Journal de la Société Française de Statistique
Vol. 152 No. 3 (2011)

Bayesian Markov model for cooperative clustering: application to robust MRI brain scan segmentation

Titre: Approche bayesienne et markovienne pour des classifications couplées coopératives : application à la segmentation d'IRM du cerveau

Florence Forbes[1], Benoit Scherrer[2] and Michel Dojat[2]

Abstract: Clustering is a fundamental data analysis step that consists of producing a partition of the observations to account for the groups existing in the observed data. In this paper, we introduce an additional cooperative aspect. We address cases in which the goal is to produce not a single partition but two or more possibly related partitions using cooperation between them. Cooperation is expressed by assuming the existence of two sets of labels (group assignments) which are not independent. We also model additional interactions by considering dependencies between labels within each label set. We propose then a cooperative setting formulated in terms of conditional Markov Random Field models for which we provide alternating and cooperative estimation procedures based on variants of the Expectation Maximization (EM) algorithm for inference. We illustrate the advantages of our approach by showing its ability to deal successfully with the complex task of segmenting simultaneously and cooperatively tissues and structures from MRI brain scans.

Résumé : La classification est une étape clef de l'analyse de données qui consiste à produire une partition des données qui traduise l'existence de groupes dans celles-ci. Dans cet article, nous introduisons la notion de classifications coopératives. Nous considérons le cas où l'objectif est de produire deux (ou plus) partitions des données de manière non indépendante mais en prenant en compte les informations que l'une des partitions apporte sur l'autre et réciproquement. Pour ce faire, nous considérons deux (ou plus) jeux d'étiquettes non indépendants. Des interactions supplémentaires entre étiquettes au sein d'un même jeu sont également modélisées pour prendre en compte par exemple des dépendances spatiales. Ce cadre coopératif est formulé à l'aide de modèles de champs de Markov conditionnels dont les paramètres sont estimés par une variante de l'algorithme EM. Nous illustrons les performances de notre approche sur un problème réel difficile de segmentation simultanée des tissus et des structures du cerveau à partir d'images de résonnance magnétique artefactées.

Keywords: Model-based clustering, Markov random fields, Bayesian analysis, EM algorithm, Generalized alternating maximization, Human Brain
Mots-clés : Classification à base de modèles, Champs de Markov, Analyse bayesienne, Algorithme EM, Maximization alternée généralisée, Cerveau humain
AMS 2000 subject classifications: 62F15, 62P10

1. Introduction

Clustering or segmentation of data is a fundamental data analysis step that has received increasing interest in recent years due to the emergence of several new areas of application. Attention has been focused on clustering various data type, regular vector data, curve data or

1. INRIA Grenoble Rhône-Alpes, Grenoble University, Laboratoire Jean Kuntzman, France.
E-mail: florence.forbes@inria.fr
2. INSERM, Grenoble Insitute of Neuroscience, Grenoble University, France.
E-mail: benoitscherrer@gmail.com and E-mail: michel.dojat@ujf-grenoble.fr

more heterogeneous data [8]. In these cases, the goal is to produce a single partition of the observations (*eg.* via a labelling of each data point) that accounts for the groups existing in the observed data. The issue we consider in this paper is that of producing more than one partition using the same data. Examples of applications in which this is relevant include tissue and structure segmentation in Magnetic Resonance (MR) brain scan analysis [26], simultaneous estimation of motion discontinuities and optical flow in motion analysis [17], consistent depth estimation and boundary [22] or depth discontinuity [28] detection, *etc.* To address this goal, we consider a probabilistic missing data framework and assume the existence of two (or more) sets of missing variables. These sets represent two (or more) sets of labels which are related in the sense that information on one of them can help in finding the other. It follows that there is a clear gain in considering the two sets of labels in a single cooperative modelling. Beyond the need for modelling cooperation, in many applications, data points are not independent and require models that account for dependencies. For this purpose, we use Markov random field (MRF) models to further specify our missing data framework. In most non trivial cases, this results in complex systems that include processes interacting on a wide variety of scales. One approach to complex processes in the presence of data is hierarchical modelling. Hierarchical modelling is based on a decomposition of the problem that corresponds to a factorization of the joint distribution

$$p(\mathbf{y},\mathbf{z},\theta) \quad = \quad p(\mathbf{y}|\mathbf{z},\theta)p(\mathbf{z}|\theta)p(\theta), \tag{1}$$

where $\mathbf{Y}$, $\mathbf{Z}$, Θ are random variables denoting respectively the data, the labels and the parameters. We use capital letters to indicate random variables while their realizations are denoted with small letters.

In our cooperative setting, we focus more particularly on situations where the $p(\mathbf{z}|\theta)$ part is made of different sub-processes which are linked and provide complementary information. We propose an approach different from the standard hierarchical modelling. Our approach consists of accounting for the joint model through a series of conditional models but which not necessarily correspond to the factors in the standard factorized decomposition (1). We refer to this alternative decomposition as the *cooperative approach* because the focus is on capturing interactions (cooperations) between the unknown quantities namely, sub-processes and parameters. More specifically, we derive a class of Bayesian joint Markov models based on the specification of a coherently linked system of conditional models that capture several level of interactions. They incorporate **1)** dependencies between variables within each label sets, which is usually referred to as spatial interactions in spatial data (*eg.* image analysis); **2)** relationships between label sets for cooperative aspects (*eg.* between brain tissues and structures in a brain MRI analysis as illustrated in Section 5) and **3)** *a priori* information for consistency with expert knowledge and to encode additional constraints on the parameters via a specific conditional model. Another strength of our approach is that the whole consistent treatment of the model is made possible using the framework of Generalized Alternating Minimization procedures [7] that generalizes the standard EM framework. The decomposition we propose is particularly well adapted to such inference techniques which are based on alternating optimization procedures in which variables of interest are examined in turn and that conditionally on the other variables. It follows a procedure made of steps that are easy to interpret and that can be enriched with additional information.

The paper is organized as follows. In the next section, we present our inference framework and more specifically how the EM algorithm can be used in a Bayesian setting. We show that for estimation purposes, a joint formulation of the model needs not to be explicit and that inference can be performed on the basis of some conditional models only. In Section 3, we present our modelling of cooperation and develop the corresponding EM framework. We show that such a setting can adapt well to our conditional models formulation and simplifies into alternating and cooperative estimation procedures. In Section 4, we further specify our missing data models by considering MRF models and show that inference reduces easily to the standard Hidden MRF models case. Eventually, we illustrate in Section 5 the advantages of this general setting by applying it to the analysis of Magnetic Resonance Imaging (MRI) brain scans. A discussion ends the paper with an appendix containing additional detailed developments.

2. Bayesian analysis of missing data models

The clustering task is addressed via a missing data model that includes a set $\mathbf{y} = \{y_1, \dots, y_N\}$ of observed variables and a set $\mathbf{z} = \{z_1, \dots, z_N\}$ of missing (also called hidden) variables whose joint distribution $p(\mathbf{y}, \mathbf{z}|\theta)$ is governed by a set of parameters denoted θ and possibly by additional hyperparameters not specified in the notation. The latter are usually fixed and not considered at first (see Section 5 for examples of such hyperparameters). Typically, the z_i's corresponding to group memberships (or equivalently label assignments), take their values in $\{e_1, \dots, e_K\}$ where e_k is a K-dimensional binary vector whose k^{th} component is 1, all other components being 0. We will denote by $\mathscr{Z} = \{e_1, \dots, e_K\}^N$ the set in which $\mathbf{z}$ takes its values and by $\underline{\Theta}$ the parameter space. The clustering task consists primarily of estimating the unknown $\mathbf{z}$. However to perform good estimation, the parameters θ values have to be available. A natural approach to estimate the parameters is based on maximum likelihood, θ is estimated as $\hat{\theta} = \arg\max_{\theta \in \underline{\Theta}} p(\mathbf{y}|\theta)$. Then an estimate of $\mathbf{z}$ can be found by maximizing $p(\mathbf{z}|\mathbf{y}, \hat{\theta})$. Note, $p(\mathbf{y}|\theta)$ is a marginal distribution over the unknown $\mathbf{z}$ variables, so that direct maximum likelihood is not in general possible. The Expectation-Maximization (EM) algorithm [21] is a general technique for finding maximum likelihood solutions in the presence of missing data. It consists of two steps usually described as the E-step in which the expectation of the so-called complete log-likelihood is computed and the M-step in which this expectation is maximized over θ. An equivalent way to define EM is the following. Let $\mathscr{D}$ be the set of all probability distributions on $\mathscr{Z}$. As discussed in [7], EM can be viewed as an alternating maximization procedure of a function F defined, for any probability distribution $q \in \mathscr{D}$, by

$$F(q, \theta) \;\; = \;\; \sum_{\mathbf{z} \in \mathscr{Z}} q(\mathbf{z}) \, \log p(\mathbf{y}, \mathbf{z} \mid \theta) + I[q], \tag{2}$$

where $I[q] = -E_q[\log q(\mathbf{Z})]$ is the entropy of q (E_q denotes the expectation with regard to q).

When prior knowledge on the parameters is available, an alternative approach is based on a Bayesian setting. It consists of replacing the maximum likelihood estimation by a maximum a posteriori (MAP) estimation of θ using the prior knowledge encoded in a distribution $p(\theta)$. The maximum likelihood estimate of θ is replaced by $\hat{\theta} = \arg\max_{\theta \in \underline{\Theta}} p(\theta|\mathbf{y})$. The EM algorithm can be used to maximize this posterior distribution. Indeed, the likelihood $p(\mathbf{y}|\theta)$ and $F(q, \theta)$ are linked through $\log p(\mathbf{y}|\theta) = F(q, \theta) + KL(q, p)$ where $KL(q, p)$ is the Kullback-Leibler divergence between q and the conditional distribution $p(\mathbf{z}|\mathbf{y}, \theta)$ and is non-negative, $KL(q, p) = \sum_{\mathbf{z} \in \mathscr{Z}} q(\mathbf{z}) \log \left(\frac{q(\mathbf{z})}{p(\mathbf{z}|\mathbf{y}, \theta)} \right)$. Using the equality $\log p(\theta|\mathbf{y}) = \log p(\mathbf{y}|\theta) + \log p(\theta) - \log p(\mathbf{y})$, it follows $\log p(\theta|\mathbf{y}) = F(q,\theta) + KL(q,p) + \log p(\theta) - \log p(\mathbf{y})$ from which, we get a lower bound $\mathscr{L}(q,\theta)$ on $\log p(\theta|\mathbf{y})$ given by $\mathscr{L}(q,\theta) = F(q,\theta) + \log p(\theta) - \log p(\mathbf{y})$. Maximizing this lower bound alternatively over q and θ leads to a sequence of estimations $\{q^{(r)},\theta^{(r)}\}_{r\in\mathbb{N}}$ satisfying $\mathscr{L}(q^{(r+1)},\theta^{(r+1)}) \geq \mathscr{L}(q^{(r)},\theta^{(r)})$. The maximization over q corresponds to the standard E-step and leads to $q^{(r)}(\mathbf{z}) = p(\mathbf{z}|\mathbf{y},\theta^{(r)})$. It follows that $\mathscr{L}(q^{(r)},\theta^{(r)}) = \log p(\theta^{(r)}|\mathbf{y})$ which means that the lower bound reaches the objective function in $\theta^{(r)}$ and that the sequence $\{\theta^{(r)}\}_{r\in\mathbb{N}}$ increases $p(\theta|\mathbf{y})$ at each step. It then appears that when considering our MAP problem, we can replace (see *eg.* [12]) the function $F(q,\theta)$ by $F(q,\theta) + \log p(\theta)$. The corresponding alternating procedure is: starting from a current value $\theta^{(r)} \in \underline{\Theta}$, set alternatively

$$q^{(r)} = \arg\max_{q\in\mathscr{D}} F(q,\theta^{(r)}) = \arg\max_{q\in\mathscr{D}} \sum_{\mathbf{z}\in\mathscr{Z}} \log p(\mathbf{z}|\mathbf{y},\theta^{(r)})\, q(\mathbf{z}) + I[q] \tag{3}$$

$$\theta^{(r+1)} = \arg\max_{\theta\in\underline{\Theta}} F(q^{(r)},\theta) + \log p(\theta) \;=\; \arg\max_{\theta\in\underline{\Theta}} \sum_{\mathbf{z}\in\mathscr{Z}} \log p(\theta|\mathbf{y},\mathbf{z})\, q^{(r)}(\mathbf{z})\,. \tag{4}$$

More generally, if prior knowledge is available only for a subpart of the parameters, say $w \in \underline{\mathscr{W}}$ where $\theta = (\psi, w) \in \underline{\Psi} \times \underline{\mathscr{W}}$, then for a constant non-informative prior $p(\psi)$ (see [12] for a justification of using improper prior), it follows from $p(w,\mathbf{y}|\psi) \;=\; p(w,\psi|\mathbf{y})\, p(\mathbf{y})\, p(\psi)^{-1}$ that $\arg\max\limits_{(w,\psi)} p(w,\psi|\mathbf{y}) = \arg\max\limits_{(w,\psi)} p(w,\mathbf{y}|\psi)$. Carrying out developments similar as before with conditioning on ψ, we can show that a lower bound on $\log p(w|\mathbf{y},\psi)$ is given by
$\mathscr{L}(q,w,\psi) = F(q,w,\psi) + \log p(w|\psi) - \log p(\mathbf{y}|\psi)$. It follows that a lower bound on $\log p(w,\mathbf{y}|\psi)$ is $F(q,w,\psi) + \log p(w|\psi)$. When w is assumed in addition to be independent of ψ so that $p(w|\psi) = p(w)$, maximizing this lower bound alternatively over q, w and ψ leads to

$$q^{(r)} = \arg\max_{q\in\mathscr{D}} F(q,w^{(r)},\psi^{(r)}) \tag{5}$$

$$w^{(r+1)} = \arg\max_{w\in\underline{\mathscr{W}}} F(q^{(r)},w,\psi^{(r)}) + \log p(w) \tag{6}$$

$$\psi^{(r+1)} = \arg\max_{\psi\in\underline{\Psi}} F(q^{(r)},w^{(r+1)},\psi)\,, \tag{7}$$

where (5) and (7) are respectively regular E and M steps.

Going back to the general case, the last equalities in (3) and (4) come from straightforward probabilistic rules and show that inference can be described in terms of the conditional models $p(\mathbf{z}|\mathbf{y},\theta)$ and $p(\theta|\mathbf{y},\mathbf{z})$. Defining these conditional models is equivalent to defining the conditional distribution $p(\mathbf{z},\theta|\mathbf{y})$. The former distributions can be deduced from the later using the product rule and the later is uniquely defined when the former distributions are given using for instance the following equality,

$$p(\mathbf{z},\theta|\mathbf{y}) = p(\mathbf{z}|\mathbf{y},\theta)\left(\sum_{\mathbf{z}\in\mathscr{Z}} \frac{p(\mathbf{z}|\mathbf{y},\theta)}{p(\theta|\mathbf{y},\mathbf{z})}\right)^{-1}.$$

It follows that for classification (or segmentation) purposes, there is no need to define a joint model $p(\mathbf{y},\mathbf{z},\theta)$, the conditional distribution $p(\mathbf{z},\theta|\mathbf{y})$ contains all useful information. Equivalently, there is no need to specify $p(\mathbf{y})$. This point of view is also the one adopted in Conditional random fields (CRF) [20] which have been widely and successfully used in applications including text processing, bioinformatics and computer vision. CRF's are *discriminative models* in the sense that they model directly the posterior or conditional distribution of the labels given the observations. Explicit models of the joint distribution of the labels and observations or of the observational process $p(\mathbf{y}|\mathbf{z},\theta)$ distribution are not required. In classification issues, the posterior distribution is the one needed and it can appear as a waste of time and computational resources to deal with the joint distribution or with complex observational processes. However, even in classification contexts, approaches that model the joint distribution of the labels and observations are considered. They are known as *generative models*. Such generative models are certainly more demanding in term of modelling but they have other advantages that we will not discuss further in this paper.

3. Cooperative clustering framework

As mentioned in the introduction, the particularity of our clustering task is to include two (or possibly more) label sets of interest which are linked and that we would like to estimate cooperatively using one to gain information on the other. In this section, we particularize the framework described in Section 2. The two label sets under consideration are denoted by $\mathbf{t} = \{t_1,\ldots,t_N\}$ and $\mathbf{s} = \{s_1,\ldots,s_N\}$. We will denoted respectively by $\mathscr{T}$ and $\mathscr{S}$ the spaces in which they take their values. Each observation y_i is now associated to two labels denoted by t_i and s_i. Denoting $\mathbf{z} = (\mathbf{t},\mathbf{s})$, we can apply the EM framework introduced in the previous section to find a MAP estimate $\hat{\theta}$ of θ using the procedure given by (3) and (4) and then generate $\mathbf{t}$ and $\mathbf{s}$ that maximize the conditional distribution $p(\mathbf{t},\mathbf{s}|\mathbf{y},\hat{\theta})$. Note that this is however not equivalent to maximizing over $\mathbf{t},\mathbf{s}$ and θ the posterior distribution $p(\mathbf{t},\mathbf{s},\theta|\mathbf{y})$. Indeed $p(\mathbf{t},\mathbf{s},\theta \mid \mathbf{y}) = p(\mathbf{t},\mathbf{s}|\mathbf{y},\theta)\; p(\theta|\mathbf{y})$ and in the modified EM setting (eq. (3) and (4)), θ is found by maximizing the second factor only. The problem is greatly simplified when the solution is determined within the EM algorithm framework.

However, solving the optimization (3) over the set $\mathscr{D}$ of probability distributions $q_{(T,S)}$ on $(\mathbf{T},\mathbf{S})$ leads for the optimal $q_{(T,S)}$ to $p(\mathbf{t},\mathbf{s}|\mathbf{y},\theta^{(r)})$ which may remain intractable for complex models. In our cooperative context, we therefore propose an EM variant in which the E-step is not performed exactly. The optimization (3) is solved instead over a restricted class of probability distributions $\tilde{\mathscr{D}}$ which is chosen as the set of distributions that factorize as $q_{(T,S)}(\mathbf{t},\mathbf{s}) = q_T(\mathbf{t})\; q_S(\mathbf{s})$ where q_T (resp. q_S) belongs to the set $\mathscr{D}_T$ (resp. $\mathscr{D}_S$) of probability distributions on $\mathbf{T}$ (resp. on $\mathbf{S}$). This variant is usually referred to as Variational EM [19]. It follows that the E-step becomes an approximate E-step,

$$(q_T^{(r)},q_S^{(r)}) = \arg\max_{(q_T,q_S)} F(q_T q_S,\theta^{(r)})\;.$$

This step can be further generalized by decomposing it into two stages. At iteration r, with current estimates denoted by $q_T^{(r-1)},q_S^{(r-1)}$ and $\theta^{(r)}$, we consider the following updating,

E-T-step: $q_T^{(r)} = \arg\max_{q_T\in\mathscr{D}_T} F(q_T\; q_S^{(r-1)},\theta^{(r)})$

E-S-step: $q_S^{(r)} = \arg\max_{q_S\in\mathscr{D}_S} F(q_T^{(r)}\; q_S,\theta^{(r)})$.

The effect of these iterations is to generate sequences of paired distributions and parameters $\{q_T^{(r)},q_S^{(r)},\theta^{(r)}\}_{r\in\mathbb{N}}$ that satisfy $F(q_T^{(r+1)}q_S^{(r+1)},\theta^{(r+1)}) \geq F(q_T^{(r)}q_S^{(r)},\theta^{(r)})$. This variant falls in the modified Generalized Alternating Minimization (GAM) procedures family for which convergence results are available [7].

We then derive two equivalent expressions of F when q factorizes as in $\tilde{\mathscr{D}}$. Expression (2) of F can be rewritten as $F(q,\theta) = E_q[\log p(\mathbf{T}|\mathbf{S},\mathbf{y},\theta)] + E_q[\log p(\mathbf{S},\mathbf{y}|\theta)] + I[q]$. Then,

$$\begin{aligned} F(q_T\, q_S,\theta) &= E_{q_T}[E_{q_S}[\log p(\mathbf{T}|\mathbf{S},\mathbf{y},\theta)]] + E_{q_S}[\log p(\mathbf{S},\mathbf{y}|\theta)] + I[q_T\, q_S] \\ &= E_{q_T}[E_{q_S}[\log p(\mathbf{T}|\mathbf{S},\mathbf{y},\theta)]] + I[q_T] + G[q_S]\ , \end{aligned}$$

where $G[q_S] = E_{q_S}[\log p(\mathbf{S},\mathbf{y}|\theta)] + I[q_S]$ is an expression that does not depend on q_T. Using the symmetry in $\mathbf{T}$ and $\mathbf{S}$, it is easy to show that similarly,

$$\begin{aligned} F(q_T\, q_S,\theta) &= E_{q_S}[E_{q_T}[\log p(\mathbf{S}|\mathbf{T},\mathbf{y},\theta)]] + E_{q_T}[\log p(\mathbf{T},\mathbf{y}|\theta)] + I[q_T\, q_S] \\ &= E_{q_S}[E_{q_T}[\log p(\mathbf{S}|\mathbf{T},\mathbf{y},\theta)]] + I[q_S] + G'[q_T]\ , \end{aligned}$$

where $G'[q_T] = E_{q_T}[\log p(\mathbf{T},\mathbf{y}|\theta)] + I[q_T]$ is an expression that does not depend on q_S. It follows that the E-T and E-S steps reduce to,

$$\textbf{E-T-step: } q_T^{(r)} = \arg\max_{q_T\in\mathscr{D}_T} E_{q_T}[E_{q_S^{(r-1)}}[\log p(\mathbf{T}|\mathbf{S},\mathbf{y},\theta^{(r)})]] + I[q_T] \tag{8}$$

$$\textbf{E-S-step: } q_S^{(r)} = \arg\max_{q_S\in\mathscr{D}_S} E_{q_S}[E_{q_T^{(r)}}[\log p(\mathbf{S}|\mathbf{T},\mathbf{y},\theta^{(r)})]] + I[q_S] \tag{9}$$

and the **M-step**

$$\theta^{(r+1)} = \arg\max_{\theta\in\Theta} E_{q_T^{(r)} q_S^{(r)}}[\log p(\theta|\mathbf{y},\mathbf{T},\mathbf{S})]\ . \tag{10}$$

More generally, we can adopt in addition, an incremental EM approach [7] which allows re-estimation of the parameters (here θ) to be performed based only on a sub-part of the hidden variables. This means that we can incorporate an M-step (4) in between the updating of q_T and q_S. Similarly, hyperparameters could be updated there too.

It appears in equations (8), (9) and (10) that for inference the specification of the three conditional distributions $p(\mathbf{t}|\mathbf{s},\mathbf{y},\theta)$, $p(\mathbf{s}|\mathbf{t},\mathbf{y},\theta)$ and $p(\theta|\mathbf{t},\mathbf{s},\mathbf{y})$ is necessary and sufficient. In practice, the advantage of writing things in terms of the conditional distributions $p(\mathbf{t}|\mathbf{s},\mathbf{y},\theta)$ and $p(\mathbf{s}|\mathbf{t},\mathbf{y},\theta)$ is that it allows to capture cooperations between $\mathbf{t}$ and $\mathbf{s}$ as will be illustrated in Section 5. Then, steps E-T and E-S have to be further specified by computing the expectations with regards to $q_S^{(r-1)}$ and $q_T^{(r)}$. In the following section, we specify a way to design such conditional distributions in a Markov modelling context.

4. Markov model clustering

We further specify our missing data model to account for dependencies between data points and propose an appropriate way to build conditional distributions for the model inference. Let V be a finite set of N sites indexed by i with a neighborhood system defined on it. A set of sites c is called a clique if it contains sites that are all neighbors. We define a Markov random field (MRF) as a collection of random variables defined on V whose joint probability distribution is a Gibbs distribution [14]. More specifically, we assume that $\mathbf{y},\mathbf{z}$ and θ are all defined on V (more general cases are easy to derive). The specification of θ as a possibly data point specific parameter,

$\theta = \{\theta_1 \dots \theta_N\}$, may seem awkward as parameters at each site is likely to yield intense problems. However, note that we are in a bayesian setting so that a prior can be defined on θ. An example is given in Section 5.3.2. We then assume in addition that the conditional distribution $p(\mathbf{z}, \theta|\mathbf{y})$ is a Markov random field with energy function $H(\mathbf{z}, \theta|\mathbf{y})$, *ie.*

$$p(\mathbf{z}, \theta|\mathbf{y}) \quad \propto \quad \exp(H(\mathbf{z}, \theta|\mathbf{y})), \tag{11}$$

with $H(\mathbf{z}, \theta|\mathbf{y}) = \sum_{c\in\Gamma}\left(U^c_{\mathbf{Z},\Theta}(\mathbf{z}_c, \theta_c|\mathbf{y}) + U^c_{\mathbf{Z}}(\mathbf{z}_c|\mathbf{y}) + U^c_{\Theta}(\theta_c|\mathbf{y})\right)$, where the sum is over the set of cliques Γ and $\mathbf{z}_c$ and θ_c denote realizations restricted to clique c. The U^c's are the clique potentials that may depend on additional parameters, not specified in the notation. In addition, in the formula above, terms that depend only on $\mathbf{z}$, resp. θ, are written explicitly and are distinguished from the first term in the sum in which $\mathbf{z}$ and θ cannot be separated. Conditions ensuring the existence of such a distribution can be found in [15].

From the Markovianity of the joint distribution it follows that any conditional distribution is also Markovian. Note that this is not true for marginals of a joint Markov field which are not necessarily Markovian [3]. For instance, $p(\mathbf{z}|\mathbf{y}, \theta)$ and $p(\theta|\mathbf{y}, \mathbf{z})$ are Markov random fields with energy functions given respectively by $H(\mathbf{z}|\mathbf{y}, \theta) = \sum_{c\in\Gamma} U^c_{\mathbf{Z},\Theta}(\mathbf{z}_c, \theta_c|y) + U^c_{\mathbf{Z}}(\mathbf{z}_c|\mathbf{y})$, and $H(\theta|\mathbf{y}, \mathbf{z}) = \sum_{c\in\Gamma} U^c_{\mathbf{Z},\Theta}(\mathbf{z}_c, \theta_c|y) + U^c_{\Theta}(\theta_c|y)$, where terms depending only on θ, resp. on $\mathbf{z}$, disappear because they cancel out between the numerator and denominator of the Gibbs form (11).

Natural examples of Markovian distributions $p(\mathbf{z}, \theta|\mathbf{y})$ are given by the standard Hidden Markov random fields referred to as HMF-IN for Hidden Markov Field with Independent Noise in [3]. HMF-IN, considering the couple of variables $(\mathbf{Z}, \Theta)$ as the unknown couple variable, are defined through two assumptions:

(i) $p(\mathbf{y}|\mathbf{z}, \theta) = \prod_{i\in V} p(y_i|z_i, \theta_i) = \prod_{i\in V} \dfrac{g(y_i|z_i, \theta_i)}{W_i(z_i, \theta_i)}$ where in the last equality the $g(y_i|z_i, \theta_i)'s$ are positive functions of y_i that can be normalized and the $W_i(z_i, \theta_i)$'s are normalizing constants that do not depend on y_i. We write explicitly the possibility to start with unnormalized quantities as this would be useful later.

(ii) $p(\mathbf{z}, \theta)$ is a Markov random field with energy function $H(\mathbf{z}, \theta)$.

It follows from **(i)** and **(ii)** that $p(\mathbf{z}, \theta|\mathbf{y})$ is a Markov random field too with energy function

$$\begin{aligned} H(\mathbf{z}, \theta|\mathbf{y}) &= H(\mathbf{z}, \theta) + \sum_{i\in V} \log g(y_i|z_i, \theta_i) - \sum_{i\in V} \log W_i(z_i, \theta_i) \\ &= H'(\mathbf{z}, \theta) + \sum_{i\in V} \log g(y_i|z_i, \theta_i)\ . \end{aligned}$$

where it is easy to see that H' still corresponds to a Markovian energy on $(\mathbf{z}, \theta)$.

Conversely, given such an energy function it is always possible to find the corresponding HMF-IN as defined by **(i)** and **(ii)**, by normalizing the g's and defining $H(\mathbf{z}, \theta) = H'(\mathbf{z}, \theta) + \sum_{i\in V} \log W_i(z_i, \theta_i)$. Therefore equivalently, we will call HMF-IN any Markov field whose energy function is $H(\mathbf{z}, \theta) + \sum_{i\in V} \log g(y_i|z_i, \theta_i)$ where $H(\mathbf{z}, \theta)$ is the energy of a MRF on $(\mathbf{z}, \theta)$ and the g's are positive functions of y_i that can be normalized. We will see in our cooperative context the advantage of using unnormalized data terms. Let us then consider MRF $p(\mathbf{z}, \theta|\mathbf{y})$ that are HMF-IN, *ie.* whose energy function can be written as

$$H(\mathbf{z}, \theta|\mathbf{y}) \quad = \quad H_{\mathbf{Z}}(\mathbf{z}) + H_{\Theta}(\theta) + H_{\mathbf{Z},\Theta}(\mathbf{z}, \theta) + \sum_{i\in V} \log g(y_i|z_i, \theta_i)\ . \tag{12}$$

The Markovian energy is separated into terms $H_{\mathbf{Z}}$, H_{Θ}, $H_{\mathbf{Z},\Theta}$ that involve respectively only $\mathbf{z}$, only θ and interactions between θ and $\mathbf{z}$.

In a cooperative framework, we assume that $\mathbf{z} = (\mathbf{t}, \mathbf{s})$ so that we can further specify

$$\begin{aligned} H_{\mathbf{Z}}(\mathbf{z}) &= H_{\mathbf{T}}(\mathbf{t}) + H_{\mathbf{S}}(\mathbf{s}) + \tilde{H}_{\mathbf{T},\mathbf{S}}(\mathbf{t}, \mathbf{s}) \qquad (13) \\ \text{and} \quad H_{\mathbf{Z},\Theta}(\mathbf{z}, \theta) &= H_{\mathbf{T},\Theta}(\mathbf{t}, \theta) + H_{\mathbf{S},\Theta}(\mathbf{s}, \theta) + \tilde{H}_{\mathbf{T},\mathbf{S},\Theta}(\mathbf{t}, \mathbf{s}, \theta) , \qquad (14) \end{aligned}$$

where we used a different notation $\tilde{H}$ to make clearer the difference between the energy terms involving interactions only (resp. $\tilde{H}_{\mathbf{T},\mathbf{S}}$ and $\tilde{H}_{\mathbf{T},\mathbf{S},\Theta}$) and the global energy terms (resp. $H_{\mathbf{Z}}$ and $H_{\mathbf{Z},\Theta}$). We will provide examples of these different terms in Section 5.2.2.

$H_{\Theta}(\theta)$ and $H_{\mathbf{Z}}(\mathbf{z})$ can be interpreted as priors resp. on Θ and $\mathbf{Z}$. In a cooperative framework, the prior on $\mathbf{Z}$ can be itself decomposed into an *a priori* cooperation term $\tilde{H}_{\mathbf{T},\mathbf{S}}(\mathbf{t}, \mathbf{s})$ and individual terms which represent *a priori* information on $\mathbf{T}$ and $\mathbf{S}$ separately. $H_{\mathbf{T},\mathbf{S},\Theta}(\mathbf{t}, \mathbf{s}, \theta)$ specifies the *process*, *ie.* the underlying model, that can also be decomposed into parts involving $\mathbf{t}$ and $\mathbf{s}$ separately or together. In what follows, we will assume that $\mathbf{t}$ and $\mathbf{s}$ are both defined on the set of sites V so that writing $z_i = (t_i, s_i)$ makes sense. With additional care, a more general situation could be considered if necessary. Eventually $\sum_{i \in V} \log g(y_i | t_i, s_i, \theta_i)$ corresponds to the data-term. An example is given in Section 5.2.1.

From such a definition of $p(\mathbf{z}, \theta | \mathbf{y})$, it follows expressions of the conditional distributions required for inference in steps (8) to (10). As already mentioned, the Markovianity of $p(\mathbf{z}, \theta | \mathbf{y})$ implies that the conditional distributions $p(\mathbf{t} | \mathbf{y}, \mathbf{s}, \theta)$ and $p(\mathbf{s} | \mathbf{y}, \mathbf{t}, \theta)$ are also Markovian with respective energy

$$\begin{aligned} H(\mathbf{t} | \mathbf{s}, \mathbf{y}, \theta) = \; & H_{\mathbf{T}}(\mathbf{t}) + \tilde{H}_{\mathbf{T},\mathbf{S}}(\mathbf{t}, \mathbf{s}) + H_{\mathbf{T},\Theta}(\mathbf{t}, \theta) + \tilde{H}_{\mathbf{T},\mathbf{S},\Theta}(\mathbf{t}, \mathbf{s}, \theta) + \\ & \sum_{i \in V} \log g(y_i | t_i, s_i, \theta_i) \qquad (15) \end{aligned}$$

$$\begin{aligned} \text{and} \quad H(\mathbf{s} | \mathbf{t}, \mathbf{y}, \theta) = \; & H_{\mathbf{S}}(\mathbf{s}) + \tilde{H}_{\mathbf{T},\mathbf{S}}(\mathbf{t}, \mathbf{s}) + H_{\mathbf{S},\Theta}(\mathbf{s}, \theta) + \tilde{H}_{\mathbf{T},\mathbf{S},\Theta}(\mathbf{t}, \mathbf{s}, \theta) \\ & + \sum_{i \in V} \log g(y_i | t_i, s_i, \theta_i), \qquad (16) \end{aligned}$$

omitting the terms that do not depend on $\mathbf{t}$ (resp. $\mathbf{s}$). Similarly,

$$H(\theta | \mathbf{t}, \mathbf{s}, \mathbf{y}) = H_{\Theta}(\theta) + H_{\mathbf{T},\mathbf{S},\Theta}(\mathbf{t}, \mathbf{s}, \theta) + \sum_{i \in V} \log g(y_i | t_i, s_i, \theta_i).$$

4.1. Inference

In (8), (9) and (10) the respective normalizing constant terms can be ignored because they are respectively independent of $\mathbf{T}$, $\mathbf{S}$ and Θ. It comes that the E-steps are equivalent to

$$\textbf{E-T-step:} \quad q_T^{(r)} = \arg\max_{q_T \in \mathcal{D}_T} E_{q_T}[E_{q_S^{(r-1)}}[H(\mathbf{T} | \mathbf{S}, \mathbf{y}, \theta^{(r)})]] + I[q_T] \qquad (17)$$

$$\textbf{E-S-step:} \quad q_S^{(r)} = \arg\max_{q_S \in \mathcal{D}_S} E_{q_S}[E_{q_T^{(r)}}[H(\mathbf{S} | \mathbf{T}, \mathbf{y}, \theta^{(r)})]] + I[q_S] \qquad (18)$$

Then, steps E-T and E-S can be further specified by computing the expectations with regards to $q_S^{(r-1)}$ and $q_T^{(r)}$. An interesting property is that if $H(\mathbf{z},\theta|\mathbf{y})$ defines an HMF-IN of the form (12), then $E_{q_S^{(r-1)}}[H(\mathbf{t}|\mathbf{S},\mathbf{y},\theta^{(r)})]$ and $E_{q_T^{(r)}}[H(\mathbf{s}|\mathbf{T},\mathbf{y},\theta^{(r)})]$ are also HMF-IN energies. Indeed denoting $H_T^{(r)}(\mathbf{t}) = E_{q_S^{(r-1)}}[H(\mathbf{t}|\mathbf{S},\mathbf{y},\theta^{(r)})]$, it follows from expression (15) that

$$\begin{aligned} H_T^{(r)}(\mathbf{t}) &= H_{\mathbf{T}}(\mathbf{t}) + H_{\mathbf{T},\Theta}(\mathbf{t},\theta^{(r)}) + \\ & \sum_{\mathbf{s}\in\mathscr{S}} q_{\mathbf{S}}^{(r-1)}(\mathbf{s}) \left(\tilde{H}_{\mathbf{T},\mathbf{S}}(\mathbf{t},\mathbf{s}) + \tilde{H}_{\mathbf{T},\mathbf{S},\Theta}(\mathbf{t},\mathbf{s},\theta^{(r)}) + \sum_{i\in V} \log g(y_i|t_i,s_i,\theta_i^{(r)}) \right), \end{aligned}$$

which can be viewed as an HMF-IN energy on $\mathbf{t}$. It appears then that step E-T is equivalent to the E-step one would get when applying EM to a standard Hidden MRF in $\mathbf{t}$. Equivalently, the same conclusion holds for $H_S^{(r)}(\mathbf{s}) = E_{q_T^{(r+1)}}[H(\mathbf{s}|\mathbf{T},\mathbf{y},\theta^{(r)})]$ when exchanging $\mathbf{S}$ and $\mathbf{T}$. Examples of such derived MRF's are given in Section 5.3.1.

The **M-step** (10) is then equivalent to

$$\theta^{(r+1)} = \arg\max_{\theta\in\underline{\Theta}} E_{q_T^{(r)} q_S^{(r)}}[H(\theta|\mathbf{y},\mathbf{T},\mathbf{S})] \tag{19}$$

which can be further specified as

$$\theta^{(r+1)} = \arg\max_{\theta\in\underline{\Theta}} H_\Theta(\theta) + E_{q_T^{(r)} q_S^{(r)}}[H_{\mathbf{T},\mathbf{S},\Theta}(\mathbf{T},\mathbf{S},\theta)] + \sum_{i\in V} E_{q_{T_i}^{(r)} q_{S_i}^{(r)}}[\log g(y_i|T_i,S_i,\theta_i)] \tag{20}$$

The key-point emphasized by these last derivations of our E and M steps is that it is possible to go from a joint cooperative model to an alternating procedure in which each step reduces to an intuitive well identified task. The goal of the above developments was to propose a well based strategy to reach such derivations. When cooperation exists, intuition is that it should be possible to specify stages where each variable of interest is considered in turn but in a way that uses the other variables current information. Interpretation is easier because in each such stage the central part is played by one of the variable at a time. Inference is facilitated because each step can be recast into a well identified (Hidden MRF) setting for which a number of estimation techniques are available. However, the rather general formulation we chose to present may fail in really emphasizing all the advantages of this technique. The goal of the following section is to further point out the good features of our approach by addressing a challenging practical issue and showing that original and very promising results can be obtained. To this end, Figure 1 shows in advance the graphical model representation of the model developed in the next section.

5. Application to MR brain scan segmentation

The framework proposed in the previous sections can apply to a number of areas. However, its description would be somewhat incomplete without some further specifications on how to set such a model in practice. In this section, we address the task of segmenting both tissues and structures in MR brain scans and illustrate our model ability to capture and integrate very naturally the desired cooperative features, and that at several levels.

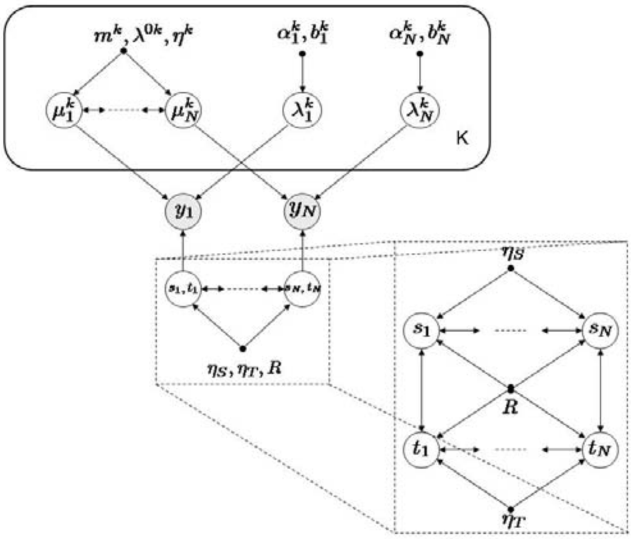

FIGURE 1. *Graphical model representation for the cooperative clustering model developed for the MRI application in section 5: two label sets are considered, resp.* $\{s_1,\dots s_N\}$ *and* $\{t_1,\dots t_N\}$*. The dashed lines show a zoom that specifies the interactions at the* $\{s_i,t_i\}$ *level: 1) within each label set with two MRFs with respective interaction parameter* η_S *and* η_T *and 2) between label sets with the intervention of a common parameter R specific to the MRI application. MRF regularization also occurs at the parameters level via smoothing of the means* $\{\mu_1^k \dots \mu_N^k\}$ *that appear in the data model (Section 5.2.1). See also Section 5.3.2 for details on the hyperparameters* m^k, λ^{0k} *and* η^k*.*

MR brain scans consist of 3D-data referred to as volumes and composed of voxels (volume elements). We generally consider the segmentation of the brain in three tissues: cephalo-spinal-fluid (CSF), grey matter (GM) and white matter (WM) (see Figure 2 (b)). Statistical based approaches usually aim at modelling probability distributions of voxel intensities with the idea that such distributions are tissue-dependent. The segmentation of subcortical structures is another fundamental task. Subcortical structures are regions in the brain (see top of Figure 2 (c)) known to be involved in various brain functions. Their segmentation and volume computation are of interest in various neuroanatomical studies such as brain development or disease stages follow-up. Difficulties in automatic MR brain scan segmentation arise from various sources (see Figure 2 (a) and (c)). The automatic segmentation of subcortical structures usually requires the introduction of *a priori* knowledge via an atlas describing anatomical structures. This atlas has to be first registered to the image to be used in the subsequent segmentation. Most of the proposed approaches share three main characteristics. First, tissue and subcortical structure segmentations are considered as two successive tasks and treated independently although they are clearly linked: a structure is composed of a specific known tissue, and knowledge about structures locations provides valuable information about local intensity distribution for a given tissue. Second, tissue models are estimated globally through the entire volume and then suffer from imperfections at a local level as illustrated in Figure 2 (a). Recently, good results have been reported using an innovative local and cooperative approach called LOCUS [25]. It performs tissue and subcortical structure segmentation by distributing through the volume a set of local Markov random field (MRF) models which better reflect local intensity distributions. Local MRF models are used alternatively for tissue and structure segmentations. Although satisfying in practice, these tissue and structure MRF's do not correspond to a valid joint probabilistic model and are not compatible in that sense. As a consequence, important issues such as convergence or other theoretical properties of the resulting local procedure cannot be addressed. In addition, a major difficulty inherent to local approaches is to ensure consistency of local models. Although satisfying in practice, the cooperation mechanisms between local models proposed in [25] are somewhat arbitrary and independent of the MRF models themselves. Third, with notable exceptions like [23], most atlas-based algorithms perform registration and segmentation sequentially, committing to the initial aligned information obtained in a pre-processing registration step. This is necessarily sub-optimal in the sense that it does not exploit complementary aspects of both problems.

In this section we show how we can use our Bayesian cooperative framework to define a joint model that links local tissue and structure segmentations but also the model parameters so that both types of cooperations, between tissues and structures and between local models, are deduced from the joint model and optimal in that sense. Our model has the following main features: 1) cooperative segmentation of both tissues and structures is encoded via a joint probabilistic model which captures the relations between tissues and structures; 2) this model specification also integrates external *a priori* knowledge in a natural way and allows to combine registration and segmentation; 3) intensity nonuniformity is handled by using a specific parametrization of tissue intensity distributions which induces local estimations on subvolumes of the entire volume and 4) global consistency between local estimations is automatically ensured by using a MRF spatial prior for the intensity distributions parameters.

We will refer to our joint model as LOCUS$^{\mathrm{B}}$, for LOcal Cooperative Unified Segmentation in a Bayesian framework. It is based on ideas partly analyzed previously in the so-called LOCUS method [25] with the addition of a powerful and elegant formalization provided by the extra Bayesian perspective.

5.1. A priori knowledge on brain tissues and structures

In this section, V is a set of N voxels on a regular 3D grid. The observations $\mathbf{y} = \{y_1, \dots, y_N\}$ are intensity values observed respectively at each voxel and $\mathbf{t} = \{t_1, \dots, t_N\}$ represents the hidden tissue classes. The t_i's take their values in $\{e_1, e_2, e_3\}$ that represents the three tissues cephalo-spinal-fluid, grey matter and white matter. In addition, we consider L subcortical structures and denote by $\mathbf{s} = \{s_1, \dots, s_N\}$ the hidden structure classes at each voxel. Similarly, the s_i's take their values in $\{e'_1, \dots, e'_L, e'_{L+1}\}$ where e'_{L+1} corresponds to an additional background class. Our approach aims at taking advantage of the relationships existing between tissues and structures. In particular, a structure is composed of an *a priori* known, single and specific tissue. We will therefore denote by T^{s_i} this tissue for structure s_i at voxel i. If $s_i = e'_{L+1}$, *ie.* voxel i does not belong to any structure, then we will use the convention that $e_{T^{s_i}} = \mathbf{0}$ the 3-dimensional null vector.

As parameters θ, we will consider $\theta = \{\psi, R\}$ where ψ are the parameters describing the intensity distributions for the $K = 3$ tissue classes and R denotes registration parameters described below. The corresponding parameter spaces are denoted by $\underline{\Psi}$ and $\mathscr{R}$. Intensity distribution parameters are more specifically denoted by $\psi = \{\psi_i^k, i \in V, k = 1, \dots, K\}$. We will write (^{t} means transpose) for all $k = 1, \dots, K$, $\psi^k = \{\psi_i^k, i \in V\}$ and for all $i \in V$, $\psi_i = {}^t(\psi_i^k, k = 1, \dots, K)$.

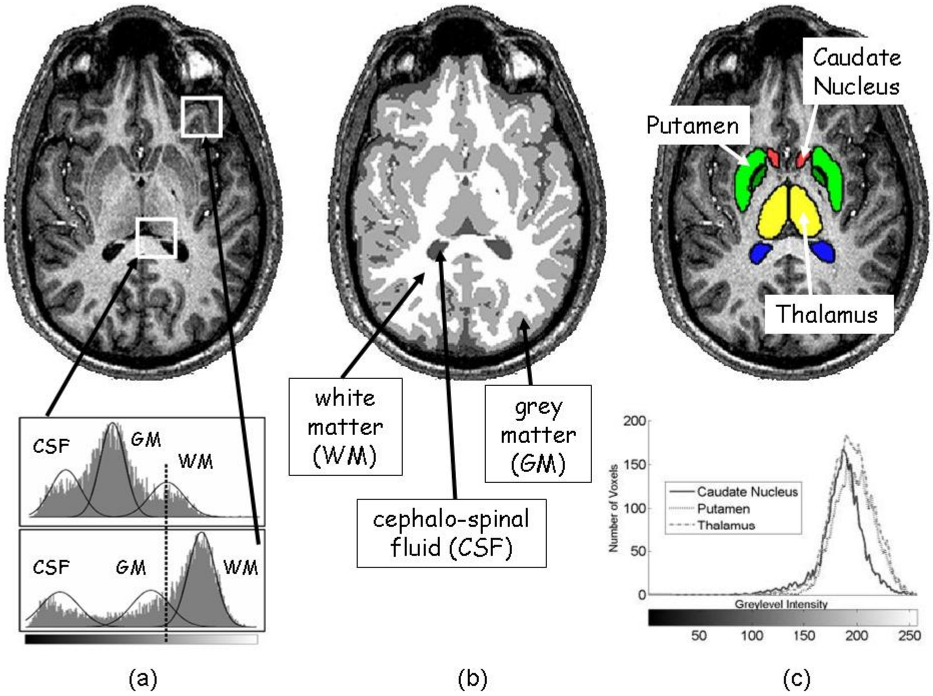

FIGURE 2. *Obstacles to accurate segmentation of MR brain scans. Image (a) illustrates spatial intensity variations: two local intensity histograms (bottom) in two different subvolumes (top) are shown with their corresponding Gaussians fitted using 3-component mixture models for the 3 brain tissues considered. The vertical line corresponds to some intensity value labelled as grey matter or white matter depending on the subvolume. Image (b) illustrates a segmentation in 3 tissues, white matter, grey matter and cephalo spinal fluid. Image (c) shows the largely overlapping intensity histograms (bottom) of 3 grey matter structures segmented manually (top), the putamen (green), the thalamus (yellow) and the caudate nuclei (red).*

Standard approaches usually consider that intensity distributions are Gaussian distributions with parameters that depend only on the tissue class. *A priori* knowledge is incorporated through fields f_T and f_S representing *a priori* information respectively on tissues and on structures. In our study, these fields correspond to prior probabilities provided by a registered probabilistic atlas on structures. They depend on the registration parameters R. We will write $f_T(R) = \{f_T(R,i), i \in V\}$ (resp. $f_S(R) = \{f_S(R,i), i \in V\}$) where $f_T(R,i) = {}^t(f_T^k(R,i), k = 1,\dots,K)$ (resp. $f_S(R,i) = {}^t(f_S^l(R,i), l = 1,\dots,L+1)$) and $f_T^k(R,i)$ (resp. $f_S^l(R,i)$) represents some prior probability that voxel i belongs to tissue k (resp. structure l). As already discussed, most approaches first register the atlas to the medical image and then segment the medical image based on that aligned information. This may induce biases caused by commitment to the initial registration. In our approach we will perform registration and segmentation simultaneously by considering that the information provided by the atlas depends on the registration parameters R that have to be estimated as well as other model parameters and whose successive values will adaptively modify the registration. More specifically, we consider a local affine non rigid registration model as in [23]. We use a hierarchical registration framework which distinguishes between global- and structure-dependent deformations. The mapping of the atlas to the image space is performed by an interpolation function $r(R,i)$ which maps voxel i into the coordinate system defined by R. We model dependency across structures by decomposing R into $R = \{R_G, R_S\}$. R_G are the global registration parameters, which describe the nonstructure-dependent deformations between atlas and image. The structure-dependent parameters $R_S = \{R_1,\dots,R_L,R_{L+1}\}$ are defined in relation to R_G and capture the residual structure-specific deformations that are not adequately explained by R_G. We refer to R_l, the l^{th} entry of R_S, as the registration parameters specific to structure l with $l \in \{1,\dots,L+1\}$. The atlas is denoted by $\phi = \{\phi_l, l = 1,\dots,L+1\}$ where the atlas spatial distribution for a single structure l is represented in our model by ϕ_l. The function ϕ_l is defined in the coordinate system of the atlas space, which is in general different from the image space. We align the atlas to the image space by making use of the interpolation function $r(R_G,R_l,i)$ where R_G and the R_l's correspond to affine non rigid transformations determined through 12 parameters each, capturing the displacement, rotation, scaling and shear 3D vectors. It follows the definition of f_S, $f_S^l(R,i) = \dfrac{\phi_l(r(R_G,R_l,i))}{\sum\limits_{l'=1}^{L+1} \phi_{l'}(r(R_G,R_{l'},i))}$. The normalization across all structures is necessary as the coordinate system of each structure is characterized by the structure-dependent registration parameters R_l. Unlike global affine registration methods, this results in structure-dependent coordinate systems that are not aligned with each other. In other words, multiple voxels in the atlas space can be mapped to one location in the image space. Although the same kind of information (atlas) is potentially available independently for tissues, in our setting we then build f_T from f_S. The quantity $f_T^k(R,i)$ is interpreted as a prior probability that voxel i belongs to tissue k. This event occurs when either voxel i belongs to a structure made of tissue k or when voxel i does not belong to any structure but in this later case we assume that, without any further information, the probability of a particular tissue k is 1/3. It follows the expression, $f_T^k(R,i) = \sum\limits_{l\ st.\ T^l=k} f_S^l(R,i) + \frac{1}{3} f_S^{L+1}(R,i)$, with the convention that the sum is null when the set $\{l\ st.\ T^l = k\} = \emptyset$ which means that there are no structure made of tissue k. This is always the case for the value of k corresponding to white matter. In practice, the global R_G transformation is estimated in a pre-processing step using some standard method such as FLIRT [18]. The other structure specific registration parameters are estimated using our joint modelling and updated as specified in Section 5.3.2.

5.2. Tissue and structure cooperative model

For our brain scan segmentation purpose, we propose to define an HMF-IN of the form (12). The specification of the energy in (12) is decomposed into three parts described below.

5.2.1. **Data term**

With $z_i = (t_i, s_i)$, the data term refers to the term $\sum\limits_{i \in V} \log g(y_i|t_i, s_i, \theta_i)$ in (12). For brain data, this data term corresponds to the modelling of tissue dependent intensity distributions and therefore does not depend on the registration parameters R. The data term reduces then to the definition of function $g(y_i|t_i, s_i, \psi_i)$. We denote by $\mathscr{G}(y|\mu, \lambda)$ the Gaussian distribution with mean μ and precision λ (the precision is the inverse of the variance). Notation $< t_i, \psi_i >$ stands for the scalar product between t_i and ψ_i seen as K-dimensional vectors, so that when $t_i = e_k$ then $< t_i, \psi_i >= \psi_i^k$. Note that we extend this convention to multi-component ψ_i^k such as $\psi_i^k = \{\mu_i^k, \lambda_i^k\}$. Therefore when $t_i = e_k$, $\mathscr{G}(y_i| < t_i, \psi_i >)$ denotes the Gaussian distribution with mean μ_i^k and precision λ_i^k. We will say that both structure and tissue segmentations agree at voxel i, when either the tissue of structure s_i is t_i or when s_i corresponds to the background class so that any value of t_i is compatible with s_i. Using our notation, agreement corresponds then to $t_i = e_{T^{s_i}}$ or $s_i = e'_{L+1}$. In this case, it is natural to consider that the intensity distribution should be $\mathscr{G}(y_i| < t_i, \psi_i >)$. Whereas, when this is not the case, a compromise such as $\mathscr{G}(y_i| < t_i, \psi_i >)^{1/2}\, \mathscr{G}(y_i| < e_{T^{s_i}}, \psi_i >)^{1/2}$ is more appropriate. It is easy to see that the following definition unifies these two cases in a single expression:

$$g(y_i|t_i, s_i, \psi_i) = \mathscr{G}(y_i| < t_i, \psi_i >)^{\frac{(1+<s_i, e'_{L+1}>)}{2}} \mathscr{G}(y_i| < e_{T^{s_i}}, \psi_i >)^{\frac{(1-<s_i, e'_{L+1}>)}{2}} .$$

Note that g as defined above is not in general normalized and would require normalization to be seen as a proper probability distribution. However, as already mentioned in Section 4 it is not required in our framework.

5.2.2. **Missing data term**

We refer to the terms $H_{\mathbf{Z}}(\mathbf{z})$ and $H_{\mathbf{Z},\Theta}(\mathbf{z}, \theta)$ involving $\mathbf{z} = (\mathbf{t}, \mathbf{s})$ in (12) as the missing data term. We will first describe the more general case involving unknown registration parameters R. We will show then for illustration how this case simplifies when registration is done beforehand as a preprocessing. Denoting by $U_{ij}^T(t_i, t_j; \eta_T)$ and $U_{ij}^S(s_i, s_j; \eta_S)$ pairwise potential functions with interaction parameters η_T and η_S, $H_{\mathbf{Z}}(\mathbf{z})$ decomposes as in (13) in three terms which are set for the MRI application as $H_{\mathbf{T}}(\mathbf{t}) = \sum_{i \in V} \sum_{j \in \mathscr{N}(i)} U_{ij}^T(t_i, t_j; \eta_T)$ and similarly $H_{\mathbf{S}}(\mathbf{s}) = \sum_{i \in V} \sum_{j \in \mathscr{N}(i)} U_{ij}^S(s_i, s_j; \eta_S)$ and then, $\tilde{H}_{\mathbf{T},\mathbf{S}}(\mathbf{t}, \mathbf{s}) = \sum_{i \in V} < t_i, e_{T^{s_i}} >$. Simple examples for $U_{ij}^T(t_i, t_j; \eta_T)$ and $U_{ij}^S(s_i, s_j; \eta_S)$ are provided by adopting a Potts model which corresponds to

$$U_{ij}^T(t_i, t_j; \eta_T) = \eta_T < t_i, t_j > \quad \text{and} \quad U_{ij}^S(s_i, s_j; \eta_S) = \eta_S < s_i, s_j > \tag{21}$$

The first two terms $H_{\mathbf{T}}(\mathbf{t})$ and $H_{\mathbf{S}}(\mathbf{s})$ capture, within each label set $\mathbf{t}$ and $\mathbf{s}$, interactions between neighboring voxels. They imply spatial interaction within each label set. Interaction between label sets is captured in the third term but in the definition above this interaction is not spatial since only singleton terms, involving one voxel at a time, appear in the sum. The expression for $\tilde{H}_{\mathbf{T},\mathbf{S}}(\mathbf{t},\mathbf{s})$ could be augmented with a *line process* term [13] to account for between label sets spatial interactions. An example of what would be the resulting energy functions is given in the Appendix.

The following terms define $H_{\mathbf{Z},\Theta}$ in (14). They are specified to integrate our *a priori* knowledge and to account for the fact that the registration parameters are estimated along the segmentation process. We set, $H_{\mathbf{T},\Theta}(\mathbf{t},\theta) = \sum_{i \in V} < t_i, \log(f_T(R,i)+1) >$ and similarly $H_{\mathbf{S},\Theta}(\mathbf{s},\theta) = \sum_{i \in V} < s_i, \log(f_S(R,i)+1) >$. The logarithm is added because $f_T(R,i)$ and $f_S(R,i)$, as defined in Section 5.1, are probability distributions whereas an energy H is homogeneous to the logarithm of a probability up to a constant. An additional 1 is added inside the logarithm to overcome the problem of its non existence at 0. The overall method does not seem to be sensitive to the exact value of the positive quantity added. It follows that at this stage the dependence on θ is only through the registration parameter R. The dependence on ψ has been already specified in the data term and no additional dependence exists in our model. In addition, as regards interactions between labels and parameters, we consider that they exist only separately for $\mathbf{t}$ and $\mathbf{s}$ so that we set $\tilde{H}_{\mathbf{T},\mathbf{S},\Theta}(\mathbf{t},\mathbf{s},\theta) = 0$.

Pre-registering the atlas beforehand is equivalent not to estimate the registration parameters but to fix them in a pre-processing step, say to R^0. Then our model is modified by setting $H_{\mathbf{Z},\Theta}(\mathbf{z},\theta) = 0$ and by adding singleton terms in $H_{\mathbf{T}}(\mathbf{t})$ and $H_{\mathbf{S}}(\mathbf{s})$ to account for pre-registered atlas anatomical information. The terms to be added would be respectively $\sum_{i \in V} < t_i, \log(f_T(R^0,i)+1) >$ and $\sum_{i \in V} < s_i, \log(f_S(R^0,i)+1) >$.

5.2.3. Parameter prior term

The last term in (12) to be specified is $H_\Theta(\theta)$. The Gaussian distribution parameters and the registration parameters are supposed to be independent and we set $H_\Theta(\theta) = H_\Psi(\psi) + H_R(R)$. The specific form of $H_\Psi(\psi)$ will be specified later. It will actually be guided by our inference procedure (see Section 5.3.2). In practice however, in the general setting of Section 5.1 which allows different values ψ_i at each i, there are too many parameters and estimating them accurately is not possible. As regards estimation then, we adopt a local approach as in [25]. The idea is to consider the parameters as constant over subvolumes of the entire volume. Let $\mathscr{C}$ be a regular cubic partionning of the volume V in a number of non-overlapping subvolumes $\{V_c, c \in \mathscr{C}\}$. We assume that for all $c \in \mathscr{C}$ and all $i \in V_c$, $\psi_i = \psi_c$ and consider an energy function on $\mathscr{C}$ denoted by $H^{\mathscr{C}}_\Psi(\psi)$ where by extension ψ now denotes the set of distinct values $\psi = \{\psi_c, c \in \mathscr{C}\}$. Outside the issue of estimating ψ in the M-step, having parameters ψ_i's depending on i is not a problem. As specified in Section 5.3.1 for the E-steps we will go back to this general setting using an interpolation step specified in Section 5.3.2. As regards $H_R(R)$, it could be specified as in [23] to favor estimation of R close to some average registration parameters computed from a training data set if available. In our case, no such data set is available and we will simply set $H_R(R) = 0$.

5.3. *Generalized Alternating Maximization*

We now derive the inference steps of Section 4.1 for the model defined above.

5.3.1. **Structure and tissue conditional E-steps**

The two-stage E-step given by (17) and (18) can be further specified by computing $H_T^{(r)}(\mathbf{t}) = E_{q_S^{(r-1)}}[H(\mathbf{t}|\mathbf{S},\mathbf{y},\theta^{(r)})]$ and $H_S^{(r)}(\mathbf{s}) = E_{q_T^{(r)}}[H(\mathbf{s}|\mathbf{T},\mathbf{y},\theta^{(r)})]$. For the expressions given in Section 5.2, it comes, omitting the terms that do not depend on $\mathbf{t}$,

$$\begin{aligned} H_T^{(r)}(\mathbf{t}) &= \sum_{i\in V}\Big(\sum_{j\in\mathcal{N}(i)} U_{ij}^T(t_i,t_j;\eta_T) + < t_i, \log(f_T(R^{(r)},i)+1) > + \\ & < t_i, \sum_{l=1}^{L} q_{S_i}^{(r-1)}(e'_l)\, e_{T^l} > + \left(\frac{1+q_{S_i}^{(r-1)}(e'_{L+1})}{2}\right) \log \mathcal{G}(y_i| < t_i, \psi_i^{(r)} >)\Big)\,, \quad (22) \end{aligned}$$

where $\sum_{l=1}^{L} q_{S_i}^{(r-1)}(e'_l) e_{T^l}$ is a 3-component vector whose k^{th} component is $\sum_{l\ st.T^l=k} q_{S_i}^{(r-1)}(e'_l)$ that is the probability, as given by the current distribution $q_{S_i}^{(r-1)}$, that voxel i is in a structure whose tissue is k. The higher this probability the more favored is tissue k. If we modified then the expression of f_T into $\tilde{f}_T^{(r)}$ defined by $\tilde{f}_T^{(r)} = {}^t(\tilde{f}_T^{k(r)}, k=1,\dots,K)$ with $\log \tilde{f}_T^{k(r)}(R,i) = \log(f_T^k(R,i) + 1) + \sum_{l\ st.T^l=k} q_{S_i}^{(r-1)}(e'_l)$, (22) can be written as

$$\begin{aligned} H_T^{(r)}(\mathbf{t}) &= \sum_{i\in V}\left(\sum_{j\in\mathcal{N}(i)} U_{ij}^T(t_i,t_j;\eta_T) + < t_i, \log(\tilde{f}_T^{(r)}(R^{(r)},i)) > + \right.\\ & \left. \log\left(\mathcal{G}(y_i| < t_i, \psi_i^{(r)} >)^{\frac{1+q_{S_i}^{(r-1)}(e'_{L+1})}{2}}\right)\right) \qquad (23) \end{aligned}$$

Then, for the E-S-step (18) we can derive a similar expression. Note that for $s_i \in \{1,\dots,L\}$, $q_{T_i}(e_{T^{s_i}}) =< s_i, \sum_{l=1}^{L} q_{T_i}(e_{T^l})\, e'_l >$ so that if we modify the expression of f_S into $\tilde{f}_S^{(r)}$ defined by $\tilde{f}_S^{(r)} = {}^t(\tilde{f}_S^{l(r)}, l=1\dots L+1)$ with $\log \tilde{f}_S^{l(r)}(R,i) = \log(f_S^l(R,i)+1) + q_{T_i}^{(r+1)}(e_{T^l})(1- < e'_l, e'_{L+1} >)$ we get,

$$\begin{aligned} H_S^{(r)}(\mathbf{s}) &= \sum_{i\in V}\sum_{j\in\mathcal{N}(i)} U_{ij}^S(s_i,s_j;\eta_S) + < s_i, \log(\tilde{f}_S^{(r)}(R^{(r)},i)) > + \qquad (24)\\ & \log\left(\left(\prod_{k=1}^{3} \mathcal{G}(y_i|\psi_i^{(r)k})^{q_{T_i}^{(r)}(e_k)}\right)^{\left(\frac{1+<s_i,e'_{L+1}>}{2}\right)} \mathcal{G}(y_i| < e_{T^{s_i}}, \psi_i^{(r)} >)^{\left(\frac{1-<s_i,e'_{L+1}>}{2}\right)}\right). \end{aligned}$$

In the simplified expressions (23) and (24), we can recognize the standard decomposition of hidden Markov random field models into three terms, from left to right, a regularizing spatial term, an external field or singleton term and a data term. This shows that at each iteration of our cooperative algorithm, solving the current E-T and E-S steps is equivalent to solving the segmentation task for standard hidden Markov models whose definition depends on the results of the previous iteration. We are not giving further details here but in our application we will use mean field like algorithms as described in [9] to actually compute $q_T^{(r)}$ and $q_S^{(r)}$.

5.3.2. M-step: Updating the parameters

We now turn to the resolution of step (20), $\theta^{(r+1)} = \arg\max_{\theta \in \underline{\Theta}} E_{q_T^{(r)} q_S^{(r)}}[H(\theta|\mathbf{y},\mathbf{T},\mathbf{S})]$. The independence of ψ and R leads to an M-step that separates into two updating stages:

$$\psi^{(r+1)} = \arg\max_{\psi \in \underline{\Psi}} H_\Psi(\psi) + \sum_{i\in V} E_{q_{T_i}^{(r)} q_{S_i}^{(r)}}[\log g(y_i|T_i,S_i,\psi_i)] \tag{25}$$

$$\text{and} \quad R^{(r+1)} = \arg\max_{R \in \underline{\mathscr{R}}} H_R(R) + E_{q_T^{(r)}}[H_{T,\Theta}(\mathbf{T},\theta)] + E_{q_S^{(r)}}[H_{S,\Theta}(\mathbf{S},\theta)] \; . \tag{26}$$

Updating the intensity distributions parameters. We first focus on the computation of the last sum in (25). Omitting, the (r) superscript, after some straightforward algebra, it comes

$$E_{q_{T_i} q_{S_i}}[\log g(y_i|T_i,S_i,\psi_i)] = \log \left(\prod_{k=1}^K \mathscr{G}(y_i|\psi_i^k)^{a_{ik}} \right) ,$$

where $a_{ik} = \frac{1}{2}\left(q_{T_i}(e_k) + q_{T_i}(e_k) q_{S_i}(e'_{L+1}) + \sum_{l \; st. T^l = k} q_{S_i}(e'_l)\right)$.

The first term in a_{ik} is the probability for voxel i to belong to tissue k without any additional knowledge on structures. The sum over k of the two other terms is one and they can be interpreted as the probability for voxel i to belong to the tissue class k when information on structure segmentation is available. In particular, the third term in a_{ik} is the probability that voxel i belongs to a structure made of tissue k while the second term is the probability to be in tissue k when no structure is present at voxel i. Then the sum of the a_{ik}'s is also one and a_{ik} can be interpreted as the probability for voxel i to belong to the tissue class k when both tissue and structure segmentations information are combined.

As mentioned in Section 5.2.3, we will now consider that the ψ_i's are constant over subvolumes of a given partition of the entire volume so that, denoting by $p(\psi)$ the MRF prior on $\psi = \{\psi_c, c \in \mathscr{C}\}$, *ie.* $p(\psi) \propto \exp(H_\Psi^{\mathscr{C}}(\psi))$, (25) can be written as,

$$\psi^{(r+1)} = \arg\max_{\psi \in \underline{\Psi}} p(\psi) \prod_{i\in V} \prod_{k=1}^K \mathscr{G}(y_i|\psi_i^k)^{a_{ik}} \; = \; \arg\max_{\psi \in \underline{\Psi}} p(\psi) \prod_{c\in\mathscr{C}} \prod_{k=1}^K \prod_{i\in V_c} \mathscr{G}(y_i|\psi_c^k)^{a_{ik}} \; .$$

Using the additional natural assumption that $p(\psi) = \prod_{k=1}^K p(\psi^k)$, it is equivalent to solve for each $k = 1,\ldots,K$,

$$\psi^{k\,(r+1)} = \arg\max_{\psi^k \in \underline{\Psi}^k} p(\psi^k) \prod_{c\in\mathscr{C}} \prod_{i\in V_c} \mathscr{G}(y_i|\psi_c^k)^{a_{ik}}. \tag{27}$$

However, when $p(\psi^k)$ is chosen as a Markov field on $\mathscr{C}$, the maximization is still intractable. We therefore replace $p(\psi^k)$ by a product form given by its *modal-field* approximation [9]. This is actually equivalent to use the ICM [5] algorithm to maximize (27). Assuming a current estimation of ψ^k at iteration ν, we consider in turn the following updating,

$$\forall c \in \mathscr{C}, \quad \psi_c^{k\,(\nu+1)} \;=\; \arg\max_{\psi_c^k \in \underline{\Psi}^k} p(\psi_c^k \mid \psi_{\mathscr{N}(c)}^{k\,(\nu)}) \prod_{i\in V_c} \mathscr{G}(y_i|\psi_c^k)^{a_{ik}} \, , \tag{28}$$

where $\mathscr{N}(c)$ denotes the indices of the subvolumes that are neighbors of subvolume c and $\psi^k_{\mathscr{N}(c)} = \{\psi^k_{c'}, c' \in \mathscr{N}(c)\}$. At convergence, the obtained values give the updated estimation $\psi^{k\,(r+1)}$. The particular form (28) above somewhat dictates the specification of the prior for ψ. Indeed Bayesian analysis indicates that a natural choice for $p(\psi^k_c \mid \psi^k_{\mathscr{N}(c)})$ has to be among conjugate or semi-conjugate priors for the Gaussian distribution $\mathscr{G}(y_i|\psi^k_c)$ [12]. We choose to consider here the latter case. In addition, we assume that the Markovian dependence applies only to the mean parameters and consider that $p(\psi^k_c \mid \psi^k_{\mathscr{N}(c)}) = p(\mu^k_c \mid \mu^k_{\mathscr{N}(c)})\, p(\lambda^k_c)$ with $p(\mu^k_c \mid \mu^k_{\mathscr{N}(c)})$ set to a Gaussian distribution with mean $m^k_c + \sum_{c' \in \mathscr{N}(c)} \eta^k_{cc'}(\mu^k_{c'} - m^k_{c'})$ and precision λ^{0k}_c, and $p(\lambda^k_c)$ set to a Gamma distribution with shape parameter α^k_c and scale parameter b^k_c. The quantities $\{m^k_c, \lambda^{0k}_c, \alpha^k_c, b^k_c, c \in \mathscr{C}\}$ and $\{\eta^k_{cc'}, c' \in \mathscr{N}(c)\}$ are hyperparameters to be specified. For this choice, we get valid joint Markov models for the μ^k's (and therefore for the ψ^k's) which are known as auto-normal models [4]. Whereas for the standard Normal-Gamma conjugate prior the resulting conditional densities fail in defining a proper joint model and caution must be exercised.

Standard Bayesian computations lead to a decomposition of (28) into two maximizations: for μ^c_k, the product in (28) has a Gaussian form and the mode is given by its mean. For λ^c_k, the product turns into a Gamma distribution and its mode is given by the ratio of its shape parameter over its scale parameter. After some straightforward algebra, we get the following updating formulas:

$$\mu^{(v+1)\,k}_c = \frac{\lambda^{(v)\,k}_c \sum_{i \in V_c} a_{ik} y_i + \lambda^{0k}_c \left(m^k_c + \sum_{c' \in \mathscr{N}(c)} \eta^k_{cc'} (\mu^{(v)\,k}_{c'} - m^k_{c'})\right)}{\lambda^{(v)\,k}_c \sum_{i \in V_c} a_{ik} + \lambda^{0k}_c} \tag{29}$$

$$\text{and} \quad \lambda^{(v+1)\,k}_c = \frac{\alpha^k_c + \sum_{i \in V_c} a_{ik}/2 - 1}{b^k_c + 1/2[\sum_{i \in V_c} a_{ik}(y_i - \mu^{(v+1)\,k}_c)^2]}\,. \tag{30}$$

In these equations, quantities similar to the ones computed in standard EM for the mean and variance parameters appear weighted with other terms due to neighbors information. Namely, standard EM on voxels of V_c would estimate μ^k_c as $\sum_{i \in V_c} a_{ik} y_i / \sum_{i \in V_c} a_{ik}$ and λ^k_c as $\sum_{i \in V_c} a_{ik} / \sum_{i \in V_c} a_{ik}(y_i - \mu^k_c)^2$. In that sense formulas (29) and (30) intrinsically encode cooperation between local models.

From these parameters values constant over subvolumes we compute parameter values per voxel by using cubic splines interpolation between θ_c and $\theta_{c'}$ for all $c' \in \mathscr{N}(c)$. We go back this way to our general setting which has the advantage to ensure smooth variation between neighboring subvolumes and to intrinsically handle nonuniformity of intensity inside each subvolume.

Updating the registration parameters. From (26), it follows that

$$\begin{aligned} R^{(r+1)} \;=\; & \arg\max_{R \in \mathscr{R}} H_R(R) + \sum_{i \in V} \sum_{k=1}^{3} q^{(r)}_{T_i}(e_k) \log(f^k_T(R,i) + 1) \\ & + \sum_{i \in V} \sum_{l=1}^{L+1} q^{(r)}_{S_i}(e'_l) \log(f^l_S(R,i) + 1) \end{aligned} \tag{31}$$

which further simplifies when $H_R(R) = 0$. It appears that the registration parameters are refined using information on structures as in [23] but also using information on tissues through the second term above. In practice, the optimization is carried out using a relaxation approach decomposing the maximization into searches for the different structure specific deformations $\{R_l, l = 1 \ldots L+1\}$.

There exists no simple expression and the optimization is performed numerically using a variant of the Powell algorithm [24]. We therefore update the 12 parameters defining each local affine transformation R_l by maximizing in turn:

$$R_l^{(r+1)} = \arg\max_{R_l \in \mathscr{R}_l} H_R(R) + \sum_{i \in V} \sum_{k=1}^{3} q_{T_i}^{(r)}(e_k) \log(f_T^k(R,i)+1) + \sum_{i \in V} \sum_{l=1}^{L+1} q_{S_i}^{(r)}(e'_l) \log(f_S^l(R,i)+1) \,. \quad (32)$$

5.4. Results

Regarding hyperparameters, we choose not to estimate the parameters η_T and η_S but fix them to the inverse of a decreasing temperature as proposed in [5]. In expressions (29) and (30), we wrote a general case but it is natural and common to simplify the derivations by setting the m_c^k's to zero and $\eta_{cc'}^k$ to $|\mathscr{N}(c)|^{-1}$ where $|\mathscr{N}(c)|$ is the number of subvolumes in $\mathscr{N}(c)$. This means that the distribution $p(\mu_c^k|\mu_{\mathscr{N}(c)}^k)$ is a Gaussian centered at $\sum_{c' \in \mathscr{N}(c)} \mu_{c'}^k / |\mathscr{N}(c)|$ and therefore that all neighbors c' of c act with the same weight. The precision parameters λ_c^{0k} is set to $N_c \lambda_g^k$ where λ_g^k is a rough precision estimation for class k obtained for instance using some standard EM algorithm run globally on the entire volume and N_c is the number of voxels in c that accounts for the effect of the sample size on precisions. The α_c^k's are set to $|\mathscr{N}(c)|$ and b_c^k to $|\mathscr{N}(c)|/\lambda_g^k$ so that the mean of the corresponding Gamma distribution is λ_g^k and the shape parameter α_c^k somewhat accounts for the contribution of the $|\mathscr{N}(c)|$ neighbors. Then, the size of subvolumes is set to $20 \times 20 \times 20$ voxels. The subvolume size is a mildly sensitive parameter. In practice, subvolume sizes from $20 \times 20 \times 20$ to $30 \times 30 \times 30$ give similar good results on high resolution images (1 mm^3). On low resolution images, a size of $25 \times 25 \times 25$ may be preferred.

Evaluation is then performed following the two main aspects of our model. The first aspect is the decomposition of the global clustering task into a set of local clustering tasks using local MRF models. The advantage of our approach is that, in addition, a way to ensure consistency between all these local models is dictated by the model itself. The second aspect is the cooperative setting which is relevant when two global clustering tasks are considered simultaneously. It follows that we first assess the performance of our model considering the local aspect only. We compare (Section 5.4.2) the results obtained with our method, restricted to tissue segmentation only, with other recent or state-of-the-art methods for tissue segmentation. We then illustrate more of the modelling ability of our approach by showing results for the joint tissue and structure segmentation (Section 5.4.3).

5.4.1. **Data**

We consider both phantoms and real 3T brain scans. We use the normal 1 mm^3 BrainWeb phantoms database from the McConnell Brain Imaging Center [10]. These phantoms are generated from a realistic brain anatomical model and a MRI simulator that simulates MR acquisition physics, in which different values of nonuniformity and noise can be added. Because these images are simulated we can quantitatively compare our tissue segmentation to the underlying tissue generative model to evaluate the segmentation performance. As in [2, 27, 29] we perform a quantitative evaluation using the Dice similarity metric [11]. This metric measures the overlap between a segmentation result and the gold standard. By denoting by TP_k the number of true positives for class k, FP_k the number of false positives and FN_k the number of false negatives the Dice metric is given by: $d_k = \frac{2\mathrm{TP}_k}{2\mathrm{TP}_k+\mathrm{FN}_k+\mathrm{FP}_k}$ and d_k takes its value in $[0,1]$ where 1 represents the perfect agreement. Since BrainWeb phantoms contain only tissue information, three subcortical structures were manually segmented by three experts: the left caudate nucleus, the left putamen and the left thalamus. We then computed our structure gold standard using STAPLE [30], which computes a probabilistic estimate of a true segmentation from a set of different manual segmentations. The results we report are for eight BrainWeb phantoms, for 3%, 5%, 7% and 9% of noise with 20% and 40% of nonuniformity for each noise level. Regarding real data, we then evaluate our method on real 3T MR brain scans (T1 weighted sequence, TR/TE/Flip = $12ms/4.6ms/8^\circ$, Recovery Time=2500ms, Acquisition Matrix=$256 \times 256 \times 176$, voxel isotropic resolution 1 mm^3) coming from the Grenoble Institut of Neuroscience (GIN).

5.4.2. A local method for segmenting tissues

Considering tissue segmentation only, we quantitatively compare our method denoted by LOCUS$^\mathrm{B}$-T to the recently proposed method LOCUS-T [25] and to two well known tissue segmentation tools, FAST [31] from FSL and SPM5 [2]. The table in Figure 3 (a) shows the results of the evaluation performed on the eight BrainWeb phantoms. The mean Dice metric over all eight experiments and for all tissues is 86% for SPM5, 88% for FAST and 89% for LOCUS-T and LOCUS$^\mathrm{B}$-T. The mean computation times for the full 3-D segmentation were 4min for LOCUS-T and LOCUS$^\mathrm{B}$-T, 8min for FAST and more than 10min for SPM5. Figure 3 (b) to (f) shows the results on a real image.

Our method shows very satisfying robustness to noise and intensity nonuniformity. On BrainWeb images, it performs better than SPM5 and similarly than LOCUS-T and FAST, for a low computational time. On real 3T scans, LOCUS-T and SPM5 also give in general satisfying results.

5.4.3. Joint tissue and structure segmentation

We then evaluate the performance of the joint tissue and structure segmentation. We consider two cases: our combined approach with fixed registration parameters (LOCUS$^\mathrm{B}$-TS) and with estimated registration parameters (LOCUS$^\mathrm{B}$-TSR). For the joint tissue and structure model (LOCUS$^\mathrm{B}$-TS) we introduce *a priori* knowledge based on the Harvard-Oxford subcortical probabilistic atlas. FLIRT was used to affine-register the atlas. For LOCUS$^\mathrm{B}$-TSR, the global registration parameters R_G are computed as in LOCUS$^\mathrm{B}$-TS as a pre-processing step. The other local registration parameters are updated at each iteration of the algorithm. Table 1 shows the evaluation on BrainWeb images using our reference segmentation of three structures. The table shows the means and standard deviations of the Dice coefficient values obtained for the eight BrainWeb images. It also shows the means and standard deviations of the relative improvements between the two models LOCUS$^\mathrm{B}$-TS and LOCUS$^\mathrm{B}$-TSR. In particular, a significant improvement of 23% is observed for the caudate nucleus. For LOCUS$^\mathrm{B}$-TSR, the mean computational time is of 10min for our three structures (45min for 17 structures) including the initial global registration

(R_G) step using FLIRT. For comparison, on one of the brainweb phantoms, the 5% noise, 40% nonuniformity image, Freesurfer leads respectively to 88%, 86%, 90%, with a computational time larger than 20 hours for 37 structures, while the results with LOCUS$^{\mathrm{B}}$-TSR on this phantom were 91%, 95% and 94%.

The BrainWeb database evaluation shows that the segmentation quality is very stable when the noise and inhomogeneity levels vary and this is one of the major difference with the algorithm in [25]. The three structures segmentations improve when registration is combined. In particular, in LOCUS$^{\mathrm{B}}$-TS the initial global registration of the caudate is largely sub-optimal but it is then corrected in LOCUS$^{\mathrm{B}}$-TSR. More generally, for the three structures we observe a stable gain for all noise and inhomogeneity levels.

Figure 4 shows the results obtained with LOCUS$^{\mathrm{B}}$-T, and LOCUS$^{\mathrm{B}}$-TSR on a real 3T brain scan. The structures emphasized in image (c) are the two lateral ventriculars (blue), the caudate nuclei (red) , the putamens (green) and the thalamus (yellow). Figure 4 (e) shows in addition a 3D reconstruction of 17 structures segmented with LOCUS$^{\mathrm{B}}$-TSR. The results with LOCUS$^{\mathrm{B}}$-TS are not shown because the differences with LOCUS$^{\mathrm{B}}$-TSR were not visible using this paper graphical resolution.

We observe therefore the gain in combining tissue and structure segmentation in particular through the improvement of tissue segmentation for areas corresponding to structures such as the putamens and thalamus. The additional integration of a registration parameter estimation step also provides some significant improvement. It allows an adaptive correction of the initial global registration parameters and a better registration of the atlas locally. These results could be however certainly further improved if *a priori* knowledge (through $H(\mathscr{R})$) on the typical deformations for each structure was used to guide these local deformations more precisely.

	CSF	GM	WM	M.C.T.
LOCUS$^{\mathrm{B}}$-T	80 % (2)	92% (1)	94% (1)	≈ 4min
LOCUS-T	80% (2)	92% (1)	94% (1)	≈ 4min
SPM5	79% (3)	89% (4)	90% (3)	≈ 12min
FAST	80% (1)	91% (1)	94% (1)	≈ 8min

(a)

(b) (c) (d) (e) (f)

FIGURE 3. *Tissue segmentation only. Table (a): mean Dice metric and mean computational time (M.C.T) values on BrainWeb over 8 experiments for different values of noise (3%, 5%, 7%, 9%) and nonuniformity (20%, 40%). The corresponding standard deviations are shown in parenthesis. Images (c) to (f): segmentations respectively by LOCUS^B-T (our approach), LOCUS-T, SPM5 and FAST of a highly nonuniform real 3T image (b). The circle in (d) points out a segmentation error which does not appear in (c).*

Structure	LOCUS^B-TS	LOCUS^B-TSR	Relative Improvement
Left Thalamus	91% (0)	94% (1)	4% (1)
Left Putamen	90% (1)	95% (0)	6% (1)
Left Caudate	74% (0)	91% (1)	23% (1)

TABLE 1. *Mean Dice coefficient values obtained on three structures using LOCUS^B-TS and LOCUS^B-TSR for BrainWeb images, over 8 experiments for different values of noise (3%, 5%, 7%, 9%) and nonuniformity (20%, 40%). The corresponding standard deviations are shown in parenthesis. The second column shows the results when registration is done as a pre-processing step (LOCUS^B-TS). The third columns shows the results with our full model including iterative estimation of the registration parameters (LOCUS^B-TSR). The last column shows the relative Dice coefficient improvement for each structure.*

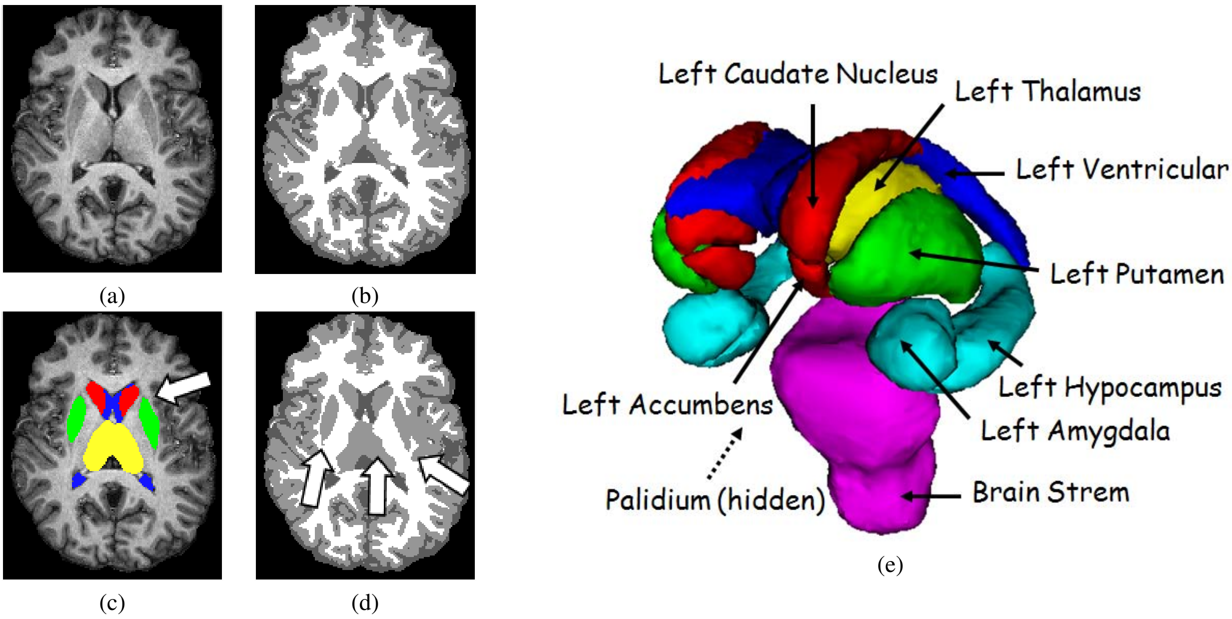

FIGURE 4. *Evaluation of LOCUS^B-TSR on a real 3T brain scan (a). For comparison the tissue segmentation obtained with LOCUS^B-T is given in (b). The results obtained with LOCUS^B-TSR are shown in the second line. Major differences between tissue segmentations (images (b) and (d)) are pointed out using arrows. Image (e) shows the corresponding 3D reconstruction of 17 structures segmented using LOCUS^B-TSR. The names of the left structures (use symmetry for the right structures) are indicated in the image.*

6. Discussion

The strength of our Bayesian joint model comes from its specification via a coherently linked system of conditional models. The whole consistent treatment and tractability of the resulting coupled clustering tasks is made possible using Generalized Alternating Minimization procedures that generalize the standard EM framework. It follows an approach made of steps that are easy to interpret and that could be enriched with additional information. These modelling abilities are illustrated on a challenging real data issue of segmenting both tissues and structures from MRI brain scans. The results obtained with our cooperative clustering approach are very satisfying and compare favorably with other existing methods. The possibility to add a conditional MRF model for the intensity distribution parameters allows to handle local estimations for robustness to nonuniformities. However, further possible investigations relate to the interpolation step that we add to increase robustness to nonuniformities at a voxel level. We believe this stage could be generalized and incorporated in the model by considering successively various degrees of locality, mimicking a multiresolution approach and refining from coarse partitions of the entire volume to finer ones. Also, our choice of prior for the intensity distribution parameters was guided by the need to define appropriate conditional specifications $p(\psi_c^k|\psi_{\mathscr{N}(c)}^{k(v)})$ in (28) that lead to a valid Markov model for the ψ^k's. Nevertheless, incompatible conditional specifications can still be used for inference, *eg.* in a Gibbs sampler or ICM algorithm with some valid justification (see [16] or the discussion in [1]). In applications, one may found that having a joint distribution is less important than incorporating information from other variables such as typical interactions. In that sense, conditional modeling allows interesting flexibility in dealing with practical problems. However, it is not clear when incompatibility of conditional distributions is an issue in practice and the theoretical properties of the procedures in this case are largely unknown and should be investigated. The tissue and structure models are also conditional MRF's that are linked and capture several level of interactions. They incorporate 1) spatial dependencies between voxels for robustness to noise, 2) relationships between tissue and structure labels for cooperative aspects and 3) *a priori* anatomical information (atlas). In most approaches, atlas registration is performed globally on the entire brain resulting in structure segmentation performance that depends crucially on the accuracy of this global registration step. Our method has the advantage of providing a way to incorporate atlas registration and to refine it locally.

More generally, the framework we propose can be adapted to other applications. It provides a strategy and guidelines to deal with complex joint processes involving more than one identified sub-processes. It is based on the idea that defining conditional models is usually more straightforward and captures more explicitly cooperative aspects, including cooperation with external knowledge. The Bayesian formulation provides additional flexibility such as the possibility to deal, in a well based manner, with some sort of non-stationarity in the parameters (as the one due to intensity nonuniformities in our MRI example). Of course, depending on the application in mind, more complex energy functions than the one given in our MRI illustration may be necessary. In particular, for our example it was enough to consider separately cooperation between label sets and spatial interactions. However, one useful extension to be investigated in future work, would be to add a spatial component in the cooperation mechanisms themselves. We describe in the Appendix a possible way to perform that.

Appendix A: Spatial cooperative interactions

In segmentation tasks, spatial regularization between neighboring labels is usually desirable and encoded via a smoothness or regularization term. We mentioned in Section 5.2.2 the use of a Potts model term (21) with a regularization parameter that we simply write here η (previously η_T or η_S). This parameter is positive when regularization is desired. A negative value of η would have the opposite effect by favoring neighboring labels to be different. Images to be segmented are however usually only locally smooth and contain discontinuities for which regularization is not necessarily appropriate. The modelling of such spatial discontinuities has been intensely studied and in particular via *Line process* models (see for instance [6, 13]). The idea we want to use here is similar to that of *Line processes*. It consists of keeping the regularizing term as the original one when no discontinuity is present while discarding the regularization term when a discontinuity is introduced or detected. The particularity in our cooperative approach is that the presence of discontinuities in one of the label sets (say **t**) is detected by considering the labels values of the other label set (say **s**) and reciprocally. As regards tissues and structures in brain MRI analysis, the key-point is then to determine for which instances of structure (resp. tissue) segmentations, neighboring tissue labels (resp. structure labels) should not be regularized.

Let i and j be two neighboring voxels. When s_i and s_j are in $\{e'_1,\ldots,e'_L\}$ and $T^{s_i} \neq T^{s_j}$, then a compatible tissue segmentation should be so that $t_i \neq t_j$. Conversely, if $t_i \neq t_j$ then $s_i \neq s_j$ should be favored and no regularization applied between i and j when segmenting structures. More than that, in both these cases, we may want to enforce repulsive interactions. Regarding the feedback of tissues on structures, note that $t_i = t_j$ does not necessarily imply $s_i = s_j$ but only $T^{s_i} = T^{s_j}$. However it is not clear whether the later condition provides enough regularization in **s**.

We first focus on structure segmentation, considering that a current tissue segmentation is available. Let $U^{S+}_{ij}(s_i,s_j,\eta_S)$ denote a regularizing local energy term between voxel i and j. Let $U^{S-}_{ij}(s_i,s_j,\eta_S)$ be the corresponding repulsive energy term. For instance if $U^{S+}_{ij}(s_i,s_j,\eta_S) = \eta_S < s_i,s_j >$, the corresponding repulsive term can simply be $U^{S-}_{ij}(s_i,s_j,\eta_S) = -\eta_S < s_i,s_j >$. Note that this is not the same as taking $U^{S-}_{ij} = 0$ as an alternative to regularization. Then, tissue discontinuities can be defined by a binary function $\delta^T_{reg}(t_i,t_j) =< t_i,t_j >$. A tissue discontinuity corresponds to $\delta^T_{reg}(t_i,t_j) = 0$ while no discontinuity corresponds to $\delta^T_{reg}(t_i,t_j) = 1$. A tissue discontinuity implies that $s_i \neq s_j$ and then an appropriate interaction term would be the repulsive energy $U^{S-}_{ij}(s_i,s_j,\eta_S)$. On the other hand, $t_i = t_j$ does not necessarily imply that $s_i = s_j$ but only $T^{s_i} = T^{s_j}$ so that an appropriate interaction term would be $U^{S+}_{ij}(e_{T^{s_i}},e_{T^{s_j}},\eta_S)$. It follows that to account for the effect of **t** on **s**, an appropriate energy term would be

$$U^S_{ij}(s_i,s_j,t_i,t_j,\eta_S) = \delta^T_{reg}(t_i,t_j) U^{S+}_{ij}(e_{T^{s_i}},e_{T^{s_j}},\eta_S) + (1-\delta^T_{reg}(t_i,t_j)) U^{S-}_{ij}(s_i,s_j,\eta_S) \ ,$$

which in the Potts model case simplifies into

$$U^S_{ij}(s_i,s_j,t_i,t_j,\eta_S) = \delta^T_{reg}(t_i,t_j)\ \eta_S < e_{T^{s_i}},e_{T^{s_j}} > -(1-\delta^T_{reg}(t_i,t_j))\ \eta_S < s_i,s_j > \ .$$

Regarding the effect of structures on tissues, the situation is slightly different since a discontinuity $s_i \neq s_j$ does not provide information on t_i and t_j in the sense that both $t_i = t_j$ and $t_i \neq t_j$ are compatible solutions in this case. A more informative choice is to consider T^{s_i} and T^{s_j} and to look for discontinuities in $T^{\mathbf{s}}$. Let us assume that s_i and s_j are not in the background (*ie.* both different from e'_{L+1}). In this case, $T^{s_i} = T^{s_j}$ implies $t_i = t_j$ and a regularizing term can be $U_{ij}^{T+}(t_i, t_j, \eta_T)$. If then $s_i = e'_{L+1}$ or $s_j = e'_{L+1}$, it means that no information is available on t_i and t_j and we may also choose to regularize using $U_{ij}^{T+}(t_i, t_j, \eta_T)$. It is more convenient then to define $\delta_{reg}^S(s_i, s_j)$ as

$$\begin{aligned}\delta_{reg}^S(s_i, s_j) &= < e_{T^{s_i}}, e_{T^{s_j}} > + \\ &\quad (1 - < e_{T^{s_i}}, e_{T^{s_j}} >)(< s_i, e'_{L+1} > + < s_j, e'_{L+1} > - < s_i, e'_{L+1} >< s_j, e'_{L+1} >) \,,\end{aligned}$$

which is one when tissues can be regularized and 0 otherwise. Indeed, the first term $< e_{T^{s_i}}, e_{T^{s_j}} >$ is one when s_i and s_j are structures made of the same tissue. When this is not the case, $\delta_{reg}^S(s_i, s_j)$ is equal to the last term $(< s_i, e'_{L+1} > + < s_j, e'_{L+1} > - < s_i, e'_{L+1} >< s_j, e'_{L+1} >)$ which is one if and only if at least one of the voxels i or j does not belong to any structure. It follows then in the energy, the spatial interaction term below

$$U_{ij}^T(t_i, t_j, s_i, s_j, \eta_T) = \delta_{reg}^S(s_i, s_j) U_{ij}^{T+}(t_i, t_j, \eta_T) + (1 - \delta_{reg}^S(s_i, s_j)) U_{ij}^{T-}(t_i, t_j, \eta_T) \,,$$

which for the Potts like term above leads to
$U_{ij}^T(t_i, t_j, s_i, s_j, \eta_T) = (2\delta_{reg}^S(s_i, s_j) - 1)\, \eta_T < t_i, t_j >$.

Eventually, one way to encode all that using the energy decomposition of Section 5.2.2, is to set, $H_T(\mathbf{t}) = \sum_{i \in V} \sum_{j \in \mathcal{N}(i)} U_{ij}^{T-}(t_i, t_j, \eta_T)$, $H_S(\mathbf{s}) = \sum_{i \in V} \sum_{j \in \mathcal{N}(i)} U_{ij}^{S-}(s_i, s_j, \eta_S)$ and

$$\begin{aligned}\tilde{H}_{T,S}(\mathbf{t}, \mathbf{s}) &= \sum_{i \in V} < t_i, e_{T^{s_i}} > \\ &\quad + \sum_{i \in V} \sum_{j \in \mathcal{N}(i)} \delta_{reg}^S(s_i, s_j)\, (U_{ij}^{T+}(t_i, t_j, \eta_T) - U_{ij}^{T-}(t_i, t_j, \eta_T)) \\ &\quad + \sum_{i \in V} \sum_{j \in \mathcal{N}(i)} \delta_{reg}^T(t_i, t_j)\, (U_{ij}^{S+}(e_{T^{s_i}}, e_{T^{s_j}}, \eta_S) - U_{ij}^{S-}(s_i, s_j, \eta_S)) \,.\end{aligned}$$

For the Potts model case, it simplifies into $H_T(\mathbf{t}) = -\eta_T \sum_{i \in V} \sum_{j \in \mathcal{N}(i)} < t_i, t_j >$, $H_S(\mathbf{s}) = -\eta_S \sum_{i \in V} \sum_{j \in \mathcal{N}(i)} < s_i, s_j >$ and

$$\begin{aligned}\tilde{H}_{T,S}(\mathbf{t}, \mathbf{s}) &= \sum_{i \in V} < t_i, e_{T^{s_i}} > \\ &\quad + \sum_{i \in V} \sum_{j \in \mathcal{N}(i)} 2\delta_{reg}^S(s_i, s_j)\, \eta_T < t_i, t_j > + \delta_{reg}^T(t_i, t_j)\, \eta_S(< e_{T^{s_i}}, e_{T^{s_j}} > + < s_i, s_j >) \,.\end{aligned}$$

References

[1] B. C. Arnold, E. Castillo, and J. M. Sarabia. Conditionally specified distributions: an introduction. *Statistical Science*, 16(3):249–274, 2001.

[2] J. Ashburner and K. J. Friston. Unified Segmentation. *NeuroImage*, 26:839–851, 2005.

[3] D. Benboudjema and W. Pieczynski. Unsupervised image segmentation using Triplet Markov fields. *Comput. Vision Image Underst.*, 99:476–498, 2005.

[4] J. Besag. Spatial interaction and the statistical analysis of lattice systems. *J. Roy. Statist. Soc. Ser. B*, 36(2):192–236, 1974.

[5] J. Besag. On the statistical analysis of dirty pictures. *J. Roy. Statist. Soc. Ser. B*, 48(3):259–302, 1986.

[6] M. J. Black and A. Rangarajan. On the unification of line processes, outlier rejection and robust statistics with application in early vision. *Int. Jour. Comput. Vision*, 19(1):57–91, 1996.

[7] W. Byrne and A. Gunawardana. Convergence theorems of Generalized Alternating Minimization Procedures. *J. Machine Learning Research*, 6:2049–2073, 2005.

[8] V. Cadez, S. Gaffney, and P. Smyth. A general probabilistic framework for clustering individuals and objects. In *6th ACM int. Conf. Knowledge Discovery and Data mining*, pages 140–149, 2000.

[9] G. Celeux, F. Forbes, and N. Peyrard. EM procedures using mean field-like approximations for Markov model-based image segmentation. *Pat. Rec.*, 36(1):131–144, 2003.

[10] D. L. Collins, A. P. Zijdenbos, V. Kollokian, J. G. Sled, N. J. Kabani, C. J. Holmes, and A. C. Evans. Design and construction of a realistic digital brain phantom. *IEEE trans. Med. Imag.*, 17(3):463–468, 1998.

[11] L. R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26:297–302, 1945.

[12] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman & Hall, 2nd edition, 2004.

[13] D. Geman and G. Reynolds. Constrained restoration and the recovery of discontinuities. *IEEE trans. Pat. Anal. Mach. Intell.*, 14(3):376–383, 1992.

[14] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE trans. Pat. Anal. Mach. Intell.*, 6:721–741, 1984.

[15] H.-O. Georgii. *Gibbs measures and phase transitions*. De Gruyter, 1988.

[16] D. Heckerman, D. M. Chickering, C. Meek, R. Rounthwaite, and C. Kadie. Dependency networks for inference, collaborative filtering and data visualization. *J. Machine Learning Research*, 1:49–75, 2000.

[17] F. Heitz and P. Bouthemy. Multimodal estimation of discontinuous optical flow using Markov random fields. *IEEE Trans. Pat. Anal. Mach. Intell.*, 15(12):1217–1232, 1993.

[18] M. Jenkinson and S. M. Smith. A global optimisation method for robust affine registration of brain images. *Medical Image Analysis*, 5(2):143–156, 2001.

[19] M.I. Jordan, Z. Ghahramani, T.S. Jaakkola, and L.K. Saul. An introduction to variational methods for graphical models. In M.I. Jordan, editor, *Learning in Graphical Models*, pages 105–162. 1999.

[20] S. Kumar and M. Hebert. Discriminative random fields. *Int. J. Comput. Vision*, 68(2):179–201, 2006.

[21] G.J. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. Wiley, 1996.

[22] R. Narasimha, E. Arnaud, F. Forbes, and R. Horaud. Cooperative disparity and object boundary estimation. In *15th IEEE Int. Conf. Imag. Proc. ICIP 08, San Diego, USA*, pages 1784–1787, 2008.

[23] K.M. Pohl, J. Fisher, E. Grimson, R. Kikinis, and W. Wells. A Bayesian model for joint segmentation and registration. *NeuroImage*, 31(1):228–239, 2006.

[24] M. J. D. Powell. An efficient method for finding the minimum of a function of several variables without calculating derivatives. *The Computer Journal*, 7, 1964.

[25] B. Scherrer, M. Dojat, F. Forbes, and C. Garbay. LOCUS: LOcal Cooperative Unified Segmentation of MRI brain scans. In *MICCAI 2007, Brisbane, Australia*, pages 219–227, 2007.

[26] B. Scherrer, F. Forbes, C. Garbay, and M. Dojat. Distributed Local MRF Models for Tissue and Structure Brain Segmentation. *IEEE trans. Med. Imag.*, 28:1296–1307, 2009.

[27] D. W. Shattuck, S. R. Sandor-Leahy, K. A. Schaper, D. A. Rottenberg, and R. M. Leahy. Magnetic resonance image tissue classification using a partial volume model. *NeuroImage*, 13(5):856–876, 2001.

[28] J. Sun, N-N. Zheng, and H-Y. Shum. Stereo matching using belief propagation. *IEEE trans. Pat. Anal. Mach. Intell.*, 25:787–800, 2003.

[29] K. Van Leemput, F. Maes, D. Vandermeulen, and P. Suetens. Automated model-based bias field correction in MR images of the brain. *IEEE trans. Med. Imag.*, 18(10):885–896, 1999.

[30] S. K. Warfield, K. H. Zou, and W. M. Wells. Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation. *IEEE trans. Med. Imag.*, 23(7):903–921, 2004.

[31] Y. Zhang, M. Brady, and S. Smith. Segmentation of brain MR images through a hidden Markov random field model and the Expectation-Maximisation algorithm. *IEEE trans. Med. Imag.*, 20(1):45–47, 2001.