Probabilistic Atlas and Geometric Variability Estimation to Drive Tissue Segmentation

Hao Xu, Bertrand Thirion, Stéphanie Allassonnière

Research Article

Statistics
in Medicine

Received XXXX

(www.interscience.wiley.com) DOI: 10.1002/sim.0000

Probabilistic Atlas and Geometric Variability Estimation to Drive Tissue Segmentation

Hao Xu,[a*†] Bertrand Thirion[b] and Stéphanie Allassonnière[a]

Computerized anatomical atlases play an important role in medical image analysis. While an atlas usually refers to a standard or mean image also called template, that presumably represents well a given population, it is not enough to charaterize the observed population in detail. A template image should be learned jointly with the geometric variability of the shapes represented in the observations. These two quantities will in the sequel form the atlas of a population. The geometric variability is modelled as deformations of the template image so that it fits the observations. In this paper, we provide a detailed analysis of a new generative statistical model based on dense deformable templates that represents several tissue types observed in medical images. Our atlas contains both an estimation of probability maps of each tissue (called class) and the deformation metric. We use a stochastic algorithm for the estimation of the probabilistic atlas given a dataset. This atlas is then used for atlas-based segmentation method to segment the new images. Experiments are shown on brain T1 MRI datasets.

Keywords: Probabilistic atlas, geometric variability, neuro-segmentation coupled with registration, atlas-based segmentation, stochastic algorithm, statistical estimation

1. Introduction

In neuroimaging, brain atlases are useful for both segmentation and registration tasks as they enable to transport known information to a new patient image to perform qualitative and quantitative comparisons. What is often referred to as an atlas actually corresponds to a mean image or *template*. This problem of estimating such an image given a population has started to be a central issue in medical imaging for the past decade. Many different methods have been proposed for the template estimation (see [1, 2] among others); they work either on gray level images, segmented data or shapes summarized by a set of landmarks. Probabilistic templates are especially attractive

[a] *CMAP Ecole Polytechnique, Route de Saclay, 91128 Palaiseau, France*
[b] *Parietal Team, INRIA Saclay-Île-de-France*
[b] *CEA, DSV, I²BM, Neurospin bât 145, 91191 Gif-Sur-Yvette, France*
[*] *Correspondence to: Hao Xu, CMAP Ecole Polytechnique, Route de Saclay, 91128 Palaiseau, France*
[†] E-mail: xuhao@cmap.polytechnique.fr

Prepared using simauth.cls *[Version: 2010/03/10 v3.00]*

[3, 4], as they make it possible to take into account the uncertainty on the underlying tissue type, which is related to PVE or to perfectible registration. In many template construction methods, pre-segmentation or pre-registration are required. In this paper, we aim at creating a probabilistic atlas, which we define as a probabilistic template together with a quantification of the population geometric variability. However this estimation does not require any pre-segmentation and pre-registration.

Atlas learning encompasses the two most fundamental problems in image analysis, namely segmentation and registration, as these are the basis of template estimation and population analysis. Concerning the segmentation issue, it is important to use automated segmentation for the sake of efficiency and reproducibility. Many different methods have already been proposed for segmentation, such as level set methods [5], model-based segmentation [6], template-based approaches [7, 8] among many others. In many cases, segmentation is coupled with registration. Indeed, performing registration and segmentation jointly is generally more effective than performing them sequentially [9, 10, 11]. The result of an accurate segmentation enable to increase the precision of a registration. On the other hand, transporting a segmentation from a template to a subject requires a accurate registration procedure. The accuracy of the registration depends on the class of deformations that are considered. Indeed, one may prefer smooth deformations that capture only the global shape changes rather than local details. On the opposite side, when the roughness of the shape is meaningful, one has to adapt the class of deformations to enable them to distinguish different border regularity. Part of this choice has to be made by the user depending on the data. However, the complexity of the deformation set has also to be constrained by the observations themselves. Some deformation models provide a metric on the space of shape through a metric on the deformation set [12] which describes geometrically the data. Another viewpoint is to propose a probabilistic approach where the probability distribution of the deformation will highlight the characteristic deformation in a population of interest. Both approaches can actually be related, in particular when considering that the deformations are multivariate normally distributed, as the covariance matrix relates to a natural metric to compute the distances between deformations [13].

As pointed in [14], estimating this probability distribution together with the template (gray level in [14]) increases the population classification accuracy, as the model better fits the observations. As for the population classification in [14], the segmentation (tissue classification of voxels) takes advantage of the registration to the template. In the sequel, the probabilistic template together with the geometric variability will be called *atlas*.

Several solutions have been proposed previously to deal with one or the other part of atlas or template estimation; we now discuss the closest works to ours. First, a problem with average templates construction is that they do not include the nonlinear deformation to align the corresponding structures. In [15], to solve this problem, a generative model was proposed to create a template using mesh-based representations endowed with a deformation model. This method computes estimates of the deformation field and the most compact mesh representation using an Expectation-Maximization (EM) algorithm. However they require the pre-segmentation of the training image. In [16], a method was proposed to do the segmentation and registration jointly, while creating an average brain template. This approach combines groupwise registration using the Kullback-Leibler divergence and the EM algorithm for segmentation, and thus demonstrates the benefit of their integration. However it does not learn the geometric variability within the estimation procedure which may reduce the accuracy of the template to match the observations with prior deformations. In [17], a probabilistic model was proposed to segment a heterogeneous data set of brain MR images simultaneously while constructing templates for each mode in the heterogeneous population using an EM algorithm. However, it performs clustering as an additional step, and does not learn the geometric variability of the population. In [14], a model was proposed to create an atlas containing the geometric variability. As the inputs are scalar images, the template is also estimated as a scalar (gray level) image. As a

consequence, the segmentation of the population is not part of the estimation process. In [18], a spherical demons algorithm with geometric variability was proposed for registering images and for creating an atlas. The registration was more accurate and this registration could be used to transfer segmentation labels onto a new image. However, the segmentation was not performed during the estimation.

In this paper, we propose to include all the aspects of the atlas estimation procedures described previously, which can improve both the estimated template image, the estimated geometric variability and the segmentation of individual data. Moreover, we propose to perform this estimation using a joint segmentation-registration. For this purpose, we propose to model the observations (gray level images) by a generative statistical model, the parameters of this model being our atlas, i.e. a probabilistic template and the geometric variability. We generalize the model proposed in [19] and use the algorithm in [14] for the estimation. We also learn the geometry as the metric on the space of deformations, which reduces the possible deformations to those that are common in the population. This takes the form of a multivariate zero mean normal distribution on the deformations, where the main parameter is the covariance matrix, which is not constrained to have a particular structure (e.g. diagonal or sparse). This captures the long distance correlations of the deformations.

To estimate the model parameters, we use a stochastic algorithm that has demonstrated good performances on real data in [20, 21] and has theoretical convergence properties [14]. We get as final output an estimation of both the probabilistic template and the geometric variability. Although the individual deformation and segmentation are not parameters of the model, the algorithm can be used to return individual deformations and segmentations of the individual images. Additional parameters are also learned by this procedure, such as the means and variances of each tissue of gray level distribution.

After the estimation is performed, we use this atlas as an anatomical prior to segment new individuals. We define the gray level template using our estimated probabilistic template and the estimated weight. Then, it is registered to the target image, the deformation being constrained by the learned geometric variability. Last, this deformation is used to define the tissue class.

As a quantitative evaluation of our method, we test our algorithm on synthetic data for which we know the ground truth. We obtain high Jaccard indices on training and test data. We perform two tests to evaluate our method on real data. At first, we tested our algorithm on 8 patients on an anatomical brain MRI dataset for which a manual segmentation is available as a quantitative segmentation evaluation. Secondly, we create our atlas on 20 patients, then we generalize to five new patients to evaluate the segmentation in new patients.

The rest of this paper is organized as follows. In Sections 2, 3, 4 and 5 we present the model, the estimation, the algorithm and segmentation method in detail. Section 6 yields the experimental results on simulated and real data. In appendix, we prove the convergence of the estimation algorithm.

2. The Observation Model

In this section, we present our statistical model, the selected set of deformations and the parametric template that we consider for the sake of computational tractability. Then, we introduce the parameters of interest and the Bayesian framework, i.e. we introduce priors to address the known issue that medical images datasets most often comprise with very few samples.

2.1. Statistical Model

We consider here n individual MR images from n patients. This set $(y_i)_{1\leqslant i\leqslant n}$ of images are observed on a grid of voxels Λ embedded in a continuous domain $D \subset \mathbb{R}^3$. We denote $x_j \in D$ the location of voxel $j \in \Lambda$. We consider that each image is composed of voxels belonging to one class among K, corresponding to K tissues types. We assume that the signal in the K tissue classes is normally distributed with class dependent means $(\mu_k)_{1\leqslant k\leqslant K}$ and variances $(\sigma_k^2)_{1\leqslant k\leqslant K}$ as proposed in [19]. Therefore the probability of observing a data with intensity y_i^j for the ith image in the jth voxel given that it belongs to the kth class ($c_i^j = k$) is defined as follows:

$$\mathbb{P}(y_i^j | c_i^j = k, \mu_k, \sigma_k^2) \sim \mathcal{N}(y_i^j; \mu_k, \sigma_k^2), \tag{2.1}$$

where $\mathcal{N}(.; \mu, \sigma^2)$ is the normal density with mean μ and variance σ^2. This expression results from the assumption that given the class, the voxels are assumed to have independent gray level. This assumption is not satisfied in real life experiment as the noise of the observation depends on the tissue type. However, this common assumption is a first approximation that is sueful for the sake of estimation.

In order to take into account the geometric variability in shape of the brain along a population, we consider that there exists a random deformation from the template to the subject that acts as follows: the *unobserved* classes of the voxels of the data y are assumed to follow the probability distribution given by the discretization on Λ of the warped probabilistic template. This template is defined by the probability maps $(P_k)_{1\leqslant k\leqslant K}$ that yield the probability of each voxel to belong to each class in the template domain. In other words, the probability maps are deformed to match the observation y (in a sense that will be detailed below) ; then they are discretized on Λ to provide, at each voxel, a voxel-dependent discrete probability measure for this point which gives the probability of each voxel to belong to each class.

As the deformation is not observed (and is actually a mathematical tool for population analysis), we assume that these deformations from the template maps to each subject are also *unobserved and random*. We define them through a random field $z : \mathbb{R}^3 \to \mathbb{R}^3$ such that for $j \in \Lambda$ the prior probability of a voxel j from subject i to be in the k^{th} class is given by:

$$\mathbb{P}(c_i^j = k) = P_k(x_j - z(x_j))\,. \tag{2.2}$$

We consider here the linearized deformation model which defines a deformation φ of the domain D as the displacement of each point x in D by a vector $v(x)$, and is thus written as:

$$\varphi(x) = x + v(x).$$

As we consider linearized deformations, we approximate φ^{-1} by $\varphi^{-1}(x) = x - v(x)$ at the first order. This makes it possible to apply the deformation to an image, here P_k, as

$$\mathbb{P}(c_i^j = k) = \varphi_i \cdot P_k(x_j) = P_k(x_j - z_i(x_j)).$$

As defined above, the deformation is an infinite dimensional object. While such a dense representation is theoretically sound, for sake of computation, we consider a subspace of deformations that will be parameterized. We assume that the deformation is controlled by the displacement of some given control points belonging to D. This reduces the problem to finite dimension. We define the deformation field as a finite linear combinations of a given kernel

K_g centered at some fixed equi-distributed control points in the domain D: $(x_g)_{1\leqslant g\leqslant k_g}$ with parameter $\boldsymbol{\beta} \in (\mathbb{R}^3)^{k_g}$

$$\forall x \in D, z_\beta(x) = (\mathbf{K_g}\boldsymbol{\beta})(x) = \sum_{k=1}^{k_g} K_g(x, x_g)\boldsymbol{\beta}(k)\,, \tag{2.3}$$

where K_g is chosen as a radial Gaussian Kernel in our experiments.

As for the deformation model, the templates $P_k : \mathbb{R}^3 \to [0,1], \forall k \in [\![1, K]\!]$, which are the tissue probability maps, should be defined on the whole domain D. However, in order to reduce their dimensions to allow for numerical computation, we pick a fixed set of control points $(p_l)_{1\leqslant l\leqslant k_p}$ which may be different from the geometric ones and parametrize the templates by the coefficients $\alpha_k \in [0,1]^{k_p}$, which satisfy $\forall l \in [\![1, k_p]\!]$, $\sum_{k=1}^{K} \alpha_k^l = 1$. Then, we write

$$\forall x \in D, P_k(x) = \mathbf{K_p}\alpha_k(x) = \sum_{l=1}^{k_p} K_p(x, p_l)\alpha_k^l, \tag{2.4}$$

where $K_p(x, p_l) = 1$ if p_l is the nearest neighbor of x among the set of points $(p_j)_j$ and 0 otherwise.

Remark 1 *The unobserved parameter $\boldsymbol{\beta}$ appears in the indicator function of the kernel. This makes it impossible to compute the gradient of classical matching energies with respect to $\boldsymbol{\beta}$ and thus precludes any algorithm based on alternative gradient descents (as e.g. in [18, 22]). However, this indicator functions make it possible to deal with the constraint on α which would appear much harder with other smoother kernels and can be handled easily with our estimation algorithm.*

The previous hypothesis provides a generative statistical model for a sample of gray level images. The random variables are the deformation vector $\boldsymbol{\beta}$, the class of each voxel c and the parameters that characterize the gray levels of the tissues $(\mu_k, \sigma_k^2)_k$. The probability distributions of the former two elements are given by Equation (2.1) and (2.2). We assume that the deformation vector follows a normal distribution with mean zero and full covariance matrix. The hierarchical model is given by: $\forall i \in [\![1, n]\!],\ \forall j \in \Lambda$

$$\begin{cases} \boldsymbol{\beta}_i \sim \mathcal{N}(0, \Gamma_g)|\Gamma_g; \\ c_i^j \sim \sum_{k=1}^{K} \delta_k P_k(x_j - z_{\boldsymbol{\beta}_i}(x_j))|\boldsymbol{\beta}_i; \\ y_i^j \sim \mathcal{N}(\mu_k, \sigma_k^2)|c_i^j = k, \mu_k, \sigma_k^2, \end{cases} \tag{2.5}$$

where δ_k is a Dirac measure on k. The covariance matrix Γ_g is not assumed to have any particular pattern of zeros. This makes it possible to model local and global correlations between control point moves, in particular, very correlated displacements can be captured such as translation of a large area of the images. The zero mean is a relevant assumption as the population is assumed to be distributed in an ellipsoid around this mean image.

2.2. Parameters and likelihood

Given this statistical model, the parameters to estimate are the covariance matrix Γ_g of the deformation coefficient (Equation (2.3)), $(\alpha_k)_{1\leqslant k\leqslant K}$ the coefficients that define the templates (Equation (2.4)), $(\mu_k)_{1\leqslant k\leqslant K}$ and $(\sigma_k^2)_{1\leqslant k\leqslant K}$ the class dependent means and variances. Let $\theta_g = \Gamma_g$, $\theta_p = ((\alpha_k)_{1\leqslant k\leqslant K})$ and $\theta_c = ((\mu_k)_{1\leqslant k\leqslant K}, (\sigma_k^2)_{1\leqslant k\leqslant K})$. We assume that $\theta = (\theta_g, \theta_p, \theta_c)$ belongs to the parameter space Θ defined as the open set

$$\Theta = \{\theta = ((\alpha_k)_{1\leqslant k\leqslant K}, (\mu_k)_{1\leqslant k\leqslant K}, (\sigma_k^2)_{1\leqslant k\leqslant K}, \Gamma_g)|\alpha_k \in]0,1[^{k_p}, \sigma_k^2 > 0, \mu_k \in \mathbb{R}, \Gamma_g \in \Sigma^+_{3k_g,*}(\mathbb{R})\} \tag{2.6}$$

Here $\Sigma^+_{3k_g,*}(\mathbb{R})$ is the set of strictly positive symmetric matrices of dimension $3k_g \times 3k_g$.

We can notice that due to the unobserved variables β and c, the observed likelihood is an integral over these random variables. This writes

$$q(y|\theta) = \int\int q(y|c,\theta_c)q(c|\boldsymbol{\beta},\theta_p)q(\boldsymbol{\beta}|\theta_g)dcd\boldsymbol{\beta} \tag{2.7}$$

where the conditional distributions are given by our model

$$q(\boldsymbol{\beta}|\theta_g) = \exp(-\frac{1}{2}\boldsymbol{\beta}^T\Gamma_g^{-1}\boldsymbol{\beta})(2\pi)^{-\frac{3}{2}kg}|\Gamma_g|^{-\frac{1}{2}} \tag{2.8}$$

$$q(c|\boldsymbol{\beta},\theta_p) = \sum_{k=1}^{K}\delta_k P_k(x_j - z_{\boldsymbol{\beta}_i}(x_j)) \tag{2.9}$$

$$q(y|c,\theta_c) = \prod_{j=1}^{|\Lambda|}(2\pi\sigma^2_{c_j})^{-1/2}\exp\left(-\frac{(y^j-\mu_{c_j})^2}{2\sigma^2_{c_j}}\right) \tag{2.10}$$

where $|\Lambda|$ is the number of voxels.
For sake of simplicity, all the likelihood functions will be denoted by q and the variables specified as arguments of this function q.

2.3. Bayesian Model

Medical images are typically high-dimensional, but usually come in small samples. To take this possibility into account, we choose to regularize the statistical model and we propose to work in a Bayesian framework. As presented in [13], we use standard conjugate priors for each parameter, i.e. an inverse-Wishart ν_g in dimension $3kg \times 3kg$ on Γ_g, a Gaussian ν_m on μ_k and inverse-Wishart ν_p in dimension 1 on σ^2_k with fixed hyper-parameters. All priors are assumed independent. These priors makes it possible to regularize when needed the estimated parameters but, when the number of observations increases, the relative prior weight decreases.

More formally we have

$$(\Gamma_g, \mu_k, \sigma^2_k) \sim \nu_g \otimes \nu_m \otimes \nu_p;$$

where

$$\begin{cases} v_g(\Gamma_g) \propto \left(\exp(-\frac{1}{2}\langle\Gamma_g^{-1},\Gamma_g^0\rangle)\frac{1}{\sqrt{|\Gamma_g|}}\right)^{a_g} d\Gamma_g, a_g \geqslant 6k_g+1, \\ v_m(\mu_k) \propto \exp\left(-\frac{(\mu_k-m_u)^2}{2\sigma^2_\mu}\right) d\mu_k \\ v_p(\sigma^2_k) \propto \left(\exp\left(-\frac{\sigma^2_0}{2\sigma^2_k}\right)\frac{1}{\sqrt{\sigma^2_k}}\right)^{a_p} d\sigma^2_k, a_p \geqslant 3. \end{cases}$$

Note that for two matrices A, B we have $\langle A, B\rangle = tr(A^TB)$ the Frobenius inner product on matrices.
Our Bayesian model can be represented by Fig. 1 where the dependencies are highlighted.

3. Estimation

3.1. Existence of the MAP estimation

Given the complete statistical model, we can learn the parameters that best fit the observations. Although real data never follow any parametric model, we try to approximate their generation so that we better understand the common and specific features of a given population. For this purpose, we consider the maximum a posteriori (MAP)

estimator: $\hat{\theta}_n = \arg\max_{\theta\in\Theta} q_B(\theta|y_1,...,y_n)$ where q_B denotes the posterior distribution of the parameters given the n observations $y_1,...,y_n$.

The following theorem proves here that given a n sample of observations, the maximum a posteriori estimator exists at finite distance in the parameter space.

Theorem 1 (Existence of the MAP estimation) *For any sample $y_1,...,y_n$, there exists $\hat{\theta}_n \in \Theta$ such that*

$$q_B(\hat{\theta}_n|y_1,...,y_n) = \sup_{\theta\in\Theta} q_B(\theta|y_1,...,y_n).$$

Remark 2 *Note that one could rely on the prior distribution to prove this property for the means $(\mu_k)_{1\leq k\leq K}$. However, as we are dealing with a Bayesian model, we introduce priors on all parameter to keep the coherence of the model. Nonetheless, it would be possible to remove the prior on these parameters thanks to the proof above. Concerning the priors, for the covariance matrix, the prior is informative as we choose the usual kernel matrix used for registration issues. The prior on the means are non-informative as the gray level of the observations change drastically when the acquisition protocols change.*

3.2. Consistency of the estimator on our model

We are interested now in the consistency property of the MAP estimator without making strong assumptions on the distribution of the observations $y_1,...,y_n$. This means that we do not assume that the observations are generated by the model described above. We denote the distribution governing the observations by π and seek to prove the convergence of the MAP estimator to the set Θ_* of model distributions *closest* to π:

$$\Theta_* = \{\theta_* \in \Theta | E_\pi(\log q(y|\theta_*)) = \sup_{\theta\in\Theta} E_\pi(\log q(y|\theta))\}$$

However, this consistency only holds for bounded variances $\forall\ c \in [\![1, K]\!]$, $\sigma_c^2 > \sigma_{min}^2$. This assumption on the admissible set is not restrictive as we have proven that the MAP estimator exists out of the boundaries. Let

$$\Theta^B = \{\theta = ((\alpha_k)_{1\leqslant k\leqslant K}, (\mu_k)_{1\leqslant k\leqslant K}, (\sigma_k^2)_{1\leqslant k\leqslant K}, \Gamma_g) | \alpha_k \in]0,1[^{k_p}, \sigma_k^2 > \sigma_{min}^2, \mu_k \in \mathbb{R}, \Gamma_g \in \Sigma^+_{3k_g,*}(\mathbb{R})\} \tag{3.1}$$

$$\Theta^B_* = \{\theta_* \in \Theta^B | E_\pi(\log q(y|\theta_*)) = \sup_{\theta\in\Theta^B} E_\pi(\log q(y|\theta))\}$$

Theorem 2 (Consistency) *Assume that Θ^B_* is non empty. Then for any compact set $K \subset \Theta^B$,*

$$\lim_{n\to+\infty} \pi(\delta(\hat{\theta}_n, \Theta^B_*) \geqslant \epsilon \wedge \hat{\theta}_n \in K) = 0,$$

where δ is the metric inherited from the Euclidean metric on $\mathbb{R}^{n_t}$ where n_t is the dimension of Θ.

Proof The theorem is an application of Wald's consistency Theorem in [23]. We only need to verify that $y \to \log q(y|\theta)$ is π *a.s.* upper semi-continuous and that for any $\theta \in \Theta$, there exists an open set $U \ni \theta$ such that $E_\pi(\sup_{\theta'\in U} \log^+(q(y|\theta'))) < \infty$ (where $\log^+$ is the positive part of log). In our setting, for any $\theta = (\alpha_k, \mu_k, \sigma_k^2, \Gamma_g) \in \Theta^B$,

we denote $U = \{]0,1[^{k_p}, \mathbb{R},]\sigma^2_{min}, +\infty[, \Sigma^+_{3k_g,*}(\mathbb{R})\}$, so that

$$\begin{aligned}\sup_{\theta' \in U} \log(q(y|\theta')) &\leqslant \sup_{\theta' \in U} \log(\sum_{k=1}^{K} q(y|c,\theta')) \\ &\leqslant \sup_{\theta' \in U} \log\left(\sum_{k=1}^{K}\left(\frac{1}{\sqrt{2\pi\sigma_k^2}}\right)^{n|\Lambda|}\right) \\ &\leqslant \log\left(K\left(\frac{1}{\sqrt{2\pi\sigma^2_{min}}}\right)^{n|\Lambda|}\right) < \infty \quad ,\end{aligned}$$

where σ^2_{min} is the lower bound of σ^2_k.

Remark 3 *Note that we only proved a limited consistency result as we have no guaranty that Θ^B_* is not empty. However, looking at our model, generalized from the Bayesian Mixed Effect Templates introduced in [13], we expect the same result. Future work will generalize the proof of consistency in [13] to the present model. We focus here on the convergence property of the estimation algorithm that we present in the following section.*

4. Estimation Algorithm using Stochastic Approximation Expectation-Maximization

We now present the estimation algorithm that we use to reach the MAP estimate of the parameters. We assume now that we observe a fixed number n of gray level images taken from a homogeneous population.

4.1. Model factorization

Despite the complex dependencies of the random variables in our statistical model, it belongs to the curved exponential family. That is to say, the complete likelihood q writes as:

$$q(y, c, \boldsymbol{\beta}, \theta) = \exp[-\psi(\theta) + \langle S(c, \boldsymbol{\beta}), \phi(\theta)\rangle] = \exp(L(\theta, S)) \tag{4.1}$$

where ψ, ϕ are two Borel functions depending on the parameters, $S(c, \boldsymbol{\beta})$ is a vector of sufficient statistics and the scalar product is the usual Euclidean one. For sake of simplicity, we have omitted the dependency with respect to the observations that are handled as a fixed input to the estimation process. The function L is called the complete log-likelihood.

Thanks to Equation (4.1), we can show that the following matrix-valued functions are the sufficient statistics of the model: $\forall k \in [\![1, K]\!]$, $\forall l \in [\![1, k_p]\!]$,

$$\begin{aligned} S_{0,k}(c,\boldsymbol{\beta}) &= \sum_{i=1}^{n}\sum_{j=1}^{|\Lambda|} \mathbb{1}_{c_i^j = k}, & S_{1,k}(c,\boldsymbol{\beta}) &= \sum_{i=1}^{n}\sum_{j=1}^{|\Lambda|} \mathbb{1}_{c_i^j = k} y_i^j, \\ S_{2,k}(c,\boldsymbol{\beta}) &= \sum_{i=1}^{n}\sum_{j=1}^{|\Lambda|} \mathbb{1}_{c_i^j = k} (y_i^j)^2, & S_3(c,\boldsymbol{\beta}) &= \sum_{i=1}^{n} \boldsymbol{\beta}_i \boldsymbol{\beta}_i^T, \\ S_{4,k,l}(c,\boldsymbol{\beta}) &= \sum_{i=1}^{n}\sum_{j=1}^{|\Lambda|} \mathbb{1}_{c_i^j = k} \mathbb{1}_{\|x_j - x_l - K_g \boldsymbol{\beta}_i(x_j)\|_1 < \frac{1}{2}}. \end{aligned} \tag{4.2}$$

We denote $S(c,\boldsymbol{\beta}) = (S_{0,k}(c,\boldsymbol{\beta}), S_{1,k}(c,\boldsymbol{\beta}), S_{2,k}(c,\boldsymbol{\beta}), S_3(c,\boldsymbol{\beta}), S_{4,k,l}(c,\boldsymbol{\beta}))_{\forall k\in[\![1,K]\!],\ \forall l\in[\![1,k_p]\!]}$ the vector of sufficient statistics and define the sufficient statistic space as

$$\mathcal{S} = \{(S_{0,k}, S_{1,k}, S_{2,k}, S_3, S_{4,k,l}) \mid \forall k \in [\![1,K]\!],\ \forall l \in [\![1,k_p]\!], \\ S_{0,k} \in \mathbb{R}_*^+, S_{1,k} \in \mathbb{R}_*^+, S_{2,k} \in \mathbb{R}_*^+, S_3 + a_g\Gamma_g^0 \in \mathrm{Sym}_{3kg}^+, S_{4,k,l} \in R_*^+)\}. \quad (4.3)$$

The set $\mathcal{S}$ can be viewed as an open set of $\mathbb{R}^s$ with $s = K + K + K + \frac{3kg(3kg+1)}{2} + Kk_p$.

Remark 4 *Note that the sufficient statistics $S_{0,k}$, $S_{1,k}$, $S_{2,k}$ and $S_{4,k}$ cannot vanish. Indeed, if for one class k_0, $S_{0,k_0} = 0$, this particular class would be empty which means that there are no voxel belonging to this tissue class. We assume that we actually know the number of expected tissues in the gray level images so that this never happens.*

The second property of our model is that there exists $\hat{\theta}$ such as $\max_{\theta\in\Theta} L(\theta, S) = \hat{\theta}(S)$. Indeed μ_k, σ_k^2, α_l^k and Γ_g are explicitly expressed with the above sufficient statistics as follows: $\forall k \in [\![1,K]\!]$, $\forall l \in [\![1,k_p]\!]$,

$$\begin{aligned} \hat{\mu}_k(S) &= \frac{S_{1,k}(c,\boldsymbol{\beta})}{S_{0,k}(c,\boldsymbol{\beta})}, & \hat{\sigma}_k^2(S) &= \frac{1}{n+a_p}\left(n\left(\frac{S_{2,k}(c,\boldsymbol{\beta})}{S_{0,k}(c,\boldsymbol{\beta})} - \frac{S_{1,k}(c,\boldsymbol{\beta})^2}{S_{0,k}(c,\boldsymbol{\beta})^2}\right) + a_p\sigma_0^2\right), \\ \hat{\Gamma}_g(S) &= \frac{S_3(c,\boldsymbol{\beta}) + a_g\Gamma_g^0}{n * |\Lambda| + a_g}, & \hat{\alpha}_k^l(S) &= \frac{S_{4,k,l}(c,\boldsymbol{\beta})}{\sum_{k'=1}^{K} S_{4,k',l}(c,\boldsymbol{\beta})}. \end{aligned} \quad (4.4)$$

These equations are well defined thanks to Remark 4. This also justifies the fact that the coefficients α_k^l belongs to $]0,1[$ for all $1 \leqslant k \leqslant K$ and for all $1 \leqslant l \leqslant k_p$.

4.2. Estimation Algorithm

As we are in an incomplete-data setting, a natural way to maximize a likelihood is to use the Expectation-Maximization (EM) algorithm or an algorithm derived from EM. We choose the Stochastic Approximation EM (SAEM) coupled with a Markov Chain Monte Carlo method thanks to its good theoretical [14] and numerical [20] performances in such settings.

We detail here the $m+1^{th}$ iteration of the SAEM-MCMC algorithm which consists of three steps:

Step 1: Simulation step. The missing data, i.e. the deformation parameters $(\boldsymbol{\beta}) = (\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_n)$ and the vector of classes $(c) = (c_1, \cdots, c_n)$, are drawn using the transition probability of a ergodic Markov chain Π_θ having the posterior distribution $q_{post}(\cdot|y,\theta)$ as its stationary distribution:

$$((\boldsymbol{\beta})_{m+1}, (c)_{m+1}) \sim \Pi_{\theta_m}(((\boldsymbol{\beta})_m, (c)_m), \cdot)$$

where we choose Π_θ to be a Metropolis-Hastings within Gibbs sampler. This particular MCMC method is well adapted for high dimensional simulation and also in our particular case where the distribution of the class depends on the deformation. The Gibbs sampler works coordinate by coordinate. Since we cannot sample from the posterior distribution of one coordinate of the vector $((\boldsymbol{\beta}), (c))$ given the others, we use a Metropolis-Hastings step inside these loops. Therefore, we simulate the coordinates one by one. We choose as the proposal of the Metropolis-Hastings method to use the probability distribution of this coordinate given the others coming from the model distributions 2.5. This way, one can estimate deformations that improve the segmentation and segmentations that improve the registration. With this choice, it is easy to calculate the acceptance rates (see in Algorithm 1). Note that it would be possible to choose others priors, however paying attention to the computational cost of the acceptance rates.

Since we have a couple of missing data, we first simulate each coordinate of $(\boldsymbol{\beta})$ knowing others coordinates of $(\boldsymbol{\beta})$ and (c), then simulate each coordinate of (c) knowing others coordinates of (c) and the new $(\boldsymbol{\beta})$. The detailed steps of the whole algorithm is given in Algorithm 1 in particular, the hybrid Gibbs sampler steps are precise.

Step 2: Stochastic approximation step. A stochastic approximation is done on the sufficient statistics using the simulated value of the missing data:

$$s_{m+1} = s_m + \Delta_m[S((c)_{m+1}, (\boldsymbol{\beta})_{m+1}) - s_m]$$

where $\Delta = (\Delta_m)_m$ is a decreasing sequence of positive step-sizes.

Step 3: Maximization step. The parameters are updated using the previous formula (4.4) where the sufficient statistics are replaced by their stochastic approximations.

$$\theta_{m+1} = \arg\min_{\theta \in \Theta} \hat{\theta}(s_{m+1}).$$

The initial values $(\boldsymbol{\beta})_0$, $(c)_0$, s_0 and θ_0 are arbitrarily chosen (see Algorithm 1).

Algorithm 1 SAEM-MCMC Algorithm (with no reprojection)

Require: $c = (c)_0$, $\boldsymbol{\beta} = (\boldsymbol{\beta})_0$, θ_0, s_0, Δ
  `Stochastic Approximation Expectation-Maximization`
  for $m = 0$ to iters **do**
    `Simulation step using Gibbs sampler:`
    for $i = 1$ to n **do**
      for $p = 1$ to $3k_g$ **do**
        `Metropolis-Hastings procedure`
        $b \sim \mathcal{N}\left(\frac{\sum_{q \neq p} R_{p,q}\boldsymbol{\beta}_i^q}{R_{p,p}}, \frac{1}{R_{p,p}}\right)$
        `Compute` $r_p(\boldsymbol{\beta}_i^p, b; \boldsymbol{\beta}_i^{-p}, c, \theta_m) = \left[\frac{q(c|\boldsymbol{\beta}_{i,b\to p})}{q(c|\boldsymbol{\beta}_i)} \wedge 1\right]$
        `With probability` $r_p(\boldsymbol{\beta}_i^p, b; \boldsymbol{\beta}_i^{-p}, c, \theta_m)$, `update` $\boldsymbol{\beta}_i^p : \boldsymbol{\beta}_i^p \leftarrow b$
      end for
      `Update` $\boldsymbol{\beta}_{i,m+1} \leftarrow \boldsymbol{\beta}_i$
      for $j = 1$ to $|\Lambda|$ **do**
        $C \sim \sum_{k=1}^{K} \delta_k P_k(x_j - z_{\boldsymbol{\beta}_{i,k+1}}(x_j))$
        `Compute` $r_j(c_i^j, C; c_i^{-j}, \theta_m) = \left[\frac{q(y|c_{i,C\to j}, \theta_m)}{q(y|c_i, \theta_m)} \wedge 1\right]$
        `With probability` $r_j(c_i^j, C; c_i^{-j}, \theta_m)$, `update` $c_i^j : c_i^j \leftarrow C$
      end for
      `Update` $c_{i,m+1} \leftarrow c_i$
    end for
    `Stochastic approximation step:`
    $s_{m+1} = s_m + \Delta_m[\mathcal{S}((c)_{m+1}, (\boldsymbol{\beta})_{m+1}) - s_m]$
    `Maximization step:`
    $\theta_{m+1} = \arg\min_{\theta \in \Theta} \hat{\theta}(s_{m+1})$.
  end for

4.3. Convergence analysis

We prove the almost sure convergence of the previous estimation algorithm towards the MAP estimator given a n-sample of observations. This proof requires to add an intermediate step in the estimation algorithm. This consists in projecting the sufficient statistics on increasing compact subsets when the stochastic approximation reaches a too large value. We refer to [24] for more details about this usual additional step. Note that in practice, no projection has been required in our experiments.

Let us first define some quantities that are required in the following Theorem.

Definition 1 *Let $\mathcal{S}$ be the open subset of $\mathbb{R}^s$ defined by Equation (4.3). We define the mean field $h : \mathcal{S} \to \mathbb{R}^s$ as $h(s) = \int \int_{\mathbb{R}^{3k_g}} H_s(c, \boldsymbol{\beta}) q_{post}(c, \boldsymbol{\beta} | \boldsymbol{y}, \hat{\theta}(s)) dc d\boldsymbol{\beta}$ where $H_s(c, \boldsymbol{\beta}) = S(c, \boldsymbol{\beta}) - s$. Let also $w : \mathcal{S} \to [0, \infty[$, $w(s) = -\mathfrak{l}(\hat{\theta}(s))$ be the corresponding Lyapunov function where $\mathfrak{l}$ is the incomplete data log-likelihood: $\mathfrak{l}(\theta) = log \int \int_{\mathbb{R}^{3k_g}} q(y, c, \boldsymbol{\beta}, \theta) dc d\boldsymbol{\beta}$. Let $\mathcal{L} \triangleq \{s \in \mathcal{S}, \langle \nabla w(s), h(s) \rangle = 0\}$ be the set of critical points of the observed likelihood.*

Theorem 3 (Convergence of our estimation algorithm for model 2.5) *Assume that there exists an $M_0 > 0$ such that $\mathcal{L} \subset \{s \in \mathcal{S}, w(s) < M_0\}$. Assume also that the sequences $\boldsymbol{\Delta} = (\Delta_m)_{m \geqslant 0}$ and $\boldsymbol{\varepsilon} = (\varepsilon_m)_{m \geqslant 0}$ are non-increasing, positive and satisfy: $\sum_{m=0}^{\infty} \Delta_m = \infty$, $\lim_{m \to \infty} \varepsilon_m = 0$ and $\sum_{m=1}^{\infty} \{\Delta_m^2 + \Delta_m \varepsilon_m^a + (\Delta_m \varepsilon_m^{-1})^p\} < \infty$, where $a \in]0, 1[$ and $p \geqslant 1$. Then there exists a compact set $\mathrm{K} \subset \mathcal{Z}$ where $\mathcal{Z} = [\![1, K]\!]^{|\Lambda|} \times \mathbb{R}^{3k_g}$ and there exists another compact subset $\mathcal{K}_0 \subset \mathcal{W}_{M_0} \triangleq \{s \in \mathcal{S},\ w(s) \leqslant M_0\}$ such that for all $((c)_0, (\boldsymbol{\beta})_0) \in \mathrm{K}$ and $s_0 \in \mathcal{K}_0$, we have $\lim_{m \to \infty} d(s_m, \mathcal{L}) = 0$ $\bar{\mathbb{P}}_{(c)_0, (\boldsymbol{\beta})_0, s_0}$-a.s, where $\bar{\mathbb{P}}_{(c)_0, (\boldsymbol{\beta})_0, s_0}$ is the probability measure associated with the chain $((c)_m, (\boldsymbol{\beta})_m, s_m)_{m \geqslant 0}$ starting at $((c)_0, (\boldsymbol{\beta})_0, s_0)$.*

We prove that the stochastic approximation sequence generated by our model and algorithm satisfies Assumptions (**A1'**), (**A2**) and (**A3'**) defined in [14]. The proof is postponed to appendix.

5. Segmentation of new individuals.

Once the atlas has been estimated, one would like to perform some posterior segmentation of new observations. This can easily be done using atlas-based segmentation methods as in [7, 8]. Our model can be used directly but this typically requires heavy computations. This complexity is not a problem when creating an atlas since this step has to be only performed once. However, the atlas based segmentation procedure has to be numerically efficient. To that purpose, we propose to use a different tool keeping all the specific aspects of the model, i.e. the parameters μ_k, σ_k^2, α_k and Γ_g.

More precisely, thanks to our estimated probabilistic template $(\hat{P}_k)_{1 \leqslant k \leqslant K}$ given by Equation (2.4) with the estimated weights $(\hat{\mu}_k)_{1 \leqslant k \leqslant K}$, we define the estimated gray level template image as

$$\hat{I} = \sum_{k=1}^{K} \hat{\mu}_k \hat{P}_k.$$

This template is defined on the whole space D. Note that this formulation of the template accounts for partial volume effect (PVE) in voxels.

Our atlas also provides the geometric variability of the population through the covariance matrix $\hat{\Gamma}_g$. We use this matrix as a metric for the space of deformations to constrain the registrations according to the learned distribution.

Given a target image y, the template $\hat{I}$ is deformed non-rigidly and registered to the target image by minimizing the classical energy:

$$E(\varphi) = \frac{1}{2}\|\varphi\|^2_{\hat{\Gamma}_g} + \int \frac{1}{2\hat{\sigma}^2_{c_{\varphi^{-1}(x)}}}(y(x) - \hat{I} \circ \varphi^{-1}(x))dx$$

The first term on the right hand side yields the cost of the deformation using the metric given by $\hat{\Gamma}_g$ whereas the second term quantifies the similarity between the observed image and the deformed template. The tradeoff between these two terms is given by the noise variances which have also been estimated to best fit the noise in the training dataset. Note that this noise also accounts for the fact that the images are not drawn by this simple approximating model.

Remark 5 *We notice that in the numerical experiment, the variances of the tissue gray levels are very close to each other so that we assume in this posterior segmentation that they are all equal to* σ^2.

Our model assumes that the deformations are linearized deformations given by control point movements. We keep this assumption and therefore the energy only depends on the vector of control point displacement $\boldsymbol{\beta}$. Using also Remark 5, we can approximate the integral by

$$E(\varphi_{\boldsymbol{\beta}}) \simeq \frac{1}{2}\boldsymbol{\beta}^T\hat{\Gamma}_g^{-1}\boldsymbol{\beta} + \frac{1}{2\sigma^2}\sum_{x\in\Lambda}\left(y(x) - \hat{I}(x - z_{\boldsymbol{\beta}}(x))\right)^2. \tag{5.1}$$

We use a gradient descent algorithm to minimize Equation (5.1) which yields

$$\boldsymbol{\beta}^* = \underset{\boldsymbol{\beta}\in\mathbb{R}^{3k_g}}{\operatorname{argmin}} \frac{1}{2}\boldsymbol{\beta}^T\hat{\Gamma}_g^{-1}\boldsymbol{\beta} + \frac{1}{2\sigma^2}\sum_{x\in\Lambda}\left(y(x) - \hat{I}(x - z_{\boldsymbol{\beta}}(x))\right)^2.$$

Then the tissue c_j^* for each voxel j of the new observation is chosen to be the class that maximizes the posterior probability of that voxel to belong to each class, given this deformation field $\boldsymbol{\beta}^*$,

$$c_j^* = \underset{c_j\in[\![1,K]\!]}{\operatorname{argmax}} \left[\log\left(q(y_j|c_j,\hat{\theta})q(c_j|\boldsymbol{\beta}^*,\hat{\theta})\right)\right].$$

The segmentation is therefore constrained by both the estimated template and the learned geometric variability.

6. Experiments and Results

We first test our algorithm on simulated data to check that it reaches our objectives, that is to say (1) recover the template image as probability maps, (2) estimate a relevant covariance matrix of deformations, (3) achieve a good estimate of the mean and variance of each class and (4) segment the observations. Then we test on real data and compare with the segmentations provided by SPM8 [25] , FAST [26] in FSL [27] and DARTEL [28] algorithms. The segmentation method in SPM8 can be used for bias correction, spatially normalizing or segmenting the data, it uses the same model as in [19]. FAST segments a 3D image of the brain into different tissue types. The underlying method is based on a hidden Markov Random Field model and an associated Expectation-Maximization algorithm. DARTEL is an algorithm for diffeomorphic image registration that registers images by computing a flow field, which can be exponentiated to generate both forward and backward deformations.

As the SAEM algorithm is an iterative procedure, we run until 250 iterations which reaches numerical convergence. We control the convergence visually on the template and numerically looking at the convergence curve of the variances. For the initialization of our algorithm, we choose $(\boldsymbol{\beta})_0 = 0$ and the initial random classification $(c)_0$.

6.1. Simulated data

In the simulated data experiment, a $24 \times 24 \times 3$ image of 4 classes is used as the reference image where the values of each class are $\{1, 2, 3, 4\}$. We generate 20 images with translations and zooms and add an independent Gaussian noise with zero mean and standard deviation 0.2 to the deformed image.

We take 64 fixed control points for the deformation model given in Equation (2.3), i.e. one control point in each $3 \times 3 \times 3$ cube and all the points in the image as the control points for the template model given in Equation (2.4) to obtain a complete probabilistic atlas. We choose (0.3×12) as the parameter of K_g, where 0.3 is the value that gives the best visual result as in [13] and 12 is a half of the largest dimension size. For the hyper-parameters, we use $a_g = 0.5$, $\Gamma_g^0 = Id$, $a_p = 0.1$ and $\sigma_0^2 = 1$. The values of a_g and a_p are especially small in practice despite the constraints of theoretical definition of the priors. Note that in Equation (4.4), they weight the priors against the data-derived terms weighted by the number of observations. Small values down-weight the priors and increase the importance of data in the estimates. Although the prior laws on Γ_g^0 and σ_0^2 are improper, the posterior laws are well defined. The main purpose of the regularization is to make Γ_g symmetric positive definite. Although it would be possible to use an informative prior on Γ_g^0 as the one used in [13], the results are similar as long as the deformations are well captured.

The results are shown in Fig.2. In the first column, each row corresponds to one slice of three exemplars of the dataset. The final estimated segmentation for each individual is shown in the second column. The most important aspect is that we get the probabilistic template in the third column, each row corresponds to one class. Each voxel belongs to one class with high probability (white) and low probability (black).

Our probabilistic maps are sharp. Most voxels in each class have a probability larger than 0.9. Only few voxels on the boundary of two classes have a non zero probability to belong to two classes. This particularly sharp template demonstrates that the deformations and segmentations have both been well captured through the simulation process. The other parameters are also well estimated upholding the theoretical convergence of our algorithm. The fourth exemplar has a large deformation compared to the others. Thanks to the coupled classification-registration, we can see that our algorithm manages to capture this large deformation and yields the corresponding classification. We calculate the Jaccard index for each class (Table 1) which demonstrates that the segmentation done during the atlas estimation is accurate.

Then we generate 20 new images, and do the segmentation using the probabilistic maps and the estimated geometric variability as presented in section 5. We calculate the Jaccard index in each class as a quantitative validation (Table 1). Only at most 5% of the voxels are misclassified in particular in class 2 which corresponds to the third classes from the center of the image. However, there is only a very thin layer of this tissue in the training images which makes it hard to detect.

In our model, we have a high dimensional parameter Γ_g that is associated with the atlas estimation and imposes to increase the number of observations to get an accurate estimate. To see whether its estimation improves the results, we also run our algorithm without estimating Γ_g. To compare different situations, we fixed different values of Γ_g, $\Gamma_g = 0.5Id,\ Id,\ 2Id$ and $4Id$. We show the estimated probabilistic templates for different values of Γ_g in Fig.3. Each column corresponds to $\Gamma_g = 0.5Id,\ Id,\ 2Id$ and $4Id$, each row corresponds to one class. The voxel belongs to one class with high probability (white) and low probability (black). We can see that the shape of the template do not fit the data as well for $\Gamma_g = Id$ and $4Id$. It seems that we get a better template for $\Gamma_g = 0.5Id$. Compred to our estimated probabilistic template in Fig.2, our maps are sharper than that obtained with any fixed Γ_g and the shape of the template fits better the data.

The segmentation results with these fixed Γ_g are shown in Fig. 4. In the first column, each row corresponds to

one slice of one exemplar of the dataset. The second to the fifth columns show the final estimated segmentation for each individual for $\Gamma_g = 0.5Id,\ Id,\ 2Id$ and $4Id$. The main problems appear on the fourth exemplar that has a large deformation. None of these values of Γ_g manages to segment this observation well. Despite it simplicity, this example shows the importance of constraining the deformations to relevant ones with respect to the population. This results confirms the classification performances presented in [14] in similar context.

As a quantitative evaluation, we calculate the Jaccard index for each class for different values of Γ_g (Table 2). We can see that we get a better value of Jaccard index for class 1 and class 2 when the value of Γ_g increase. However, for class 3 and class 4, the value of Jaccard index increases first and then decreases when the value of Γ_g becomes too large. By considering both the template and the Jaccard index, we get a better template for $\Gamma_g = 0.5Id$, however we get a poor Jaccard index. From the first row of Table 1 and Table 2, we can see that our model always gets a better Jaccard index than the model with fixed values of Γ_g.

In summary, our model always gets a better result than the model with fixed values of Γ_g. However, it would be difficult to choose the optimal value of Γ_g if we want to fix it. Furthermore, we do not know the dependence between the motion of control points, which may be very complex. Therefore, this justifies to estimate the high dimensional parameter Γ_g in our model. Another argument is that the atlas estimation has to be performed only once in each population. Therefore, it could be interesting to spend some time to get an accurate estimation so that the following tasks based on these parameters reach a better performance.

6.2. Real data

The proposed method was also tested on real MRI data, derived from manual annotations that are publicly available at the Internet Brain Segmentation Repository (IBSR) [29]. Eight images are available. Each image is the size of $160 \times 160 \times 128$ with resolution $0.9735 \times 0.9735 \times 1.5\ mm^3$. The images were considered to have 3 tissue classes: gray matter (GM), white matter (WM) and CSF+background. Each tissue class follows a gaussian distribution. The variances are class dependent rather than homogeneous.

We take 800 fixed control points for the deformation model given in Equation (2.3), corresponding to one control points in each $16 \times 16 \times 16$ cube and $80 \times 80 \times 64$ points in the image as the control points for the template model given in Equation (2.4), corresponding to one control point in each $2 \times 2 \times 2$ cube. We choose $(0.3 \times 80)^2$ as the parameter of K_g as for synthetic images. For the hyper parameters, we choose $a_g = 0.5$, $\Gamma_g^0 = Id$, $a_p = 0.1$ and $\sigma_0^2 = 1$ for the same reasons as above. For comparison purpose, we always present the same image slice for all methods in these experiments.

For the first experiment, we run our algorithm with 8 patient images as training data. These images are provided with their segmentation, allowing for the validation of our online segmentation of the training images.

At first, we compare our estimated template with DARTEL template which use the SPM's segmentation as input and the average template without deformation. The first three columns in Fig. 5 show one slice of DARTEL template (first column), our probabilistic template with 8 subjects (second column) and the average template without deformation (third column), the first and second rows correspond to GM and WM respectively. Because of the smoothing step that creates regular contours, the DARTEL template is smoother than our SAEM template. Moreover, the anatomical prior template used in DARTEL makes the output very contrasted (almost binary). For DARTEL template, zone 1 shows wide CSF digitations and zone 2 shows large primary visual cortex (V1) pattern, which are much thinner in the data. Also notice that DARTEL requires pre-segmentation of the data and does not provide the geometric distribution of the population. Our model only takes into account the training data and thus avoids creating these biases. The weakly contrasted template may also be an advantage as it explains the uncertainty on voxels coming from both the PVE and the registration level of details. Another bias is shown in

zone 3 where our probabilistic maps capture the presence of the cerebral white matter. Thanks to the deformation estimated along the atlas estimation, the presence of the cerebral white matter in zone 3 for our template is sharper than the one for the average template and the shape of the V1 in zone 2 fits better the data (first column in Fig. 7). The demonstrates that incorporating the deformation metric improves the atlas estimation.

In order to evaluate the estimated geometric variability, we use our generative model to resample some images that should be representative of the population (Fig. 6). Our model manages to capture the global and local deformations. The second brain has a more round shape, the forth one a more elliptical shape and the last one is larger than the others. This global shape changes are therefore well captured. The deformations of the ventricles are realistic as well as the cortex foldings which look like some training ones. Moreover, the cortex thickness changes (highlighted in red). This shows that even with a small sample, we manage to capture the population geometric variability accurately.

The segmentation results are shown in Fig. 7. Each row of the first column corresponds to the same slice (128×160 voxels) of three training data. The manual segmentation for each individual is shown in the second column. Our algorithm gives the final classification for each exemplar in the third column. We also get the segmentation from SPM8 and FAST algorithm in the last two columns. SPM8 and FAST outputs are probabilistic, hence we define the deterministic tissue class of each voxel by the tissue which has the maximum probability in this voxel. However, the SPM8 tissue probability maps are sharp. Therefore, there are almost no difference between the probability maps and the binary result (deterministic) we used in our comparisons. These methods assume that there exists three classes, the class of CSF is considered as CSF+background in our case. Our segmentation looks accurate as it shows the folds of the V1 as in zone 4. Moreover, thanks to the class dependent variance, which is estimated along the algorithm iterations, there is no misclassification of voxels that creates holes with both SPM8 and FAST. This can be seen both in zone 4 on both sides of the cortex foldings. Furthermore, the segmented cortical thickness by both SPM8 and FAST is much smaller than that given by the manual segmentation. This may come from two different aspects of these algorithms. First, they rely on a template that is not estimated with the observations and therefore may create a bias on the cortical thickness. Moreover, the registration is not done simultaneously with the segmentation. This may also create this bias as the deformation is crucial as already noticed for the synthetic examples. The FAST tissue probability maps are fuzzy, either registration or segmentation is not well done, therefore it exists too much uncertainty which creates these facts.

Our method fails to segment the subcortical structures. The voxels belonging to these structures have the values between the GM and WM means in the training set. Therefore, they are either classified by GM or by WM. For example, the putamen's gray level is closer to WM than GM mean in zone 5a, therefore it is misclassified by our method. On the other hand, in zone 5b, as the gray level of the thalamsus proper reaches a value closer to GM, our algorithm performs better. We can notice that both SPM8 and FAST capture these structures (however not entirely, see zone 5b). This is made possible thanks to their prior templates used for segmentation which contains these structures and thus guide the segmentation around these positions. In our model, there is no informative anatomical prior set on the template and on the segmentations. Hence, the algorithm fails to fully classify these parts as GM.

To quantify the visual performance, we calculate the Jaccard index for each class for different methods (Table 3). We perform much better for the gray matter, as we succeed in segmenting the cortex with the right thickness whereas SPM8 and FAST reduce it. However the Jaccard index for the white matter is a little worse, this is a result of the misclassified subcortical structures. To compensate for the misclassified subcortical structures, we try to use DARTEL template as the informative prior on the probability maps $(\alpha_k)_{1 \leqslant k \leqslant K}$. However the gray level plays a greater role than the prior in the process, we lose the prior gradually and still fail to classify these structures.

For evaluating the segmentation of new individual, we create the probabilistic atlas with 20 images and then segment 5 new images using the estimated atlas. The database IBSR has not enough images, therefore, we mix 3 images from IBSR and 17 images from Open Access Series of Imaging Studies (OASIS) [30] to obtain 20 images. There are 416 subjects aged from 18 to 96 with resolution $1 \times 1 \times 1\ mm^3$ in the OASIS dataset. The 17 images aged from 28 to 64 are chosen randomly. For the pre-processing, we use BET [31] to delete the non-brain tissue from the images of the database OASIS. The segmentation (Fig. 8) looks quite accurate, except the subcortical structures. We get a Jaccard index (Table 4) around 75% for each tissue type. The rate for GM is better than SPM8 and FAST for the reason we already noticed before. The rate for WM is similar to those of other algorithms. The segmentation accuracy increases with the number of training images. The fourth column in Fig. 5 shows one slice of our probabilistic maps with 20 subjects. The template obtained with 20 subjects captures more details on the boundary of two types than the one with 8 subjects. This appears in particular on both right and left cerebral cortex areas where the 8 subjects template classifies voxels only belonging to GM whereas the 20 subjects template captures the presence of WM voxels. Although our model has high dimensional parameters, we obtain a reasonable estimate with 20 images. The computation time is about 10 days for 8 images and almost a month for 20 images. Since our algorithm can be parallelized for the simulation step, we are working on a parallel C++ version of our code to make it possible to increase the training set and decrease the computation time. With the parallel version, it should cost about 1 day for 8 images and 3 days for 20 images.

7. Conclusion and discussion

In this study, we proposed a statistical model and used a stochastic algorithm to perform a probabilistic atlas estimation. This model opens the way to performing registration and segmentation simultaneously along the probabilistic atlas estimation. We also provide a proof of convergence of what toward a critical point of the observed likelihood. Our algorithm has several advantages. First, the probabilistic atlas contains both the templates and the geometric variability of the population. Second, we do not need any pre-registration to perform the segmentation which is automatically obtained as an output. Third, the estimated atlas can be used for segmenting new individuals. The experiments show that the proposed approach compares well with state-of-the art tools.

Our experiments also show that our model does not manage to segment the subcortical structures. This is easily explained by the fact that the image grey level provides ambiguous information in these regions, and that the segmentation is an ill-posed problem in the absence of prior information. One possible solution is thus to use the anatomical prior as SPM8 and FAST. Another solution is to use multimodal registration and segmentation. Multimodal images enable to take advantage of the different information given by different imaging modalities. In a recent generalisation of this model [33], we manage to segment these structures using T1- and functional MRI. Another improvement would be to consider diffeomorphic deformations as in [12] or [34]. This control on the deformations would help to respect anatomical constraints. However, one should keep a parametric description in order to be able to sample these deformations easily.

A. Proof of Theorem 1

From Equation (2.7), we have $q(y|\theta) = \sum_{k=1}^{K} q(y|c=k,\theta_c) \int q(c|\boldsymbol{\beta},\theta_g) q(\boldsymbol{\beta}|\theta_p) d\boldsymbol{\beta}$.
Since the right hand side term of Equation (2.9) is bounded by 1 (as it is a probability distribution),

$$\begin{aligned} q(y|\theta) &\leqslant \sum_{k=1}^{K} q(y|c=k,\theta_c) \int q(\boldsymbol{\beta}|\theta_p) d\boldsymbol{\beta} \\ &\leqslant \sum_{k=1}^{K} q(y|c=k,\theta_c) \\ &= \sum_{k=1}^{K} \frac{1}{\sqrt{2\pi\sigma_k^2}^{n|\Lambda|}} \exp\left(-\frac{\sum_{i=1}^{n}\sum_{j=1}^{|\Lambda|}(y_i^j - \mu_k)^2}{2\sigma_k^2} \right) \end{aligned}$$

where n is the number of images and $|\Lambda|$ is the number of voxels on the grid Λ.
We denote

$$f(\sigma_c^2) = \frac{1}{\sqrt{2\pi\sigma_c^2}^{n|\Lambda|}} \exp\left(-\frac{S_c^2}{2\sigma_c^2} \right)$$

where $S_c = \sum_{i=1}^{n}\sum_{j=1}^{|\Lambda|}(y_i^j - \mu_c)^2$. We want to bound f on $\mathbb{R}_*^+$, let f' be its derivative:

$$f'(\sigma_c^2) = \frac{1}{\sqrt{2\pi}^{n|\Lambda|}} (\sigma_c^2)^{-\frac{n|\Lambda|}{2}-1} \exp\left(-\frac{S_c^2}{2\sigma_c^2} \right) \left(\frac{S_c^2}{2\sigma_c^2} - \frac{n|\Lambda|}{2} \right).$$

For $\tilde{\sigma}_c^2 = \frac{S_c^2}{n|\Lambda|}$, $f'(\tilde{\sigma}_c^2) = 0$ and $f''(\tilde{\sigma}_c^2) < 0$, so that for all $\sigma_c^2 > 0$,

$$f(\sigma_c^2) \leqslant f(\tilde{\sigma}_c^2) = \frac{1}{\sqrt{2\pi}^{n|\Lambda|}} \left(\frac{S_c^2}{n|\Lambda|} \right)^{-\frac{n|\Lambda|}{2}} \exp\left(-\frac{n|\Lambda|}{2} \right).$$

Therefore

$$\begin{aligned} \log(q_B(\theta|y_1,...y_n)) \leqslant & \log \sum_{k=1}^{K} \left(\frac{\sum_i \sum_j (y_i^j - \mu_k)^2}{n|\Lambda|} \right)^{-\frac{n|\Lambda|}{2}} + \frac{a_g}{2}\log|R_g| \\ & - \frac{a_g}{2}\langle R_g, \Gamma_g^0 \rangle - \sum_{k=1}^{K} \left(\frac{a_p \sigma_0^2}{2\sigma_k^2} + \frac{a_p}{2}\log\sigma_k^2 \right) - \frac{(\mu_k - m_u)^2}{2\sigma_u^2} + C \end{aligned}$$

where $R_g = \Gamma_g^{-1}$, and C is a constant which does not depend on the parameters. If we denote η_g^0 the smallest eigenvalue of Γ_g^0 and $\|R_g\|$ the operator norm of R_g (which is also its largest eigenvalue), we get

$$\langle R_g, \Gamma_g^0 \rangle \geqslant \eta_g^0 \|R_g\| \text{and } \log(|R_g|) \leqslant (3k_g - 1)\log\|R_g\| - \log\|\Gamma_g\|$$

so that

$$\lim_{\|R_g\|+\|\Gamma_g\|\to\infty} -\frac{a_g}{2}\langle R_g, \Gamma_g^0 \rangle + \frac{a_g}{2}\log|R_g| = -\infty.$$

Similarly, we can show that

$$\forall k \in [\![1, K]\!], \lim_{\sigma_k^2+\sigma_k^{-2}\to\infty} \frac{a_p\sigma_0^2}{2\sigma_k^2} + \frac{a_p}{2}\log\sigma_k^2 = -\infty.$$

Moreover, for all $k \in [\![1, K]\!]$, there exists at least one voxel j_k in one image i_k such that $y_{i_k}^{j_k} \neq \mu_k$, otherwise all μ_k would be equal and all the images would be constant. Thus

$$\log \sum_{k=1}^{K} \left(\frac{\sum_{i=1}^{n}\sum_{j=1}^{|\Lambda|}(y_i^j - \mu_k)^2}{n|\Lambda|} \right)^{-\frac{n|\Lambda|}{2}} \leqslant \log \left(\sum_{k=1}^{K} \left(\min_{\substack{(i_k,j_k),\\ y_{i_k}^{j_k}\neq\mu_k}} \left(\frac{(y_{i_k}^{j_k} - \mu_k)^2}{n|\Lambda|} \right) \right)^{-\frac{n|\Lambda|}{2}} \right).$$

Furthermore,

$$\forall k \in [\![1, K]\!], \lim_{|\mu_k|\to\infty} \left(\frac{(y_{i_k}^{j_k} - \mu_k)^2}{n|\Lambda|} \right)^{-\frac{n|\Lambda|}{2}} = 0.$$

which implies that

$$\lim_{|\mu_k|\to\infty} \log \left(\sum_{k=1}^{K} \left(\min_{\substack{(i_k,j_k),\\ y_{i_k}^{j_k}\neq\mu_k}} \left(\frac{(y_{i_k}^{j_k} - \mu_k)^2}{n|\Lambda|} \right) \right)^{-\frac{n|\Lambda|}{2}} \right) = -\infty.$$

So that

$$\forall k \in [\![1, K]\!], \lim_{|\mu_k|\to\infty} \log \sum_{k=1}^{K} \left(\frac{\sum_{i=1}^{n}\sum_{j=1}^{|\Lambda|}(y_i^j - \mu_k)^2}{n|\Lambda|} \right)^{-\frac{n|\Lambda|}{2}} - \frac{(\mu_k - m_u)^2}{2\sigma_u^2} = -\infty.$$

Now considering the Alexandrov one-point compactification $\Theta \cup \{\infty\}$ of Θ, we have

$$\lim_{\theta\to\infty} \log(q_B(\theta|y_1, ..., y_n)) \to -\infty.$$

Since $\theta \to \log(q_B(\theta|y_1, ...y_n))$ is smooth on Θ, we get the result.

B. Proof of Theorem 3

In this Section, we prove Theorem 3. To this purpose, we will follow the path of proof in [14], i.e. prove that the stochastic approximation sequence satisfies assumptions (**A1'**)(**ii**), (**iii**), (**iv**), (**A2**) and (**A3'**). The fact that the critical points remain in a level set of the Lyapunov function remains an assumption because of the complexity of our model. We detail only the crucial steps and arguments of the proof which differ from the previously mentioned one and refer to [14] when it is identical.

The sufficient statistic vector S, the set $\mathcal{S}$ as well as the explicit expression of $\hat{\theta}(s)$ have been given in Subsection 3.2. As noted, $\hat{\theta}$ is a smooth function of $\mathcal{S}$.

*B.1. Proof of assumption (**A1'**).*

We recall that the functions H, h and w are defined in Subsection 4.3. Thanks to these particular forms, we satisfy (**A1'(iii)**) and (**A1'(iv)**) as proved in [32].

Moreover, since the interpolation kernel K_p is bounded, there exist $A > 0$, $B > 0$, $C > 0$, $D > 0$ such that for any $(c, \boldsymbol{\beta}) \in \mathcal{Z}$, we have

$$0 < S_{0,k}(c, \boldsymbol{\beta}) \leqslant A, \|S_{1,k}(c, \boldsymbol{\beta})\| \leqslant B, 0 \leqslant S_{2,k}(c, \boldsymbol{\beta}) \leqslant C, 0 \leqslant S_3(c, \boldsymbol{\beta}) \text{ and } 0 \leqslant S_{4,k,l}(c, \boldsymbol{\beta}) \leqslant D.$$

We define the set $\mathcal{S}_a$ by

$$\mathcal{S}_a \triangleq \{S \in \mathcal{S} | 0 \leqslant S_{0,k} \leqslant A, \|S_{1,k}\| \leqslant B, 0 \leqslant S_{2,k} \leqslant C, 0 \leqslant S_3 \text{ and } 0 \leqslant S_{4,k,l} \leqslant D\}.$$

Since the constraints are obviously convex and closed, we get that $\mathcal{S}_a$ is a closed convex subset of $\mathbb{R}^s$ such that

$$\mathcal{S}_a \subset \mathcal{S} \subset \mathbb{R}^s$$

and satisfying

$$s + \rho H_s(c, \boldsymbol{\beta}) \in \mathcal{S}_a \text{ for any } \rho \in [0, 1] \text{ any } s \in \mathcal{S}_a \text{ any } (c, \boldsymbol{\beta}) \in \mathcal{Z}.$$

We now focus on the first two points. As $\mathtt{l}$ and $\hat{\theta}$ are continuous functions, we only need to prove that $\mathcal{W}_M \cap \mathcal{S}_a$ is a bounded set for a constant $M \in \mathbb{R}_*^+$ with:

$$\mathcal{W}_M = s \in \mathcal{S}, w(s) \leqslant M,$$

where $w(s)$ is defined in Definition 1.

On $\mathcal{S}_a$, s_0, s_1, s_2 and s_4 are bounded; writing $\hat{\theta}(s) = (\alpha_k(s))_{1 \leqslant k \leqslant K}, (\mu_k(s))_{1 \leqslant k \leqslant K}, (\sigma_k^2(s))_{1 \leqslant k \leqslant K}, \Gamma_g(s))$, we deduce from Equation (4.4) that $(\alpha_k(s))_{1 \leqslant k \leqslant K}, (\mu_k(s))_{1 \leqslant k \leqslant K}, (\sigma_k^2(s))_{1 \leqslant k \leqslant K}$ are bounded on $\mathcal{S}_a$. Considering the sufficient statistic s_3, thus

$$w(s) \geqslant -\log\left(\int\int q(y, c, \boldsymbol{\beta}, \hat{\theta}(s)) dc d\boldsymbol{\beta}\right) \geqslant -\log(q_B(\hat{\theta}(s))) + C \geqslant -\log(q_{B|_\Gamma}(\Gamma(s))) + C,$$

where C is a constant independent of $s \in \mathcal{S}_a$. Since

$$-\log(q_{B|_\Gamma}(\Gamma(s))) = \frac{a_g}{2}(\langle \Gamma_g^{-1}, \Gamma_g^0 \rangle_F + \log|\Gamma_g|) \geqslant \frac{a_g}{2} \log|\Gamma_g|$$

and

$$\lim_{\|s\| \to +\infty, s \in \mathcal{S}_a} \log(|\Gamma_g(s)|) = \lim_{\|s\| \to +\infty, s \in \mathcal{S}_a} \log(|(s_3 + a_g \Gamma_g^0)/(n * |\Lambda| + a_g)|) = +\infty,$$

we deduce that

$$\lim_{\|s\| \to +\infty, s \in \mathcal{S}_a} w(s) = +\infty.$$

Since w is continuous and $\mathcal{S}_a$ is closed, this proves **(A1'(ii))**.

*B.2. Proof of assumption (**A2**)*

We prove the same condition (**DRI1**) defined in [14] which will imply (**A2**) under the condition that H and V are related. We, in fact, have the following property : $\exists\ C > 0$ such that :

$$\sup_{s \in \mathcal{K}} |H_s(c, \boldsymbol{\beta})| \leqslant C\ V(c, \boldsymbol{\beta}),$$

where we set $V : [\![1, K]\!]^{|\Lambda|} \times \mathbb{R}^{3k_g} \to [1, +\infty[$ as the following function

$$V(c, \boldsymbol{\beta}) = 1 + \|\boldsymbol{\beta}\|^2. \tag{B.1}$$

We now prove the following lemma which gives the existence of the small set $\mathtt{C}$ required by condition **(DRI1)**:

Lemma 1 *Any compact set of* $\mathcal{Z} = [\![1, K]\!]^{|\Lambda|} \times \mathbb{R}^{3k_g}$ *is a small set for* $(\Pi_{\hat{\theta}(s)})_{s\in\mathcal{K}}$

Proof Let A be a Borel set of $\mathcal{Z}$ and $x \in \mathtt{C}$ a compact subset of $\mathcal{Z}$, then we have

$$\begin{aligned}
\Pi_{\hat{\theta}(s),t}(\boldsymbol{x}, A) &\geqslant \int_{A^t} \left(1 \wedge \frac{\pi_{\hat{\theta}(s),t}(z^t)}{q(z^t, \hat{\theta}(s))} \frac{q(x^t, \hat{\theta}(s))}{\pi_{\hat{\theta}(s),t}(x^t)}\right) q(z^t, \hat{\theta}(s)) dz^t \\
&\geqslant \int_{A^t} \left(\frac{q(z^t, \hat{\theta}(s))}{\pi_{\hat{\theta}(s),t}(z^t)} \wedge \frac{q(x^t, \hat{\theta}(s))}{\pi_{\hat{\theta}(s),t}(x^t)}\right) \pi_{\hat{\theta}(s),t}(z^t) dz^t \,, \\
&\geqslant \int_{A^t} \left(\frac{q(z^t, \hat{\theta}(s))}{\pi_{\hat{\theta}(s),t}(z^t)} \wedge \frac{q(x^t, \hat{\theta}(s))}{\pi_{\hat{\theta}(s),t}(x^t)}\right) \pi_{\hat{\theta}(s),t}(z^t) \mathbb{1}_{\mathtt{C}}(\boldsymbol{z}) dz^t \,.
\end{aligned}$$

If we can prove that for any compact set $\mathcal{K} \in \mathcal{S}$, there exists a constant $C_{\mathcal{K},\mathtt{C}}$ such that

$$\frac{\pi_{\hat{\theta}(s),t}(z^t)}{q(z^t, \hat{\theta}(s))} \leqslant C_{\mathcal{K},\mathtt{C}} \,, \tag{B.2}$$

then:

$$\begin{aligned}
\Pi_{\hat{\theta}(s),t}(\boldsymbol{x}, A) &\geqslant \int_{A^t} \frac{1}{C_{\mathcal{K},\mathtt{C}}} \pi_{\hat{\theta}(s),t}(z^t) \mathbb{1}_{\mathtt{C}}(\boldsymbol{z}) dz^t &\text{(B.3)} \\
&\geqslant \int_{A^t} \frac{1}{C_{\mathcal{K},\mathtt{C}}} \pi_{\mathcal{K},t}(z^t) \mathbb{1}_{\mathtt{C}}(\boldsymbol{z}) dz^t \,, &\text{(B.4)}
\end{aligned}$$

where $\forall \boldsymbol{z} \in \mathtt{C},\ \pi_{\mathcal{K},t}(z^t) = \min_{s\in\mathcal{K}} \pi_{\hat{\theta}(s),t}(z^t)$ is a positive measure thanks to the smoothness of the probability measure $q(z^t|\boldsymbol{z}^{-t}, \boldsymbol{y}, \hat{\theta}(s))$ in its parameter s for all $\boldsymbol{z} \in \mathtt{C}$.

Let us now prove (B.2).

$$\begin{aligned}
\pi_{\hat{\theta}(s),t}(\boldsymbol{z}) &= q(z^t|\boldsymbol{z}^{-t}, \boldsymbol{y}, \hat{\theta}(s)) \\
&= \frac{q(\boldsymbol{y}|\boldsymbol{z}, \hat{\theta}(s)) q(z^t, \hat{\theta}(s))}{q(\boldsymbol{y}|\boldsymbol{z}^{-t}, \hat{\theta}(s))} \\
\frac{\pi_{\hat{\theta}(s),t}(\boldsymbol{z})}{q(z^t, \hat{\theta}(s))} &\leqslant \frac{1}{\int \prod_{j=1}^{|\Lambda|} \exp\left(-\frac{1}{2\hat{\sigma}^2_{c_j}(s)}(y_j - \mu_{c_j})^2\right) q(z^t, \hat{\theta}(s)) \mathbb{1}_{\mathtt{C}}(\boldsymbol{z}) dz^t}
\end{aligned}$$

Since there exists $a > 0$ such that $\forall j, (y_j - \mu_{c_j})^2 \leqslant a$ and since $\sigma^2_c \geqslant \sigma^2_{min}$, then we have

$$\exp\left(-\frac{1}{2\sigma^2_{c_j}}(y_j - \mu_{c_j})^2\right) \geqslant \exp\left(-\frac{a}{2\sigma^2_{min}}\right)$$

So that

$$\begin{aligned}\frac{\pi_{\hat{\theta}(s),t}(\boldsymbol{z})}{q(z^t,\hat{\theta}(s))} &\leqslant \frac{1}{\exp\left(-\frac{a}{2\sigma^2_{min}}\right)^{|\Lambda|}\int q(z^t,\hat{\theta}(s))\mathbb{1}_{\mathtt{C}}(\boldsymbol{z})dz^t} \\ &\leqslant \exp\left(\frac{a}{2\sigma^2_{min}}\right)^{|\Lambda|} \quad \text{car} \int q(z^t,\hat{\theta}(s))\mathbb{1}_{\mathtt{C}}(\boldsymbol{z})dz^t = 1.\end{aligned}$$

It is bounded by $C_{\mathcal{K},\mathtt{C}} = \exp\left(\frac{a}{2\sigma^2_{min}}\right)^{|\Lambda|}$ on $\mathcal{K}$ for any $z \in \mathtt{C}$. The complete transition kernel is a composition of the previous kernel for t from 1 to $3k_g + |\Lambda|$. Since the coordinate of z are independent we get:

$$\Pi_{\hat{\theta}(s)}(\boldsymbol{x},A) \geqslant \int_A \left(\frac{1}{C_{\mathcal{K},\mathtt{C}}}\right)^{3k_g+|\Lambda|}\left(\prod_{t=1}^{3k_g+|\Lambda|}\pi_{\mathcal{K},t}(z^t)\right)\mathbb{1}_{\mathtt{C}}(\boldsymbol{z})d\mathbf{z}$$

This yields the existence of the small set and the third condition of (**DRI1**) with

$$\delta\nu(A) = \int_A \left(\frac{1}{C_{\mathcal{K},\mathtt{C}}}\right)^{3k_g+|\Lambda|}\left(\prod_{t=1}^{3k_g+|\Lambda|}\pi_{\mathcal{K},t}(z^t)\right)\mathbb{1}_{\mathtt{C}}(\boldsymbol{z})d\mathbf{z}$$

and ends the proof.

For proving the first conditions of (**DRI1**), we need to ensure that our acceptance rates are always strictly positive. We notice that $\tilde{r}_p(\boldsymbol{\beta}^p, b; \boldsymbol{\beta}^{-p}, c, \theta) = \frac{q(c|\boldsymbol{\beta}_{b\to p})}{q(c|\boldsymbol{\beta})} \geqslant q(c|\boldsymbol{\beta}_{b\to p}) > 0$, because for $\forall\ (c,\boldsymbol{\beta}) \in \mathcal{Z}$, $q(c|\boldsymbol{\beta}) \in]0,1[$ which is justified in Remark 4. Therefore, for any compact set $\mathcal{K} \subset \mathcal{S}$, $\exists\ a_{\mathcal{K}} > 0 :\ \forall\ r_p(\boldsymbol{\beta}^p, b; \boldsymbol{\beta}^{-p}, c, \theta) \geqslant a_{\mathcal{K}}$. The other proof of the first and second conditions of (**DRI1**) is similar with the proof in [14] with the function V defined in Equation (B.1).

B.3. Proof of assumption (A3')

For proving the Hölder condition **(A3'(ii))**. We will use the lemma 6.4 and lemma 6.5 in [14] which state Lipschitz conditions on the transition kernel and its iterates. If we can prove that the derivative of the acceptance rates in our model are Lipschitz functions, we get the result of lemma 6.4.

Proof Concerning the derivative of $r_p(\boldsymbol{\beta}^p, b; \boldsymbol{\beta}^{-p}, c, \theta)$, since

$$\log(\tilde{r}_p(\boldsymbol{\beta}^p, b; \boldsymbol{\beta}^{-p}, c, \theta)) = \sum_{j=1}^{|\Lambda|}\log\alpha_k(c_j = k|\boldsymbol{\beta}_{b\to p}) - \sum_{j=1}^{|\Lambda|}\log\alpha_k(c_j \neq k|\boldsymbol{\beta}),$$

we have $|\frac{d}{d\epsilon}\tilde{r}_p(\boldsymbol{\beta}^p, b; \boldsymbol{\beta}^{-p}, c, \theta))| \leqslant C_{\mathcal{K}}\|s' - s\|$.

Concerning the derivative of $r_j(c^j, k; c^{-j}, \theta)$,

$$\begin{aligned}\left|\frac{d}{d\epsilon}r_j(c^j,k;c^{-j},\theta)\right| &= \left|-\frac{1}{2}\frac{d}{d\epsilon}\log\sigma_k^2 - \frac{d}{d\epsilon}\frac{(y_j-\mu_k)^2}{2\sigma_k^2}\right| \\ &= \left|-\frac{1}{2\sigma_k^2}\frac{d}{d\epsilon}\sigma_k^2 + \frac{(y_i-\mu_k)^2}{2\sigma_k^4}\frac{d}{d\epsilon}\sigma_k^2 - \frac{y_i-\mu_k}{2\sigma_k^2}\frac{d}{d\epsilon}\mu_k\right| \\ &\leqslant C_1\left|\frac{d}{d\epsilon}\sigma_k^2\right| + C_2\left|\frac{d}{d\epsilon}\mu_k\right|\end{aligned}$$

where $C_1, C_2 > 0$ and

$$\left|\frac{d}{d\epsilon}\mu_k\right| = \left|\frac{s_{1,k}}{s_{0,k}} - \frac{s'_{1,k}}{s'_{0,k}}\right| = \left|\frac{s_{1,k}(s'_{0,k} - s_{0,k})}{s_{0,k}s'_{0,k}} - \frac{(s'_{1,k} - s_{1,k})s_{0,k}}{s_{0,k}s'_{0,k}}\right| \leqslant C_{\mathcal{K}}\|s - s'\|.$$

Similarly,

$$\left|\frac{d}{d\epsilon}\sigma_k^2\right| = \left|\frac{n}{n + a_p}\left(\frac{s_{2,k}}{s_{0,k}} - \frac{s'_{2,k}}{s'_{0,k}} - \frac{s_{1,k}^2}{s_{0,k}^2} + \frac{s_{1,k}'^2}{s_{0,k}'^2}\right)\right| \leqslant C_{\mathcal{K}}\|s - s'\|.$$

Thus, we have

$$\left|\frac{d}{d\epsilon}r_j(c^j, k; c^{-j}, \theta)\right| \leqslant C_{\mathcal{K}}\|s' - s\|.$$

The next proof for (**A3'(ii)**) and the proofs for **(A3')(i)(iii)** are the same as the proofs in [14].

References

1. Seghers D, Agostino E, Maes F, Vandermeulen D, Suetens P: Construction of a brain template from MR images using state-of-the art registration and segmentation technique. *MICCAI* 2004.
2. Shattuck DW, Mirza Mubeena, Adisetiyo V, Hojatkashani C, Salamon G, Narr KL, Poldrack RA, Bilder RM, Toga AW: Construction of a 3D probabilistic atlas of humain cortical structures. *NeuroImage* 2008; **39**:1064-1080.
3. Gouttard S, Styner M, Joshi S, Smith RG, Cody H, Gerig G: Subcortical structure segmentation using probabilistic atlas priors. *Proc. SPIE* 2007; **6512**.
4. Van Leemput K, Maes F, Vandermeulen D, Colchester A, Suetens P: Automated segmentation of multiple sclerosis lesions by model outlier detection. *IEEE TMI* 2001; **20**(8):667-688.
5. Drapacaa C, Cardenasb V, Studholme C: Segmentation of tissue boundary evolution from brain MR image sequences using multi-phase level sets. *CVIU* 2005; **100**:312-329.
6. Hu S, Collins DL: Joint level-set Shape Modeling and Appearance Modeling for Brain Structure Segmentation. *NeuroImage* 2007; **36**(1):672-683.
7. Shan L, Charles C, Niethammer M: Automatic Atlas-based Three-Label Cartilage Segmentation from MR Knee Images. *MMBIA* 2011.
8. Zhang D, Wu G, Jia H, Shen D: Confidence-Guided Sequential Label Fusion for Multi-atlas Based Segmentation. *MICCAI* 2011.
9. Wyatt P, Noble A: MAP MRF joint segmentation and registration of medical images. *MIA* 2003; **7**(4):539-552.
10. Pohl K, Fisher J, Grimson W, Kikinis R, Wells W: A bayesian model for joint segmentation and registration. *NeuroImage* 2006; **31**(1):228-239.
11. Hachama M, Desolneux A, Richard F: A bayesian technique for image classifying registration. *IEEE TIP* 2012; **21**(9):4080-4091.
12. Trouvé A: Diffeomorphic groups and pattern matching in image analysis. *Int. J. Computer Vision* 1998; **28**:213-221.
13. Allassonnière S, Amit Y, Trouvé A: Toward a coherent statistical framework for dense deformable template estimation. *the Journal of the Royal Statistical Society* 2007; **69**(1):3-69.
14. Allassonnière S, Kuhn E, Trouvé A: Construction of Bayesian deformable models via stochastic approximation algorithm: A convergence study. *Bernoulli Journal* 2010; **16**(3):641-678.
15. Van Leemput K: Encoding Probabilistic Brain Atlases Using Bayesian Inference. *IEEE TMI* 2009; **28**(6):822-837.
16. Bhatia K, Aljabar P, Boardman JP, Srinivasan L, Murgasova M, Counsell SJ, Rutherford MA, Hajnal J, Edwards AD, Rueckert D: Groupwise combined segmentation and registration for atlas construction. *MICCAI* 2007.
17. Ribbens A, Hermans J, Maes F, Vandermeulen D, Suetens P: SPARC: Unified framework for automatic segmentation, probabilistic atlas construction, registration and clustering of brain MR images. *IEEE ISBI* 2011;856-859.
18. Yeo BTT, Sabuncu MR, Vercauteren T, Ayache N, Fischl B, Golland P: Spherical Demons: Fast Diffeomorphic Landmark-free Surface Registration. *IEEE TMI* 2010; **29**(3):650-668.
19. Ashburner J, Friston KJ: Unified segmentation. *NeuroImage* 2005; **26**(3):839-851.
20. Allassonnière S, Kuhn E, Ratnanather JT, Trouvé A: Consistent Atlas Estimation on BME Template Model: Applications to 3D Biomedical Images. *PMMIA* 2009.
21. Richard F, Samson A, Cuénod C: A SAEM algorithm for the estimation of template and deformation parameters in medical image sequences. *Statistics and Computing* 2009; **19**(4):465-478.

22. Joshi SC, Miller MI: Landmark matching via large deformation diffeomorphisms. *IEEE Image Processing* 2000; **9**(8):1357-1370.
23. van der Vaart AW: Asymptotic statistics. Cambridge series in statistical and probabilistic mathematics. *Cambridge University Press, Cambridge* 1998.
24. Andrieu, C, Moulines, Eric and Priouret, P: Stability of stochastic approximation under verifiable conditions. *SIAM J. Control Optim.* 2005; **44**(1):283–312 (electronic).
25. Statistical Parametric Mapping (SPM8) [Online]. Available: www.fil.ion.ucl.ac.uk/spm/software/spm8/.
26. Zhang Y, Smith S, Brady M: Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE TMI* 2001; **20**(1):45-57.
27. FMRIB Software Library (FSL) [Online]. Available: http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/.
28. Ashburne J: A fast diffeomorphic image registration algorithm. *NeuroImage* 2007; **38**(1):95-113.
29. Internet brain segmentation repository (IBSR) [Online]. Available: http://www.cma.mgh.harvard.edu/ibsr.
30. Open Access Series of Imaging Studies (OASIS) [Online]. Available: http://www.oasis-brains.org/.
31. Jenkinson M, Pechaud M, Smith S: BET2: MR-based estimation of brain, skull and scalp surfaces. *In Eleventh Annual Meeting of the Organization for Human Brain Mapping* 2005.
32. Delyon B, Lavielle M, Moulines E: Convergence of a stochastic approximation version of the EM algorithm. *Ann. Statist* 1999; **27**(1):94-128.
33. Xu H, Thirion B, Allassonnière S: Bayesian estimation of probabilistic atlas for anatomically-informed functional MRI group analyses. *MICCAI* 2013.
34. Vercauteren T, Pennec X, Perchant A, Ayache N: Diffeomorphic Demons: Efficient Non-parametric Image Registration. *NeuroImage* 2009: **45**(1, Supp.1); S61-72.

	Class 1	Class 2	Class 3	Class 4
Training data	99.8%	98.6%	99.2%	99.4%
Test data	99.0%	94.4%	97.6%	97.3%

Table 1. Experiments on synthetic data. The Jaccard Index for the training data and test data average across 20 subjects.

Γ_g	Class 1	Class 2	Class 3	Class 4
0.5Id	99.4%	96.3%	98.0%	96.6%
Id	99.6%	97.2%	98.3%	97.2%
2Id	99.6%	97.7%	**98.9%**	**98.8%**
4Id	**99.7%**	**98.1%**	98.8%	97.2%

Table 2. Experiments on synthetic data. The Jaccard Index for the training data using the model with fixed Γ_g. Compared with the first row of Table 1, our model with estimated Γ_g gets the best result.

	FAST	SPM	SAEM
GM	58.6%	65.2%	**79.9%**
WM	**76.6%**	76.0%	68.9%

Table 3. Experiments on real data. The Jaccard Index for different methods average across 8 subjects. Our method gives a much higher value of Jaccard index for GM. However a little worse for WM, it is because our method does not manage to segment the subcortical structures as GM which is even difficult to segment manually. FAST and SPM8 use an anatomical prior, therefore they segment successfully these structures.

	SAEM	FAST	SPM8
GM	**75.4%**	56.9%	64.0%
WM	75.3%	77.2%	**77.3%**

Table 4. Experiments on real data. The Jaccard Index for different methods (our segmentation method for new individual using the estimated atlas with 20 subjects, FAST and SPM8) averaged across 5 subjects.

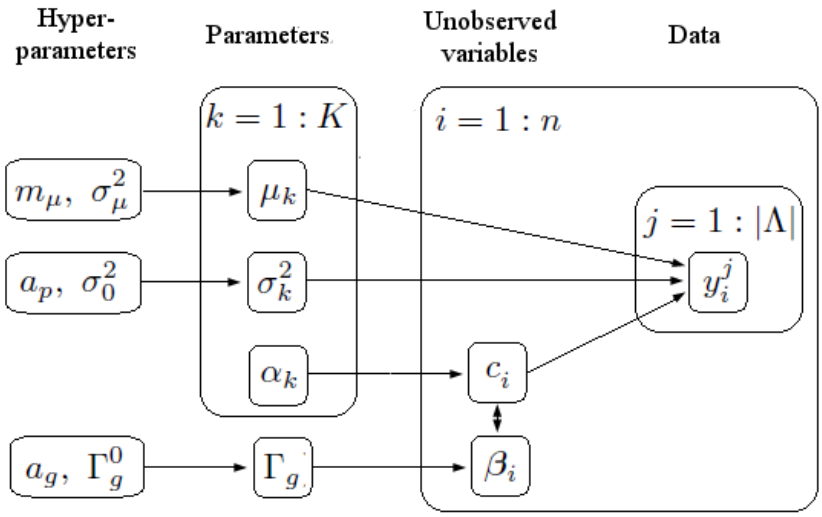

Figure 1. Our Bayesian model.

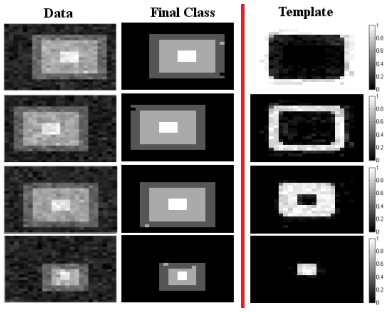

Figure 2. Experiments on simulated data. The first two columns correspond to one slice of four data images and their final segmentations. The last column correspond to the first slice of the probabilistic template, each row corresponds to a class. The white color corresponds to high probability and the black color to low one.

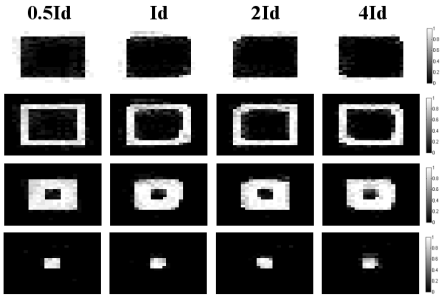

Figure 3. The estimated probabilistic templates for the model with fixed Γ_g. Each column corresponds to $\Gamma_g = 0.5Id,\ Id,\ 2Id$ and $4Id$, each row corresponds to one class. The voxel belongs to one class with high probability (white) and low probability (black).

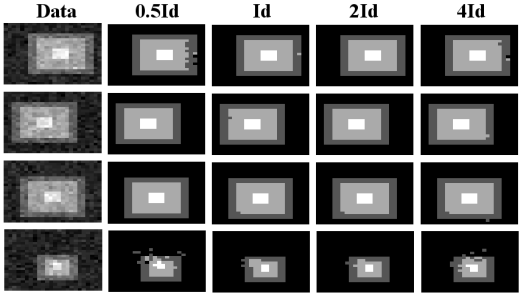

Figure 4. The segementation results for the model with fixed Γ_g. In the first column, each row corresponds to one slice of one exemplar of the dataset. The second to the fifth columns show the final estimated segmentation for each individual for $\Gamma_g = 0.5Id,\ Id,\ 2Id$ and $4Id$.

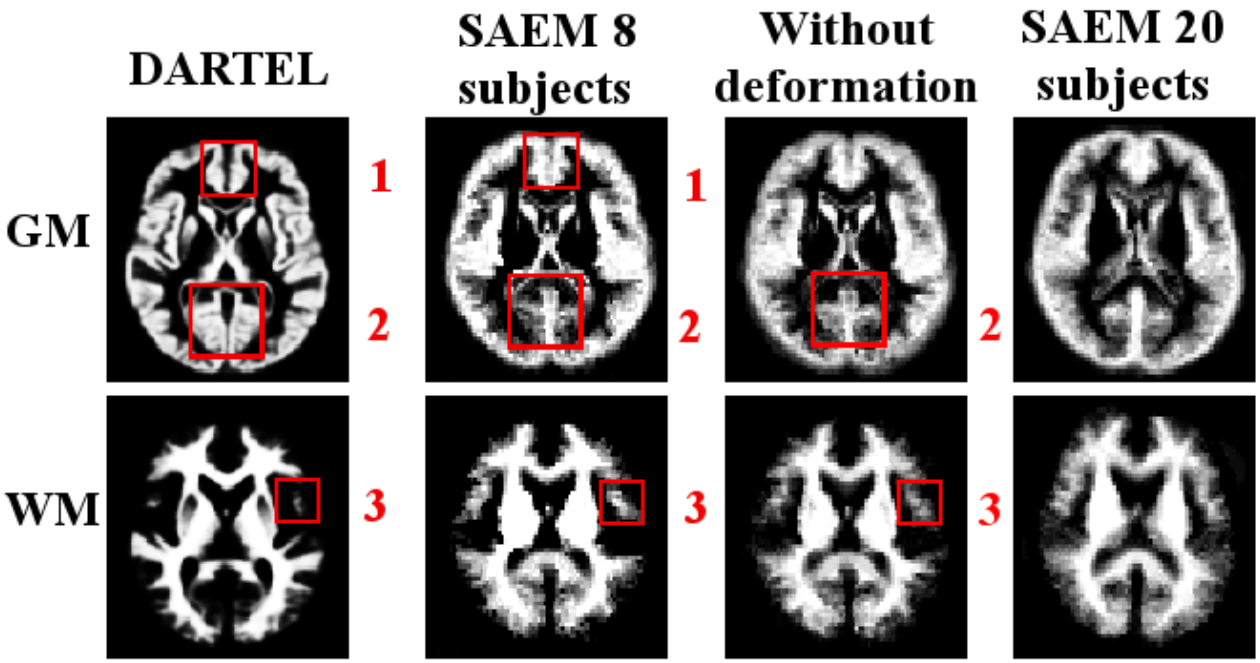

Figure 5. The template obtained by DARTEL (first column), our method with 8 subjects (second column), the average template without deformation (third column) and 20 subjects (fourth column). GM for the first row and WM for the second. DARTEL template shows a wide line of CSF in zone 1 and a large pattern of the V1 in zone 2 which are thin in the data images. Our probabilistic maps capture the presence of WM in zone 3, however DARTEL only detects a small region. The shape of the V1 in zone 2 for our template fits better the data than the one for the average template. The presence of WM in zone 3 for the average template is fuzzy. Our estimated template with 8 subjects is good. With 20 subjects, we get more details on the boundary of two types.

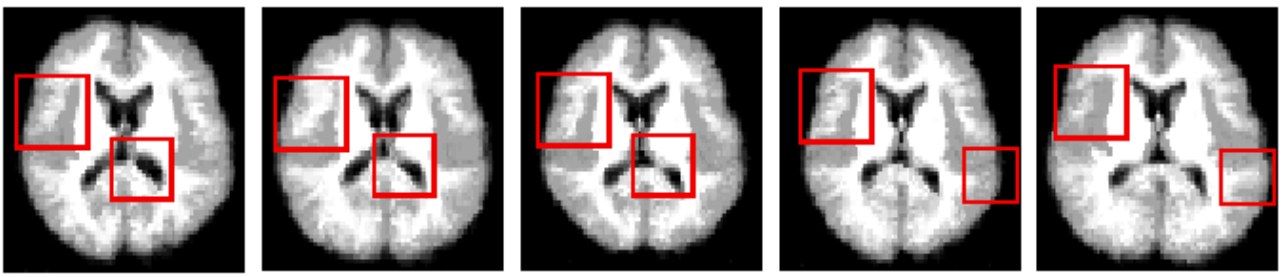

Figure 6. Five simulated images using the estimated template with 8 subjects. The deformations of the ventricles are realistic as well as the cortex foldings which look like some training ones. Moreover, the cortex thickness changes.

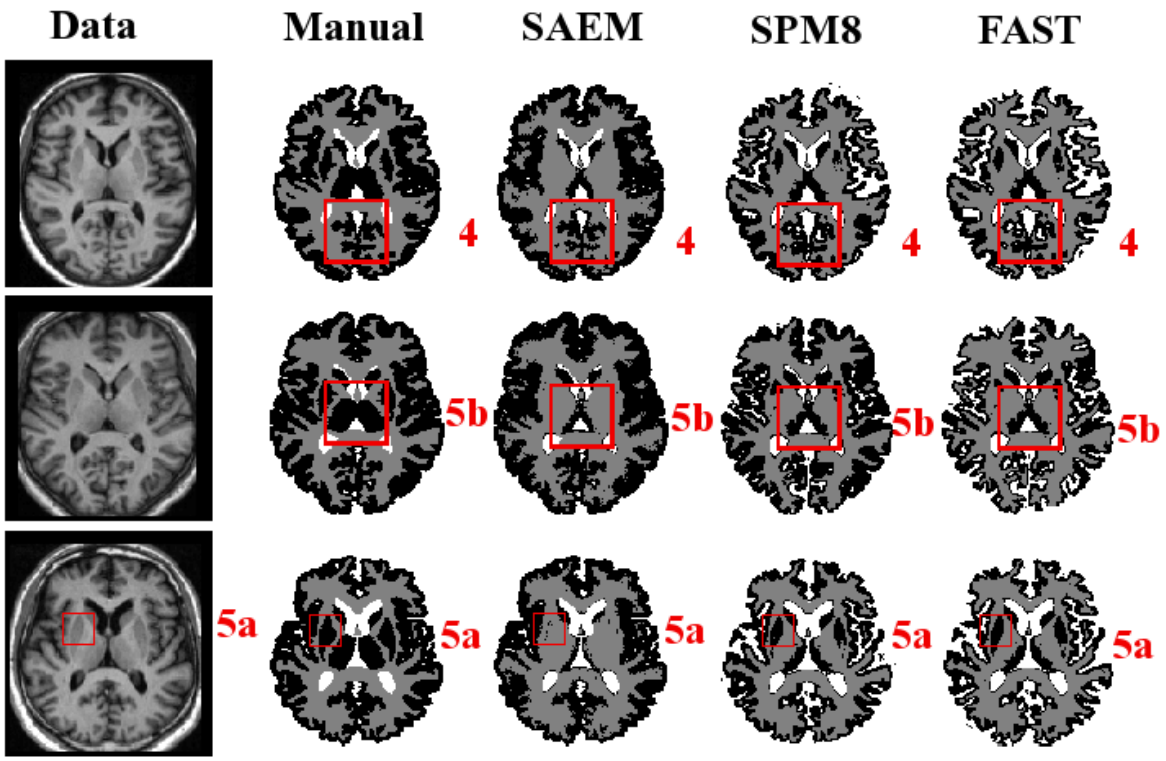

Figure 7. Experiments on real data. Each column corresponds to the one slice of 3 data images, the manual segmentation and the segmentation obtained by our method, SPM8 and FAST. Our methods shows each fold of the V1 (zone 1). Our method does not manage to segment the subcortical structures (zone 5a), others segment successfully with the strong prior (however not entirely, see zone 5b).

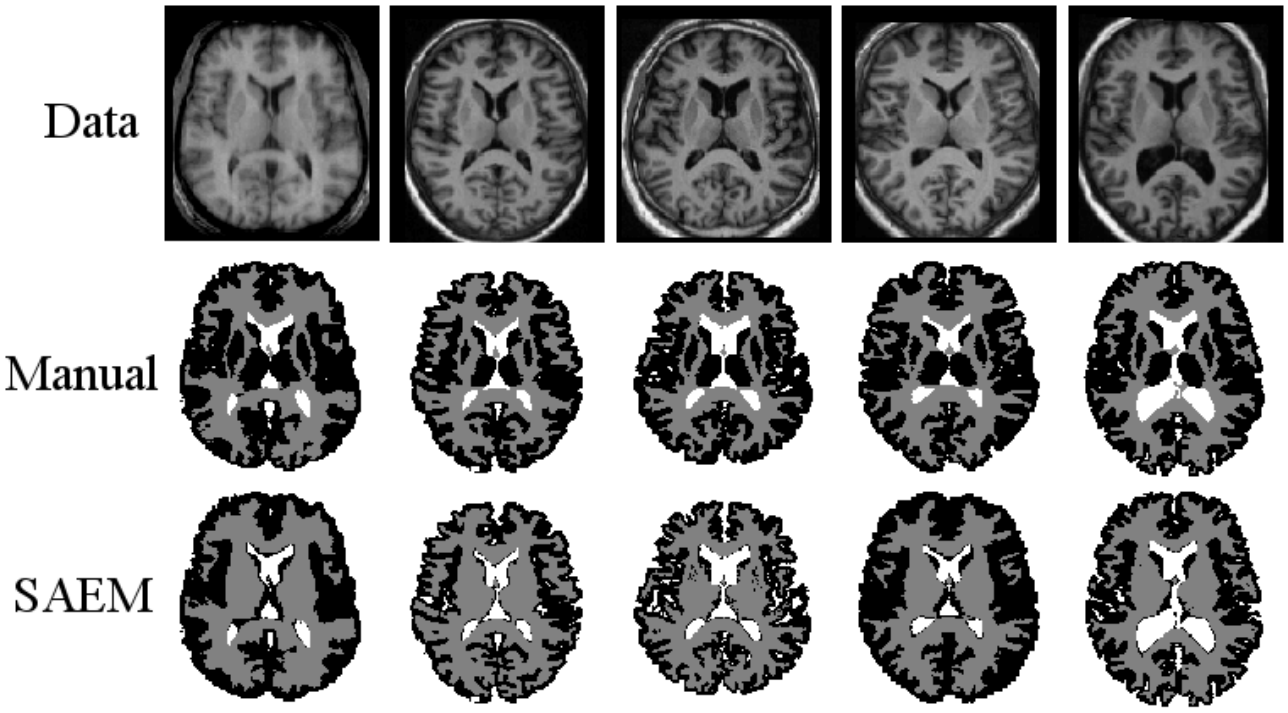

Figure 8. Segmentation for 5 new individuals using the atlas created by 20 individuals. Each row corresponds to the one slice of 5 data images, the manual segmentation and the segmentation obtained by our method.