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

1Copyright (c) 2010 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending a request to pubs-permissions@ieee.org.
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 (Bayesian Model Selection) [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 three-fold: 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 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 three-fold: 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.

considered Gaussian components of the model.

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, \( \mathcal{Y} = \{y_1, \ldots, y_N\} \) denotes the set of log-transformed normalized intensities indexed by \( n, N \) being the total number of observations. The feature vector \( y_n \) 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 \( \mathcal{G}(\theta_j) \). Considering \( J \) Gaussian instances in the mixture, the set of parameters to optimize is composed of \( \Theta_j = \{\theta_1, \ldots, \theta_J\} \), where \( \theta_j = \{\mu_j, \Lambda_j\} \), with \( \mu_j \) and \( \Lambda_j \) being respectively the mean and covariance matrix of the \( j^{th} \) component of the mixture. The vector of weights for the mixture is \( \pi \) whose \( j^{th} \) component \( \pi_j \) 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 \( \Xi_j = \{\Theta_j, \pi\} \) 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 inner layer (I), modelling the healthy tissues, and to the outer layer (O), related to the unexpected observations, such that
   \[
   f(y_n|\Xi_K) = b_I f(y_n|\Xi_K) + b_O f(y_n|\Xi_K),
   \]
   with \( b_I + b_O = 1 \) and \( b_I \geq 0 \), and introducing \( b \) the vector formed with these parameters. Note that \( O(y_n|\Xi_K) \) 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 inner or outer voxel belongs to WM, GM, CSF or other non-brain (NB) tissues). The number of anatomical classes \( J \) is considered the same for both the inner and outer classes since the model is built under an assumption of symmetry, simplifying \( J_J = J_O = J \). The distribution is thus:
   \[
   f(y_n|\Xi_K) = \sum_{l=1}^{J} b_l \sum_{j=1}^{J} a_{lj} \Phi(y_n|\Theta_{lj}),
   \]
   where \( b_l, a_{lj} \) and \( \Phi(y_n|\Theta_{lj}) \) are respectively the mixing weight of \( l \), the class weight for \( j \) and the likelihood of the data at voxel \( n \) for the tissue class \( j \). At this point, \( a \) denotes the vector of mixing weights \( a_{lj} \) satisfying \( \sum_{j=1}^{J} a_{lj} = 1 \) and \( a_{lj} \geq 0 \), \( \forall l \in \{I,O\} \).

3) The third level (Level 3), indexed by \( k \), characterizes the multiple intensity clusters of each inner or outer 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 \( M \), that can be Gaussian (\( \mathcal{G} \)) and/or uniform (\( \mathcal{U} \)) such that
   \[
   \Phi(y_n|\Theta_{lj}) = \sum_{k=1}^{K_{lj}+1} w_{ljk} M(y_n|\theta_{ljk}),
   \]
   where \( K_{lj} \) is the number of Gaussian components in class \( j \), \( w_{ljk} \) is the mixing proportion (\( \geq 0 \)) of class \( j \) and \( \theta_{ljk} \) are the corresponding Gaussian parameters. The uniform distribution in each class \( j \) is only parameterized by the mixing coefficient \( w_{ljk} \). The mixing coefficients for class \( j \) are gathered in the vector \( w_{lj} \), with \( W \) being the set of all such vectors that satisfy \( \sum_{k=1}^{K_{lj}+1} w_{ljk} = 1 \), \( \forall \in \{I,O\} \) and \( \forall \in \{1,\cdots,J\} \).

Adopting the notation \( \pi_{ljk} = b_l a_{lj} w_{ljk} \) with \( \pi = \{b,a,W\} \) the set of \textit{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:

\[
\begin{align*}
  f(\mathcal{Y}|\Xi_K) &= \prod_{t=1}^{n} \prod_{j=1}^{J} \sum_{k=1}^{K_{lj}+1} \left( \pi_{ljk} \cdot M(y_n|\theta_{ljk}) \right),
  \\
  f(\mathcal{Y},\mathcal{Z}|\Xi_K) &= f(\mathcal{Y}|\Xi_K) f(\mathcal{Z}|\Xi_K) f(\mathcal{Z})
\end{align*}
\]

An example of a possible hierarchical model is displayed in Figure 1, where at Level 1 \( l \) takes the values for \( O \) 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 \( \theta_{lj} \); secondly, the contribution \( \pi_{ljk} \) of each distribution to the overall observation model; and thirdly and most importantly, the number of Gaussian components \( K_{lj} \) necessary to describe the underlying distribution of tissue class \( j_l \).

3) EM algorithm: The Expectation-Maximization algorithm introduced in [27] is commonly used for the optimization of \( \Xi_K \). Introducing the labelling configuration, denoted by \( \mathcal{L} = \{z_1,\cdots,z_N\} \), the complete data is then defined as \( \mathcal{X} = \{\mathcal{Y},\mathcal{Z}\} \). Here, \( z_n \), supporting label \( \ell \), is defined as \( e_\ell \) vector of the canonical basis, i.e. the unity vector with component \( \ell \) equal to 1 and all the others to 0. The conditional distribution of the complete data given the parameters is expressed as:

\[
\begin{align*}
  f(\mathcal{X}|\Xi_K) &= f(\mathcal{Y},\mathcal{Z}|\Xi_K) f(\mathcal{Y}|\mathcal{Z}) \\
  &= \prod_{n=1}^{N} \prod_{l \in I} \prod_{j=1}^{J} \prod_{k=1}^{K_{lj}+1} \left[ \pi_{ljk} \cdot M(y_n|\theta_{ljk}) \right]^{z_{nlk}}
\end{align*}
\]

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 (\( \mathbb{E} \)) of the log of the conditional distribution of the complete data given the parameters obtained at iteration \( t \), i.e.:
The maximization step or M-step, consists in the maximization of the $D \left( \Xi_K \parallel \Xi_K^{(i)} \right)$ with respect to $\Xi_K$.

**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 $\chi_m$ given the spatial location $\text{pos}_n$, $\sum_{m=1}^{M} c_m \chi_m (\text{pos}_n)$. When considering the log-intensities $\mathcal{Y}$, the bias field becomes additive and the parameters $c_m$, vectors of size $D$, can then be progressively optimized within the EM framework. In the following, $y_n^{(t)}$ denotes the corrected intensity feature at voxel $n$ at iteration $t$ with:

$$y_n^{(t)} = y_n - \sum_{m=1}^{M} c_m \chi_m (\text{pos}_n)$$

Denoting $C = \{ c_1, \ldots, c_M \}$ the set of vectors of linear coefficients used to model the magnetic field inhomogeneity, the set of parameters to optimize becomes $\Xi_K = \{ \Theta_K, \pi, C \}$

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 $\pi_j$ 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, $\pi$ 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 $b_n$ (resp. $a_n$). The conditional distribution becomes:

$$f(\mathcal{Y} \parallel \Xi_K) = \prod_{n=1}^{N} \prod_{j=1}^{J} \prod_{k=1}^{\kJ} \prod_{l=1}^{\kJ} \left[ \pi_{nkj} \mathcal{M} \left( y_n, \theta_{nkj} \right) \right]^{\pi_{nkj}}$$

with $\pi_{nkj} = b_n a_{nj} w_{jk}$. Considering a symmetric model at Level 2, the atlases are defined so that $\forall n, a_{nj} = a_{nOj} = a_{nj}$. The responsibilities are now defined as:

$$P_{nj}^{(t+1)} = \frac{b_n a_{nj} w_{jk}^{(t)} \mathcal{M} \left( y_n, \theta_{nkj}^{(t)} \right)}{\sum_{j' \in J, \O} \sum_{k' = 1}^{K_{j'}} \sum_{l' = 1}^{L} b_{n'} a_{nj}^{(t)} \mu_{n}^{(t)} \mathcal{M} \left( y_n, \theta_{nkj}^{(t)} \right)}$$

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 $\mathcal{D}$. Now denoting $\pi = \{ \mathbf{B}, \mathbf{A}, \mathbf{W} \}$, the mixing coefficients follow the distribution:

$$f(\pi) = f(\mathbf{B}) f(\mathbf{A})$$

$$f(\pi) = \prod_{n=1}^{N} \mathcal{D} \left( b_n, \beta_n \right) \mathcal{D} \left( \alpha_n \right)$$

where $\mathcal{B}$ is a Beta function and $\beta_n$, $\alpha_n$ the vectors of Dirichlet prior parameters for voxel $n$ such that $b_{nj} = 1 + \delta_1 b_{nj}$ and $\alpha_{nj} = 1 + \delta_2 a_{nj}$, where $\delta_1$ and $\delta_2$ 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:

$$f(\mathcal{Y} \parallel \Theta_K, \pi, C) \cdot f(\Theta_K, \pi, C)$$

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:

$$\tilde{b}_{nl}^{(t+1)} = \frac{\delta_1 b_{nl} + P_{nl}^{(t+1)}}{\delta_1 + 1}$$

$$\tilde{a}_{nj}^{(t+1)} = \frac{\delta_2 a_{nj} + P_{nj}^{(t+1)}}{\delta_2 + 1}$$

with:

$$P_{nl}^{(t+1)} = \sum_{j=1}^{K_{j}+1} \sum_{k=1}^{K_{k}} P_{nk}^{(t+1)}$$

$$P_{nj}^{(t+1)} = \sum_{l=1, \O} \sum_{k=1}^{K_{k}} P_{nk}^{(t+1)}$$

Since the probabilistic atlases are supposed to be smooth but the responsibilities are not, a Gaussian Kernel $G_{\sigma}$ with standard deviation $\sigma$ 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:

$$\tilde{b}_{nl}^{(t+1)} = (1 - \kappa_1) b_{nl} + \kappa_1 (G_{\sigma} * P_{nl}^{(t+1)})$$

$$\tilde{a}_{nj}^{(t+1)} = (1 - \kappa_2) a_{nj} + \kappa_2 (G_{\sigma} * P_{nj}^{(t+1)})$$
where $\kappa = 1/\delta + 1$ and $\ast$ 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 $\mathcal{A}_n$ represents the set of von Neumann neighbors of voxel $n$:

$$f(\mathbf{z} | H, \pi) \propto e^{-U_{\text{MRF}}(\mathbf{z} | H)} = \prod_{n=1}^{N} \prod_{j=1}^{J} \prod_{k=1}^{K_{j}} \prod_{l=1}^{L_{j}} \pi_{\mathbf{e}_{n}}^{L_{j}} e^{-U_{\text{MRF}}(\mathbf{e}_{n} | H)}$$

where $\pi_{\mathbf{e}_{n}}^{L_{j}}$ represents the set of von Neumann neighbors of voxel $n$:

$$f(\mathbf{z} | \pi) \propto e^{-U_{\text{MRF}}(\mathbf{z} | H)}$$

leading to the following responsibilities

$$p_{nl_{jk}}^{(t+1)} \approx \frac{\phi_{nl_{jk}}^{(t)} \pi_{nl_{jk}}^{(t)} \psi_{nl_{jk}}^{(t)}}{\sum_{l' \in I} \sum_{k' \in K} \phi_{nl'_{jk}}^{(t)} \pi_{nl'_{jk}}^{(t)} \psi_{nl'_{jk}}^{(t)}}$$

where we adopt the notations

$$\phi_{nl_{jk}}^{(t)} = f \left( \mathbf{y}_{n} | \mathbf{z}_{n} = \mathbf{e}_{l_{jk}}, \mathbf{Z} \right)$$

$$\psi_{nl_{jk}}^{(t)} = \exp \left( -U_{\text{MRF}}(\mathbf{e}_{n} | \mathbf{p}_{nl_{jk}}, H) \right)$$

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

$$f_{\Psi} \left( A_{l_{jk}} \right) \propto \frac{|\Psi|^{-\frac{N}{2}}}{|A_{l_{jk}}|^{-\frac{N+d+1}{2}}} \exp \left[ -\frac{1}{2} \text{Tr} \left( A_{l_{jk}}^{-1} \sum_{j=1}^{J} \Omega_{jk} \Psi \sum_{j=1}^{J} \Omega_{jk} \right) \right]$$

where $\Omega_{jk} = \text{tr} \left( \mathbf{Y}_{jk} \right)$ is the trace of the matrix, $\Psi$ is a positive definite matrix and $\mathbf{Y}_{jk}$ is a scaling diagonal matrix, such that $\mathbf{Y}_{jk}(d,d) = 1/\exp(\mu_{jk}^{(d)})$ 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 $\Psi$ is optimized as a model parameter and the update of the covariance matrix is modified into

$$\Omega_{jk}^{(t+1)} = \frac{\sum_{n \in I} \sum_{j=1}^{J} \sum_{k=1}^{K_{j}} \mathbf{Y}_{jk} \Omega_{jk}^{(t+1)} \mathbf{Y}_{jk}}{N \cdot \sum_{n \in I} \sum_{j=1}^{J} \sum_{k=1}^{K_{j}} 1}$$

$$\Lambda_{l_{jk}}^{(t+1)} = \frac{\sum_{n \in I} \sum_{j=1}^{J} \sum_{k=1}^{K_{j}} p_{nl_{jk}}^{(t+1)}}{\sum_{n \in I} \sum_{j=1}^{J} \sum_{k=1}^{K_{j}} 1}$$

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

This work is licensed under a Creative Commons Attribution 3.0 License. For more information, see http://creativecommons.org/licenses/by/3.0/.
5) Summary of the modified EM: Using the adaptation of the population atlases described in II-B2, \( \tilde{\pi} = \{ \tilde{B}, \tilde{A}, \tilde{W} \} \) represents the sets of adapted atlases (\( \tilde{B} \) and \( \tilde{A} \)) for the first and the second level and the set of weights (\( \tilde{W} \)) attributed to the classical mixtures in Level 3. The conditional definition of the labelling \( f(\mathcal{Z}|H, \tilde{\pi}) \) is then proportional to

\[
N \prod_{n=1}^{N} \prod_{k=1}^{K} \prod_{j=1}^{J} \prod_{l_j=1}^{L_j} \left[ \tilde{b}_{nl} \tilde{a}_{nkj} \tilde{w}_{lkj} \right]^{\tilde{c}_{nlj}} \psi_{nlj}
\]

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

\[
P_{nlkj}^{(t+1)} = \frac{\phi_{nlj}^{(t)} \tilde{a}_{nkj} \tilde{w}_{lkj}^{(t)} \tilde{w}_{nl}^{(t)}}{\sum_{l_j' \in L_j \setminus l_j} \sum_{k=1}^{K} \phi_{nlj}^{(t)} \tilde{a}_{nkj} \tilde{w}_{lkj}^{(t)} \psi_{nlj}'^{(t)}}
\]

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

\[
\begin{align*}
\tilde{w}_{lkj}^{(t)} &= \frac{\sum_{n=1}^{N} P_{nlkj}^{(t)} \tilde{c}_{nlj}^{(t)}} {\sum_{k=1}^{K} \sum_{n=1}^{N} P_{nlkj}^{(t)} \tilde{c}_{nlj}^{(t)}} \\
\tilde{\mu}_{lkj}^{(t)} &= \frac{\sum_{n=1}^{N} P_{nlkj}^{(t)} \tilde{y}_{n}^{(t)}} {\sum_{n=1}^{N} P_{nlkj}^{(t)} \tilde{y}_{n}^{(t)}} \\
\tilde{\Omega}_{lkj}^{(t)} &= \frac{\sum_{n=1}^{N} P_{nlkj}^{(t)} \left( \tilde{y}_{n}^{(t)} - \tilde{\mu}_{lkj}^{(t)} \right) \left( \tilde{y}_{n}^{(t)} - \tilde{\mu}_{lkj}^{(t)} \right)^T} {\sum_{n=1}^{N} P_{nlkj}^{(t)}}
\end{align*}
\]

where \( \tilde{y}_{n}^{(t)} \) stands for the IIH-corrected intensities and \( \tilde{\Omega}_{lkj}^{(t)} \) is the weighted covariance matrix used in the update of the Gaussian covariance \( \tilde{\Lambda}_{lkj}^{(t)} \).

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 \( K_j \) 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, \( l_{j_1} \) and \( l_{j_2} \), into a single Gaussian distribution \( l_j \). 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:

\[
\text{KLD}(P_1 \parallel P_2) = \sum_{y} P_1(y) \log \left( \frac{P_1(y)}{P_2(y)} \right) + P_2(y) \log \left( \frac{P_2(y)}{P_1(y)} \right)
\]

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 \( l_{j_1} \) and \( l_{j_2} \) into a single component \( l_j \), the initial Gaussian parameters are expressed as:

\[
\begin{align*}
\tilde{w}_{l_j} &= \tilde{w}_{l_{j_1}} + \tilde{w}_{l_{j_2}} \\
\tilde{\mu}_{l_j} &= \frac{\tilde{w}_{l_{j_1}} \tilde{\mu}_{l_{j_1}} + \tilde{w}_{l_{j_2}} \tilde{\mu}_{l_{j_2}}}{\tilde{w}_{l_{j_1}} + \tilde{w}_{l_{j_2}}} \\
\tilde{\Lambda}_{l_j} &= \frac{\tilde{w}_{l_{j_1}} \tilde{\Lambda}_{l_{j_1}} + \tilde{w}_{l_{j_2}} \tilde{\Lambda}_{l_{j_2}}}{\tilde{w}_{l_{j_1}} + \tilde{w}_{l_{j_2}}}
\end{align*}
\]

When splitting a Gaussian component \( l_j \) into two components \( l_{j_1} \) and \( l_{j_2} \), we adopt the implementation described by Li et al. [48], which follows Richardson et al. [49]. Introducing \( v_{l_{j}} \) as the vector corresponding to the highest eigenvalue from the orthogonal factorization of the covariance matrix \( \tilde{\Lambda}_{l_{j}} \) and setting the free parameters to 0.5, the initial parameters for the new Gaussian components are:

\[
\begin{align*}
\tilde{w}_{l_{j_1}} &= \tilde{w}_{l_{j_2}} = 0.5 \cdot \tilde{w}_{l_j} \\
\tilde{\mu}_{l_{j_1}} &= \tilde{\mu}_{l_{j_2}} - 0.5 \cdot v_{l_{j}} \\
\tilde{\mu}_{l_{j_2}} &= \tilde{\mu}_{l_{j_1}} + 0.5 \cdot v_{l_{j}} \\
\tilde{\Lambda}_{l_{j_1}} &= \tilde{\Lambda}_{l_{j_2}} = \tilde{\Lambda}_{l_j} - 0.25 \cdot v_{l_{j}} v_{l_{j}}^T
\end{align*}
\]

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 unmodelled 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 \( \mathcal{N} \). 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 \( \hat{K} \) and expressed by

\[
\text{BIC}(K) = \nu \log(f(\mathcal{Y}|\Xi_K)) - \mathcal{P}(K),
\]

is here used as an objective function. The BIC penalizes the log-likelihood of the model according to the penalization function

\[
\mathcal{P}(K) = \sum \sum J_{ij} \left( \frac{(D+1)D}{2} + 1 \right) \log(N \cdot \nu),
\]

which depends on the number of free parameters being optimized. The decimation factor \( \nu \) is introduced into the BIC cost function to compensate for the lack of spatial independence between the observed samples, i.e. \( \nu \) represents the proportion of voxels that can be considered as independent [50]. The decimation factor is defined as:

\[
\nu = \prod_{r \in \{x,y,z\}} 0.9394^{\text{FWHM}_r} \text{ with } \text{FWHM}_r^2 = \frac{-2 \log 2}{\log(\text{corr}_r)}
\]

where \( \text{corr}_r \) 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 \( \text{BIC}(K) \) 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 \( \Xi_K \), 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 \( \text{List}_{\text{SM}} \): 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 \( \text{List}_{\text{SM}} \) 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 \( \text{List}_{\text{SM}} \) 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.

**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 correc-
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2015.2419072, IEEE Transactions on Medical Imaging

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:

\[
l_{jk} \in L \text{ if } \begin{cases} l_j = O_{\text{WM}} \\
\text{or } \mu_{l_j}^{\text{Patho}} > \mu_{O_{\text{WM}}}^{\text{Patho}}, \end{cases}
\]

with \(\text{Patho} = \{\text{FLAIR, T2, PD}\}\) 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

\[
DL = \left\{ s \in L \mid \mu_s^{(T1)} < \mu_{O_{\text{GM}}}^{(T1)} \right\}
\]

\(TL = L \backslash DL\)

The refined set of lesion \(RL\) is then defined as:

\[
RL = TL + \left\{ s \in DL \mid \mu_s^{(T2)} \geq \mu_{TL}^{(T2)} \right\}
\]

where \(\mu_{TL}\) 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 \textit{a priori} statistical information is known about the location of the outliers. Therefore, constant priors over the image are initially used so that \(\forall n, b_{nO} = b_0\) and

This work is licensed under a Creative Commons Attribution 3.0 License. For more information, see http://creativecommons.org/licenses/by/3.0/.
\[ b_{\text{nl}} = 1 - b_{\text{O}}. \]

The value for \( b_{\text{O}} \) 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 \( \mathbf{K}^{(\text{ini})} \) is initialized as \( K_j = \mathbf{1} \) and \( K_{\text{O}} = 0 \), 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 \( \mathbf{B} \) and \( \mathbf{A} \) 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 \( \mathbf{H} \) containing the neighborhood energy cliques for the MRF is defined as:

\[
H(l_j, l_j') = \begin{cases} 
0 & \text{if } j = j' \\
0.15 & \text{otherwise}
\end{cases}
\]

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](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/](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://www.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

<table>
<thead>
<tr>
<th>Name</th>
<th>Equation</th>
<th>Best</th>
</tr>
</thead>
<tbody>
<tr>
<td>DSC</td>
<td>$100 \times \frac{2 \times</td>
<td>\text{Ref} \cap \text{Seg}</td>
</tr>
<tr>
<td>VD</td>
<td>$100 \times \left</td>
<td>1 - \frac{</td>
</tr>
<tr>
<td>FPR</td>
<td>$100 \times \frac{</td>
<td>\text{Seg}</td>
</tr>
<tr>
<td>TPR</td>
<td>$100 \times \frac{</td>
<td>\text{Ref}</td>
</tr>
<tr>
<td>FNR</td>
<td>$\sum_{s \in \text{Seg} \cap \text{Ref}} \min d(s,r) + \sum_{r \in \text{Ref} \cap \text{Seg}} \min d(s,r)$</td>
<td>0(%)</td>
</tr>
<tr>
<td>AvDist</td>
<td>$\frac{\sum_{s \in \text{Seg} \cap \text{Ref}} \min d(s,r)}{</td>
<td>\text{Ref}</td>
</tr>
<tr>
<td>DE</td>
<td>$\sum_{F \subseteq \text{FP} \cup \text{FN}}</td>
<td>F</td>
</tr>
<tr>
<td>OER</td>
<td>$100 \times \frac{\sum_{T \cap \text{TP}}</td>
<td>(\text{Ref} \cap \text{Seg})</td>
</tr>
</tbody>
</table>

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 $\partial$ denotes the cardinality of a set considered in a voxelwise manner, the subscript $c$ indicating the connected set, and $\partial$ denoting the border of a set defined in the 18-neighborhood connectivity scheme. With this denomination, considering $\text{Seg} \cup \text{Ref}$, $F_P$ denotes the set of connected components wrongly considered as positive, while $\text{FN}_c$ corresponds to the set of connected components wrongly considered as negative. In turn, $\text{TP}_c$ represents the true positive connected components, that is the connected components for which $\text{Seg} \cap \text{Ref} \neq \emptyset$. 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.

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 levels 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

This work is licensed under a Creative Commons Attribution 3.0 License. For more information, see http://creativecommons.org/licenses/by/3.0/.
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.

<table>
<thead>
<tr>
<th>Assessment method</th>
<th>Mild</th>
<th>Moderate</th>
<th>Severe</th>
</tr>
</thead>
<tbody>
<tr>
<td>DSC (%)</td>
<td>VD (%)</td>
<td>FPR (%)</td>
<td>TPR (%)</td>
</tr>
<tr>
<td>BaMoS-NoCov</td>
<td>36.9</td>
<td>94.3</td>
<td>62.9</td>
</tr>
<tr>
<td>BaMoS</td>
<td><strong>51.4</strong></td>
<td>46.4</td>
<td>44.9</td>
</tr>
<tr>
<td>BaMoS-static</td>
<td>48.6</td>
<td><strong>43.1</strong></td>
<td><strong>25.6</strong></td>
</tr>
<tr>
<td>EMS-D</td>
<td>2.6</td>
<td>4851.6</td>
<td>4893.3</td>
</tr>
<tr>
<td>EMS-C</td>
<td>20.4</td>
<td>582.6</td>
<td>6060.0</td>
</tr>
<tr>
<td>LST-WML</td>
<td>24.9</td>
<td>120.9</td>
<td>173.1</td>
</tr>
<tr>
<td>LST-MS</td>
<td>22.8</td>
<td>70.6</td>
<td>15.0</td>
</tr>
<tr>
<td>TOADS</td>
<td>0.8</td>
<td>983.8</td>
<td>1078.8</td>
</tr>
<tr>
<td>DSC (%)</td>
<td>VD (%)</td>
<td>FPR (%)</td>
<td>TPR (%)</td>
</tr>
<tr>
<td>BaMoS-NoCov</td>
<td>55.7</td>
<td>75.3</td>
<td>91.6</td>
</tr>
<tr>
<td>BaMoS</td>
<td><strong>70.2</strong></td>
<td>22.9</td>
<td>33.6</td>
</tr>
<tr>
<td>BaMoS-static</td>
<td>65.0</td>
<td>33.9</td>
<td><strong>20.8</strong></td>
</tr>
<tr>
<td>EMS-D</td>
<td>20.4</td>
<td>582.6</td>
<td>6060.0</td>
</tr>
<tr>
<td>EMS-C</td>
<td>44.9</td>
<td>91.8</td>
<td>124.8</td>
</tr>
<tr>
<td>LST-WML</td>
<td>62.9</td>
<td>30.2</td>
<td>16.2</td>
</tr>
<tr>
<td>LST-MS</td>
<td>62.6</td>
<td>34.7</td>
<td>13.4</td>
</tr>
<tr>
<td>TOADS</td>
<td>15.0</td>
<td>64.2</td>
<td>144.3</td>
</tr>
<tr>
<td>DSC (%)</td>
<td>VD (%)</td>
<td>FPR (%)</td>
<td>TPR (%)</td>
</tr>
<tr>
<td>BaMoS-NoCov</td>
<td>77.2</td>
<td>37.5</td>
<td>46.3</td>
</tr>
<tr>
<td>BaMoS</td>
<td><strong>81.2</strong></td>
<td>23.2</td>
<td>32.6</td>
</tr>
<tr>
<td>BaMoS-static</td>
<td>79.7</td>
<td><strong>20.1</strong></td>
<td><strong>16.4</strong></td>
</tr>
<tr>
<td>EMS-D</td>
<td>39.9</td>
<td>195.5</td>
<td>217.1</td>
</tr>
<tr>
<td>EMS-C</td>
<td>64.1</td>
<td>42.9</td>
<td>52.6</td>
</tr>
<tr>
<td>LST-WML</td>
<td>77.3</td>
<td>13.1</td>
<td>14.6</td>
</tr>
<tr>
<td>LST-MS</td>
<td>77.1</td>
<td>15.8</td>
<td>13.0</td>
</tr>
<tr>
<td>TOADS</td>
<td>64.6</td>
<td>6.1</td>
<td>31.5</td>
</tr>
</tbody>
</table>

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).

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

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.

### 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.

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

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...
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 re-
mainning 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 as-
essment 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 signifi-
cantly 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 seg-
mentation 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
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 overall 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 × 0.958 × 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 samples.
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.

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.
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.

**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
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).

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 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.
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. 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
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 insufficient 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 somewhat 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 WML.
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.
TLL-manual Manually segmented TLL.
TPR True Positive Rate.
VD Volume difference.
WM White matter.
WMH White matter hyper-intensity.

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.
\( || \) Applied on a matrix, denotes the determinant of the matrix.
\( \mathcal{E} \) Applied on a random variable, denotes the expectation of this variable.
\( e_\ell \) The \( \ell \) vector of the canonical basis.
\( \ast \) 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 \( l \) 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 \( K_j \) Used for the Level 3 of the hierarchy with index \( k \), and \( K_j \) representing the number of Gaussian distributions used to model class \( l_j \).

3) Sets and vectors:

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

4) GMM density distributions and parameters:

\( \mu \) Mean (vector) for a Gaussian distribution.
\( \Lambda \) Covariance matrix for a Gaussian distribution.
\( \theta \) Set of Gaussian parameters \( \{ \mu, \Lambda \} \), generally indexed by \( l_j \).
\( \Theta \) Parameters of the Gaussian components of a mixture indexed by \( l_j \).
\( K \) Model under consideration, characterising the number of Gaussian components \( K_j \) per mixture \( l_j \).
Denotes the complete set of parameters used for the model.

\( l \) and \( O \) Denotes the density distribution function for the inlier and outlier part of the model respectively.

\( \Phi \) Density distribution for a mixture at Level 2 of the hierarchy, indexed by \( l_j \).

\( \mathcal{G} \) Notation adopted for a Gaussian density distribution

\( \mathcal{U} \) Notation adopted for a uniform distribution.

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

\( \Xi \) Denotes the complete set of parameters used for the model.

\( \psi \) Indexed by \( m \), corresponds to a IIH polynomial basis function.

\( \text{pos} \) Indexed by \( n \), corresponds to the spatial location of voxel \( n \).

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

\( W \) Denotes the set of all vectors \( \{w_l\} \) 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.

\( \mathcal{A} (\mathcal{B}) \) In the context of statistical atlases, corresponds to the vector built from \( \mathcal{A} (\mathcal{B}) \), indexed by \( n \). Otherwise, corresponds to the vector of global mixing weights.

\( \mathcal{X} \) Set of vectors \( \{a_1, \cdots, a_N\} \) (\( \{b_1, \cdots, b_N\} \)) representing the statistical atlases.

\( \pi \) When indexed by \( n l_j \), is defined as \( \pi_{nl_j} = b_{nl} a_{nj} w_{lj} \).

\( \delta \) Depending on the context, denotes either the set \( \{b, a, W\} \) of global mixing weights or the set \( \{\mathcal{B}, \mathcal{A}, \mathcal{W}\} \) using the statistical atlases for Level 1 and Level 2.

\( \mathcal{D} \) Denotes the Dirichlet distribution.

\( \mathcal{B} \) Denotes the Beta distribution.

\( \delta \) 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.

\( \kappa \) Directly related to \( \delta \). The highest \( \kappa \), the strongest the relaxation.

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

\( \beta_{n_j}, \beta_n \) Same as \( \alpha_{n_j} \) and \( \alpha_n \) but at Level 1 of the hierarchy.

\( p_{nl} \) Responsibility marginalized over the second level.

\( p_{nj} \) Responsibility marginalized over the first level.

\( G_{\sigma} \) Gaussian kernel with standard deviation \( \sigma \).

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

\( \mathcal{P}_n \) Denotes the vector or responsibilities for all \( l_{jk} \) components. When indexed by \( \mathcal{N}_n \), denotes the set of such vectors for the voxels in \( \mathcal{N}_n \).

\( U_{\mathcal{MRF}} \) Energy function related to the current labelling configuration.

\( \phi_{nl_j} \) Abbreviation of \( f \left( y_n, z_n = e_{l_j}, \xi \right) \).

\( \psi_{nl_j} \) Abbreviation of \( \exp \left( -U_{\mathcal{MRF}} \left( e_{l_j}, p_{lj}^{(i)}, H \right) \right) \).

\( H \) MRF inter-class energy matrix.

\( \Omega_{l_j} \) Weighted covariance matrix for the Gaussian component \( l_j \) at Level 3.

\( \Psi \) Prior over the model covariances.

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

\( \mathcal{BIC}(\mathcal{K}) \) Bayesian Information Criterion on model \( \mathcal{K} \).

\( \mathcal{DK}(\mathcal{K}) \) Penalization function over the model \( \mathcal{K} \) used in BIC.

\( \text{corr}_r \) Value of correlation between adjacent voxels in the \( r \)-direction.

\( \mathcal{L} \) Set of components potentially considered as lesions.

\( \mathcal{TL} \) Set of components considered as lesions.

\( \mathcal{DL} \) Set of possible lesion-related components that need further refinement.

\( \mathcal{RL} \) Final set of lesion-related components refining \( \mathcal{DL} \).

\( \mathcal{Patho} \) Modalities that can be used as indicator of pathology in the case of WMH, that is FLAIR, T2 and PD.

\( \mathcal{L} \) Set of components potentially considered as lesions.

\( \mathcal{TL} \) Set of components considered as lesions.

\( \mathcal{DL} \) Set of possible lesion-related components that need further refinement.

\( \mathcal{RL} \) Final set of lesion-related components refining \( \mathcal{DL} \).

\( \mathcal{Patho} \) Modalities that can be used as indicator of pathology in the case of WMH, that is FLAIR, T2 and PD.

\( \partial \) Denotes the border of a binarized object.

\( \# \) Denotes the cardinality of a set.

\( \mathcal{Seg} \) Refers to the binary lesion segmentation evaluated.

\( \mathcal{FP}_c \) Set of connected elements that belong only to \( \mathcal{Seg} \).

\( \mathcal{FN}_c \) Set of connected elements that belong only to \( \mathcal{Ref} \).

\( \mathcal{TP}_c \) Set of connected elements for which at least one voxel is a true positive.
SegRef 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 Alzheimer’s Research UK (ARUK). The Dementia Research Centre is supported by ARUK, Brain Research Trust and Alzheimer’s Research UK (ARUK). The Dementia Research Centre is supported by ARUK, Brain Research Trust and Alzheimer’s Research UK (ARUK). The Dementia Research Centre is supported by ARUK, Brain Research Trust and Alzheimer’s Research UK (ARUK).

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


