Bayesian model selection for pathological neuroimaging data applied to white matter lesion segmentation

Carole H. Sudre*†, M. Jorge Cardoso*†, Willem Bouvy‡,
Geert J. Biessels‡, Josephine Barnes†, Sebastien Ourselin*†

*Centre for Medical Image Computing (CMIC), University College London, UK,
†Dementia Research Centre (DRC), University College London, UK,
‡Department of Neurology and Neurosurgery, UMC Utrecht, Netherlands

***Abstract*—In neuroimaging studies, pathologies can present themselves as abnormal intensity patterns. Thus, solutions for detecting abnormal intensities are currently under investigation. As each patient is unique, an unbiased and biologically plausible model of pathological data would have to be able to adapt to the subject's individual presentation. Such a model would provide the means for a better understanding of the underlying biological processes and improve one's ability to define pathologically meaningful imaging biomarkers. With this aim in mind, this work proposes a hierarchical fully unsupervised model selection framework for neuroimaging data which enables the distinction between different types of abnormal image patterns without pathological *a priori* knowledge. Its application on simulated and clinical data demonstrated the ability to detect abnormal intensity clusters, resulting in a competitive to improved behavior in white matter lesion segmentation when compared to three other freely-available automated methods.**

Index Terms: Bayesian inference criterion (BIC), brain segmentation, Gaussian mixture model (GMM), magnetic resonance imaging (MRI), split-and-merge (SM) strategy, white matter lesion (WML).

I. Introduction

[1] Measures of pathological load or proportion of abnormal *vs.* healthy tissue within the brain can be used to ascertain clinical correlations and infer disease progression in multiple sclerosis (MS), Alzheimer's disease, as well as other neurological conditions [1]. As a consequence of the direct relationship between biological properties and signal intensity in structural magnetic resonance imaging (MRI), the presence of pathology commonly leads to the observation of unusual intensity patterns in the image. Once a protocol has been developed, these patterns can be manually outlined by trained operators. However manual segmentations can suffer from inter- and intra-rater variability and tend to be very time consuming [2]. Therefore, there is a crucial need for reliable automated methods to identify and delineate pathology-related observations on MRI. Furthermore, as human delineation relies mostly on anatomical knowledge and image contrast rather than on quantitative features, inter-subject discrepancies and biases can occur due to lesion size, neighboring intensities, and/or their spatial location. Moreover, in the context of drug trials and in observational studies where serial images are obtained, the threshold of visual rating for longitudinal change is too high to observe subtle changes [3] further motivating the need for automated solutions. In the case of automated segmentation of pathological data, different strategies have been proposed, either focusing on the independent delineation of lesions or by jointly estimating healthy and pathological tissues. Since the relationships between the presence of abnormal intensities, such as lesions, and surrogate biomarkers are increasingly investigated in clinical studies, a combination of appropriate tissue delineation and lesion measurement is necessary.

Using MRI, three main anatomical components of the intracranial space can be detected: grey matter (GM), white matter (WM) and corticospinal fluid (CSF). In addition to these, voxels with unexpected intensities can be observed. These have different causes: blood vessels, pathologies of different kinds, imaging artefacts or confounding non-brain structures. From a modelling perspective, two main problems arise when dealing with these observations: first, bias is introduced in the estimation of model parameters by the presence of outlier intensities when segmenting non-pathological tissues [4], [5]; secondly, there is a need for prior knowledge in order to design better pathology-specific segmentation algorithms. Due to pathology-specific tuning and the reliance on knowledge-based heuristic rules [6], [7], pathology segmentation methods are not easily applicable to multiple pathologies [3]. Heterogeneity in the expression of the pathology of interest further complicates the analysis. For instance, pathological tissues exhibit different signal patterns according to the stage of the disease in patients with multiple sclerosis (MS) or stroke. The existence of diffuse pathology, such as the dirty appearing white matter in the case of age-related white matter hyper-intensities (WMH) might further hinder the segmentation results. The tuning of methods toward specific applications introduces heuristic constraints to avoid false positive and false negative classification [8] or to separate lesions from other outliers (such as imaging artefacts) [6]. Task-specific models also obviate situations in which different types of pathologies or lesions are present in the same subject [9], a problem which is seldom addressed [10], [11].

In the field of lesion segmentation, numerous methods have been designed to automatically delineate lesions. These can be classified as supervised or unsupervised methods. The performance of supervised methods was reported to be high [12] but the main caveat of these methods is the choice of the training set which may not include the full extent of pathology which exists in the population, thus producing a less-than-optimal representation of pathological variability [1]. Caution is further warranted: with problems such as lesion segmentation, for which inter- and intra-rater variability is high [13], clinical definition can be disputed [14]. Alternative strategies, whose strengths and weaknesses have been discussed in [9], include the use of atlases [15], the application of tissue segmentation to drive the lesion detection [16] or the direct application of empirical rules [17]. Most of the methods however make use of complementary information obtained using different structural MR acquisition sequences. For example, the T1-weighted (T1) images are known to provide a good contrast between the healthy tissues while FLuid Attenuated Inversion Recovery (FLAIR) sequences are widely used to distinguish pathologies present in the white matter. In addition, FLAIR images offer a good contrast between the CSF and the lesions but is known to suffer from acquisition artefacts such as pulsatile flow in the CSF and overestimation of the demyelination as previously reported in [1], [18], [19]. Multimodal information may be used directly in a joint manner [6], [20] or following a multi-steps scheme [21], [22], making use of specific modalities to obtain either the tissue [3], [23] or the final lesion segmentation [8]. It has been noted that, since the different acquisition sequences represent different physical behaviors, their combined use makes the definition of the lesions even more complex [21]. A spatially weighted model characterising the amount of information provided by each modality has also been proposed [24]. Finally, some methods avoid problems related to the registration of multiple acquisition sequences by using one unique acquisition sequence [25], [26].

Using a mixture of Gaussian distributions also known as Gaussian Mixture Model (GMM), is a classical and elegant way of modelling the observed intensities in MR images, where the optimization of the model parameters can be done through the Expectation-Maximization (EM) algorithm [27]. However, the EM algorithm is known to be very sensitive to initialization and to the presence of outliers: a single unexpected observation can strongly bias the parameters estimation and consequently the segmentation outcome [5], [28]. Various algorithms have been designed to provide a robust estimation of the tissue parameters in presence of outliers. These approaches consist of the down-weighting of the responsibility of the voxels considered as outliers when estimating the parameters [5], [7], [29] or the introduction of an outlier-specific class, thus reducing the bias introduced by the outliers into the parameter estimates of the normal tissue classes [5], [6].

One should note that when the signal to noise ratio (SNR) of an MR image is above 2 [3], [30], the Rician noise in magnitude MR images can be approximated as Gaussian, justifying the Gaussian assumption used in the GMM. However, GMM models which consider a single Gaussian component per tissue class have been challenged and are especially controversial in the case of the CSF, as mentioned in [23]. Even in the absence of pathology, additional components modelling the presence of partial volume effect have been added [31]. The method developed in [32] considered a fixed number of Gaussian subcomponents per tissue class. As previously reported, intensities are widely spread in the CSF, leading to an increase in the variance estimates [20]. This aspect is particularly challenging for white matter hyper-intensities observed in T2-weighted (T2) images since the variance overestimation may lead to an overlap between the CSF class and the hyper-intense lesions. A combination of non parametric methods and Gaussian models for the GM and the WM have therefore been proposed to tackle specifically the CSF problem [23]. Another way to circumvent the limitations of the GMM is to increase the number of parametric components in the model. An adaptive mixture model with up to 25 Gaussians was detailed in [33] to delineate the lesions, while models using many Gaussians have also been developed in a more local framework [24]. Lastly Freifeld *et al.* proposed to use many local Gaussians, which are then classified as one of the four labels (GM, WM, CSF or lesion) [34].

Task-specific methods are developed toward a certain target application, making them prone to a modelling bias. This bias can hinder their performance in pathologically normal subjects and in cases with multiple types of outliers. The lack of generalisation capability has been in fact put forward as a reason for the absence of a standard in clinical practice [9]. To tackle this problem, this work proposes a novel adaptive framework for data modelling in the presence of multiple types of outlier observations, named BaMoS (**Ba**yesian **Mo**del **S**election) [35]. In this framework, the data is modelled hierarchically by first dividing the model into an inlier and an outlier part. Each one of these parts is then modelled as a mixture of multiple anatomical classes, with each one of these classes modelled as a Gaussian mixture model. As the number of Gaussians necessary to characterize each tissue class is not known a priori, we propose to use the Bayesian Information Criterion (BIC) for model selection and a split-and-merge (SM) strategy for the optimization of the model complexity [36]. The EM algorithm developed in this framework applies additional improvements presented earlier in the field of medical imaging. Intensity inhomogeneities (IIH), also known as bias field, are corrected according to the method detailed in [37], anatomical spatial knowledge is introduced through probabilistic atlases that are then relaxed [38], [39], and spatial context constraints are enforced through the use of a Markov Random Field (MRF) [37], [40]. Thanks to its non-pathology-specific formulation, the proposed algorithm is then subsequently targeted toward different applications. As a further refinement of the model developed in [35], constraints are added to the Gaussian covariances. The validation of the proposed algorithm is performed on both simulated images with pathological lesions in different conditions of noise and bias field, and on clinical data in the context of both multiple sclerosis and age-related white matter hyperintensities (WMH). The purpose of this validation is threefold: first to show the improvements in lesion segmentation brought by considering more than one Gaussian per tissue in a GMM, second the generic applicability of the proposed work to different contexts, and third the ability of the proposed method to compete with other available lesion segmentation algorithms while providing a finer definition of lesion severity and potentially other types of outliers.

II. Methods

This section introduces the methodological contributions of the proposed work. Note that, due to the large number of variables and notations used in the proposed model, abbreviations and acronyms are listed in Appendix V-A while a summary table of mathematical notations is provided in Appendix V-B.

The model selection method presented in this work relies on an hierarchical mixture of Gaussian mixture models. The proposed three-tiered hierarchy is first modeled as an inlier/outlier mixture, with each component being split into four tissue types (GM, WM, CSF and Non-Brain (NB)). Each of this tissue is in turn modelled as a GMM, whose number of components is automatically determined through a split and merge strategy. Figure 1 displays an example of such a model.

After a short review of GMM (II-A1) and the mathematical description of the three-layered hierarchical model (II-A2), the use of the Expectation-Maximization algorithm is detailed II-A3. Previously published approaches to control for the intensity inhomogeneities (II-B1), morphological variability in statistical atlases (II-B2) and the addition of spatial neighboring smoothing constraints through a Markov Random Field (II-B3) are then presented in the proposed hierarchical mixture scheme. In section II-B4 we then present a Gaussian-model derived prior over the covariance of the considered Gaussian components of the model.

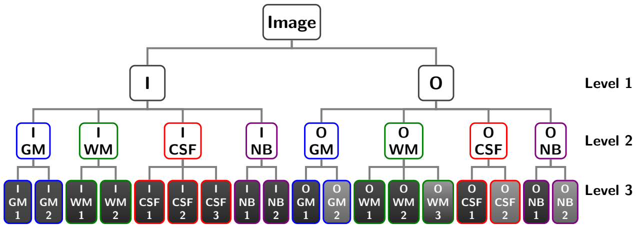

Fig. 1: Example of a possible hierarchical model, where each level is denoted by l (Level 1), lj (Level 2), and ljk (Level 3). The elements in Level 1 and 2 represent a mixture distribution, whereas the elements in Level 3 are either a Gaussian or a uniform distribution. Lighter shaded elements in Level 3 correspond to a hypothetical uniform distribution.

Afterwards, the model selection process that occurs at the third level of the hierarchy and enables the automatic selection of the number of needed components is detailed in II-C. The split and merge (SM) strategy used in this framework is expanded (II-C1) and the algorithm steps are further explained (II-C3). Lastly, the application of this model selection in the context of white matter lesion segmentation is detailed in II-D.

A. Hierarchical Gaussian mixture model and EM algorithm

1) Gaussian mixtures: In the following, $\mathscr{Y} = \{\mathbf{y}_1, \cdots, \mathbf{y}_N\}$ denotes the set of log-transformed normalized intensities indexed by n, N being the total number of observations. The feature vector $\mathbf{y}_n$ is of dimension D, representing the number of channels used for the segmentation. The log transformation of the intensities is further used to model the IIH as additive as detailed in II-B1. The log-transformed intensities are still assumed to be Gaussian-distributed and this assumption with this modelling of the bias field has been shown to yield good results [41]. In GMMs, the intensity of each voxel is considered as stemming in different mixing proportions from J Gaussian density distribution functions $\mathscr{G}\left(\mathscr{Y}\middle|\theta_j\right)$. Considering J Gaussian instances in the mixture, the set of parameters to optimize is composed of $\Theta_J = \{\theta_1, \cdots, \theta_J\}$, where $\theta_j = \left\{\boldsymbol{\mu}_j, \Lambda_j\right\}$, with $\boldsymbol{\mu}_j$ and Λ_j being respectively the mean and covariance matrix of the j^{th} component of the mixture. The vector of weights for the mixture is $\boldsymbol{\pi}$ whose j^{th} component π_j is the weight attributed to the component j under the constraint $\sum_{j=1}^{J} \pi_j = 1$ and $\forall j,\ \pi_j > 0$. For mathematical simplicity we introduce $\Xi_J = \{\Theta_J, \boldsymbol{\pi}\}$ the complete set of parameters. The parameters are optimized in order to maximize the log-likelihood of the model and consequently enable an accurate labelling of the voxels of the image.

2) Hierarchical model: Adopting a hierarchical treatment of pathological images is common in the literature. Many lesion segmentation methods operate in a stepwise fashion, by first extracting the class containing the lesions and then refining this segmentation according to some heuristic rules in order to extract the lesions ([1], [6], [20], [23]).

In this work, BaMoS follows a three-level hierarchical architecture:

1) At the first level (Level 1), indexed by l, the model is robustly separated into two density functions I and O, that correspond respectively to the inlier part (I), modelling the healthy tissues, and to the outlier part (O), related to the unexpected observations, such that

$$f\left(\mathbf{y}_n|\Xi_{\mathbf{K}}\right) = b_I \cdot I\left(\mathbf{y}_n|\Xi_{\mathbf{K}}\right) + b_O \cdot O\left(\mathbf{y}_n|\Xi_{\mathbf{K}}\right),$$

with $b_I + b_O = 1$ and $b_l \geq 0$, and introducing $\mathbf{b}$ the vector formed with these parameters. Note that $O\left(\mathbf{y}_n|\Xi_{\mathbf{K}}\right)$ is a full mixture model, contrary to previously proposed models which assumed a uniform distribution for O [42].

2) The second level (Level 2), indexed by j characterizes the anatomical tissue classes (*i.e.* if an inlier or outlier voxel belongs to WM, GM, CSF or other non-brain (NB) tissues). The number of anatomical classes J_l is considered the same for both the inlier and outlier classes since the model is built under an assumption of symmetry, simplifying $J_I = J_O = J$. The distribution is thus:

$$f\left(\mathbf{y}_n|\Xi_{\mathbf{K}}\right) = \sum_{l\in I,O} b_l \sum_{j=1}^{J} a_{l_j} \Phi\left(\mathbf{y}_n\middle|\Theta_{l_j}\right),$$

where b_l, a_{l_j} and $\Phi\left(\mathbf{y}_n\middle|\Theta_{l_j}\right)$ are respectively the mixing weight of l, the class weight for l_j and the likelihood of the data at voxel n for the tissue class l_j. At this point, $\mathbf{a}$ denotes the vector of mixing weights a_{l_j} satisfying $\sum_{j=1}^{J} a_{l_j} = 1$ and $a_{l_j} \geq 0$, $\forall l \in \{I, O\}$.

3) The third level (Level 3), indexed by k, characterizes the multiple intensity clusters of each inlier or outlier tissue class and models the acquisition noise in the observations from the expected biological mean signal. Each anatomical class density distribution is modelled by a mixture of multiple components with distribution $\mathscr{M}$, that can be Gaussian ($\mathscr{G}$) and/or uniform ($\mathscr{U}$) such that

$$\begin{aligned}\Phi\left(\mathbf{y}_n\middle|\Theta_{l_j}\right) &= \sum_{k=1}^{K_{l_j}+1} w_{l_{j_k}} \mathscr{M}\left(\mathbf{y}_n\middle|\theta_{l_{j_k}}\right)\\ &= \sum_{k=1}^{K_{l_j}} w_{l_{j_k}} \mathscr{G}\left(\mathbf{y}_n\middle|\theta_{l_{j_k}}\right) + w_{l_{j_{K_{l_j}+1}}} \mathscr{U}_{l_j}\end{aligned}$$

where K_{l_j} is the number of Gaussian components in class l_j, $w_{l_{j_k}}$ is the mixing proportion (≥ 0) of class l_{j_k} and $\theta_{l_{j_k}}$ are the corresponding Gaussian parameters. The uniform distribution in each class l_j is only parameterized by the mixing coefficient $w_{l_{j_{K_{l_j}+1}}}$. The mixing coefficients for class l_j are gathered in the vector $\mathbf{w}_{l_j}$, with $\mathbf{W}$ being the set of all such vectors that satisfy $\sum_{k=1}^{K_{l_j}+1} w_{l_{j_k}} = 1$, $\forall l \in \{I, O\}$ and $\forall j \in \{1, \cdots, J\}$.

Adopting the notation $\pi_{l_{j_k}} = b_l a_{l_j} w_{l_{j_k}}$ with $\boldsymbol{\pi} = \{\mathbf{b}, \mathbf{a}, \mathbf{W}\}$ the set of *a priori* mixing weights at the different hierarchical levels and considering the observations as independent and identically distributed (*iid*), the multi-layered mixture model can finally be expressed as follows:

$$f\left(\mathscr{Y}|\Xi_{\mathbf{K}}\right) = \prod_{i=1}^{N} \sum_{l\in I,O} \sum_{j=1}^{J} \left[\sum_{k=1}^{K_{l_j}+1} \pi_{l_{j_k}} \mathscr{M}\left(\mathbf{y}_n\middle|\theta_{l_{j_k}}\right)\right].$$

An example of a possible hierarchical model is displayed in Figure 1, where at Level 1 l takes the values Ior O and j the values in $\{$GM, WM, CSF, NB$\}$. At this point, three main types of parameters have to be optimized: first, the parameters of each Gaussian distribution $\theta_{l_{j_k}}$; secondly, the contribution $\pi_{l_{j_k}}$ of each distribution to the overall observation model; and thirdly and most importantly, the number of Gaussian components K_{l_j} necessary to describe the underlying distribution of tissue class l_j.

3) EM algorithm: The Expectation-Maximization algorithm introduced in [27] is commonly used for the optimization of $\Xi_{\mathbf{K}}$. Introducing the labelling configuration, denoted by $\mathscr{Z} = \{\mathbf{z}_1, \cdots, \mathbf{z}_N\}$, the complete data is then defined as $\mathscr{X} = \{\mathscr{Y}, \mathscr{Z}\}$. Here, $\mathbf{z}_n$, supporting label ℓ, is defined as $\mathbf{e}_\ell$ vector of the canonical basis, *i.e.* the unity vector with component ℓ equal to 1 and all the others to 0. The conditional distribution of the complete data given the parameters is expressed as:

$$\begin{aligned}f\left(\mathscr{X}|\Xi_{\mathbf{K}}\right) &= f\left(\mathscr{Y}, \mathscr{Z}|\Xi_{\mathbf{K}}\right)\\ &= f\left(\mathscr{Y}|\mathscr{Z}, \Xi_{\mathbf{K}}\right) f\left(\mathscr{Z}|\boldsymbol{\pi}\right)\\ &= \prod_{n=1}^{N} \prod_{l\in I,O} \prod_{j=1}^{J} \prod_{k=1}^{K_{l_j}+1} \left[\pi_{l_{j_k}} \mathscr{M}\left(\mathbf{y}_n\middle|\theta_{l_{j_k}}\right)\right]^{z_{nl_{j_k}}}\end{aligned}$$

The EM algorithm consists in alternating between two steps, ensuring the increase of the log-likelihood. The Expectation step or E-step consists in finding the expectation ($\mathscr{E}$) of the log of the conditional distribution of the complete data given the parameters obtained at iteration t, *i.e*:

$$\begin{aligned}\mathscr{Q}\left(\Xi_{\mathbf{K}}\middle|\Xi_{\mathbf{K}}^{(t)}\right) &= \mathscr{E}_{\Xi_{\mathbf{K}}^{(t)}}\left[\log\left(f\left(\mathscr{X}|\Xi_{\mathbf{K}}\right)\right)\right]\\ &= \mathscr{E}_{\Xi_{\mathbf{K}}^{(t)}}\left[\log \prod_{n=1}^{N} \prod_{l\in I,O} \prod_{j=1}^{J} \prod_{k=1}^{K_{l_j}+1} \left[\pi_{l_{j_k}} \mathscr{M}\left(\mathbf{y}_n\middle|\theta_{l_{j_k}}\right)\right]^{z_{nl_{j_k}}}\right]\\ &= \sum_{n=1}^{N} \sum_{l\in I,O} \sum_{j=1}^{J} \sum_{k=1}^{K_{l_j}+1} \mathscr{E}_{\Xi_{\mathbf{K}}^{(t)}}\left[z_{nl_{j_k}}\right] \log\left(\pi_{l_{j_k}} \mathscr{M}\left(\mathbf{y}_n\middle|\theta_{l_{j_k}}\right)\right)\\ &= \sum_{n=1}^{N} \sum_{l\in I,O} \sum_{j=1}^{J} \sum_{k=1}^{K_{l_j}} p_{nl_{j_k}}^{(t+1)} \left[\log\left(\mathscr{M}\left(\mathbf{y}_n\middle|\theta_{l_{j_k}}\right)\right) + \log(\pi_{l_{j_k}})\right]\end{aligned}$$

where $p_{nl_{j_k}}^{(t+1)}$, also called responsibility, is obtained by applying the Bayes' Rule as:

$$\begin{aligned}p_{nl_{j_k}}^{(t+1)} &= f(\mathbf{z}_n = \mathbf{e}_{l_{j_k}} | \mathbf{y}_n, \Xi_{\mathbf{K}}^{(t)})\\ &= \frac{\pi_{l_{j_k}}^{(t)} \cdot \mathscr{M}\left(\mathbf{y}_n\middle|\theta_{l_{j_k}}^{(t)}\right)}{\sum_{l'\in I,O} \sum_{j'=1}^{J} \sum_{k'=1}^{K_{l'_{j'}}+1} \pi_{l'_{j'_{k'}}}^{(t)} \mathscr{M}\left(\mathbf{y}_n\middle|\theta_{l'_{j'_{k'}}}^{(t)}\right)}\end{aligned}$$

The maximization step or M-step, consists in the maximization of the $\mathscr{Q}\left(\Xi_{\mathbf{K}}|\Xi_{\mathbf{K}}^{(t)}\right)$ with respect to $\Xi_{\mathbf{K}}$.

B. Variations for applications to medical imaging

1) Intensity Inhomogeneity correction (IIH): Intensity inhomogeneities, also called bias field, may appear in MR images due to spatial inhomogeneity of the scanner main magnetic field during the acquisition of the image [43]. It causes smooth variations in the intensity observed for the same tissue throughout the image and might lead to misclassifications if left uncorrected. When pathological tissues are present, this correction is of special interest and often performed as a preprocessing step. In this work, the estimation of the bias field follows the model detailed in [37]. The bias field, multiplicative in the original MRI observation space, is modelled as a linear combination of M polynomial basis functions χ_m given the spatial location pos_n, $\sum_{m=1}^{M} \boldsymbol{c}_m \chi_m(\text{pos}_n)$. When considering the log-intensities $\mathscr{Y}$, the bias field becomes additive and the parameters $\mathbf{c}_m$, vectors of size D, can then be progressively optimized within the EM framework. In the following, $\mathbf{y}_n^{\text{c}\,(t)}$ denotes the corrected intensity feature at voxel n at iteration t with:

$$\mathbf{y}_n^{\text{c}\,(t)} = \mathbf{y}_n - \sum_{m=1}^{M} \boldsymbol{c}_m^{(t)} \chi_m(\text{pos}_n)$$

Denoting $\boldsymbol{C} = \{\mathbf{c}_1, \cdots, \mathbf{c}_M\}$ the set of vectors of linear coefficients used to model the magnetic field inhomogeneity, the set of parameters to optimize becomes $\Xi_{\mathbf{K}} = \{\Theta_{\mathbf{K}}, \boldsymbol{\pi}, \mathbf{C}\}$

2) EM algorithm with spatial adaptive a priori knowledge: The classical mixture model presented above has been modified to consider spatially varying mixing coefficients, where the parameters π_j become spatially variant [44]. As previously reported [39], this model cannot distinguish structures with similar intensities from one another. For instance, very hyper-intense lesions in the white matter cannot be distinguished from flow artefacts occurring in the CSF. To overcome this problem, a popular solution has been to adopt *a priori* knowledge stemming from statistical atlases [45]. Adopting such form of *a priori* knowledge at Levels 1 and 2 of the hierarchical model, $\boldsymbol{\pi}$ denotes now $\{\mathbf{B}, \mathbf{A}, \mathbf{W}\}$, where $\mathbf{B}$ (resp. $\mathbf{A}$) refers to the set of N vectors of *a priori* mixing probabilities $\mathbf{b}_n$ (resp. $\mathbf{a}_n$). The conditional distribution becomes:

$$f(\mathscr{X}|\Xi_{\mathbf{K}}) = \prod_{n=1}^{N} \prod_{l \in I,O} \prod_{j=1}^{J} \prod_{k=1}^{K_{l_j}+1} \left[\pi_{nl_{j_k}} \mathscr{M}\left(\mathbf{y}_n, \theta_{l_{j_k}}\right)\right]^{z_{nl_{j_k}}}$$

with $\pi_{nl_{j_k}} = b_{nl} a_{nl_j} w_{l_{j_k}}$. Considering a symmetric model at Level 2, the atlases are defined so that $\forall n$, $a_{nI_j} = a_{nO_j} = a_{nj}$. The responsibilities are now defined as:

$$p_{nl_{j_k}}^{(t+1)} = \frac{b_{nl} a_{nj} w_{l_{j_k}}^{(t)} \mathscr{M}\left(\mathbf{y}_n \middle| \boldsymbol{\theta}_{l_{j_k}}^{(t)}\right)}{\sum_{l' \in I,O} \sum_{j'=1}^{J} \sum_{k'=1}^{K_{l'_{j'}}+1} b_{nl'} a_{nj'} w_{l'_{j'_{k'}}}^{(t)} \mathscr{M}\left(\mathbf{y}_n \middle| \boldsymbol{\theta}_{l'_{j'_{k'}}}^{(t)}\right)}$$

This solution is subject to some difficulties related to both the choice of an appropriate population to build the statistical atlases and the choice of a suitable coordinate mapping [25]. This problem is especially important when pathological features are not present in the statistical atlases. For example, the study of white matter hyper-intensities (WMH) in elderly subjects with enlarged ventricles can be biased by the use of atlases that originate from a population of young healthy volunteers [3]. A framework for statistical atlas adaptation has been described in [38], [39]. This adaptation enables the handling of pathological morphologies while still preserving some prior spatial information constraints. This model assumes that the spatially varying mixing priors are derived from a Dirichlet distribution, noted $\mathscr{D}$. Now denoting $\tilde{\boldsymbol{\pi}} = \{\tilde{\mathbf{B}}, \tilde{\mathbf{A}}, \mathbf{W}\}$, the mixing coefficients follow the distribution:

$$\begin{aligned} f(\tilde{\boldsymbol{\pi}}) &= f\left(\tilde{\mathbf{B}}\right) f\left(\tilde{\mathbf{A}}\right) \\ &= \prod_{n=1}^{N} \mathscr{D}\left(\tilde{\mathbf{b}}_n, \boldsymbol{\beta}_n\right) \mathscr{D}\left(\tilde{\mathbf{a}}_n, \boldsymbol{\alpha}_n\right) \\ &= \prod_{n=1}^{N} \frac{\prod_{l \in I,O} b_{nl}^{(\beta_{nl}-1)}}{\mathscr{B}(\boldsymbol{\beta}_n)} \frac{\prod_{j=1}^{J} a_{nj}^{(\alpha_{nj}-1)}}{\mathscr{B}(\boldsymbol{\alpha}_n)} \end{aligned}$$

where $\mathscr{B}$ is a Beta function and $\boldsymbol{\beta}_n$, $\boldsymbol{\alpha}_n$ the vectors of Dirichlet prior parameters for voxel n such that $\beta_{nl} = 1 + \delta_1 b_{nl}$ and $\alpha_{nj} = 1 + \delta_2 a_{nj}$, where δ_1 and δ_2 are positive parameters assessing the strength of the relaxation applied at Level 1 and Level 2 of the hierarchy. Note that the choice for an asymmetric modelling in the inlier and outlier part would prevent the decoupling in the adaptation of the different levels. The E-step is kept unchanged but the M-step consists now in the optimization of:

$$f(\mathscr{X}|\Theta_{\mathbf{K}}, \tilde{\boldsymbol{\pi}}, \mathbf{C}) \cdot f(\Theta_{\mathbf{K}}, \tilde{\boldsymbol{\pi}}, \mathbf{C})$$

with respect to the parameters. Using the Lagrange multipliers method as detailed in [39] in order to enforce the constraint that the mixing coefficients must sum to 1, the M-step leads to an update of the mixing coefficients which related to the probabilistic atlases, such that

$$\tilde{b}_{nl}^{(t+1)} = \frac{\delta_1 b_{nl} + p_{nl}^{(t+1)}}{\delta_1 + 1} \qquad \tilde{a}_{nl}^{(t+1)} = \frac{\delta_2 a_{nj} + p_{nj}^{(t+1)}}{\delta_2 + 1}$$

with

$$p_{nl}^{(t+1)} = \sum_{j=1}^{J} \sum_{k=1}^{K_{l_j}+1} p_{nl_{j_k}}^{(t+1)} \qquad p_{nj}^{(t+1)} = \sum_{l \in I,O} \sum_{k=1}^{K_{l_j}+1} p_{nl_{j_k}}^{(t+1)}.$$

Since the probabilistic atlases are supposed to be smooth but the responsibilities are not, a Gaussian Kernel G_σ with standard deviation σ is convolved with the responsibilities as a form of spatial regularization, similarly to previously described methods [38], [39]. Eventually, the update for the *a priori* mixing coefficients is:

$$\begin{aligned} \tilde{b}_{nl}^{(t+1)} &= (1-\kappa_1) b_{nl} + \kappa_1 (G_\sigma \star p_{nl}^{(t+1)}) \\ \tilde{a}_{nj}^{(t+1)} &= (1-\kappa_2) a_{nj} + \kappa_2 (G_\sigma \star p_{nj}^{(t+1)}) \end{aligned}$$

where $\kappa_i = 1/\delta_i + 1$ and $\star$ represents the convolution operator. The smoothing mentioned here is related only to the relaxation of the statistical atlases and does not concern the posterior that is used to produce the final segmentation and optimize most model parameters. This atlas relaxation contributes greatly towards the detection of outliers in the proposed model. One can loosely see this step as placing probabilistic seeds on outlier regions as has been done in [22].

3) EM algorithm with neighborhood consistency constraints: Spatial constraints are not only added on a global scale with probabilistic atlases but can be also reinforced at a more local scale. Neighborhood context constraints are introduced here to improve the consistency between neighboring voxels. By introducing a Markov Random Field, one can promote the propensity of neighboring voxels to be classified under the same class. Usually, one adopts a Potts model with a unique value expressing the energy needed for two voxels to be classified under different labels. However, additional anatomical information can be added if the neighborhood relationships between tissue classes are known [38]. This leads to the introduction of a symmetric matrix H of parameters, stating the energy relationships between neighboring tissues. Adding these parameters into the model results in the following expression for the prior over the labelling, where $\mathscr{N}_n$ represents the set of von Neumann neighbors of voxel n:

$$\begin{aligned} f\left(\mathscr{Z}|H,\boldsymbol{\pi}\right) &\propto e^{\left(-U_{\text{MRF}}(\mathscr{Z}|H)\right)} \prod_{n=1}^{N} \prod_{l\in I,O} \prod_{j=1}^{J} \prod_{k=1}^{K_{l_j}+1} \pi_{nl_{j_k}}^{z_{nl_{j_k}}} \\ U_{\text{MRF}}(\mathscr{Z}|H) = & \quad \sum_{n=1}^{N} U_{\text{MRF}}\left(\mathbf{z}_n|\mathbf{z}_{\mathscr{N}_n},H\right) \\ U_{\text{MRF}}\left(\mathbf{z}_n|\mathbf{z}_{\mathscr{N}_n},H\right) = & \quad \sum_{l\in\mathscr{N}_n} \mathbf{z}_n^T H \mathbf{z}_l \end{aligned} .$$

In case of anisotropy in the resolution of the data, scaling factors might be used on the matrix H to take into account the distance between voxels, increasing the constraint with the increase in voxel proximity [31]. Adding an MRF spatial context constraint results in the invalidity of the independence assumption, making the calculation of the E-step intractable. The vector $\mathbf{p}_n$ is built from the set of responsibilities associated to each component l_{j_k} at voxel n and the set of such vectors at locations $\mathscr{N}_n$ is denoted $\mathbf{p}_{\mathscr{N}_n}$. The mean field approximation described in [40] is used here, such that

$$\begin{aligned} & f\left(\mathbf{z}_n = \mathbf{e}_{l_{j_k}} \middle| \mathbf{z}_{\mathscr{N}_n}^{(t)}, H, \boldsymbol{\pi}^{(t)}\right) \\ & \qquad \approx f\left(\mathbf{z}_n = \mathbf{e}_{l_{j_k}} \middle| \mathscr{E}_{\boldsymbol{\Xi}_{\mathbf{K}}^{(t)}}\left[\mathbf{z}_{\mathscr{N}_n}\right], H, \boldsymbol{\pi}^{(t)}\right) \\ & \qquad = f\left(\mathbf{z}_n = \mathbf{e}_{l_{j_k}} \middle| \mathbf{p}_{\mathscr{N}_n}^{(t)}, H, \boldsymbol{\pi}^{(t)}\right) \\ & \qquad = \frac{\pi_{nl_{j_k}}^{(t)} e^{\left(-U_{\text{MRF}}\left(\mathbf{e}_{l_{j_k}} \middle| \mathbf{p}_{\mathscr{N}_n}^{(t)}, H\right)\right)}}{\sum_{l'\in I,O} \sum_{j'=1}^{J} \sum_{k'}^{K_{l'_{j'}}+1} \pi_{nl'_{j'_{k'}}}^{(t)} e^{\left(-U_{\text{MRF}}\left(\mathbf{e}_{l'_{j'_{k'}}} \middle| \mathbf{p}_{\mathscr{N}_n}^{(t)}, H\right)\right)}} \end{aligned}$$

leading to the following responsibilities

$$p_{nl_{j_k}}^{(t+1)} \approx \frac{\phi_{nl_{j_k}}^{(t)} \pi_{nl_{j_k}}^{(t)} \psi_{nl_{j_k}}^{(t)}}{\sum_{l'\in I,O} \sum_{j'=1}^{J} \sum_{k'=1}^{K_{l_j}+1} \phi_{nl'_{j'_{k'}}}^{(t)} \pi_{nl'_{j'_{k'}}}^{(t)} \psi_{nl'_{j'_{k'}}}^{(t)}}$$

where we adopt the notations

$$\begin{aligned} \phi_{nl_{j_k}}^{(t)} &= f\left(\mathbf{y}_n|\mathbf{z}_n = \mathbf{e}_{l_{j_k}}, \boldsymbol{\Xi}_{\mathbf{K}^{(t)}}\right) \\ \psi_{nl_{j_k}}^{(t)} &= \exp\left(-U_{\text{MRF}}\left(\mathbf{e}_{l_{j_k}}|\mathbf{p}_{\mathscr{N}_n}^{(t)}, H\right)\right) \end{aligned}$$

4) Constraint over the covariance matrix: If the same noise model is used over all the observations, it makes sense to consider that if the covariance matrix of the Gaussian parameters is only describing the acquisition noise, then all covariances should be related. A way of constraining the noise covariance is to introduce a prior distribution over these matrices. Here we choose the Inverse Wishart distribution [46] as a prior form. Thus, the *a priori* distribution of the covariances is expressed as

$$f_{\Psi}\left(\Lambda_{l_{j_k}}\right) \propto \frac{|\Psi|^{\frac{N}{2}}}{|\Lambda_{l_{j_k}}|^{\frac{N+D+1}{2}}} \exp\left[-\frac{1}{2}\text{Tr}\left(\Lambda_{l_{j_k}}^{-1} \Upsilon_{l_{j_k}} \Psi \Upsilon_{l_{j_k}}\right)\right]$$

where Tr is the trace of the matrix, Ψ is a positive definite matrix and $\Upsilon_{l_{j_k}}$ is a scaling diagonal matrix, such that $\Upsilon_{l_{j_k}}(d,d) = 1/\exp(\mu_{l_{j_k}}^{(d)})$ is taking into account the relation between the mean and the covariance matrix under log-transformed image observations. Within the Maximum a Posteriori EM (MAP-EM) algorithm framework, the constraint Ψ is optimized as a model parameter and the update of the covariance matrix is modified into

$$\Psi^{(t+1)\,-1} = \frac{\sum_{l\in I,O} \sum_{j=1}^{J} \sum_{k=1}^{K_j} \Upsilon_{l_{j_k}}^{(t+1)} \Omega_{l_{j_k}}^{(t+1)\,-1} \Upsilon_{l_{j_k}}^{(t+1)}}{N \cdot \sum_{l\in I,O} \sum_{j=1}^{J} \sum_{k=1}^{K_{l_j}} 1}$$

$$\Lambda_{l_{j_k}}^{(t+1)} = \frac{\Omega_{l_{j_k}}^{(t+1)} + \dfrac{\Upsilon_{l_{j_k}} \Psi^{(t+1)} \Upsilon_{l_{j_k}}}{\sum_n p_{nl_{j_k}}^{(t+1)}}}{1 + \dfrac{N+d+1}{\sum_n p_{nl_{j_k}}^{(t+1)}}}$$

where $\Omega_{l_{j_k}}^{(t+1)}$ is the weighted covariance matrix at iteration $(t+1)$.

5) Summary of the modified EM: Using the adaptation of the population atlases described in II-B2, $\tilde{\boldsymbol{\pi}} = \left\{\tilde{\mathbf{B}}, \tilde{\mathbf{A}}, \mathbf{W}\right\}$ represents the sets of adapted atlases ($\tilde{\mathbf{B}}$ and $\tilde{\mathbf{A}}$) for the first and the second level and the set of weights ($\mathbf{W}$) attributed to the classical mixtures in Level 3. The conditional definition of the labelling $f\left(\mathscr{Z}|H, \tilde{\boldsymbol{\pi}}\right)$ is then proportional to

$$\prod_{n=1}^{N} \prod_{l \in I,O} \prod_{j=1}^{J} \prod_{k=1}^{K_{l_j}+1} \left[\tilde{b}_{nl} \tilde{a}_{nl_j} w_{l_{j_k}}\right]^{z_{nl_{j_k}}} \psi_{nl_{j_k}}$$

The advantage of using a matrix of neighborhood parameters for the spatial context constraint (through the MRF) appears here, where the rules can lower the energy needed for subclasses of the same tissue j to be neighbors.

Within this framework, the E-step of the EM contributes to the update of the responsibilities so that

$$p_{nl_{j_k}}^{(t+1)} = \frac{\phi_{nl_{j_k}}^{\mathrm{c}(t)} \tilde{b}_{nl}^{(t)} \tilde{a}_{n_j}^{(t)} w_{l_{j_k}}^{(t)} \psi_{nl_{j_k}}^{(t)}}{\sum\limits_{l' \in I,O} \sum\limits_{j'=1}^{J} \sum\limits_{k'=1}^{K_{j'}} \phi_{nl'_{j'_{k'}}}^{\mathrm{c}(t)} \tilde{b}_{nl'}^{(t)} \tilde{a}_{nl'_{j'}}^{(t)} w_{l'_{j'_{k'}}}^{(t)} \psi_{nl'_{j'_{k'}}}^{(t)}}$$

The M-step of the described MAP-EM with the prior over the covariance matrices (Section II-B4), IIH correction (Section II-B1) and atlas adaptation (Section II-B2) enables the update of the Gaussian parameters and the weights for the third level mixture of Gaussian and uniform components such that

$$w_{l_{j_k}}^{(t)} = \frac{\sum\limits_{n=1}^{N} p_{nl_{j_k}}^{(t)}}{\sum\limits_{k'=1}^{K_{l_j}+1} \sum\limits_{n=1}^{N} p_{nl_{j_{k'}}}^{(t)}} \qquad \boldsymbol{\mu}_{l_{j_k}}^{(t)} = \frac{\sum\limits_{n=1}^{N} p_{nl_{j_k}}^{(t)} \mathbf{y}_n^{\mathrm{c}(t)}}{\sum\limits_{n=1}^{N} p_{nl_{j_k}}^{(t)}}$$

$$\Omega_{l_{j_k}}^{(t)} = \frac{\sum\limits_{n=1}^{N} p_{nl_{j_k}}^{(t)} \left(\mathbf{y}_n^{\mathrm{c}(t)} - \boldsymbol{\mu}_{l_{j_k}}^{(t)}\right) \left(\mathbf{y}_n^{\mathrm{c}(t)} - \boldsymbol{\mu}_{l_{j_k}}^{(t)}\right)^T}{\sum\limits_{n=1}^{N} p_{nl_{j_k}}^{(t)}}$$

where $\mathbf{y}_n^{\mathrm{c}}$ stands for the IIH-corrected intensities and $\Omega_{l_{j_k}}^{(t)}$ is the weighted covariance matrix used in the update of the Gaussian covariance $\Lambda_{l_{j_k}}^{(t)}$.

C. Model selection

1) Split and merge strategy: The flexibility of the proposed model lies in the automatic selection of the appropriate number of components K_{l_j} needed to model each mixture l_j. Ideally, this parameter could be optimized using a Markov Chain Monte Carlo algorithm. In fact, due to the computational complexity of such an approach, here, a split-and-merge (SM) strategy is used for model optimization [47], [48]. The SM operations were introduced in order to deal with the initialization problem of the EM algorithm by redistributing the Gaussian components over the observation space. However, incorrect initialization may lead to the convergence of the log-likelihood toward a local maximum [28], resulting in errors in the final segmentation.

In BaMoS, the number of Gaussian components per tissue class enables the modelling of complex biological phenomena, further justifying the prior constraint over the covariance matrices introduced in Section II-B4. In this SM approach, a merge operation consists of transforming two Gaussian distributions, $l_{j_{k_1}}$ and $l_{j_{k_2}}$, into a single Gaussian distribution l_{j_k}. A split operation is the transformation of a single distribution, Gaussian or uniform, into two subcomponents. The symmetric Kullback-Leibler Divergence (KLD) is used to define which component(s) should be modified in the model. Considering two probabilistic distributions P_1 and P_2 over the variable y, the discretized symmetric KLD is expressed as follows:

$$\mathrm{KLD}(P_1 \parallel P_2) = \sum_{y} P_1(y) \log\left(\frac{P_1(y)}{P_2(y)}\right) + P_2(y) \log\left(\frac{P_2(y)}{P_1(y)}\right)$$

When looking for which component to split, the most likely candidate is the one whose model distribution explains the corresponding observations most poorly, *i.e* the one whose KLD compared to the underlying observations is the largest. When merging two components of the model, the selected pair is the one that produces the smallest KLD between the two model distributions. The initialization of the newly formed component(s) follows the strategy previously described in [49], such that when merging components $l_{j_{k_1}}$ and $l_{j_{k_2}}$ into a single component l_{j_k}, the initial Gaussian parameters are expressed as:

$$\begin{aligned}
w_{l_{j_k}} &= w_{l_{j_{k_1}}} + w_{l_{j_{k_2}}} \\
\boldsymbol{\mu}_{l_{j_k}} &= \frac{w_{l_{j_{k_1}}} \boldsymbol{\mu}_{l_{j_{k_1}}} + w_{l_{j_{k_2}}} \boldsymbol{\mu}_{l_{j_{k_2}}}}{w_{l_{j_{k_1}}} + w_{l_{j_{k_2}}}} \\
\Lambda_{l_{j_k}} &= \frac{w_{l_{j_{k_1}}} \tilde{\Lambda}_{l_{j_{k_1}}} + w_{l_{j_{k_2}}} \tilde{\Lambda}_{l_{j_{k_2}}}}{w_{l_{j_{k_1}}} + w_{l_{j_{k_2}}}} \\
\tilde{\Lambda}_{l_{j_{k_i}}} &= \Lambda_{l_{j_{k_i}}} + \left(\boldsymbol{\mu}_{l_{j_{k_i}}} - \boldsymbol{\mu}_{l_{j_k}}\right) \left(\boldsymbol{\mu}_{l_{j_{k_i}}} - \boldsymbol{\mu}_{l_{j_k}}\right)^T
\end{aligned}$$

When splitting a Gaussian component l_{j_k} into two components $l_{j_{k_1}}$ and $l_{j_{k_2}}$, we adopt the implementation described by Li *et al.* [48], which follows Richardson *et al.* [49]. Introducing $\mathbf{v}_{l_{j_k}}$ as the vector corresponding to the highest eigenvalue from the orthogonal factorization of the covariance matrix $\Lambda_{l_{j_k}}$ and setting the free parameters to 0.5, the initial parameters for the new Gaussian components are:

$$w_{l_{j_{k_1}}} = w_{l_{j_{k_2}}} = 0.5 \cdot w_{l_{j_k}}$$

$$\boldsymbol{\mu}_{l_{j_{k_1}}} = \boldsymbol{\mu}_{l_{j_k}} - 0.5 \cdot \mathbf{v}_{l_{j_k}} \qquad \boldsymbol{\mu}_{l_{j_{k_2}}} = \boldsymbol{\mu}_{l_{j_k}} + 0.5 \cdot \mathbf{v}_{l_{j_k}}$$

$$\Lambda_{l_{j_{k_1}}} = \Lambda_{l_{j_{k_2}}} = \Lambda_{l_{j_k}} - 0.25 \cdot \mathbf{v}_{l_{j_k}} \mathbf{v}_{l_{j_k}}{}^T$$

With the aim of modelling the observed data as Gaussian distributions, a uniform distribution is split into one Gaussian component modelling a sub-cluster of the data under the

original uniform distribution, and one new uniform distribution modelling the rest of the data. Splitting a uniform into a Gaussian plus a remaining uniform ensures that unmodeled outliers can still be accurately captured by the probabilistic model, thus maintaining the stability of the model. As no closed-form solution exists, a 2-class k-means algorithm is used to estimate the 2 main sub-clusters of the samples under $\mathscr{U}_{l_j}$. The mean and covariance of the cluster with the smallest variance are used to initialize the new Gaussian class.
Every time a new model is initialized, the EM optimization described above is run until convergence.

2) Penalty function - Acceptance criterion: The proposed model is optimized using an iterative conditional modes (ICM) approach, where it switches between the optimization of the model parameters and the model selection. In order to provide a bias-variance trade-off between accuracy and complexity of the model, the Bayesian Information Criterion (BIC), testing the current model $\mathbf{K}$ and expressed by

$$\text{BIC}(\mathbf{K}) = \upsilon \log\left(f\left(\mathscr{Y}|\Xi_{\mathbf{K}}\right)\right) - \mathscr{P}(\mathbf{K}),$$

is here used as an objective function. The BIC penalizes the log-likelihood of the model according to the penalization function $\mathscr{P}(\mathbf{K}) = \left[\sum_l \sum_j K_{l_j}\left(\frac{(D+1)D}{2}+1\right) - J\right]\cdot \log(N\cdot\upsilon)$, which depends on the number of free parameters being optimized. The decimation factor υ is introduced into the BIC cost function to compensate for the lack of spatial independence between the observed samples, *i.e.* υ represents the proportion of voxels that can be considered as independent [50]. The decimation factor is defined as:

$$\upsilon = \prod_{r\in(x,y,z)} \frac{0.9394}{\text{FWHM}_r} \text{ with } \text{FWHM}_r^2 = \frac{-2\log 2}{\log\left(\text{corr}_r\right)}$$

where corr_r is the correlation between adjacent voxels in the r-direction.

3) Iterative model selection scheme: For each ICM iteration, the model evolves given the most probable model. If the selected model fails to increase the objective function $\text{BIC}(\mathbf{K})$ after convergence of the EM, the next most probable model is tested. The SM search stops when all possible SM models have been tested. Given a model and its parameters Ξ_K, the model selection process is performed in an iterative five steps sheme as summarized in Figure2:

Step 1 Computation of the list of possible SM operations List_{SM}: given the current model, an ordered list of possible operations is defined by an alternating sequence of split and merge operations. Merge operations are ordered by increasing KLD and split operations by decreasing KLD as detailed in Section II-C. As a hard constraint, and mostly for computational reasons, Gaussian components with a relative weight to class l_j inferior to 0.01 are not allowed to split. Merging can only occur between Gaussian components from the same mixture l_j.

Step 2 Initialization of a new model: the first element of List_{SM} is used to define the transformation on the current model. The parameters for the changed components are initialized as detailed in Section II-C1. The matrix H, used to define the MRF neighborhood rules, is updated.

Step 3 Optimization of the parameters of the tested model with the EM algorithm.

Step 4 Test of the new model using the BIC: the new model is accepted only if the relative change in the objective function is above 10^{-4}. In order to avoid instability in the model, components with a relative weight below 0.01 are removed from the model.

Step 5 Check for evolution: if the model is accepted, the process restarts at Step 1. If not, the first element of List_{SM} is removed from the list and the process restarts at Step 2, thus testing the next operation. If all the elements in the list have been tested, the algorithm terminates.

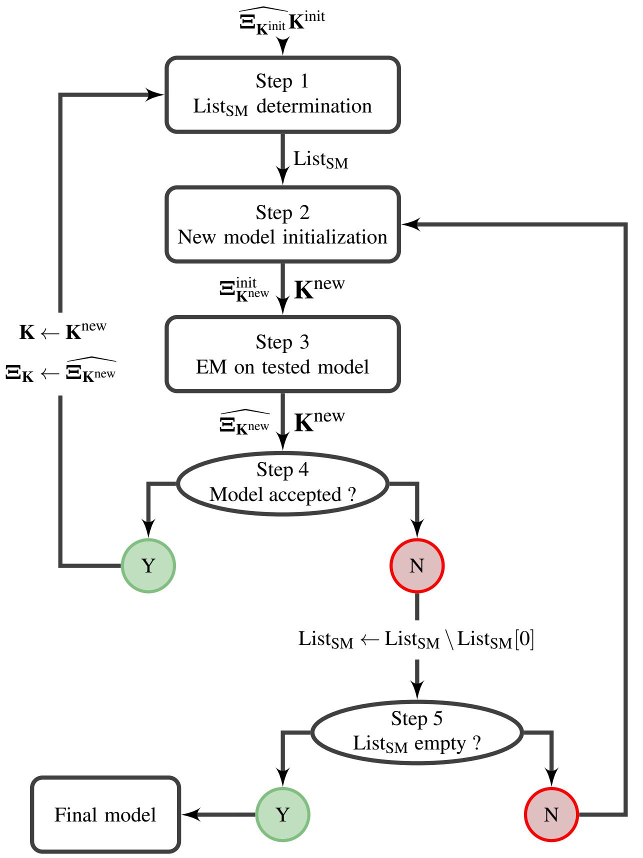
$\widehat{\Xi_{\mathbf{K}^{\text{init}}}}\mathbf{K}^{\text{init}}$
Step 1
List_{SM} determination
List_{SM}
Step 2
New model initialization
$\Xi^{\text{init}}_{\mathbf{K}^{\text{new}}}$
$\mathbf{K}^{\text{new}}$
Step 3
EM on tested model
$\widehat{\Xi_{\mathbf{K}^{\text{new}}}}$
$\mathbf{K}^{\text{new}}$
$\mathbf{K} \leftarrow \mathbf{K}^{\text{new}}$
$\Xi_{\mathbf{K}} \leftarrow \widehat{\Xi_{\mathbf{K}^{\text{new}}}}$
Step 4
Model accepted ?
Y
N
$\text{List}_{\text{SM}} \leftarrow \text{List}_{\text{SM}} \setminus \text{List}_{\text{SM}}[0]$
Step 5
List_{SM} empty ?
Final model
Y
N

Fig. 2: Graphic scheme of the model selection process performed in BaMoS.

D. Application to white matter lesion segmentation

Once BaMoS has converged to a final model, the obtained hierarchical model can be further used to gather components that share a similar biological meaning. For instance the resulting model can be analysed to gather the components related to the lesion and after a minor false positive correction produce a delineation of the white matter lesions that can be observed clinically in multiple sclerosis patients or in elderly.

1) Components selection: To obtain the lesion segmentation, the meaningful mixture components must be extracted from the model. As the lesions segmented with BaMoS can belong to more than one of the model's clusters, the relevant components of the final model were automatically selected and combined by adding the corresponding probability maps inversely weighted with their Mahalanobis distance towards the mean of the grey matter if this distance was below 3. The model components relevant for lesion segmentation were selected from the outlier part of the final model. The following heuristic rules were used to select the pathology relevant modalities:

$$l_{j_k} \in L \text{ if} \begin{cases} l_j = O_{\mathrm{WM}} \\ \text{or} \\ l_j = O_{\mathrm{GM}} \end{cases} \text{and } \mu_{l_{j_k}}^{(Patho)} > \mu_{I_{\mathrm{WM}}}^{(Patho)} ,$$

with $Patho = \{\text{FLAIR, T2, PD}\}$ where PD refers to proton density weighted images. The inclusion of the GM outlier is necessary due to the smoothness of the atlases used in this work and the presence of lesions near the cortical sheet as well as close to the deep gray matter.

When the FLAIR modality is not available, a further refinement on the selected components is required in order to avoid the inclusion of voxels related to partial volume effect at the GM-CSF border. Such components present themselves as very hypo-intense on the T1 modality and slightly but not strikingly hyper-intense on the T2 image. Since the change in intensities is monotonic with lesion severity for both T1 and T2 intensities (respectively decreasing and increasing) but with a much stronger slope for the T2 compared to the T1 [51], [52], it can be assumed that very low intensities on the T1, would only correspond to a very severe lesion and thus an even higher hyper-intensity level on the T2 image. To address this correspondence, in cases where a potential lesion-related component with a very hypo intense mean on T1 (*i.e.* a mean lower that the mean observed for the GM inliers), the corresponding mean on the T2 was checked to be hyper-intense compared to the mean of undisputed lesion components. Such components must present a hyper-intense mean on T2 (lesion-like) and a mean on T1 higher than the mean of the GM inliers.

Mathematically we have the set L of components initially considered as related to the lesions split into two groups: TL, standing for true lesion (or undisputed lesion), and DL, standing for disputed lesion such that

$$\begin{aligned} DL =& \left\{ s \in L \;\middle|\; \mu_s^{(\mathrm{T1})} < \mu_{I_{GM}}^{(\mathrm{T1})} \right\} \\ TL =& L \setminus DL \end{aligned}$$

The refined set of lesion RL is then defined as:

$$RL = TL + \left\{ s \in DL \;\middle|\; \mu_s^{(T2)} \geq \mu_{TL}^{(T2)} \right\}$$

where μ_{TL} denotes the mean intensity of the TL set.

It must be emphasized that this refinement is not needed and would be erroneous in the case where the FLAIR modality is available. Indeed, at the CSF-GM interface, intensities do not appear hyper-intense on the FLAIR image. Furthermore the monotonic evolution of intensities with lesion severity does not hold in the case of the FLAIR modality due to the inversion recovery process of the acquisition [52]. A similar selection process is performed voxelwise for the remaining uniform distributions. The main strength of such a selection process relies on its post-processing characteristic. As it occurs after obtaining the final model, the complete model is independent of the definition of the outliers of interest. Furthermore, this post-processing step remains flexible and adaptable to different acquisition protocols and subtleties in the clinical definition of the pathology [53].

2) Correction for false positives: In the final lesion segmentation, a constraint on the minimum lesion size of 3 voxels was chosen to define a lesion. The inclusion of GM lesions led to false positives in the cortical sheet, septum pellucidum and in the outer border of the brain as well as the shine through artefacts of the third and fourth ventricles. To correct for those errors, the lesion connected elements are defined and neighboring rules based on the surrounding segmentation results are used. To correct for the inclusion of elements at the outer border of the brain, distance to the NB class and to the mask border is for instance the main indicator. For the other types of FP correction, the elements considered for correction were those for which the proportion of neighboring voxels belonging to WM was lower than the GM neighboring proportion or presented very mixed proportion between GM, CSF and WM neighbors. Among those, the elements that were not in the potential deep grey matter area as defined by the ICBM statistical atlas [54] were discarded and considered as belonging to the cortical sheet. The position relative to the mask centre and the midline of gravity was further used to eliminate at most areas related to the septum pellucidum and shine through effects on the inferior third and fourth ventricles.

E. Implementation details - Choice of parameters

In order to obtain a final lesion segmentation, four steps are required:

1) Preprocessing
2) Initialization
3) Model selection
4) Application to lesion segmentation

Hereafter are the details of the parameters chosen for these different steps.

1) Preprocessing: The preprocessing of the data used in this study consists first in the spatial co-registration of the different modalities. As detailed in Section II-B2, statistical atlases are used at Level 1 and 2 of the hierarchy. For the first level, no *a priori* statistical information is known about the location of the outliers. Therefore, constant priors over the image are initially used so that $\forall n$, $b_{nO} = b_O$ and

$b_{nI} = 1 - b_O$. The value for b_O has been set to 0.01 in all our experiments, shown to be probabilistically equivalent to a Mahalanobis distance of 3 in previous work [55]. Conversely, for the second level, tissue-specific probabilistic maps are used to describe the prior probability of the four main tissues (GM, WM, CSF and NB). The probabilistic maps used in this study (ICBM452) are aligned with the observed data and re-normalized between 0 and 1 before use. For computational and complexity purposes, skull stripping is performed on the image [56]. The masks obtained for brain extraction are filled to include the ventricles and sulcal CSF using morphological operations (dilation of 2 voxels, filling and erosion by 1 voxel).

2) Initialization: The initial model $\mathbf{K}^{(\text{init})}$ is initialized as $K_{I_j} = 1$ and $K_{O_j} = 0 \ \forall j$, meaning that each inlier tissue component is being modelled by a single Gaussian, and the outlier tissue components are modelled by a uniform distribution. As the inlier mixtures are assumed to be governed only by Gaussians, the inlier classes' uniform distribution mixing weight is set to 0. To avoid instability and overfit in the presence of a large number of classes, the IIH is only optimized on the initial model $\mathbf{K}^{(0)}$, while the atlases $\boldsymbol{B}$ and $\boldsymbol{A}$ are relaxed after convergence of the EM on $\mathbf{K}^{(0)}$. They are considered static thereafter. The parameters that control the relaxation of the statistical atlases have been set to $\kappa = 1$ and $\sigma = 1$ and the EM is considered to have converged once the relative increase in the log-likelihood is less than 10^{-4}.

3) Model selection: The MRF spatial constraints and the constraint over the covariance are applied throughout the model selection process detailed in Section II-C whereas the correction for IIH is only performed once before the beginning of the model selection process. In order to behave similarly to the MRF weighting of 0.15 used in the VBM segmentation tool of SPM8 and also in [57], the symmetric matrix H containing the neighborhood energy cliques for the MRF is defined as:

$$H(l_{jk}, l'_{j'k'}) = \begin{cases} 0 & \text{if } j = j' \\ 0.15 & \text{otherwise} \end{cases}$$

Enforcing spatial consistency only at Level 2 of the hierarchical model should avoid the smoothing of relatively small but very hyper intense lesions that will be detected thanks to the inlier/outlier separation.

III. Data Experiments and Results

To validate the wide applicability of BaMoS, its performance was evaluated in various contexts: first it was applied to simulated data with presence of pathology at different levels of lesion load, with various degree of noise and bias field, as provided by the BrainWeb project [58]. The provided fuzzy lesion class membership was considered as ground truth (GT) for segmentation comparison. BaMoS application in the context of multiple sclerosis and age-related WMH was further evaluated on clinical datasets made available during the MICCAI challenge 2008 in the case of MS [59] and during the MICCAI BrainS challenge for the age-related WMH. Lesion manual segmentations were used as reference and considered as gold standard (GS). In those experiments, BaMoS was compared to two simpler versions of itself and to three other freely-available automated lesion segmentation algorithms, demonstrating both the importance of the adaptability of the model as well as its competitive performance when compared to other validated algorithms. Three versions of BaMoS have been tested: the first one, called BaMoS-static consists in the static solution obtained without allowing for a change in the number of Gaussian components, *i.e.* that performs the first initial EM, the atlas adaptation and a final EM refinement before exiting the process. The second version of BaMoS, denoted BaMoS-NoCov has been detailed in [35], *i.e.* without any prior over the covariances. Finally, the final full version of BaMoS is simply denoted BaMoS. In terms of computation, for instance for simulated data, about 250 EM related to model selection (Step 3 of the model selection process Section II-E3, Figure 2) were required for BaMoS, which amounted to about 12 hours of computation (single core desktop computer). Among the numerous methods described for lesion segmentation [9], we selected the ones which were made available online and maintained as listed in [60] as well as being able to handle data with artefacts. The first one was the classical EMS algorithm [6] that belongs to the same family of methods as BaMoS and thus enables a very similar set-up. The EMS code, available at (https://mirc.uzleuven.be/MedicalImageComputing/downloads/ems.phpEMSsection=download&pagePath=2) allows for a similar choice in the parameters (atlases, MRF), thus decreasing the comparison bias otherwise induced by preprocessing and parameter choice. The default value of 3 for the Mahalanobis distance, noted to be the most suitable [6] and used for comparison in [1] was chosen in all experiments. The second algorithm chosen, part of the SPM8 package and available for download at http://www.applied-statistics.de/lst_download.html is the Lesion Segmentation Tool (LST) detailed in [22]. Among the variety of proposed methods in the literature, this method has been validated both for application in the context of both MS and age-related WMH with a difference on a single threshold parameter. According to [61], the default value of 0.3 is to be chosen for MS applications (LST-MS) whereas the value of 0.25 is more appropriate when applied to age-related WMH (LST-WML). The third comparison point was the Lesion-TOADS (TOADS) algorithm [15], (http://www.nitrc.org/projects/toads-cruise/) that belongs to the family of non-parametric methods but also corrects for IIH within its scheme. It is available for installation at http://wwww.nitrc.org/projects/toads-cruise/. Two variants of the MRF energy matrix were used when comparing EMS to BaMoS: the H matrix defined for BaMoS (EMS-C) and the automatically adapted MRF detailed in [6], with the latter being the default set up of EMS algorithm (EMS-D). To ensure consistency in the comparisons, the same masked data were used for EMS, TOADS and BaMoS.

A. Assessment of lesion segmentation

Name	Equation	Best
DSC	$100\times\frac{2\cdot\sharp(\mathrm{Ref}\cap\mathrm{Seg})}{\sharp\mathrm{Ref}+\sharp\mathrm{Seg}}$	100 (%)
VD	$100\times\left\lvert 1-\frac{\sharp\mathrm{Seg}}{\sharp\mathrm{Ref}}\right\rvert$	0 (%)
FPR	$100\times\frac{\sharp\mathrm{Seg}-\sharp(\mathrm{Ref}\cap\mathrm{Seg})}{\sharp\mathrm{Ref}}$	0(%)
TPR	$100\times\frac{\sharp(\mathrm{Ref}\cap\mathrm{Seg})}{\sharp\mathrm{Ref}}$	100 (%)
FNR	$100\times\frac{\sharp\mathrm{Ref}-\sharp(\mathrm{Ref}\cap\mathrm{Seg})}{\sharp\mathrm{Ref}}$	0(%)
AvDist	$\frac{\sum\limits_{s\in\partial\mathrm{Seg}}\min\limits_{r\in\partial\mathrm{Ref}}d(s,r)+\sum\limits_{r\in\partial\mathrm{Ref}}\min\limits_{s\in\partial\mathrm{Seg}}d(s,r)}{\sharp\partial\mathrm{Ref}+\sharp\partial\mathrm{Seg}}$	0 (mm)
DE	$\sum_{\mathrm{F}\in\mathrm{FP}_c\cup\mathrm{FN}_c}\sharp\mathrm{F}$	0 (mL)
OER	$100\times\frac{\sum_{\mathrm{T}\in\mathrm{TP}_c}\sharp(\mathrm{Ref}_T\cup\mathrm{Seg}_T)-\sharp(\mathrm{Ref}_T\cap\mathrm{Seg}_T)}{\sharp\mathrm{Ref}}$	0 (%)

TABLE I: Table of lesion segmentation evaluation measures.

As noted in [9], the use of a unique method for the assessment of the quality of the lesion segmentation is insufficient and to better understand the strengths and weaknesses of different methods as well as the origin of the errors. Thus it can be useful to study jointly different segmentation assessment measurements. In this work, the popular Dice similarity coefficient (DSC) was combined with the true positive rate (TPR), and the false positive rate (FPR) as defined in [15] as well as the false negative rate (FNR). The other evaluation measures used here are the volume difference (VD) and the average distance (AvDist) also defined in [15]. To better assess the origin of errors, two measures, that have been shown to be less dependent on the total lesion load (TLL) according to [13] were used as a complement, respectively the detection error (DE) and the outline error rate (OER). The definitions of these quality assessment measures are gathered in Table I where $\sharp$ denotes the cardinality of a set considered in a voxelwise manner, the subscript $_c$ indicating the connected set, and ∂ denoting the border of a set defined in the 18-neighborhood connectivity scheme. With this denomination, considering $\mathrm{Seg}\cup\mathrm{Ref}$, FP_c denotes the set of connected components wrongly considered as positive, while FN_c corresponds to the set of connected components wrongly considered as negative. In turn, TP_c represents the true positive connected components, that is the connected components for which $\mathrm{Seg}\cap\mathrm{Ref}\neq\varnothing$. When possible, statistical tests between methods were performed using a non parametric Wilcoxon-Mann-Whitney test. Results were considered significant for p-values below 0.01.

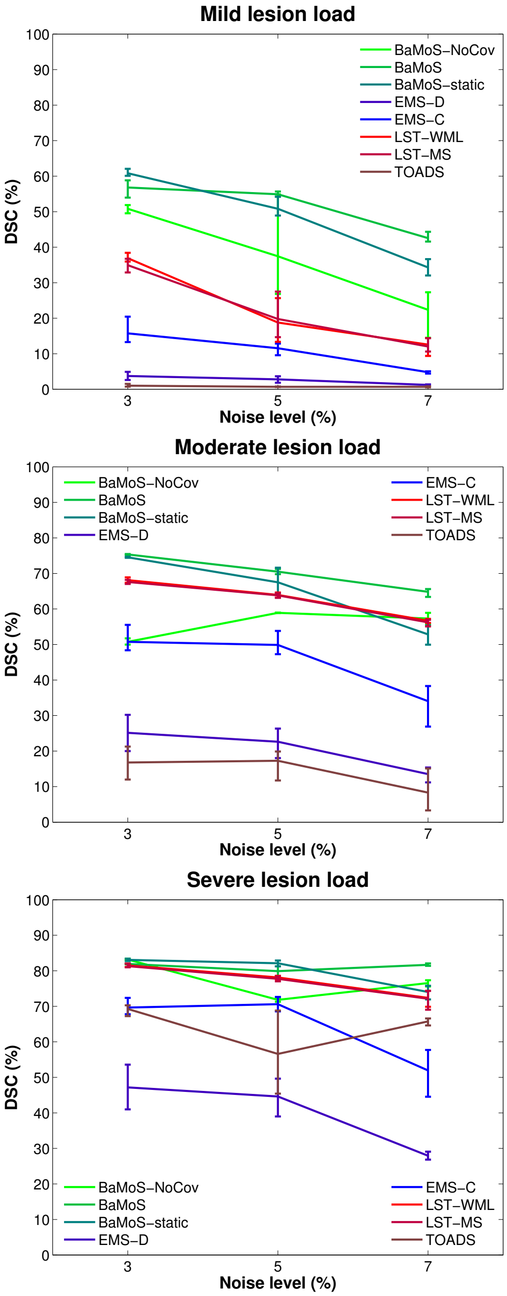

Fig. 3: Comparison of DSC results for the automated methods with noise level variation at mild (top), moderate (middle) and severe (bottom) lesion load. The errorbars refer to the minimum and maximum obtained when varying the intensity inhomogeneity level.

B. Validation on simulated images

As a first validation, the segmentation framework described earlier was applied to the simulated BrainWeb brain images available at http://brainweb.bic.mni.mcgill.ca in order to assess its performance for various level of image quality. A simulated model with multiple sclerosis lesions is available for three different lesion loads (Mild, Moderate, Severe). The ground truth segmentations are provided as maps of fuzzy membership. BaMoS was used on all different combinations of the available modalities (T1, T2 and PD) for different levels of noise (3%, 5% and 7%) at different

		Assessment method							
		DSC (%)	VD (%)	FPR (%)	TPR (%)	FNR (%)	AvDist	DE	OER (%)
Mild	BaMoS-NoCov	36.9	94.3	62.9	44.2	55.8	8.3	84.3	101.7
	BaMoS	**51.4**	46.4	44.9	52.0	48.0	**2.7**	**40.9**	86.5
	BaMoS-static	48.6	**43.1**	**25.6**	42.9	57.1	4.5	45.8	74.5
	EMS-D	2.6	4851.6	4893.3	**58.3**	**41.7**	40.0	13207.9	1063.9
	EMS-C	10.7	724.3	775.2	49.1	50.9	37.4	2600.4	82.8
	LST-WML	22.8	56.1	27.9	16.0	84.0	13.4	186.8	67.2
	LST-MS	22.3	70.6	15.0	14.5	85.5	13.5	146.8	**66.7**
	TOADS	0.8	983.8	1078.8	4.9	95.1	24.9	4190.6	137.3
Moderate	BaMoS-NoCov	55.7	75.3	91.6	72.1	27.9	1.3	507.6	104.8
	BaMoS	**70.2**	**22.9**	33.6	**72.8**	27.2	**0.8**	**173.9**	55.3
	BaMoS-static	65.0	33.9	**20.8**	60.2	39.8	1.3	309.2	51.3
	EMS-D	20.4	582.6	606.0	76.5	**23.5**	16.5	13197.8	188.0
	EMS-C	44.9	91.8	124.8	67.0	33.0	15.8	2935.0	65.8
	LST-WML	62.9	30.2	16.2	53.6	46.4	3.9	693.9	**42.1**
	LST-MS	62.6	34.7	13.4	51.9	48.1	4.0	648.9	42.4
	TOADS	15.0	64.2	144.3	19.9	80.1	7.8	1362.3	185.0
Severe	BaMoS-NoCov	77.2	37.5	46.3	**91.2**	**8.8**	0.5	71.6	54.3
	BaMoS	**81.2**	23.2	32.6	90.5	9.5	**0.4**	**33.6**	41.7
	BaMoS-static	79.7	**20.1**	**16.4**	77.9	22.1	**0.4**	50.1	**38.0**
	EMS-D	39.9	195.5	217.1	78.4	21.6	10.0	13105.7	88.7
	EMS-C	64.1	42.0	52.6	74.1	25.9	8.7	2596.9	49.6
	LST-WML	77.3	13.1	14.6	72.5	27.5	0.9	310.8	38.8
	LST-MS	77.1	15.8	13.0	71.2	28.8	0.9	286.6	38.9
	TOADS	64.6	6.1	31.5	62.6	37.4	2.1	838.2	60.6

TABLE II: Comparison for the different assessment measures of the segmentation method for the T1T2 combination modality. The results are taken as the mean over all level noise and IIH at the three different lesion loads.

Load	Noise	DSC (%)	VD (%)	FPR (%)	TPR (%)	FNR (%)	AvDist (mm)	DE (μL)	OER (%)
Mild	3	56.8	64.9	89.6	75.4	24.6	1.3	42.7	108.4
	5	54.9	17.7	32.3	50.0	50.0	2.1	32.0	77.5
	7	42.6	56.6	12.9	30.5	69.5	4.7	48.0	73.7
	Mean	51.4	46.4	44.9	52.0	48.0	2.7	40.9	86.5
Moderate	3	75.3	10.3	31.1	79.2	20.8	0.5	44.0	50.1
	5	70.5	33.5	51.2	82.3	17.7	0.6	59.3	66.9
	7	64.8	24.9	18.3	56.8	43.2	1.4	418.3	49.1
	Mean	70.2	22.9	33.6	72.8	27.2	0.8	173.9	55.3
Severe	3	81.9	28.7	35.0	93.7	6.3	0.4	45.3	40.9
	5	79.9	34.6	40.9	93.7	6.3	0.4	30.0	46.8
	7	81.7	6.2	22.0	84.2	15.8	0.4	25.3	37.4
	Mean	81.2	23.2	32.6	90.5	9.5	0.4	33.6	41.7

TABLE III: Assessment of BaMoS for various measures at different noise level and lesion load for the T1T2 modality combination (mean over various intensity inhomogeneity levels).

severity of intensity inhomogeneity (0%, 20% and 40%) for the three available lesion loads (0.4 mL, 3.5 mL, 10.1 mL). For the T1T2 combination that is the standard choice of modalities for the tested segmentation methods, the various lesion segmentation assessment measures are gathered in Table II as the mean of each measure across noise and IIH level for each lesion load separately.

1) Behavior with noise level: Since the quality of imaging data is very heterogenous, assessing the robustness of a given method against the level of noise is an important validation step. The range of noise between 3% and 7% (as defined in BrainWeb) was found to be comparable to the range of noise of 3T and 1.5 T clinical scans, and was thus used for comparison. The noise model in BrainWeb consists in adding Gaussian white noise on both the real and imaginary components of the image with a standard deviation chosen based on a reference tissue signal such that the ratio between the standard deviation and the signal is the percentage value of the noise model. The obtained noise on the magnitude image thus follows a Rician distribution. As the signal for the reference tissue varies through modalities, the observed effect of noise is also modality-dependent. The robustness to the noise level for the different lesion loads for the compared methods is presented using the DSC in

Figure 3 for which the methods are applied on the T1T2 modality combination. At each noise level, the extremities of the errorbar present the minimum and maximum result obtained when varying the intensity inhomogeneity level from 0 to 40%. For the T1T2 combination, Table III gathers the various lesion segmentation evaluation results obtained when varying level of noise and lesion load for BaMoS supporting the notion that BaMoS is robust to noise. When increasing the noise level, we generally observed a logical decrease in FP related to the boundaries of the lesions and an increase in FN. As shown by the related changes in DE and OER, mis-detection of lesions increases only for the change between 5 and 7% noise on the moderate case. The number of subclasses necessary to model the data was negatively correlated with the noise level, as the wider class variance makes it hard to justify the need to more Gaussian classes under the BIC model.

		Modality combination			
Load	Method	T1PD	T1T2	T1T2PD	T2PD
Mild	BaMoS-NoCov	10.3	36.9	35.8	17.9
	BaMoS	11.1	51.4	47.5	18.0
	BaMoS-static	10.0	48.6	40.9	16.7
	EMS-D	0.7	2.6	3.0	1.2
	EMS-C	5.7	10.7	7.1	4.9
	LST-WML	0.0	22.8	/	/
	LST-MS	0.0	22.3	/	/
	TOADS	0.8	0.8	/	/
Moderate	BaMoS-NoCov	29.1	55.7	56.7	28.4
	BaMoS	27.5	70.2	63.5	29.3
	BaMoS-static	23.8	65.0	59.9	32.2
	EMS-D	7.4	20.4	24.9	14.4
	EMS-C	22.8	44.9	43.0	22.3
	LST-WML	0.0	62.9	/	/
	LST-MS	0.0	62.6	/	/
	TOADS	14.0	15.0	/	/
Severe	BaMoS-NoCov	40.0	77.2	72.3	60.6
	BaMoS	42.2	81.2	77.3	60.2
	BaMoS-static	32.8	79.7	77.8	57.9
	EMS-D	13.9	39.9	46.2	34.0
	EMS-C	26.3	64.1	61.2	37.7
	LST-WML	0.0	77.3	/	/
	LST-MS	0.0	77.1	/	/
	TOADS	65.1	64.6	/	/

TABLE IV: Mean DSC (%) results over noise and IIH levels for the compared methods for various modality combinations at the three lesion loads. The slash (/) sign indicates that the combination was not possible to use for the given method.

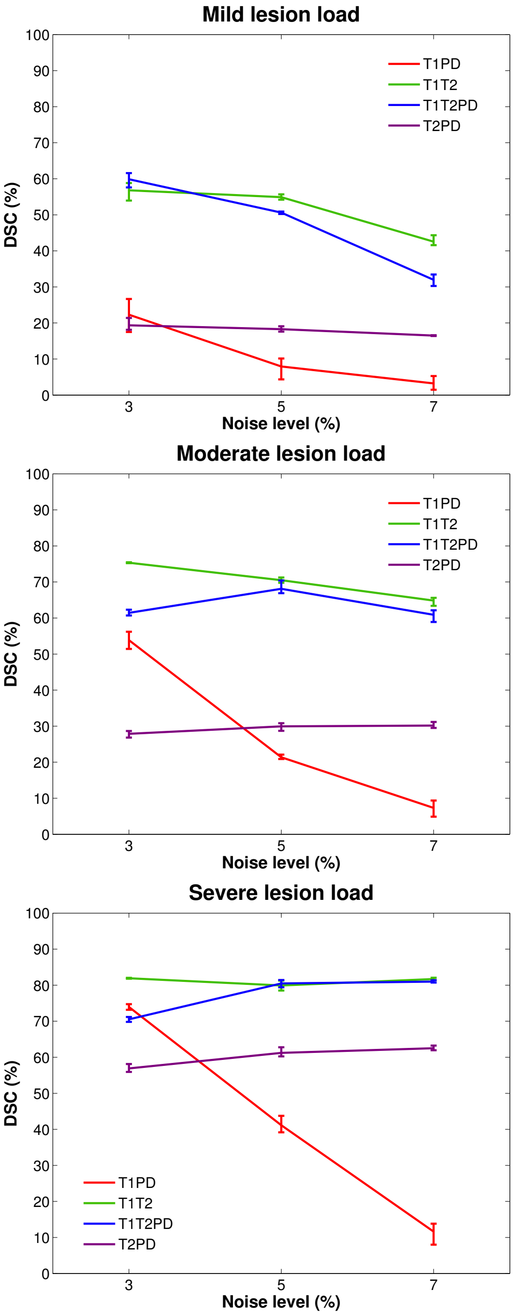

Fig. 4: Comparison of the DSC results for BaMoS at different noise levels for different modalities combinations for mild (top), moderate (middle) and severe (bottom) lesion load. The errorbars indicate the minimum and maximum obtained when varying the IIH level.

2) Impact of modality combination: It has been suggested in [62], that the variability in the choice of imaging modalities for clinical studies might be a cause for discrepancies between conclusions as the association between white matter lesion location and clinical outcome. Moreover, in clinical trials where multiple modalities are acquired, knowing the performance of different modality combinations may help in the appropriate choice of sequences. In this perspective, the graphs in Figure 4 show the impact of the noise level on the result for BaMoS at different lesion loads for the different modality combinations, while Table IV gathers the mean DSC for the compared methods and for the different modality combinations.

As neither TOADS nor LST are able to handle the T1T2PD and T2PD combinations, the results are not provided. The visual comparison between the segmentations

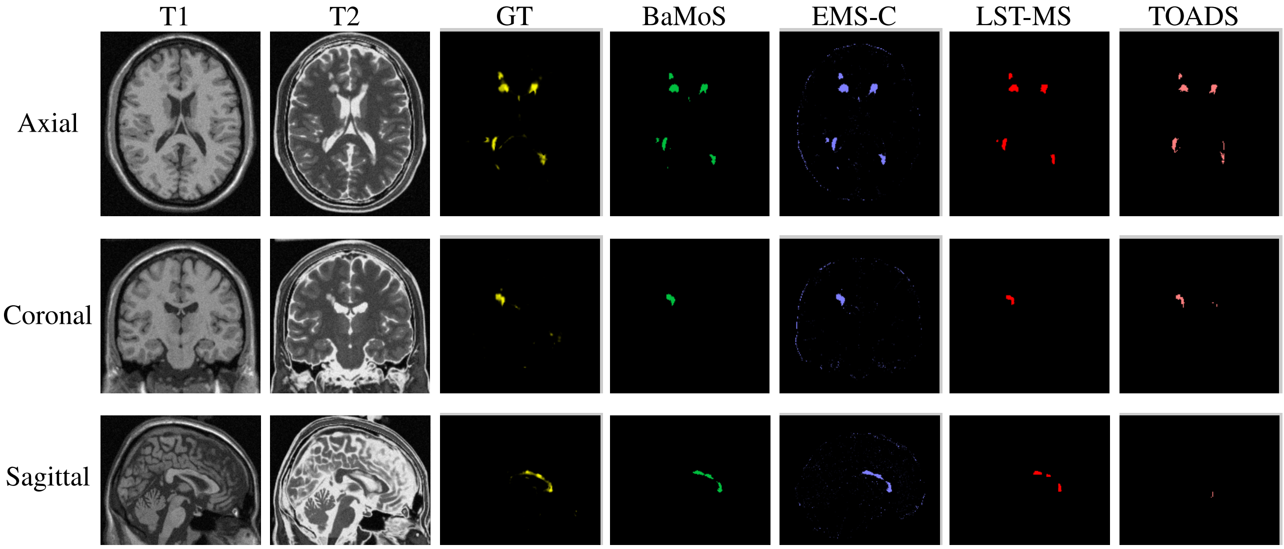

Fig. 5: Simulated BrainWeb multiple sclerosis model with severe lesion load case. Each row displays in a different orientation (axial, coronal and sagittal) from left to right the T1 image, the T2 image, the ground truth (GT) for the lesion segmentation and the corresponding results for BaMoS, EMS-C, LST-MS and TOADS.

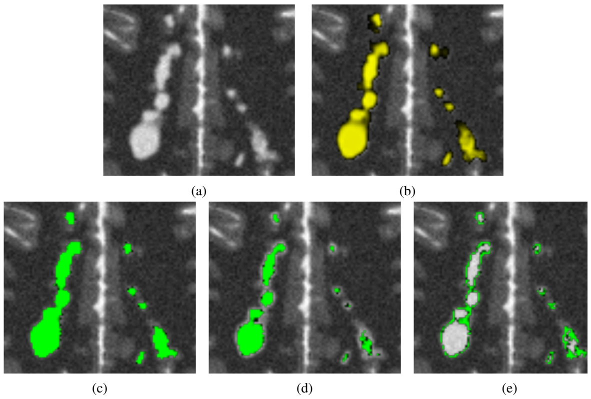

Fig. 6: Enlarged section on an axial slice of the T2-weighted (a) simulated image with severe lesion load. Overlayed with the lesion segmentation ground truth (b), the total segmentation obtained with BaMoS (c) and the two separated components lesion-related (d-e). Note that the separation between the lesion-related components is linked to the outlierness of the lesion.

obtained for the automated methods (BaMoS, EMS-C, LST-MS and TOADS) is presented in Figure 5. The lesion-related sub-components from the BaMoS model are displayed as an enlarged axial section in Figure 6. Note that different clusters are formed according to the lesion severity of the underlying voxel.

C. Clinical data

Due to its low performance over all experiments, the results obtained for EMS-D are not presented in the remaining of the figures. The suggested LST-MS parameter set (0.3) was used for the MS experiment and the LST-WML parameter set (0.25) was used for the WMH experiment.

1) MS lesion segmentation: The MICCAI Challenge 2008 (http://www.ia.unc.edu/MSseg/) data set was used to further validate BaMoS for MS. For this dataset, 20 T1 T2 and FLAIR images are provided along with the manual segmentation. All images were resampled isotropically to the space of the T1 image. Comparison between methods was performed using T1 and FLAIR images. For the assessment methods described in Table I, statistical results are gathered in Figure 7 where each reference method (in rows), is compared against all other methods using all assessment measures. In this infographic, green corresponds to a significantly better performance, grey to a non stastically significant difference in performance and red to a significantly worse performance. For each measure, the diagonals are kept white. For this dataset, BaMoS and LST both appear to perform better than TOADS and EMS. Figure 8 presents an example of the obtained segmentations for the different automated methods. Also, when comparing the three versions of BaMoS, the only significant differences observed were related to the DSC and the TPR for which BaMoS performed significantly better than BaMoS-static.

Among the 30 available datasets used for testing, the publically available results concern only 23 of them. The obtention of the scores relative to the quality of the segmentation is described in [59]. When comparing on these 23 images, the overall mean scores, for BaMoS, EMS, LST and TOADS were respectively 79.9, 62.7, 80.0 and 69.9. Note that the current version of TOADS appeared to perform worse than the one tested in 2010, for which the obtained score was 79.9. For the sake of consistency across the experiments the same currently freely-available online

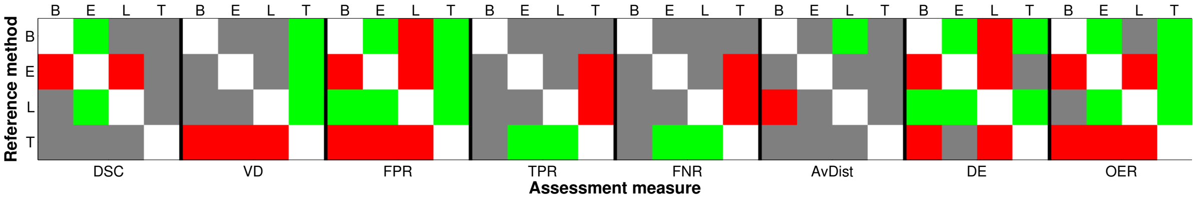

Fig. 7: Color-coded statistical difference significance summary for each assessment measure on the MS dataset, where each automated reference method: BaMoS (B), EMS (E), LST (L) and TOADS (T) is tested against another method (column) for a specified assessment measure. Green relates to a significantly better performance, Red to a significantly worse performance and Grey to a non statistically significant difference

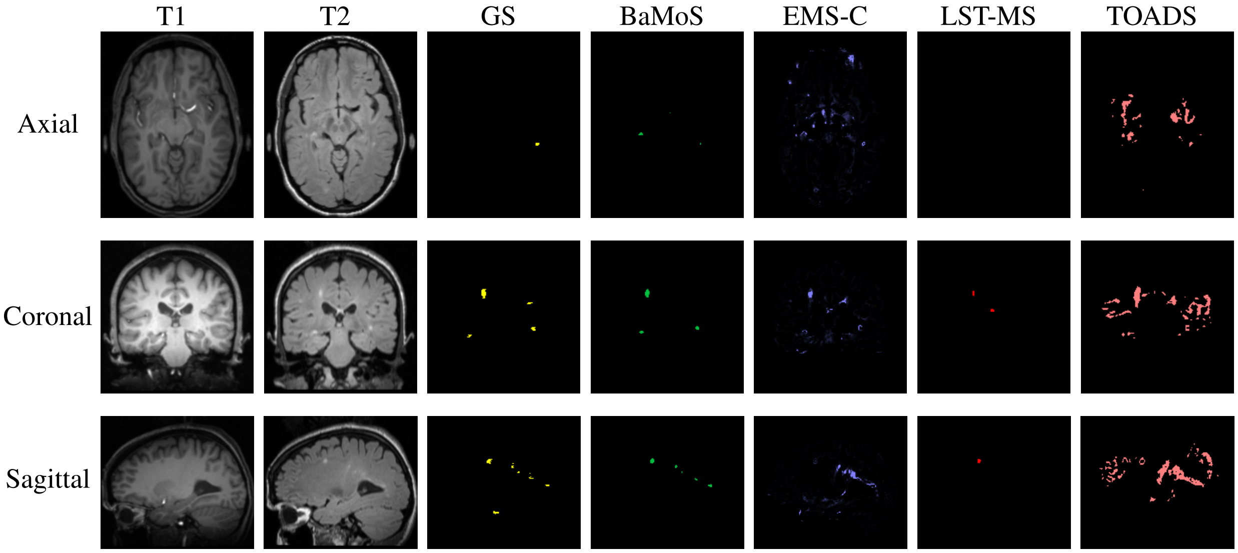

Fig. 8: Comparison of segmentation results for an MS patient. Each row displays in a different orientation (axial, coronal and sagittal) from left to right the T1 image, the T2 image, the gold standard (GS) manual lesion segmentation and the corresponding results for BaMoS, EMS-C, LST-MS and TOADS.

version of TOADS was used in all the experiments. The comparison was limited to the available algorithms in order to keep a consistent preprocessing of the data and therefore limit its influence over the final segmentation result. Thus, in a generic framework, in which the model selection is not finely tuned towards the specific detection of white matter lesions, BaMoS appeared to be competitive compared to other methods.

2) WMH in population with cardiovascular risk and diabetes: WML can present themselves in many other contexts than MS and are for example known to appear as part of the aging process. Age-related WMH and MS lesions, though arising from different pathological pathways may have similar appearance on T2 and FLAIR modalities [63] and additional shape and location criteria have been put forward to increase the diagnostic specificity [64]. In age-related white matter lesions, one of the aetiological explanations for the existence of such lesions is the partial ischemia of the white matter. As the myelin degrades, the fat/water ratio in the white matter changes, leading to the signal change observed in the MRI images. Such white matter lesions are thought to correlate with cognitive decline and disability in the elderly [65] and have been associated with risk factors such as Type 2 diabetes (T2DB) [66] and cardiovascular risk factors [67]. Compared to the general population, the decrease of contrast between WM and GM or the enlargement of the ventricles related to the aging process may further affect the detection of WML [3]. In this context, the behavior of BaMoS was assessed on images part of the MICCAI MRBrainS2013 Challenge.

For this study, brain images from T2DB patients and matched controls with increased cardiovascular risk (age > 50) were acquired on a 3T Philips scanner. Multi-slice FLAIR images ($0.958 \times 0.958 \times 3$ mm) and T1-weighted 3D registered images were used. Further details about the acquisition and preprocessing can be found at http://mrbrains13.isi.uu.nl. WMH were manually segmented on twenty FLAIR images giving a total lesion load (TLL) range between 0 mL and 35.48 mL (median 6.02 mL, interquartile range 9.22 mL) and used as gold standard for the evaluation of the automated methods. No WMH was detected by the human rater for one of the twenty patients. The obtained lesion segmentations were evaluated using the various lesion segmentation assessments defined in Section III-A for the nineteen subjects in which TLL > 0.

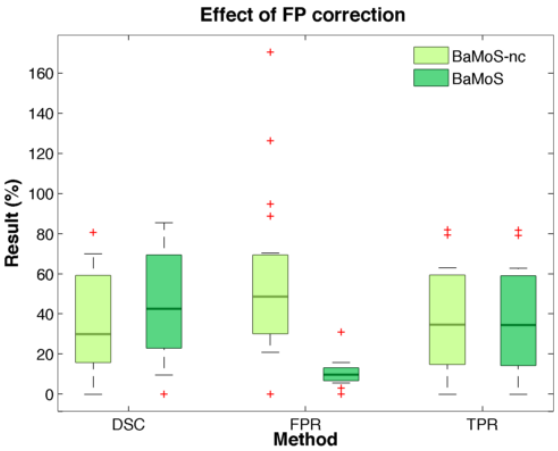

Fig. 9: Effect of the FP correction on BaMoS in terms of DSC, FPR and TPR. The correction reduces the FPR but does not affect the TPR. BaMoS-nc refers to the result of BaMoS uncorrected for FP.

In addition, the impact of the correction for false positives after selection of the lesion-related components is shown in Figure 9 for the DSC, the TPR and the FPR. The stability in TPR supports the appropriateness of the FP detection.

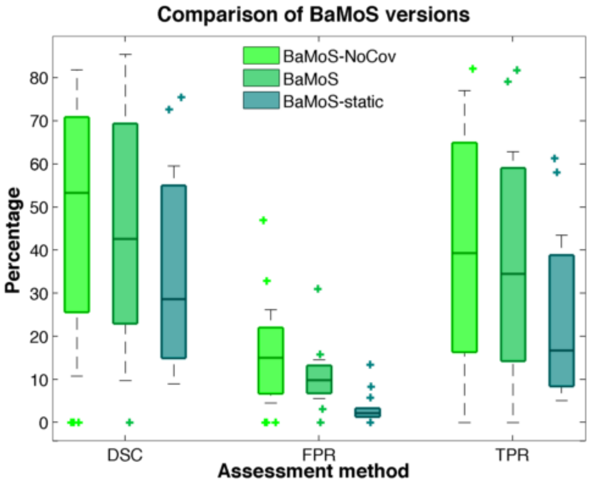

Fig. 10: Comparison of the three versions of BaMoS in terms of DSC, FPR and TPR. Note the existence of low outliers for the DSC in the BaMoS-NoCov version. The only outlier for BaMoS corresponds to the case with only 0.06 mL of lesion.

The comparison between the three versions of BaMoS, is presented in Figure 10. The apparently surprising observation that BaMoS-static did not contain any DSC of 0 is caused mostly by the fact that the lesion segmentation is performed on a voxelwise basis for this method. The only DSC of value 0 observed for BaMoS corresponds to the case where only 24 voxels were manually segmented as lesion. Overall, BaMoS performed significantly better than BaMoS-static for global assessment measures (DSC, VD, TPR) but no significant difference was observed for the Average distance, DE and OER. BaMoS-NoCov performed significantly better than BaMoS-static for VD, TPR but not for DSC. BaMoS-static performed significantly better in terms of FPR compared to BaMoS and BaMoS-NoCov but this observation is directly linked to the low TPR. The seeming discrepancy in results showing a higher median for DSC and TPR of BaMoS-NoCov compared to BaMoS, but a significant improvement of the DSC results of only BaMoS over BaMoS-static is related to the existence of outliers in the results of BaMoS-NoCov. This outlines the positive impact on the robustness of the method when including the prior over the covariance.

The comparison between the other automated methods is summarized in terms of statistical significance in Figure 11 showing that for this application, BaMoS outperformed EMS and LST and performed similarly to TOADS.

Since the measure of global TLL has been related with cognitive decline, TLL correlations between the automatic and manual segmentations were studied using both the Pearson's R^2 correlation coefficient and the slope of the linear regression for the twenty cases. The quantitative results for both lesion segmentation assessment measures and TLL regression are presented in Table V. The TLL linear regression, whose results are presented in the last couple of lines of this table, was performed over the 20 subjects whereas all the other assessments are summarized for subjects with a positive TLL (19). BaMoS slightly underestimated the lesion volume (linear coefficient of 0.88) but the correlation was high when compared to the other methods (R^2=0.96). The volume related study is presented visually in Figures 12 and 13. The Bland-Altman plot shows less bias for BaMoS. The plot of automated TLL per patient ordered by increasing manually segmented TLL highlights a potential problem regarding lesion overestimation in TOADS for very mild cases.

The analysis of the errors observed in BaMoS with respect to the manual segmentation showed that 12% of the FN corresponded to missed lesions, the rest being related to the outline of the lesions (*i.e.* border disagreement). Among those missed lesions, 87% of this amount corresponded to lesions with a volume lower than 0.1 mL. When comparing the number of missed lesions in the automated methods, no significant difference was observed between BaMoS, EMS and TOADS, and all were able to detect significantly more lesions than LST. The periventricular region, known to be prone to partial volume effect due to resampling and shining through effects on the ventricular lining was the most prone to FN outline errors. In turn, the occurrence of FP, that represent 20% of the errors concerned the outline for 65% of its volume. For the erroneously detected FP lesions, 66% of the volume corresponded to lesions of less than 0.1mL.

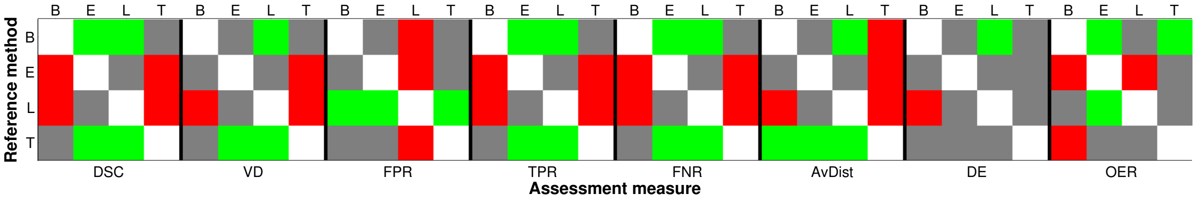

Fig. 11: Summary of statistical differences observed between the automated methods for each assessment measure in a Green Grey Red code on the T2DB dataset. Each method used as a reference (row) is compared to the other three (column). Significantly better and worse performances for a specific assessment are coded in green and red respectively. No statistically significant difference is coded in grey. Diagonals stay white.

Method	BaMoS	EMS-C	LST-WML	TOADS
DSC	46.2 26.5	27.92 18.77	30.9 25.0	52.1 17.6
VD	52.0 29.0	105.6 193.8	75.7 21.4	183.2 644.0
FPR	10.2 6.5	68.7 222.7	2.6 3.0	177.4 666.9
TPR	37.7 25.8	20.9 13.9	21.7 20.1	44.5 12.9
FNR	62.3 25.8	79.1 13.9	78.3 20.1	55.5 12.9
AvDist	6.8 11.6	7.7 5.6	10.2 12.4	3.8 6.9
DE	1.0 0.6	1.5 1.5	1.9 1.1	1.1 0.7
OER	53.3 23.2	69.2 15.8	49.7 0.7	86.1 100.2
R^2 **Lin**	0.96 0.88	0.86 0.41	0.94 0.63	0.90 0.53

TABLE V: Comparison of the different methods according to the various lesion segmentation assessment measures for the T2DB dataset of 19 subjects with positive TLL. All the measures are given in a two lines format with the mean on the first and the standard deviation on the second. The last set of lines gives the Pearson's R^2 correlation coefficient for the twenty subject and the corresponding linear coefficient (Lin).

Those FP lesions were mostly located close to the ventricular lining. Similarly to what was observed on MS simulated data in Section III-B, BaMoS was able to separate different types of lesion in clinical data. Other types of outliers, such as areas of iron deposition in the basal ganglia with much darker intensities on FLAIR images were also assigned their own cluster. A visual example of BaMoS' ability to stratify different types of WM and deep GM sub-clusters is presented in Figure 14. Thanks to the BIC constraint, the number of subclasses observed for the inlier classes was stable across the clinical dataset.

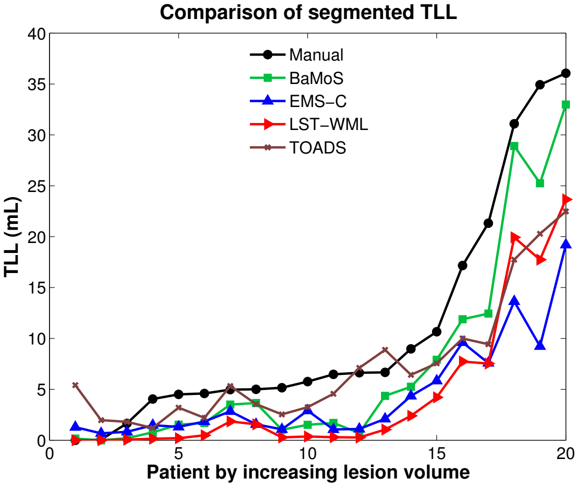

Fig. 12: Comparison of TLL per patient for the four automated methods against the manual segmentation (black line).

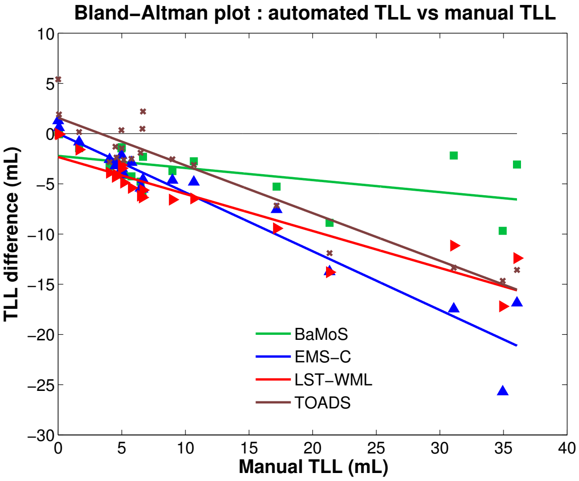

Fig. 13: Bland-Altman plot of the automated methods against the manual gold standard segmentation. The markers represent the twenty cases from the T2DB dataset and the line the corresponding linear fit TLL_{auto} - $\text{TLL}_{\text{manual}}$.

D. Specific caution on the use of manually segmented gold standard for lesion segmentation assessment

It is a well known fact that especially in the case of white

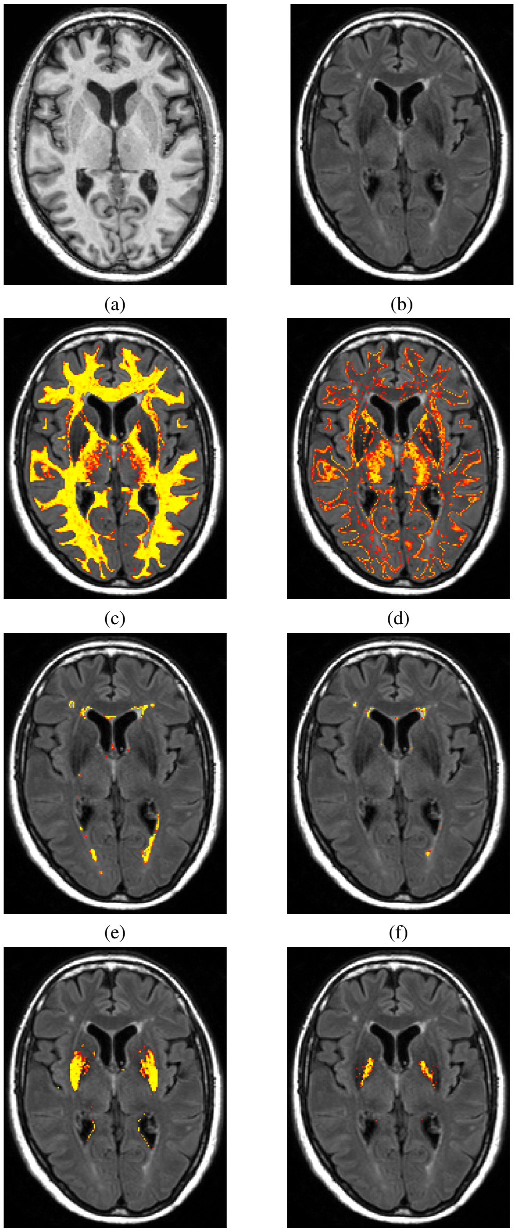
(a) (b) (c) (d) (e) (f) (g) (h)

Fig. 14: WM segmentation for one case of the clinical T2DB dataset. First row: Images of two modalities used T1 (a) and FLAIR (b). Second row: two subclasses obtained for the inliers of the WM (c-d). Third and fourth row: 4 subclasses classified as outliers, spatially related to WM presenting hyper- (third row: e-f) and hypo-intensities (fourth row: g-h).

matter lesions, the tremendous work required to manually delineate the lesions is subjected to a high inter- and intra-rater variability [13]. Furthermore, as the clinical definition of such lesions remains unclear [14], the protocols designed to manually segment those lesions might result in a lack of consistency that can for instance be expressed in terms of the intensity characteristics of the lesions. As those manual segmentations are the basis of any validation of automated methods, methods that tend to reproduce human behavior will appear to perform better compared to methods more focused on the understanding of the underlying signal and its biological classification.

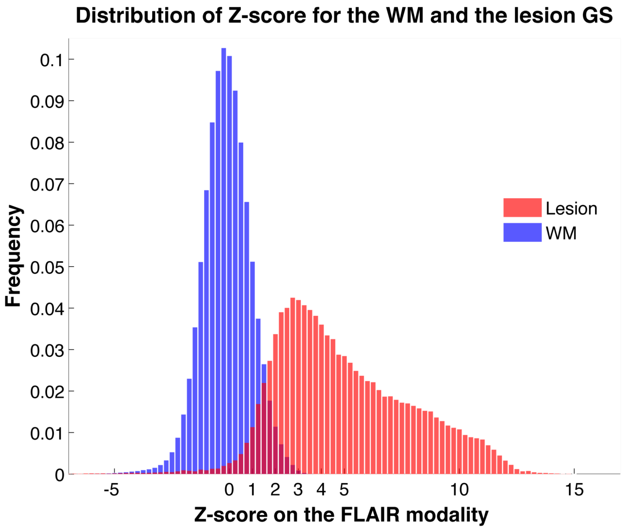

Fig. 15: Compared Z-score distribution of manually segmented lesion and WM with respect to WM on the FLAIR modality.

Based on the age-related dataset (aging population with diabetes and/or cardiovascular risk) presenting WML for which the data provided was already corrected for intensity inhomogeneities, the intensity distribution of the segmented WM was compared to that of the manually segmented lesions. Note that in MS datasets, iso-/hypo-intense FLAIR regions surrounded by an hyper-intense rim are considered as lesions [68] but this is not the case for the age-related dataset. The white matter mask used as reference for the distribution of normal intensities was obtained as the intersection of the resultant WM obtained by the four automated methods. Voxels that belonged to the manual segmentation were removed from this mask. Since in MRI the absolute signal value is not quantitative, Z-scores and Mahalanobis distances with respect to the mean of the WM were used here to assess the relative signal distribution. Figure 15 presents as an example the distribution of intensity Z-scores with respect to the WM on FLAIR modality for the manually segmented lesions and the WM mask across the dataset.The overlap observed and the fact that the mode of the lesion segmentation is located on this overlapped region, highlights the difficulty to define consistently the limit between normal appearing white matter, dirty appearing white matter and lesions.

Such uncertainty impacts directly the assessment of automated lesion segmentation methods. As an example Figure 16, presents the correlation between the proportion per manually segmented lesion of voxels whose intensities falls below the threshold of 2 in terms of Mahalanobis distance compared to the WM *versus* the lesion DSC for BaMoS. As expected, a negative correlation is observed. The DSC decreases when the proportion of normal appearing voxels considered as lesion increases (R^2 0.48, Linear coefficient -0.96). Thus, comparing automated methods to manual segmentations does not suffice to validate the algorithms' degree of systematism in the definition of lesion intensity characteristics.

Figure 17 illustrates such behaviors in terms of two possible assessment measures:

PropLes : Proportion of lesion intensities that overlap with WM intensities.

DistQuant : Difference in Z-score between the first quartile of lesion intensity and the third quartile of WM intensity.

A good consistency across the dataset would correspond to a low variance for these measurements. A low proportion overlap between intensities of normal appearing white matter and segmented lesions was observed for LST-WML and BaMoS. This effect can be explained by a more conservative segmentation algorithm. For example, such a behavior for LST-WML is directly related to the seed-growing principle on which this segmentation method is based. In a lesser extent, this can also be observed for BaMoS, as the evolving outlier atlas can be seen to provide ”seeds” for outlier segmentation. The high variance observed for TOADS in terms of PropLes, can be explained by the inclusion of false positives in CSF-containing regions when performing a filling of the lesions. In turn, EMS-C presented a low variance in the intensity characteristic that however corresponded to an important intensity overlap between lesions and normal white matter.

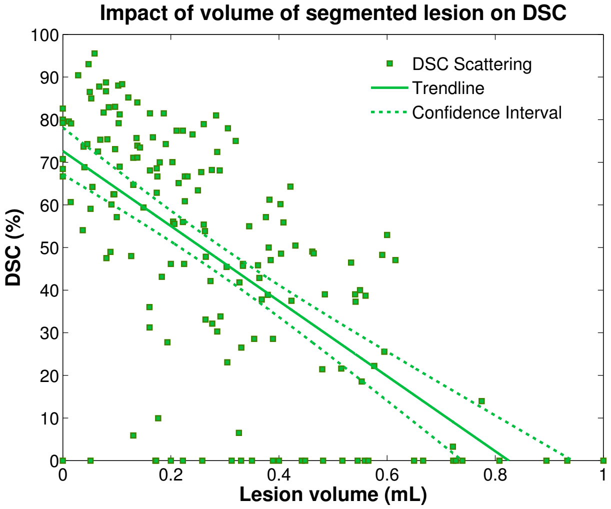

Fig. 16: Illustration of the impact of the intensity overlap between manually segmented lesions with normal appearing white matter. The DSC per lesion (with volume >0.05 mL) is plotted (scattered points) with respect to the proportion of manually segmented voxels that present an intensity at a Mahalanobis distance inferior to 2 compared to the normal WM. The bold and dashed lines represent respectively the trendline of the correlation and the corresponding 95% confidence interval.

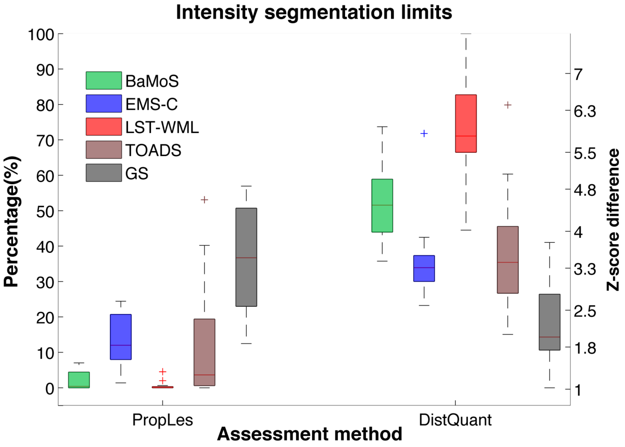

Fig. 17: Comparison of consistency in terms of lesion intensity limits for the four automated methods and the gold standard in terms of PropLes (left axis) and DistQuant (right axis).

IV. Discussion

In this work, we developed a comprehensive method to dynamically determine the most appropriate model to describe multimodal data in presence of outliers. Based on a complexity criterion, BaMoS automatically estimates the necessary number of components as well as the corresponding model parameters needed to model the inlier and outlier components of the data simultaneously, according to anatomical prior knowledge introduced as statistical atlases. The main advantage of this model is its ability to describe different types of outliers at the same time, thus avoiding any bias in the segmentation of the normal tissues.

The application chosen for the validation has been the extraction of white matter lesions (WML) in the case of multiple sclerosis (BrainWeb simulated data and clinical data) and age-related WML with type 2 diabetes or increased cardiovascular risk (clinical data) with a large range of lesion loads. The components of the model corresponding to the clinical definition of WML were extracted automatically after the model optimization, providing the final lesion segmentation described in Section II-D. As the model selection is independent of the observed pathology, BaMoS has a large clinical flexibility to different modalities and pathological contexts.

BaMoS separates the lesions into different components according to their intensity, opening the door to a deeper understanding of the underlying lesion's pathophysiology and highlighting the need for further investigation into the problem of partial volume between lesions and surrounding healthy tissue. Since a consistent and systematic intensity cut-off is difficult to draw between normal appearing white matter, dirty appearing white matter and true lesions, the ability to introduce fuzzyness and uncertainty in the lesion segmentation result may prove useful. Further investigation is still necessary to account for these effects locally. Such investigation would also benefit the final healthy tissue segmentation but is out of the scope of this work. These intensity levels may be of further interest when studying the evolution of the WMH longitudinally since in the elderly it has been shown that such lesions evolve from pre-existing WM damages [69].

Validation of automated segmentation methods is usually performed by comparison to a gold standard (GS) that is difficult to acquire for clinical data and is commonly based on manual segmentation. As the total volume load is currently the clinical standard for the assessment of lesions, correlations between the automated and manual lesion volumes, here considered as a clinical gold standard, is a common form of validating and comparing segmentation strategies [6], [61], [22]. Contrary to the lesion count, important when looking at early stages of MS, lesion volume has been found to be related to the clinical outcome and to cognitive decline [70] when investigating age-related white matter changes. In the age-related WML dataset, R^2 of the correlation was 0.96 for BaMoS compared to 0.88 for EMS-C, 0.94 for LST-WML and 0.90 for TOADS. Application of BaMoS to larger clinical datasets would however be needed to properly assess the relevance of the detected lesion burden in terms of disease staging and progression. As the TLL is insuficient to evaluate segmentation accuracy, lesion shape, localization accuracy and overlap [22] [9] [71], the eight segmentation evaluation measures defined in Table I were used to assess the automated segmentations and better analyse the origin of the segmentation errors. The study of the lesion count version of these measures would also be of interest for a more complete evaluation of the methods.

When compared to the validation on clinical data, the main advantage of using synthetic images is the availability of a ground truth. However, the validation using the BrainWeb dataset was limited by multiple factors as previously underlined in [1]. As only one phantom is available, no statistical analysis is possible. Also, synthetic images cannot be considered truly realistic. Moreover the range of lesion load is limited compared to the amount that can be found in clinical cases. Thus, the BrainWeb data was used here to test the algorithmic stability to different imaging modality combinations, different degrees of image quality with varying noise level and intensity inhomogeneity. BaMoS was more stable across noise level when compared to the other methods especially TOADS and EMS. Naturally, the decrease in TPR with image quality was most important in the mildest case since subtle and small lesions are more affected by the noise level than more prominent lesions as observed in the severe case. Comparing three versions of the proposed methodology: BaMoS-static, BaMoS-NoCov and BaMoS in its full version, BaMoS was observed to be more robust to IIH than BaMoS-NoCov and more robust to noise than both BaMoS-static and BaMoS-NoCov. BaMoS-static appeared to perform reasonably well as the BrainWeb data can be modelled by a limited number of Gaussian components. However, the TPR was clearly lower for BaMoS-static than for the other BaMoS versions. The constraint over the covariance matrix for the lesions contributed to promote smaller covariances and thus encourage a more detailed characterization of difference in levels of lesion severity and partial volume effect. Compared to the other three families of methods, BaMoS appeared more robust to noise, a feature that is particularly important for milder cases of lesion load. Naturally, the number of Gaussian components needed to model the lesions decreased with an increase in the noise level. BaMoS was reasonably stable when presented with different combinations of modalities and when compared to the other automated techniques that are optimized towards specific combinations. This can be of real interest when considering clinical studies for which some imaging modalities might not be available for certain subjects. Further investigation would be needed to better understand the biological correlates between various modalities and observed signal as well as the direct impact on lesion detection and segmentation.

As far as the clinical data is concerned, using the same assessment methods, when comparing the sub versions of BaMoS, the static version of BaMoS, appeared to be less sensitive to the presence of lesion and BaMoS appeared to perform significantly better than BaMoS-static in terms of DSC for both clinical datasets. This tends to illustrate the improvements brought by the use of a higher number of Gaussian components in the data modelling. In turn, despite performing marginally worse than BaMoS-NoCov in terms of median (BaMoS 42.6% vs BaMoS-NoCov 53.2%), BaMoS was found to be more stable as shown in Figure 10 (illustrated by a higher mean BaMoS 46.2% *versus* BaMoS-NoCov 45.2%) for the age-related WMH dataset. The lower sensitivity to lesion in BaMoS-static expressed itself both in terms of a significantly lower TPR and a significantly better FPR when compared to BaMoS and BaMoS-NoCov for the age-related dataset. The higher robustness of BaMoS compared to BaMoS-NoCov was further exemplified in the case of the clinical MS dataset for which significant differences between subversions was only observed for BaMoS which performed significantly better than BaMoS-static in terms of DSC, FNR and TPR. BaMoS obtained comparable, and sometimes improved results when compared to automated methods in both MS and age-related WML contexts with a tendency to slightly underestimate the lesion volume. When looking at the rate metrics, such as the somehow low TPR obtained for the WMH elderly population, one should keep in mind that rates are prone to large variations when the rate normaliser is small. Therefore, a few missegmented voxels in subjects with low lesion load result in a low average TPR. We noted however that BaMoS had consistently good results when compared to the heterogeneous performance of competing methods.

The analysis of the origin of possible false negatives for BaMoS showed that the undetected regions were mostly related to small lesions of less than 0.01 mL. Lesion segmentation accuracy was also found to be negatively correlated with the proportion of lesion voxels with WM-like intensities. Given the strength of BaMoS model selection process, improvements in sensitivity could be obtained through the use of shape or texture features in the post-processing step. This observation relates strongly to the segmentation protocol defined in [68] for the segmentation of MS lesions in which a conservative segmentation of lesions promoting false negatives over false positives for mildly hyper-intense regions is encouraged to enable the observation of change over time.

Good scores in ground-truth based assessment mostly characterizes the algorithm's ability to reproduce human behavior rather than its true ability to detect abnormal biological signal. One should thus be cautious when assessing an algorithm entirely based on manual segmentations. Due to the extreme difficulty of the manual segmentation task, clinical expertise might be encouraged to provide careful lesion definitions that can be translated into automatic methods. In addition, the variations in the degree of confidence in what should be considered as lesion or not, should further support the argument for non-binary lesion segmentations. Surrogate measurements, such as the ability to predict clinical outcome or measures of intensity consistency proposed in Section III-D can be used to further validate automated algorithms. Indeed, as far as treatment effects are concerned, consistency in the segmentation may appear of greater importance than sensitivity since a predictable bias is commonly much less detrimental to the power of longitudinal studies than a larger variance.

In many methods specially dedicated to the segmentation of white matter lesions, a post processing step is often needed to avoid taking into account the hyper-intense voxels due to flow artefacts in FLAIR images [72] [73] or voxels at the border between GM and WM [74], or other types of false positives [75]. The post-processing FP correction heuristics applied are mostly needed due to the presence of imaging artifacts and minor anatomical variations and do not affect the generality of the BaMoS model that consider the brain as a mixture of normal appearing (inliers) and unexpected (outliers) tissues. The FP correction performed in this work is partly related to the smoothness of the statistical atlases. The spatial separation in the modelling of the outliers enables indeed to naturally avoid the CSF-related flow artefacts. The use of different atlases could lead to a modification in the post-processing selection process that would reduce the FP correction. Besides, in this framework, BaMoS does not rely on a prior T1 segmentation of the white matter that can be biased toward the presence of hypo-intense lesion regions [74] [22].

In addition to false negatives undetected in small lesions, the ventricular lining appeared to be the most prone to false negatives. As this region has been highlighted as of special interest when dichotomizing the clinical impact of white matter lesions [76], further work is needed to avoid confusion between deep gray matter and WML.

As each split operation is mostly intensity driven, the biological interpretation of the resulting clusters remains difficult. Causes other than the underlying biology, such as partial volume effect and remaining bias field inhomogeneities, may induce similar changes in the images. The robustness of the method could be also tested using same-day repeatability scans. It should be also noted that the prior over the covariance matrix might hinder the detection of hypo-intense outliers.

The use of multiple channels appears to be important to stratify different levels and/or types of lesions. When FLAIR images are available, the process of anatomical prior relaxation improves the ventricular segmentation, enabling the correct classification of periventricular lesions as a subclass of the WM outliers. Even in difficult conditions such as anisotropic clinical data, BaMoS performed well. Nonetheless, BaMoS performance should be further improved by the reduction of flow and resampling artefacts [18] in 3D isotropic FLAIR images as encouraged in [60]. The generic framework that governs BaMoS makes it applicable for all possible range of modality combinations contrary to TOADS or LST. As the lesion selection process is performed after selection of the hierarchical Gaussian mixture model, the final segmentation remains independent from the lesion definition that can be adapted to the variability in lesion definition contrary to EMS in which the lesion definition is included in the initial description of the model. This flexibility allows for more variation in the clinical description of lesions and further emphasize the generability of BaMoS.

In our validation, we focused our efforts on the segmentation of white matter lesions as they are of biological and clinical interest in improving our understanding of the aging process and cognitive decline and assessing the evolution of multiple sclerosis. The main strength of the proposed method is its ability to model different types of outliers in a consistent and unified manner. This could enable further correlative studies between white matter disease and related outcomes. Owing to the generic outlier modelling obtained through BaMoS, associations of different types of outliers such as white matter lesions and extent of iron deposition could be further investigated since iron deposition is known to be associated with cognitive decline in aging [77] and observed in the course of MS [78]. Future work will explore the ability to characterize different degrees of hyper-intensity, if related to lesion severity, which might prove useful in the longitudinal assessment of certain pathologies. Also, as the definition of WML in FLAIR images is still far from consensual [79], future applications will also explore the usefulness of certain subclasses of the inlier WM to characterize normal *vs.* dirty appearing white matter.

V. Appendixes

A. Acronyms and abbreviations

The following acronyms are ordered alphabetically and not by order of appearance.

AvDist	Average distance.
BaMoS	Bayesian Model Selection.
BaMoS-No-Cov	Version of BaMoS without the covariance constraint.
BaMoS-static	Version of BaMoS without Model evolution.
BIC	Bayesian Inference Criterion.
CSF	Corticospinal fluid..
DE	Detection error
DSC	Dice Similarity Coefficient.
EM	Expectation-Maximization.
EMS	Expectation Maximization Segmentation tool
EMS-C	Clinical version of EMS (with MRF as defined in Section II-E3)
EMS-D	Default version of EMS (with adaptive MRF).
E-step	Expectation step.
FLAIR	FLuid Attenuation Inversion Recovery.
FPR	False Positive Rate.
FWHM	Full Width at Half Maximum.
GM	Grey matter.
GMM	Gaussian mixture model.
GT	Ground Truth.
GS	Gold standard.
ICM	Iterative Conditional Mode.
IIH	Intensity inhomogeneity.
KLD	Kullback-Leibler Divergence.
LST	Lesion Segmentation Tool.
LST-WML	Clinical version of LST (0.25) optimized for WMH.
LST-MS	Default version of LST (0.3) optimized for MS.
MAP-EM	Maximum a Posteriori Expectation-Maximization.
MRF	Markov Random Field.
MR(I)	Magnetic Resonance (Imaging).
MS	Multiple sclerosis.
M-step	Maximization step.
NB	Non-Brain.
OER	Outline Error Rate.
PD	Proton density weighted.
SM	Split and merge.
T1	T1-weighted.
T2	T2-weighted.
T2DB	Type 2 Diabetes.
TLL	Total lesion load.
TLL_{auto}	TLL obtained by an automated method.
$\text{TLL}_{\text{manual}}$	Manually segmented TLL.
TPR	True Positive Rate.
VD	Volume difference.
WM	White matter.
WMH	White matter hyper-intensity.
WML	White matter lesion.

B. Mathematical notations

In this work, a non bold lower case symbol corresponds generally to a scalar. Depending on the context, the notation f can either refer to a distribution density function or to a probability function.

1) Superscripts and generalities:

$^{(t)}$	Iteration t.
^{c}	Corrected for IIH.
^{T}	Applied to a vector or a matrix, denotes the transposition operation.
$^{(d)}$	Taking the value d or Patho, T1, T2, denotes a specific channel.
$\lvert\,\rvert$	Applied on a matrix, denotes the determinant of the matrix.
$\mathscr{E}$	Applied on a random variable, denotes the expectation of this variable.
$\boldsymbol{e}_\ell$	The ℓ vector of the canonical basis .
$\star$	Indicates a convolution operation.

2) Indexing and counters:

n and N	Applied to the voxels of the image, with index n and total number N.
m and M	Applied to the IIH polynomial basic functions, with index m and total number M.
d and D	Applied to the number of image modalities, with index d and total number D.
l	Used for the Level 1 of the model hierarchy and assume the value I or O.
j and J	Used for the Level 2 of the hierarchy with index j, and J representing the total number of anatomical classes.
k and K_{l_j}	Used for the Level 3 of the hierarchy with index k, and K_{l_j} representing the number of Gaussian distributions used to model class l_j.

3) Sets and vectors:

$\mathbf{y}$	Indexed by n and of size D, refers to the vector of normalized log-intensities.
$\mathscr{Y}$	Set of vectors $\mathbf{y}_n$, with n varying from 1 to N.
$\mathbf{z}$	Indexed by n, represents the unity vector of the canonical basis characterising labelling configuration for voxel n.
$\mathscr{Z}$	Set of vectors $\mathbf{z}$, represents the full labelling configuration for the images (hidden data).

4) GMM density distributions and parameters:

$\boldsymbol{\mu}$	Mean (vector) for a Gaussian distribution.
Λ	Covariance matrix for a Gaussian distribution.
$\boldsymbol{\theta}$	Set of Gaussian parameters $\{\boldsymbol{\mu},\Lambda\}$, generally indexed by l_{j_k}.
Θ	Parameters of the Gaussian components of a mixture indexed by l_j.
$\mathbf{K}$	Model under consideration, characterising the number of Gaussian components K_{l_j} per mixture l_j.

Ξ Denotes the complete set of parameters used for the model.

I and O Denotes the density distribution function for the inlier and outlier part of the model respectively.

Φ Density distribution for a mixture at Level 2 of the hierarchy, indexed by l_j.

$\mathscr{G}$ Notation adopted for a Gaussian density distribution

$\mathscr{U}$ Notation adopted for a uniform distribution.

$\mathscr{M}$ Generic notation for a distribution at Level 3: can be either uniform ($\mathscr{U}$)) or Gaussian (G)

5) IIH correction:

χ Indexed by m, corresponds to a IIH polynomial basis function.

pos Indexed by n, corresponds to the spatial location of voxel n.

$\mathbf{c}$, $\mathbf{C}$ Denotes respectively the vector, indexed by m (of size D), of basis coefficients and the corresponding set of such vectors (size M).

6) Mixing weights and atlases:

w Indexed by l_{j_k}, denotes the mixing weight of the component l_{j_k} in mixture l_j.

$\mathbf{w}$ Indexed by l_j, represents the vector of mixing weights $w_{l_{j_k}}$ used to model mixture l_j.

$\mathbf{W}$ Denotes the set of all vectors $\mathbf{w}_{l_j}$ in the model.

a (b) When simply indexed by l_j (l), corresponds to the global mixing weight of class l_j (l) at Level 2 (Level 1). When also indexed by n, corresponds to a voxelwise *a priori* probability.

$\mathbf{a}$ ($\mathbf{b}$) In the context of statistical atlases, corresponds to the vector built from a_{nl_j} (b_{nl}), indexed by n. Otherwise, corresponds to the vector of global mixing weights.

$\mathbf{A}$ ($\mathbf{B}$) Set of vectors $\{\mathbf{a_1}, \cdots, \mathbf{a_N}\}$ ($\{\mathbf{b_1}, \cdots, \mathbf{b_N}\}$) representing the statistical atlases.

$\tilde{}$ When used as a diacritic mark, denotes the relaxed version of the coefficient/atlas.

π When indexed by nl_{j_k}, is defined as $\pi_{nl_{j_k}} = b_{nl}a_{nj}w_{l_{j_k}}$,

$\boldsymbol{\pi}$ Depending on the context, denotes either the set $\{\mathbf{b}, \mathbf{a}, \mathbf{W}\}$ of global mixing weights or the set $\{\mathbf{B}, \mathbf{A}, \mathbf{W}\}$ using the statistical atlases for Level 1 and Level 2.

7) Atlas adaptation and parameters:

$\mathscr{D}$ Denotes the Dirichlet distribution.

$\mathscr{B}$ Denotes the Beta distribution.

δ Indexed by the level of the hierarchy (1 or 2), defines the strength of the prior relaxation process. A lower value represents a stronger relaxation.

κ Directly related to δ. The highest κ, the strongest the relaxation.

$\alpha_{nj}, \boldsymbol{\alpha_n}$ Dirichlet prior parameters at voxel n for the anatomical tissue j, and the corresponding vector of all gathered anatomical features (Level 2).

$\beta_{nj}, \boldsymbol{\beta_n}$ Same as α_{nj} and $\boldsymbol{\alpha_n}$ but at Level 1 of the hierarchy.

p_{nl} Responsibility marginalized over the second level.

p_{nj} Responsibility marginalized over the first level.

G_σ Gaussian kernel with standard deviation σ.

8) MRF notations:

$\mathscr{N}_n$ Denotes the set of von Neumann neighbors of voxel n, *i.e.* the 6 nearest neighbors (east, west, north, south, top and bottom).

$\mathbf{p}_n$ Denotes the vector or responsibilities for all l_{j_k} components. When indexed by $\mathscr{N}_n$, denotes the set of such vectors for the voxels in $\mathscr{N}_n$.

U_{MRF} Energy function related to the current labelling configuration.

$\phi_{nl_{j_k}}$ Abbreviation of $f\left(\mathbf{y}_n \middle| \mathbf{z}_n = \mathbf{e}_{l_{j_k}}, \Xi_\mathbf{K}\right)$.

$\psi_{nl_{j_k}}$ Abbreviation of $\exp\left(-U_{\text{MRF}}\left(\mathbf{e}_{l_{j_k}} \middle| \mathbf{p}^{(t)}_{\mathscr{N}_n}, H\right)\right)$.

H MRF inter-class energy matrix.

9) Constraint over the covariance matrix:

$\Omega_{l_{j_k}}$ Weighted covariance matrix for the Gaussian component l_{j_k} at Level 3.

Ψ Prior over the model covariances.

$\Upsilon_{l_{j_k}}$ Scaling diagonal matrix used to compensate for the log-transformation of the intensities in the prior over the covariances.

10) Model selection:

$\mathbf{v}$ Indexed by l_{j_k}, denotes the vector in the orthogonal decomposition of the covariance matrix $\Lambda_{l_{j_k}}$ associated with the highest eigenvalue.

υ Decimation factor accounting for the proportion of truly independent voxels.

$\text{BIC}(\mathbf{K})$ Bayesian Information Criterion on model $\mathbf{K}$.

$\mathscr{P}(\mathbf{K})$ Penalization function over the model $\mathbf{K}$ used in BIC

corr_r Value of correlation between adjacent voxels in the r-direction.

11) Lesion definition:

$Patho$ Modalities that can be used as indicator of pathology in the case of WMH, that is FLAIR, T2 and PD.

L Set of components potentially considered as lesions.

TL Set of components considered as lesions.

DL Set of possible lesion-related components that need further refinement.

RL Final set of lesion-related components refining DL.

12) Lesion assessment:

∂ Denotes the border of a binarized object.

$\sharp$ Denotes the cardinality of a set.

Ref Refers to the binary lesion segmentation used as reference.

Seg Refers to the binary lesion segmentation evaluated.

FP_c Set of connected elements that belong only to Seg.

FN_c Set of connected elements that belong only to Ref.

TP_c Set of connected elements for which at least one voxel is a true positive.

Seg$_T$ (Ref$_T$)	Seg (Ref) restricted to the connected element T.

Acknowledgments

This work was supported by the Wolfson Foundation, the Faculty of Engineering Science UCL, EPSRC (EP/H046410/1, EP/J020990/1, EP/K005278), the MRC (MR/J01107X/1), the EU-FP7 project VPH-DARE@ IT (FP7-ICT-2011-9-601055), NIHR Queen Square Dementia Biomedical Research Unit, the NIHR UCLH/UCL Biomedical Research Centre (High Impact Initiative) and Alzheimer's Research UK (ARUK). The Dementia Research Centre is supported by ARUK, Brain Research Trust and the Wolfson Foundation. The authors would like to thank the staff and patients of the Vascular Cognitive Impairment Group at UMC Utrecht without whom collection and generation of the diabetes-related data would not be possible.

References

[1] Daniel García-Lorenzo, Simon Francis, Sridar Narayanan, Douglas L Arnold, and D Louis Collins, "Review of automatic segmentation methods of multiple sclerosis white matter lesions on conventional magnetic resonance imaging," *Medical Image Analysis*, vol. 17, pp. 1–18, 2013.

[2] Marko Wilke, Bianca de Haan, Hendrik Juenger, and Hans-Otto Karnath, "Manual, semi-automated, and automated delineation of chronic brain lesions: a comparison of methods," *NeuroImage*, vol. 56, pp. 2038–2046, 2011.

[3] Faiza Admiraal-Behloul, Dominique M J M van den Heuvel, Hans Olofsen, Matthias J P van Osch, Jeroen van der Grond, Mark a. Van Buchem, and Johan H C Reiber, "Fully automatic segmentation of white matter hyperintensities in {MR} images of the elderly," *NeuroImage*, vol. 28, pp. 607–617, 2005.

[4] Marco Battaglini, Mark Jenkinson, and Nicola De Stefano, "Evaluating and Reducing the Impact of White Matter Lesions on Brain Volume Measurements," *Human Brain Mapping*, vol. 33, pp. 2062–2071, 2012.

[5] Philippe Shroeter, Jean-Marc Vesin, Thierry Langenberger, and Reto Meuli, "Robust parameter estimation of intensity distributions for brain magnetic resonance images," *IEEE Transactions on Medical Imaging*, vol. 17, no. 2, pp. 172–186, Apr. 1998.

[6] Koen Van Leemput, Frederik Maes, Dirk Vandermeulen, Alan Colchester, and Paul Suetens, "Automated segmentation of multiple sclerosis lesions by model outlier detection," *IEEE Transactions on Medical Imaging*, vol. 20, no. 8, pp. 677–688, Aug. 2001.

[7] Daniel García-Lorenzo, Sylvain Prima, Arnold L Douglas, D Louis Collins, and Christian Barillot, "Trimmed-Likelihood Estimation for Focal Lesions and Tissue Segmentation in Multisequence {MRI} for Multiple Sclerosis," *IEEE Transactions on Medical Imaging*, vol. 30, no. 8, pp. 1455–1467, Aug. 2011.

[8] Renske de Boer, Henri a. Vrooman, Fedde van der Lijn, Meike W. Vernooij, M. Arfan Ikram, Aad van der Lugt, Monique M B Breteler, and Wiro J. Niessen, "White matter lesion extension to automatic brain tissue segmentation on {MRI}," *NeuroImage*, vol. 45, no. 4, pp. 1151–1161, 2009.

[9] Xavier Lladó, Arnau Oliver, Mariano Cabezas, Jordi Freixenet, Joan C Vilanova, Ana Quiles, Laia Valls, Lluís Ramió-Torrentà, and Àlex Rovira, "Segmentation of multiple sclerosis lesions in brain {MRI} : a review of automated approaches," *Information Sciences*, vol. 186, pp. 164–185, 2012.

[10] Ying Wu, Simon K Warfield, I Leng Tan, William M Wells III, Dominik S Meier, Ronald A van Schijndel, Frederik Barkhof, and Guttmann Charles R.G., "Automated segmentation of multiple sclerosis lesion subtypes with multichannel {MRI}," *NeuroImage*, vol. 32, pp. 1205–1215, 2006.

[11] Yanbo Wango, Joseree Ann Catindig, Saima Hilal, Hock Wei Soon, Eric Ting, Tien Yin Wong, Narayanaswamy Venketasubramanian, Christopher Chen, Anqi Qiu, Yanbo Wang, Joseree Ann Catindig, Saima Hilal, Hock Wei Soon, Eric Ting, Tien Yin Wong, Narayanaswamy Venketasubramanian, Christopher Chen, and Anqi Qiu, "Multi-stage segmentation of white matter hyperintensity, cortical and lacunar infarcts," *Neuroimage*, vol. 60, no. 4, pp. 2379–2388, 2012.

[12] Vamsi Ithapu, Vikas Singh, Christopher Lindner, Benjamin P Austin, Chris Hinrichs, M Carlsson Cynthia, Barbara B Bendlin, and Sterling C Johnson, "Extracting and summarizing white matter hyperintensities using supervised segmentation methods in {A}lzheimer's disease risk and aging studies," Jan. 2014.

[13] David S Wack, Michael G Dwyer, Niels Bergsland, Carol Di Perri, Laura Ranza, Sara Hussein, Deepa Ramasamy, Guy Poloni, and Robert Zivadinov, "Improved assessment of multiple sclerosis lesion segmentation agreement via detection and outline error estimates," *BMC Medical imaging*, vol. 12, no. 17, 2012.

[14] Martha E Payne, Denise L Fetzer, James R MacFall, James M Provenzale, Chistopher E Byrum, and K Ranga R Krishnan, "Development of a semi-automated method for quantification of {MRI} gray and white matter lesions in geriatric subjects," *Psychiatry Research : Neuroimaging*, vol. 115, pp. 63–77, 2002.

[15] Navid Shiee, Pierre-Louis Bazin, Arzu Ozturk, Daniel S Reich, Peter A Calabresi, and Dzung L Pham, "A topology-preserving approach to the segmentation of brain images with multiple sclerosi lesions," *NeuroImage*, vol. 49, no. 2, pp. 1524–1535, 2010.

[16] Laure S Aït-Ali, Sylvain Prima, P Hellier, B Carsin, G Edan, and Christian Barillot, "{STREM}: A robust multidimensional parametric method to segment {MS} lesions in {MRI}," in *MICCAI 2005*. 2005, LNCS 3749, pp. 409–416, Springer International.

[17] Abdel-Ouahab Boudraa, Sidi Mohammed Réda Dehak, Yue-Min Zhu, Chahin Pachai, Yong-Gang Bao, and Jérôme Grimaud, "Automated segmentation of multiple sclerosis lesions in multispectral {MR} imaging using fuzzy clustering," *Computers in Biology and Medicine*, 2000.

[18] Eleftherios Lavdas, Ioannis Tsougos, Stella Kogia, Georgios Gratsias, Patricia Svolos, Violetta Roka, Ioannis V Fezoulidis, and Eftychia Kapsalaki, "T2 {FLAIR} artifacts at 3{T} brain magnetic resonance imaging," *Clinical Imaging*, vol. In press, 2013.

[19] Daryoush Mortazavi, Abbas Z Kouzani, and Hamid Soltanian-Zaheh, "Segmentation of multiple sclerosis lesions in {MR} images : a review," *Diagnostic Neuroradiology*, vol. 54, pp. 299–320, 2012.

[20] Yulian Wolff, Shmuel Miron, Anat Achiron, and Hayit Greespan, "Improved {CSF} classification and lesion detection in {MR} brain images with multiple sclerosis," in *Proceedings of SPIE Medical Imaging 2007 : Image Processing*. 2007, vol. 6512, Pluim, Josien P. W. and Reinhardt, Joseph M.

[21] Guillaume Dugas-Phocion, Miguel Angel Gonzalez Ballester, Christine Lebrun, Stéphane Chanalet, Caroline Bensa, Grégoire Malandain, and Nicolas Ayache, "Hierarchical Segmentation of Multiple Sclerosis Lesions in Multi-Sequence {MRI}," in *International Symposium on Biomedical Imaging: From Nano to Macro (ISBI'04)*, Apr. 2004.

[22] Paul Schmidt, Christian Gaser, Milan Arsic, Dorothea Buck, Annette Förschler, Achim Berthele, Muna Hoshi, Rüdiger Ilg, Volker J Schmid, Claus Zimmer, Bernhard Hemmer, and Mark Mühlau, "An automated tool for detection of {FLAIR}-hyperintense white matter lesions in multiple sclerosis," *NeuroImage*, vol. 59, pp. 3774–3783, 2012.

[23] Balasrinivasa Rao Sajja, Sushmita Datta, Renjie He, Meghana Mehta, Rakesh K. Gupta, Jerry S. Wolinsky, Narayana Ponnada A., and Ponnada a. Narayana, "Unified approach for Multiple Sclerosis Lesion Segmentation on Brain {MRI}," *Ann. Biomed Eng*, vol. 34, no. 1, pp. 142–151, Jan. 2006.

[24] Florence Forbes, Senan Doyle, Daniel Garcia-Lorenzo, Christian Barillot, and Michel Dojat, "Adaptive weighted fusion of multiple {MR} sequences for brain lesion segmentation," in *IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, 2010.

[25] Rita Simoes, Monninghoff Christoph, Martha Dlugaj, Christian Weimar, Isabel Wanke, Anne-Marie van Cappellen van Walsum, and Cornelis Slump, "Automatic segmentation of cerebral white matter hyperintensities using only 3{D FLAIR} images," *Magnetic resonance imaging*, vol. In press, 2013.

[26] April Khademi, Anastasios Venetsanopoulos, and Alan R Moody, "Robust white matter lesion segmentation in {FLAIR} {MRI}," *IEEE*

Transactions on biomedical engineering, vol. 59, no. 3, pp. 860–871, Mar. 2012.

[27] Arthur P Dempster, Nan M Laird, and Donald B Rubin, "Maximum Likelihood from Incomplete Data via the {EM} {A}lgorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977.

[28] Chris Fraley and E Raftery Adrian, "Model-based clustering, discriminant analysis, and density estimation," *Journal of the American Statistical Association*, vol. 97, no. 458, pp. 611–631, June 2002.

[29] José V. Manjón, Jussi Tohka, Gracian García-Martí, José Carbonell-Caballero, Juan J. Lull, Luís Martí-Bonmatí, and Montserrat Robles, "Robust {MRI} Brain Tissue parameter estimation by multistage outlier rejection," *Magnetic resonance in medicine : official journal of the Society of Magnetic Resonance in Medicine / Society of Magnetic Resonance in Medicine*, vol. 59, pp. 866–873, 2008.

[30] Hakon Gudbjartsson and Samuel Patz, "The {R}ician distribution of noisy {MRI} data.," *Magnetic resonance in medicine : official journal of the Society of Magnetic Resonance in Medicine / Society of Magnetic Resonance in Medicine*, vol. 34, no. 6, pp. 910–914, Dec. 1995.

[31] M Jorge Cardoso, Matthew J Clarkson, Gerard R Ridgway, Marc Modat, Nick C Fox, Sebastien Ourselin, and The Alzheimer's Disease Neuroimaging Initiative, "Lo{A}d: A locally adaptive cortical segmentation algorithm," *Neuroimage*, vol. 56, no. 3, pp. 1386–1397, June 2011.

[32] John Ashburner and Karl J Friston, "Unified segmentation," *Neuroimage*, vol. 26, no. 3, pp. 839–851, July 2005.

[33] Rasoul Khayati, Mansur Vafadust, Farzad Towdhidkhah, and S Massood Nabavi, "Fully automatic segmentation of multiple sclerosis lesions in brain {MR FLAIR} images using adaptive mixtures method and {M}arkov random field model," *Computers in Biology and Medicine*, vol. 38, pp. 379–390, 2008.

[34] Oren Freifeld, Hayit Greespan, and Jacob Goldberger, "Multiple Sclerosis Lesion Detection Using Constrained {GMM} and Curve Evolution," *International {J}ournal of {B}iomedical {I}maging*, vol. 2009, pp. 13, 2009.

[35] Carole H Sudre, M Jorge Cardoso, Willem Bouvy, Geert J Biessels, Josephine Barnes, and Sébastien Ourselin, "{B}ayesian Model Selection for Pathological Data," in *MICCAI 2014*, P Golland Et al., Ed. 2014, LNCS 8673, pp. 323–330, Springer International.

[36] Jia Li, "Clustering based on a multilayer mixture model," *Journal of Computational and Graphical Statistics*, vol. 14, no. 3, pp. 547–566, 2004.

[37] Koen Van Leemput, Frederik Maes, Dirk Vandermeulen, and Paul Suetens, "Automated model-based bias field correction of {MR} images of the brain," *Medical Imaging, IEEE Transactions on*, vol. 18, no. 10, pp. 885–896, Oct. 1999.

[38] M Jorge Cardoso, Andrew Melbourne, Giles S Kendall, Marc Modat, Nicola J Robertson, Neil Marlow, and S Ourselin, "Ada{PT}: an adaptive preterm segmentation algorithm for neonatal brain {MRI}," *NeuroImage*, vol. 65, pp. 97–108, 2013.

[39] Navid Shiee, Pierre-Louis Bazin, Jennifer Cuzzocreo, Ari Blitz, and Dzung L Pham, "Segmentation of Brain Images using adaptive atlases with application to ventriculomegaly," in *Informatic Processing Medical Imaging*, 2011, vol. 22, pp. 1–12.

[40] Jun Zhang, "The Mean Field Theory in {EM} procedures for {M}arkov {R}andom {F}ields," *IEEE Transactions on signal processing*, vol. 40, no. 10, pp. 2570–2583, Oct. 1992.

[41] Sylvain Prima, Nicolas Ayache, Tom Barrick, and Neil Roberts, "Maximum Likelihood estimation of the bias field in {MR} brain images: investigating different modelings of the imaging process," in *MICCAI 2001*, 2001, pp. 811–819.

[42] Xinhua Zhuang, Yan Huang, K Palaniappan, and Yunxin Zhao, "Gaussian mixture density modeling, decomposition, and applications," in *IEEE Transactions on Image Processing*, Sept. 1996, vol. 5, pp. 1293–1302.

[43] Zujun Hou, "A review on {MR} Image Intensity Inhomogeneity Correction," *International {J}ournal of {B}iomedical {I}maging*, vol. 2006, pp. 1–11, Feb. 2006.

[44] S Sanjay Gopal and Thomas J Hebert, "Bayesian pixel classification using spatially variant finite mixtures and the generalized {EM} algorithm," in *IEEE Transactions on Image Processing*, July 1998, vol. 7, pp. 1014–1027.

[45] Mariano Cabezas, Arnau Oliver, Xavier Lladó, and Jordi Freixenet, "A review of atlas-based segmentation for magnetic resonance brain images," *Computer Methods and Programs in Biomedicine*, vol. 104, pp. 158–177, Dec. 2011.

[46] Hichem Snoussi and Ali Mohammad-Djafari, "Penalized maximum likelihood for multivariate {G}aussian mixture," in *AIP Conference proceedings*, 2002, vol. 617, pp. 36–46.

[47] Naonori Ueda, Ryohei Nakano, Zoubin Ghahramani, and Geoffrey E Hinton, "{SMEM} {A}lgorithm for {M}ixture {M}odels," *Neural Comput.*, vol. 12, no. 9, pp. 2109–2128, Sept. 2000.

[48] Yan Li and Lei Li, "A split and merge {EM} algorithm for color image segmentation," in *Intelligent Computing and Intelligent Systems, 2009. ICIS 2009. IEEE International Conference on*, Nov. 2009, vol. 4, pp. 395–399.

[49] Sylvia Richardson and Peter J Green, "On {B}ayesian {A}nalysis of {M}ixtures with an {U}nknown {N}umber of {C}omponents," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 59, no. 4, pp. pp. 731–792, 1997.

[50] Adrian R Groves, Christian F Beckmann, Steve M Smith, and Mark W Woolrich, "Linked independent component analysis for multimodal data fusion.," *NeuroImage*, vol. 54, no. 3, pp. 2198–2217, 2011.

[51] Detlev Uhlenbrock and S Sehlen, "The value of {T1}-weighted images in the differentiation between {MS}, white matter lesions, and subcortical arteriosclerotic encephalopathy (SAE)," *Neuroradiology*, 1989.

[52] Wang Zhan, Yu Zhang, Suzanne G Mueller, Peter Lorenzen, Stathis Hadjidemetriou, Norbert Schuff, and Michael W Weiner, "Characterization of white matter degeneration in elderly subjects by magnetic resonance diffusion and FLAIR imaging correlation," *NeuroImage*, vol. 48, pp. 758–765, 2009.

[53] Joanna M Wardlaw, Eric E Smith, G J Biessels, Charlotte Cordonnier, Franz Fazekas, Richard Frayne, Richard I Lindley, John T O'Brien, Frederik Barkhof, Oscar R Benavente, Sandra E Black, Carol Brayne, Monique M B Breteler, Hugues Chabriat, Charles DeCarli, Frank-Erik de Leeuw, Fergus Doubal, Marco Duering, Nick C Fox, Steven Greenberg, Vladimir Hachinski, Ingo Kilimann, Vincent Mok, Robert van Oostenbrugge, Leonardo Pantoni, Oliver Speck, Blossom C M Stephan, Stefan Teipel, Viswanathan Anand, David Werring, Christopher Chen, Colin Smith, Mark A van Buchem, Bo Norrving, Philip B Gorelick, and Martin Dichgans, "Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration," *Lancet Neurology*, vol. 12, pp. 822–838, 2013.

[54] J Mazziotta, A Toga, A Evans, P Fox, J Lancaster, K Zilles, R Woods, T Paus, G Simpson, B Pike, C Holmes, L Collins, P Thompson, D MacDonald, M Iacoboni, T Schormann, K Amunts, N Palomero-Gallagher, S Geyer, L Parsons, K Narr, N Kabani, G Le Goualher, D Boomsma, T Cannon, R Kawashima, and B Mazoyer, "A probabilistic atlas and reference system for the human brain: International Consortium for Brain Mapping (ICBM).," *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, vol. 356, no. 1412, pp. 1293–1322, Aug. 2001.

[55] Guillermo Gallego, Carlos Cuevos, Raul Mohedano, and Narciso Garcia, "On the {M}ahalanobis distance classification criterion for multidimensional normal distribution," *Signal Processing, IEEE Transactions on*, vol. 61, no. 17, pp. 4387–4396, 2013.

[56] M Jorge Cardoso, Kelvin Leung, Marc Modat, Shiva Keihaninejad, David Cash, Josephine Barnes, Nick C Fox, and Sebastien Ourselin, "{STEPS}: Similarity and Truth Estimation for Propagated Segmentations and its application to hippocampal segmentation and brain parcelation," *Medical Image Analysis*, vol. 17, pp. 671–684, 2013.

[57] Andrés Ortiz, Juan M Górriz, Javier Ramírez, and Francisco J Martinez-Murcia, "Automatic {ROI} selection in structural brain {MRI} using {SOM} 3{D} projection," *PLoS ONE*, vol. 9, no. 4, 2014.

[58] Remi K S Kwan, Alan C. Evans, and Bruce Pike, "MRI simulation-based evaluation of image-processing and classification methods," *IEEE Transactions on Medical Imaging*, vol. 18, no. 11, pp. 1085–1097, 1999.

[59] Martin A Styner, Joohwi Lee, Brian Chin, Matthew S Chin, Olivier Commowick, Hoai-Huong Tran, Silva Markovic-Plese, Valerie Jewells, and Simon K Warfield, "{3D} segmentation in the clinic: a grand challenge {II}: {MS} lesion segmentation," in *The MIDAS Journal - MS Lesion Segmentation*, 2008.

[60] Hugo Vrenken, Mark Jenkinson, M A Horsfield, Marco Battaglini, Ronald A van Schijndel, Egill Rostrup, Jeroen J G Geurts, E Fisher, A Zijdenbos, John Ashburner, David H Miller, Massimo Filippi, Franz Fazekas, Marco Rovaris, Àlex Rovira, Frederik Barkhof, Nicola

De Stefano, and MAGNIMS Study Group, "Recommendations to improve imaging and analysis of brain lesion load and atrophy in longitudinal studies of multiple sclerosis," *J. Neurology*, vol. 260, pp. 2458–2471, 2013.

[61] Joseph a. Maldjian, Christopher T. Whitlow, Baidya N. Saha, G. Kota, C. Vandergriff, Emma M. Davenport, J. Divers, Barry I. Freedman, and Donald W. Bowden, "Automated white matter total lesion volume segmentation in diabetes," *American journal of neuroradiology*, vol. 34, pp. 2265–2270, 2013.

[62] Niousha Bolandzadeh, Jennifer C Davis, Roger Tam, Todd C Handy, and Teresa Liu-Ambrose, "The association between cognitive function and white matter lesion location in older adults: a systematic review," *BioMed Central Neurology*, vol. 12, no. 126, 2012.

[63] Peter Kapeller, Stefan Ropele, Christian Enzinger, Theresa Lahousen, Siegrid Strasser-Fuchs, Reinhold Schmidt, and Franz Fazekas, "Discrimination of white matter lesions and multiple sclerosis plaques by short echo quantitative {1H}-magnetic resonance spectroscopy," *J. Neurology*, vol. 252, pp. 1229–1234, 2005.

[64] Frederik Barkhof, Massimo Filippi, David H Miller, Philip Scheltens, Adriana Campi, Chris H Polman, Giancarlo Comi, Herman J Ader, Nick Losseff, and Jacob Valk, "Comparison of {MRI} criteria at first presentation to predic conversion to clinically definite multiple sclerosis," *Brain*, vol. 120, pp. 2059–2069, 1997.

[65] The LADIS Study group, "2001-2011: a decade of the {LADIS} ({L}eukoaraiosis {A}nd {DIS}ability) Study: what have we learned about white matter changes and small-vessel disease," *Cerebrovascular diseases*, vol. 32, pp. 577–588, 2011.

[66] J de Bresser, Audrey M Tiehuis, Esther van den Berg, Yaël D Reijmer, Cynthia Jongen, L Jaap Kappelle, W P Mali, M A Viergever, and Geert J Biessels, "Progression of cerebral atrophy and white matter hyperintensities in patients with type 2 diabetes," *Diabetes Care*, vol. 33, no. 6, pp. 1309–1314, June 2010.

[67] Faith M Gunning-Dixon, Adam M Brickman, Janice C Cheng, and George S Alexopoulos, "Aging of cerebral white matter : a review of {MRI} findings," *International Journal of Geriartric Psychiatry*, 2008.

[68] Massimo Filippi, M I Gawne-Cain, C Gasperini, J H van Waesberghe, Jérôme Grimaud, Frederik Barkhof, M P Sormani, and David H Miller, "Effect of training and different measurement strategies on the reproducibility of brain {MRI} lesion load measurements in multiple sclerosis," *Neurology*, vol. 50, pp. 238–244, 1998.

[69] Marius de Groot, Benjamin F J Verhaaren, Renske de Boer, Stefan Klein, Albert Hofman, Aad van der Lugt, M Arfan Ikram, Wiro J Niessen, and Meike W Vernooij, "Changes in normal appearing white matter precede development of white matter lesions," *Stroke : Journal of the american heart association*, vol. 44, pp. 1037–1042, 2013.

[70] Elisabeth C W Van Straaten, Franz Fazekas, Egill Rostrup, Philip Scheltens, Reinhold Schmidt, Leonardo Pantoni, Domenico Inzitari, Gunhild Waldemar, Timo Erkinjuntti, Riita Mäntylä, Lars-Olof Olof Wahlund, and Frederik Barkhof, "Impact of white matter hyperintensities scoring method on correlations with clinical data: the {LADIS} study," *Stroke : Journal of the american heart association*, vol. 37, pp. 836–840, 2006.

[71] Rubén Cárdenes, Rodrigo de Luis-García, and Meritxell Bach-Cuadra, "A multidimensional segmentation evaluation of medical image data," *Computer Methods and Programs in Biomedicine*, vol. 96, pp. 108–124, 2009.

[72] Keith M Hulsey, Mohit Gupta, Kevin S King, Ronald M Peshok, Anthony R Whittermore, and Roderick W McColl, "Automated quantification of white matter disease extent at {3T} : Comparison with volumetric readings," *Journal of magnetic resonance imaging*, vol. 36, pp. 305–311, 2012.

[73] Kok Haur Ong, Dhanesh Ramachandram, Rajeswari Mandava, and Ibrahim Lutfi Shuaib, "Automatic white matter lesion segmentation using an adaptive outlier detection method," *Magnetic resonance imaging*, vol. 30, no. 6, pp. 807–823, 2012.

[74] Thomas Samaille, Olivier Colliot, Didier Dormont, and Marie Chupin, "Automatic segmentation of age-related white matter changes on {FLAIR} images: methods and multicentre validation," in *IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, 2011.

[75] Bilwaj Gaonkar, Guray Erus, Nick Bryan, and Christos Davatzikos, "Automated segmentation of brain lesions by combining intensity and spatial information," in *ISBI'10 Proceedings*, 2010.

[76] Ki Woong Kim, James R MacFall, and Martha E Payne, "Classification of white matter lesions on magnetic resonance imaging in the elderly," *Biol. Psychiatry*, vol. 64, no. 4, pp. 273–290, Aug. 2008.

[77] Jesper Hagemeier, Jeroen J G Geurts, and Robert Zivadinov, "Brain iron accumulation in aging and neurodegenerative disorders," *Expert Reviews Neurotherapeutics*, vol. 12, pp. 1467–1480, 2012.

[78] Stefan Ropele, Wolter de Graaf, Michael Khalil, Mike P Wattjes, Christian Langkammer, Maria A Rocca, Àlex Rovira, Jacqueline Palace, Frederik Barkhof, Massimo Filippi, and Franz Fazekas, "{MRI} asssessment of iron deposition in multiple sclerosis," *Journal of magnetic resonance imaging*, vol. 34, pp. 13–21, 2011.

[79] Pauline Maillard, Owen Carmichael, Danielle Harvey, Evan Fletcher, Bruce Reed, Dan Mungas, and Charles DeCarli, "{FLAIR} and Diffusion {MRI} signals are independent predictors of white matter hyperintensities," *American journal of neuroradiology*, vol. 34, pp. 54–61, Jan. 2013.