Variational inference for medical image segmentation

Claudia Blaiotta[a,*], M. Jorge Cardoso[b], John Ashburner[a]

[a] *Wellcome Trust Centre for Neuroimaging, University College London, London WC1N 3BG, UK*
[b] *Translational Imaging Group, CMIC, University College London, London WC1E 6BT, UK*

Abstract

Variational inference techniques are powerful methods for learning probabilistic models and provide significant advantages over 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. Finally, we demonstrate how suitable intensity priors, that can be used in combination with the presented algorithm, can be learned from large imaging data sets by adopting an empirical Bayes approach.

Keywords: Image segmentation, Bayesian inference, Variational Bayes, Neuroimaging, MRI

1. Introduction

When analysing neuroimaging data, it is often necessary or helpful to partition brain tissues into different types. This represents indeed the primary stage for performing brain volumetry, which is extremely valuable both in research and for clinical practice [17, 3]. In fact, quantifying brain structure volume not only has a major role for unraveling the mechanisms underlying neurodegenerative and psychiatric disorders, but can also significantly help in disease diagnosis and treatment planning or monitoring [30].

For healthy subjects the tissues of interest are typically gray matter, white matter and cerebrospinal fluid, while for patients, additional classes may be defined, such as tumor, edema or necrosis [35, 31]. In this framework, magnetic resonance imaging (MRI) is usually the most convenient imaging modality to work with, as it provides simultaneously high spatial resolution, excellent soft tissue contrast and good signal to noise ratio.

Many widely used image segmentation algorithms rely on probabilistic modelling techniques to fit the intensity distributions of images. These methods commonly operate by means of unsupervised clustering algorithms and assume that the data are drawn from mixture distributions, with different mixture components being associated to different tissue types. In particular, Gaussian mixture models (GMM) have been extensively adopted as they provide a flexible and computationally efficient framework that can be easily applied to solve the problem of automatically partitioning images into homogeneous regions [50, 44, 46, 19, 32, 18, 13, 31, 42].

Intensity based segmentation tools of this sort have been developed profusely over the past twenty years. Most of them either rely directly on an explicit Bayesian formulation, or exhibit an implicit probabilistic interpretation. Nevertheless almost all of them are based on maximum likelihood (ML) or maximum a posteriori (MAP) estimation of the model parameters [4, 27, 37, 28, 47, 50, 24, 43, 48, 18], without exploiting the potential of full Bayesian inference.

Indeed, the choice of ML or MAP techniques ensures mathematical tractability and sufficient segmentation accuracy for many applications. Nonetheless there is still a crucial theoretical point that makes these methods somehow suboptimal, regardless of their mathematical and computational convenience: the fact that they just provide point estimates of the model parameters instead of full posterior probability distributions. In other words, information is missing on the posterior uncertainty in estimating unobserved variables, and this often results in the occurrence of overfitting as well as in the difficulty to perform model comparison [5].

On the other hand, full Bayesian inference has been poorly explored in the field of medical image segmentation, in spite of a promising potential, which was shown for example by Woolrich and Behrens [45] and Tian et al. [40]. The reason for this is most probably related to the computational challenges that arise when trying to evaluate the model evidence or the posterior probability distributions over the model parameters. In fact, very often and also

*Corresponding author
Email address: claudia.blaiotta.13@ucl.ac.uk (Claudia Blaiotta)

for relatively simple models, integrating out all the unobserved variables turns out to be intractable in analytical form. On the other hand, numerical integration is generally impractical because either the dimensionality, or the complexity, of the problem would make the computational resources necessary to integrate over all possible parameter configurations too large for real world applications.

One approach for dealing with the mathematical difficulties that arise in Bayesian inference is to make use of stochastic techniques to sample from the probability distributions that are of interest [1]. In particular, Markov Chain Monte Carlo methods can provide rather accurate solutions at the expenses of a long processing time. As to be expected, the time required to reach convergence increases with the size of the data set. The result of this being the fact that, for large-scale problems, sampling techniques can become computationally prohibitive. The work of Iglesias et al. [21] is among the very few attempts to exploit MCMC sampling methods to integrate out model parameters in the context of Bayesian medical image segmentation. Their atlas based segmentation method takes into account the uncertainty in the estimates of the deformations that bring the individual images in alignment with the reference anatomical space. However, they report a running time of the sampling of approximately three hours, which might still be non-viable for some applications, even if this additional computational time ensures higher segmentation accuracy. Related work, on solving image segmentations problems making use of stochastic techniques, has been done also by Fan et al. [15], Kato [22] and da Silva [12].

A second family of approaches is based on introducing analytical approximations. For instance, one possibility is to approximate an unknown posterior probability distribution by a Gaussian, centered at the mode of the posterior, or at one of the modes, if the distribution is multimodal. Such a method, known as the Laplace approximation, overcomes many of the limitations of sampling techniques, since the number of required computations is much lower in this case. Nevertheless, depending on how different the actual posterior distribution is from a Gaussian, the method might provide a poor approximation. In particular the underlying Gaussian assumption might become inadequate for samples that are far from the mode of the density.

Variational Bayes (VB) represents an alternative way of obtaining approximate solutions to inference problems. It often relies on analytical approximations, as the Laplace method, and likewise it is much less computationally expensive than MCMC. However, the VB framework is more general and flexible than the Laplacian approach. In fact, even if for computational reasons it is often necessary to constrain the posterior distributions to have a specific form or factorization, they are not necessarily forced to be Gaussian. In other words, variational Bayesian inference permits finding a trade off between allowing sufficient complexity and accuracy of the estimated posteriors and ensuring computational tractability. Stochastic variational algorithms have also been proposed [20].

Even if the estimated posteriors will almost never be exact, variational methods have proved to be more convenient than standard ML or MAP techniques, since, at a substantially similar computational cost, they significantly alleviate 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 [5, 11, 8]. 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 significant limitation of the ML formulation (for mixture models) is the presence of singular points of the likelihood function, which have to be avoided during the optimization process to assure numerical stability.

So far, very few authors have explored the applicability of the variational Bayes framework to perform medical image segmentation. In particular, Woolrich and Behrens [45] exploited variational inference to fit spatial mixture models to medical imaging data while automatically tuning the parameter controlling spatial regularization. Tian et al. [40] proposed a variational algorithm for segmenting brain MRI data, which combines variational Bayes techniques with a genetic algorithm to initialize the priors on tissue intensities.

In this paper, we present an extension of the tissue classification algorithm presented by Ashburner and Friston [4] and publicly distributed as part of the SPM12 software. Specifically, we replace the maximum likelihood approach adopted in Ashburner and Friston [4] to estimate the Gaussian mixture parameters that model the distribution of image intensities, with a fully Bayesian inference scheme relying on variational approximations.

This greatly increases the robustness of the method if suitable intensity priors are introduced, thus reducing significantly the chance of the algorithm failing due to the mismatch or misregistration of the tissue probability maps with the individual scans. Additionally we demonstrate that, in principle, having tissue specific intensity priors yields fairly accurate segmentations also in a completely atlas- (and registration-) free setting.

Secondly, we illustrate how the fundamental problem of determining the optimal model complexity, i.e. the number of Gaussian components that are necessary to model the distributions of the different tissues, can be effectively addressed in a variational setting. Such a framework, in fact, implicitly implements an automatic relevance determination scheme, where redundant mixture components are automatically pruned out of the model.

Finally we present a parametric empirical Bayes approach to learn informative intensity priors from sufficiently large data sets and demonstrate how the priors estimated in this fashion can increase the robustness of

the presented segmentation algorithm. We also address the common problem of different MRI images having different intensity values by incorporating a free global rescaling parameter that is optimized, within the same Bayesian framework, so as to increase the consistency of intensities across scans.

2. Background on variational Bayes

In this section we summarize the underpinnings of variational Bayesian inference. We highlight the advantages of VB over point estimation techniques and illustrate some of the challenges that arise in the variational framework.

Variational Bayesian inference can be formulated as a maximization (or minimization) problem.

Let us consider the marginal log likelihood $\log p(\mathbf{X})$ given by

$$\log p(\mathbf{X}) = \log \int p(\mathbf{X}, \boldsymbol{\Upsilon}) \, \mathrm{d}\boldsymbol{\Upsilon} \; , \tag{1}$$

where $\mathbf{X}$ indicates the observed data and $\boldsymbol{\Upsilon}$ the set of unobserved variables (model parameters and latent variables).

If we introduce a distribution $q(\boldsymbol{\Upsilon})$ over the unobserved variables, the log evidence in (1) can be re-expressed as

$$\begin{aligned} \log p(\mathbf{X}) &= \int q(\boldsymbol{\Upsilon}) \log p(\mathbf{X}) \, \mathrm{d}\boldsymbol{\Upsilon} \\ &= \int q(\boldsymbol{\Upsilon}) \log \left\{ \frac{p(\mathbf{X}, \boldsymbol{\Upsilon})}{q(\boldsymbol{\Upsilon})} \right\} \mathrm{d}\boldsymbol{\Upsilon} \\ &+ \int q(\boldsymbol{\Upsilon}) \log \left\{ \frac{q(\boldsymbol{\Upsilon})}{p(\boldsymbol{\Upsilon}|\mathbf{X})} \right\} \mathrm{d}\boldsymbol{\Upsilon} \; , \end{aligned} \tag{2}$$

which is a decomposition of $\log p(\mathbf{X})$ that holds for any $q(\boldsymbol{\Upsilon})$.

The second integral in the last line of (2) is the Kullback-Leibler divergence $D_{KL}(q\|p)$ between $q(\boldsymbol{\Upsilon})$, which is a variational approximating posterior, and $p(\boldsymbol{\Upsilon}|\mathbf{X})$, which is the true posterior distribution [8].

Since $D_{KL}(q\|p) \geq 0$, the first integral in the last line of (2) defines a lower bound $\mathcal{L}(q)$ on the logarithm of the model evidence

$$\log p(\mathbf{X}) \geq \mathcal{L}(q) = \int q(\boldsymbol{\Upsilon}) \log \left\{ \frac{p(\mathbf{X}, \boldsymbol{\Upsilon})}{q(\boldsymbol{\Upsilon})} \right\} \mathrm{d}\boldsymbol{\Upsilon} \; . \tag{3}$$

The previous statement can also be derived from (1) by applying Jensen's inequality.

In summary equation (2) can be rewritten as [41]

$$\log p(\mathbf{X}) = \mathcal{L}(q) + D_{KL}(q\|p) \; . \tag{4}$$

As anticipated, $D_{KL}(q\|p)$ is always non negative and, in particular, it is equal to zero if and only if $q(\boldsymbol{\Upsilon}) = p(\boldsymbol{\Upsilon}|\mathbf{X})$. In such a case the variational posterior is an exact solution and the lower bound is exactly equal to the evidence.

In all the other cases, $D_{KL}(q\|p) > 0$ and $\mathcal{L}(q) < \log p(\mathbf{X})$, which means that $q(\boldsymbol{\Upsilon})$ is an approximate posterior.

In summary, the inference problem can be solved by maximizing the functional $\mathcal{L}(q)$ with respect to the distribution $q(\boldsymbol{\Upsilon})$, which is equivalent to minimizing the Kullback-Leibler divergence between the variational and the true posterior distribution.

The lower bound on the model evidence (negative variational free energy) can be further decomposed as

$$\mathcal{L}(q) = \int q(\boldsymbol{\Upsilon}) \log p(\mathbf{X}|\boldsymbol{\Upsilon}) \mathrm{d}\boldsymbol{\Upsilon} + \int q(\boldsymbol{\Upsilon}) \log \left\{ \frac{p(\boldsymbol{\Upsilon})}{q(\boldsymbol{\Upsilon})} \right\} \mathrm{d}\boldsymbol{\Upsilon} \; . \tag{5}$$

This shows that the lower bound comprises a likelihood term which is equal to the expected value of the log likelihood $\log p(\mathbf{X}|\boldsymbol{\Upsilon})$ under the variational posterior $q(\boldsymbol{\Upsilon})$

$$\mathcal{L}_1 = \int q(\boldsymbol{\Upsilon}) \log p(\mathbf{X}|\boldsymbol{\Upsilon}) \mathrm{d}\boldsymbol{\Upsilon} = \mathbb{E}_{\boldsymbol{\Upsilon}} \big[\log p(\mathbf{X}|\boldsymbol{\Upsilon}) \big] \; , \tag{6}$$

and a regularizing term which is the negative Kullback-Leibler divergence beetween the approximating posterior $q(\boldsymbol{\Upsilon})$ and the prior distribution over the unobserved variables $p(\boldsymbol{\Upsilon})$ [5]

$$\mathcal{L}_2 = \int q(\boldsymbol{\Upsilon}) \log \left\{ \frac{p(\boldsymbol{\Upsilon})}{q(\boldsymbol{\Upsilon})} \right\} \mathrm{d}\boldsymbol{\Upsilon} = -D_{KL}(q\|p_0) \; . \tag{7}$$

This last term penalizes overly complex or implausible models (Occam factor) [5].

While in principle arbitrary variational posterior distributions $q(\boldsymbol{\Upsilon})$ can be used, a commonly adopted strategy to solve the inference problem consists in restricting the space of $q(\boldsymbol{\Upsilon})$ so as to ensure mathematical tractability, which also means that $D_{KL}(q\|p) > 0$, or, in other words, that $q(\boldsymbol{\Upsilon}) \neq p(\boldsymbol{\Upsilon}|\mathbf{X})$. In particular, it is often convenient to assume that $q(\boldsymbol{\Upsilon})$ factorizes into a product of terms, each one involving just a subset of $\boldsymbol{\Upsilon}$ (mean field theory): $q(\boldsymbol{\Upsilon}) = \prod_{s=1}^{S} q_s(\boldsymbol{\Upsilon}_s)$.

In such a case, the lower bound depends on the generic factor $q_{\hat{s}}(\boldsymbol{\Upsilon}_{\hat{s}})$ as follows [8]

$$\mathcal{L}(q_{\hat{s}}) = -D_{KL}(q_{\hat{s}} \,\|\, \hat{p}(\mathbf{X}, \boldsymbol{\Upsilon}_{\hat{s}})) + \text{const} \; , \tag{8}$$

with

$$\hat{p}(\mathbf{X}, \boldsymbol{\Upsilon}_{\hat{s}}) \propto \exp(\mathbb{E}_{s \neq \hat{s}}[\log p(\mathbf{X}, \boldsymbol{\Upsilon})]) \; . \tag{9}$$

Equation (8) shows that the optimal form of the factor $q_{\hat{s}}(\boldsymbol{\Upsilon}_{\hat{s}})$ corresponds to the one that minimizes the Kullback-Leibler divergence between $q_{\hat{s}}(\boldsymbol{\Upsilon}_{\hat{s}})$ and $\hat{p}(\mathbf{X}, \boldsymbol{\Upsilon}_{\hat{s}})$ which is defined in (9). Therefore $q_{\hat{s}}(\boldsymbol{\Upsilon}_{\hat{s}}) \equiv \hat{p}(\mathbf{X}, \boldsymbol{\Upsilon}_{\hat{s}})$.

Note that this solution is not analytical, 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 [8, 41].

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. One other advantage from adopting a VB approach is that its generality allows it to be combined with standard MAP and ML techniques in a unified and principled framework. If one of the subsets $\{\boldsymbol{\Upsilon}_s\}_{s=1,\ldots,S}$ of the unobserved variables cannot be treated in a fully Bayesian manner, it is still possible to obtain MAP point estimates for the corresponding parameters. 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 $\mathbf{X}$ and $\boldsymbol{\Upsilon}$, $\mathbb{E}[\log p(\mathbf{X}, \boldsymbol{\Upsilon})]$. The main difference from the EM algorithm for ML (or MAP) estimation is that in the VBEM case, expectations are computed not only over the latent variables of the model (as in the EM), but also over all the model parameters that are described in terms of a full posterior distribution.

3. Data model

This section describes the mathematical model adopted for this work. We introduce all variables, illustrate their conditional dependencies and, finally, we derive a variational objective function (lower bound).

Let $\mathbf{X}$ 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 $\mathbf{x}_j \in \mathbb{R}^D$, with $j \in \{1, \ldots, N\}$.

We can model the distribution of $\mathbf{x}_j$ as a multivariate Gaussian mixture consisting of K clusters parametrized by mean vectors $\{\boldsymbol{\mu}_k\}_{k=1,\ldots,K}$ and covariance matrices $\{\boldsymbol{\Sigma}_k\}_{k=1,\ldots,K}$. The mixing proportions of the different components are given by $\Theta_\pi = \{\pi_{jk}\}$ with $\pi_{jk} \in [0, 1]$ and $\sum_k \pi_{jk} = 1$. Essentially π_{jk} indicates the prior probability of signal at spatial location j being drawn from cluster k.

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

The prior probability of voxel j belonging to tissue t is considered to be given a priori (through a probabilistic atlas) and is indicated by τ_{jt}. Furthermore, these tissue priors are allowed to be rescaled by a set of weights $\{w_t\}_{t=1,\ldots,T}$ to accommodate individual differences in tissue composition. Finally it is necessary to 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

$$\forall t \in \{1, \ldots, T\} : \sum_{k \in C_t} g_k = 1. \tag{10}$$

As a result the mixing proportions Θ_π of the presented GMM can be expressed as

$$\pi_{jk} = g_k \, \frac{\tau_{jt} \, w_t}{\sum_{t'}^{T} \tau_{jt'} \, w_{t'}} \, , \tag{11}$$

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 $\mathbf{X}$.

To correct for intensity non-uniformity artifacts, a multiplicative D-dimensional bias field, denoted by $\{\mathbf{b}_j(\Theta_\beta)\}_{j=1,\ldots,N}$, is introduced in the model, where Θ_β 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 [4].

Finally, to account for the variability of anatomical shapes among subjects, the probabilistic atlas given by $\{\boldsymbol{\tau}_t\}_{t=1,\ldots,T}$ is allowed to be deformed according to a displacement field parametrized by the set of vectors $\Theta_\alpha = \{\boldsymbol{\alpha}_j\}_{j=1,\ldots,N}$. The warped tissue priors can therefore be expressed as $\{\boldsymbol{\tau}_t(\varphi(\Theta_\alpha))\}_{t=1,\ldots,T}$, 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 $\{\boldsymbol{\alpha}_j\}_{j=1,\ldots,N}$, so that

$$\tilde{\mathbf{y}}_j = \mathbf{y}_j + \boldsymbol{\alpha}_j \, , \tag{12}$$

where the vector $\mathbf{y}_j$ encodes the coordinates of the centre of voxel j.

If we introduce a set of binary latent variables $\mathbf{Z}$ denoting the class memberships of the observed data $\mathbf{X}$, the probability of $\mathbf{Z}$ given the mixing proportions Θ_π and the deformation parameters Θ_α is given by

$$p(\mathbf{Z}|\Theta_\pi, \Theta_\alpha) = \prod_{j=1}^{N} \prod_{k=1}^{K} \left(\frac{\tau_{jt}\big(\varphi(\Theta_\alpha)\big) \, w_t}{\sum_{t'}^{T} \tau_{jt'}\big(\varphi(\Theta_\alpha)\big) \, w_{t'}} \right)^{z_{jk}} \, , \tag{13}$$

where we have assumed that all data are independent.

The conditional distribution (class conditional density) of the observed intensities given the latent variables, the Gaussian means Θ_μ and covariances Θ_Σ and the bias field parameters Θ_β can be expressed as

$$p(\mathbf{X}|\mathbf{Z}, \Theta_\mu, \Theta_\Sigma, \Theta_\beta) = \prod_{j=1}^{N} \prod_{k=1}^{K} \left(|\mathbf{B}_j| \, \mathcal{N}(\mathbf{B}_j \mathbf{x}_j | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right)^{z_{jk}} \, , \tag{14}$$

with $\mathbf{B}_j = \mathrm{diag}(\mathbf{b}_j)$, modelling the bias field.

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

$$p(\mathbf{X}, \mathbf{Z}, \Theta_\mu, \Theta_\Sigma, \Theta_\beta, \Theta_\alpha | \Theta_\pi) = p(\mathbf{X}|\mathbf{Z}, \Theta_\mu, \Theta_\Sigma, \Theta_\beta) p(\mathbf{Z}|\Theta_\pi, \Theta_\alpha) p(\Theta_\mu, \Theta_\Sigma) p(\Theta_\alpha) p(\Theta_\beta) \, . \tag{15}$$

The voxel specific mixing proportions Θ_π are treated here as deterministic parameters depending on the available anatomical atlas, on the tissue weights $\{w_t\}_{t=1,\dots,T}$ and on the within-tissue mixing proportions $\{g_k\}_{k=1,\dots,K}$, therefore they are determined via ML estimation.

It should be noted that we have kept the formal distinction between latent variables $\mathbf{Z}$ and model parameters $\boldsymbol{\Theta}$ for clarity, even if the treatment of these is essentially the same with variational inference techniques.

The priors on the means and covariances of the different classes are modelled as Gaussian-Wishart distributions

$$p(\Theta_\mu, \Theta_\Sigma) = \prod_{k=1}^{K} p(\boldsymbol{\mu}_k|\boldsymbol{\Sigma}_k^{-1}) p(\boldsymbol{\Sigma}_k^{-1}) \;, \tag{16}$$

with

$$p(\boldsymbol{\mu}_k|\boldsymbol{\Sigma}_k^{-1}) = \mathcal{N}(\boldsymbol{\mu}_k|\boldsymbol{m}_{0k}, b_{0k}^{-1}\boldsymbol{\Sigma}_k) \;, \tag{17}$$

$$p(\boldsymbol{\Sigma}_k^{-1}) = \mathcal{W}(\boldsymbol{\Sigma}_k^{-1}|\boldsymbol{W}_{0k}, \nu_{0k}) \;, \tag{18}$$

where $\mathcal{W}(\boldsymbol{W}, \nu)$ indicates the probability density function of a Wishart distribution with ν degrees of freedom and scale matrix $\boldsymbol{W}$ (see Appendix A for a more detailed description of Gaussian-Wishart priors).

Such a choice is algebraically convenient, as it leads to posterior distributions having the same functional form of the priors (conjugate priors).

The parameters governing the priors will be indicated as

$$\Phi_0 = \{\beta_{0k}, \boldsymbol{m}_{0k}, \nu_{0k}, \boldsymbol{W}_{0k}\}_{k=1,\dots,K} \;. \tag{19}$$

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 the same regularization scheme described in [4] is adopted. 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 [39].

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

$$\begin{aligned}\mathcal{L} = &\sum_{\mathbf{Z}} \iint q(\mathbf{Z}, \Theta_\mu, \Theta_\Sigma) \\ &\times \log\left\{\frac{p(\mathbf{X}, \mathbf{Z}, \Theta_\mu, \Theta_\Sigma, \Theta_\beta, \Theta_\alpha|\Theta_\pi)}{q(\mathbf{Z}, \Theta_\mu, \Theta_\Sigma)}\right\} \mathrm{d}\Theta_\mu \mathrm{d}\Theta_\Sigma \;.\end{aligned} \tag{20}$$

To make the problem tractable we assume that the variational distribution $q(\mathbf{Z}, \Theta_\mu, \Theta_\Sigma)$ factorizes as $q(\mathbf{Z}, \Theta_\mu, \Theta_\Sigma) = q(\mathbf{Z}) q(\Theta_\mu, \Theta_\Sigma)$, so that

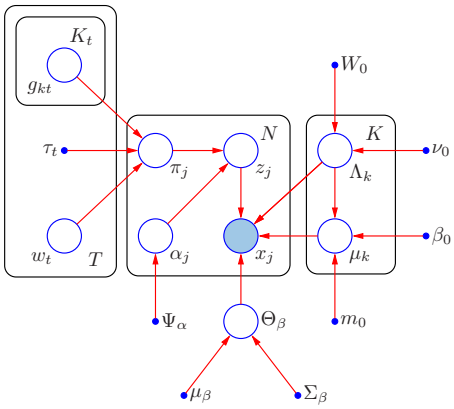


Figure 1: Directed acyclic graph representing the generative Gaussian mixture model adopted in this work for the purpose of segmenting neuroimaging data into tissue types. Large filled circles indicate the observed data (image intensities $\mathbf{X}$). Unfilled circles represent unobserved random variables (latent variables $\mathbf{Z}$, which encode class memberships, and model parameters Θ). Solid dots denote fixed hyperparameters. The observed intensities are assumed to be drawn from a Gaussian mixture distribution consisting of K components with means $\{\boldsymbol{\mu}_k\}$ and covariance matrices $\{\boldsymbol{\Sigma}_k\}$. Intensity non-uniformities are modelled through a multiplicative bias field parametrized by Θ_β. The mixing proportions of our model $\{\pi_{jk}\}$ vary locally according to a smooth anatomical atlas $\{\tau_t\}$, which is mapped onto the individual data by the deformation field encoded in $\{\boldsymbol{\alpha}_j\}$.

$$\begin{aligned}\mathcal{L} = &\sum_{\mathbf{Z}} \iint q(\mathbf{Z}) q(\Theta_\mu, \Theta_\Sigma) \log p(\mathbf{X}|\mathbf{Z}, \Theta_\mu, \Theta_\Sigma, \Theta_\beta) \mathrm{d}\Theta_\mu \mathrm{d}\Theta_\Sigma \\ &+ \sum_{\mathbf{Z}} \iint q(\mathbf{Z}) q(\Theta_\mu, \Theta_\Sigma) \\ &\times \log\left\{\frac{p(\mathbf{Z}|\Theta_\pi, \Theta_\alpha) p(\Theta_\mu, \Theta_\Sigma)}{q(\mathbf{Z}) q(\Theta_\mu, \Theta_\Sigma)}\right\} \mathrm{d}\Theta_\mu \mathrm{d}\Theta_\Sigma \\ &+ p(\Theta_\beta) + p(\Theta_\alpha) \;.\end{aligned} \tag{21}$$

The described probabilistic model can be represented by a directed acyclic graph, as shown in figure 1. It should be noted that such a model is generative. In fact equation (15) allows to generate synthetic observations $\mathbf{X}$ via sampling from the joint distribution of all random variables conditioned on the mixing proportions [7].

Once the model has been learned, it is directly possible to infer the tissue labels by examining the posterior distribution $p(\mathbf{Z}|\mathbf{X})$. In fact, the decision on which class each voxel belongs to should be taken according to Bayes decision rule (BDR), which states that the class to be chosen is the one that minimizes the probability ("risk") of error. In practice,

$$\mathbf{x}_j \in \text{class } k \iff p(z_{jk} = 1|\mathbf{x}_j) > p(z_{jk'} = 1|\mathbf{x}_j) \quad \forall k' \neq k \;. \tag{22}$$

4. Model learning

The statistical model descried in Section 3 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 $q(\mathbf{Z})q(\Theta_\mu, \Theta_\Sigma)$, to estimate MAP values for the parameters $\{\Theta_\alpha, \Theta_\beta\}$ and ML values for $\{g_k\}_{k=1,\dots,K}$ and $\{w_t\}_{t=1,\dots,T}$.

4.1. 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 (20) is maximized with respect to the posterior factor $q(\mathbf{Z})$ over the latent variables [8]. Making use of (9) we find that

$$\begin{aligned} q(\mathbf{Z}) \propto & \exp\big(\log p(\mathbf{Z}|\Theta_\pi, \Theta_\alpha) \\ & + \mathbb{E}_{\mu,\Sigma}\left[\log p(\mathbf{X}|\mathbf{Z}, \Theta_\mu, \Theta_\Sigma, \Theta_\beta)\right]\big) \ . \end{aligned} \tag{23}$$

If we define

$$\log \rho_{jk} = \log p(\mathbf{Z}|\Theta_\pi, \Theta_\alpha) + \mathbb{E}_{\mu,\Sigma}\left[\log p(\mathbf{X}|\mathbf{Z}, \Theta_\mu, \Theta_\Sigma, \Theta_\beta)\right] \ , \tag{24}$$

it follows that

$$q(\mathbf{Z}) \propto \prod_{j=1}^{N}\prod_{k=1}^{K} (\rho_{jk})^{z_{jk}} \ . \tag{25}$$

By normalizing the variational distribution $q(\mathbf{Z})$ we obtain

$$q(\mathbf{Z}) = \prod_{j=1}^{N}\prod_{k=1}^{K} \left(\frac{\rho_{jk}}{\sum_{c=1}^{K}\rho_{jc}}\right)^{z_{jk}} = \prod_{j=1}^{N}\prod_{k=1}^{K} (\gamma_{jk})^{z_{jk}} \ . \tag{26}$$

The quantity $\log \rho_{jk}$ can be computed from (24) to give

$$\begin{aligned} \log \rho_{jk} = & \log \pi_{jk}(\varphi(\Theta_\alpha)) \\ & - \frac{D}{2}\log(2\pi) + \frac{1}{2}\mathbb{E}_{\mathbf{\Sigma}_k}\left[\log|(\mathbf{\Sigma}_k)^{-1}|\right] \\ & - \frac{1}{2}\mathbb{E}_{\boldsymbol{\mu}_k,\mathbf{\Sigma}_k}\left[(\mathbf{B}_j\mathbf{x}_j - \boldsymbol{\mu}_k)^T\mathbf{\Sigma}_k^{-1}(\mathbf{B}_j\mathbf{x}_j - \boldsymbol{\mu}_k)\right] \ . \end{aligned} \tag{27}$$

The expectations that appear in (27) have to be computed with respect to the current estimates of the variational posterior distributions over $\{\boldsymbol{\mu}_k\}_{k=1,\dots,K}$ and $\{\mathbf{\Sigma}_k\}_{k=1,\dots,K}$ (Appendix A).

The terms $\{\gamma_{jk}\}$ which are computed during the VE-step are equal to the expectations of the latent variables with respect to their posterior variational distribution (responsibilities) [8]. They can be used to compute the following sufficient statistics of the observed data, which will serve to update the posterior distributions of $\{\boldsymbol{\mu}_k\}_{k=1,\dots,K}$ and $\{\mathbf{\Sigma}_k\}_{k=1,\dots,K}$, during the VM-step

$$\begin{aligned} s_{0k} &= \sum_{j=1}^{N}\gamma_{jk} \ , \\ \boldsymbol{s}_{1k} &= \sum_{j=1}^{N}\gamma_{jk}\mathbf{B}_j\mathbf{x}_j \ , \\ \boldsymbol{S}_{2k} &= \sum_{j=1}^{N}\gamma_{jk}(\mathbf{B}_j\mathbf{x}_j)(\mathbf{B}_j\mathbf{x}_j)^T \ . \end{aligned} \tag{28}$$

It should be noted that the computational complexity of this VE-step is identical to that of the E-step in the standard EM algorithm.

4.2. Variational M-step

In the following VM-step we can derive approximate solutions for the posterior distributions over the cluster means and covariance matrices [8]. Making again use of (9) we obtain

$$\begin{aligned} q(\Theta_\mu, \Theta_\Sigma) \propto \exp\Bigg\{ & \sum_{j=1}^{N}\sum_{k=1}^{K}\gamma_{jk}\log\mathcal{N}(\mathbf{B}_j\mathbf{x}_j|\boldsymbol{\mu}_k, \mathbf{\Sigma}_k) \\ & + \sum_{k=1}^{K}\log p(\Theta_\mu, \Theta_\Sigma)\Bigg\} \ . \end{aligned} \tag{29}$$

It can be proved (Appendix B) that the posterior distributions on the means and covariances of the different Gaussians take the same form as the corresponding priors [8], that is

$$q(\Theta_\mu, \Theta_\Sigma) = \prod_{k=1}^{K} q(\boldsymbol{\mu}_k|\mathbf{\Sigma}_k^{-1})q(\mathbf{\Sigma}_{ik}^{-1}) \ , \tag{30}$$

with

$$q(\boldsymbol{\mu}_k|\mathbf{\Sigma}_k^{-1}) = \mathcal{N}(\boldsymbol{\mu}_k|\boldsymbol{m}_k, b_k^{-1}\mathbf{\Sigma}_k) \ , \tag{31}$$

$$q(\mathbf{\Sigma}_k^{-1}) = \mathcal{W}(\mathbf{\Sigma}_k^{-1}|\boldsymbol{W}_k, \nu_k) \ . \tag{32}$$

The parameters that govern these posterior distributions

$$\Phi = \{\beta_k, \boldsymbol{m}_k, \nu_k, \boldsymbol{W}_k\}_{k=1,\dots,K} \ , \tag{33}$$

can be computed as a function of the prior hyperparameters and the sufficient statistics obtained in the previous VE-step, as follows (Appendix B)

$$\begin{aligned} b_k &= b_{0k} + s_{0k} \ , \\ \boldsymbol{m}_k &= \frac{b_{0k}\boldsymbol{m}_{0k} + \boldsymbol{s}_{1k}}{b_{0k} + s_{0k}} \ , \\ \boldsymbol{W}_k^{-1} &= \boldsymbol{W}_{0k}^{-1} + \boldsymbol{S}_{2k} + \frac{b_{0k}s_{0k}\boldsymbol{m}_{0k}\boldsymbol{m}_{0k}^T}{b_{0k} + s_{0k}} - \frac{\boldsymbol{s}_{1k}\boldsymbol{s}_{1k}^T}{b_{0k} + s_{0k}} \\ & \quad - \frac{b_{0k}\boldsymbol{s}_{1k}\boldsymbol{m}_{0k}^T}{b_{0k} + s_{0k}} - \frac{b_{0k}\boldsymbol{m}_{0k}\boldsymbol{s}_{1k}^T}{b_{0k} + s_{0k}} \ , \\ \nu_k &= \nu_{0k} + s_{0k} \ . \end{aligned} \tag{34}$$

The point estimates of the mixing proportions $\{g_k\}_{k=1,\dots,K}$ within each tissue type can instead be updated by

$$g_k = \frac{s_{0k}}{\sum_{c \in C_t} s_{0c}} \,, \tag{35}$$

while for the tissue weights $\{w_t\}_{t=1,\dots,T}$ we obtain the following

$$w_t = \frac{\sum_{k \in C_t} s_{0k}}{\sum_{j=1}^{N} \frac{\tau_{jt}(\varphi(\Theta_\alpha))}{\sum_{t'=1}^{T} \tau_{jt'}(\varphi(\Theta_\alpha)) w_{t'}}} \,. \tag{36}$$

A brief discussion on how to solve the bias and deformation optimization problems is provided in Appendix C and Appendix D.

4.3. Empirical Bayes learning of GMM priors

The hyperparameters Φ_0 reflect prior beliefs on how signal intensities should be distributed within each tissue type. With a Gaussian-Wishart parametrization, the following hyperparameter setting ensures minimally informative and non improper priors

$$\beta_{0k} \simeq 0 \wedge \nu_{0k} \simeq D - 1 \implies p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \simeq \text{const} \,. \tag{37}$$

With such a choice, the posterior distributions of the Gaussian parameters would be essentially determined by fitting the data, in a similar way to the maximum likelihood framework, and the regularization term of the lower bound would reduce to the entropy of the posterior distributions.

On the contrary, choosing more informative priors can potentially increase the robustness of the algorithm by enforcing meaningfulness and plausibility of the estimated posteriors and, at the same time, ensure faster convergence. However, defining pertinent priors is a non trivial problem, as ideally such priors should express information derived from previously acquired data, rather than simple subjective beliefs. Therefore an appropriate hyperparameter configuration should be learned from large population data. Essentially, informative empirical priors should allow transferring the posterior information inferred from a training data set onto new unseen testing data [26, 36, 38].

Interestingly, the model described so far can be further extended to represent a data set comprising scans of different subjects and, therefore, it provides a natural framework for estimating empirical priors. In fact, a lower bound on the marginal likelihood can be expressed, for a population of M subjects, as follows

$$\begin{aligned} \mathcal{L} = & \sum_{i=1}^{M} \sum_{\mathbf{Z}} \iint q_i(\mathbf{Z}_i, \Theta_\mu, \Theta_\Sigma) \\ & \times \log \left\{ \frac{p_i(\mathbf{X}_i, \mathbf{Z}_i, \Theta_\mu, \Theta_\Sigma, \Theta_\beta, \Theta_\alpha | \Theta_\pi)}{q_i(\mathbf{Z}_i, \Theta_\mu, \Theta_\Sigma)} \right\} \mathrm{d}\Theta_\mu \mathrm{d}\Theta_\Sigma \end{aligned} \tag{38}$$

Supposing that the posteriors $\{q_i(\Theta_\mu, \Theta_\Sigma)\}_{i=1,\dots,M}$ have been estimated, equation (38) can be maximized with respect to $p(\Theta_\mu, \Theta_\Sigma)$. Since we are assuming that the functional form of this distribution is parametric and known (Gaussian-Wishart), standard non-linear optimization techniques can be exploited to find maximum likelihood estimates of the hyperparameters Φ_0.

Indeed, the lower bound in (38) can be expressed as a function of Φ_0

$$\begin{aligned} \mathcal{L}(\Phi_0) = & \sum_{i=1}^{m} \int \int q_i(\Theta_\mu, \Theta_\Sigma) \log p(\Theta_\mu, \Theta_\Sigma) \, \mathrm{d}\Theta_\mu \mathrm{d}\Theta_\Sigma + \text{const} \\ = & \frac{1}{2} \sum_{i=1}^{M} \sum_{k=1}^{K} \Big\{ \mathbb{E}\big[\log |\boldsymbol{\Sigma}_{ik}^{-1}| \big] (\nu_{0k} - D) \\ & - \nu_k \operatorname{Tr}(\boldsymbol{W}_{0k}^{-1} \boldsymbol{W}_{ik} + \beta_{0k} (\boldsymbol{m}_{ik} - \boldsymbol{m}_{0k})(\boldsymbol{m}_{ik} - \boldsymbol{m}_{0k})^T \boldsymbol{W}_k) \Big\} \\ & + \frac{M}{2} \sum_{k=1}^{K} D \log \frac{\beta_{0k}}{2\pi} - D \sum_{i=1}^{M} \sum_{k=1}^{K} \frac{\beta_{0k}}{\beta_{ik}} \\ & + 2M \sum_{k=1}^{K} \log B_W(\boldsymbol{W}_{0k}, \nu_{0k}) + \text{const} \,, \end{aligned} \tag{39}$$

where B_W indicates the normalizing constant of the Wishart distribution.

We also derived the first and second derivatives of $\mathcal{L}(\Phi_0)$, which are useful to solve this optimization problem using gradient based techniques. Such derivatives are reported in Appendix F.

In summary, a convenient strategy for learning Gaussian mixture priors consists in, first, initializing the hyperparameters so as to obtain weak priors, secondly, estimating the posterior distributions for a population of M subjects, finally, optimizing $\mathcal{L}$ with respect to Φ_0. The estimates of the hyperparameters (Φ_0) can then be further refined by using these empirical priors to reestimate the posteriors and so on, thus leading to an iterative learning scheme.

5. Experimental results

In this section we present a series of experiments that were performed to assess the validity of our approach and to explore some of its properties and potential applications. The results presented in 5.1 were produced making use of synthetic data while the ones described in 5.2 were obtained on real, publicly available, MRI data.

5.1. Experiments on synthetic data

The performance of our variational algorithm was first evaluated making use of simulated data produced by the Brainweb MRI simulator [9, 25, 10].

To assess the accuracy of brain tissue classification performed by our method we employed twenty synthetic T1-weighted scans of healthy adult subjects [6]. The volumes

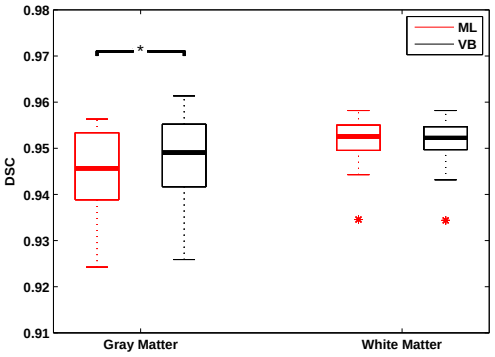


Figure 2: Dice similarity coefficients between the gray and white matter segmentations produced by our algorithm (VB) and the underlying ground truth for twenty simulated T1-weighted scans. We report as well the DSC obtained with the ML algorithm provided with SPM12. For each boxplot, the central mark indicates the median, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points, while outliers are indicated by red stars.

were generated with the following MR simulation parameters: Spoiled Fast Low Angle Shot (SFLASH) sequence with repetition time (TR) of 22 ms, echo time (TE) of 9.2 ms, flip angle=30 deg and 1 mm isotropic voxel size.

We segmented these images using the algorithm presented in the previous section. We set the following hyperparameters values so as to obtain weakly informative intensity priors (WIP). This choice in fact permits quantifying the accuracy of our method in the most general case that no reliable information is available on the distribution of tissue intensities.

$$\begin{aligned} \beta_{0k} &= 0.1, \\ \boldsymbol{m}_{0k} &= \frac{1}{N}\sum_{j=1}^{N}\mathbf{x}_j, \\ \nu_{0k} &= D - 0.9, \\ \boldsymbol{W}_{0k}^{-1} &= \frac{1}{N}\sum_{j=1}^{N}(\mathbf{x}_j - \boldsymbol{m}_{0k})(\mathbf{x}_j - \boldsymbol{m}_{0k})^T. \end{aligned} \tag{40}$$

The resulting segmentations were compared to the anatomical models used to generate the data by computing Dice similarity coefficients (DSC). Results, which are reported in figure 2, indicate that our method can segment gray and white matter with an accuracy that is at least equal to that of some widely used, state-of the-art segmentation tools, such as the ones provided with SPM [4], FSL [50] and Freesurfer [16], whose performance was assessed in [23].

The Brainweb database also provides multimodality data, even if, in this case, only one anatomical model is available. We made use of these synthetic scans to test the performance of our algorithm in segmenting multispectral data. Specifically we simulated T1-weighted and T2-weighted volumes, with the pulse sequence parameters reported in table 1 and then segmented the data with the same hyperparameter setting used for the previous experiment.

To examine the behaviour of the algorithm with respect to noise, we repeated the analyses for three different levels of noise in the data (3%, 5% and 9% of the brightest intensity).

Results are summarized in table 2. They clearly indicate that our method can successfully handle multimodality data sets and that, even if the use of a single modality (in this case T1-weighted) already ensures very accurate segmentations, the availability of scans with different contrast can provide additional robustness to noise.

A similar behaviour is exhibited by the ML algorithm provided with the SPM software (table 2). However, by comparing the accuracy achieved by the two methods, we find that the variational implementation provides significantly better results.

With this simulated data we could also assess the validity of bias field correction performed by our method. To do so we computed Pearson's correlation coefficients between the estimated non-uniformity fields and the ground truth. Results are shown in table 3, where we report as well the correlation coefficients achieved by SPM ML based segmentation algorithm. As to be expected the two methods perform quite similarly in estimating the non-uniformity field. In fact, they rely on the same parametrization and optimization of the bias. Nevertheless, because the accuracy in correcting intensity inhomogeneities depends heavily on how reliable the estimates of the Gaussian parameters are, our algorithm, which takes into account the posterior uncertainty of such estimates, can outperform the maximum likelihood implementation when noise in the data increases.

With our method, run time for each individual segmentation was approximately 3 min 30 s, on a Quad-Core PC at 3.19 GHz with 12 GB RAM.

5.1.1. Learning GMM priors

Among the advantages of the variational framework that we present here is the fact that it allows incorporating priors on the parameters modelling the intensity distributions of brain (and potentially non brain) tissues. This form of a priori knowledge acts conjointly with the shape information carried by the tissue probability maps, thus ensuring additional robustness. The use of different intensity priors leads to differences in the estimated posteriors and segmentations, in the sense that the algorithm tries to simultaneously maximize the model fit (i.e. the likelihood of the data) while minimizing the divergence between the prior and posterior probability distributions.

Determining suitable priors for each application, or imaging modality, is a fundamental question. However, it should also be noted that the need to define priors does not limit the applicability of the method if compared to

Table 1: Simulation parameters selected to generate the synthetic data which was used to evaluate the accuracy of the presented VB algorithm in segmenting multispectral data. SFLASH and DSE indicate respectively a spoiled fast low angle shot amd a dual spin echo sequence. A bias intensity of 20% corresponds to values of the non-uniformity field in the range [0.9, 1.1] .

		Sequence	TR (ms)	Flip angle (deg)	TE (ms)	Bias field
Modality	T1w	SFLASH	18	30	10	20%
	T2w	DSE_LATE	3300	90	35,120	20%

Table 2: Dice similarity coefficients between the ground truth tissue labels and the segmentations produced by the presented algorithm (VB) and by the ML implementation provided with the SPM software. The experiments were performed on simulated normal brain scans (T1- and T2-weighted) for three different noise levels.

Maximum Likelihood (ML)							
Noise level		**3%**		**5%**		**9%**	
	Modality	**T1w**	**T1w and T2w**	**T1w**	**T1w and T2w**	**T1w**	**T1w and T2w**
Tissue	GM	0.93	0.90	0.91	0.90	0.87	0.88
	WM	0.95	0.95	0.93	0.94	0.88	0.89
Variational Bayes (VB)							
Noise level		**3%**		**5%**		**9%**	
	Modality	**T1w**	**T1w and T2w**	**T1w**	**T1w and T2w**	**T1w**	**T1w and T2w**
Tissue	GM	0.92	0.92	0.92	0.92	0.87	0.89
	WM	0.95	0.96	0.94	0.94	0.89	0.90

standard maximum likelihood techniques. In fact, whenever no information is available on what priors it is most convenient or correct to use, it is always possible to resort to minimally informative priors, which would simply let the algorithm determine the posterior distributions that explain the data best, given the assumption that all parameter settings within the admissible parameter space are equally (or almost equally) probable a priori.

As explained in section 4.3 the variational framework presented in this paper can be exploited to learn empirical priors over the Gaussian mixture parameters from large population data. To demonstrate the efficacy of this procedure we used the same set of simulated T1-weighted scans employed for the previous experiments. We learned priors over the intensities of gray and white matter by first collecting the posterior probability distributions for all the subjects in the data set and then maximizing the functional in (39) with respect to Φ_0. This optimization problem was solved making use of a Gauss-Newton scheme. In particular we iterated over optimizing the priors and updating the posteriors so as reduce the chance of finding suboptimal solutions.

Results are depicted in figure 3. We report the estimated Gaussian priors over the mean intensity of gray (3a) and white (3b) matter. These should be compared to the modes of the corresponding posteriors (red crosses).

Our empirical Bayes learning scheme captures very precisely the information encoded in the variational posteriors. In particular the more the posteriors are peaked and the more they overlap, the more the priors will be informative. If one or more posteriors have relatively high variance, this uncertainty will be immediately reflected in the empirical priors, which will become less informative. In the case of the synthetic data set used for this experiment, all the volumes have the same contrast and intensity mapping, therefore the resulting priors are highly informative.

The true means (blue crosses) are also shown in figure 3. For white matter, they are extremely consistent with the estimated posterior means. In fact, for this data set, our algorithm exhibits higher accuracy in segmenting white matter than gray matter (see figure 2). A slightly higher discrepancy emerges between the true and estimated gray matter mean intensities, which also explains the relatively lower accuracy in classifying gray matter.

5.1.2. Robustness to misregistration and atlas-free segmentation

Atlas based segmentation methods rely heavily on the accuracy in estimating the deformations mapping from the atlas to the individual volumes. Solving the segmentation and registration problems within a single modelling and computational framework has been widely accepted as a powerful and effective strategy in order to assure the success of both processing tasks, additionally to being a theoretically principled approach [34, 4, 14, 48, 49]. Nonetheless, it is possible to encounter cases in which aligning the template to an individual scan turns out to be particularly difficult, due for example to a poor initialization of the deformations or to the presence of anatomical features (especially pathological ones) that the atlas does not capture correctly. In such cases segmentation accuracy can be strongly affected by misregistration errors.

Introducing priors over the intensity distribution parameters is a convenient and reliable solution to cope with

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

	Noise level	**3%**		**5%**		**9%**	
	Algorithm	**VB**	**ML**	**VB**	**ML**	**VB**	**ML**
Modality	T1w	0.83	0.83	0.84	0.82	0.68	0.61
	T2w	0.88	0.88	0.89	0.89	0.88	0.70

these difficulties. In fact, it can help to prevent implausible parameter estimates, whenever registration errors are misleading the model fitting process.

To demonstrate this property, the synthetic data set consisting of twenty T1-weighted scans was split into a training and a testing subset, of ten volumes each. The first ten images were processed by our variational algorithm to learn population representative intensity priors, as explained in 5.1.1. Secondly, the remaining test images were segmented making use of these priors, while registration failure was simulated by imposing a 7.5 mm shift of the atlas from its optimal alignment position in each of the three spatial directions (figure 4).

The accuracy of the resulting segmentations was finally assessed by computing Dice overlap coefficients. Results are illustrated in figure 5. Here the performance of the presented method, used in combination with the empirical priors, is compared to that of the same algorithm with uninformative priors, as well as to that of a maximum likelihood method (implemented in SPM12).

As to be expected the maximum likelihood method and the variational method with uninformative priors do not perform very differently, except for the fact that the ML algorithm shows higher variance in the results. On the contrary, when using the priors learned from the training data, the accuracy in segmenting gray and white matter increases significantly, with a relatively low variance of the overlap measures. This confirms that Bayesian inference can augment the robustness of standard maximum likelihood algorithms, while providing a general and flexible framework that can be applied to many real world problems, by learning appropriate priors from available data.

As an additional proof of validity, we implemented an atlas free version of the algorithm presented in this paper, which we tested on the same synthetic data. We don't expect this purely intensity based framework to achieve segmentation accuracy and reliability comparable to that of the full, atlas driven method. However by showing that, even in the absence of tissue probability maps, we obtain fairly accurate segmentations (figure 6) we demonstrate again the soundness of the algorithm presented in this paper.

5.2. Experiments on real data

The previous experiments, performed on simulated data, have demonstrated and quantified the accuracy of the presented method in segmenting brain tissues from MRI volumes. In fact, due to the availability the underlying ground truth, working with synthetic data is especially convenient for the objective testing of new techniques and for the comparison of their performance to that of the methods that have become established as current state-of-the-art.

Nonetheless, simulated data is intrinsically less complex than the data encoded in any real scan, from a biological point of view, as well as in terms of signal and noise properties. For this reason it is quite important to asses the behaviour of image processing tools also on real data.

In this section we present a series of experiments that we performed on real MRI data from two publicly available data sets: the OASIS (Open Access Series of Imaging Studies) [29] and the IXI (Information eXtraction from Images) databases. Such experiments provide evidence regarding the accuracy of our method for segmenting brain tissues and illustrate some of its distinctive properties, which derive from adopting a variational inference scheme.

5.2.1. Assessing segmentation accuracy

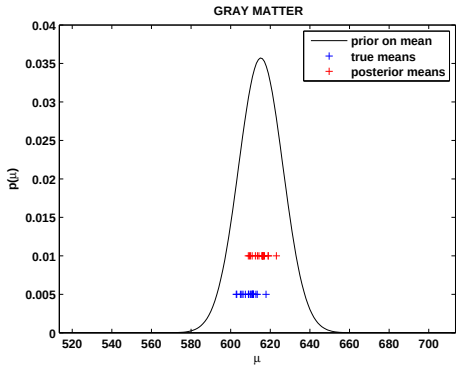
To assess the performance of our segmentation algorithm we used data from the cross-sectional OASIS database (`http://www.oasis-brains.org`), which includes T1-weighted scans of young, middle aged, nondemented and demented older adults, all acquired in one site with the same scanning protocol. Manual labels for a subset of this data set (35 subjects) are provided by Neuromorphometrics, Inc. (`http://Neuromorphometrics.com`) under academic subscription, thus allowing quantitative evaluation of image segmentation tools.

We processed this data with our segmentation algorithm and compared its performance to that of the SPM software.

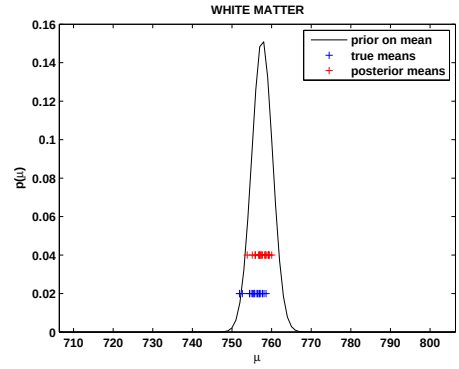
In figure 7 we summarize the distributions of Dice coefficients for gray and white matter, which were obtained by comparing the manual labels with the segmentations produced by our VB method using minimally informative priors and by SPM ML algorithm.

For both tissue types we observe a statistically significant increase in segmentation accuracy, when we use our variational algorithm, compared to the maximum likelihood implementation.

As to be expected, the Dice scores are generally lower, compared to the experiments performed on synthetic data. This is due to the more complex nature of real MRI signals. Additionally, the subset of the OASIS database that

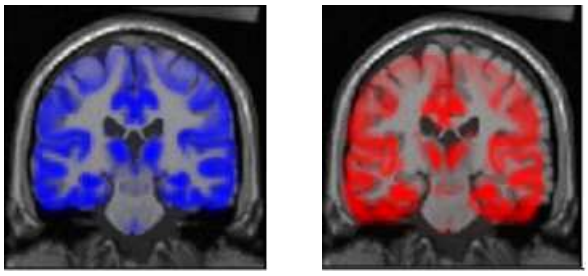


(a)



(b)

Figure 3: Priors over the mean intensities of gray (a) and white (b) matter. The priors were learned from a synthetic data set consisting of 20 T1-weighted scans generated with the Brainweb MR simulator. The Gaussian curves show the estimated priors, while crosses represent the true (blue) and estimated (red) tissue means. The estimated means correspond to the modes of the posterior distributions computed by our VB algorithm.



(a) Aligned template (b) Misaligned template

Figure 4: To test the robustness of the algorithm to misregistration, the tissue probability maps were voluntarily shifted from their optimal position (a) by imposing a 7.5 mm translation in each direction, as illustrated in (b).

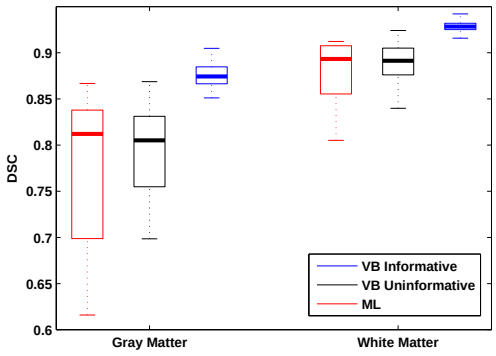


Figure 5: Accuracy of the presented variational algorithm obtained on synthetic data in the presence of registration errors. The performance of the VB algorithm with population based priors (blue) is compared to that of the same algorithm with uninformative priors (red) and to the ML approach implemented in SPM12 (black).

was used for this experiment comprises few scans of elderly subjects with severe atrophy and abnormal signal intensities, which explains the presence of negative outliers in the distribution of accuracy scores.

We performed additional validation experiments using the freely available IXI brain database (`http://www.ixi.org.uk`), which, as opposed to the OASIS data sets, includes multiple modalities, in particular T1-, T2- and PD-weighted images of healthy adult subjects, acquired in three different sites, with different scanning systems. Ground truth segmentations are not available for such a data set. However, in this case, as opposed to the experiments performed on the OASIS data, our aim is to illustrate some of the properties and advantages of our method, rather than providing explicit accuracy measures.

5.2.2. Determining model complexity

One of the most significant advantages of variational inference over maximum likelihood estimation is its intrinsic capability of containing overfitting [5, 8]. In the case of mixture models this allows, for instance, determining the optimal number of components (K) without performing cross-validation, which is usually rather demanding for the amount of computations and data that it requires [8].

Indeed, the question of selecting model's complexity has often been overlooked in the framework of medical image segmentation: throughout the literature, the most common way of handling the choice on the number of classes, is to manually tune K, based on visual inspection of the segmentations and/or intensity histograms. Clearly, this is too arbitrary and subjective for even being considered as a model selection strategy.

Instead, our method implements an implicit automated relevance determination scheme, where, 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 [11, 41], as their responsibilities γ_{jk} are quickly driven to zero by the algorithm. This follows from adopting a variational lower bound to approximate the marginal likelihood, which causes overly complex models (that is to say models with additional clusters that do not significantly help to explain the observed data) to be implicitly penalized [5]. A similar behaviour is inherently impossible to reproduce within model fitting strategies that do not take into account estimation uncertainty,

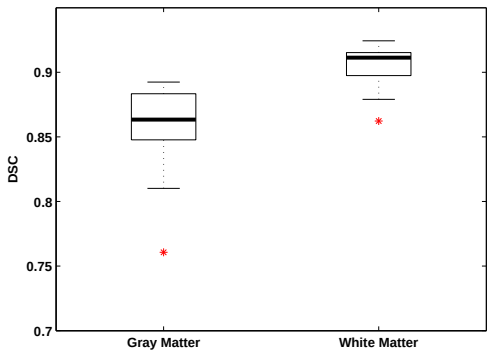


Figure 6: Dice similarity coefficients between the gray and white matter segmentations produced by our algorithm in an atlas free setting and the underlying ground truth, for twenty simulated T1-weighted scans.

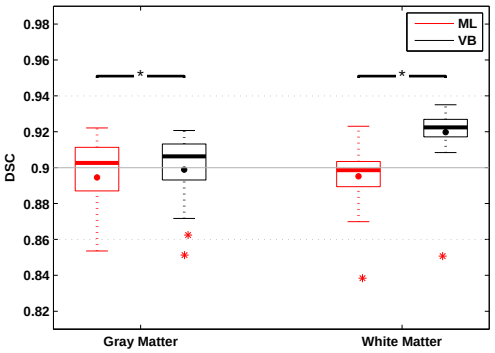
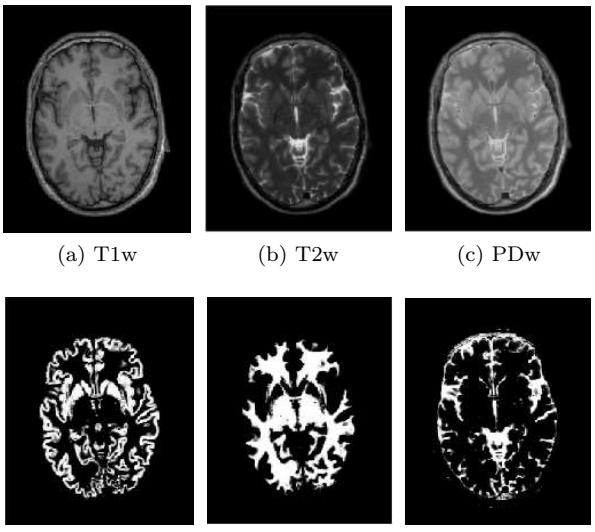


Figure 7: Dice scores computed between the manual labels provided by Neuromorphometrics for a subset of the OASIS data set and the gray and white matter segmentations obtained with our VB method (using minimally informative priors) and with SPM ML algorithm.

such as the maximum likelihood framework.

To illustrate this property of our algorithm we use the images of one of the subjects included in the IXI database. In particular, we processed the data illustrated in figures 8a, 8b and 8c with our method, after having set 5 Gaussians for each of the tissue types of interest. At convergence of the VBEM algorithm, we observed only two components surviving for gray matter, one for white matter, three for CSF, two for bone and four for soft tissues, as shown in figure 9. The plots reported in figure 10 illustrate how the posterior densities over the means of white matter evolve during model learning and, in particular, how four irrelevant components are reverted to their prior distributions, which in this case are uninformative. In a similar setting a ML or MAP 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.



(a) T1w (b) T2w (c) PDw

(d) GM (e) WM (f) CSF

Figure 8: Axial slices of T1-weighted (a), T2-weighted (b) and PD-weighted (c) scans and resulting gray matter (d), white matter (e) and cerebrospinal fluid (f) segmentations obtained with the variational algorithm described in this paper.

5.2.3. Learning informative GMM priors via intensity normalization

One of the difficulties of working with MRI data is the lack of a standardized intensity scale [33]. With respect to our work, this makes it difficult to define, or learn, intensity priors that can effectively generalize to unseen data. Indeed, even for images of a single data set (comprising volumes acquired with the same scanner and protocol) the distribution of intensities across subjects might be poorly consistent.

Unsurprisingly, when we tried to learn intensity priors using real data from the IXI database, we were directly confronted with the problem of normalizing MR signal intensities. Initially we randomly selected 50 T1-weighted scans acquired in the same site and with the same scanner. We processed such data with our variational algorithm and made use of the resulting posterior distributions to estimate intensity priors as described in 4.3. Results are reported in figure 11, where we illustrate the collection of individual posteriors over the means of gray (a) and white (b) matter, together with the resulting empirical priors (black curves).

It is easy to observe how, for this non-quantitative data, the empirical priors turn out to be weakly informative. In fact, they properly reflect the uncertainty due to the variability of the intensity scales. The situation would have been even worse if we had selected volumes acquired with different scanners.

Nonetheless, our generative model can also be exploited to address the problems associated with the non-standardized nature of MRI signals. In particular we can introduce a linear global scaling parameter of the intensities, which can be optimized in the same learning scheme

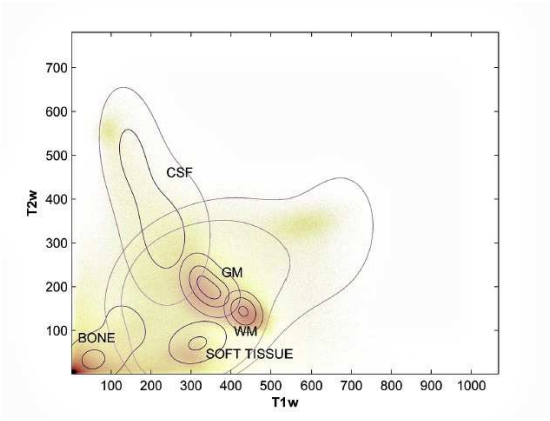


Figure 9: Contour plot of the intensity distributions of gray matter, white matter, cerebrospinal fluid, bone and soft tissue obtained for one subject included in the IXI dataset, overlaid on the joint histogram of the T1- and T2-weighted images. The optimal number of components is determined automatically by our VB algorithm.

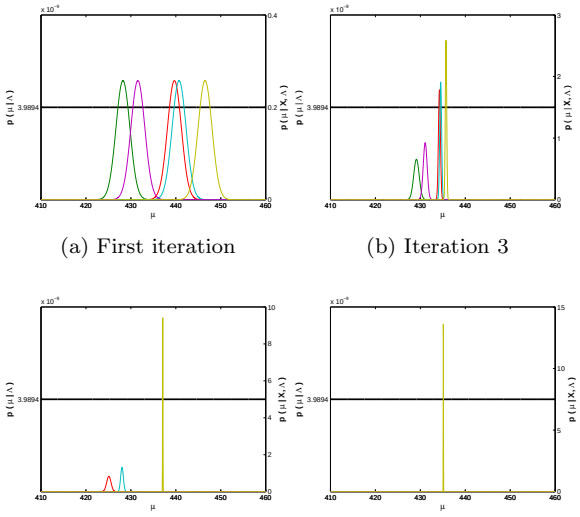
presented in this paper.

We can formulate this as a maximization problem, where our aim is to maximize the following term ($\mathcal{L}_2$) contributing to the lower bound

$$\begin{aligned}\mathcal{L}_2 &= \iint q(\Theta_\mu, \Theta_\Sigma|\Theta_{gs}) \log\left\{\frac{p(\Theta_\mu, \Theta_\Sigma)}{q(\Theta_\mu, \Theta_\Sigma|\Theta_{gs})}\right\} \mathrm{d}\Theta_\mu \mathrm{d}\Theta_\mu \\ &= -D_{KL}(q(\Theta_\mu, \Theta_\Sigma|\Theta_{gs}) \| p(\Theta_\mu, \Theta_\Sigma)) \ ,\end{aligned} \tag{41}$$

which corresponds to minimizing the KL divergence between the intensity priors and the approximating posteriors. The problem can be solved using non linear, gradient based optimization techniques, by computing the first and second derivatives of $\mathcal{L}_2(\Theta_{gs})$ with respect to the global scaling parameters Θ_{gs}. If we iterate over updating the empirical priors and estimating the scaling factors for the individual scans, we manage to learn informative intensity priors, as illustrated in figure 12, while automatically compensating for the inconsistency of MRI signal intensities.

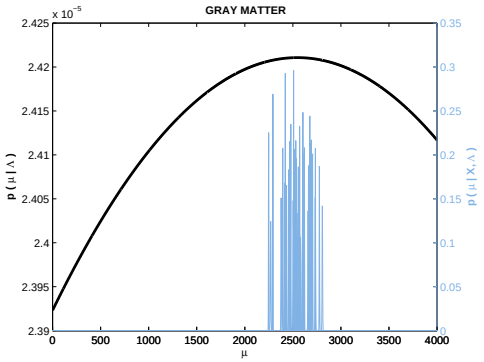
Naturally, such a procedure requires accurate estimates of the intensity distribution, bias and deformation parameters for each individual, that is to say, the problems of learning priors and estimating individual posteriors are inherently related in a circular manner. As a result, for particularly critical data sets, i.e. pathological data, which often exhibit larger anatomical variability, the Bayesian framework described in this manuscript might not be able to provide informative priors, due to the lack of a sufficient number of samples or to poor initial estimates of the model parameters. Nonetheless, in such cases, the presented computational framework, which represents a coherent generalization of some state-of-the-art segmentation algorithms that rely on ML model fitting, could be applied with minimally informative intensity priors and yet it would outperform ML estimation (as indicated by the experiments that we performed on the OASIS data).



(a) First iteration (b) Iteration 3

(c) Iteration 7 (d) Last iteration

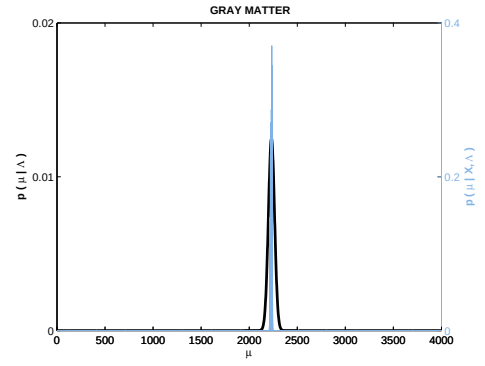
Figure 10: Posterior densities over the means of white matter, at different iterations of our algorithm, showing non relevant components being reverted to their prior distributions.



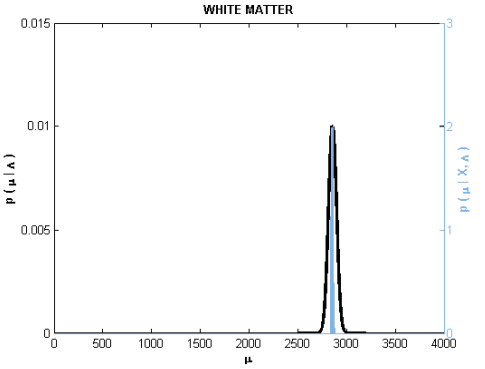
(a)

(b)

Figure 11: Collection of individual posteriors over the means of gray (a) and white (b) matter, obtained from 50 subjects included in the IXI database. Without performing any intensity normalization the resulting empirical priors (black curves) are poorly informative.



(a)



(b)

Figure 12: Collection of individual posteriors over the means of gray (a) and white (b) matter, after including a global rescaling parameter, that is optimized within our generative modelling framework. The estimated priors (black curves) are now much more informative than the ones depicted in figure 11.

6. 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.

In general, finding exact solutions involves the computation of integrals whose treatment in analytical form is very difficult or impossible, while numerical integration is often unfeasible too, due to the volume of data and the complexity of the equations to be solved [8]. In such circumstances, stochastic approximation techniques, like MCMC, have turned out to be impracticable, at least so far, since they require large computational resources, which makes the resulting algorithms rather slow for dealing with real life applications.

Variational Bayes techniques instead formulate Bayesian inference as an optimization problem, where the objective function is constructed to be a lower bound on the marginal likelihood. 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.

When tested on both synthetic and real data, the presented method provided very accurate results, at an equivalent computational cost compared to ML or MAP implementations. We also illustrated some of advantages deriving from adopting a fully Bayesian formulation, such as, the possibility of automatically determining optimal model complexity and performing model comparison by evaluating the model evidence. Finally we described an empirical Bayes learning scheme, that can serve to estimate informative intensity priors. Such priors can be used to ensure even greater robustness, for example in the presence of misalignment between the tissue probability maps and the individual scans, or whenever the available atlases are not best representative of the population of interest.

All of these properties can compensate for the corresponding limitations of standard maximum likelihood techniques, which are inherently prone to overfitting, inappropriate for comparing models and, in the context of atlas based image segmentation, highly sensitive to registration accuracy.

Appendix A. Gaussian-Wishart priors

The Gaussian-Wishart distribution is the conjugate prior of a multivariate D-dimensional normal distribution with unknown mean $\boldsymbol{\mu}$ and precision matrix $\boldsymbol{\Lambda}$. Its probability density function is

$$\begin{aligned} p(\boldsymbol{\mu}, \boldsymbol{\Lambda} | \boldsymbol{m}, \beta, \boldsymbol{W}, \nu) &= p(\boldsymbol{\mu} | \boldsymbol{\Lambda}, \boldsymbol{m}, \beta) p(\boldsymbol{\Lambda} | \boldsymbol{W}, \nu) \\ &= \mathcal{N}(\boldsymbol{\mu} | \boldsymbol{m}, (\beta \boldsymbol{\Lambda})^{-1}) \mathcal{W}(\boldsymbol{\Lambda} | \boldsymbol{W}, \nu) \,, \end{aligned} \tag{A.1}$$

with

$$\mathcal{N}(\boldsymbol{\mu} | \boldsymbol{m}, (\beta \boldsymbol{\Lambda})^{-1}) = \frac{|\beta \boldsymbol{\Lambda}|^{1/2}}{(2\pi)^{D/2}} \exp \left\{ -\frac{1}{2} (\boldsymbol{\mu} - \boldsymbol{m})^T \boldsymbol{\Lambda} (\boldsymbol{\mu} - \boldsymbol{m}) \right\} \,, \tag{A.2}$$

and

$$\mathcal{W}(\boldsymbol{\Lambda} | \boldsymbol{W}, \nu) = B_W(\boldsymbol{W}, \nu) |\boldsymbol{\Lambda}|^{\frac{\nu - D - 1}{2}} \exp \left\{ -\frac{1}{2} \operatorname{Tr} \left(\boldsymbol{W}^{-1} \boldsymbol{\Lambda} \right) \right\} \,. \tag{A.3}$$

The normalizing constant B_W is given by

$$B_W(\boldsymbol{W}, \nu) = |\boldsymbol{W}|^{-\nu/2} \left(2^{\nu D/2} \pi^{D(D-1)/4} \prod_{i=1}^{D} \Gamma \left(\frac{\nu + 1 - i}{2} \right) \right)^{-1} \,, \tag{A.4}$$

where $\Gamma(\cdot)$ is the Gamma function

$$\Gamma(x) = \int_0^\infty u^{x-1} e^{-u} \mathrm{d}u \; . \tag{A.5}$$

The expectation of the determinant of the precision matrix, which appears in equation 27 (VE-step), is equal to [8]

$$\mathbb{E}[\log|\mathbf{\Lambda}|] = \sum_{i=1}^{D} \psi\left(\frac{\nu+1-i}{2}\right) + D\log 2 + \log|\boldsymbol{W}| \; , \tag{A.6}$$

where $\psi(\cdot)$ indicates the digamma function, which is the logarithmic derivative of the Gamma function

$$\psi(x) = \frac{d}{dx}\log\Gamma(x) = \frac{\Gamma'(x)}{\Gamma(x)} \; . \tag{A.7}$$

The following expectation has also to be computed during the VE-step

$$\mathbb{E}_{\boldsymbol{\mu},\mathbf{\Lambda}}\left[(\mathbf{x}-\boldsymbol{\mu})^T \mathbf{\Lambda}\,(\mathbf{x}-\boldsymbol{\mu})\right] = D\beta^{-1} + \nu(\mathbf{x}-\boldsymbol{m})^T \boldsymbol{W}(\mathbf{x}-\boldsymbol{m}) \; . \tag{A.8}$$

Appendix B. Updating the posterior hyperparameters

From equation 29 we obtain

$$\begin{aligned}
&\log q(\boldsymbol{\mu}_k, \mathbf{\Sigma}_k^{-1}) = \\
&+ \log \mathcal{N}(\boldsymbol{\mu}_k|\boldsymbol{m}_{0k}, b_{0k}^{-1}\mathbf{\Sigma}_k) \\
&+ \log \mathcal{W}(\mathbf{\Sigma}_k^{-1}|\boldsymbol{W}_{0k}, \nu_{0k}) \\
&+ \sum_{j=1}^{N} \gamma_{jk} \log \mathcal{N}(\mathbf{B}_j\mathbf{x}_j|\boldsymbol{\mu}_k, \mathbf{\Sigma}_k) + \text{const},
\end{aligned} \tag{B.1}$$

which can be expanded as

$$\begin{aligned}
&\log q(\boldsymbol{\mu}_k, \mathbf{\Sigma}_k^{-1}) = \\
&- \frac{\beta_{0k}}{2}(\boldsymbol{\mu}_k - \boldsymbol{m}_{0k})^T \mathbf{\Sigma}_k^{-1}(\boldsymbol{\mu}_k - \boldsymbol{m}_{0k}) \\
&- \frac{1}{2}\sum_{j=1}^{N} \gamma_{jk}(\mathbf{B}_j\mathbf{x}_j - \boldsymbol{\mu}_k)^T \mathbf{\Sigma}_k^{-1}(\mathbf{B}_j\mathbf{x}_j - \boldsymbol{\mu}_k) \\
&- \frac{1}{2}\operatorname{Tr}\left((\mathbf{\Sigma}_k \boldsymbol{W}_{0k})^{-1}\right) \\
&+ \frac{1}{2}\log|\mathbf{\Sigma}_k^{-1}| \\
&+ \frac{\nu_{0k}-D-1}{2}\log|\mathbf{\Sigma}_k^{-1}| \\
&+ \frac{1}{2}\sum_{j=1}^{N}\gamma_{jk}\log|\mathbf{\Sigma}_k^{-1}| + \text{const} \; .
\end{aligned} \tag{B.2}$$

Let us first consider the terms containing $\boldsymbol{\mu}_k$

$$\begin{aligned}
&\log q(\boldsymbol{\mu}_k|\mathbf{\Sigma}_k^{-1}) = \\
&-\frac{1}{2}\sum_{j=1}^{N}\gamma_{jk}\left(\boldsymbol{\mu}_k^T\mathbf{\Sigma}_k^{-1}(\mathbf{B}_j\mathbf{x}_j - \boldsymbol{\mu}_k) - (\mathbf{B}_j\mathbf{x}_j)^T\mathbf{\Sigma}_k^{-1}\boldsymbol{\mu}_k\right) \\
&-\frac{\beta_{0k}}{2}\left(\boldsymbol{\mu}_k^T\mathbf{\Sigma}_k^{-1}(\boldsymbol{\mu}_k - \boldsymbol{m}_{0k}) - \boldsymbol{m}_{0k}^T\mathbf{\Sigma}_k^{-1}\boldsymbol{\mu}_k\right) \\
&+\text{const} \; .
\end{aligned} \tag{B.3}$$

By rearranging and grouping terms we obtain

$$\begin{aligned}
&\log q(\boldsymbol{\mu}_k|\mathbf{\Sigma}_k^{-1}) = \\
&+ \boldsymbol{\mu}_k^T\mathbf{\Sigma}_k^{-1}\left(\beta_{0k}\boldsymbol{m}_{0k} + \sum_{j=1}^{N}\gamma_{jk}\mathbf{B}_j\mathbf{x}_j\right) \\
&- \frac{1}{2}\left(\beta_{0k} + \sum_{j=1}^{N}\gamma_{jk}\right)\boldsymbol{\mu}_k^T\mathbf{\Sigma}_k^{-1}\boldsymbol{\mu}_k + \text{const} \; .
\end{aligned} \tag{B.4}$$

By completing the square we obtain

$$q(\boldsymbol{\mu}_k|\mathbf{\Sigma}_k^{-1}) = \mathcal{N}(\boldsymbol{\mu}_k|\boldsymbol{m}_k, \beta_k^{-1}\mathbf{\Sigma}_k) \; , \tag{B.5}$$

with

$$\beta_k = \beta_{0k} + s_{0k} \; , \tag{B.6}$$

and

$$\boldsymbol{m}_k = \frac{b_{0k}\boldsymbol{m}_{0k} + \sum_{j=1}^{N}\gamma_{jk}\mathbf{B}_j\mathbf{x}_j}{b_{0k} + s_{0k}} \; . \tag{B.7}$$

The posterior $q(\mathbf{\Sigma}_k^{-1})$ can be computed by

$$\log q(\mathbf{\Sigma}_k^{-1}) = \log q(\boldsymbol{\mu}_k, \mathbf{\Sigma}_k^{-1}) - \log q(\boldsymbol{\mu}_k|\mathbf{\Sigma}_k^{-1}) \; . \tag{B.8}$$

Thus we obtain

$$\begin{aligned}
&\log q(\mathbf{\Sigma}_k^{-1}) = \\
&- \frac{\beta_{0k}}{2}(\boldsymbol{\mu}_k - \boldsymbol{m}_{0k})^T\mathbf{\Sigma}_k^{-1}(\boldsymbol{\mu}_k - \boldsymbol{m}_{0k}) \\
&- \frac{1}{2}\sum_{j=1}^{N}\gamma_{jk}(\mathbf{B}_j\mathbf{x}_j - \boldsymbol{\mu}_k)^T\mathbf{\Sigma}_k^{-1}(\mathbf{B}_j\mathbf{x}_j - \boldsymbol{\mu}_k) \\
&- \frac{1}{2}\operatorname{Tr}\left((\mathbf{\Sigma}_k\boldsymbol{W}_{0k})^{-1}\right) + \frac{\nu_{0k}-D-1}{2}\log|\mathbf{\Sigma}_k^{-1}| \\
&+ \frac{\beta_k}{2}(\boldsymbol{\mu}_k - \boldsymbol{m}_k)^T\mathbf{\Sigma}_k^{-1}(\boldsymbol{\mu}_k - \boldsymbol{m}_k) \\
&+ \frac{1}{2}\sum_{j=1}^{N}\gamma_{jk}\log|\mathbf{\Sigma}_k^{-1}| + \text{const} \; .
\end{aligned} \tag{B.9}$$

Making use of the property $\boldsymbol{u}^T\boldsymbol{A}\boldsymbol{u} = \operatorname{Tr}(\boldsymbol{A}\boldsymbol{u}\boldsymbol{u}^T)$ we can

write

$$q(\boldsymbol{\Sigma}_k^{-1}) = \frac{1}{2}\sum_{j=1}^{N}(\gamma_{jk} + \nu_{0k} - D - 1)\log|\boldsymbol{\Sigma}_k^{-1}| - \frac{1}{2}\operatorname{Tr}\Big\{\big(\boldsymbol{W}_{0k}^{-1} + \beta_{0k}(\boldsymbol{\mu}_k - \boldsymbol{m}_{0k})(\boldsymbol{\mu}_k - \boldsymbol{m}_{0k})^T + \sum_{j=1}^{N}\gamma_{jk}(\mathbf{B}_j\mathbf{x}_j - \boldsymbol{\mu}_k)(\mathbf{B}_j\mathbf{x}_j - \boldsymbol{\mu}_k)^T - \beta_k(\boldsymbol{\mu}_k - \boldsymbol{m}_k)(\boldsymbol{\mu}_k - \boldsymbol{m}_k)^T\big)\boldsymbol{\Sigma}_k^{-1}\Big\} + \text{const}\ . \tag{B.10}$$

Substituting B.6 and B.7 into B.10 we obtain

$$q(\boldsymbol{\Sigma}_k^{-1}) = \mathcal{W}(\boldsymbol{\Sigma}_k^{-1}|\boldsymbol{W}_k, \nu_k)\ , \tag{B.11}$$

where

$$\nu_k = \nu_{0k} + s_{0k}\ , \tag{B.12}$$

and

$$\boldsymbol{W}_k^{-1} = \boldsymbol{W}_{0k}^{-1} + \sum_{j=1}^{N}\gamma_{jk}(\mathbf{B}_j\mathbf{x}_j)(\mathbf{B}_j\mathbf{x}_j)^T - \frac{\boldsymbol{s}_{1k}\boldsymbol{s}_{1k}^T}{b_{0k} + s_{0k}} + \frac{b_{0k}s_{0k}\boldsymbol{m}_{0k}\boldsymbol{m}_{0k}^T}{b_{0k} + s_{0k}} - \frac{b_{0k}\boldsymbol{s}_{1k}\boldsymbol{m}_{0k}^T}{b_{0k} + s_{0k}} - \frac{b_{0k}\boldsymbol{m}_{0k}\boldsymbol{s}_{1k}^T}{b_{0k} + s_{0k}}\ . \tag{B.13}$$

Appendix C. Estimating the bias field

In order to estimate optimal parameters to represent the bias field we need, at each iteration of the algorithm, to maximize the lower bound (E.1) on the objective function with respect to the parameters Θ_β. 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

$$\hat{\Theta}_\beta = \arg\max_{\Theta_\beta}\ \Big\{\mathbb{E}_{\mathbf{Z},\Theta_\mu,\Theta_\Sigma}\big[\log p(\mathbf{X}|\mathbf{Z},\Theta_\mu,\Theta_\Sigma,\Theta_\beta)\big] + \log p(\Theta_\beta)\Big\}\ . \tag{C.1}$$

A convenient and fast converging strategy to estimate the bias field, consists in adopting a Gauss-Newton iterative scheme. This involves computing the first and second derivatives of $\mathcal{L}$ with respect to Θ_β. A very similar approach for finding MAP estimates of the bias field parameters has already been described in [4] for the same parameterization of the non-uniformity field adopted in this work. Therefore, further details on how this optimization problem can be solved are omitted.

Appendix D. Estimating the deformations

The deformation field that best matches the population based atlas $\{\boldsymbol{\tau}_t\}_{t=1,\dots,T}$ to an individual image can be estimated by maximizing $\mathcal{L}$ with respect to Θ_α. This is equivalent to finding

$$\hat{\Theta}_\alpha = \arg\max_{\Theta_\alpha}\ \Big\{\mathbb{E}_{\mathbf{Z}}\big[\log p(\mathbf{Z}|\Theta_\pi,\Theta_\alpha)\big] + \log p(\Theta_\alpha)\Big\}\ . \tag{D.1}$$

Solutions can again be obtained making use of a Gauss-Newton scheme. This involves applying the following update rule

$$\Theta_\alpha^{(n+1)} = \Theta_\alpha^{(n)} - \boldsymbol{H}^{-1}\boldsymbol{f}\ , \tag{D.2}$$

where $\boldsymbol{f}$ indicates the gradient and $\boldsymbol{H}$ the Hessian of the objective function. Numerical techniques are required to solve the term $\boldsymbol{H}^{-1}\boldsymbol{f}$ since a very local, and therefore highly dimensional, parameterization of the deformations is adopted in this work. A very efficient strategy consist in using multigrid algorithms. In particular the method presented in this work employs a multigrid scheme with the same implementation described in [2].

Appendix E. Computing the lower bound

The lower bound can be easily evaluated, once the sufficient statistics and the variational posterior distributions have been computed [8], by

$$\begin{aligned}\mathcal{L} = {} & \mathbb{E}_{\mathbf{Z},\Theta_\mu,\Theta_\Sigma}[\log p(\mathbf{X}|\mathbf{Z},\Theta_\mu,\Theta_\Sigma,\Theta_\beta)] \\ & + \mathbb{E}_{\mathbf{Z}}[\log p(\mathbf{Z}|\Theta_\pi,\Theta_\alpha)] \\ & + \mathbb{E}_{\Theta_\mu,\Theta_\Sigma}[\log p(\Theta_\mu,\Theta_\Sigma)] \\ & + \log p(\Theta_\alpha) + \log p(\Theta_\beta) \\ & - \mathbb{E}_{\mathbf{Z}}[\log q(\mathbf{Z})] \\ & - \mathbb{E}_{\Theta_\mu,\Theta_\Sigma}[\log q(\Theta_\mu,\Theta_\Sigma)]\ ,\end{aligned} \tag{E.1}$$

with

$$\begin{aligned}&\mathbb{E}_{\mathbf{Z},\Theta_\mu,\Theta_\Sigma}\big[\log p(\mathbf{X}|\mathbf{Z},\Theta_\mu,\Theta_\Sigma,\Theta_\beta)\big] = \\ &\frac{1}{2}\sum_{k=1}^{K}s_{0k}\,\mathbb{E}\big[\log|\boldsymbol{\Sigma}_k^{-1}|\big] - D\log(2\pi) - \frac{D}{\beta_k} \\ &- \frac{1}{2}\sum_{k=1}^{K}s_{0k}\nu_k\boldsymbol{m}_k^T\boldsymbol{W}_k\boldsymbol{m}_k \\ &- \frac{1}{2}\sum_{k=1}^{K}\nu_k\operatorname{Tr}(\boldsymbol{W}_k\boldsymbol{S}_{2k} - 2\boldsymbol{s}_{1k}\boldsymbol{m}_k^T\boldsymbol{W}_k) \\ &+ \sum_{j=1}^{N}\sum_{k=1}^{K}\gamma_{jk}\log|\mathbf{B}_j|\ .\end{aligned} \tag{E.2}$$

$$\mathbb{E}_{\mathbf{Z}}[\log p(\mathbf{Z}|\Theta_\pi,\Theta_\alpha)] = \sum_{j=1}^{N}\sum_{k=1}^{K}\gamma_{jk}\log\pi_{jk}(\varphi(\Theta_\alpha))\ . \tag{E.3}$$

$$\mathbb{E}_{\Theta_\mu,\Theta_\Sigma}[\log p(\Theta_\mu,\Theta_\Sigma)] =$$
$$+\frac{1}{2}\sum_{k=1}^{K}\left\{D\log\frac{\beta_{0k}}{2\pi}-D\frac{\beta_{0k}}{\beta_k}\right\}+2K\log B_W(\boldsymbol{W}_{0k},\nu_{0k})$$
$$-\sum_{k=1}^{K}\left\{\frac{\nu_k}{2}\mathrm{Tr}\big((\boldsymbol{W}_{0k}^{-1}+\beta_{0k}(\boldsymbol{m}_k-\boldsymbol{m}_{0k})(\boldsymbol{m}_k-\boldsymbol{m}_{0k})^T)\boldsymbol{W}_k\big)\right.$$
$$\left.+\mathbb{E}\big[\log|\boldsymbol{\Sigma}_k^{-1}|\big](\nu_{0k}-D)\right\}. \tag{E.4}$$

$$\mathbb{E}_{\mathbf{Z}}[\log q(\mathbf{Z})] = \sum_{j=1}^{N}\sum_{k=1}^{K}\gamma_{jk}\log\gamma_{jk}\ . \tag{E.5}$$

$$\mathbb{E}_{\Theta_\mu,\Theta_\Sigma}[\log q(\Theta_\mu,\Theta_\Sigma)] =$$
$$\sum_{k=1}^{K}\left\{\frac{1}{2}D\left(\log\frac{\beta_k}{2\pi}-1-\nu_k\right)+\log B_W(\boldsymbol{W}_k,\nu_k)\right.$$
$$\left.+\,\mathbb{E}\big[\log|\boldsymbol{\Sigma}_k^{-1}|\big]\Big(\frac{1}{2}\nu_k-D\Big)\right\}. \tag{E.6}$$

The term $B_W(\boldsymbol{W},\nu)$ in equations (E.4) and (E.6) indicates the normalizing constant for a Wishart distribution parametrized by $\boldsymbol{W}$ and ν.

Appendix F. Derivatives of the lower bound with respect to the intensity prior hyperparameters

Given a data set of M individual scans, the lower bound on the marginal likelihood depends on the hyperparameters $\{\beta_{0k}\}_{k=1,\dots,K}$ as follows

$$\mathcal{L}(\beta_{0k}) = \frac{MD}{2}\log\left(\frac{\beta_{0k}}{2\pi}\right)-\frac{1}{2}\sum_{i=1}^{M}\left\{D\frac{\beta_{0k}}{\beta_{ik}}\right.$$
$$\left.-\,\beta_{0k}\nu_{ik}(\boldsymbol{m}_{ik}-\boldsymbol{m}_{0k})^T\boldsymbol{W}_{ik}(\boldsymbol{m}_{ik}-\boldsymbol{m}_{0k})\right\}$$
$$+\,\text{const}\ , \tag{F.1}$$

and the corresponding gradient and Hessian are given by

$$g_\beta=\frac{MD}{\beta_{0k}}-\frac{1}{2}\sum_{i=1}^{M}\left\{\frac{D}{\beta_{ik}}-\nu_{ik}(\boldsymbol{m}_{ik}-\boldsymbol{m}_{0k})^T\boldsymbol{W}_{ik}(\boldsymbol{m}_{ik}-\boldsymbol{m}_{0k})\right\},$$
$$\boldsymbol{H}_\beta=-\frac{MD}{2\beta_{0k}^2}. \tag{F.2}$$

Similarly for $\{\boldsymbol{m}_{0k}\}_{k=1,\dots,K}$ we find that $\mathcal{L}(\boldsymbol{m}_{0k})$ can be expressed as

$$\mathcal{L}(\boldsymbol{m}_{ok})=\frac{1}{2}\sum_{i=1}^{M}\beta_{0k}\nu_{ik}(\boldsymbol{m}_{ik}-\boldsymbol{m}_{0k})^T\boldsymbol{W}_{ik}(\boldsymbol{m}_{ik}-\boldsymbol{m}_{0k})+\text{const}. \tag{F.3}$$

The first and second derivatives are instead

$$g_{\boldsymbol{m}} = -\sum_{i=1}^{M}\beta_{0k}\nu_{ik}(\boldsymbol{m}_{ik}-\boldsymbol{m}_{0k})^T\boldsymbol{W}_{ik}\ ,$$
$$\boldsymbol{H}_{\boldsymbol{m}} = \sum_{i=1}^{M}\beta_{0k}\nu_{ik}\boldsymbol{W}_{ik}\ . \tag{F.4}$$

The following indicates the dependency of $\mathcal{L}$ on the degrees of freedom of the Wishart priors

$$\mathcal{L}(\nu_{0k}) = \sum_{i=1}^{M}\frac{\nu_{0k}}{2}\,\mathbb{E}\big[\log|\boldsymbol{\Sigma}_{ik}^{-1}|\big]+M\log|\boldsymbol{W}_{0k}|^{-\frac{\nu_{0k}}{2}}$$
$$+\,M\log\left(2^{\frac{D\nu_{0k}}{2}}\,\pi^{\frac{D(D-1)}{4}}\prod_{d=1}^{D}\Gamma\left(\frac{\nu_{0k}+1-d}{2}\right)\right)^{-1}$$
$$+\,\text{const}\ . \tag{F.5}$$

In this case the gradient and Hessian can be computed by

$$g_\nu = \frac{1}{2}\sum_{i=1}^{M}\mathbb{E}\big[\log|\boldsymbol{\Sigma}_{ik}^{-1}|\big]$$
$$-\,\frac{M}{2}\left\{\log|\boldsymbol{W}_{0k}|+D\log 2+\sum_{d=1}^{D}\varphi\left(\frac{\nu_{0k}+1-d}{2}\right)\right\}\ ,$$
$$\boldsymbol{H}_\nu = M\varphi_1\left(\frac{\nu_{0k}+1-d}{2}\right)\ . \tag{F.6}$$

Finally for the Wishart scale matrices we find that

$$\mathcal{L}(\boldsymbol{W}_{0k}) = M\nu_{0k}\log|\boldsymbol{C}_{0k}|$$
$$-\,\frac{1}{2}\sum_{i=1}^{M}\nu_{ik}\,\mathrm{Tr}(\boldsymbol{C}_{0k}^T\boldsymbol{W}_{ik}\boldsymbol{C}_{0k})+\text{const}\ , \tag{F.7}$$

where $\boldsymbol{C}_{0k}$ is the Cholesky factor of $\boldsymbol{W}_{0k}^{-1}$

$$\boldsymbol{W}_{0k}^{-1} = \boldsymbol{C}_{0k}\boldsymbol{C}_{0k}^T\ . \tag{F.8}$$

The first and second derivatives are given by

$$g_{\boldsymbol{W}} = M\nu_{0k}\,\mathrm{diag}(1/C_{11},\dots,1/C_{DD})-\sum_{i=1}^{M}\nu_{ik}\boldsymbol{W}_{ik}\boldsymbol{C}_{0k}\ ,$$
$$\boldsymbol{H}_{\boldsymbol{W}} = -M\nu_{0k}\,\mathrm{diag}(1/C_{11}^2,\dots,1/C_{DD}^2)-\sum_{i=1}^{M}\nu_{ik}\boldsymbol{W}_{ik}\ . \tag{F.9}$$

Acknowledgments

Claudia Blaiotta is jointly funded by a UCL Impact Studentship and Balgrist University Hospital, Zurich. The Wellcome Trust Centre for Neuroimaging is supported by core funding from the Wellcome Trust [grant number 091593/Z/10/Z].

References

[1] Andrieu, C., De Freitas, N., Doucet, A., Jordan, M. I., 2003. An introduction to MCMC for machine learning. Machine learning 50 (1-2), 5–43.

[2] Ashburner, J., 2007. A fast diffeomorphic image registration algorithm. Neuroimage 38 (1), 95–113.

[3] Ashburner, J., Friston, K. J., 2000. Voxel-based morphometry: the methods. Neuroimage 11 (6), 805–821.

[4] Ashburner, J., Friston, K. J., 2005. Unified segmentation. Neuroimage 26 (3), 839–851.

[5] Attias, H., 1999. Inferring parameters and structure of latent variable models by variational Bayes. In: Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence. Morgan Kaufmann Publishers Inc., pp. 21–30.

[6] Aubert-Broche, B., Griffin, M., Pike, G. B., Evans, A. C., Collins, D. L., 2006. Twenty new digital brain phantoms for creation of validation image data bases. Medical Imaging, IEEE Transactions on 25 (11), 1410–1416.

[7] Bishop, C. M., Lasserre, J., et al., 2007. Generative or discriminative? getting the best of both worlds. Bayesian statistics 8, 3–24.

[8] Bishop, C. M., et al., 2006. Pattern Recognition and Machine Learning. Vol. 1. Springer New York.

[9] Cocosco, C. A., Kollokian, V., Kwan, R. K.-S., Pike, G. B., Evans, A. C., 1997. Brainweb: Online interface to a 3D MRI simulated brain database. In: NeuroImage. Citeseer.

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

[11] Corduneanu, A., Bishop, C. M., 2001. Variational Bayesian model selection for mixture distributions. In: Artificial intelligence and Statistics. Vol. 2001. Morgan Kaufmann Waltham, MA, pp. 27–34.

[12] da Silva, A. R. F., 2009. Bayesian mixture models of variable dimension for image segmentation. computer methods and programs in biomedicine 94 (1), 1–14.

[13] Dugas-Phocion, G., Ballester, M. A. G., Malandain, G., Lebrun, C., Ayache, N., 2004. Improved EM-based tissue segmentation and partial volume effect quantification in multi-sequence brain MRI. In: Proc. Medical Image Computing and Computer Assisted Intervention, MICCAI 2004. Springer, pp. 26–33.

[14] DAgostino, E., Maes, F., Vandermeulen, D., Suetens, P., 2006. A unified framework for atlas based brain image segmentation and registration. In: Biomedical Image Registration. Springer, pp. 136–143.

[15] Fan, A. C., Fisher III, J. W., Wells III, W. M., Levitt, J. J., Willsky, A. S., 2007. MCMC curve sampling for image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2007. Springer, pp. 477–485.

[16] Fischl, B., van der Kouwe, A., Destrieux, C., Halgren, E., Ségonne, F., Salat, D. H., Busa, E., Seidman, L. J., Goldstein, J., Kennedy, D., et al., 2004. Automatically parcellating the human cerebral cortex. Cerebral cortex 14 (1), 11–22.

[17] Giorgio, A., De Stefano, N., 2013. Clinical use of brain volumetry. Journal of Magnetic Resonance Imaging 37 (1), 1–14.

[18] Greenspan, H., Ruf, A., Goldberger, J., 2006. Constrained Gaussian mixture model framework for automatic segmentation of MR brain images. Medical Imaging, IEEE Transactions on 25 (9), 1233–1245.

[19] Guillemaud, R., Brady, M., 1997. Estimating the bias field of MR images. Medical Imaging, IEEE Transactions on 16 (3), 238–251.

[20] Hoffman, M. D., Blei, D. M., Wang, C., Paisley, J., 2013. Stochastic variational inference. The Journal of Machine Learning Research 14 (1), 1303–1347.

[21] Iglesias, J. E., Sabuncu, M. R., Van Leemput, K., 2012. Incorporating parameter uncertainty in bayesian segmentation models: Application to hippocampal subfield volumetry. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2012. Springer, pp. 50–57.

[22] Kato, Z., 2008. Segmentation of color images via reversible jump MCMC sampling. Image and Vision Computing 26 (3), 361–371.

[23] Klauschen, F., Goldman, A., Barra, V., Meyer-Lindenberg, A., Lundervold, A., 2009. Evaluation of automated brain MR image segmentation and volumetry methods. Human brain mapping 30 (4), 1310–1327.

[24] Kovacevic, N., Lobaugh, N., Bronskill, M., Levine, B., Feinstein, A., Black, S., 2002. A robust method for extraction and automatic segmentation of brain images. Neuroimage 17 (3), 1087–1100.

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

[26] Lawrence, N. D., Platt, J. C., 2004. Learning to learn with the informative vector machine. In: Proceedings of the twenty-first international conference on Machine learning. ACM, p. 65.

[27] Liang, Z., MacFall, J. R., Harrington, D. P., 1994. Parameter estimation and tissue segmentation from multispectral MR images. Medical Imaging, IEEE Transactions on 13 (3), 441–449.

[28] Lorenzo-Valdés, M., Sanchez-Ortiz, G. I., Elkington, A. G., Mohiaddin, R. H., Rueckert, D., 2004. Segmentation of 4D cardiac MR images using a probabilistic atlas and the EM algorithm. Medical Image Analysis 8 (3), 255–265.

[29] Marcus, D. S., Wang, T. H., Parker, J., Csernansky, J. G., Morris, J. C., Buckner, R. L., 2007. Open access series of imaging studies (oasis): cross-sectional mri data in young, middle aged, nondemented, and demented older adults. Journal of cognitive neuroscience 19 (9), 1498–1507.

[30] Mazzara, G. P., Velthuizen, R. P., Pearlman, J. L., Greenberg, H. M., Wagner, H., 2004. Brain tumor target volume determination for radiation treatment planning through automated MRI segmentation. International Journal of Radiation Oncology* Biology* Physics 59 (1), 300–312.

[31] Moon, N., Bullitt, E., Van Leemput, K., Erig, G., 2002. Model-based brain and tumor segmentation. In: Pattern Recognition, 2002. Proceedings. 16th International Conference on. Vol. 1. IEEE, pp. 528–531.

[32] Noe, A., Gee, J. C., 2001. Partial volume segmentation of cerebral MRI scans with mixture model clustering. In: Information Processing in Medical Imaging. Springer, pp. 423–430.

[33] Nyúl, L. G., Udupa, J. K., Zhang, X., 2000. New variants of a method of MRI scale standardization. Medical Imaging, IEEE Transactions on 19 (2), 143–150.

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

[35] Prastawa, M., Bullitt, E., Ho, S., Gerig, G., 2004. A brain tumor segmentation framework based on outlier detection. Medical image analysis 8 (3), 275–283.

[36] Raina, R., Ng, A. Y., Koller, D., 2006. Constructing informative priors using transfer learning. In: Proceedings of the 23rd international conference on Machine learning. ACM, pp. 713–720.

[37] Rajapakse, J. C., Kruggel, F., 1998. Segmentation of MR images with intensity inhomogeneities. Image and Vision Computing 16 (3), 165–180.

[38] Seeger, M., 2002. Covariance kernels from bayesian generative models. Advances in neural information processing systems 2, 905–912.

[39] Simpson, I. J., Schnabel, J. A., Groves, A. R., Andersson, J. L., Woolrich, M. W., 2012. Probabilistic inference of regularisation in non-rigid registration. NeuroImage 59 (3), 2438–2451.

[40] Tian, G., Xia, Y., Zhang, Y., Feng, D., 2011. Hybrid genetic and variational expectation-maximization algorithm for Gaussian-mixture-model-based brain MR image segmentation. Information Technology in Biomedicine, IEEE Transactions on 15 (3), 373–380.

[41] Tzikas, D. G., Likas, A. C., Galatsanos, N. P., 2008. The variational approximation for bayesian inference. Signal Processing Magazine, IEEE 25 (6), 131–146.

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

[43] Van Leemput, K., Maes, F., Vandermeulen, D., Suetens, P., 2003. A unifying framework for partial volume segmentation of brain MR images. Medical Imaging, IEEE Transactions on 22 (1), 105–119.

[44] Wells III, W. M., Grimson, W. E. L., Kikinis, R., Jolesz, F. A., 1996. Adaptive segmentation of MRI data. Medical Imaging, IEEE Transactions on 15 (4), 429–442.

[45] Woolrich, M. W., Behrens, T. E., 2006. Variational Bayes inference of spatial mixture models for segmentation. Medical Imaging, IEEE Transactions on 25 (10), 1380–1391.

[46] Woolrich, M. W., Jbabdi, S., Patenaude, B., Chappell, M., Makni, S., Behrens, T., Beckmann, C., Jenkinson, M., Smith, S. M., 2009. Bayesian analysis of neuroimaging data in FSL. Neuroimage 45 (1), S173–S186.

[47] Wyatt, P. P., Noble, J. A., 2003. MAP MRF joint segmentation and registration of medical images. Medical Image Analysis 7 (4), 539–552.

[48] Xiaohua, C., Brady, M., Rueckert, D., 2004. Simultaneous segmentation and registration for medical image. In: Proc. Medical Image Computing and Computer Assisted Intervention, MICCAI 2004. Springer, pp. 663–670.

[49] Yezzi, A., Zollei, L., Kapur, T., 2001. A variational framework for joint segmentation and registration. In: Mathematical Methods in Biomedical Image Analysis, MMBIA 2001, IEEE Workshop on. IEEE, pp. 44–51.

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