On the convergence of EM-like algorithms for image segmentation using Markov random fields

4 authors, including:

Alexis Roche
Didimo
159 PUBLICATIONS   **5,242** CITATIONS

SEE PROFILE

Delphine Ribes Lemay
École Polytechnique Fédérale de Lausanne
15 PUBLICATIONS   **221** CITATIONS

SEE PROFILE

Meritxell Bach Cuadra
Lausanne University Hospital
245 PUBLICATIONS   **2,909** CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Autoradiographic Mouse Brain Sections Analysis View project

Brain morphometry View project

On the Convergence of EM-Like Algorithms for Image Segmentation using Markov Random Fields

Alexis Roche[a,c], Delphine Ribes[a], Meritxell Bach-Cuadra[b], Gunnar Krüger[a]

[a]*CIBM-Siemens, Ecole Polytechnique Fédérale (EPFL), CH-1015 Lausanne, Switzerland*
[b]*Signal Processing Laboratory 5, Ecole Polytechnique Fédérale (EPFL), CH-1015 Lausanne, Switzerland*
[c]*Computer Vision Laboratory, Eidgenössische Technische Hochschule (ETHZ), CH-8092 Zurich, Switzerland*

Abstract

Inference of Markov random field images segmentation models is usually performed using iterative methods which adapt the well-known expectation-maximization (EM) algorithm for independent mixture models. However, some of these adaptations are ad-hoc and may turn out numerically unstable. In this paper, we review three EM-like variants for Markov random field segmentation and compare their convergence properties both at the theoretical and practical levels. We specifically advocate a numerical scheme involving *asynchronous* voxel updating, for which general convergence results can be established. Our experiments on brain tissue classification in magnetic resonance images provide evidence that this algorithm may achieve significantly faster convergence than its competitors while yielding at least as good segmentation results.

Keywords: segmentation, Markov random field, expectation-maximization, mean field, convergence

1. Introduction

Nowadays, Markov Random Field (MRF) models are of widespread use in a variety of image segmentation tasks as they provide a natural approach to the general problem of partitioning an image into clusters of homogeneous signal value, while controlling the spatial smoothness of the partition. Following several authors (Van Leemput et al., 1999b; Kapur, 1999; Zhang et al., 2001; Bach Cuadra et al., 2005; Forbes and Fort, 2007), this paper uses MRF models for the automatic segmentation of brain magnetic resonance (MR) images into white matter (WM), gray matter (GM), and cerebrospinal fluid (CSF). The focus here is on the numerical convergence of MRF segmentation algorithms.

A customary, but restrictive, way of formulating MRF-based image segmentation is to search for the labeling with maximum a posteriori (MAP) probability given an image. In his seminal work, Besag (1974) proposed a simple MAP tracking algorithm called iterated conditional modes (ICM), which was later criticized for being prone to convergence to local maxima. A variety of alternative optimization algorithms were proposed, including simulated annealing (Geman and Geman, 1984), graph cuts (Boykov et al., 2001), max-product loopy belief propagation (Weiss and Freeman, 2001), gradient projection descent (Marroquin et al., 2003), or tree reweighted message passing (Wainwright et al., 2005). Some of these methods were recently compared on a variety of benchmark vision problems by Szeliski et al. (2008).

However, all MAP estimation methods have in common that they output a pointwise estimate of the label image, hence they overlook estimation uncertainty. A more comprehensive approach is to compute, or at least approximate, the posterior probability distribution of the label image. This strategy is known as approximate inference in the machine learning literature (Bishop, 2006) and has long been implemented in image segmentation using "EM-like" algorithms. A strong motivation for approximate inference in MR tissue classification is to provide clinicians with error bars on anatomical measures derived from segmentation results, such as volumes of gray or white matter. Such measures are prone to uncertainty due to image noise and partial volume effects, although deterministic models can be used to account for the latter (Van Leemput et al., 2003; Bach Cuadra et al., 2005; Liang and Wang, 2009).

"EM-like" refers to the classical expectation-maximization (EM) algorithm for finite mixture models used, e.g., in Wells et al. (1996); Prima et al. (2001); Ashburner and Friston (2005), which assumes spatially independent labels and is therefore not suited to any but the simplest MRF model. When spatial dependencies are involved, the EM algorithm

becomes analytically intractable. While Monte Carlo methods can then be used to accurately approximate posterior probabilities, but at heavy computational cost (Van Leemput et al., 2003), several authors have proposed analytical adaptations of the finite mixture EM algorithm to MRF image segmentation models. To our best knowledge, the first such attempt dates back to Zhang (1992), who proposed an ad-hoc modification of the E-step based on a computational trick from statistical physics known as the mean-field (MF) approximation, which essentially averages interactions between variables as if they were mutually independent. The resulting MF-EM algorithm approximates the posterior distribution of the labels as a product of voxelwise marginals, yet without assuming a similar factorization for the prior.

In the MF-EM scheme proposed by Zhang (1992), the voxelwise posterior distributions are updated in parallel. The same author realized that this is not guaranteed to converge for models with strong prior spatial smoothness, as characterized by the "damping factor" denoted β in Section 2. He later proposed an extension to work around this issue (Zhang, 1996), but Fessler (1998) pointed a flaw in the convergence proof and concluded: "the convergence of the mean field iteration remains an important open question for large β". Nevertheless, the synchronous voxel updating scheme underlying the MF-EM algorithm appears to have been successfully applied to MR tissue classification (Van Leemput et al., 1999b; Kapur, 1999; Forbes and Fort, 2007).

Some neuroimaging software packages such as FSL (http://www.fmrib.ox.ac.uk/fsl) or the VBMtools extension for SPM (http://dbm.neuro.uni-jena.de/vbm) implement another numerical scheme proposed by an homonym author (Zhang et al., 2001), which can be seen as a variant of MF-EM where the mean-field approximation is replaced with a pseudo-likelihood maximization step as in the ICM method of Besag (1974). This scheme, denoted ICM-EM in the sequel, is computationally comparable to MF-EM but requires less memory storage. It was proved applicable to a variety of tissue classification models (Bach Cuadra et al., 2005). Similarly to MF-EM, however, no general convergence result is available, and we will see in Section 4 that ICM-EM may actually fail to converge even when carefully initialized.

This paper advocates a third EM variant for inference in MRF models that enjoys both theoretical convergence properties and faster convergence in practice than both the MF-EM and ICM-EM schemes, with no counterpart computational overhead and even less memory requirements. This scheme was previously used by Weiss (2001) in toy problems of low level vision, but its application to medical image segmentation does not seem to be documented anywhere. It is a simple variant of the MF-EM algorithm in which voxels are updated asynchronously rather than all at once. Convergence is ensured by the fact that each of these incremental updates decreases a lower-bounded objective function.

The algorithm can be interpreted as an instance of the variational expectation-maximization (VEM) algorithm, also known as variational Bayes, a generalization of the EM algorithm covered in recent handbooks of computational statistics (Bishop, 2006). VEM was previously considered for image segmentation in Penny et al. (2005); Woolrich and Behrens (2006); Forbes and Fort (2007) but these references do not discuss asynchronous voxel updating, and thus miss the issue of convergence.

2. Markov random field segmentation

2.1. Statistical model

The classical hidden MRF image segmentation model is a two-level hierarchical probabilistic model. The first level specifies the conditional probability of the observed image Y given an unknown label image X which conventionally takes on values in $\{1, \ldots, K\}$, where K is the number of classes. In the basic model, voxels are assumed to be mutually independent given the labels and normally distributed within each class:

$$P(Y|X, \mu, \sigma) = \prod_i N(y_i; \mu_{x_i}, \sigma_{x_i}),$$

where $\mu = (\mu_1, \ldots, \mu_K)$ and $\sigma = (\sigma_1, \ldots, \sigma_K)$ are the unknown vectors of class-specific means and standard deviations, respectively.

The second level of the model describes the prior probability on the label image, and is conveniently formulated using a MRF model. As a consequence of the famous Hammersley-Clifford theorem (Li, 1995), this is equivalent to

defining a Gibbs potential $U(X)$, which ranks possible configurations proportionally to a positive-valued parameter β:

$$P(X|\beta) = \frac{1}{Z(\beta)} e^{-\beta U(X)}, \tag{1}$$

where $Z(\beta)$ is a normalization constant as long as β is tuned beforehand and held fixed throughout segmentation, as will be the case in our experiments. See, e.g., Forbes and Fort (2007); Woolrich and Behrens (2006) for adaptive tuning strategies. Except in trivial cases, $Z(\beta)$ lacks a closed-form expression.

In MR segmentation applications, the Gibbs potential is usually defined via cliques of order 2:

$$U(X) = \sum_{i,j \in N_i} U_{ij}(x_i, x_j), \tag{2}$$

where U_{ij} is a pairwise interaction energy, and N_i is the set of voxels which are neighbors of voxel i according to a given discrete topology. In this work, we used the 26-neighborhood. Note that the Gibbs potential (2) may involve an additional order-one potential, sometimes called "external field", in order to represent marginal voxelwise tissue priors as obtained, e.g. by registering a probabilistic atlas onto the image to be segmented.

As pointed out by Geman and Geman (1984), the Markov-Gibbs prior (1) maximizes entropy among all distributions such that the expected potential is upper bounded by a known constant. This means that the MRF model is, in some sense, least informative about the labeling X, and is thus justified as a general segmentation prior under vague anatomical knowledge.

2.2. Exact vs. approximate Bayesian inference

In the following, we merge the unknown parameters into a single parameter $\theta = (\mu, \sigma)$. Given an image Y, the canonical Bayesian approach to inference about the labeling X involves computing the conditional distribution:

$$P(X|\theta, Y) = \frac{P(Y|X, \theta)P(X)}{P(Y|\theta)},$$

where $P(Y|\theta)$ is the so-called integrated likelihood:

$$P(Y|\theta) = \sum_{X \in \mathcal{X}} P(Y|X, \theta)P(X), \tag{3}$$

and integrating $P(X|\theta, Y)$ with respect to some prior distribution $\pi(\theta)$, yielding the posterior $P(X|Y)$. The general goal of Bayesian inference is to "summarize" the posterior by the expectation of well-chosen functions of X. In (3), the sum takes place on the set $\mathcal{X} = \{1, \ldots, K\}^n$ of all possible configurations for X, and becomes intractable as soon as more than a dozen voxels are involved. This curse of dimensionality makes it impossible to compute posterior expectations, not to mention the fact that the parameter θ needs to be further marginalized out. An exact implementation of the Bayesian paradigm is therefore hopeless in this case.

Alternatively, we may resort to approximating the posterior by a tractable distribution, for instance one that factorizes as a product of voxelwise distributions:

$$P(X|Y) \approx Q(X) = \prod_i q_i(x_i). \tag{4}$$

This form is sometimes referred to as a "fully factorized" approximation. It may be useful as long as one is only interested in estimating quantities that sum up functions of single voxels, such as tissue volumes, although it is irrelevant to estimate quantities that involve posterior correlations between voxels.

3. Approximate inference algorithms

3.1. VEM algorithm

A general method of fitting a simplified distribution is the VEM algorithm, an extension of the well-known EM algorithm (Dempster et al., 1977). The VEM algorithm seems to have originated in the pioneering work of Csiszár and

Tusnády (1984) and has progressively become a popular tool in machine learning (Neal and Hinton, 1998; Bishop, 2006). Its underlying principle is to minimize the so-called free energy function:

$$F(Q, \theta) = \sum_{X \in \mathcal{X}} Q(X) \log \frac{Q(X)}{P(X, Y|\theta)} \tag{5}$$

$$= \sum_i \sum_{x_i} q_i(x_i) \Big[\log q_i(x_i) - \log N(y_i, \mu_{x_i}, \sigma_{x_i})$$

$$+ \beta \sum_{j \in N_i} \sum_{x_j} q_j(x_j) U_{ij}(x_i, x_j) \Big] + \log Z(\beta) \tag{6}$$

Note that it is not necessary to evaluate the partition function $Z(\beta)$ which plays no role when β is fixed. The rational for this minimization appears when rewriting free energy in terms of the Kullback-Leibler divergence $D(Q\|P_{X|Y,\theta})$ from $Q(X)$ to $P(X|Y, \theta)$:

$$F(Q, \theta) = D(Q\|P_{X|Y,\theta}) - \log P(Y|\theta).$$

Hence, minimizing free energy with respect to Q, at fixed θ, boils down to finding the approximating distribution $Q(X)$ closest to $P(X|\theta, Y)$ in the sense of the Kullback-Leibler divergence. To understand the more general minimization with respect to both Q and θ, we may assume for technical reasons that θ takes on discrete values and assign it a flat prior probability so that $P(\theta|Y) = cP(Y|\theta)$, where c is independent from θ. Free energy may then be rewritten:

$$F(Q, \theta) = D(Q.\delta_\theta\|P_{X,\theta|Y}) + \log c,$$

where δ_θ denotes the Kronecker delta. Therefore, $F(Q, \theta)$ is essentially (up to an additive constant) the Kullback-Leibler divergence from the *joint* factorized distribution $Q(X, \theta') = Q(X)\delta_\theta(\theta')$ to the posterior joint distribution $P(X, \theta'|Y)$. This result extends to continuous-valued parameters as a limiting case of large discrete parameter spaces. In short, free energy is nothing but a Kullback-Leibler divergence in the labels × parameters product space.

In its academic form, the VEM algorithm minimizes (5) alternatively with respect to Q and θ, yielding updating rules called the VE-step and the VM-step, respectively. As seen above, each step may be thought of as adjusting one component of the product approximation $Q.\delta_\theta$, hence implicitly assuming that $P(X|\theta, Y) \approx P(X|Y)$ resembles a fully factorized distribution and that $P(\theta|X, Y) \approx P(\theta|Y)$ is sharply peaked around a certain θ value, which should then be close to the maximum likelihood estimate. We shall note that the VM-step is formally equivalent to the M-step of the classical independent mixture EM algorithm (indep EM), yielding a unique solution in our case. For each $k \in \{1, \dots, K\}$,

$$\hat{\mu}_k = \frac{\sum_i q_i(k) y_k}{\sum_i q_i(k)}, \qquad \hat{\sigma}_k^2 = \frac{\sum_i q_i(k)(y_k - \hat{\mu}_k)^2}{\sum_i q_i(k)}.$$

There is some technical difficulty in implementing the VE-step which amounts to a fix point equation with no closed-form solution (Forbes and Fort, 2007). However, the mean-field approximation (4) suggests an incremental alternative to the standard VE-step whereby the global minimization of the free energy with respect to Q is replaced with successive, simpler minimizations along each marginal q_i with all $q_{j \neq i}$ held fixed. Each such minimization yields a strictly convex problem:

$$\min_{q_i} \Big\{ \sum_k q_i(k) \Big[\log q_i(k) - \log N(y_i, \mu_k, \sigma_k)$$

$$+ 2\beta \sum_{j \in N_i} \sum_{x_j} q_j(x_j) U_{ij}(k, x_j) \Big] \Big\},$$

the solution of which is found explicitly using differential calculus:

$$q_i(k) \propto N(y_i; \mu_k, \sigma_k) e^{-2\beta \sum_{j \in N_j} \sum_{x_j} q_j(x_j) U_{ij}(k, x_j)}. \tag{7}$$

The proportionality constant is tuned so that $\sum_k q_i(k) = 1$. From an implementation perspective, this simply involves correcting the E-step of the indep EM algorithm with a "neighborhood agreement" multiplicative factor.

3.2. Convergence of asynchronous VEM

By construction, the sequential voxel updating in VEM guarantees that only moves which decrease free energy are accepted, so that the system can never return to a previous state of higher free energy. Asynchronous voxel updating is perhaps not intuitively appealing as it makes the algorithm dependent on the voxel space traversal. In our implementation, voxels are visited by default in the order of the image buffer in memory, which is obviously arbitrary. At this stage, we have limited expertise as to optimally traverse the voxel space. It should however be kept in mind that *any* traversal schedule warrants monotonic decrease in free energy.

This monotonicity property almost immediately implies that the sequence of free energy updates converges to a finite value. As a technicality, we need to add a constraint to the VM-step to force the σ_k updates to stay above a very small threshold (this plays no role in practice), thereby ensuring that the free energy is lower bounded.

Additional convergence properties follow from a general result on alternating optimization given in Bezdek and Hathaway (2002), Theorem 2. The above non-zero variance constraint implies that θ is searched in a compact subset of $\mathbb{R}^K \times \mathbb{R}_+^{*K}$, without affecting the existence and uniqueness properties of the VM-step. The theorem then states that the limit of every convergent subsequence of parameters $(q_1, \ldots, q_n, \theta)$ is a local minimizer or a saddle point of the free energy. This means that, in the worst-case scenario, there might be several accumulation points with the very same level of free energy. Except in such very unlikely pathological situations, the algorithm is therefore guaranteed to converge, albeit not necessarily towards the global minimizer of free energy.

3.3. MF-EM algorithm

It turns out that the mean-field algorithm introduced by Zhang (1992), abbreviated MF-EM in this paper, is closely related to the above-described VEM scheme. The sole difference is that the MF-EM algorithm updates the approximating probability map $Q(X)$ all at once rather than sequentially, meaning that (7) is computed from a *copy* of $Q(X)$. This implies that MF-EM is independent from voxel space traversal contrary to VEM. Beside doubling memory load, the MF-EM scheme may fail to converge (Zhang, 1996; Weiss, 2001).

The analogy between MF-EM and the VEM algorithm was already noted by Forbes and Fort (2007), however we should stress that MF-EM is not, strictly speaking, a VEM algorithm because the synchronous updating of $Q(X)$ may not decrease free energy. Specifically, Forbes and Fort (2007) present the synchronous scheme as a derivation of VEM in which the approximating distribution $Q(X)$ is optimized all at once (as in the standard VEM formulation presented in Section 3.1), leading to a fix point equation. Whether this fix point equation can be solved by the synchronous iteration is precisely the topic of Zhang (1996) and is not addressed in Forbes and Fort (2007). However, if the iteration converges, then we know that it converges towards one of the above mentioned fix points, which contain all the local minima of free energy by construction. Hence, if convergence occurs and if there are few fix points, as is reasonable to expect, then the MF-EM is likely to converge to the same solution as the asynchronous VEM when both algorithms are initialized in the same way.

3.4. ICM-EM algorithm

Another popular ad-hoc extension of indep EM proposed in (Zhang et al., 2001) involves modifying the classical E-step based on a pseudo-likelihood heuristic:

$$q_i(x_i) \propto N(y_i; \mu_{x_i}, \sigma_{x_i}) P(x_i | \hat{X}_{N_i}),$$

where $\hat{X}$ is a current MAP estimate of the labeling which is re-computed at the beginning of the E-step via $\hat{X} = \arg\max_X Q(X)$, yielding $\hat{x}_i = \arg\max_k q_i(k)$, $\forall i$. The update schedule is synchronous similarly to MF-EM, and amounts to a simple variant of (7) where $\sum_{x_j} q_j(x_j) U_{ij}(k, x_j)$ is replaced with $U_{ij}(k, \hat{x}_j)$. In other words, each neighbor contributes a single tissue class in a winner-take-all fashion, instead of releasing an "opinion poll" as in both VEM and MF-EM. This may be the main reason for the much weaker convergence properties of ICM-EM demonstrated experimentally in the sequel.

4. Experiments

This section presents an experimental convergence study of the three above-described algorithms (VEM, MF-EM, ICM-EM) applied to brain MR segmentation from both simulated and real data. For comparison, we also included the independent mixture EM algorithm using a *fixed* uniform prior on tissue classes, which corresponds to the limiting case of each of the three above algorithms when β goes to zero. Note that it is more common to use independent mixture models with adjustable proportions for MR segmentation, but this would be inconsistent with the chosen MRF prior, which involves no parameter beside β (see Section 2.1).

4.1. Datasets

An artificial dataset of 9 images was generated from a unique anatomical skull-stripped brain model using the BrainWeb MRI simulator provided by the McConnell Brain Imaging Center (Collins et al., 1998; Kwan et al., 1999), see `http://www.bic.mni.mcgill.ca/brainweb`. All simulated images have a $217 \times 181 \times 181$ matrix with isotropic 1 mm^3 voxel size, 0% RF nonuniformity, and 5% noise level. The noise level was roughly estimated from a few of the real images described below, by computing intensity statistics in manually delineated homogeneous regions. We used another dataset, referred to as the multi-subject BrainWeb dataset in the following, consisting of 20 T1-weighted images with $256 \times 256 \times 181$ matrix and isotropic 1 mm^3 voxel size simulated from normal brains (Aubert-Broche et al., 2006). The images provided by BrainWeb have 20% RF nonuniformity and 3% noise level, and it is currently impossible to simulate other images. RF nonuniformity was corrected using the method in (Sled et al., 1998). Segmentations were restricted to masks of the total intracranial volume (TIV) computed by summing up the fuzzy models (Kwan et al., 1999) corresponding to WM, GM and CSF, respectively, and thresholding the result at 0.5.

Real MR images from 76 normal subjects (49 ± 27.5 yrs) with slightly anisotropic voxel size $1 \times 1 \times 1.2$ mm^3 were acquired using the same protocol as in the Alzheimer's Disease Neuroimaging Initiative (`http://adni.loni.ucla.edu/research/protocols/mri-protocols`) on a 3 Tesla scanner (Magnetom Trio a Tim System, Siemens Healthcare, Erlangen, Germany) at Lausanne University Hospital, Switzerland. Acquisition involved a 32-head channel coil and a three dimensional (3D) T1-weighted magnetization-prepared rapid acquisition gradient echo (MPRAGE) with TR = 2300 ms, TE = 2.98 ms, FOV = $256 \times 240 \times 160$ mm^3, BW = 240 Hz/pixel, 160 slices, slice thickness = 1.2 mm and GRAPPA = 3. Images were submitted to a series of prepossessing involving successive corrections for B1 receive (Narayana et al., 1988), gradient distortion (Schmitt, 1985; Wald et al., 2001) and intensity nonuniformity (Sled et al., 1998). Finally, TIV masks were computed using SPM8 (Ashburner and Friston, 2005) to exclude non-brain tissues such as fat, vessels and eyes.

4.2. Model specification

Both artificial and real images were submitted, together with their associated TIV masks, to automatic segmentation in three classes labeled as CSF, GM and WM. We used the same distance-weighted Potts model as in Zhang et al. (2001) and Bach Cuadra et al. (2005) as a special instance of the Markov random field prior model described in Section 2. In this model, the pairwise potentials read:

$$U_{ij}(x_i, x_j) = \frac{w_{ij}}{2}[1 - \delta(x_i, x_j)],$$

where $\delta(.,.)$ is the Kronecker delta and w_{ij} is the inverse Euclidean distance between voxels i and j. As discussed in the aforementioned references, distance weighting is safe for anisotropic images, although in our case weights are uniform for the artificial data and close to uniform for the real data.

For simplicity of presentation, we report results obtained with the basic sampling model in which the intensity distribution within each class is represented by a single Gaussian. We tested other models, including mixtures of generalized Gaussians (Prasad et al., 2007) and the 5-class mixed tissue model proposed in Shattuck et al. (2001), which all yielded qualitatively similar convergence results. Other strategies to deal with partial volume include oversampling the label image (Van Leemput et al., 2003), or estimating proportions of tissue types in each voxel (Liang and Wang, 2009). We also direct the reader to (Wells et al., 1996; Van Leemput et al., 1999a; Prima et al., 2001; Van Leemput et al., 2003) for segmentation models that incorporate bias field correction.

4.3. Evaluated algorithms

We used an in-house implementation of the four different algorithms (indep EM, VEM, MF-EM and ICM-EM) written in both Python and C languages based on Neuroimaging in Python (`http://www.nipy.org`). Each algorithm was run for 75 iterations for each image. Many iterations were deliberately chosen for proper assessment of convergence performances. The computation time on a standard PC (single processor Intel Core i7 720QM running at 1.60GHz) was roughly 1'30 minute per image for indep EM and 3 minutes for each of the three other algorithms, plus another 3 minutes to compute the quality and convergence measures described below. There was little difference in timings between datasets since TIV masks were similar in size.

Algorithms were always initialized with uniform tissue membership probabilities: $q_i(k) = 1/3$ for every voxel i and class k. Initial tissue means and standard deviations were tuned using a fast automatic moment matching method based on an additional Brainweb case. For this reference case, tissue means $\mu^\star = [813.9, 1628.4, 2155.8]$ and standard deviations $\sigma^\star = [215.6, 173.9, 130.9]$ for CSF, GM and WM, respectively, were estimated once and for all by running indep EM from a careful manual initialization. A simple proportionality rule was then used to guess the tissue class parameters of an incoming image: the global intensity mean and standard deviation m and s were computed on the whole TIV, and compared with the corresponding global mean and standard deviation $m^\star = 1643.1$ and $s^\star = 502.8$ found in the reference Brainweb case, yielding:

$$\mu_k = a\mu_k^\star + b, \qquad \sigma_k = a\sigma_k^\star,$$

where $a = s/s^\star$ and $b = m - am^\star$. This method provides a fairly accurate alternative to atlas-based initialization while avoiding the need for image registration.

4.4. Quality measures

Since a ground truth tissue classification is available for the BrainWeb data, we may evaluate the segmentation accuracy via some overlap measure. Specifically, we use the following coefficient computed between the BrainWeb fuzzy tissue membership maps (Kwan et al., 1999) and the posterior probability maps output by the algorithm under evaluation:

$$D_k = 2\frac{\sum_i \sqrt{p_i(k)q_i(k)}}{\sum_i p_i(k) + q_i(k)}, \tag{8}$$

where $p_i(k)$ is the BrainWeb probability of tissue k at voxel i and $q_i(k)$ the corresponding probability output by the algorithm. In the following, we refer to (8) as the "fuzzy Dice coefficient" given that it indeed generalizes the traditional Dice coefficient to non-binary probabilities. The definition follows from interpreting the traditional Dice coefficient as a normalized dot product between binary vectors.

In the real data case, there is no direct measure of segmentation quality. As suggested in Section 3, we may use free energy (5) to assess how close an approximation to the actual posterior of tissue labels. Free energy is the objective function being explicitly minimized by the VEM algorithm. While no such property is established for the other algorithms, a general empirical observation is that they tend to decrease free energy across iterations, possibly up to small fluctuations as illustrated below.

4.5. Convergence measure

In order to assess the convergence of the different algorithms, we study how the volume of each tissue class varies across iterations, where the volume V_k of class k is defined by summing up the output posterior probabilities $q_i(k)$ over the TIV. A natural convergence test is then based on the maximum relative tissue change between the current and previous iterations:

$$\varepsilon_V = \max_{k \in \{GM, WM, CSF\}} \frac{|V_k^{(r)} - V_k^{(r-1)}|}{V_k^{(r-1)}},$$

where r stands for the iteration number. Convergence may be declared as soon as ε_V falls below a pre-specified threshold, $\varepsilon_V < 10^{-\ell}$ or, equivalently, $\log_{10}(\varepsilon_V) < -\ell$, where ℓ is typically a positive integer representing the severity of the convergence test.
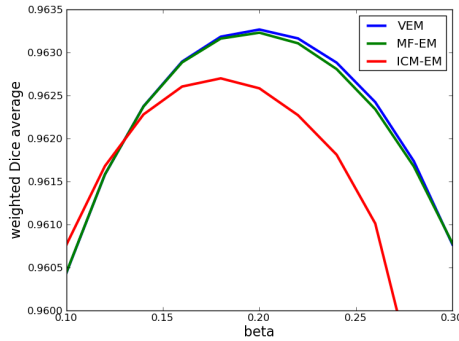
Figure 1: Weighted fuzzy Dice coefficients as a function of the β parameter for the VEM, MF-EM and ICM-EM algorithms.

4.6. Tuning of β

Automatic segmentations of the same left-out BrainWeb image as used in the parameter initialization method (see Section 4.3) were performed using the VEM, MF-EM and ICM-EM algorithms for different values of the β parameter (1) varying from 0 to 0.5 by increments of 0.02. For each algorithm, we plotted the sum of fuzzy Dice coefficients weighted by the ground truth normalized tissue volumes (0.18, 0.48, 0.34 for CSF, GM, WM, respectively). The curves of both VEM and MF-EM are seen in Figure 1 to attain their maximum at $\beta = 0.2$, while that of ICM-EM peaks at $\beta = 0.18$. In all experiments reported below, β was set to 0.2, except for indep EM which uses $\beta = 0$ by definition. We verified that setting $\beta = 0.18$ had negligible impact on the results.

4.7. Results on simulated data

Quality measures indicate that the three algorithms using the Markov random field model give very similar results after 75 iterations on both the single-subject and multi-subject datasets. As seen in Table 1, they yield significantly higher fuzzy Dice coefficients for both CSF and GM than the indep EM algorithm. We note however, as reflected by Table 2 and Figure 4, that the final value of free energy obtained with VEM was the smallest compared to the other algorithms in most cases (note that the comparison with indep EM is somewhat irrelevant since free energy was computed at $\beta = 0.2$). In 6 cases of the multi-subject dataset, the MF-EM ended up with lower free energy than VEM only by tiny amounts of the order of machine precision.

Dataset	Tissue	indep EM	VEM	MF-EM	ICM-EM
	CSF	$0.86 \pm 2.9 \cdot 10^{-4}$	$0.96 \pm 0.9 \cdot 10^{-4}$	$0.96 \pm 0.9 \cdot 10^{-4}$	$0.96 \pm 1.0 \cdot 10^{-4}$
Single-subject	GM	$0.93 \pm 0.8 \cdot 10^{-4}$	$0.96 \pm 1.0 \cdot 10^{-4}$	$0.96 \pm 1.0 \cdot 10^{-4}$	$0.96 \pm 1.0 \cdot 10^{-4}$
	WM	$0.97 \pm 0.7 \cdot 10^{-4}$	$0.97 \pm 1.8 \cdot 10^{-4}$	$0.97 \pm 1.8 \cdot 10^{-4}$	$0.97 \pm 2.0 \cdot 10^{-4}$
	CSF	$0.91 \pm 5.6 \cdot 10^{-2}$	$0.96 \pm 1.2 \cdot 10^{-2}$	$0.96 \pm 1.2 \cdot 10^{-2}$	$0.96 \pm 1.2 \cdot 10^{-2}$
Multi-subject	GM	$0.95 \pm 0.7 \cdot 10^{-2}$	$0.96 \pm 0.4 \cdot 10^{-2}$	$0.96 \pm 0.4 \cdot 10^{-2}$	$0.96 \pm 0.4 \cdot 10^{-2}$
	WM	$0.97 \pm 0.5 \cdot 10^{-2}$	$0.97 \pm 0.9 \cdot 10^{-2}$	$0.97 \pm 0.9 \cdot 10^{-2}$	$0.97 \pm 0.8 \cdot 10^{-2}$

Table 1: BrainWeb data: means and standard deviations of the fuzzy Dice coefficients after 75 iterations.

Visual assessment confirms that there are little differences between the final results of VEM, MF-EM and ICM-EM while, not surprisingly, indep EM tends to produce more "noisy" classifications. This is illustrated in Figure 2 which compares hard tissue classifications obtained by assigning each voxel in the TIV to its most probable label.

The convergence analysis, however, reveals significant differences between algorithms, as illustrated in Table 3 by the number of iterations needed to reach a given tolerance on the relative tissue volume change. Especially in the single-subject dataset, these are seen to vary from one algorithm to the other, with VEM requiring the fewest iterations, followed by MF-EM, ICM-EM, and indep EM. These findings reflect the curves in Figure 3 where the

8

Dataset	indep EM	MF-EM	ICM-EM
Single-subject	$4.5 \cdot 10^{-1}$ (100%)	$4.9 \cdot 10^{-9}$ (100%)	$1.2 \cdot 10^{-4}$ (100%)
Multi-subject	$6.7 \cdot 10^{-1}$ (100%)	$5.2 \cdot 10^{-14}$ (70%)	$7.0 \cdot 10^{-5}$ (100%)

Table 2: BrainWeb data: mean of final relative free energies and percentage of cases where the final relative free energy is positive. Relative free energy is defined by subtracting the free energy obtained with VEM and dividing the result by the number of voxels in the TIV.
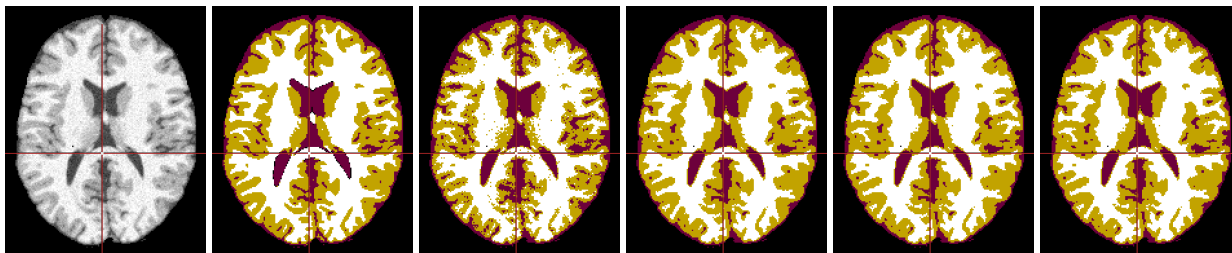

Figure 2: Hard tissue classifications obtained from a BrainWeb image after 75 iterations. From left to right: raw slice, ground truth, results of indep EM, ICM-EM, MF-EM, and VEM.

base-10 logarithm of ε_V is plotted against the iteration number. Convergence speed may be characterized by how fast $\log_{10}(\varepsilon_V)$ goes to $-\infty$ over iterations. The curves associated with VEM are the fastest to decrease, only with two borderline cases for which convergence is hardly faster than with MF-EM. On the other hand, the curves of ICM-EM show that ε_V stabilizes around 10^{-6} after 45-60 iterations, meaning that the algorithm fails to converge in all 9 single-subject cases. Figure 4 shows that ICM-EM actually oscillates between two states of slightly different free energies. Such behavior is impossible with VEM since free energy cannot increase over iterations.

Discrepancies between MRF-based algorithms are less spectacular in the multi-subject dataset where VEM and MF-EM converge at very similar speeds, probably due to the fairly low amount of simulated noise (3%) in the T1-weighted BrainWeb images, which made it easier for both MF-EM and ICM-EM to converge. This observation suggests that convergence performances are more critically dependent on image quality than patient anatomy.

Dataset	Test	indep EM	VEM	MF-EM	ICM-EM
Single-subject	$\varepsilon_V < 10^{-2}$	15 ± 0	1 ± 0	2 ± 0	11 ± 0
	$\varepsilon_V < 10^{-3}$	28 ± 0	8 ± 0	12 ± 0	23 ± 0
	$\varepsilon_V < 10^{-4}$	41 ± 0	18 ± 0	25 ± 0	36 ± 1
Multi-subject	$\varepsilon_V < 10^{-2}$	7.5 ± 1.5	4 ± 0	4 ± 0	5 ± 0
	$\varepsilon_V < 10^{-3}$	28.5 ± 9.5	10 ± 0.5	10 ± 1	13 ± 1.5
	$\varepsilon_V < 10^{-4}$	52.5 ± 22	17 ± 1	16 ± 1	22.5 ± 2.5

Table 3: BrainWeb data: median and median of absolute deviations of the minimal number of iterations needed to reach a given tolerance on the relative tissue volume change ε_V.

4.8. Results on real data

The 76 real images were segmented using the same initialization strategy and damping factor value ($\beta = 0.2$) as for the BrainWeb database. A major difference with BrainWeb data is that the final classifications obtained with ICM-EM are visually quite different from those obtained with both VEM and MF-EM, which turn out similar again. A typical effect of ICM-EM is to misclassify cortical gray matter as CSF, like in the example of Figure 5. While it is generally possible to improve results by restarting the algorithm from a more accurate initialization, this again points out the lack of robustness of this scheme.
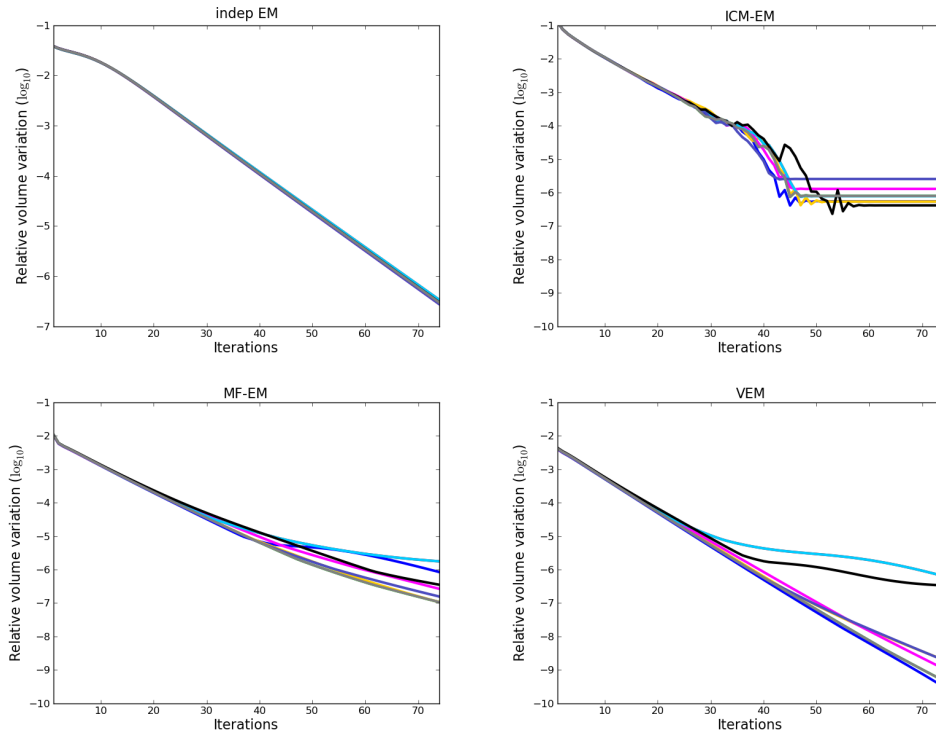
9

Figure 3: Single-subject BrainWeb data: convergence speed study. The maximum relative volume variations are plotted in log scale $\log_{10}(\varepsilon_V)$ against iterations. Each color represents a different case.
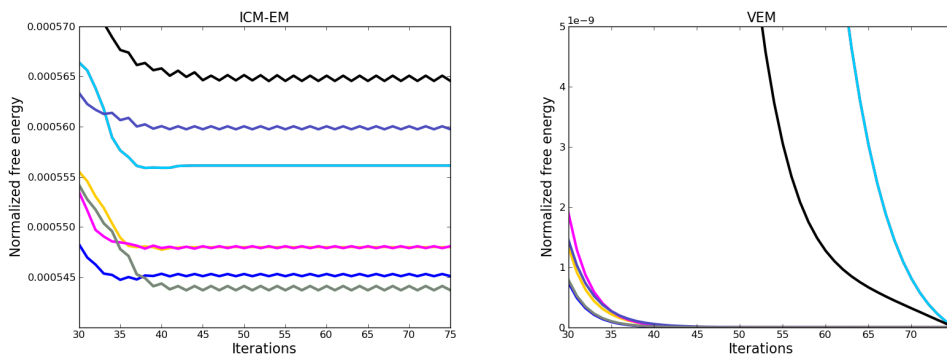

Figure 4: Single-subject BrainWeb data: comparison of ICM-EM and VEM regarding free energy variations in late iterations. Colors code for the same cases as in Figure 3. While free energy decreases continuously with VEM, as predicted by theory, it exhibits oscillatory behavior with ICM-EM.
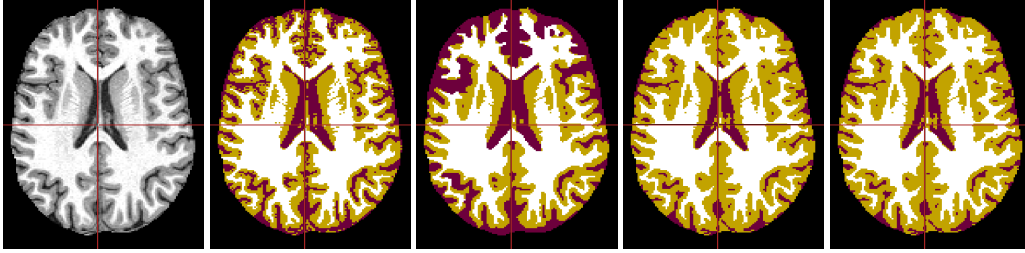
Figure 5: Hard tissue classifications obtained from a real MR image after 75 iterations. From left to right: raw slice, results of indep EM, ICM-EM, MF-EM, and VEM.

These findings are reflected by the free energy curves in Figure 6, which exhibit more variability across subjects for both indep EM and ICM-EM than for both their competitors. For the sake of visualization, a subset of 9 randomly selected cases is represented. Table 4 shows that the final free energy values obtained with ICM-EM are significantly higher than with both VEM and MF-EM in all 76 cases. The comparison of final free energies between the two latter methods indicates lower energy for VEM in 97% cases (74 out of 76), yet with much smaller differences in cases all cases.

indep EM	MF-EM	ICM-EM
$6.0 \cdot 10^{-1}$ (100%)	$5.6 \cdot 10^{-6}$ (97%)	$7.0 \cdot 10^{-3}$ (100%)

Table 4: Real data: mean of final relative free energies and percentage of cases where the final relative free energy is positive. Relative free energy is defined by subtracting the free energy obtained with VEM and dividing the result by the number of voxels in the TIV.

As for convergence, VEM is again significantly faster than both MF-EM and ICM-EM (see Table 5), although we note that the number of necessary iterations is much larger than in the BrainWeb datasets for all three methods. Neither MF-EM nor ICM-EM achieves convergence at 10^{-4} within 75 iterations.

Surprisingly enough, indep EM converges faster than with simulated data and is now the fastest method (although it always converged to sub-optimal free energy configurations). Careful inspection of the results shows that indep EM consistently overestimated the mean intensity of CSF, and underestimated its variance. This is easily understood looking at the image histograms in Figure 7 which reveal stronger dispersion of CSF intensities in real data than in BrainWeb data, resulting in the CSF mode being "swamped" beneath the GM. In such a situation, indep EM tends to converge faster as it erroneously recruits voxels in the GM to stabilize CSF parameter estimates. This effect can be hold in using spatial regularization, as shown by the results of both VEM and MF-EM, but also turns out to be amplified by ICM-EM in cases such as in Figure 5.

Figure 8 compares the different algorithms in terms of the relative volume variations ε_V across iterations, for another 9 randomly selected subjects. The different convergence speeds are highlighted and confirm all the trends observed on the BrainWeb data, except for the slower convergence of indep EM.

Test	indep EM	VEM	MF-EM	ICM-EM
$\varepsilon_V < 10^{-2}$	9 ± 1	10.5 ± 2.5	12 ± 4	21 ± 11.5
$\varepsilon_V < 10^{-3}$	18 ± 2	29.5 ± 4.5	39 ± 6	54 ± 18
$\varepsilon_V < 10^{-4}$	29 ± 3	51 ± 7.5	> 75	> 75

Table 5: Real data: median and median of absolute deviations of the minimal number of iterations needed to reach a given tolerance on the relative tissue volume change ε_V.
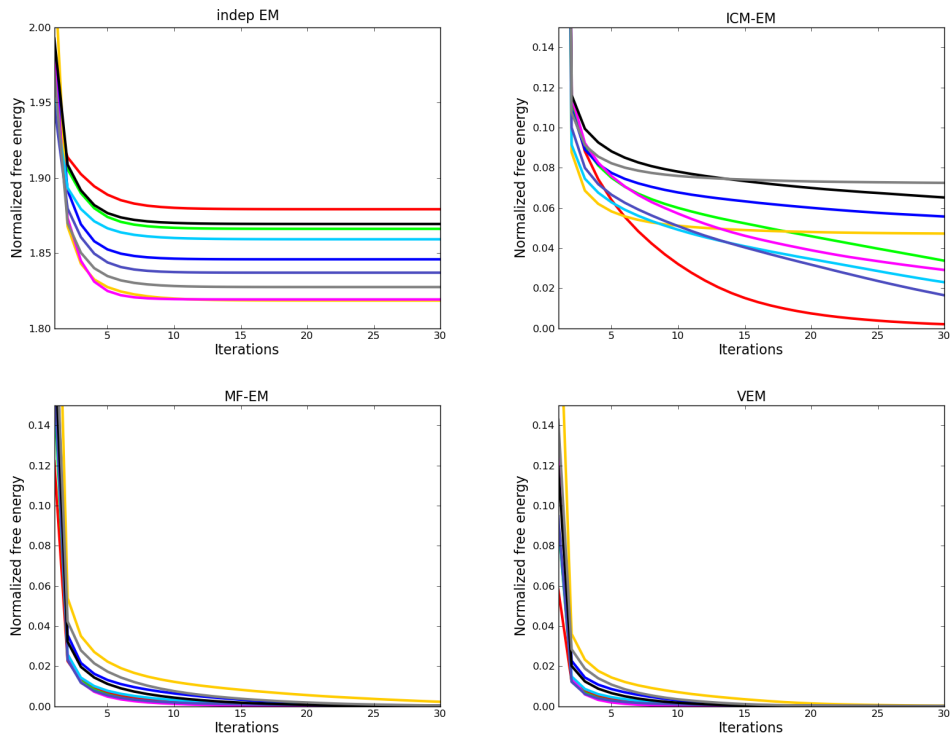
Figure 6: Real data: free energy values across iterations for a random subset of 9 subjects (each color represents a different subject). For better visualization, only the first 30 iterations are represented. Free energy is normalized by subtracting the final free energy obtained with VEM and dividing the result by the number of voxels in the TIV. Note that the indep EM plot is not at the same scale as the others.
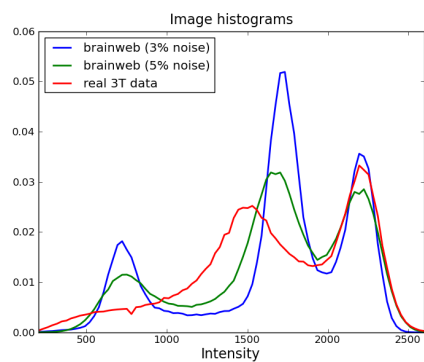

Figure 7: Histograms of intensities within TIV masks for 3 different T1-weighted images: a BrainWeb image with 3% noise (multi-subject dataset), a BrainWeb image with 5% noise (single-subject dataset), and a real image. Intensities are rescaled to the reference BrainWeb case (see Section 4.3) using linear regression.
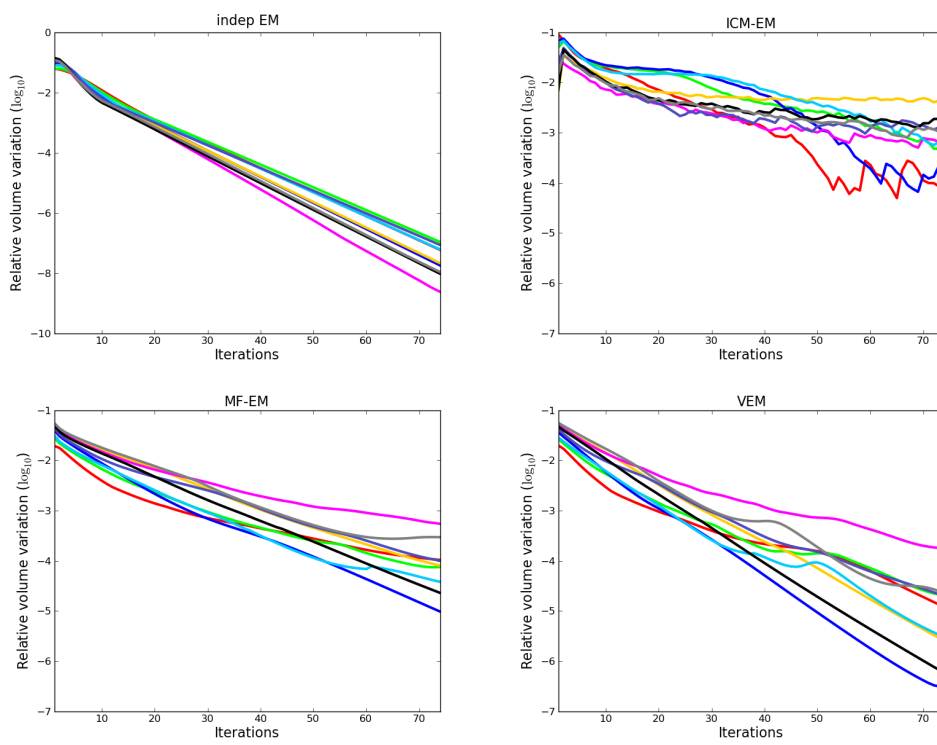
Figure 8: Real data: convergence speed study. The maximum relative volume variations are plotted in log scale $\log_{10}(\varepsilon_V)$ against iterations, for another 9 randomly selected subjects (each color represents a different subject).

5. Discussion

Our results show significant differences in convergence performances between approximate inference algorithms for MRF-based image segmentation. The VEM and MF-EM schemes were found to yield very similar results on both artificial and real datasets, with the former generally converging faster to the solution. For instance, in real data, VEM required about 25% iterations less than MF-EM to reach a 10^{-3} relative tissue volume variation. Both schemes achieved significantly faster convergence than the ICM-EM scheme implemented in neuroimaging software such as FSL and VBM8, which also systematically found solutions of higher free energy (5) than the other two schemes.

We observed a striking match between algorithm rankings regarding final free energy values and segmentation quality, as evaluated both by visual inspection and tissue overlap measures in the case of artificial data. Therefore, free energy appears as a reliable blind indicator of segmentation quality, although it is a relative measure which should not be compared across images. This observation provides additional support for the VEM algorithm which explicitly considers free energy as its objective function.

In real data, the classical indep EM algorithm was found to converge in fewer iterations than any of the MRF-based schemes (with each iteration requiring roughly half computation time), but tended to overestimate CSF and produce noisy segmentations. In our experience, initializing MRF-based segmentation from the outcome of indep EM does not significantly speed up convergence, nor does it warrant a better segmentation result. Note, however, that we did not present the indep EM variant with adjustable tissue proportions, which tends to have less bias towards CSF.

There are no significant differences in timing per iteration between the VEM, MF-EM and ICM-EM algorithms, meaning that the computational performance is determined by the number of iterations in all three approaches. As suggested in Section 3, it takes little efforts to switch from one implementation to the other as they only differ by subtle technical aspects. Last but not least, the VEM algorithm enjoys a general convergence result unlike both its competitors, and provides at least as accurate segmentation results. For all these reasons, we recommend VEM as the default algorithm to use in practice. There are, however, some remaining issues to be tackled in the future.

One of them is that VEM is dependent upon voxel space traversal, which is a rather unusual property for an image processing method. While this does not affect theoretical convergence, the extent to which it may affect both convergence speed and segmentation accuracy is still unclear. There might be better traversal strategies than others in this regard. Moreover, asynchronous voxel updating makes a parallel implementation non trivial. This suggests that there is still room for improvement.

Moreover, a limitation of VEM is that it is not guaranteed to find the global minimum of free energy. From practical experience, we consider that the algorithm works "well" in that segmentation results are visually appealing provided that the tissue class parameters are correctly initialized. Fairly accurate parameter initialization is easy to achieve in practice. However, there is no evidence as to how frequently the algorithm converges to the global free energy minimizer, even when accurately initialized. This could be a motivation for considering alternative inference algorithms.

The general VEM algorithm turns out to be one case of an even more general family of "message passing" algorithms described by Minka (2005). This family also includes the classical sum-product belief propagation algorithm (Weiss, 2001; Bishop, 2006), whose "max-product" version has been used previously for hard image segmentation via the MAP (Weiss and Freeman, 2001; Szeliski et al., 2008). A comprehensive comparison of message passing algorithms in approximate inference of MR tissue classification models would be useful to assess global convergence performances. At this stage, however, the VEM algorithm proved empirically accurate in addition to being numerically stable and presenting computational performance compatible with clinical applications.

References

Ashburner, J., Friston, K., 2005. Unified segmentation. NeuroImage 26 (3), 839–851.

Aubert-Broche, B., Evans, A., Collins, L., 2006. A new improved version of the realistic digital brain phantom. NeuroImage 32 (1), 138–145.

Bach Cuadra, M., Cammoun, L., Butz, T., Cuisenaire, O., Thiran, J.-P., 2005. Comparison and validation of tissue modelization and statistical classification methods in T1-weighted MR brain images. IEEE Transactions on Medical Imaging 24 (12), 1548–1565.

Besag, J., 1974. Spatial interaction and the statistical analysis of lattice systems. Journal of the Royal Statistical Society: Series B 36 (2), 192–236.

Bezdek, J. C., Hathaway, R. J., 2002. Some Notes on Alternating Optimization. In: Advances in Soft Computing – AFSS International Conference on Fuzzy Systems. Vol. 2275 of Lecture Notes in Computer Science. pp. 288–300.

Bishop, C., 2006. Pattern Recognition and Machine Intelligence. John Wiley & Sons, Inc.

Boykov, Y., Veksler, O., Zabih, R., 2001. Fast approximate energy minimization via graph cuts. IEEE Transactions on Pattern Analysis and Machine Intelligence 23 (11), 1222–1239.

Collins, D. L., Zijdenbos, A. P., Kollokian, V., Sled, J. G., Kabani, N. J., Holmes, C. J., Evans, A. C., Jun. 1998. Design and construction of a realistic digital brain phantom. IEEE Transactions on Medical Imaging 17 (3), 463–468.

Csiszár, I., Tusnády, G., 1984. Information Geometry and Alternating Minimization Procedures. Statistics & Decisions 1, 205–237, supplement Issue.

Dempster, A. P., Laird, N. M., Rubin, D. B., 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society: Series B 39, 1–38.

Fessler, J., 1998. On the Convergence of Mean Field Procedures for MRFs. IEEE Transactions on Image Processing 7 (6), 917.

Forbes, F., Fort, G., 2007. Combining Monte Carlo and Mean-Field-Like Methods for Inference in Hidden Markov Random Fields. IEEE Transactions on Image Processing 16 (3), 824–837.

Geman, S., Geman, D., 1984. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. IEEE Transactions on Pattern Analysis and Machine Intelligence 6 (6), 721–741.

Kapur, T., 1999. Model based three dimensional Medical Image Segmentation. Ph.D. thesis, Massachusetts Institute of Technology.

Kwan, R.-S., Evans, A., Pike, G., 1999. MRI simulation-based evaluation of image-processing and classification methods. IEEE Transactions on Medical Imaging 18 (11), 1085–1097.

Li, S. Z., 1995. Markov random field modeling in computer vision. Springer, Tokyo, New-York, Berlin.

Liang, Z., Wang, S., 2009. An EM Approach to MAP Solution of Segmenting Tissue Mixtures: A Numerical Analysis. IEEE Transactions on Medical Imaging 28 (2), 297–310.

Marroquin, J., Santana, E., Botello, S., 2003. Hidden Markov measure field models for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence 25 (11), 1380–1387.

Minka, T. P., Dec. 2005. Divergence measures and message passing. Tech. Rep. MSR-TR-2005-173, Microsoft Research Ltd., Cambridge, UK.

Narayana, P., Brey, W., Kulkarni, M., Sievenpiper, C., 1988. Compensation for surface coil sensitivity variation in magnetic resonance imaging. Magnetic Resonance Imaging 6, 271–274.

Neal, R., Hinton, G., 1998. A view of the EM algorithm that justifies incremental, sparse, and other variants. In: Jordan, M. (Ed.), Learning in Graphical Models. Kluwer Academic Publishers, pp. 355–368.

Penny, W., Trujillo-Bareto, N., Friston, K., 2005. Bayesian fMRI time series analysis with spatial priors. NeuroImage 24 (2), 350–362.

Prasad, R., Srinivas, R., Srinivas, Y., 2007. Unsupervised Image Segmentation Method based on Finite Generalized Gaussian Distribution with EM & K-Means Algorithm. International Journal of Computer Science and Network Security 7 (4), 317–321.

Prima, S., Ayache, N., Barrick, T., Roberts, N., 2001. Maximum Likelihood Estimation of the Bias Field in MR Brain Images: Investigating Different Modelings of the Imaging Process. In: Medical Image Computing And Computer-Assisted Intervention (MICCAI). Vol. 2208 of LNCS. Springer Verlag, pp. 811–819.

Schmitt, F., 1985. Correction of geometric distortion in MR Images. In: Proceedings of the computer assisted radiology (CAR). Springer Verlag, pp. 15–25.

Shattuck, D., Sandor-Leahy, S., Schaper, K., Rottenberg, D., Leahy, R., 2001. Magnetic resonance image tissue classification using a partial volume model. NeuroImage 13 (5), 856–876.

Sled, J., Zijdenbos, A., Evans, A., 1998. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. IEEE Transactions on Medical Imaging 17, 87–97.

Szeliski, R., Zabih, R., Scharstein, D., Veksler, O., Kolmogorov, V., Agarwala, A., Tappen, M., Rother, C., 2008. A comparative study of energy minimization methods for Markov random fields with smoothness-based priors. IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (6), 1068–1080.

Van Leemput, K., Maes, F., Vandermeulen, D., Suetens, P., Oct. 1999a. Automated model-based bias field correction of MR images of the brain. IEEE Transactions on Medical Imaging 18 (10), 885–896.

Van Leemput, K., Maes, F., Vandermeulen, D., Suetens, P., Oct. 1999b. Automated model-based tissue classification of MR images of the brain. IEEE Transactions on Medical Imaging 18 (10), 897–908.

Van Leemput, K., Maes, F., Vandermeulen, D., Suetens, P., 2003. A Unifying Framework for Partial Volume Segmentation of Brain MR Images. IEEE Transactions on Medical Imaging 22 (1), 105–119.

Wainwright, M., Jaakkola, T., Willsky, A., 2005. MAP estimation via agreement on (hyper)trees: Message-passing and linear-programming approaches. IEEE Transactions on Information Theory 51 (11), 3697–3717.

Wald, L., Schmitt, F., Dale, A., 2001. Systematic spatial distortion in MRI due to gradient nonlinearities. In: Proceedings of the 7th Annual Human Brain Mapping Meeting. Brighton, UK, p. 203.

Weiss, Y., 2001. Advanced Mean Field Methods. MIT Press, Ch. Comparing the mean field method and belief propagation for approximate inference in MRFs, pp. 229–239.

Weiss, Y., Freeman, W., 2001. On the optimality of solutions of the max-product belief-propagation algorithm in arbitrary graphs. IEEE Transactions on Information Theory 47 (2), 723–735.

Wells, W. M., Grimson, W. E. L., Kikinis, R., Jolesz, F. A., Aug. 1996. Adaptive Segmentation of MRI Data. IEEE Transactions on Medical Imaging 15 (4), 429–442.

Woolrich, M., Behrens, T., 2006. Variational Bayes Inference of Spatial Mixture Models for Segmentation. IEEE Transactions on Medical Imaging 25 (10), 1380–1391.

Zhang, J., 1992. The Mean Field Theory in EM Procedures for Markov Random Fields. IEEE Transactions on Signal Processing 40 (10), 2570–2583.

Zhang, J., 1996. The Convergence of Mean-Field Procedures for MRFs. IEEE Transactions on Image Processing 5 (12), 1662–1665.

Zhang, Y., Brady, J., Smith, S., 2001. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. IEEE Transactions on Medical Imaging 20 (1), 45–57.

15