

# Joint inference on structural and diffusion MRI for sequence-adaptive Bayesian segmentation of thalamic nuclei with probabilistic atlases

Juan Eugenio Iglesias<sup>1,2,3</sup>, Koen Van Leemput<sup>2,4</sup>, Polina Golland<sup>3</sup>, Anastasia Yendiki<sup>2</sup>

<sup>1</sup> Centre for Medical Image Computing, Department of Medical Physics and Biomedical Engineering, University College London, United Kingdom

<sup>2</sup> Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital and Harvard Medical School, USA

<sup>3</sup> Computer Science and Artificial Intelligence Laboratory, MIT, USA

<sup>4</sup> Department of Health Technology, Technical University of Denmark

**Abstract.** Segmentation of structural and diffusion MRI (sMRI/dMRI) is usually performed independently in neuroimaging pipelines. However, some brain structures (e.g., globus pallidus, thalamus and its nuclei) can be extracted more accurately by fusing the two modalities. Following the framework of Bayesian segmentation with probabilistic atlases and unsupervised appearance modeling, we present here a novel algorithm to jointly segment multi-modal sMRI/dMRI data. We propose a hierarchical likelihood term for the dMRI defined on the unit ball, which combines the Beta and Dimroth-Scheidegger-Watson distributions to model the data at each voxel. This term is integrated with a mixture of Gaussians for the sMRI data, such that the resulting joint unsupervised likelihood enables the analysis of multi-modal scans acquired with any type of MRI contrast, b-values, or number of directions, which enables wide applicability. We also propose an inference algorithm to estimate the maximum-a-posteriori model parameters from input images, and to compute the most likely segmentation. Using a recently published atlas derived from histology, we apply our method to thalamic nuclei segmentation on two datasets: HCP (state of the art) and ADNI (legacy) – producing lower sample sizes than Bayesian segmentation with sMRI alone.

## 1 Introduction

Automated segmentation of MRI scans is a prerequisite for most human neuroimaging studies. Most of the algorithms commonly used for this task rely solely on structural MRI (sMRI) scans, and belong to one of three categories: Bayesian segmentation with a probabilistic atlas (e.g., [1, 2]); multi-atlas segmentation [3]; and, more recently, convolutional neural networks (e.g., [4]). Typically, these techniques segment the brain into tissue types (i.e., gray matter, white matter, and cerebrospinal fluid), or into finer anatomical structures (e.g., hippocampus, ventricle). Bayesian methods drive the primary segmentation modules of the most widespread neuroimaging packages, like FreeSurfer [2], FSL [5], or SPM [1].

The aforementioned approaches rely mostly on T1 contrast to distinguish between gray and white matter. However, some boundaries between structures are nearly invisible in T1 (and other structural MR contrasts) due to insufficient difference in proton density and relaxation times. This is exacerbated by lower contrast-to-noise ratio in deep-brain structures, due to greater distance from the head coil. Two examples from the state-of-the-art Human Connectome Project (HCP) dataset [6] are shown in Fig. 1. In the first example, the lateral boundary of the thalamus appears very faint (Fig. 1a). In the second, the lateral boundary of the globus pallidus is visible thanks to the contrast with the neighboring putamen, but the medial boundary is not (Fig. 1c).

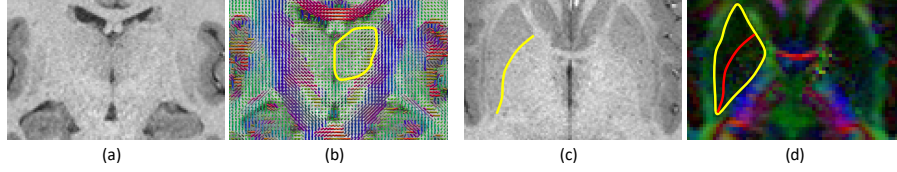
These issues create a need for fusing data from several MR modalities to better delineate structure boundaries. A natural complement to sMRI is diffusion MRI (dMRI), which may help discriminate between certain tissue types, despite its lower resolution. For example, in Fig. 1b, the lateral boundary of the thalamus is clearly discernible in the principal diffusion direction map obtained from dMRI. The diffusion data also complement the T1 scan in the pallidum, which can be delineated by combining contours obtained from the two modalities (medial from dMRI, lateral from sMRI, see Fig. 1d).

Most prior work on segmentation of dMRI focuses on delineating white matter structures, using tractography [7–9] or volumetric segmentation [10, 11]. Tractography has also been used to subdivide subcortical structures (e.g., thalamus [12], amygdala [13]) based on long-range connections. Surprisingly, the literature on *joint* modeling of multimodal sMRI/dMRI is sparse. When sMRI and dMRI are used by the same tool, this is most often done serially, e.g., a segmentation derived from sMRI is used to analyze the dMRI (e.g., to derive priors for Bayesian tractography [9]). To the best of our knowledge, the only works analyzing sMRI and dMRI simultaneously have been on thalamic nuclei segmentation with random forests [14, 15]. The main concern with such discriminative techniques is their generalization ability to other datasets, which is limited by differences in MRI acquisition. For sMRI segmentation, this problem can be ameliorated with data augmentation and pretraining [4]. However, this is harder to do in dMRI, where acquisition protocols are much less standardized.

The ability to generalize across datasets is critical when software is released publicly and few assumptions can be made on the acquisition. In such scenarios, Bayesian segmentation methods that automatically estimate appearance models from input images remain very popular, as they are agnostic to the MRI contrast of the input scan, and thus robust to acquisition differences. These methods are used for tissue segmentation by major neuroimaging packages (e.g., Unified Segmentation [1] in SPM, and FAST [16] in FSL). However, they can be inaccurate when segmenting structures with poor sMRI contrast (see Fig. 1).

Here we propose a sequence-adaptive Bayesian algorithm that uses a probabilistic atlas to segment sMRI and dMRI data *simultaneously*<sup>5</sup>. This is achieved via a novel dMRI likelihood term, which relies on a hierarchical model for the

<sup>5</sup> Henceforth, we use “Bayesian segmentation” to refer to this specific family of Bayesian methods, using probabilistic atlases and unsupervised appearance modeling.



**Fig. 1.** (a) Coronal plane across the thalami of a T1 scan from the HCP. (b) Corresponding map of principal diffusion directions, and manual delineation of the left thalamus. (c) Axial plane of the T1 scan across the basal ganglia, and manual delineation of the boundary between the putamen and globus pallidus. (d) Corresponding principal diffusion directions (weighted by FA), with the joint boundary of the putamen and pallidum (in yellow, visible in this map) and the contour from the T1 (in red).

fractional anisotropy (FA) and principal diffusion orientation. Combined with a Gaussian likelihood for sMRI, this model of image intensities is flexible enough to produce accurate segmentations, while keeping dimensionality low. We also propose a novel inference algorithm to automatically segment scans by fitting the model to multi-modal sMRI/dMRI data. Thanks to unsupervised intensity modeling, applicability across a wide range of acquisition protocols is achieved, which is demonstrated experimentally on two considerably different datasets.

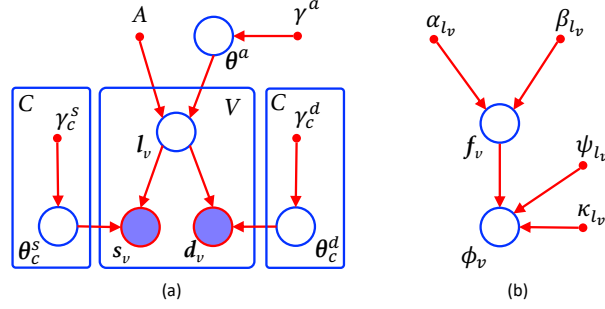
## 2 Methods

### 2.1 Forward probabilistic model

The graphical model of our framework is shown in Fig. 2a. The observed variables are a bias field corrected (possibly multispectral) sMRI scan  $\mathbf{S} = [\mathbf{s}_1, \dots, \mathbf{s}_V]$  defined on  $V$  voxels, a dMRI scan  $\mathbf{D} = [\mathbf{d}_1, \dots, \mathbf{d}_V]$  defined at the same voxel coordinates (which might require resampling), and a probabilistic atlas  $\mathbf{A}$ , which provides the probabilities of observing one of  $C$  neuroanatomical classes at every location across a reference spatial coordinate system. The model is governed by three sets of deterministic hyperparameters specified by the user:  $\gamma^a$ ,  $\gamma^s$  and  $\gamma^d$ .

At the top of the generative model we find the atlas  $\mathbf{A}$ , along with a set of related parameters  $\theta^a$  that deform this atlas into the space of the MRI data. These parameters are a sample of a distribution that regularizes the deformation field by penalizing, e.g., its bending energy. The strength of the regularization is controlled by the set of hyperparameters  $\gamma^a$ .

Given the deformed atlas, a labeling (segmentation)  $\mathbf{L} = [l_1, \dots, l_V]$ , with  $l_v \in \{1, \dots, C\}$ , is obtained by independently sampling the categorical distribution defined by the deformed atlas at each voxel location. Given  $\mathbf{L}$ , the observed sMRI and dMRI data are assumed to be conditionally independent from each other and across voxels. The sMRI data  $\mathbf{s}_v$  at  $v$  follows a distribution (typically a Gaussian) whose parameters  $\theta_c^s$  depend on the corresponding label  $c = l_v$ . Any prior knowledge on these parameters is encoded in their priors, which are governed by hyperparameter vectors  $\{\gamma_c^s\}$ . Similarly,  $\mathbf{d}_v$  is also assumed to be



**Fig. 2.** (a) Graphical model of proposed framework. (b) Hierarchical model of dMRI likelihood. Circles represent random variables (open if hidden, shaded if observed). Smaller solid circles are deterministic parameters. Plates indicate replication.

a mixture conditioned on the segmentation, described by parameters  $\{\theta_c^d\}$  and hyperparameters  $\{\gamma_c^d\}$ , which yields a symmetric likelihood model (Fig. 2a). The joint probability density function (PDF) of the model is therefore:

$$\begin{aligned}
 & p(\mathbf{S}, \mathbf{D}, \mathbf{L}, \boldsymbol{\theta}^a, \boldsymbol{\theta}^s, \boldsymbol{\theta}^d | \mathbf{A}, \boldsymbol{\gamma}^a, \boldsymbol{\gamma}^s, \boldsymbol{\gamma}^d) \\
 &= p(\mathbf{S} | \mathbf{L}, \boldsymbol{\theta}^s) p(\mathbf{D} | \mathbf{L}, \boldsymbol{\theta}^d) p(\mathbf{L} | \mathbf{A}, \boldsymbol{\theta}^a) p(\boldsymbol{\theta}^s | \boldsymbol{\gamma}^s) p(\boldsymbol{\theta}^d | \boldsymbol{\gamma}^d) p(\boldsymbol{\theta}^a | \boldsymbol{\gamma}^a) \\
 &= \left( \prod_{v=1}^V p(s_v | \theta_{l_v}^s) p(d_v | \theta_{l_v}^d) p_v(l_v | \mathbf{A}, \boldsymbol{\theta}^a) \right) \left( \prod_{c=1}^C p(\theta_c^s | \gamma_c^s) p(\theta_c^d | \gamma_c^d) \right) p(\boldsymbol{\theta}^a | \boldsymbol{\gamma}^a), \quad (1)
 \end{aligned}$$

where  $\boldsymbol{\theta}^s = \{\theta_c^s\}_{c=1, \dots, C}$ ,  $\boldsymbol{\gamma}^s = \{\gamma_c^s\}_{c=1, \dots, C}$ ,  $\boldsymbol{\theta}^d = \{\theta_c^d\}_{c=1, \dots, C}$ ,  $\boldsymbol{\gamma}^d = \{\gamma_c^d\}_{c=1, \dots, C}$ .

## 2.2 Model instantiation

**Probabilistic atlas:** We follow the framework of the thalamic atlas [17] that we use in the experiments in Section 3, in which the atlas is encoded as a tetrahedral mesh. Deforming the mesh is penalized by a regularizer  $R$ , weighted by the mesh stiffness  $\lambda$ . The prior is given by (see further details in [18]):

$$p(\boldsymbol{\theta}^a | \boldsymbol{\gamma}^a) \propto \exp[-\lambda R(\boldsymbol{\theta}^a)], \quad \text{and} \quad l_v \sim \text{Cat}[\mathbf{A}_v(\boldsymbol{\theta}^a)], \quad (2)$$

where  $\mathbf{A}_v(\boldsymbol{\theta}^a) = [A_{v1}, \dots, A_{vC}]^t$  is simply the vector of  $C$  label probabilities provided by the atlas at voxel  $v$  when deformed with parameters  $\boldsymbol{\theta}^a$ ;  $\text{Cat}[\cdot]$  is the categorical distribution; and hyperparameters  $\boldsymbol{\gamma}^a$  comprise just  $\lambda$ , i.e.,  $\boldsymbol{\gamma}^a = \{\lambda\}$ .

**Likelihood of sMRI:** In order to model the sMRI intensities given the segmentation  $\mathbf{L}$ , we follow the Bayesian brain MR segmentation literature (e.g., [1, 16, 17]) and use Gaussian intensity distributions, such that  $\boldsymbol{\theta}_c^s = \{\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c\}$ , i.e., the mean and covariance of the intensities of class  $c$ . We place a Normal Inverse Wishart (NIW) distribution on these parameters (i.e., the conjugate prior), with zero degrees of freedom for the covariance, as we found it difficult in practice to

inform such parameter a priori. Therefore, we have:  $\gamma_c^s = \{\mathbf{M}_c, n_c\}$ , where  $\mathbf{M}_c$  is the hypermean and  $n_c$  is the scale. The sMRI likelihood is thus:

$$\begin{aligned} \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c \mid \mathbf{M}_c, n_c &\sim \text{NIW}(\mathbf{M}_c, n_c, 0, 0\mathbf{I}), \\ \mathbf{s}_v \mid l_v, \{\boldsymbol{\mu}_c\}, \{\boldsymbol{\Sigma}_c\} &\sim \mathcal{N}(\boldsymbol{\mu}_{l_v}, \boldsymbol{\Sigma}_{l_v}), \end{aligned} \quad (3)$$

where  $\mathcal{N}(\cdot, \cdot)$  is the Gaussian distribution and  $\mathbf{I}$  is the identity matrix.

**Likelihood of dMRI:** We have two requirements for the likelihood function of the dMRI: low demands on gradient directions and b-values to accommodate legacy data; and low number of parameters to facilitate unsupervised clustering (yet sufficient to separate the classes). To satisfy the first requirement, we adopt the diffusion tensor imaging (DTI) model, which can be fit from virtually all available dMRI data. Rather than modeling the tensors directly (e.g., with a Wishart distribution, which we found in pilot experiments to fade too quickly from its mode), we use a hierarchical model (Fig. 2b) that only considers the FA  $f_v$  and the principal eigenvector  $\boldsymbol{\phi}_v$  at each voxel, i.e.,  $\mathbf{d}_v = \{f_v, \boldsymbol{\phi}_v\}$ .

At the first level, we model the FA conditioned on the class, with Beta distributions parameterized by  $\{\alpha_c, \beta_c\}$ . We chose the Beta because it can model location and dispersion of signals defined on the  $[0,1]$  interval with two parameters. At the second level, we model the principal eigenvector with the Dimroth-Scheidegger-Watson (DSW) distribution, which is axial (i.e., antipodally symmetric), accommodating the directional invariance of dMRI [19]. This distribution is also rotationally symmetric around a mean direction  $\boldsymbol{\psi}$  and its opposite  $-\boldsymbol{\psi}$  ( $\|\boldsymbol{\psi}\| = 1$ ), with a dispersion around the mean parameterized by a concentration  $\kappa$ . It has fewer parameters than other axial distributions, such as the (non rotationally symmetric) Bingham. Its PDF is given by [20]:

$$f(\boldsymbol{\phi}; \boldsymbol{\psi}, \kappa) = [Z(\kappa)]^{-1} \exp[\kappa(\boldsymbol{\psi}^t \boldsymbol{\phi})^2], \quad (4)$$

with domain  $\|\boldsymbol{\phi}\| = 1$ , and where the partition function is the Kummer function in 3D [20]:  $Z\kappa = \int_0^1 \exp(\kappa t^2) dt$ . We further assume that the concentration is modulated (multiplied) by the FA. This is a simple yet effective way of modeling the higher directional dispersion in voxels with low FA (e.g., in areas of unrestricted diffusion or fiber crossings), without having to resort to mixtures or additional parameters. The overall model for the dMRI likelihood is thus:

$$\begin{aligned} f_v \mid l_v, \{\alpha_c\}, \{\beta_c\} &\sim \text{Beta}(\alpha_{l_v}, \beta_{l_v}), \\ \boldsymbol{\phi}_v \mid l_v, f_v, \{\boldsymbol{\psi}_c\}, \{\kappa_c\} &\sim \text{DSW}(\boldsymbol{\psi}_{l_v}, f_v \kappa_{l_v}), \end{aligned} \quad (5)$$

and the set of parameters is thus:  $\boldsymbol{\theta}_c^d = \{\alpha_c, \beta_c, \boldsymbol{\psi}_c, \kappa_c\}$ , with  $\|\boldsymbol{\psi}_c\| = 1, \forall c$ . We decided not to inform these parameters, such that  $p(\boldsymbol{\theta}_c^d) \propto 1$ , and  $\gamma_c^d = \emptyset$ . We note that this likelihood model defines a PDF on the unit ball for vector  $f_v \boldsymbol{\phi}_v$ .

### 2.3 Segmentation as Bayesian inference

Within our joint generative model of sMRI and dMRI, we pose segmentation as an optimization problem, seeking to maximize the posterior probability of the

labeling, given the known hyperparameters and observed input data:

$$\begin{aligned} \operatorname{argmax}_{\mathbf{L}} \int \int \int p(\mathbf{L}|\boldsymbol{\theta}^a, \boldsymbol{\theta}^s, \boldsymbol{\theta}^d, \mathbf{S}, \mathbf{D}, \mathbf{A}) p(\boldsymbol{\theta}^a, \boldsymbol{\theta}^s, \boldsymbol{\theta}^d | \mathbf{S}, \mathbf{D}, \mathbf{A}, \boldsymbol{\gamma}^a, \boldsymbol{\gamma}^s) d\boldsymbol{\theta}^a d\boldsymbol{\theta}^s d\boldsymbol{\theta}^d, \\ \approx \operatorname{argmax}_{\mathbf{L}} p(\mathbf{L}|\hat{\boldsymbol{\theta}}^a, \hat{\boldsymbol{\theta}}^s, \hat{\boldsymbol{\theta}}^d, \mathbf{S}, \mathbf{D}, \mathbf{A}), \end{aligned} \quad (6)$$

where we have made the standard approximation that the posterior distribution of the parameters is heavily peaked around point estimates  $\hat{\boldsymbol{\theta}}^a, \hat{\boldsymbol{\theta}}^s, \hat{\boldsymbol{\theta}}^d$  given by:

$$\{\hat{\boldsymbol{\theta}}^a, \hat{\boldsymbol{\theta}}^s, \hat{\boldsymbol{\theta}}^d\} = \operatorname{argmax}_{\{\boldsymbol{\theta}^a, \boldsymbol{\theta}^s, \boldsymbol{\theta}^d\}} p(\boldsymbol{\theta}^a, \boldsymbol{\theta}^s, \boldsymbol{\theta}^d | \mathbf{S}, \mathbf{D}, \mathbf{A}, \boldsymbol{\gamma}^a, \boldsymbol{\gamma}^s). \quad (7)$$

Therefore, we segment a scan by first estimating the parameters with Eq. 7, and then obtaining the (approximate) most likely labeling with Eq. 6.

Applying Bayes rule to Eq. 7, marginalizing over the hidden segmentation  $\mathbf{L}$ , and considering the structure of the model and our design choices, we obtain:

$$\{\hat{\boldsymbol{\theta}}^a, \hat{\boldsymbol{\theta}}^s, \hat{\boldsymbol{\theta}}^d\} = \operatorname{argmax}_{\{\boldsymbol{\theta}^a, \boldsymbol{\theta}^s, \boldsymbol{\theta}^d\}} p(\boldsymbol{\theta}^a | \boldsymbol{\gamma}^a) p(\boldsymbol{\theta}^s | \boldsymbol{\gamma}^s) \sum_{\mathbf{L}} p(\mathbf{S} | \mathbf{L}, \boldsymbol{\theta}^s) p(\mathbf{D} | \mathbf{L}, \boldsymbol{\theta}^d) p(\mathbf{L} | \boldsymbol{\theta}^a, \mathbf{A}).$$

Expanding and taking logarithm, we obtain the following objective function:

$$\begin{aligned} O(\boldsymbol{\theta}^a, \{\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c, \alpha_c, \beta_c, \boldsymbol{\psi}_c, \kappa_c\}) = \log p(\boldsymbol{\theta}^a | \boldsymbol{\gamma}^a) + \sum_{c=1}^C \log p(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c | \mathbf{M}_c, n_c) \\ + \sum_{v=1}^V \log \left[ \sum_{c=1}^C p(\mathbf{s}_v | \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c) p(f_v | \alpha_c, \beta_c) p(\boldsymbol{\phi}_v | f_v, \boldsymbol{\psi}_c, \kappa_c) p(l_v = c | \mathbf{A}, \boldsymbol{\theta}^a) \right]. \end{aligned} \quad (8)$$

We maximize Eq. 8 with a Generalized Expectation Maximization (GEM) algorithm [21], iterating between expectation (E) and maximization (M) steps:

**E step:** In the E step, we use Jensen's inequality to build a lower bound for the objective function, which touches it at the current value of the parameters:

$$\begin{aligned} O \geq Q = \sum_{v=1}^V \sum_{c=1}^C w_{vc} \log [p(\mathbf{s}_v | \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c) p(f_v | \alpha_c, \beta_c) p(\boldsymbol{\phi}_v | f_v, \boldsymbol{\psi}_c, \kappa_c) p(l_v = c | \mathbf{A}, \boldsymbol{\theta}^a)] \\ + \log p(\boldsymbol{\theta}^a | \boldsymbol{\gamma}^a) + \sum_{c=1}^C \log p(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c | \mathbf{M}_c, n_c) - \sum_{v=1}^V \sum_{c=1}^C w_{vc} \log w_{vc}, \end{aligned} \quad (9)$$

where  $\{w_{vc}\}$  a soft segmentation according to the current parameter estimates:

$$\begin{aligned} w'_{vc} &= p(\mathbf{s}_v | \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c) p(f_v | \alpha_c, \beta_c) p(\boldsymbol{\phi}_v | f_v, \boldsymbol{\psi}_c, \kappa_c) p(l_v = c | \mathbf{A}, \boldsymbol{\theta}^a) \\ &= |\boldsymbol{\Sigma}_c|^{-1/2} \exp[-\frac{1}{2}(\mathbf{s}_v - \boldsymbol{\mu}_c)^t \boldsymbol{\Sigma}_c^{-1}(\mathbf{s}_v - \boldsymbol{\mu}_c)] f_v^{\alpha_c-1} (1 - f_v)^{\beta_c-1} [B(\alpha_c, \beta_c)]^{-1} \\ &\quad \times [Z(f_v \kappa_c)]^{-1} \exp[f_v \kappa_c (\boldsymbol{\psi}_c^t \boldsymbol{\phi}_v)^2] A_{vc}(\boldsymbol{\theta}^a), \quad \text{and } w_{vc} = w'_{vc} / \sum_{c'=1}^C w'_{vc'}, \end{aligned} \quad (10)$$

where  $B$  is the Beta function.

**M step:** In the generalized M step, we seek to improve the lower bound Q in Eq. 9. While optimizing the bound with respect to all parameters simultaneously is difficult, optimizing different subsets each time (coordinate ascent) is feasible.

*Optimizing  $\theta^a$ :* Fixing all other parameters and switching signs, we obtain:

$$\operatorname{argmin}_{\theta^a} \sum_{v=1}^V \sum_{c=1}^C w_{vc} \log [w_{vc}/A_{vc}(\theta^a)] + \lambda R(\theta^a). \quad (11)$$

This is a registration problem combining the regularizer  $R$  with a data term: the Kullback–Leibler (KL) divergence between the deformed atlas and the current soft segmentation. We solve it numerically with the conjugate gradient method.

*Optimizing  $\{\mu_c, \Sigma_c\}$ :* Setting derivatives to zero yields a closed-form solution:

$$\mu_c = \frac{n_c M_c + \sum_{v=1}^V w_{vc} s_v}{n_c + \sum_{v=1}^V w_{vc}}, \quad (12)$$

$$\Sigma_c = \frac{n_c(\mu_c - M_c)(\mu_c - M_c)^t + \sum_{v=1}^V w_{vc}(s_v - \mu_c)(s_v - \mu_c)^t}{1 + \sum_{v=1}^V w_{vc}}. \quad (13)$$

*Optimizing  $\{\alpha_c, \beta_c\}$ :* Substituting the expression of the Beta distribution into Eq. 9, the problem decouples across classes:

$$\operatorname{argmax}_{\alpha_c, \beta_c} (\alpha_c - 1) \sum_{v=1}^V w_{vc} \log f_v + (\beta_c - 1) \sum_{v=1}^V w_{vc} \log(1 - f_v) - \log B(\alpha_c, \beta_c) \sum_{v=1}^V w_{vc}. \quad (14)$$

This is a simple 2D optimization problem, which we solve with conjugate gradient. In the first iteration, we use the method of moments for initialization.

*Optimizing  $\{\psi_c\}$ :* This optimization can also be carried out one  $c$  at the time:

$$\operatorname{argmax}_{\psi_c: \|\psi_c\|=1} \sum_v w_{vc} f_v (\psi_c^t \phi_v)^2 = \operatorname{argmax}_{\psi_c: \|\psi_c\|=1} \psi_c^t \left[ \sum_v w_{vc} f_v \phi_v \phi_v^t \right] \psi_c,$$

with closed-form solution given by the leading eigenvector of:  $\sum_v w_{vc} f_v \phi_v \phi_v^t$ .

*Optimizing  $\{\kappa_c\}$ :* This optimization problem also decouples across classes:

$$\operatorname{argmax}_{\kappa_c} \kappa_c \sum_{v=1}^V w_{vc} f_v (\psi_c^t \phi_v)^2 - \sum_{v=1}^V w_{vc} \log Z(f_v \kappa_c), \quad (15)$$

which we solve with conjugate gradient, initializing  $\kappa_c = 10$  in the first iteration.

**Final Segmentation:** It is straightforward to show that the approximate posterior probability of the segmentation from Eq. 6 factorizes across voxels and is given by  $p(L|\hat{\theta}^a, \hat{\theta}^s, \hat{\theta}^d, S, D, A, \gamma^a, \gamma^s) = \prod_{v=1}^V \hat{w}_{v,l_j}$ , where  $\hat{w}_{v,l_j}$  is obtained by evaluating Eq. 10 at the optimal parameter values  $\hat{\theta}^a, \hat{\theta}^s, \hat{\theta}^d$ . Therefore, the optimal segmentation can be computed independently at each location  $v$  as:

$$\hat{l}_v = \operatorname{argmax}_c \hat{w}_{vc}, \quad (16)$$

and the expected value of the volume of class  $c$  is given by:  $\sum_{v=1}^V \hat{w}_{vc}$  (in voxels).

**Algorithm 1** Bayesian segmentation with sMRI and dMRI

---

**Require:**  $\mathbf{A}, \mathbf{S}, \mathbf{D}, \gamma^a, \gamma^s, \{\mathbf{M}_c, n_c\}$   
**Ensure:**  $\hat{\theta}^a, \{\hat{\mu}_c, \hat{\Sigma}_c, \hat{\alpha}_c, \hat{\beta}_c, \hat{\psi}_c, \hat{\kappa}_c\}$   
Initialize  $\theta^a$ , with affine registration and mutual information  
Initialize  $w_{vc} \leftarrow p(l_v = c | \mathbf{A}, \theta^a), \forall v, c$   
Initialize  $\kappa_c = 10, \forall c$   
Initialize  $\alpha_c, \beta_c$  with method of moments,  $\forall c$   
 $it \leftarrow 0$   
**while**  $\theta^a, \{\mu_c, \Sigma_c, \alpha_c, \beta_c, \psi_c, \kappa_c\}$  change AND  $it \leq it_{\max}$  **do**  
   $it \leftarrow it + 1$   
  **for**  $c = 1, \dots, C$  **do**  
    Update  $\mu_c, \Sigma_c$  with Eqns. 12 and 13  
    Update  $\alpha_c, \beta_c$  by numerically optimizing Eq. 14 with conjugate gradient  
    Update  $\psi_c \leftarrow$  leading eigenvector of:  $\sum_v w_{vc} f_v \phi_v \phi_v^t$   
    Update  $\kappa_c$  by numerically optimizing Eq. 15 with conjugate gradient  
  **end for**  
  **if**  $\text{mod}(it, 5) = 0$  **then**  
    Update  $\theta^a$  by numerically optimizing Eq. 11 with conjugate gradient  
  **end if**  
  Update  $w_{vc}$  with Eq. 10,  $\forall v, c$   
**end while**  
Compute final segmentation with Eq. 16.

---

**Implementation details:** Since GEM only requires *improving* the bound at each iteration, we follow a schedule in which all the model parameters except for  $\theta^a$  are updated once in the M step. Since updating  $\theta^a$  requires solving a more computationally expensive registration problem, we only update  $\theta^a$  in the M step every five GEM iterations. The method is summarized in Algorithm 1.

In practice, we also force some parameters  $\{\theta_c^s\}$  and  $\{\theta_c^d\}$  to be shared across classes, for increased robustness of the algorithm. For the sMRI parameters ( $\{\mu_c, \sigma_c^2\}$ ), we follow [17] and force parameter sharing across: cortex, hippocampus and amygdala; reticular nucleus and white matter; mediodorsal and pulvinar nuclei; rest of thalamic nuclei; and contralateral structures. For the FA, parameters  $\{\alpha_c, \beta_c\}$  are shared across each of the six groups of thalamic nuclei in Table 2 of [17], and across contralateral structures. The same grouping – but without contralateral constraints – is used for the directional parameters  $\{\psi_c, \kappa_c\}$ .

### 3 Experiments and results

#### 3.1 Data

We evaluate our method with a recent probabilistic atlas of 25 thalamic nuclei and surrounding regions derived from histology [17]. The thalamus is an excellent target region, due to its faint lateral boundaries in sMRI (as explained in Section 1), and its set of nuclei with different connectivity. We use two considerably different datasets in evaluation: HCP (state of the art) and ADNI (legacy).



*HCP*: Isotropic T1 and dMRI scans from 100 healthy subjects (age  $29.1 \pm 3.3$ , 44 males), at 0.7 mm (T1) and 1.25 mm resolution (dMRI). We fit the DTI model to the  $b=1000$  s/mm<sup>2</sup> shell (180 directions) and 12 scans with  $b=0$  (details in [22]).

*ADNI*: T1 and dMRI scans from 77 subjects from ADNI2: 39 Alzheimer’s disease (AD) and 38 age-matched controls ( $74.1 \pm 8.1$  years; 40 females total). T1 resolution:  $1.2 \times 1 \times 1$  mm (sagittal); dMRI resolution:  $1.35 \times 1.35 \times 2.7$  mm (axial); 5 volumes with  $b=0$ , 41 directions ( $b=1000$  s/mm<sup>2</sup>; details at [adni-info.org](http://adni-info.org)).

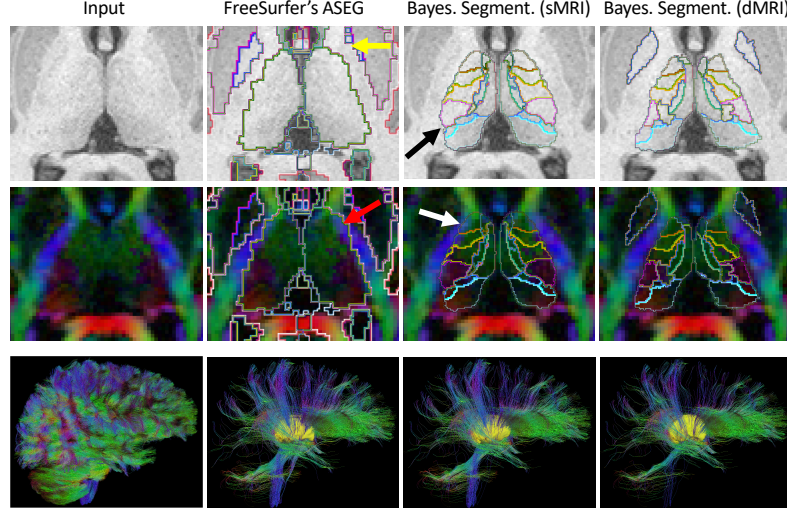
### 3.2 Experimental setup

We evaluate three competing methods: *(i)* Segmentation of the whole thalamus with FreeSurfer [2]; *(ii)* Segmentation of thalamic nuclei using Bayesian segmentation on T1 only [17]; and *(iii)* Segmentation of thalamic nuclei with the full model, including dMRI. We compare these approaches in three experiments: *(i)* Qualitative assessment of segmentation and tractography in HCP; *(ii)* Correlation between thalamic and total intracranial volume (ICV) in HCP; and *(iii)* Ability to discriminate AD and control subjects based on volumes in ADNI. The sMRI and dMRI data are resampled to 0.5 mm isotropic in a bounding box around the thalami (DTI is interpolated in a log-euclidean framework [23]). We set  $\lambda = 0.05$  as in [17],  $M_c$  to the median T1 intensity in class  $c$  according to the main FreeSurfer segmentation, and  $n_c$  to the volume of the class in mm<sup>3</sup>.

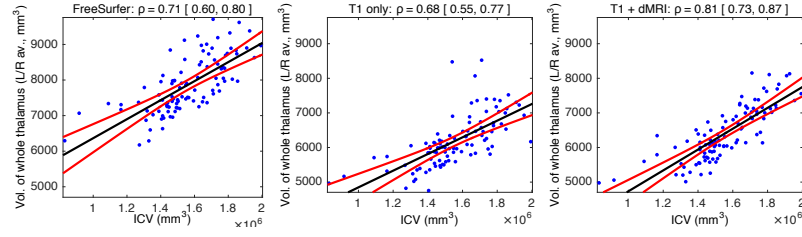
### 3.3 Results

Figure 3 shows qualitative results on an HCP subject. FreeSurfer almost completely misses the left pallidum (yellow arrow in the figure) and oversegments the thalami. We test the effects of the latter on tractography by reconstructing the full dMRI data with generalized q-sampling [24], performing whole-brain tractography, and isolating the tracts that intersect the whole thalami, as automatically segmented by the three competing methods. The FreeSurfer thalamus yields many false positive tracts, mostly due to overlap with the internal capsule (red arrow). Aggregating the nuclei produced by Bayesian segmentation on the T1 produces a more accurate boundary, but still oversegments the anterior thalamus (white arrow), and undersegments the pulvinar nucleus (black arrow). Our multi-modal method yields less false positive tracts, and segments thalamic nuclei that are more homogeneous in terms of diffusion orientation and FA.

We also evaluate segmentation performance quantitatively on HCP, in an indirect fashion, by computing the correlation of total thalamic volume obtained by each method (left-right averaged) with the ICV estimated by FreeSurfer; noisy thalamic segmentations are expected to degrade this correlation. Scatter plots and regression lines are shown in Fig. 4. The FreeSurfer volumes are quite large on average, and their correlation with ICV is  $\rho=0.71$ . Bayesian segmentation with T1 yields  $\rho=0.68$  (not significantly different, with  $p=0.37$  on a two-tailed Steiger test). The proposed algorithm produces fewer outliers than the other two, and yields the highest correlation ( $\rho=0.81$ ), significantly higher than those of FreeSurfer ( $p=0.006$ ) and T1-only segmentation ( $p=0.001$ ).



**Fig. 3.** Top two rows: axial slice of a T1 scan and principal diffusion directions of an HCP subject, with segmentations from FreeSurfer, Bayesian segmentation (T1 only), and the proposed method (T1+dMRI). Bottom row (left to right): Whole-brain tractography (25,000 tracts); subset of tracts going through the thalami (in yellow) as segmented by: FreeSurfer (2,602 tracts); Bayesian segmentation of T1 (2,193 tracts); and proposed method (1,676 tracts). See Section 3.3 for a description of the arrows.



**Fig. 4.** Scatter plot for intracranial vs. whole thalamic volumes (left-right averaged), and regression lines (black) with 95% confidence intervals (in red).

Finally, we evaluate the ability of the segmented volumes to classify the ADNI subjects into AD and controls. We use a simple linear discriminant analysis, whose performance is mostly determined by the quality of the volumes. We project the volumes (left-right averaged, corrected for ICV and age) onto the normal to the discriminant hyperplane in a leave-one-out fashion. We use the projections to compute the area under the ROC curve (AUC), accuracy at its elbow, and sample sizes ( $\alpha = 0.05, \beta = 0.2$ ). Results are shown in Table 1. Our method yields a fair improvement compared with T1-only Bayesian segmentation (increase of 7 points in accuracy and AUC, and reduction of 6 samples). Compared with FreeSurfer, our algorithm reduces the sample size by 60%.

**Table 1.** AUC, accuracy at elbow, and sample size for the AD classification experiment.

Method	FreeSurfer (whole)	Bayes. Seg. T1	Proposed
AUC	60.3%	66.5%	73.6%
Accuracy at elbow	61.0%	67.5%	74.0%
Sample size	50	26	20

## 4 Conclusion

We have presented a Bayesian method for joint segmentation of sMRI and dMRI, which is robust to changes in acquisition platform and protocol – as shown with two substantially different datasets. Compared with Bayesian segmentation using sMRI alone, our method produces more accurate boundaries for subcortical structures, and yields smaller sample sizes in an AD classification task.

Future work at the methodological level will follow five main directions: *(i)* Modeling partial voluming in the dMRI, which may be important for smaller structures; *(ii)* Exploring other axial PDFs, as well as mixtures; *(iii)* Placing a prior on the dMRI likelihood parameters, e.g., to utilize prior knowledge on the FA; *(iv)* Modeling the bias field in the sMRI data, e.g., as in [1]; and *(v)* Adding connectivity derived from tractography to the dMRI likelihood, which may be challenging because tractography results depend largely on the MR acquisition.

We also plan to manually trace structures some of the HCP and AD data, with three purposes. First, to include white matter bundles in the atlas, as modeling the whole cerebral white matter with a single Beta-DSW is not realistic (not even within a bounding box). Second, to enable direct evaluation of our segmentations. And third, to help us explain discrepancies in AD classification accuracy between our results and those presented in [17], which may be due to the their larger dataset, their different ADNI sample, or some other factor.

As high-resolution dMRI becomes more common in neuroimaging, we believe that segmentation techniques that jointly model gray and white matter with sMRI and dMRI – like the one in this paper – will become increasingly important.

**Acknowledgement:** Supported by Horizon 2020 (ERC Starting Grant 677697, Marie Curie grant 765148), Danish Council for Independent Research (DFF-6111-00291), NIH (R21AG050122, P41EB015902), Wistron Corp., SIP, and AWS.

## References

1. Ashburner, J., Friston, K.: Unified segmentation. *Neuroimage* **26** (2005) 839–851
2. Fischl, B., Salat, D.H., Busa, E., Albert, M., Dieterich, M., et al.: Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron* **33**(3) (2002) 341–355
3. Iglesias, J.E., Sabuncu, M.R.: Multi-atlas segmentation of biomedical images: a survey. *Med. Im. Anal.* **24**(1) (2015) 205–219
4. Roy, A.G., Conjeti, S., Navab, N., Wachinger, C.: Quicknat: Segmenting MRI neuroanatomy in 20 seconds. *NeuroImage* **186**(1) (2019) 713–727

5. Patenaude, B., Smith, S., Kennedy, D., Jenkinson, M.: A bayesian model of shape and appearance for subcortical brain segmentation. *NeuroImage* **56** (2011) 907–922
6. Van Essen, D.C., Smith, S.M., Barch, D.M., Behrens, T.E., Yacoub, E., et al.: The WU-Minn human connectome project: an overview. *NeuroImage* **80** (2013) 62–79
7. O'Donnell, L.J., Westin, C.F.: Automatic tractography segmentation using a high-dimensional white matter atlas. *IEEE Trans. Med. Im.* **26**(11) (2007) 1562–1575
8. Wassermann, D., Bloy, L., Kanterakis, E., Verma, R., Deriche, R.: Unsupervised white matter fiber clustering and tract probability map generation: applications of a Gaussian process framework for white matter fibers. *NeuroImage* **51** (2010) 228
9. Yendiki, A., Panneck, P., Srinivasan, P., Stevens, A., Zöllei, L., et al.: Automated probabilistic reconstruction of white-matter pathways in health and disease using an atlas of the underlying anatomy. *Front. Neuroinform.* **5** (2011) 23
10. Awate, S.P., Zhang, H., Gee, J.C.: A fuzzy, nonparametric segmentation framework for DTI and MRI analysis: with applications to DTI-tract extraction. *IEEE Trans. Med. Im.* **26**(11) (2007) 1525–1536
11. Hagler, D.J., Ahmadi, M.E., Kuperman, J., Holland, D., McDonald, C.R., et al.: Automated white-matter tractography using a probabilistic diffusion tensor atlas: Application to temporal lobe epilepsy. *Hum. Brain Map.* **30**(5) (2009) 1535–1547
12. Behrens, T.E., Johansen-Berg, H., Woolrich, M., Smith, S., Wheeler-Kingshott, C., et al.: Non-invasive mapping of connections between human thalamus and cortex using diffusion imaging. *Nat. Neurosci.* **6**(7) (2003) 750
13. Saygin, Z.M., Osher, D.E., Augustinack, J., Fischl, B., Gabrieli, J.D.: Connectivity-based segmentation of human amygdala nuclei using probabilistic tractography. *Neuroimage* **56**(3) (2011) 1353–1361
14. Stough, J.V., Glaister, J., Ye, C., Ying, S.H., Prince, J.L., Carass, A.: Automatic method for thalamus parcellation using multi-modal feature classification. *Proc. of MICCAI (Pt 3)* (2014) 169–176
15. Glaister, J., Carass, A., Stough, J.V., Calabresi, P.A., Prince, J.L.: Thalamus parcellation using multi-modal feature classification and thalamic nuclei priors. *Proc SPIE Int Soc Opt Eng* **9784** (2016)
16. Zhang, Y., Brady, M., Smith, S.: Segmentation of brain MR images through a hidden markov random field model and the expectation-maximization algorithm. *IEEE Trans.Med.Im.* **20**(1) (2001) 45–57
17. Iglesias, J.E., Insausti, R., Lerma-Usabiaga, G., Bocchetta, M., Van Leemput, K., et al.: A probabilistic atlas of the human thalamic nuclei combining ex vivo MRI and histology. *NeuroImage* **183** (2018) 314–326
18. Van Leemput, K.: Encoding probabilistic brain atlases using bayesian inference. *IEEE Trans. Med. Im.* **28**(6) (2009) 822
19. Zhang, H., Schneider, T., Wheeler-Kingshott, C.A., Alexander, D.C.: NODDI: practical in vivo neurite orientation dispersion and density imaging of the human brain. *Neuroimage* **61**(4) (2012) 1000–1016
20. Mardia, K., Jupp, P.: Directional statistics. Volume 494. John Wiley & Sons (2009)
21. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. B* (1977) 1–38
22. Sotiropoulos, S.N., Jbabdi, S., Xu, J., Andersson, J.L., Moeller, S., et al.: Advances in diffusion MRI acquisition and processing in the human connectome project. *Neuroimage* **80** (2013) 125–143
23. Arsigny, V., Fillard, P., Pennec, X., Ayache, N.: Log-euclidean metrics for fast and simple calculus on diffusion tensors. *Magn. Reson. Med.* **56**(2) (2006) 411–421
24. Yeh, F.C., Wedeen, V.J., Tseng, W.Y.I.: Generalized q-sampling imaging. *IEEE Trans. Med. Im.* **29**(9) (2010) 1626–1635