsymbol{y}_{\{i,j,\cdot\}}; i,j\}$, the set of all 1D restrictions of $\boldsymbol{y}$. Choosing $\boldsymbol{x}$ as the auxiliary variable of the EM algorithm and applying the result of Paragraph III-C yields

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^0 \,;\, \boldsymbol{y}) = \sum_{r \in \mathcal{R}_1} Q^{(r)}(\boldsymbol{\theta}, \boldsymbol{\theta}^0 \,;\, \boldsymbol{y}_{(r)}) \,, \quad (22)$$

$$= \sum_{r \in \mathcal{R}_1} E[\ln(f(\boldsymbol{y}_{(r)} \,|\, \boldsymbol{X}_{(r)} \,;\, \boldsymbol{\theta}) \Pr(\boldsymbol{X}_{(r)} \,;\, \boldsymbol{\theta})) \,|\, \boldsymbol{y}_r \,;\, \boldsymbol{\theta}^0] \,. \quad (23)$$

The process $\boldsymbol{y}_{(r)}$ has the structure of a 1D hidden Markov model with hidden process $\boldsymbol{X}_{(r)}$, and (23) shows that functions $Q^{(r)}$ are identical to those obtained for EM estimation of the parameters of 1D hidden Markov models. In other words, the reestimation formulas essentially operate on 1D quantities, which is the key to a tractable numerical implementation. We now precisely define these quantities and derive the corresponding EM algorithm, keeping in mind that parameter vector $\boldsymbol{\theta}$ can be partitioned into $\{\boldsymbol{\theta}_{Y|X}, \boldsymbol{\theta}_X\}$ which allows decoupling of the maximization step.

V. Telegraph model

In this section, we introduce the telegraph model whose purpose is to reduce the computational cost of parameter estimation and to ensure that the necessary condition of stationarity of the 1D restrictions of PRF $\boldsymbol{X}$ are fulfilled. As indicated by (20) and (21), the prior model needs only to specify the distribution of 1D quantities. Therefore the process we consider, i.e., the telegraph model (TM), is strictly a 1D Markov chain model, the 3D nature of the problem being accounted for through the aforementioned equations. We now define the TM and its parameter vector $\boldsymbol{\theta}_X$ and then derive the corresponding EM reestimation formulas.

A. Telegraph model definition

The TM is a straightforward generalization of a class of discrete-valued Markov chains proposed in [13] for segmentation of seismic signals. The transition probability matrix $\boldsymbol{P} = (P_{mn})$ of the model is defined by

$$\boldsymbol{P} = \boldsymbol{\Lambda} + (\boldsymbol{1} - \boldsymbol{\lambda})\boldsymbol{\mu}^{\mathrm{t}} \,, \quad (24)$$

with $\boldsymbol{\lambda} \triangleq \mathrm{vect}(\lambda_n)$, $\boldsymbol{\Lambda} \triangleq \mathrm{diag}(\lambda_n)$, $\mathbf{1} = (1, \ldots, 1)^{\mathrm{t}}$.

From an intuitive ground, the telegraphic parameterization $\boldsymbol{\theta}_X = \{\boldsymbol{\mu}, \boldsymbol{\lambda}\}$ can be interpreted as follows. The transition from one state to another is the result of a two-stage sampling experiment. On the basis of the first toss, the decision of keeping the current state m is made with probability λ_m. Otherwise, a new state n is chosen with probability μ_n, independently from the previous state. Since, in the latter case, n may be equal to m with probability μ_m, the probability of keeping the current state m is actually $\lambda_m + \mu_m - \lambda_m\mu_m$. According to such values, typical trajectories of the TM are more or less "blocky". This is a one-dimensional counterpart to well-known spatial Gibbsian models available for unordered colors [1].

In order to ensure that the resulting Markov chain is well defined and irreducible, it is straightforward to check that the following constraints form a set of sufficient conditions:

$$\sum_{n=1}^{N} \mu_n = 1, \tag{25}$$

$$\forall n = 1, \ldots, N, \mu_n > 0, \tag{26}$$

$$\forall n = 1, \ldots, N, \lambda_n < 1, \tag{27}$$

$$\forall n = 1, \ldots, N, \lambda_n > -\mu_n/(1-\mu_n). \tag{28}$$

Note that λ_n is not necessarily positive, although $\lambda_n > 0, n = 1, \ldots, N$ was understood in the above interpretation of the TM.

The stationary probability vector of the TM is readily obtained as

$$\boldsymbol{p} = \left(\boldsymbol{I} - \boldsymbol{\Lambda} + \boldsymbol{\mu}\boldsymbol{\lambda}^{\mathrm{t}}\right)^{-1} \boldsymbol{\mu}\,, \tag{29}$$

where $\boldsymbol{I}$ is the identity matrix. Componentwise, such a vector also reads

$$p_n = \frac{\mu_n}{1-\lambda_n} \Big/ \sum_{m=1}^{N} \frac{\mu_m}{1-\lambda_m}\,. \tag{30}$$

Moreover, it can be verified that matrix $\mathrm{diag}(\boldsymbol{p})\boldsymbol{P}$ is symmetric, so the TM is reversible in its stationary state. Therefore, as long as the initial state probability vector is equal to $\boldsymbol{p}$ and that constraints (25)-(28) are fulfilled, (24) defines a stationary and reversible Markov chain that we choose to parameterize with $\boldsymbol{\theta}_X = \{\boldsymbol{\lambda}, \boldsymbol{\mu}\}$. The resulting number of degrees of freedom is $2N-1$, which is linear w.r.t. the number of states, as opposed to the standard HMM case, which yields $N^2 - 1$ free parameters.

B. Reestimation formulas for $\boldsymbol{\theta}_X$

One of the reasons for introducing the TM is to simplify the forward-backward algorithm used to evaluate marginal likelihood values $\Pr(X_t = n \,|\, \boldsymbol{y}\,;\, \boldsymbol{\theta})$. As seen in Table II (evaluation of quantities $\mathcal{F}_t$, $\mathcal{B}_t$ and p_t^0), each of the $T-1$ recursions of the algorithm requires matrix products involving transition matrix $\boldsymbol{P}^0$. As seen in the sequel, expressing $\boldsymbol{P}$ according to (24) allows us to bring the computational complexity of each recursion down from $O(N^2)$ to $O(N)$.

B.1 E-step

From the definition of Q_X (13), we have

$$\begin{aligned} Q_X(\boldsymbol{\theta}_X, \boldsymbol{\theta}^0\,;\boldsymbol{y}) &= \sum_{\boldsymbol{x}} \Pr(\boldsymbol{x} \,|\, \boldsymbol{y}\,;\, \boldsymbol{\theta}^0) \log \Pr(\boldsymbol{x}\,;\, \boldsymbol{\theta}_X) \\ &= \sum_{n=1}^{N} p_{1,n}^0 \log p_n + \sum_{m,n=1}^{N} \sum_{t=2}^{T} p_{t,mn}^0 \log p_{mn}\,, \end{aligned}$$

where

$$\begin{aligned} p^0_{t,n} &\triangleq \Pr(X_t = n \,|\, \boldsymbol{y}\,;\, \boldsymbol{\theta}^0), \\ p^0_{t,mn} &\triangleq \Pr(X_{t-1} = m, X_t = n \,|\, \boldsymbol{y}\,;\, \boldsymbol{\theta}^0)\,. \end{aligned}$$

Then, expressions (24) and (30) allow us to express the explicit dependence of Q_X on $\boldsymbol{\lambda}, \boldsymbol{\mu}$:

$$Q_X(\boldsymbol{\theta}_X, \boldsymbol{\theta}^0\,;\,\boldsymbol{y}) = \sum_{n=1}^N \alpha^0_n \log\mu_n + \beta^0_n \log(1-\lambda_n) + s^0_n \log\left(1 + \frac{\lambda_n}{\mu_n(1-\lambda_n)}\right) - \log\sum_{n=1}^N \frac{\mu_n}{1-\lambda_n}\,, \quad (31)$$

with

$$\alpha^0_n \triangleq \sum_{t=1}^T p^0_{t,n}\,,\ \beta^0_n \triangleq \sum_{t=2}^{T-1} p^0_{t,n}\,,\ s^0_n \triangleq \sum_{t=2}^{T} p^0_{t,nn}\,. \quad (32)$$

B.2 Approximate M-step

The major difficulty lies in the M step, which consists of maximizing Q_X under constraints (25)-(28). Because of the last term in (31), explicit maximization is intricate. On the other hand, relative simplification occurs if Q_X is approximated by

$$\begin{aligned} \widetilde{Q}_X(\boldsymbol{\theta}_X, \boldsymbol{\theta}^0\,;\,\boldsymbol{y}) &\triangleq Q_X(\boldsymbol{\theta}_X, \boldsymbol{\theta}^0\,;\,\boldsymbol{y}) - \mathrm{E}\left[\log P(X_1\,;\,\boldsymbol{\theta}_X) P(X_T\,;\,\boldsymbol{\theta}_X) \,|\, \boldsymbol{y}\,;\,\boldsymbol{\theta}^0\right]/2 \\ &= \mathrm{E}\left[\log P(X_2, \ldots, X_T \,|\, X_1\,;\,\boldsymbol{\theta}_X) P(X_1, \ldots, X_{T-1} \,|\, X_T\,;\,\boldsymbol{\theta}_X) \,|\, \boldsymbol{y}\,;\,\boldsymbol{\theta}^0\right]/2 \\ &= \mathrm{E}\left[\log P(X_2, \ldots, X_T \,|\, X_1\,;\,\boldsymbol{\theta}_X) \,|\, y_1, \ldots, y_T\,;\,\boldsymbol{\theta}^0\right]/2 \\ &\quad + \mathrm{E}\left[\log P(X_2, \ldots, X_T \,|\, X_1\,;\,\boldsymbol{\theta}_X) \,|\, y_T, \ldots, y_1\,;\,\boldsymbol{\theta}^0\right]/2\,. \end{aligned}$$

Apart from the fact that the difference between Q_X and $\widetilde{Q}_X$ is moderate, it is not difficult to check that $\widetilde{Q}_X$ itself is an exact auxiliary function associated to a modified likelihood function. The latter reads

$$f_{\boldsymbol{\pi},\boldsymbol{\lambda},\boldsymbol{\mu}}(y_1, \ldots, y_T) f_{\boldsymbol{\pi},\boldsymbol{\lambda},\boldsymbol{\mu}}(y_T, \ldots, y_1)\,, \quad (33)$$

where $f_{\boldsymbol{\pi},\boldsymbol{\lambda},\boldsymbol{\mu}}$ is the probability density function of the data when the initial probability vector of the TM is an arbitrary vector $\boldsymbol{\pi}$, while the transition matrix is parameterized by $(\boldsymbol{\lambda}, \boldsymbol{\mu})$ according to (24). The latter property ensures that the fixed-point EM procedure based on $\widetilde{Q}_X$ does converge (towards a stationary point of (33)).

First, let us express $\widetilde{Q}_X$ as an explicit function of $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$:

$$\widetilde{Q}_X(\boldsymbol{\theta}_X, \boldsymbol{\theta}^0\,;\boldsymbol{y}) = \sum_{n=1}^N \widetilde{Q}_n\,,$$

with

$$\widetilde{Q}_n \triangleq \eta^0_n \log\mu_n(1-\lambda_n) + s^0_n \log\left(1 + \frac{\lambda_n}{\mu_n(1-\lambda_n)}\right)\,, \quad (34)$$

and

$$\eta^0_n \triangleq (\alpha^0_n + \beta^0_n)/2\,. \quad (35)$$

It is easy to maximize $\widetilde{Q}_X$ w.r.t. $\boldsymbol{\lambda}$ when $\boldsymbol{\mu}$ is held constant, since each function $\widetilde{Q}_n$ depends on λ_n only, and its maximum is reached at a unique point

$$\widehat{\lambda}_n = \frac{s_n^0/\eta_n^0 - \mu_n}{1-\mu_n}\,. \tag{36}$$

Moreover, constraints (27), (28) are fulfilled by $\widehat{\boldsymbol{\lambda}} = (\widehat{\lambda}_n)$ since $s_n^0 < \alpha_n^0$ and $s_n^0 < \beta_n^0$ according to (32), provided that $\boldsymbol{\theta}^0$ meets (25)-(28). Substituting (36) into (34) allows us to express $\widetilde{Q}_n$ as a function of μ_n to within an additive constant factor:

$$\widetilde{Q}_n = \gamma_n^0 \log \frac{\mu_n}{1-\mu_n}\,, \tag{37}$$

with $\gamma_n^0 \triangleq \eta_n^0 - s_n^0 \geq 0$. The Lagrange multiplier technique is used for maximization of $\widetilde{Q}_X$ with respect to $\boldsymbol{\mu}$ under constraints (25) and (26). Equating the gradient of the corresponding criterion to zero yields:

$$\forall n, \quad \nu\widehat{\mu}_n^2 - \nu\widehat{\mu}_n + \gamma_n^0 = 0\,, \tag{38}$$

where ν denotes the Lagrange multiplier. When $\nu > 4\gamma_n^0$, the above equation has two distinct roots, $\mu_n^+(\nu)$ and $\mu_n^-(\nu)$, located in $(0,1)$ on either side of $1/2$:

$$\mu_n^{\pm}(\nu) = \frac{1}{2}\left(1 \pm \sqrt{1 - 4\gamma_n^0/\nu}\right)\,.$$

At first glance, the set of all possible combinations of μ_n^- and μ_n^+ provides 2^N different forms for $\widehat{\boldsymbol{\mu}} = (\widehat{\mu}_n)$. However, according to (25) and (26), $\widehat{\boldsymbol{\mu}}$ may only contain one μ_n^+. This brings the number of possible combinations down to $N+1$. Furthermore, among the N combinations that include one μ_n^+, $\widetilde{Q}_X$ is maximized if and only if the corresponding state n is chosen among the maximizers of (γ_n^0): $\forall m, \gamma_m^0 \leq \gamma_n^0$. Such a result stems from the following property: let us assume that constraint (25) is fulfilled by

$$\boldsymbol{\mu}(\nu) = \left(\mu_1^-(\nu), \ldots, \mu_{n-1}^-(\nu), \mu_n^+(\nu), \mu_{n+1}^-(\nu), \ldots, \mu_N^-(\nu)\right)$$

for some value of ν, and, for instance, that $\gamma_1^0 > \gamma_n^0$. Then, for the same value of ν, constraint (25) is still fulfilled after the permutation of $\mu_1^-(\nu)$ and $\mu_n^+(\nu)$ in $\boldsymbol{\mu}(\nu)$, while it is easy to check from (37) that Q_X is increased by the positive amount

$$(\gamma_1^0 - \gamma_n^0) \log \frac{\mu_n^+(\nu)\left(1-\mu_1^-(\nu)\right)}{\mu_1^-(\nu)\left(1-\mu_n^+(\nu)\right)}\,.$$

Only two possible forms of combination remain:

$$\begin{aligned}
&\boldsymbol{\mu}^-, \text{ defined by:} && \forall m, \mu_m = \mu_m^-\,,\\
&\boldsymbol{\mu}_n^+, \text{ defined by:} && \begin{cases} \forall m \neq n, \mu_m = \mu_m^-\,,\\ \mu_n = \mu_n^+\,,\\ \forall m, \gamma_m^0 \leq \gamma_n^0\,. \end{cases}
\end{aligned}$$

Note that there is as much different combinations $\boldsymbol{\mu}_n^+$ as maximizers of (γ_n^0). Further analysis of the properties of the remaining combinations brings the following existence and uniqueness result: the maximum of $\widetilde{Q}_X = \sum_{n=1}^N \widetilde{Q}_n$ (where $\widetilde{Q}_n$ is given by (37)), under constraints (25) and (26), is reached by a unique vector $\widehat{\boldsymbol{\mu}}(\widehat{\nu})$, where $\widehat{\nu}$ is uniquely determined by $\sum_{n=1}^N \widehat{\mu}_n(\widehat{\nu}) = 1$, and

$$\widehat{\boldsymbol{\mu}} = \begin{cases} \boldsymbol{\mu}^- & \text{if } \sum_{n=1}^N \omega_n^0 \leq N-2\,,\\ \boldsymbol{\mu}^+_{\arg\max_n \gamma_n^0} & \text{otherwise}\,, \end{cases} \tag{39}$$

with

$$\omega_n^0 = \sqrt{1 - \gamma_n^0 / \max_n \gamma_n^0}\,.$$

Since $0 \le \omega_n^0 < 1$ for all n, and $\omega_m^0 = 0$ if $\gamma_m^0 = \max_n \gamma_n^0$, it is not difficult to check that $\sum_{n=1}^N \omega_n^0 \le N-2$ if γ_n^0 admits more than one maximizer. Hence, the (unique) maximizer $\arg\max_n$ is well defined in (39).

In practice, $\widehat{\nu}$ cannot be expressed in closed form, but tight lower and upper bounds can be easily derived and classical numerical interpolation techniques can then be employed to refine the approximation. A summary of the forward-backward algorithm and of the reestimation formulas for $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$ is given in Table III.

for $n = 1, \dots, N$: $\quad p_n^0 = \dfrac{\mu_n^0/(1-\lambda_n^0)}{\sum_{m=1}^N \mu_m^0/(1-\lambda_m^0)}, \quad P_{nn}^0 = \lambda_n^0 + \mu_n^0 - \lambda_n^0 \mu_n^0.$

- Forward step:

$$\mathcal{N}_1 = \textstyle\sum_{n=1}^N p_n^0\, \mathcal{G}_n,$$

for $n = 1, \dots, N$: $\quad \mathcal{F}_{1,n} = p_n^0\, \mathcal{G}_n / \mathcal{N}_1,$

for $t = 2, \dots, T$:

$$\mathcal{N}_t = \textstyle\sum_{n=1}^N \big(\lambda_n^0\, \mathcal{F}_{t-1,n} + \big(1 - \sum_{m=1}^N \lambda_m^0\, \mathcal{F}_{t-1,m}\big)\mu_n^0\big)\mathcal{G}_n,$$

for $n = 1, \dots, N$: $\quad \mathcal{F}_{t,n} = \big(\lambda_n^0\, \mathcal{F}_{t-1,n} + \big(1 - \sum_{m=1}^N \lambda_m^0\, \mathcal{F}_{t-1,m}\big)\mu_n^0\big)\mathcal{G}_n / \mathcal{N}_t,$

- Backward step:

for $n = 1, \dots, N$: $\quad \mathcal{B}_{T,n} = 1,$

for $t = T-1, \dots, 1$:

for $n = 1, \dots, N$: $\quad \mathcal{B}_{t,n} = \Big(\lambda_n^0\, \mathcal{B}_{t+1,n}\, \mathcal{G}_n + \big(\sum_{m=1}^N (1-\lambda_m^0)\, \mathcal{B}_{t+1,m}\, \mathcal{G}_m\big)\mu_n\Big) / \mathcal{N}_{t+1}.$

- For $t = T-1, \dots, 1$:

for $n = 1, \dots, N$: $\quad p_{t,n}^0 = \mathcal{F}_{t,n} \mathcal{B}_{t,n},$

$$p_{t,nn}^0 = \mathcal{F}_{t-1,n} P_{nn}^0 \mathcal{B}_{t,n} \mathcal{G}_n / \mathcal{N}_t.$$

- Reestimation step:

approximate $\widehat{\nu}$ s.t. $\sum_{n=1}^N \widehat{\mu}_n(\widehat{\nu}) = 1$ by interpolation, where, for $n = 1, \dots, N$:

$$\widehat{\mu}_n(\nu) = \begin{cases} \big(1 + \sqrt{1 - 4\gamma_n^0/\nu}\big)/2 & \text{if } \gamma_n^0 = \max\limits_m \gamma_m^0 \text{ and } \sum_{m=1}^N \sqrt{1 - \gamma_m^0/\gamma_n^0} > N-2, \\ \big(1 - \sqrt{1 - 4\gamma_n^0/\nu}\big)/2 & \text{otherwise}; \end{cases}$$

for $n = 1, \dots, N$: $\quad \mu_n = \widehat{\mu}_n(\nu), \qquad \lambda_n = (s_n^0/\eta_n^0 - \mu_n)/(1-\mu_n),$

for $n = 1, \dots, N$: $\quad u_n = \sum_t p_{t,n}^0\, y_t / \alpha_n^0, \;\; v_n = \big(2a + \sum_t p_{t,n}^0\, (y_t - u_n)^2\big)/(2b + \alpha_n^0).$

TABLE III

PENALIZED EM FORMULAS FOR A TELEGRAPHIC HMM WITH GAUSSIAN OBSERVATIONS.

VI. Mixture of Gaussians

We now address the question of the degeneracy of the likelihood with respect to parameters $\boldsymbol{\theta}_{Y|X}$. Maximizing $f(\boldsymbol{y};\boldsymbol{\theta})$ with respect to

$$\boldsymbol{\theta}_{Y|X} = (\boldsymbol{u},\boldsymbol{v}) = (u_1,\dots,u_N,v_1,\dots,v_N) \in \Theta = \mathbb{R}^N \times {\mathbb{R}_+^*}^N$$

is indeed a degenerate problem since $f(\boldsymbol{y};\boldsymbol{\theta})$ is not bounded above: for an arbitrary state n and an arbitrary data sample y_t, it is clear that $f(\boldsymbol{y};\boldsymbol{\theta})$ can take arbitrary large values as v_n comes close to 0, when $u_n = y_t$ and every other unknowns are held fixed to arbitrary constants. This is a well known problem for the maximum likelihood approach to the identification of some mixture models [10], [3]. In order to cope with the degeneracy in the case of an independent identically distributed (i.i.d.) mixture model, Hathaway proposed to restrict the admissible domain, and he showed that an EM strategy could still be implemented to solve the resulting constrained maximization problem [14], [15].

Here, we adopt a slightly different approach, based on the maximization on Θ of a *penalized* version of the likelihood function:

$$F(\boldsymbol{y};\boldsymbol{\theta}) = f(\boldsymbol{y};\boldsymbol{\theta})\, G(\boldsymbol{v})$$

where $G(\boldsymbol{v})$ is an *ad hoc* prior distribution for $\boldsymbol{v}$ that compensates for the degeneracy at $v_n \searrow 0, n = 1,\dots,N$. For this purpose, the solution of choice is the i.i.d. inverted gamma model:

$$G(\boldsymbol{v}) = \prod_{n=1}^{N} g(v_n), \tag{40}$$

with

$$g(v_n) = \frac{a^{b-1}}{\Gamma(b-1)} \frac{1}{v_n^b} \exp\Big\{-\frac{a}{v_n}\Big\} 1_{[0,+\infty)}, \tag{41}$$

which is ensured to be proper if $b > 1$ and $a > 0$. The justification is twofold:

- For small values of v, $g(v)$ vanishes fast enough to compensate for the corresponding degeneracy of $f(\boldsymbol{y};\boldsymbol{\theta})$. More precisely, it can be established that F is a bounded function on Θ, which tends to zero when v vanishes. Thus, the global maximum of F is finite and it is reached for strictly positive components of $\boldsymbol{v}$, whereas the degeneracy points of f are not even local maxima for F. In the case of independent Gaussian mixtures, it has been recently shown that the global maximizer of F is a strongly consistent estimator [16].
- Substituting F for f allows us to maintain explicit reestimation equations in the classical EM scheme for Gaussian mixtures. The underlying reason is that the inverse gamma distribution $G(\boldsymbol{v})$ is *conjugate* for the complete-data distribution $f(\boldsymbol{y}\,|\,\boldsymbol{x};\boldsymbol{u},\boldsymbol{v})\Pr(\boldsymbol{x})$. Contrarily to Hathaway's constrained formulation, our penalized version is as simple to derive and to implement as the original EM scheme. The resulting reestimation formula for each v_n is

$$v_n = \frac{2a + \sum_{t=1}^{T} \Pr\left(X_t = n \,|\, \boldsymbol{y};\boldsymbol{\theta}^0\right)(y_t - u_n)^2}{2b + \sum_{t=1}^{T} \Pr\left(X_t = n \,|\, \boldsymbol{y};\boldsymbol{\theta}^0\right)},$$

while the other reestimation equations are unaltered.

The equations of the complete EM algorithm are given in Table III. Note that the MMAP segmentation stage directly follows from (20) and forward-backward evaluation of quantity $p^0_{t,n}$.

VII. Results

The unsupervised segmentation method described above was successfully tested on synthetic and real 2D and 3D images. In this section, we present a limited set of results in order to illustrate two points: the ability of the penalized approach to cope with the degeneracy of the likelihood and the performance of the method in real-size 3D data processing.

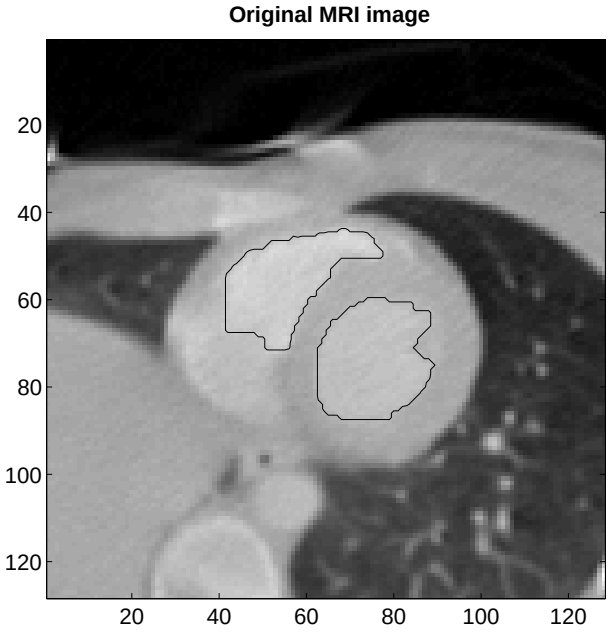

Fig. 1. Original magnetic resonance image of the heart region, size 128×128. The structures of interest are the ventricles whose approximate boundaries have been superimposed on the image.

A. *Likelihood degeneracy*

The unsupervised segmentation method was applied to the 128×128 2D magnetic resonance image of the heart region[1] presented in Fig. 1. The structures of interest are the ventricles whose approximate boundaries have been superimposed on the image. The model parameters were initialized in the following manner: the histogram of the original image was separated into N equal quantiles, and the values of $\boldsymbol{\theta}_{Y|X} = \{\boldsymbol{u}, \boldsymbol{v}\}$ were set to the empirical mean and variance of each quantile; all elements of $\boldsymbol{\mu}$ were set to $1/N$, and all elements of $\boldsymbol{\lambda}$ were set to the same value $\lambda_0 < 1/2$.

Without penalization of the likelihood, the method diverged after 12 iterations of the EM algorithm. The trajectories of the elements of $\boldsymbol{\theta}_X$ and $\boldsymbol{\theta}_{Y|X}$ are shown in Fig. 2. As can be observed, the divergence occurs when one of the components of the variance parameters $\boldsymbol{v}$ approaches zero. This result is consistent with the analysis of the degeneracy presented in Section VI.

The penalized method was applied to the magnetic resonance image with the same initial conditions. The parameters of the inverse gamma distribution were set to $a = 25, b = 1.01$. The trajectories of the estimated parameters and the resulting MMAP segmented image are shown in Figs. 3 and 4, respectively. It can be observed that convergence was reached after about 150 iterations and that even though several components of variance vector $\boldsymbol{v}$ were small, none of them approached zero closely enough for divergence to occur, thanks to the penalization term. More complete simulation results about likelihood degeneracy can be found in [17].

Regarding implementation issues, it should be underlined that with $N = 15$ labels, the TM induces a reduction of the computational cost of about one order of magnitude with respect to a standard parameterization of the Markov chains.

It should also be noted that for the particular example of Fig. 1, the best results were obtained with $N = 13$ labels. In these conditions, even the non penalized method happens to converge. The results obtained with the penalized algorithm are presented in Figs. 5 and 6. It can be observed that convergence takes place after less than 100 iterations and that in the MMAP segmented images, the two structures of interest can be clearly identified.

[1]Data courtesy of Dr. Alain Herment, INSERM U494, Hôpital de la Pitié-Salpêtrière, Paris, France

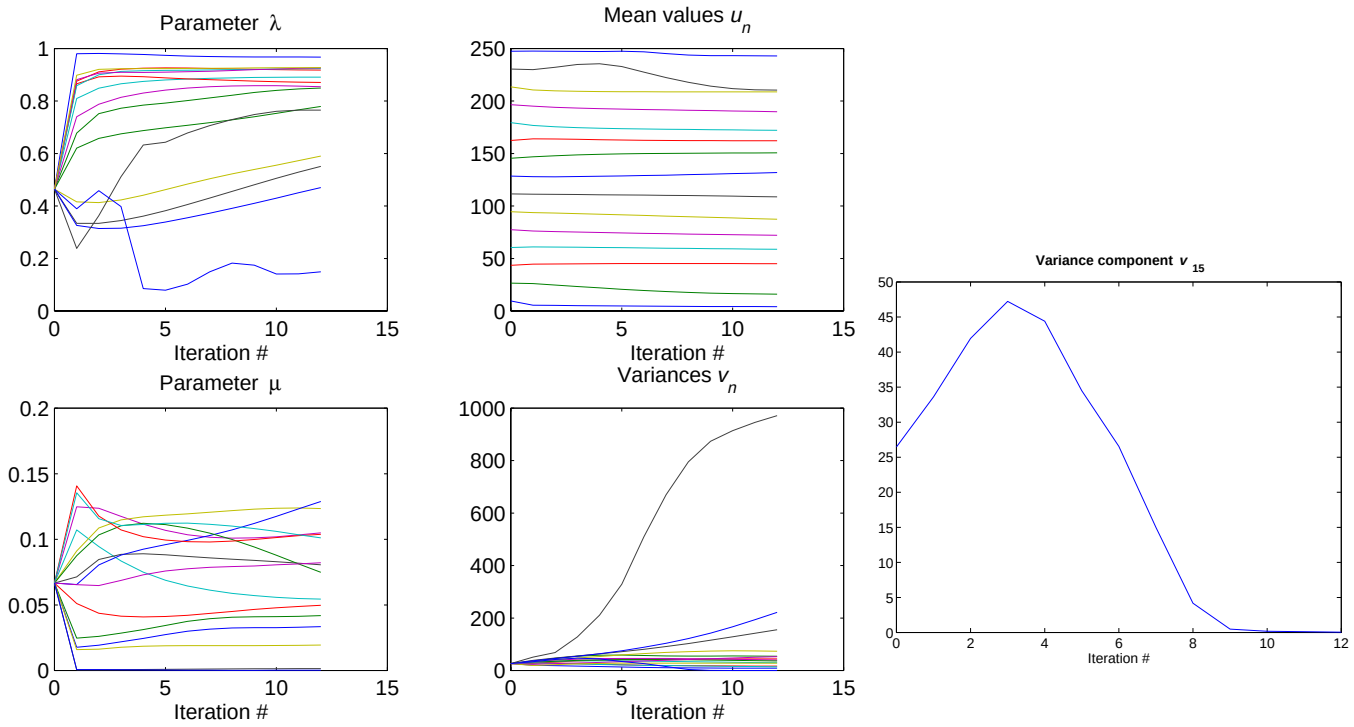

Fig. 2. Trajectories of the components of parameters $\boldsymbol{\theta}_X$ and $\boldsymbol{\theta}_{Y|X}$ without penalization of the likelihood, $N = 15$. Divergence occurs after 12 iterations of the EM algorithm, as component 15 of variance vector $\boldsymbol{v}$ approaches zero.

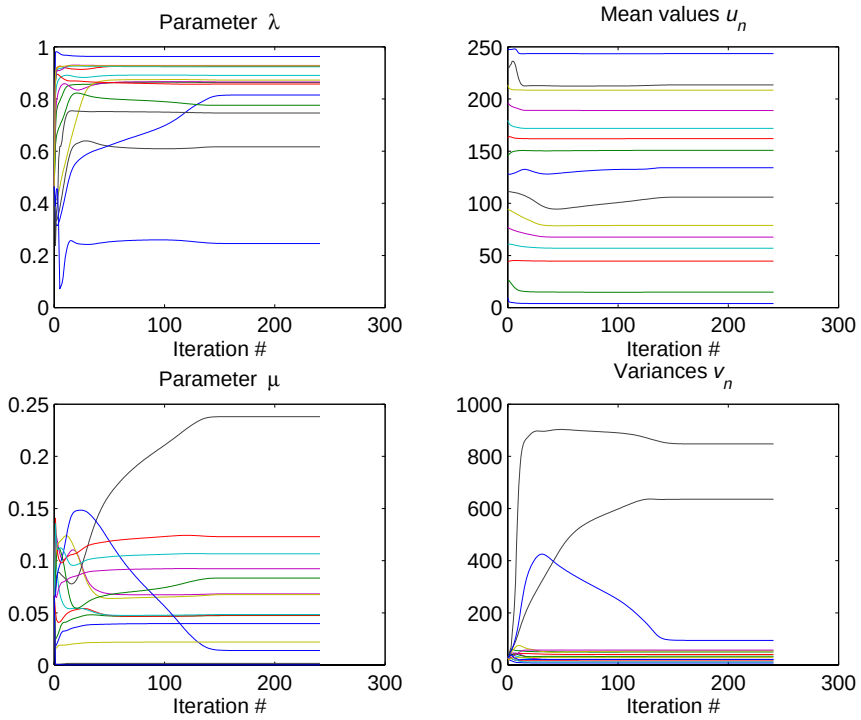

Fig. 3. Trajectories of the components of parameters $\boldsymbol{\theta}_X$ and $\boldsymbol{\theta}_{Y|X}$ with a penalized likelihood, $N = 15$. Convergence takes place after about 150 iteration of the EM procedure.

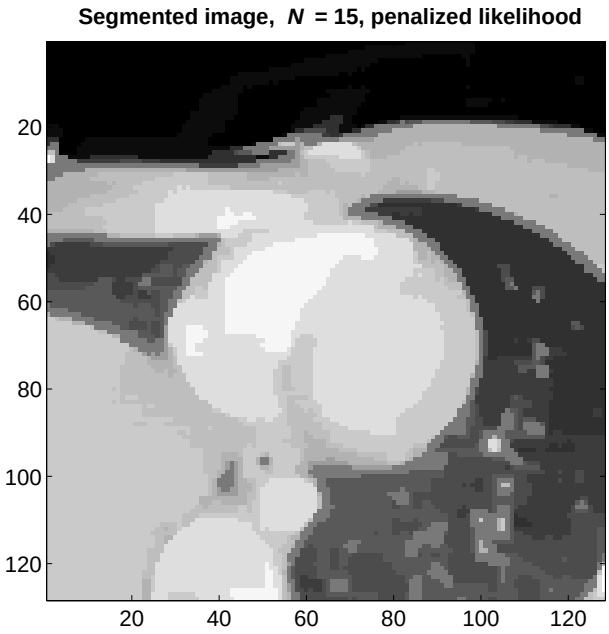

Fig. 4. MMAP unsupervised segmentation result, $N = 15$. The parameters were obtained with a penalized likelihood.

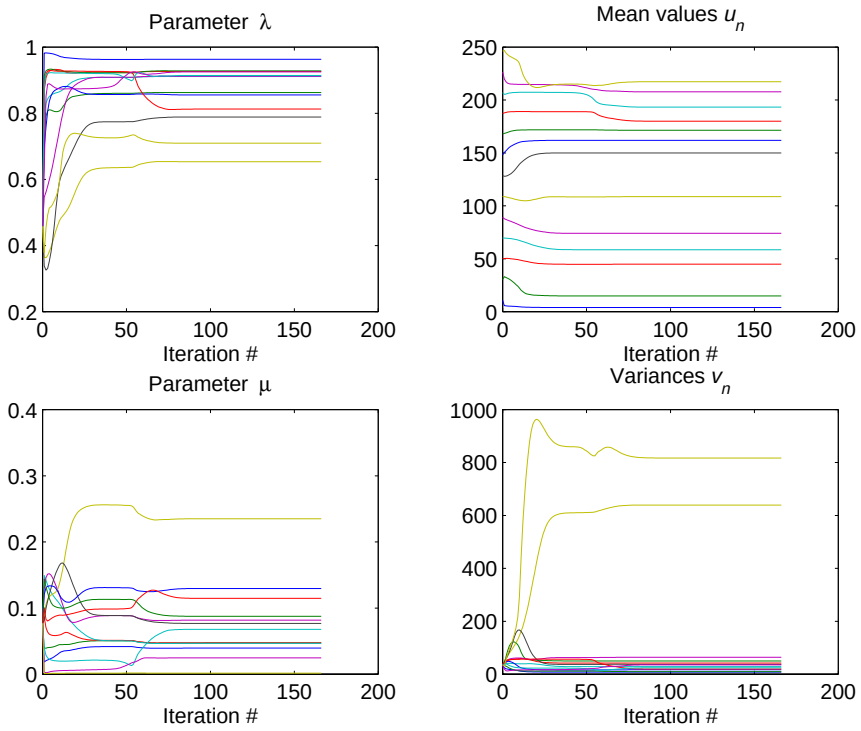

Fig. 5. Trajectories of the components of parameters $\boldsymbol{\theta}_X$ and $\boldsymbol{\theta}_{Y|X}$ with a penalized likelihood, $N = 13$. Convergence takes place after less than 100 iteration of the EM procedure.

B. Segmentation of 3D data

The 3D data to be segmented were obtained with a power Doppler ultrasound echograph. This imaging modality is used for analysis of blood flow. The data set[2] presented here was collected on a synthetic blood vessel which exhibits a strongly stenosed area. It consisted of 80 frames of size

[2]Data courtesy of Dr. Guy Cloutier, Institut de recherches cliniques de Montréal, Montreal, Quebec, Canada.

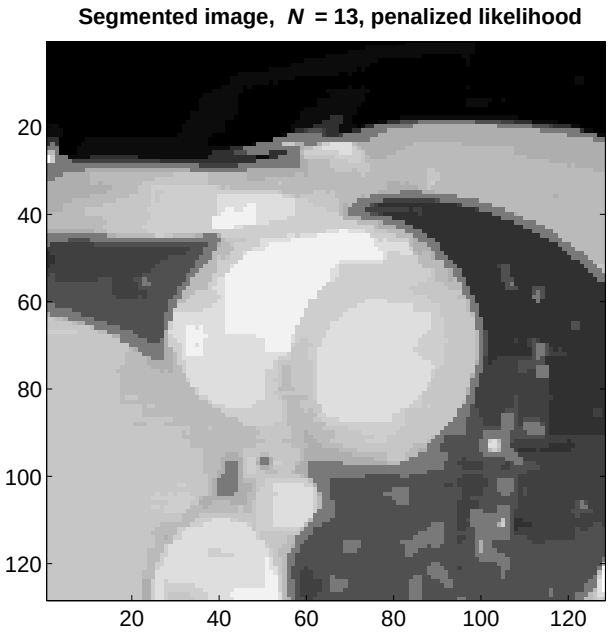

Fig. 6. MMAP unsupervised segmentation result, $N = 13$. The parameters were obtained with a penalized likelihood. The two ventricles can be clearly identified in the segmented image.

166×219, and segmentation was used to assess the dimension of the stenosis.

The number N of labels was set to four as such a number is sufficient to separate the various velocity regions present in the data, and the penalized version of the method was used.

Fig. 7 shows a longitudinal slice of the original power Doppler data and of the segmented data. This representation is adopted because of the approximately cylindrical symmetry of the medium. The trajectories of the estimated parameters is presented in Fig. 8. It can be observed that convergence occurs after a number of iterations that is much smaller than in the 2D case. This can be interpreted as the consequence of smaller number of labels and of the larger number of observations. The value of the stenosis diameter inferred from the segmented data was closer to the actual value than results provided by conventional methods of the field, thereby indicating a satisfactory behavior of our technique. It should be underlined that each iteration of the EM algorithm took approximately 100 seconds on a desktop Pentium II/300 computer, which shows that the proposed 3D method is usable even with a moderate computing power.

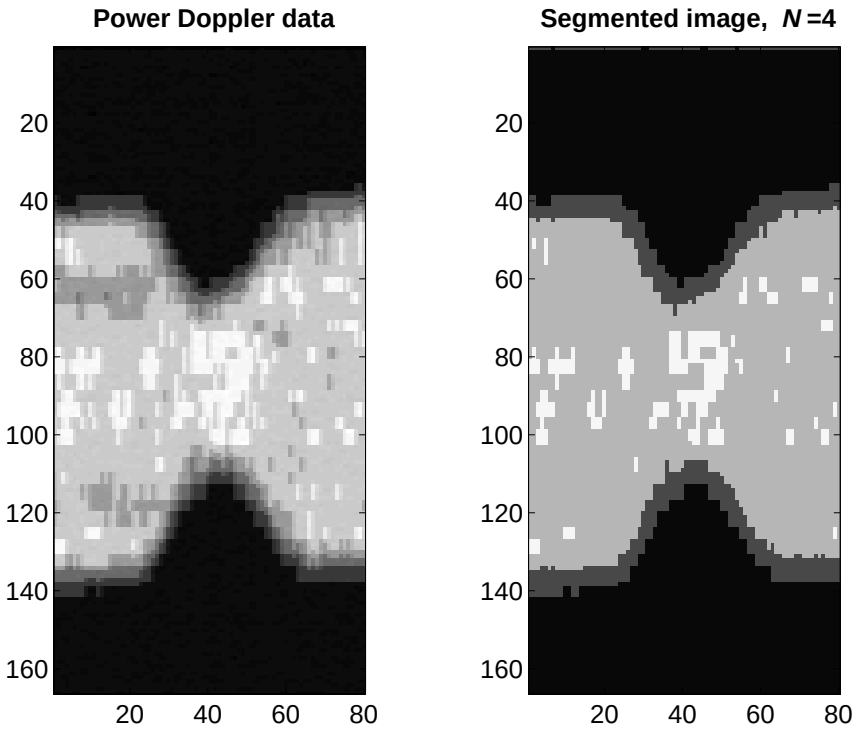

Fig. 7. Longitudinal slice of 3D data. Left: original power Doppler data; right: MMAP segmented data, $N = 4$.

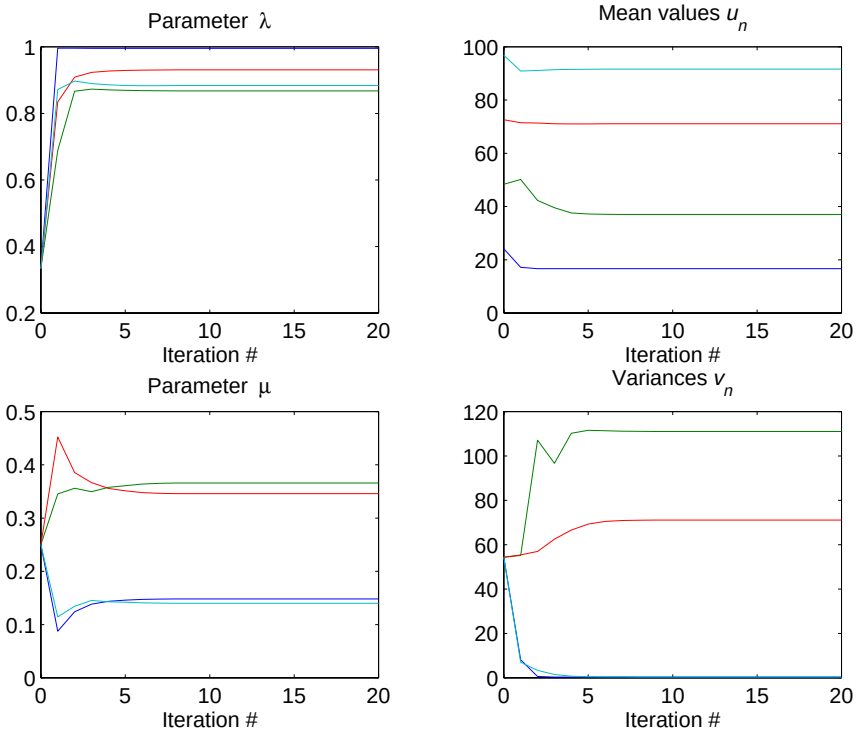

Fig. 8. Trajectories of the components of parameters $\boldsymbol{\theta}_X$ and $\boldsymbol{\theta}_{Y|X}$ with a penalized likelihood, $N = 4$. Convergence takes place after less than 10 iteration of the EM procedure.

VIII. Conclusion

In this paper, we have presented a fully unsupervised method for segmenting 2D and 3D images, provided that the number N of levels is known. Following Devijver and Dekesel [2], local dependencies were taken into account through a unilateral hidden Markov model with a conditionally Gaussian distribution of the observed pixels.

In [2], an EM strategy was introduced in order to carry out ML estimation of the model parameters. However, because of its heavy computational cost and of hazardous behavior [3], the authors finally preferred to perform joint estimation of model parameters and of the image, even though this technique presents controversial statistical properties [18].

Compared to [2], our contribution makes the EM strategy truly practicable for parameter estimation. On the one hand, we adopted a more parsimonious description of the hidden Markov model. It is a generalized version of the telegraph Markov chain model found in [13], whose number of parameters is of order $O(N)$ instead of $O(N^2)$. On the other hand, we introduced a penalized maximum likelihood approach that avoids the degeneracy of the usual likelihood function. Moreover, our penalized version is as simple to derive and to implement as the standard EM scheme.

Implementation of image processing methods based on Markov modeling usually requires heavy computations, even in supervised contexts. In this respect, the proposed segmentation method is a noticeable exception. This low numerical cost is obtained at the expense of a clear coarseness of the prior model, mostly due to its unilateral structure. It was also necessary to neglect diagonal interactions in the segmentation stage. As a consequence, the proposed method can be placed in the category of general purpose techniques best suited for automatic batch processing of big data sets. For specific types of images, more accurate segmentation results can probably be obtained using computationally more intensive Markov methods.

Acknowledgments

The authors wish to thank Dr. Alain Herment, INSERM U494, Hôpital de la Pitié-Salpêtrière, Paris, France and Dr. Guy Cloutier, Institut de recherches cliniques de Montréal, Montreal, Quebec, Canada for providing the ultrasound data used in Section VII-B. Partial support for this work was provided by the Natural Sciences and Engineering Research Council of Canada (Research Grant # OGP0138417) and by the ministère des Relations internationales du Québec (Cooperative Program 5.1.4, Project # 7) and the French ministère des Affaires Étrangères (Coopération scientifique franco-québécoise, Projet I.1.2.1.7).

References

[1] J. E. Besag, "On the statistical analysis of dirty pictures (with discussion)", *Journal of the Royal Statistical Society B*, vol. 48, no. 3, pp. 259–302, 1986.

[2] P. A. Devijver and M. Dekesel, "Champs aléatoires de Pickard et modélisation d'images digitales", *Traitement du Signal*, vol. 5, no. 5, pp. 131–150, 1988.

[3] A. Nádas, "Hidden Markov chains, the forward-backward algorithm, and initial statistics", *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. ASSP-31, no. 2, pp. 504–506, April 1983.

[4] D. K. Pickard, "A curious binary lattice process", *Journal of Applied Probability*, vol. 14, pp. 717–731, 14 1977.

[5] D. K. Pickard, "Unilateral Markov fields", *Advances in Applied Probability*, vol. 12, pp. 655–671, 12 1980.

[6] F. Champagnat, J. Idier, and Y. Goussard, "Stationary Markov random fields on a finite rectangular lattice", *IEEE Transactions on Information Theory*, vol. 44, pp. 2901–2916, 1998.

[7] J. Idier and Y. Goussard, "Champs de Pickard 3D", Tech. Rep., IGB / GPI-LSS, 1999.

[8] J. E. Besag, "Spatial interaction and the statistical analysis of lattice systems (with discussion)", *Journal of the Royal Statistical Society B*, vol. 36, no. 2, pp. 192–236, 1974.

[9] X. Guyon, *Champs aléatoires sur un réseau : modélisations, statistique et applications*, Techniques stochastiques. Masson, Paris, 1992.

[10] R. A. Redner and H. F. Walker, "Mixture densities, maximum likelihood and the EM algorithm", *SIAM Review*, vol. 26, no. 2, pp. 195–239, April 1984.

[11] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occuring in the statistical analysis of probabilistic functions of Markov chains", *Annals of Mathematical Statistics*, vol. 41, no. 1, pp. 164–171, 1970.

[12] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm", *Journal of the Royal Statistical Society B*, vol. 39, pp. 1–38, 1977.

[13] R. Godfrey, F. Muir, and F. Rocca, "Modeling seismic impedance with Markov chains", *Geophysics*, vol. 45, no. 9, pp. 1351–1372, September 1980.
[14] R. Hathaway, "A constrained formulation of maximum-likelihood estimation for normal mixture distributions", *The Annals of Statistics*, vol. 13, pp. 1, 1985.
[15] R. J. Hathaway, "A constrained EM algorithm for univariate normal mixtures", *J. Statist. Comput. Simul.*, vol. 23, pp. 211–230, 1986.
[16] G. Ciuperca, A. Ridolfi, and J. Idier, "Penalized maximum likelihood estimator for normal mixtures", Tech. Rep., Université Paris-sud, 2000.
[17] A. Ridolfi and J. Idier, "Penalized maximum likelihood estimation for univariate normal mixture distributions", in *Actes du 17^e Colloque GRETSI*, 1999, pp. 259–262.
[18] R. J. A. Little and D. B. Rubin, "On jointly estimating parameters and missing data by maximizing the complete-data likelihood", *The American Statistician*, vol. 37, pp. 218–220, Aug. 1983.