Variational Inference for Image Segmentation

Claudia Blaiotta\(^1\), M. Jorge Cardoso\(^2\), and John Ashburner\(^1\)

\(^1\) Wellcome Trust Centre for Neuroimaging, University College London, London, UK
\(^2\) Centre for Medical Image Computing, University College London, London, UK

**Abstract.** Variational inference techniques are powerful methods for learning probabilistic models and provide significant advantages with respect to maximum likelihood (ML) or maximum a posteriori (MAP) approaches. Nevertheless they have not yet been fully exploited for image processing applications. In this paper we present a variational Bayes (VB) approach for image segmentation. We aim to show that VB provides a framework for generalizing existing segmentation algorithms that rely on an expectation-maximization formulation, while increasing their robustness and computational stability. We also show how optimal model complexity can be automatically determined in a variational setting, as opposed to ML frameworks which are intrinsically prone to overfitting.

**Keywords:** Variational Bayes, Gaussian Mixture Model, Image Segmentation

1 Introduction

Many of the most widely used image segmentation algorithms rely on probabilistic modelling techniques to fit the intensity distributions of images. Nevertheless, the majority of currently available tools are based on maximum likelihood (ML) or maximum a posteriori (MAP) estimation of the model parameters, without recourse to a fully Bayesian formulation. In fact, the latter approach has been poorly explored in the field of medical imaging, in spite of a promising potential which was shown for example in [1–3].

A fully Bayesian treatment of probabilistic models involves the introduction of prior distributions over all unobserved variables \(Z\) (model parameters and latent variables) and the evaluation of the posterior distribution \(p(Z|X)\), as well as of the model evidence \(p(X)\). Very often and also for relatively simple models, computing the full posterior distributions over the hidden variables turns out to be infeasible. In such cases, variational techniques represent a computationally effective way of evaluating approximated solutions to the inference problem by postulating a specific form, or factorization, of the posterior distributions. As a result, the estimated posteriors will never be exact. Nevertheless variational methods have proved to be more convenient than standard ML or MAP techniques, since, at a substantially similar computational cost, they avoid the problems related to overfitting which are intrinsic to the other methods. In other words, variational techniques open up the possibility of learning the optimal model structure (the one with highest generalization capability) without
performing ad-hoc cross validation analyses [4–6]. Another interesting aspect of working within a VB framework is that it leads to a more general formulation of the EM algorithm, which has the same convergence properties and higher computational stability. In fact, one important limitation of the ML formulation is the presence of singular points of the likelihood function, which have to be avoided during the optimization process to assure numerical stability.

A convenient way of restricting the space of the approximating posterior distribution \( q(Z) \) consists in assuming that it factorizes into a product of terms, each one involving just a subset of \( Z \) (mean field theory): 
\[
q(Z) = \prod_{s=1}^{S} q_s(Z_s).
\]

The optimal solution for the different factors \( \{q_s\}_{s=1,...,S} \) can be obtained by maximizing a functional that defines a lower bound on the log marginal probability \( \log p(X) \). Such a lower bound can be derived from the following decomposition which holds for any \( q(Z) \)
\[
\log p(X) = \int q(Z) \log \left\{ \frac{p(X,Z)}{q(Z)} \right\} dZ + \int q(Z) \log \left\{ \frac{q(Z)}{p(Z|X)} \right\} dZ. \tag{1}
\]

The first integral in (1) defines a lower bound \( L(q) \) on the logarithm of the model evidence while the second integral is the Kullback-Leibler divergence \( D_{KL}(q\|p) \) between the variational approximating posterior and the true posterior distribution. It can be proved that the optimal form of each factor \( q_s(Z_s) \) is
\[
q_s(Z_s) \propto \exp(\mathbb{E}_{s \neq \hat{s}}[\log p(X, Z)]) . \tag{2}
\]

Equation (2) is not an analytical solution anyway, since the different factors have optimal forms that depend on one another. As a result, the natural approach for solving this variational optimization problem consist in iteratively updating each factor given the most recent forms of the other ones. This leads to a scheme that turns out to be very similar to the structure of the EM algorithm [6].

For some complex models, a fully Bayesian treatment of all unobserved variables might still be extremely impractical, if not impossible, even when variational techniques are used. Anyway, in this case it is still possible to obtain MAP point estimates for the parameters which are not treated in a fully Bayesian manner. Such values are obtained in a way that is a generalization of the M-step in the EM algorithm. In particular, the function that needs to be optimized is the expectation of the logarithm of the joint probability of \( X \) and \( Z \), 
\[
\mathbb{E}[\log p(X, Z)].
\]

In this paper we show that VB techniques can be effectively applied to solve the problem of image segmentation, providing extra robustness, stability and flexibility over existing methods, but without significantly affecting the computational cost. Moreover we demonstrate that a fully Bayesian formulation of image segmentation can be exploited for automatically determining the optimal model complexity. The framework described in the following sections is built on the segmentation method implemented in the SPM12 \(^1\) software (Wellcome Trust Centre for Neuroimaging, UCL, London, UK) and our aim is to extend this widely used method by incorporating priors on tissue intensities, so as to

\(^1\) http://www.fil.ion.ucl.ac.uk/spm
increase its robustness. In fact our future work will focus on learning, from large data sets, informative intensity priors that could be used in combination with the proposed method.

2 Data Model

Let $X$ denote the observed data, that is to say the intensities corresponding to $D$ images of the same subject acquired with different modalities. The signal at voxel $j$ can then be represented by a $D$-dimensional vector $x_j \in \mathbb{R}^D$, with $j \in \{1, \ldots, N\}$.

We model the distribution of $x_j$ as a multivariate Gaussian mixture consisting of $K$ clusters parametrized by mean vectors $\{\mu_k\}_{k=1, \ldots, K}$ and covariance matrices $\{\Sigma_k\}_{k=1, \ldots, K}$. The mixing proportions of the different components are given by $\Theta = \{\pi_{jk}\}$ with $\pi_{jk} \in [0, 1]$ and $\sum_k \pi_{jk} = 1$. Essentially $\pi_{jk}$ indicates the prior probability of signal at spatial location $j$ being drawn from cluster $k$.

Moreover we assume that the $K$ Gaussians can be partitioned into $T$ subsets, corresponding to different tissue types. We denote these subsets by $\{C_t\}_{t=1, \ldots, T}$, with $\bigcup_{t=1}^{T} C_t = \{1, \ldots, K\}$. This means that each tissue $t \in \{1, \ldots, T\}$ is itself represented by a GMM consisting of $K_t$ components with $\sum_{t=1}^{T} K_t = K$.

We consider the prior probability of voxel $j$ belonging to tissue $t$ to be given a priori (through a probabilistic atlas) and we indicate this probability as $\tau_{jt}$. Furthermore we allow these tissue priors to be rescaled by a set of weights $\{w_t\}_{t=1, \ldots, T}$ to accommodate for individual differences in tissue composition.

Finally we introduce a set of parameters $\{g_k\}_{k=1, \ldots, K}$ denoting the normalized weights of the different Gaussians associated with one tissue type, so that

$$\forall t \in \{1, \ldots, T\} : \sum_{k \in C_t} g_k = 1. \quad (3)$$

As a result we can express the mixing proportions $\Theta$ of our GMM as

$$\pi_{jk} = g_k \cdot \frac{\tau_{jt} \cdot w_t}{\sum_{t'} \tau_{jt'} \cdot w_{t'}}, \quad (4)$$

where $\{\tau_{jk}\}$ are known parameters, while $\{w_t\}_{t=1, \ldots, T}$ and $\{g_k\}_{k=1, \ldots, K}$ have to be estimated from the observed data $X$.

To correct for intensity non-uniformity artifacts, we introduce a multiplicative $D$-dimensional bias field denoted by $\{b_j(\Theta_\beta)\}_{j=1, \ldots, N}$, where $\Theta_\beta$ is a vector of parameters. Each of the $D$ components of the bias is modelled as the exponential of a linear combination of discrete cosine transform basis functions.

Finally, to account for the variability of anatomical shapes among subjects, we allow the probabilistic atlas given by $\{\tau_t\}_{t=1, \ldots, T}$ to be deformed according to a displacement field parametrized by the set of vectors $\Theta_\alpha = \{\alpha_j\}_{j=1, \ldots, N}$. The warped tissue priors can therefore be expressed as $\{\tau_t(\varphi(\Theta_\alpha))\}_{t=1, \ldots, T}$, where $\varphi(\Theta_\alpha)$ is a coordinate mapping from the individual image space into the atlas space. The parametrization adopted here consists in adding to the identity
transform a small displacement field \( \{ \alpha_j \}_{j=1,...,N} \), so that \( \tilde{y}_j = y_j + \alpha_j \), where the vector \( y_j \) encodes the coordinates of the centre of voxel \( j \).

If we introduce a set of binary latent variables \( Z \) denoting the class memberships of the observed data \( X \), the probability of \( Z \) given the mixing proportions \( \Theta_\pi \) and the deformation parameters \( \Theta_\alpha \) is equal to

\[
p(Z|\Theta_\pi,\Theta_\alpha) = \prod_{j=1}^N \prod_{k=1}^K (\pi_{jk}(\varphi(\Theta_\alpha)))^{z_{jk}},
\]

where we have assumed that all data are independent.

The conditional distribution of the observed intensities given the latent variables, the Gaussian means \( \Theta_\mu \) and covariances \( \Theta_\Sigma \) and the bias field parameters \( \Theta_\beta \) can be expressed as follows, with \( B_j = \text{diag}(b_j) \), modelling the bias field

\[
p(X|Z,\Theta_\mu,\Theta_\Sigma,\Theta_\beta) = \prod_{j=1}^N \prod_{k=1}^K (\det(B_j) N(B_j x_j|\mu_k,\Sigma_k))^{z_{jk}},
\]

The joint probability of all the random variables conditioned on the mixing proportions is given, for our model, by

\[
p(X, Z, \Theta_\mu, \Theta_\Sigma, \Theta_\beta, \Theta_\alpha|\Theta_\pi) = p(X|Z,\Theta_\mu,\Theta_\Sigma,\Theta_\beta)p(Z|\Theta_\pi,\Theta_\alpha)p(\Theta_\alpha)p(\Theta_\beta).
\]

The voxel specific mixing proportions \( \Theta_\pi \) are treated here as deterministic parameters depending on the available anatomical atlas, on the tissue weights \( \{ w_t \} \) and on the within-tissue mixing proportions \( \{ g_k \} \), therefore they are determined via ML estimation. The priors on the means and covariances of the different classes are modelled as Gaussian-Wishart distributions

\[
p(\Theta_\mu, \Theta_\Sigma) = \prod_{k=1}^K N(\mu_k|m_{0k},b_{0k}^{-1}\Sigma_k)W(\Sigma_k^{-1}|W_{0k},\nu_{0k}),
\]

The terms \( p(\Theta_\alpha) \) and \( p(\Theta_\beta) \) represent prior probability distributions over the deformation and bias field parameters. Their function is to regularize the solution obtained through model fitting by penalizing improbable parameters values. In doing so, they assure greater physical plausibility of the resulting non-uniformity and deformation fields, while also improving numerical stability within the optimization process. Here we adopt the same regularization scheme described in [7]. The question of how to determine optimal forms for the regularization terms is beyond the scope of this work and therefore is not addressed here. Interestingly, such a problem could also be solved in a variational inference framework, as shown in [8].

A lower bound on the marginal likelihood \( p(X, \Theta_\alpha, \Theta_\beta|\Theta_\pi) \) is given by

\[
\mathcal{L} = \sum_Z \int \int q(Z,\Theta_\mu,\Theta_\Sigma) \log \left( \frac{p(X, Z, \Theta_\mu, \Theta_\Sigma, \Theta_\beta, \Theta_\alpha|\Theta_\pi)}{q(Z,\Theta_\mu,\Theta_\Sigma)} \right) d\Theta_\mu d\Theta_\Sigma.
\]
To make the problem tractable we assume that the variational distribution \( q(Z, \Theta_\mu, \Theta_\Sigma) \) factorizes as
\[
q(Z, \Theta_\mu, \Theta_\Sigma) = q(Z)q(\Theta_\mu, \Theta_\Sigma).
\]
A graphical representation of the described generative model is provided in figure 1.

3 Model Learning

The statistical model described in Section 2 can be fit to data adopting a variational version of the standard EM algorithm for MLE. The objective of this optimization procedure is to learn optimal solutions for the variational posterior distribution \( q(Z)q(\Theta_\mu, \Theta_\Sigma) \), to estimate MAP values for the parameters \( \{\Theta_\alpha, \Theta_\beta\} \) and ML values for \( \{g_k\} \) and \( \{w_t\} \). Finally, inferring tissue labels can be done determining, voxel by voxel, the tissue type that has the highest posterior probability of having generated the data.

3.1 Variational E-step

In the variational generalization of the EM algorithm (VBEM) we can still distinguish two steps: VE-step and VM-step. In the variational E-step the functional \( \mathcal{L} \) in (9) is maximized with respect to the posterior factor \( q(Z) \) over the latent variables \([6]\). Making use of (2) we find that
\[
q(Z) \propto \exp \left( \log p(Z|\Theta_\pi, \Theta_\alpha) + \mathbb{E}_{\mu, \Sigma} [\log p(X|Z, \Theta_\mu, \Theta_\Sigma, \Theta_\beta)] \right). \tag{10}
\]
If we define
\[
\log \rho_{jk} = \log p(Z|\Theta_\pi, \Theta_\alpha) + \mathbb{E}_{\mu, \Sigma} [\log p(X|Z, \Theta_\mu, \Theta_\Sigma, \Theta_\beta)], \tag{11}
\]
it follows that
\[
q(Z) = \prod_{j=1}^{N} \prod_{k=1}^{K} \left( \frac{\rho_{jk}}{\sum_{c=1}^{K} \rho_{jc}} \right)^{z_{jk}} = \prod_{j=1}^{N} \prod_{k=1}^{K} (\gamma_{jk})^{z_{jk}}. \tag{12}
\]
The quantity \( \log \rho_{jk} \) can be computed from (11) to give
\[
\log \rho_{jk} = \log \pi_{jk}(\varphi(\Theta_\alpha)) - \frac{D}{2} \log(2\pi) + \frac{1}{2} \log \left( \frac{1}{\| \Sigma_k \|} \right) - \frac{1}{2} \mathbb{E}_{\mu_k, \Sigma_k} \left[ (B_j x_j - \mu_k)^T \Sigma_k^{-1} (B_j x_j - \mu_k) \right].
\]
(13)
The terms \( \{ \gamma_{jk} \} \) can then be used to compute the following sufficient statistics of the observed data, which will serve to update the posterior distributions of \( \{ \mu_k \}_{k=1, \ldots, K} \) and \( \{ \Sigma_k \}_{k=1, \ldots, K} \), during the VM-step
\[
s_{0k} = \sum_{j=1}^N \gamma_{jk}, \quad s_{1k} = \sum_{j=1}^N \gamma_{jk} B_j x_j, \quad S_{2k} = \sum_{j=1}^N \gamma_{jk} (B_j x_j)(B_j x_j)^T.
\]
(14)

### 3.2 Variational M-step

In the following VM-step we can derive approximate solutions for the posterior distributions over the cluster means and covariance matrices [6]. Making again use of (2) we obtain
\[
q(\Theta_\mu, \Theta_\Sigma) \propto \exp \left( \sum_{j=1}^N \sum_{k=1}^K \gamma_{jk} \log \mathcal{N}(B_j x_j | \mu_k, \Sigma_k) + \sum_{k=1}^K \log p(\Theta_\mu, \Theta_\Sigma) \right).
\]
(15)

It can be proved that the posterior distributions on the means and covariances of the different Gaussians take the same form as the corresponding priors [6], that is
\[
q(\Theta_\mu, \Theta_\Sigma) = \prod_{k=1}^K \mathcal{N}(\mu_k | m_k, b_k^{-1} \Sigma_k)W(\Sigma_k^{-1} | W_k, \nu_k),
\]
(16)
The parameters that govern these posterior distributions can be computed as a function of the prior hyperparameters and the sufficient statistics obtained in the previous VE-step, as follows
\[
b_k = b_{0k} + s_{0k}, \quad m_k = \frac{b_{0k} m_{0k} + s_{1k}}{b_{0k} + s_{0k}},
\]
\[
W_k^{-1} = W_{0k}^{-1} + S_{2k} + \frac{b_{0k} s_{0k} m_{0k} m_{0k}^T}{b_{0k} + s_{0k}} - \frac{s_{1k} s_{1k}^T}{b_{0k} + s_{0k}} - \frac{b_{0k} s_{1k} m_{0k}^T}{b_{0k} + s_{0k}} - \frac{b_{0k} m_{0k} s_{1k}^T}{b_{0k} + s_{0k}},
\]
\[
\nu_k = \nu_{0k} + s_{0k} + 1.
\]
(17)
The point estimates of the mixing proportions \( \{ g_k \}_{k=1, \ldots, K} \) and of the tissue weights \( \{ w_t \}_{t=1, \ldots, T} \) can instead be updated by
\[
g_k = \frac{s_{0k}}{\sum_{c \in C_t} s_{0c}}, \quad w_t = \frac{\sum_{k \in C_t} s_{0k} \tau_{jt}(\varphi(\Theta_\alpha))}{\sum_{j=1}^T \sum_{t'=1}^T \tau_{jt'}(\varphi(\Theta_\alpha))w_{t'}}.
\]
(18)
3.3 Computing the Lower Bound

The lower bound can be easily evaluated, once the sufficient statistics and the variational posterior distributions have been computed [6], by

\[
\mathcal{L} = \mathbb{E}_{Z, \Theta, \theta_{\Sigma}} \left[ \log p(X|Z, \Theta_{\mu}, \Theta_{\Sigma}, \Theta_{\beta}) \right] + \mathbb{E}_{Z} \left[ \log p(Z|\Theta_{\pi}, \Theta_{\alpha}) \right] \\
+ \mathbb{E}_{\Theta, \theta_{\Sigma}} \left[ \log p(\Theta_{\mu}, \Theta_{\Sigma}) \right] + \log p(\Theta_{\alpha}) + \log p(\Theta_{\beta}) \\
- \mathbb{E}_{Z} \left[ \log q(Z) \right] - \mathbb{E}_{\Theta, \theta_{\Sigma}} \left[ \log q(\Theta_{\mu}, \Theta_{\Sigma}) \right],
\]

(19)

with

\[
\mathbb{E}_{Z, \Theta, \theta_{\Sigma}} \left[ \log p(X|Z, \Theta_{\mu}, \Theta_{\Sigma}, \Theta_{\beta}) \right] = \\
\frac{1}{2} \sum_{k=1}^{K} s_{0k} \left( \mathbb{E} \left[ \log |\Sigma_{k}^{-1}| \right] - D \log(2\pi) - \frac{D}{\beta_k} \right) - \frac{1}{2} \sum_{k=1}^{K} s_{0k} \left( \nu_k m_k^T W_k m_k \right) \\
- \frac{1}{2} \sum_{k=1}^{K} \nu_k \left( \text{Tr}(W_k S_{2k} - 2s_{1k} m_k^T W_k) \right) + \sum_{j=1}^{N} \sum_{k=1}^{K} \gamma_{jk} \log |B_j|,
\]

(20)

\[
\mathbb{E}_{Z} \left[ \log p(Z|\Theta_{\pi}, \Theta_{\alpha}) \right] = \sum_{j=1}^{N} \sum_{k=1}^{K} \gamma_{jk} \log \pi_{jk}(\varphi(\Theta_{\alpha})).
\]

(21)

\[
\mathbb{E}_{\Theta, \theta_{\Sigma}} \left[ \log p(\Theta_{\mu}, \Theta_{\Sigma}) \right] = \\
\frac{1}{2} \sum_{k=1}^{K} D \log \frac{\beta_{0k}}{2\pi} - \frac{D \beta_{0k}}{\beta_k} + 2 \log BW(W_{0k}, \nu_{0k}) + \frac{1}{2} \sum_{k=1}^{K} \mathbb{E} \left[ \log |\Sigma_{k}^{-1}| \right] (\nu_{0k} - D) \\
- \nu_k \text{Tr}(W_{0k}^{-1} W_k + \beta_{0k}(m_k - m_{0k})(m_k - m_{0k})^T W_k) .
\]

(22)

\[
\mathbb{E}_{Z} \left[ \log q(Z) \right] = \sum_{j=1}^{N} \sum_{k=1}^{K} \gamma_{jk} \log \gamma_{jk}.
\]

(23)

\[
\mathbb{E}_{\Theta, \theta_{\Sigma}} \left[ \log q(\Theta_{\mu}, \Theta_{\Sigma}) \right] = \\
\mathbb{E} \left[ \log |\Sigma_{k}^{-1}| \right] \left( \frac{1}{2} \nu_k - D \right) + \frac{1}{2} D \left( \log \frac{\beta_k}{2\pi} - 1 - \nu_k \right) + \log BW(W_k, \nu_k).
\]

(24)

The term \( BW(W, \nu) \) in equations (22) and (24) indicates the normalizing constant for a Wishart distribution parametrized by \( W \) and \( \nu \).

3.4 Estimating Bias Field and Deformations

In order to estimate optimal parameters to represent the bias field we need, at each iteration of the algorithm, to maximize the lower bound (19) on the objective function with respect to the parameters \( \Theta_{\beta} \). A closed form solution does
not exist in this case; therefore, recourse to numerical optimization techniques cannot be avoided. The optimization problem can be formulated as follows

$$\hat{\Theta}_\beta = \arg \max_{\Theta_\beta} \{ \mathbb{E}_{Z, \Theta_\mu, \Theta_\Sigma} [\log p(X|Z, \Theta_\mu, \Theta_\Sigma, \Theta_\beta)] + \log p(\Theta_\beta) \} .$$  \hspace{1cm} (25)$$

Similarly, the deformation field that best matches the population based atlas \(\{\tau_t\}_{t=1,...,T}\) to an individual image can be estimated solving the following optimization problem

$$\hat{\Theta}_\alpha = \arg \max_{\Theta_\alpha} \{ \mathbb{E}_Z[\log p(Z|\Theta_\pi, \Theta_\alpha)] + \log p(\Theta_\alpha) \} .$$  \hspace{1cm} (26)$$

In both cases we adopt a Gauss-Newton optimization scheme.

4 Experimental Results

4.1 Experiments on Synthetic Data

The accuracy of the presented variational algorithm was first assessed making use of synthetic data produced by the Brainweb MRI simulator [9]. In particular, synthetic T1- and T2-weighted scans were generated from a healthy adult brain model, with the parameters indicated in table 1

Table 1: Simulation parameters selected to generate the synthetic data which was used to evaluate the accuracy of the presented VB algorithm. A bias intensity of 20% corresponds to values of the non uniformity field in the range \([0.9, 1.1]\).

<table>
<thead>
<tr>
<th>Sequence</th>
<th>TR (ms)</th>
<th>Flip angle (deg)</th>
<th>TE (ms)</th>
<th>Bias field</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1w</td>
<td>SFLASH</td>
<td>18</td>
<td>30</td>
<td>10</td>
</tr>
<tr>
<td>T2w</td>
<td>DSE,LATE</td>
<td>3300</td>
<td>90</td>
<td>35,120</td>
</tr>
</tbody>
</table>

The images were first segmented using the method described in this paper; the resulting gray matter and white matter segmentations were then compared to the ground truth provided by the underlying model of the data. The similarity between ground truth and estimated segmentations was quantified by computing Dice similarity coefficients. To examine the behaviour of the algorithm with respect to noise, the same analyses were repeated for three different levels of noise in the data (3%, 5% and 9% of the brightest intensity). For this experiment we used the probabilistic atlas provided by the SPM12 software which includes gray matter, white matter, CSF, bone, soft tissue and air. Moreover we set the following hyperparameter values (weakly informative priors)

$$\beta_0 = 1, \ m_0 = \frac{\sum_{j=1}^N x_j}{N}, \ \nu_0 = D, \ W_0^{-1} = \frac{\sum_{j=1}^N (x_j - m_0)(x_j - m_0)^T}{N}.$$  \hspace{1cm} (27)$$

In addition, equivalent accuracy measures were computed on the segmentations obtained processing the same data with the Maximum Likelihood algorithm implemented in SPM12. Results are shown in table 2 and indicate that
our method provides very accurate segmentations while never performing worse than the SPM ML approach.

Table 2: Dice similarity coefficients between the estimated segmentations (of gray and white matter) and the ground truth labels, for our VB algorithm and for SPM ML approach.

<table>
<thead>
<tr>
<th>Noise level</th>
<th>VB</th>
<th>ML</th>
<th>VB</th>
<th>ML</th>
<th>VB</th>
<th>ML</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tissue</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GM</td>
<td>0.91</td>
<td>0.90</td>
<td>0.91</td>
<td>0.89</td>
<td>0.87</td>
<td>0.87</td>
</tr>
<tr>
<td>WM</td>
<td>0.95</td>
<td>0.94</td>
<td>0.94</td>
<td>0.94</td>
<td>0.89</td>
<td>0.89</td>
</tr>
</tbody>
</table>

For this simulated data we could also evaluate the accuracy of the presented algorithm with respect to bias field correction. To do so we computed Pearson’s correlation coefficients between the estimated non-uniformity field and the ground truth. Results are shown in table 3, where we report as well the correlation coefficients achieved by SPM segmentation algorithm. Both methods exhibit very good agreement between the estimated bias field and the “true” one, which was used to simulate the data.

Table 3: Pearson’s correlation coefficients between estimated and ground truth bias fields for the presented VB method and for SPM ML method.

<table>
<thead>
<tr>
<th>Noise level</th>
<th>VB</th>
<th>ML</th>
<th>VB</th>
<th>ML</th>
<th>VB</th>
<th>ML</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tissue</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T1w</td>
<td>0.83</td>
<td>0.83</td>
<td>0.84</td>
<td>0.82</td>
<td>0.68</td>
<td>0.61</td>
</tr>
<tr>
<td>T2w</td>
<td>0.88</td>
<td>0.88</td>
<td>0.89</td>
<td>0.89</td>
<td>0.88</td>
<td>0.70</td>
</tr>
</tbody>
</table>

Run time was 3 min 23 s for our algorithm and 3 min 26 s for SPM, on a Quad-Core PC at 3.19 GHz with 12 GB RAM.

4.2 Experiments on Real Data

To test the performance of the presented method on real data we used T1-, T2- and PD-weighted scans of 15 randomly selected subjects from the freely available IXI brain database. Images were all acquired with the same scanner, a Philips Medical Systems Gyroscan Intera 1.5 Tesla. We applied the algorithm described in Section 3 to segment such a multispectral data set with the same settings adopted for the previous experiments.

Unfortunately no ground truth is available for this data. Anyway we compared our results with those obtained, on the same images, with the widely used segmentation algorithm implemented in SPM12 and described in [7]. A quantitative assessment of the accuracy and reliability of such a method can be found in [10].

http://www.brain-development.org
Fig. 2: A T1-weighted image included in the IXI data set is shown in panel (a). The corresponding gray matter (GM) segmentations obtained with the ML method implemented in SPM12 and with our VB approach are illustrated in subfigures (b) and (c) respectively. In panel (d) we illustrate how the gray matter atlas, which is depicted in panel (e), is warped by our algorithm to match the individual image.

Fig. 3: (a) Dice similarity coefficients between the results produced by our algorithm and by the segmentation algorithm implemented in SPM12. The bars indicate the average scores for gray matter (GM) white matter (WM) and cerebrospinal fluid (CSF). Error bars indicate standard deviations. Scatter plots of the coefficients are also shown, superimposed to the bar plot. (b) Contour plot of the intensity distributions of gray matter, white matter, cerebrospinal fluid, bone and soft tissue obtained for one subject included in the IXI dataset, overlaid on the joint histogram of the T1- and T2-weighted images. The optimal number of components is determined automatically by our VB algorithm therefore irrelevant model components are not shown here.
For simple illustrative purposes we report in Figure 2 the gray matter segmentations obtained, for one of the selected subjects, with SPM ML approach (b) and with our VB method (c).

To quantitatively compare the two algorithms, we computed Dice similarity coefficients between the set of resulting gray matter, white matter and cerebrospinal fluid segmentations. Results are reported in Figure 3(a). We obtained very high values of the overlap measure, for all tissue classes (DSC $> 0.9$). This confirms our hypothesis that the presented VB framework could represent a powerful generalization of existing methods for image segmentation, such as the one provided with SPM12. The results presented in this section were obtained using weakly informative priors, therefore the high similarity between the performance of ML and VB approaches is to be expected, nonetheless it will be possible to achieve even greater robustness, compared to ML approaches, by learning informative priors of tissue intensities from available data sets, which will be part of our future work.

In addition, it should be noted that variational techniques have an inherent capability of avoiding overfitting, which represents a fundamental advantage over ML model learning [4,6]. In the case of mixture models, this allows for example to determine the most appropriate number of components ($K$) without performing cross-validation, which is usually rather demanding in terms of computations and amount of data required [6].

For the model presented here, if the number of Gaussians is set to a value that is higher than the optimal one, the redundant components will be automatically pruned out of the model [5] as their responsibilities $\gamma_{jk}$ are quickly driven to zero by the algorithm. This follows from the fact that the variational objective function contains an implicit penalty for complex models [4].

We illustrate such a behavior using the images of one subject in the IXI database. In particular we set 5 Gaussians for each of the tissue types of interest and, after convergence of the VBEM algorithm, we observed only 2 components surviving for gray matter, 1 for white matter, 3 for CSF, 2 for bone and 4 for soft tissues, as shown in Figure 3(b). In a similar setting a ML algorithm would have simply found the best fit to the data, making use of all the available components, but the optimal number of Gaussians would have had to be determined a priori, through some form of model comparison.

5 Discussion

Evaluating posterior probability distributions over model parameters and latent variables is often a very demanding task in the context of probabilistic modelling problems, especially when working with large-scale data sets. Unfortunately this is usually the case for image processing applications.

VB techniques belong to the family of deterministic approaches for finding approximate solutions to inference problems. Despite not providing exact results, they allow learning of fully Bayesian models without the computational drawbacks of sampling techniques. The resulting algorithms have the interesting
property of generalizing ML or MAP approaches, while automatically addressing the overfitting issues associated with ML estimation.

In this paper we have shown that VB represents a viable and effective framework for performing medical image segmentation, in spite of not having been exploited so far in such a field. We demonstrated the accuracy of our variational segmentation algorithm on synthetic data and we compared its performance on real data to that the widely used SPM image segmentation algorithm which we believe our method could be a powerful extension of. We have also provided evidence that the VBEM algorithm presented in Section 3 can be used to automatically determine the optimal number of components (model complexity) of the underlying GMM, without incurring overfitting as the standard EM approach. Further increases in the robustness of the framework presented here could be achieved by learning informative intensity priors from a population data set, a problem that we aim to address in our future work.

References