NeuroImage 124 (2016) 1031–1043

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Stratified mixture modeling for segmentation of white-matter lesions in brain MR images

Alfiia Galimzianova [a,*], Franjo Pernuš [a], Boštjan Likar [a,b], Žiga Špiclin [a,b]

[a] University of Ljubljana, Faculty of Electrical Engineering, Tržaška 25, 1000 Ljubljana, Slovenia
[b] Sensum, Computer Vision Systems, Tehnološki park 21, 1000 Ljubljana, Slovenia

ARTICLE INFO

Article history:
Received 2 May 2015
Accepted 20 September 2015
Available online 30 September 2015

Keywords:
Magnetic resonance
Brain imaging
Lesion segmentation
Intensity model
Stratified mixture model
Robust estimation

ABSTRACT

Accurate characterization of white-matter lesions from magnetic resonance (MR) images has increasing importance for diagnosis and management of treatment of certain neurological diseases, and can be performed in an objective and effective way by automated lesion segmentation. This usually involves modeling the whole-brain MR intensity distribution, however, capturing various sources of MR intensity variability and lesion heterogeneity results in highly complex whole-brain MR intensity models, thus their robust estimation on a large set of MR images presents a huge challenge. We propose a novel approach employing stratified mixture modeling, where the main premise is that the otherwise complex whole-brain model can be reduced to a tractable parametric form in small brain subregions. We show on MR images of multiple sclerosis (MS) patients with different lesion loads that robust estimators enable accurate mixture modeling of MR intensity in small brain subregions even in the presence of lesions. Recombination of the mixture models across strata provided an accurate whole-brain MR intensity model. Increasing the number of subregions and, thereby, the model complexity, consistently improved the accuracy of whole-brain MR intensity modeling and segmentation of normal structures. The proposed approach was incorporated into three unsupervised lesion segmentation methods and, compared to original and three other state-of-the-art methods, the proposed modeling approach significantly improved lesion segmentation according to increased Dice similarity indices and lower number of false positives on real MR images of 30 patients with MS.

Introduction

Magnetic resonance (MR) imaging is by far the most sensitive imaging technique for detection of white-matter lesions (Rocca et al., 2013), a pathological presence of which is highly associated with the clinical outcome of certain neurodegenerative and mental disorders, and cerebrovascular diseases (Rovira et al., 2015; Debette and Markus, 2010; Prins and Scheltens, 2015). Quantification of the number, size and spatial distribution of the lesions, which are valuable biomarkers, requires accurate segmentation of three-dimensional MR images of several conventional sequences, like T1-weighted (T1w), T2-weighted (T2w), proton density weighted (PD), and fluid attenuated inversion recovery (FLAIR). Segmentation can be performed manually by delineating each lesion on every two-dimensional (2D) slice of an MR image. However, this task is cumbersome and time-consuming, but most of all subjective and thus rather unreliable. Especially in large clinical trials that involve processing of a large number of MR images, there is a need for efficient, accurate and reliable automated lesion segmentation so as to deliver timely and consistent measurements. Although automated segmentation is becoming a general routine in large clinical trials, none of the methods has yet been widely accepted as the standard method (Lladó et al., 2012; Vrenken et al., 2013; García-Lorenzo et al., 2013).

Automated methods, in general, apply learned discriminative or generative models of normal and/or pathological brain structures for lesion segmentation. Supervised learning uses training datasets with reference segmentations to combine various MR-intensity derived features with different classifiers. The features can be multi-sequence MR (msMR) intensities normalized across datasets (Shah et al., 2011), voxel spatial locations (Anbeek et al., 2004), aggregative features of msMR intensity, shape, location, and neighborhood derived from image subregions (Akselrod-Ballin et al., 2009), and sagittal brain symmetry features (Geremia et al., 2011), while the classifiers can be *k* nearest neighbor (*k*-NN) (Cocosco et al., 2003; Anbeek et al., 2004; Warfield et al., 2000; Steenwijk et al., 2013), random decision forests (Akselrod-Ballin et al., 2009; Geremia et al., 2011), Parzen window classifiers (Datta and Narayana, 2013), support vector machines (Lao et al., 2008), relevance vector machines (Karimaghaloo et al., 2012), and regression models (Sweeney et al., 2013). Unsupervised learning, on the other hand, does not require training datasets as it searches for natural clusters of image features formed by the theoretical sources of the

* Corresponding author.
E-mail address: alfiia.galimzianova@fe.uni-lj.si (A. Galimzianova).
URL: http://lit.fe.uni-lj.si (A. Galimzianova).

http://dx.doi.org/10.1016/j.neuroimage.2015.09.047
1053-8119/

imaging processes, e.g., in msMR intensity distributions (Van Leemput et al., 2001; García-Lorenzo et al., 2011; Khayati et al., 2008; Sudre et al., 2014), combined space-intensity distributions (García-Lorenzo et al., 2008; Freifeld et al., 2009), and image patch distributions (Kadoury et al., 2012; Weiss et al., 2013). Both supervised and unsupervised learning strategies can benefit from incorporation of prior anatomical knowledge in the form of statistical atlases (Warfield et al., 2000; Van Leemput et al., 2001), topological atlases (Shiee et al., 2010), disease-related rules (García-Lorenzo et al., 2011), physical models of lesion growth (Prastawa and Gerig, 2008), or healthy population intensity distributions (Roy et al., 2014; Tomas-Fernandez and Warfield, 2015).

As the msMR intensities are the core feature of brain segmentation methods, the generative models are most commonly the msMR intensity models of the brain. Under the assumption of intensity homogeneity of major normal-appearing brain structures (NABS), such as cerebrospinal fluid (CSF), gray matter (GM) and white matter (WM), their joint intensity probability distributions are particularly convenient to be modeled by finite mixture models (McLachlan and Peel, 2005). These are usually estimated from the msMR intensities of the target MR image set in an unsupervised manner through (efficient) likelihood maximization (Dempster et al., 1977). Besides the variability of NABS and lesion proportions from subject to subject, other considerable sources of MR intensity variability are the lack of MR intensity signal standardization (Shah et al., 2011), partial volume effect (PVE) at interfaces of brain structures (Cuadra et al., 2005), non-stationary Rician noise (Gudbjartsson and Patz, 1995; Manjón et al., 2010), presence of spatial intensity variations due to MR bias field (Vovk et al., 2007) and spatially-varying structure properties (Xiao et al., 2010). To deal with the aforementioned sources, various modeling approaches have been proposed, including modeling of NABS as Gaussian mixture models (GMM) combined with robust parameter estimation (Van Leemput et al., 2001; García-Lorenzo et al., 2011), modeling NABS and lesions as a mixture of Gaussian and uniform distributions (Rouaïnia et al., 2006), modeling explicitly the PVE (Souplet et al., 2008), modeling NABS mixtures of generalized Gaussian and Rician distribution (Wu et al., 2011), modeling the normal-appearing subcortical structures and lesions as a mixture of unimodal intensity clusters (Shiee et al., 2010), modeling NABS and lesions by mixtures of GMMs (Khayati et al., 2008; Xiao et al., 2010; Elliott et al., 2013), or by mixtures of Gaussian and uniform mixture distributions (Sudre et al., 2014).

Accounting for as many sources of MR intensity variability and lesion heterogeneity as possible results in increased complexity of generative models and, therefore, robust estimation of these models in an unsupervised setting presents a very challenging problem. As demonstrated by Xiao et al. (2010), the within-structure spatial intensity variability is better addressed by modeling each structure as a multimodal distribution and consequently the whole-brain NABS as a mixture of GMMs. While in supervised learning the optimal number of mixture components and their parameters can be learned from the training datasets (Xiao et al., 2010; Elliott et al., 2013), in unsupervised learning in order to avoid over-fitting of the NABS components, it is crucial that all the sources of msMR intensities are captured (Khayati et al., 2008; Sudre et al., 2014). However, explicitly modeling the outliers of the NABS model, such as lesions, vessels, and iron deposition, further increases the model complexity and might thus result in false positives when the lesion volume is small or absent. Moreover, none of the modeling approaches referred above employed explicitly the main source of intensity variability, i.e., the spatial variation of a structure's property.

Spatially-adaptive msMR models were introduced for supervised learning of the intensity distributions of NABS and lesions (Harmouche et al., 2015), and for unsupervised learning of the intensity distributions of NABS in healthy brains (Scherrer et al., 2009; Tohka et al., 2010). In the latter approaches, however, both model estimation and segmentation are performed locally and, therefore, additional spatial regularization of the local models is needed to obtain consistent segmentations. The regularization may substantially increase model complexity and model estimation.

In this paper, we propose to use stratified mixture modeling for robust unsupervised estimation of the distribution of msMR intensities of NABS and then detect lesions as outliers of the generative model. Two different stratified models were evaluated, one based on parameter- and the other on model-wise recombination over strata into a whole-brain generative model of NABS intensity distribution. The latter better captured the NABS intensity distributions according to better goodness-of-fit and provided more accurate segmentation of normal structures, thus it was employed for upgrading lesion segmentation methods. The main premise of the proposed approach is that an otherwise complex generative model of the whole-brain distribution of msMR intensities reduces to a tractable parametric form in small enough local brain subregions, as shown previously by other researchers (Scherrer et al., 2009; Shattuck et al., 2001). However, because lesions may represent a substantial and variable fraction of observations in small brain subregions, they can adversely impact the estimation of local msMR intensity models. Our solution to this problem is an effective spatial stratification of the brain into plausible subregions such that they contain a certain minimal and maximal fraction of NABS and outliers, respectively. For this purpose, results of a tentative whole brain segmentation are used. In this way, the minimal requirements of the robust unbalanced mixture model estimator (Galimzianova et al., 2015) are satisfied and, therefore, a good estimation of the whole-brain stratified mixture can be obtained.

The proposed stratified mixture modeling approach improves the estimation of the distributions of msMR intensities of NABS on 30 real MR image datasets of patients with multiple sclerosis (MS). As such, it was incorporated into three unsupervised lesion segmentation methods based on the model outlier detection paradigm (Van Leemput et al., 2001; García-Lorenzo et al., 2009, 2011). Compared to the original implementations of these methods based on GMM of the whole-brain msMR intensities, the upgraded implementations with stratified mixture models significantly improved lesion segmentation in patients with mild and moderate lesion loads. The improvements were mainly due to considerably lower number of false positives. A comparison to three other state-of-the-art methods showed that one of the upgraded implementations was superior in segmentation performance on the majority of the MS patient datasets.

MR intensity modeling

The generative model of msMR intensities of major normal-appearing brain structures that we propose aims to account for sources of msMR intensity variability, while keeping the model complexity low to enable efficient and robust estimation of its parameters.

The proposed generative model is based on the idea of stratified sampling (Thompson, 2012), which provides means to collect the statistics of an inhomogeneous sample through collecting and recombining the statistics across more homogeneous subsamples (strata). The collection of statistics, for our purposes, refers to the estimation of the model parameters. The strata should be mutually exclusive and collectively exhaustive, i.e., every observation in the sample must be assigned to only one stratum, and the strata should cover the whole sample. Compared to the approach, in which the generative model is estimated from the original inhomogeneous sample, the advantage of stratified model estimation and recombination of the models, or model parameters, over strata is that it provides more accurate and robust model estimates.

In the context of modeling intensities of brain MR images, for example, inhomogeneity of MR intensities may originate from a spatially smooth multiplicative MR bias field (Vovk et al., 2007). Relying on the spatial smoothness of the bias field, the MR intensities are generally considered homogeneous over small enough local brain subregions (Shattuck et al., 2001). Similarly, the natural variability of brain

structures across the lobes (Xiao et al., 2010) can be considered negligible within local subregions smaller than the sizes of the lobes. Based on these observations, we consider local subregions of the brain as strata and apply the principles of stratified sampling to the modeling and estimation of the joint probability distribution of brain msMR intensities.

Stratified mixture models

Let $Y = \{y_1, \ldots, y_N\}$ be a collection of M-dimensional vectors in the real domain $\mathbb{R}^M$, which are composed of msMR intensities (of conventional T1w, T2w and PD or FLAIR sequences) sampled from the spatial domain Ω of the brain with coordinates $X = \{x_1, \ldots, x_N\}$ in the real domain $\mathbb{R}^3$. Let $Z = \{z_1, \ldots, z_N\}$ be a collection of indicator variables $z_j = (z_j^1, \ldots, z_j^L)$ with z_j^l taking the value 1 *iff* the j-th voxel corresponds to a hidden label $l \in \{1, \ldots, L\}$ that indicates one of the major normal-appearing brain structures (i.e., CSF, GM or WM).

Although conventional MR imaging sequences generally produce observations y_j that form L separated clusters in M-dimensional space for each of these structures, the imperfections of the physical MR acquisition process such as MR bias field, varying structural properties, PVE and non-stationary noise can strongly influence the form of these clusters. The joint probability distribution of this process is captured by a compact but very general representation of a finite mixture model $p(y_j|\Theta) = \sum_{l=1}^{L} \pi_l p(y_j|\theta_l)$, in which each cluster is a mixture of R distributions of the same parametric form (Xiao et al., 2010):

$$p\left(y_j|\theta_l\right) \simeq \sum\nolimits_r \pi_{l,r} p\left(y_j|\theta_{l,r}\right) \tag{1}$$

where, for the l-th structure and its r-th component, $\theta_{l,r}$ is the vector of parameters and $\pi_{l,r}$ is the mixture weight. When approximating the component distributions by a Gaussian as $p(y_j|\theta_{l,r}) = g(y_j|\mu_{l,r}, \Lambda_{l,r})$ with mean $\mu_{l,r}$ and covariance matrix $\Lambda_{l,r}$, the l-th structural model is defined by the set of parameters $\theta_l = \{\mu_{l,r}, \Lambda_{l,r}\}_{r=1,\ldots,R}$.

To obtain strata that are homogeneous, mutually exclusive and collectively exhaustive, we parcellate (stratify) the spatial domain Ω of the brain image into R non-overlapping subregions Ω_r, $r = 1, \ldots, R$. The joint probability distribution is obtained by a weighted sum of stratum-specific distributions (Breunig, 2008; Wakimoto, 1971) as:

$$p\left(y_j|\Theta\right) = \sum_{r=1}^{R} w_r p\left(y_j|\Theta_r\right) \tag{2}$$

where $w_r = |\Omega_r|/|\Omega|$, $|\cdot|$ is the number of observations in a region used for estimation, and $\Theta_r = \{\pi_{l,r}, \mu_{l,r}, \Lambda_{l,r}\}_{l=1,\ldots,L}$ is a set of parameters for region r. Based on the above formulations, the priors of the l-th structure are $\pi_L = \sum_{r=1}^{R} w_r \pi_{l,r}$ and the corresponding *stratified mixture model* (SM-GMM) has L components of the following form:

$$p\left(y_j|\Theta_l\right) = \pi_l^{-1} \sum_{r=1}^{R} w_r \pi_{l,r}\, g\left(y_j\middle|\mu_{l,r}, \Lambda_{l,r}\right). \tag{3}$$

Note that the traditional stratification of the statistics such as means and covariances of the clusters, i.e. as in Wakimoto (1971):

$$\begin{aligned} \mu_l &= \textstyle\sum_{r=1}^{R} w_r \mu_{l,r}, \\ \Lambda_l &= \textstyle\sum_{r=1}^{R} w_r \left(\Lambda_{l,r} + \left(\mu_{l,r} - \mu_l\right)\left(\mu_{l,r} - \mu_l\right)^T \right) \end{aligned} \tag{4}$$

results in a recombination of the mixture model parameters over strata (*stratified model parameters* (SP-GMM)) into a conventional L-component GMM with weights $\pi_l = \sum_{r=1}^{R} w_r \pi_{l,r}$. Such a GMM is expected to have parameter estimates affected by different brain subregions, however, it is generally over-simplified, as will be shown in Section MR intensity modeling.

Robust model estimation

Estimation of the stratified mixture models (Eq. 3) proceeds by spatial stratification of the brain region into R strata and performing R independent three-component ($L = 3$) GMM parameter estimations. For robust estimation of a Gaussian mixture with parameters Θ_r, each of the subregions Ω_r, $r = 1, \ldots, R$ should contain a sufficient number of observations $N_r = |\Omega_r|$, whereas the required minimum number of observations depends on the employed estimator.

We use an estimator of unbalanced mixtures (Galimzianova et al., 2015), a trimming-based approach that is robust to outliers at an a priori specified trimming fraction α. Parameter α represents some high, marginal value of an expected outlier fraction and can take values in the range [0, 0.5]. Besides, the characteristics of each subregion Ω_r should meet the following three criteria for a solution (Neykov et al., 2007): 1) the sample size N_r must be larger than $L(M+1)$; 2) the actual fraction of outliers h^* must be less than $h_{max} = \frac{1}{N_r}(N_r - L(M+1))/2 \approx 0.5$; and 3) α should be set higher than h^*. In order to meet the last two criteria, we need to roughly estimate the outlier map. Since it is sufficient to find an upper limit of the actual outlier fraction h, we compute a tentative over-estimated outlier map $O = \{o_1, \ldots, o_N\}$, $o_j \in \{0, 1\}$ by fitting a three-component GMM using the robust estimator with high trimming fraction on the msMR intensities of the whole brain volume Ω.

It can be easily verified that for any region with the fraction h of outliers, a set of subregions will have a maximal fraction of outliers not lower than h. Therefore, spatial stratification of the brain region is preferably performed by top-down approaches that, given an initial distribution of outliers, can assess at each subregion the fraction of outliers and thus guarantee that a subregion does not have to be further subdivided unless it has an outlier fraction lower than h. We spatially stratified the brain region as described in Algorithm 1, which finds strata such that they: 1) meet the stratification principles by being mutually exclusive, collectively exhaustive, and containing more homogeneous subsamples, 2) allow local mixture estimation by verification of the representativeness of all the NABS components, and 3) allow reliable robust estimation by meeting the estimator requirements. An example of a spatial stratification is shown in Fig. 1. Note how highly unbalanced and contaminated by various fractions of outliers are the mixtures in

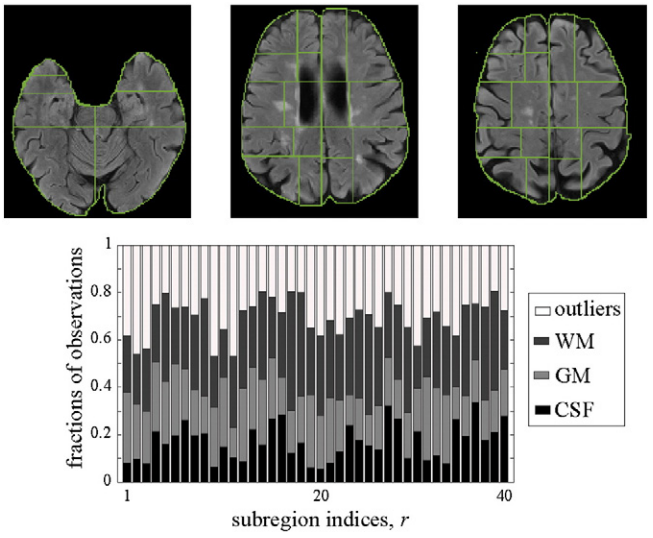

Fig. 1. Example of spatial brain stratification: (upper row) boundaries of subregions (*green*) superimposed over axial slices of the FLAIR image, and (lower row) the cardinalities of the NABS clusters and outliers over the obtained subregions. Spatial stratification is performed in 3D and by construction ensures that each strata contains observations from each of the three main structures (CSF, GM and WM).

$R = 40$ strata. Nevertheless, the stratification algorithm ensures that each strata contains observations from each of the three main structures (CSF, GM and WM) with corresponding mixture weight no less than π_{min}.

After spatial stratification, the stratum-specific parameters of the estimator α_r should be set to a value above the true outlier fraction in region r (Neykov et al., 2007; Galimzianova et al., 2015). As $h_r = \frac{1}{N}\sum_{j\in\Omega} o_j$ is an over-estimated fraction of outliers, we set $\alpha_r = h_r$. The weights w_r of each stratum are also updated according to their contribution to the whole-brain model estimate as $w_r = (1-\alpha_r)N_r/(1-\alpha)N$, i.e., according to the fraction of inliers in a stratum. The parameters Θ_r compactly encode the msMR intensity variability within each subregion (Fig. 2) and, following (Eq. 3), form the stratified mixture model parameters.

Algorithm 1. Spatial brain stratification

Input: whole-brain region Ω with spatial coordinates X, tentative NABS segmentation $Z = \{z_1, \ldots z_N\}$, initial map of outliers $O = \{o_1, \ldots, o_N\}$, minimal NABS component weight π_{min}, maximal outlier fraction h_{max}, minimal size of the subregions N_{min}.

Output: a set of R non-overlapping subregions or strata $\{\Omega_r\}_{r=1,\ldots,R}$, $\cup_r\Omega_r = \Omega$.

Initialization:Set the whole-brain region as the initial subregion, i.e., $\Omega_1 := \Omega$.

Stratification:

1. If the number of the brain voxels in the subregion $N_r \le N_{min}$ or the outlier fraction is not lower than a threshold, $\frac{1}{N_r}\sum_{j\in\Omega} o_j \ge h_{max}$, return the subregion Ω_r.
2. Find the longest side of the subregion $i_{max} := \mathrm{argmax}_{i\in\{1,2,3\}}(\max_{j\in\Omega_r} x^i_j - \min_{j\in\Omega_r} x^i_j)$.
3. Make a linear split at the median of outlier distribution along i_{max}, i.e., obtain $\Omega_{r1} \cup \Omega_{r2} = \Omega_r$ such that $\Omega_{r1} = \{j \in \Omega_r : x^{i_{min}}_j < x^{i_{min}}_*\}$, $\Omega_{r2} = \{j \in \Omega_r : x^{i_{min}}_j > x^{i_{min}}_*\}$ and $x^{i_{min}}_* : \sum_{j\in\Omega_{r1}} o_j \approx \sum_{j\in\Omega_r 2} o_j$.
4. If at one of the two obtained regions $r' \in \{r1, r2\}$ a) the outlier fraction is higher than a threshold, $\frac{1}{N_{r'}}\sum_{j\in\Omega_{r'}} o_j > h_{max}$, or b) the weight of one of the NABS components is smaller than a threshold, $\min_l \sum_{j\in\Omega_{r'}} z^l_j/(N_{r'} - \sum_{j\in\Omega_{r'}} o_j) < \pi_{min}$, or c) the number of observations is smaller than a threshold, $N_{r'} < N_{min}$, return the current subregion Ω_r; else recursively stratify (i.e., go to step 1) the two obtained regions Ω_{r1} and Ω_{r2}.

Lesion segmentation

The proposed stratified mixture modeling approach was incorporated into three unsupervised lesion segmentation methods (Van Leemput et al., 2001; García-Lorenzo et al., 2009, 2011). All three methods perform estimation of the generative model of msMR intensities of NABS, but mainly differ in the formulation of the objective function and in the way lesions are detected as model outliers. In these methods, the generative model of brain voxel intensities was originally represented by a three-component GMM with $p(y_j|\theta_l) = g(y_j|\mu_l, \Lambda_l)$, $l \in \{CSF, GM, WM\}$, which we have upgraded with the proposed stratified mixture model (Eq. 3). A summary of the three methods is given in Table 1, while more detailed descriptions of the methods and adaptations required to incorporate the stratified mixture model are given below.

Segmentation by model outlier detection: VL'2001

The method VL'2001 (Van Leemput et al., 2001) estimates mixture parameters Θ through maximization of the following objective function:

$$Q_w\left(\Theta\middle|\Theta^{(it-1)}\right) = \sum_{j=1}^{N}\sum_{l=1}^{L} p^{(it)}_{jl}\, t^{(it)}_{jl} \log p\left(y_j\middle|\theta^{(it-1)}_l\right) \tag{5}$$

where it denotes iteration, $p^{(it)}_{jl} = p(y_j|\theta^{(it-1)}_l)\pi^{Atlas}_{jl} / \sum_k p(y_j|\theta^{(it-1)}_l)\pi^{Atlas}_{jl}$ are the posterior likelihoods and π^{Atlas}_{jl} the priors of the l-th brain structure at voxel j based on the MNI305 brain atlas (Evans et al., 1993) co-registered to the msMR images. The robustness of model estimation was improved by typicality weights $t^{(it)}_{jl}$ (Van Leemput et al., 2001) computed for each of the L structures of NABS, i.e:

$$t^{(it)}_{jl} = \frac{p\left(y_j|\theta^{(it-1)}_l\right)\omega^{(it)}_{jl}}{p\left(y_j|\theta^{(it-1)}_l\right)\omega^{(it)}_{jl} + p\left(y^l_\kappa|\theta^{(it-1)}_l\right)\left(u^{(it)}_j\omega^{(it)}_{j,WM} + v^{(it)}_j\omega^{(it)}_{j,CSF}\right)} \tag{6}$$

where $\omega^{(it)}_{jl}$ are Markov random field (MRF) priors computed by mean-field approximation, $u^{(it)}_j$ and $v^{(it)}_j$ are the hyper- and hypo-intensity constraints, respectively, and $p(y^l_\kappa|\theta^{(it-1)}_l)$ is the likelihood at a confidence level (CL) corresponding to the value of parameter κ.

In the original method κ represented a threshold on Mahalanobis distance to determine $y^l_\kappa : (y^l_\kappa - \mu_l)\Lambda_l^{-1}(y^l_\kappa - \mu_l)^T = \kappa$, and subsequently

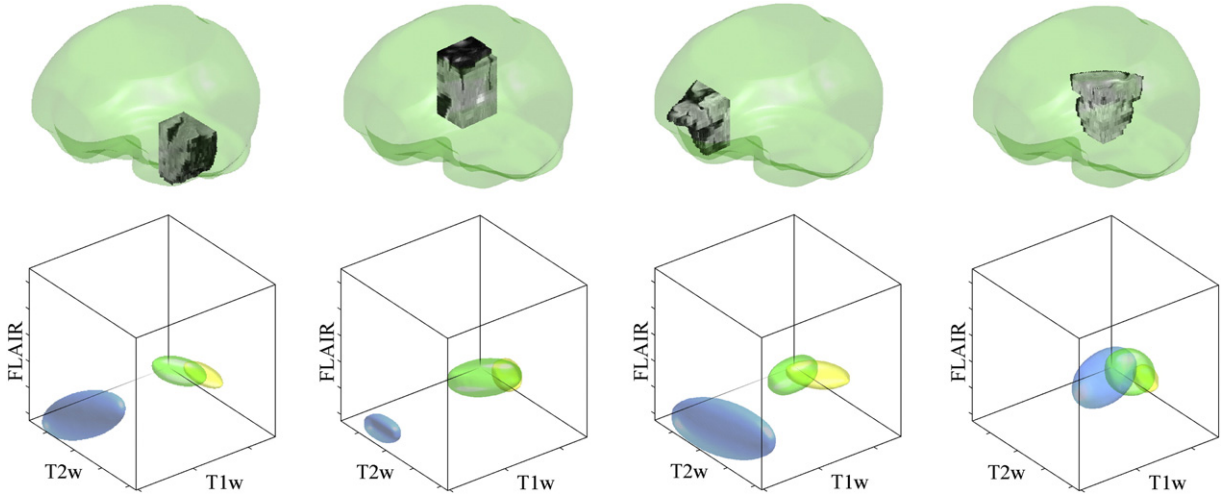

Fig. 2. An example of local msMR intensity model estimation over different brain subregions at a proximity of temporal, parietal, frontal, and occipital lobes (*top row*), and corresponding local mixture estimates shown at 0.9-confidence levels with components corresponding to CSF (*cyan*), GM (*green*), and WM (*yellow*).

Table 1
Summary of three unsupervised lesion segmentation methods tested with the proposed stratified mixture model.

Method	Objective function	Spatial constraints	Lesion detection	Parameters
VL'2001	Expectation weighted by structure typicality (Eq. 5)	MRF prior	Low model likelihood and structure typicality	$\kappa = 3.0$
GL'2009	Graph energy (Eq. 8)	Ising model	Fuzzy maps on model confidence level (Eq. 7)	$\lambda = 5$, $\alpha = 0.3$, $\kappa_b = 2.5$, $\kappa_e = 3.5$
GL'2011	Trimmed likelihood (Eq. 11)	Connected components of lesions	Threshold on model confidence level	$\alpha = 0.3$, $\delta_c = 0.3$, $p_h = 10^{-2}$

$p(y_\kappa^l|\theta_l^{(it-1)})$, while here parameter κ determines the CL as $\delta_\kappa = P_{\chi^2_M}(\kappa^2)$, for which the corresponding y_κ^l is found by numerical integration of the l-th component likelihood over domain Ω as:

$$\begin{aligned} &y_\kappa^l : CL_\kappa = \delta_\kappa, \\ &CL_\kappa = \int_{\Omega(y_\kappa^l)} p(\omega|\theta_l)d\omega, \\ &\Omega\left(y^l\right) = \left\{\omega \in \Omega : p(\omega|\theta_l) \geq p\left(y_\kappa^l|\theta_l\right)\right\}. \end{aligned} \tag{7}$$

The resulting y_κ^l is used to compute the likelihood function $p(y_\kappa^l|\theta_l^{(it-1)})$ in (Eq. 6).

At each iteration, the brain stratification was performed according to the current segmentation $z_j^l = p_{jl}t_{jl}$ and the estimates of the model parameters $\Theta^{(it)}$ were found using the original W-estimator proposed by Van Leemput et al. (2001). In order to increase robustness, the brain region was stratified as in Section Robust model estimation wherein the outlier map O was replaced by the soft atypicality weights found at iteration it as $o^{(it)} = (1 - \sum_l^L p_{jl}t_{jl})$.

The hyper-intense outliers (e.g., lesions) are defined with respect to GM mean intensity by the indicator variable $u_j^{(it)} = \wedge_{m \in \{T2w,FLAIR\}}(P_m(y_j|\theta_{GM,m}^{(it-1)}) > 0.5)$, where $\wedge(\cdot)$ is the logical conjunction (AND) operation and P_m is the marginal cumulative probability function computed on MR sequences m. The hypo-intense outliers (e.g., vessels) are defined analogously by the indicator variable $v_j^{(it)} = \wedge_{m \in \{T2w,FLAIR\}}(P_m(y_j|\theta_{GM,m}^{(it-1)}) < 0.5)$.

Given the estimated mixture parameters Θ, the probability maps of hyper- and hypo-intense outliers are given by $u_j^{(it)}(1 - \sum_l^L p_{jl}t_{jl})$ and $v_j^{(it)}(1 - \sum_l^L p_{jl}t_{jl})$, respectively. The final lesion segmentation is obtained by thresholding the hyper-intense outlier probability map at 0.5.

Segmentation by multi-modal graph cuts: GL'2009

Lesion segmentation in GL'2009 is formulated as a two-label graph-cut problem with the following total energy function (García-Lorenzo et al., 2009):

$$E(Z) = \lambda \sum_{j=1}^{N} U_v(z_j) + \sum_{\{i,j\} \in \mathcal{N}, i \neq j} B\left(y_i, y_j\right) I(z_i \neq z_j) \tag{8}$$

where U_v is the intensity potential, $B(y_i, y_j)$ the spectral gradient of msMR images computed in neighborhood $\mathcal{N}$, and $I(\cdot)$ is the indicator function. The linear weight λ is a user-defined parameter balancing the influence of the first term. The intensity potentials are computed by a fuzzy conjunction operator between FLAIR and T2w hyper-intensity fuzzy maps ($W_{j,FLAIR}$ and $W_{j,T2w}$, respectively) and minimal confidence levels mCL_j, i.e.:

$$U_v(z_j) = -\log(\wedge\{W_{j,FLAIR}, W_{j,T2w}, mCL_j\}). \tag{9}$$

Analogously to CL in (Eq. 7), the minimal confidence levels $mCL_j = \min_l(\int_{\Omega(y_j)} p(\omega|\theta_l)d\omega)$ were computed by numerical integration over corresponding confidence regions $\Omega(y_j) = \{\omega \in \Omega : p(\omega|\theta_l) \geq p(y_j|\theta_l)\}$. Mixture parameters θ_l were obtained by the stratified modeling approach as described in Section Robust model estimation.

The hyper-intensity fuzzy maps for $m \in \{T2w, FLAIR\}$ are defined as:

$$W_{j,m} = \begin{cases} 0, \ if \ x < \delta_b \\ 1, \ if \ x > \delta_e \\ \frac{x - \delta_b}{\delta_e - \delta_b}, \ otherwise \end{cases} \tag{10}$$

where $x = \text{sgn}(P_m(y_j|\theta_{WM,m}) - 0.5)\,[-\log(p(y_j \mid \theta_{WM,m}) / \max_j p(y_j|\theta_{WM,m}))]^{1/2}$ and $P_m(y_j|\theta_{WM,m})$ is a value of marginal cumulative probability function, while $\delta_b = \delta(\kappa_b)$ and $\delta_e = \delta(\kappa_e)$ are the hyper-intensity constraints on confidence level, where $\delta(\kappa) = (1 + \text{sgn}(\kappa) P_{\chi^2_1}(\kappa^2))/2$ with $\text{sgn}(\cdot)$ being the signum function.

The obtained lesion segmentation is post-processed to eliminate groups of connected voxels that are adjacent to the brain mask border. In our implementations, the segmentation of lesions was obtained by minimizing the energy function in (Eq. 8) by an efficient approximate optimizer (Komodakis et al., 2008).

Segmentation by trimmed likelihood estimation: GL'2011

In GL'2011, the trimmed likelihood (TL) objective function is formulated as (García-Lorenzo et al., 2011):

$$TL\left(\Theta \middle| \Theta^{(it-1)}\right) = \prod_{j=1}^{N-\alpha N} p\left(y_{\nu^{(it)}(j)} \middle| \Theta^{(it-1)}\right) \tag{11}$$

where $\nu^{(it)}(j)$ is a permutation of indices j that ensures a decreasing order of likelihoods $p(y_{\nu^{(it)}(j)}|\Theta^{(it-1)})$ and α represents the trimming fraction. Parameters Θ are obtained by maximizing TL, in which αN observations with lowest likelihood are trimmed from the objective function. Note that the permutation $\nu^{(it)}(j)$ is updated at each iteration. The mixture parameters Θ were obtained by the stratified modeling approach as described in Section Robust model estimation.

Given the estimated mixture parameters Θ, the set of voxels representing candidate lesions is found by thresholding minimal confidence levels $mCL_j \leq \delta_c$ computed as in (Eq. 7) with δ_c as the input parameter. Furthermore, the set of voxels is post-processed by applying three heuristic rules, specific for MS lesions: 1) elimination of voxels that are not hyper-intense according to $P_m(y_j|\theta_{WM,m}) > 1 - p_h$, where $P_m(y_j|\theta_{WM,m})$ is the marginal cumulative probability function computed for MR sequences $m \in \{T2w, FLAIR\}$ and p_h is a fixed parameter, 2) elimination of groups of connected voxels forming a volume smaller than 9 mm^3, and 3) elimination of groups of connected voxels that are adjacent to the border of the brain mask and those that are not adjacent to the WM mask.

Experimental results

Experiments involved validation of the proposed MR intensity modeling and validation of lesion segmentation methods, which involved the original method implementations and the implementations based on the proposed stratified mixture model of msMR intensities of normal-appearing brain structures. Since the comparison with respect to other state-of-the-art methods is an important aspect of validation, additional three lesion segmentation methods were evaluated. Validation was performed on clinical image datasets with manual annotations

of normal structures and pathology. The descriptions of validation datasets, validation experiments and results are given in the following subsections.

Validation datasets

Clinical datasets consisted of conventional MR images of 30 patients with MS. For each patient T1w, T2w and FLAIR sequences were acquired on a Siemens 3 T MR machine in axial multi-slice no-gap acquisition mode with 0.4×0.4 mm^2 in-plane sampling and 3.3 mm slice thickness. Automated brain mask extraction (Iglesias et al., 2011), with manual corrections where necessary, was performed on T1w image, followed by rigid registration of T1w and T2w images to corresponding FLAIR images. Each of the images was corrected for intensity inhomogeneities using N4 (Tustison et al., 2010) and downsampled to the in-plane resolution of 1×1 mm^2. To validate the methods, the WM lesions were manually segmented in all 30 MR image datasets independently by two neuroradiology experts, which then cross-validated and updated their segmentations until a consensus on the final lesion segmentations was reached. In a similar way, manual segmentations of the normal brain structures (CSF, GM, WM) were performed on 7 msMR image datasets.

The MS patients had different disease severity, which is characterized by total lesion load (TLL), where higher TLL correlates with higher patient disability. To analyze the performance of methods with respect to TLL the MR image datasets of 30 patients were divided into three groups according to TLL: mild (10 patients, $TLL \leq 5 \cdot 10^3$ mm^3), moderate (10 patients, $5\ 10^3$ mm^3 $< TLL < 20 \cdot 10^3$ mm^3) and severe (10 patients, $TLL \geq 20 \cdot 10^3$ mm^3).

MR intensity modeling

Validation of MR intensity modeling involved the estimation of generative models of msMR intensities of NABS and the assessment of the obtained models based on goodness-of-fit to the histograms of the msMR intensities and corresponding NABS segmentation performance. The initial model parameters per stratum were obtained from a simple tentative segmentation based on Otsu's thresholding of the T1w MR image. The samples for model estimation were comprised of all of the msMR intensities in a brain mask. The initial parameters Θ were found as maximum-likelihood estimates and a fraction $\alpha = 0.3$ of intensities with highest *CL*s were marked as outliers *O*.

Within the framework of stratified sampling, the generative model was obtained either by the estimation of model parameters on strata and recombination of the mixture models (Eq. 3) (i.e., SM-GMM) or by estimation and recombination of mixture model parameters over strata (i.e., SP-GMM) into the conventional three-component GMM (Eq. 4). Here we study the performance of these two models and compare them to two other intensity models estimated from the whole-brain msMR intensity observations: the three-component GMM and the mixture of GMMs (M-GMM). The first model used one component per CSF, GM and WM structures, while the M-GMM was implemented similar to the intensity model described in (Ashburner and Friston, 2005) using two components for CSF, three for GM, and two for WM. Both models were estimated using the robust estimator (Galimzianova et al., 2015) at the same trimming fraction parameter value $\alpha = 0.3$.

The two proposed approaches were tested for different numbers of strata (subregions of the MR image), obtained by the spatial brain stratification (Algorithm 1) with parameters $\pi_{min} = 0.01$ and varying

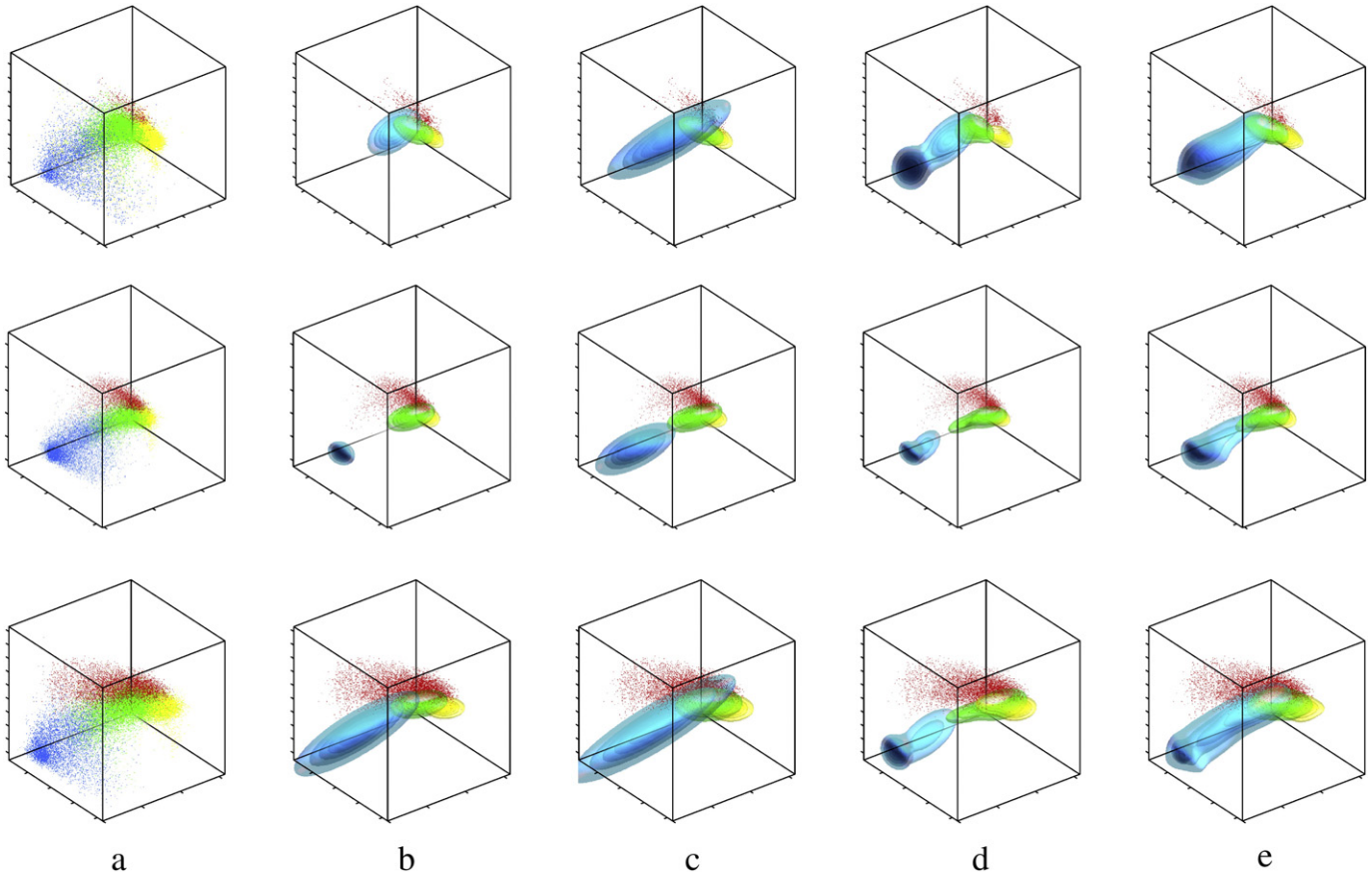

Fig. 3. Examples of model estimation on three selected datasets of images with different lesion loads. (a) Multivariate MR intensity observations colored in accordance with the manual segmentations as CSF (*cyan*), GM (*green*), WM (*yellow*), and lesions (*red*); and distribution estimates according to (b) GMM, (c) SP-GMM, (d) M-GMM, and (e) SM-GMM shown at 0.9, 0.7,0.5,0.3 and 0.1 confidence levels.

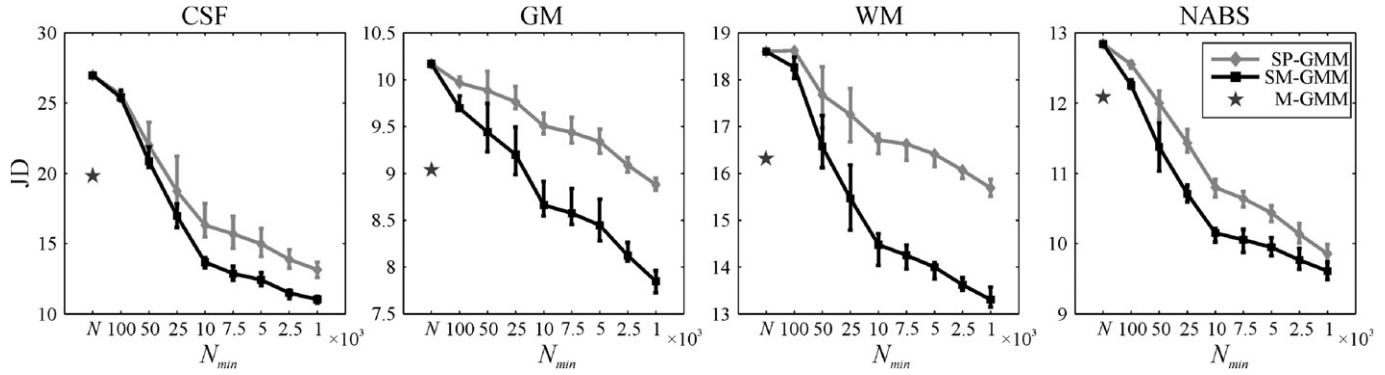

Fig. 4. The goodness-of-fit of the estimated SP-GMM (*gray line*), SM-GMM (*black line*), and M-GMM (*dark gray star*) measured by Jeffrey's divergence (JD) with respect to the msMR intensity histograms of brain structures CSF, GM, WM, and overall NABS, based on reference segmentations. The JD values were averaged over 7 MR image datasets with brain structure segmentations tested at different values of parameter N_{min}. The bars indicate the first and third quartiles of JD values.

$N_{min} \in \{100, 50, 25, 10, 7.5, 5, 2.5, 1\} \times 10^3$, that resulted in a different median number of subregions $R \in \{2, 4, 10, 21, 27, 40, 74, 147\}$.

The impact of cut direction in the spatial stratification was also tested. For this purpose, the spatial coordinates X of the brain voxels were found in a coordinate system with image-aligned axis and at its nine rotations, $\{\pi/8, \pi/4, 3\pi/8\}$ radian angles about each of the three axes of the 3D image. Together with the image-aligned coordinates, there were overall 10 spatial orientations tested. Similarly to the work of Xiao et al. (2010), the goodness-of-fit was measured by computing the Jeffrey's divergence (JD) with respect to the three-dimensional histograms of msMR intensities of CSF, GM and WM brain structures, and their union, NABS, based on reference segmentations using $\sqrt{N}$ as the number of histogram bins. Lower values of JD indicate better modeling of msMR intensities of NABS. The segmentation performance was measured by computing the overlap in terms of Dice similarity index (DSI) (Dice, 1945) between the maximum a posteriori classifications of the estimated GMM, M-GMM, SP-GMM and SM-GMM models and the manual segmentations of NABS.

The main advantage of stratified mixture modeling compared to conventional mixture modeling approach is the capability to capture arbitrarily complex joint probability distributions as shown in Fig. 3, while remaining robust to outliers, e.g., lesion intensity observations. Since the msMR intensity distributions are comparable across datasets, the obtained mixture models should also be similar. Fig. 3 demonstrates that more accurate and stable models are obtained by the two stratification approaches as compared to whole-brain model estimation (cf. the CSF components across the three datasets).

The goodness-of-fit based on JDs and segmentation performance of NABS based on DSI are shown in Figs. 4 and 5, respectively, over 7 clinical MR image datasets for which manual NABS segmentations were available. Note that $N_{min} = N$ corresponds to the whole-brain GMM estimation, which had the highest value of JD and lowest value of DSI, and, thus, the worst model fit. In general, stratification (i.e., $N_{min} < N$) improved the goodness-of-fit and segmentation performance of the models of msMR intensities of NABS with decreasing N_{min} (and thus increasing R). Cut direction had a minor impact on the performance. The approach based on SM-GMM consistently provided a better fit compared to the approach based on SP-GMM. Furthermore, based on Wilcoxon signed rank test, the improvements for JD and DSI metrics were statistically significant ($p < 0.01$) for $N_{min} < 50 \cdot 10^3$ for all the structures considered (CSF, GM, WM and NABS) and for the weighted average of DSI values (WA). The absolute differences in goodness-of-fit (JD) were especially prominent for the CSF. The reason is that CSF has two large interfaces to GM and WM and is thus most affected by PVE, which cannot be captured by a single Gaussian in the three-component GMM. Using SP-GMM model the DSI increased notably for CSF, but remained the same for GM and WM regardless of the stratification parameter N_{min}. Conversely, using SM-GMM model the DSI increased notably and consistently for all three structures, indicating overall improved segmentation performance for lower N_{min}.

Clearly, the three-component GMM is inadequate for accurate modeling of the whole-brain msMR intensity distributions (Fig. 3b,c). The higher degree-of-freedom M-GMM and SM-GMM models enable better modeling of msMR intensity distributions of normal structures (Fig. 3d,e). The results obtained for M-GMM provided higher goodness-of-fit (i.e., lower JD score) compared to the simple three-component GMM and also to the SP-GMM and SM-GMM for $N_{min} \geq 50 \cdot 10^3$ (Fig. 4). Although the obtained DSI of NABS segmentation

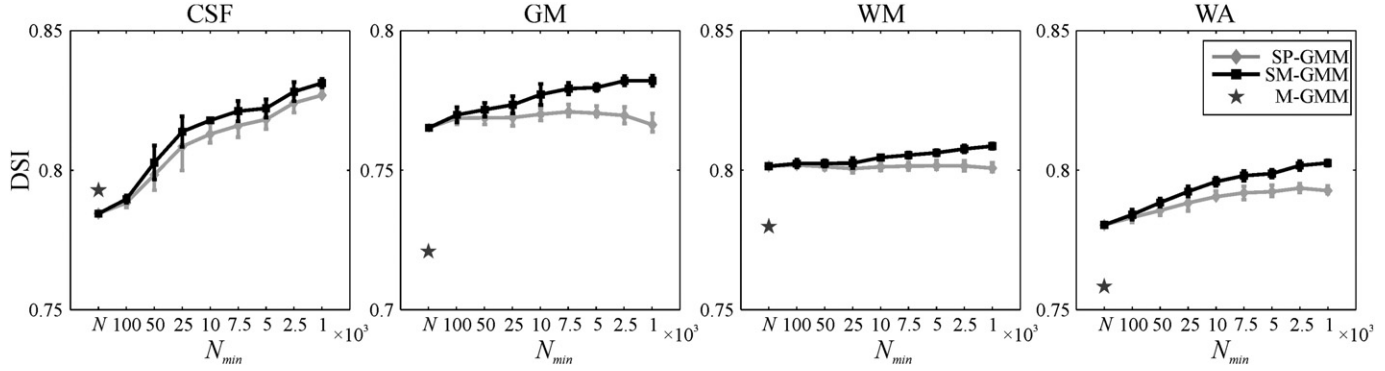

Fig. 5. The DSI values of the estimated SP-GMM (*gray line*), SM-GMM (*black line*), and M-GMM (*dark gray star*) with respect to CSF, GM, WM, and their volume weighted average (WA), based on reference segmentations. The DSI values were averaged over 7 MR image datasets with brain structure segmentations tested at different values of parameter N_{min}. The bars indicate the first and third quartiles of the DSI values.

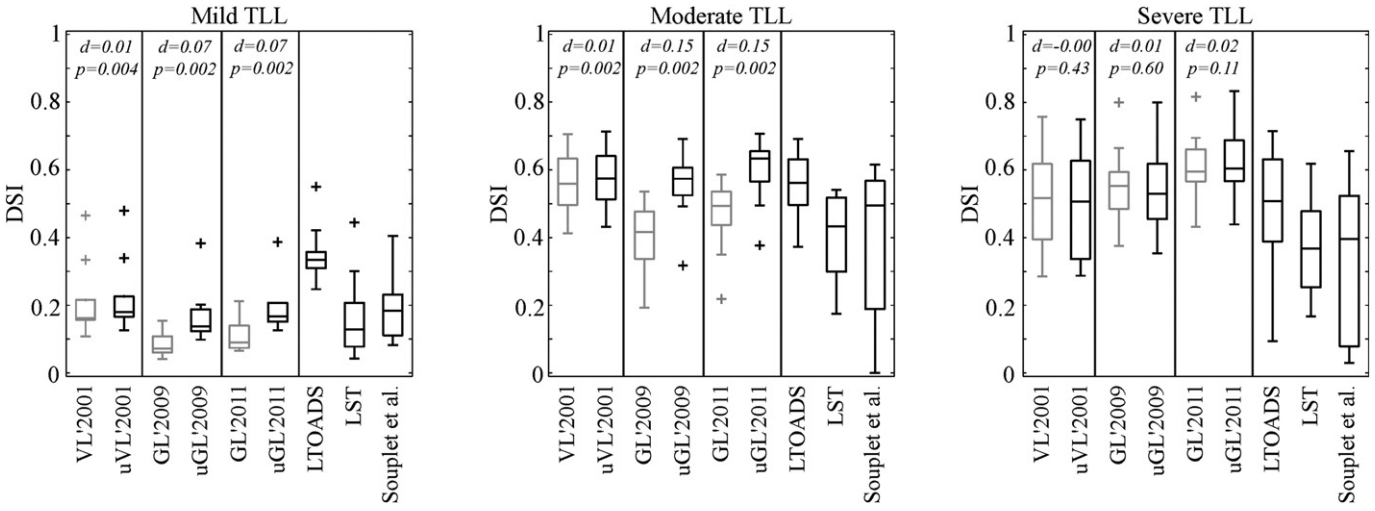

Fig. 6. Box-whisker diagrams of DSI values grouped according to reference TLL for the three lesion segmentation methods (Table 1) based on original (gray) and the proposed updated (u; black) generative models of msMR intensity distributions and for three state-of-the-art methods. The values above the updated three methods are the results of the signed rank test over the TLL groups, where *d* are the median differences between the upgraded and the original methods, and *p* are the *p*-values.

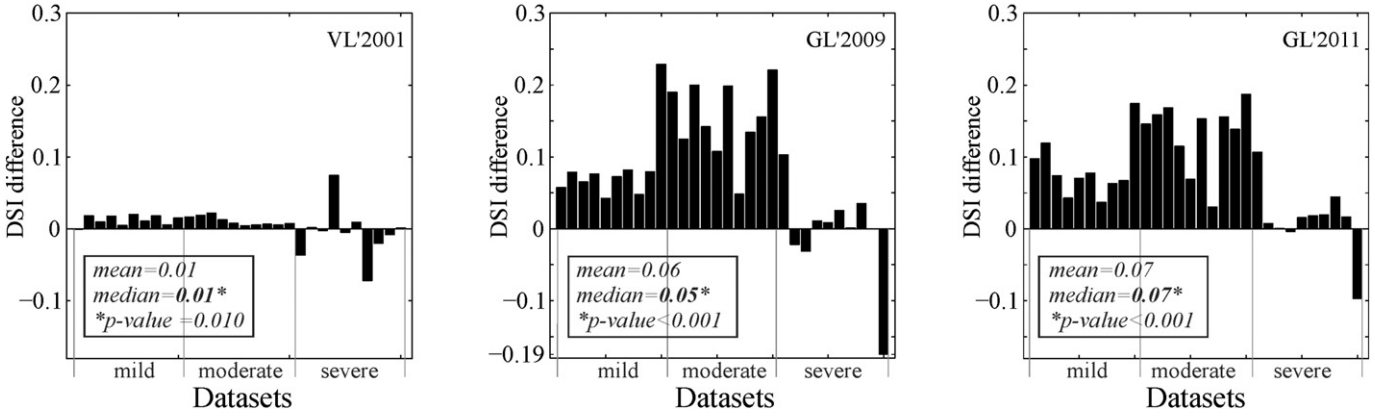

Fig. 7. The DSI differences for each of the 30 MR image datasets of MS patients for three lesion segmentation methods (Table 1) tested with original and the proposed stratified mixture model of msMR intensity distributions. For each method we report the means and medians of DSI differences and *p*-values of Wilcoxon's signed rank test.

was slightly higher for CSF with the M-GMM method compared to the three-component GMM and to the SP-GMM and SM-GMM for $N_{min} \geq 100 \cdot 10^3$, the DSIs for GM and WM segmentations, and overall WA were substantially lower with the M-GMM compared to all other tested methods. The M-GMM might require case-specific selection of optimal number of components per structure, however, high number of Gaussians could result in overfitting. Conversely, when using SM-GMM model, both goodness-of-fit and the segmentation performance improved notably and consistently for all three structures for lower N_{min} (i.e., more strata), the parameter which is intuitive and easier to adjust, and, more importantly, presents lower risk of overfitting. Therefore, we choose to employ the SM-GMM for upgrading lesion segmentation methods.

Lesion segmentation

Validation of lesion segmentation methods involved comparison of the performances of three unsupervised lesion segmentation methods (Van Leemput et al., 2001; García-Lorenzo et al., 2009, 2011), implemented either in their original form or upgraded with the stratified mixture modeling (Section Stratified mixture models). To serve as baseline for comparison, methods developed by Shiee et al. (2010) and Schmidt et al. (2012), both of which have publicly available implementations, and are referred to as LTOADS[1] and LST[2] were evaluated. The third method (Souplet et al., 2008) was from the winners of 2008 MS lesion segmentation challenge,[3] for which the implementation was devised from the descriptions in the literature. Validation involved execution of the methods on 30 clinical image datasets of patients with MS, followed by evaluation of DSI between the obtained and reference manual lesion segmentations.

In the first experiment, we study the dependence of lesion segmentation performance on the choice of parameter N_{min}. As in the previous experiment, the spatial brain stratification (Algorithm 1) was applied with parameters $\pi_{min} = 0.01$ and varying $N_{min} \in \{100, 50, 25, 10, 7.5, 5, 2.5, 1\} \times 10^3$ and also the whole-brain estimation was performed at $N_{min} = N$. The intrinsic parameters of the methods were set to values indicated in Table 1, which were chosen in accordance with the recommendations by their authors in Van Leemput et al. (2001) and García-Lorenzo et al. (2009, 2011). The obtained DSI values are shown in Fig. 10. Compared to the original

[1] LTOADS: https://www.nitrc.org/projects/toads-cruise/.
[2] LST: http://dbm.neuro.uni-jena.de/software/lst/.
[3] MS challenge 2008: http://www.ia.unc.edu/MSseg/.

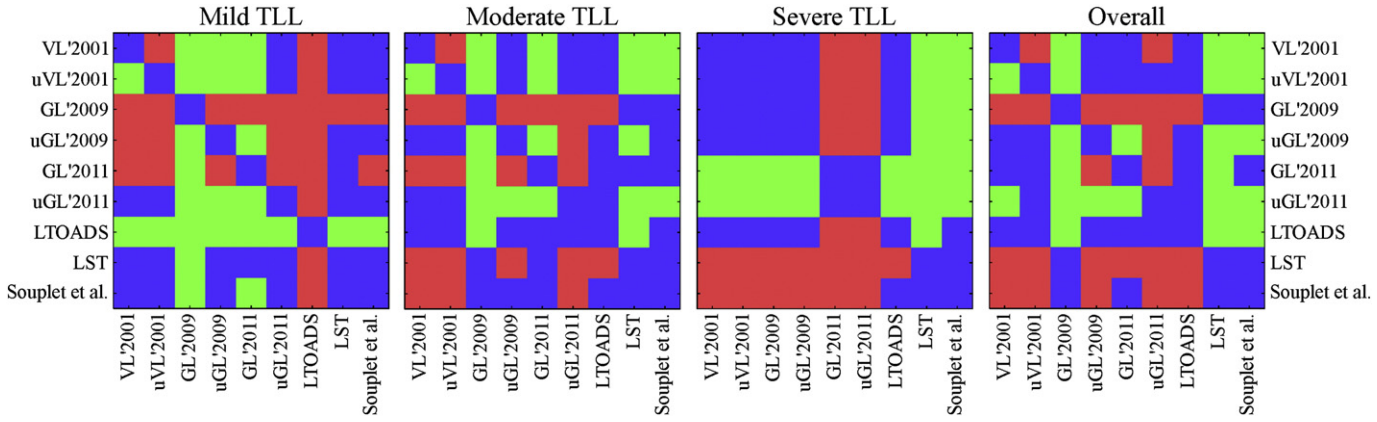

Fig. 8. Significance of DSI differences for three patient groups w.r.t. TLL and overall. According to Wilcoxon signed rank test at $p < 0.05$, the green, red and blue colors of row-column pairs indicate significant improvement, significant deterioration or no significant difference, respectively.

implementations of segmentation methods, the implementations upgraded with SM-GMM showed significant ($p < 0.05$) improvement of lesion segmentation performance for $N_{min} \leq 25 \cdot 10^3$. Moreover, with decreasing N_{min} the changes in DSI were increasingly more significant (note the number of asterisks (*) at graph nodes), while the relative improvement between consecutive values of N_{min} was the highest and significant up to $N_{min} = 5 \cdot 10^3$.

In the following experiment, we study the performance of methods at fixed minimal size of the subregions $N_{min} = 5 \cdot 10^3$. The obtained DSIs with respect to three groups of TLL, are shown in Fig. 6 for the three original and upgraded lesion segmentation methods, while the differences of DSIs of the two implementations are shown in Fig. 7 for each of the 30 MR image datasets. The Wilcoxon's signed rank test was used to indicate whether a difference of DSIs was significant ($p < 0.01$). The improvement of DSI was found significant on datasets with mild and moderate TLLs for all three methods. In general, the median DSIs improved for all methods and over all groups of TLL (Fig. 6), except for the VL'2001 on datasets with severe TLL where a small and insignificant decrease in performance was observed (difference: −0.004; p-value: 0.4). The proposed implementations of VL'2001, GL'2009 and GL'2011 significantly improved the overall median DSIs by 0.01, 0.05and 0.07 (Fig. 7), respectively, at p-values equal or below 0.01.

Table 2 illustrates the differences between upgraded and original methods by nine performance measures (see table for abbreviations). The results indicate that the significant decrease in the number of false positives (low FP) was consistently considerably higher than decrease in true positive voxels (TP). This provided improved overlap between the reference and the automated segmentations (higher DSI and LPPV) due to significantly higher PPV, while the sensitivity (TPR, LTPR) reduced slightly and at a lower rate than PPV, except on datasets with mild lesion loads. There the TPR is more affected even if only a few lesion voxels are not segmented, since the lesion volume is very small. Additional improvements were achieved in terms of volumetric measurements (lower VDR) and lower SD, thus indicating that the use of stratified mixture model stabilized the performance of lesion segmentation.

Fig. 9 shows lesion segmentations by the original and upgraded implementations, where the latter generally exhibits a much lower number of FPs. Furthermore, comparison of the three methods to other three state-of-the-art methods in Fig. 6 indicates that in terms of DSIs the LST and method by Souplet et al. (2008) were generally inferior to original and upgraded implementation of the three tested methods (VL'2001, GL'2009 and GL'2011), while the LTOADS method outperformed all other methods on datasets with mild TLL. On datasets

Table 2
Median differences of the number of false positives (FP), number of true positives (TP), true positive rate (TPR), positive predictive value (PPV), Dice similarity index (DSI), volume difference rate (VDR), average symmetric surface distance (SD, in mm), lesion-wise true positive rate (LTPR) and lesion-wise positive predictive value (LPPV) between the upgraded and original lesion segmentation methods, applied to images of patients with mild, moderate and severe lesion loads. The *FN* indicates the number of false negative voxels, *TPL*, *FPL* and *FNL* indicate number of true positive, false positive and false negative lesions, respectively. ∂S and ∂R indicate sets of the border voxels for automated and reference segmentations, and $d_m(v, V)$ is the minimal of the Euclidean distances between a voxel v and voxels in a set V.

Criterion	Segmentation method								
	VL'2001 (original vs. upgraded)			GL'2009 (original vs. upgraded)			GL'2011 (original vs. upgraded)		
	Mild	Moderate	Severe	Mild	Moderate	Severe	Mild	Moderate	Severe
FP	−239*	−221*	−223.5	−7183*	−6817*	−3948.5	−4724*	−4500*	−2173
TP	−8	−20	−112	−141.5	−572†	−1018.5	−112.5†	−350†	−585
$PPV = \frac{TP}{TP+FP}$	0.01*	0.02*	0.01	0.05*	0.23*	0.18*	0.05*	0.22*	0.13*
$TPR = \frac{TP}{TP+FN}$	−0.01	−0.00	−0.01	−0.20†	−0.13†	−0.13	−0.17†	−0.08†	−0.08
$DSI = \frac{2 \times PPV \times TPR}{PPV+TPR}$	0.01*	0.01*	−0.00	0.07*	0.15*	0.01	0.07*	0.15*	0.02
$VDR = \frac{\lvert FP-FN \rvert}{TP+FN}$	−0.35*	−0.04	0.03	−10.83	−1.66*	0.00	−7.30*	−1.10*	0.00
$LPPV = \frac{TPL}{TPL+FPL}$	0.01*	0.02*	0.03*	0.02*	0.05*	0.05	0.05*	0.19*	0.21*
$LTPR = \frac{TPL}{TPL+FNL}$	−0.02	0.00	−0.02†	−0.08†	−0.15†	−0.07†	−0.16†	−0.15†	−0.09
$SD = \frac{\sum_{s \in \partial S} d_m(s, \partial R) + \sum_{r \in \partial R} d_m(r, \partial S)}{\lvert \partial S \rvert + \lvert \partial R \rvert}$	−0.53*	−0.21*	−0.03	−3.44*	−3.73*	−0.43	−3.12	−2.70*	−0.20

* Statistically significant ($p < 0.01$) improvement.
† Statistically significant ($p < 0.01$) deterioration of the performance.

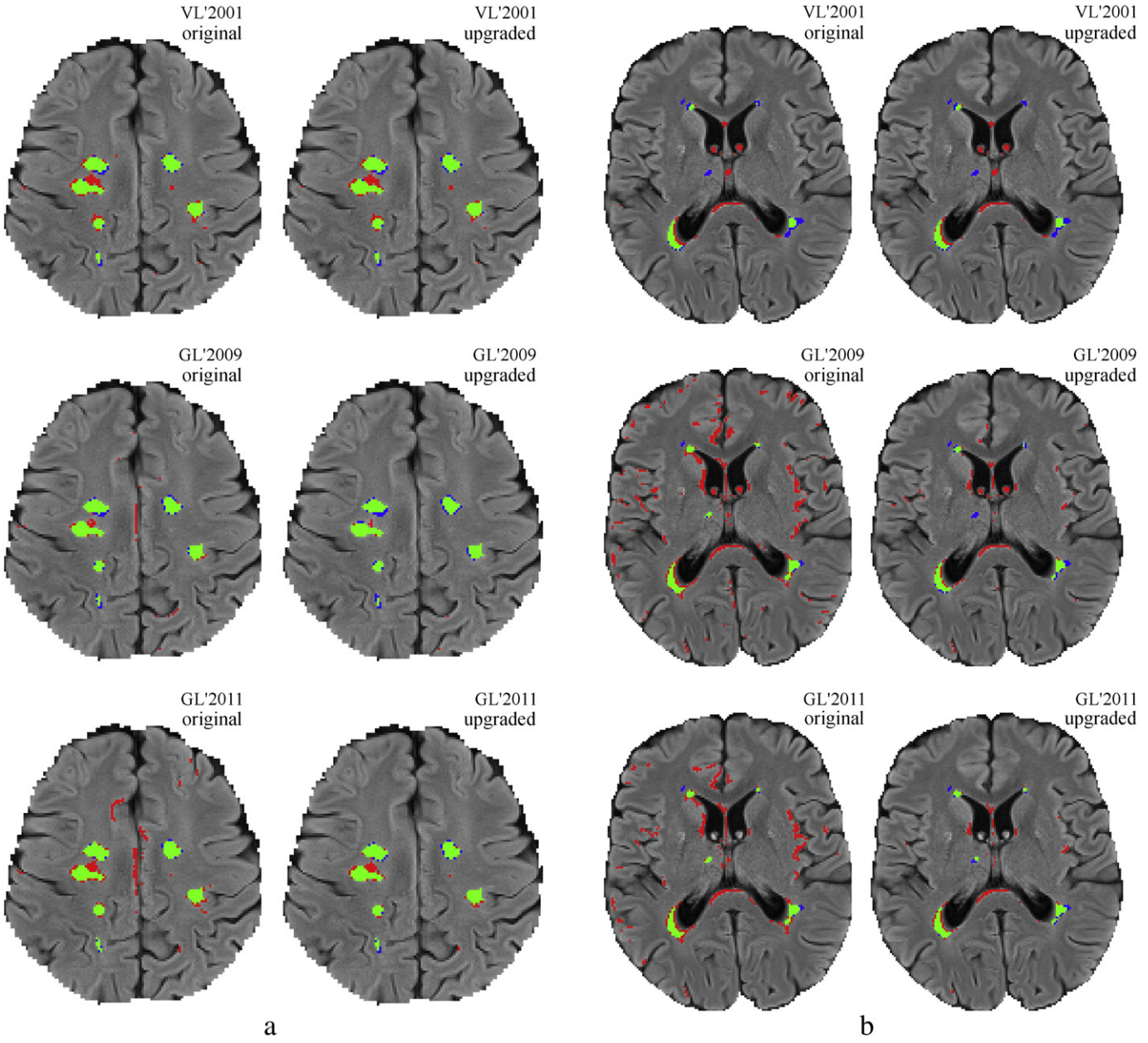

Fig. 9. Examples of segmentations by the original and upgraded (by the proposed stratified mixture model) lesion segmentation methods, applied on datasets with (a) moderate and (b) mild TLLs. Segmentation results labeled as true positives (*green*), false negatives (*blue*) or false positives (*red*) are superimposed on the FLAIR images.

with moderate and severe TLL and across all datasets the upgraded implementation of uGL'2011 (the GL'2011 upgraded by the proposed SM-GMM) clearly had superior performance. This was also verified by Wilcoxon's signed rank tests, which showed that the corresponding changes were also statistically significant (cf. Fig. 8).

Discussion

In this paper, we proposed a novel modeling approach for unsupervised estimation of complex distributions of msMR intensities of brain structures. Modeling has a crucial impact on the performance of unsupervised, and also some supervised, methods for automated segmentation of normal brain structures and pathological structures such as white-matter lesions (Lladó et al., 2012; García-Lorenzo et al., 2013). In practice, multisequence MR images of brain structures are observations of a great variety of intensity sources, affected by both acquisition imperfections and natural or pathological anatomical variations. Attempts to incorporate into generative models as many intensity sources as possible result in an increased complexity of such models and thus their robust estimation, especially in an unsupervised learning, presents a huge challenge.

As recently demonstrated by Xiao et al. (2010), the whole-brain structure intensities are better modeled by mixture models, primarily due to anatomy-specific intensity variations across different parts of the brain. Several approaches that iterate between model selection and parameter estimation (Khayati et al., 2008; Sudre et al., 2014) were proposed to facilitate the application of such models in unsupervised learning, but they are more complex as they require explicit modeling of abnormal structures like lesions. Although such approaches model intratissue spatial variability, the spatial relationship is not explicitly modeled and is not considered during the estimation.

The proposed approach to msMR intensity modeling employs stratified mixture models, wherein the central assumption is that the otherwise complex generative model of the whole-brain distribution

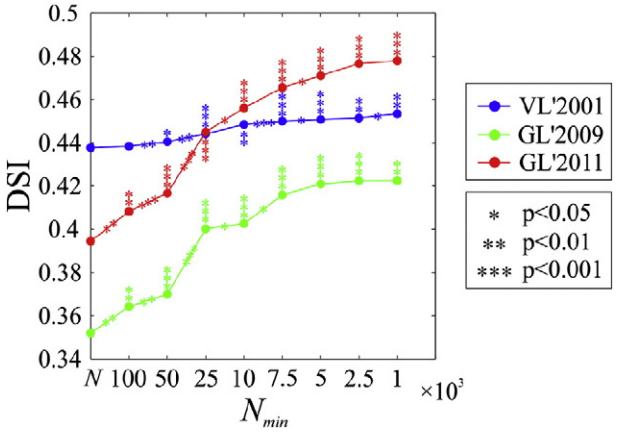

Fig. 10. The Dice similarity indices (DSI) for the three lesion segmentation methods (Table 1) tested at different values of parameter N_{min}. The DSI values were averaged over 30 MR datasets. The asterisks (*) at the graph nodes indicate significance of the difference from the original implementation, and the asterisks at the graph edges indicate significance of the difference from the previous value of the parameter according to Wilcoxon's signed rank test.

of msMR intensities reduces to a tractable parametric form at a small enough brain subregion. This is because various sources of MR intensity variability, like spatially smooth multiplicative MR bias field (Vovk et al., 2007) or natural variability of brain structures (Xiao et al., 2010), etc., can be considered negligible at small brain subregions (Shattuck et al., 2001). Although such local estimation of msMR intensity distribution has been introduced before (Scherrer et al., 2009; Tohka et al., 2010) for segmentation of MR images of normal brains, application of these methods to segmentation of MR brain images containing lesions is not straightforward. The reason is that in different small brain subregions the varying sizes and number of lesions impose a varying and possibly substantial fraction of outliers, which could adversely impact model estimation. Besides, the samples of NABS intensity observations can be highly unbalanced, which is even more prominent at local brain subregions (Fig. 1).

Our approach involves the use of a spatial stratification of the brain into subregions or strata, which are mutually exclusive, collectively exhaustive, and contain more homogeneous subsamples. Many algorithms can be proposed for this purpose and incorporated into the stratified mixture modeling approach. In this paper, we used a new spatial stratification method (Algorithm 1) which subdivides a tentative segmentation of CSF, GM and WM structures by performing rectilinear splits in a hierarchical manner. Another very important goal of the algorithm is to ensure that each stratum contains some minimal number of observations (parameter N_{min}) and some minimal fraction (parameter π_{min}) of observations for each of the three structures. This is required to obtain good mixture estimates. Herein we used the recently developed robust estimator of unbalanced mixtures (Galimzianova et al., 2015) to estimate the stratified mixture models, which provided accurate generative models of the whole-brain distribution of msMR intensities of NABS on 30 real msMR images of MS patients.

The main advantage of the proposed stratified mixture modeling is the stratum-wise estimation of simple three-component mixtures of normal-appearing structures. Such a model was assumed valid within a stratum that captures a small region of the brain, since the intensity variations may be neglected. Hence, underestimation is not a problem. The main benefit is that the model is able to accurately capture the overlap of the intensity distributions of normal structures, whereas overfitting within stratum is avoided by using such a simple but constrained model and robust estimation (Galimzianova et al., 2015). We tested two models, one based on parameter- and the other on model-wise recombination over strata into a whole-brain model, i.e., SP-GMM and SM-GMM, respectively. Of the two proposed models, the SM-GMM better captures the MR intensity distributions according to better goodness-of-fit and more accurate segmentation results (Figs. 4 and 5, respectively). A general observation was that with higher number of strata the estimated mixture model better adapts to the actual MR intensity distribution.

Other similarly complex models like the M-GMM were previously used to capture spatially-varying structural properties implicitly (Ashburner and Friston, 2005; Xiao et al., 2010). In addition to rendering the problem of selecting optimal structure-specific number of components (or model order), the use of M-GMM has another important deficiency. The M-GMM attempts to model the overall MR intensity distribution of a certain structure, in which several intensity distributions from various locations and structure interfaces may substantially overlap. Clearly, the estimation of model parameters based on the spatial stratification of observed multi-sequence MR intensities is a more robust approach compared to a direct estimation of model parameters based solely on the MR intensities. This is apparent when comparing the performance of M-GMM vs. SP-GMM, which is a simple three-component mixture model estimated through the proposed spatial stratification approach (Figs. 4 and 5). For high number of strata the SP-GMM even outperformed the M-GMM. The increased flexibility of SM-GMM, achieved by recombining local models over strata, only further improved the modeling of msMR intensities.

To analyze dependence of the performance on the implementation of spatial stratification, the MR intensity modeling experiments in Section MR intensity modeling were performed with *k*-means based spatial stratification. The obtained JD and DSI values were comparable to the spatial stratification in Algorithm 1, showing similar patterns of JD and DSI with respect to N_{min} as in Figs. 4 and 5. Hence, as long as the stratification criteria (cf. Section Robust model estimation) are met, a particular implementation of spatial stratification does not seem to have a large impact on the performance of the proposed stratified mixture modeling.

The proposed stratified mixture model was implemented into three unsupervised lesion segmentation methods (Van Leemput et al., 2001; García-Lorenzo et al., 2009, 2011), which detect lesions as model outliers, and was validated against the original implementations and three other state-of-the-art methods on 30 msMR image datasets of MS patients with various TLLs. Compared to original implementations, the upgraded implementations using the SM-GMM generally exhibited improved performance of lesion segmentation, measured as the overlap between the reference manual and obtained automated segmentations (Figs. 6 and 7). Comparison to the state-of-the-art methods LST and Souplet et al. showed inferior performance with respect to both original and upgraded implementation of the three tested methods (VL'2001, GL'2009 and GL'2011). On the other hand, LTOADS method outperformed all methods on datasets with mild TLL, while on datasets with moderate and severe TLL and across all datasets the uGL'2011 based on SM-GMM had superior performance (Fig. 8).

In general, the DSIs were much lower on datasets with mild than on those with moderate and severe TLL. The main reason is that on datasets with small TLL, the usually large amount of FPs has a much higher impact on the overall DSI score. The value of DSI may thus be misleading in terms of usefulness of the obtained segmentation, however, as Fig. 9 demonstrates, the use of the stratified mixture model substantially reduced the amount of FPs, especially on datasets with mild TLL (Table 2). The DSIs either remained the same or improved for all three tested methods.

The use of the stratified mixture model also resulted in higher consistency of lesion segmentation across patient datasets and among the three tested methods (Fig. 9). Consistent performance is one of the critical requirements for the application of lesion segmentation methods for diagnosis and management of treatment of diseases causing white-matter lesions (Vrenken et al., 2013), and for large multi-center clinical studies, as it ensures that quantitative measurements of lesions are consistent between different patients and on the same patient over time.

Conclusions

The novel stratified mixture modeling approach results in accurate and robust unsupervised estimation of the whole-brain MR intensity model. By reducing the otherwise complex whole-brain model to a tractable parametric mixture model through spatial stratification of the brain into subregions and performing robust local model estimation, the whole-brain model can be accurately recombined from the local models. The stratified mixture modeling was incorporated into three unsupervised lesion segmentation methods and, compared to the original modeling approaches, significantly improved lesion segmentation on 30 real msMR images of patients with MS according to increased Dice similarity indices and lower number of FPs.

Acknowledgments

This research was supported by the Slovenian Research Agency under grants J2-5473, L2-5472, and J7-6781.

References

Akselrod-Ballin, A., Galun, M., Gomori, J.M., Filippi, M., Valsasina, P., Basri, R., Brandt, A., 2009. Automatic segmentation and classification of multiple sclerosis in multichannel MRI. IEEE Trans. Biomed. Eng. 56, 2461–2469.

Anbeek, P., Vincken, K.L., van Osch, M.J.P., Bisschops, R.H.C., van der Grond, J., 2004. Probabilistic segmentation of white matter lesions in MR imaging. NeuroImage 21, 1037–1044.

Ashburner, J., Friston, K.J., 2005. Unified segmentation. NeuroImage 26, 839–851.

Breunig, R., 2008. Nonparametric density estimation for stratified samples. Stat. Probab. Lett. 78, 2194–2200.

Cocosco, C.A., Zijdenbos, A.P., Evans, A.C., 2003. A fully automatic and robust brain MRI tissue classification method. Med. Image Anal. 7, 513–527.

Cuadra, M., Cammoun, L., Butz, T., Cuisenaire, O., Thiran, J., 2005. Comparison and validation of tissue modelization and statistical classification methods in T1-weighted MR brain images. IEEE Trans. Med. Imaging 24, 1548–1565.

Datta, S., Narayana, P.A., 2013. A comprehensive approach to the segmentation of multichannel three-dimensional MR brain images in multiple sclerosis. NeuroImage Clin. 2, 184–196.

Debette, S., Markus, H.S., 2010. The clinical importance of white matter hyperintensities on brain magnetic resonance imaging: systematic review and meta-analysis. BMJ 341 (c3666–c3666).

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Ser. B Methodol. 39, 1–38.

Dice, L., 1945. Measures of the amount of ecologic association between species. Ecology 26, 297–302.

Elliott, C., Arnold, D., Collins, D., Arbel, T., 2013. Temporally consistent probabilistic detection of new multiple sclerosis lesions in brain MRI. IEEE Trans. Med. Imaging 32, 1490–1503.

Evans, A., Collins, D., Mills, S., Brown, E., Kelly, R., Peters, T., 1993. 3D statistical neuroanatomical models from 305 MRI volumes. Proc. Nuclear Science Symposium and Medical Imaging Conference vol. 3, pp. 1813–1817.

Freifeld, O., Greenspan, H., Goldberger, J., 2009. Multiple sclerosis lesion detection using constrained GMM and curve evolution. Int. J. Biomed. Imaging 2009, 715124.

Galimzianova, A., Pernuš, F., Likar, B., Špiclin, Z., 2015. Robust estimation of unbalanced mixture models on samples with outliers. IEEE Trans. Pattern Anal. Mach. Intell. 37, 2273–2285.

García-Lorenzo, D., Prima, S., Collins, L., Arnold, D.L., Morrissey, S.P., Barillot, C., 2008. Combining robust expectation maximization and mean shift algorithms for multiple sclerosis brain segmentation. Proc. MICCAI Workshop on Medical Image Analysis on Multiple Sclerosis (MIAMS'08), pp. 82–91.

García-Lorenzo, D., Lecoeur, J., Arnold, D.L., Collins, D.L., Barillot, C., 2009. Multiple sclerosis lesion segmentation using an automatic multimodal graph cuts. Proc. 12th Int. Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI 2009), pp. 584–591.

García-Lorenzo, D., Prima, S., Arnold, D.L., Collins, D.L., Barillot, C., 2011. Trimmed-likelihood estimation for focal lesions and tissue segmentation in multisequence MRI for multiple sclerosis. IEEE Trans. Med. Imaging 30, 1455–1467.

García-Lorenzo, D., Francis, S., Narayanan, S., Arnold, D.L., Collins, D.L., 2013. Review of automatic segmentation methods of multiple sclerosis white matter lesions on conventional magnetic resonance imaging. Med. Image Anal. 17, 1–18.

Geremia, E., Clatz, O., Menze, B.H., Konukoglu, E., Criminisi, A., Ayache, N., 2011. Spatial decision forests for MS lesion segmentation in multi-channel magnetic resonance images. NeuroImage 57, 378–390.

Gudbjartsson, H., Patz, S., 1995. The Rician distribution of noisy MRI data. Magn. Reson. Med. 34, 910–914.

Harmouche, R., Subbanna, N., Collins, D., Arnold, D., Arbel, T., 2015. Probabilistic multiple sclerosis lesion classification based on modelling regional intensity variability and local neighbourhood information. IEEE Trans. Biomed. Eng. 62, 1281–1292.

Iglesias, J., Liu, C.Y., Thompson, P., Tu, Z., 2011. Robust brain extraction across datasets and comparison with publicly available methods. IEEE Trans. Med. Imaging 30, 1617–1634.

Kadoury, S., Erus, G., Zacharaki, E.I., Paragios, N., Davatzikos, C., 2012. Manifold-constrained embeddings for the detection of white matter lesions in brain MRI. Proc. 9th IEEE Int. Symposium on Biomedical Imaging (ISBI 2012), pp. 562–565.

Karimaghaloo, Z., Shah, M., Francis, S., Arnold, D., Collins, D., Arbel, T., 2012. Automatic detection of gadolinium-enhancing multiple sclerosis lesions in brain MRI using conditional random fields. IEEE Trans. Med. Imaging 31, 1181–1194.

Khayati, R., Vafadust, M., Towhidkhah, F., Nabavi, M., 2008. Fully automatic segmentation of multiple sclerosis lesions in brain MR FLAIR images using adaptive mixtures method and Markov random field model. Comput. Biol. Med. 38, 379–390.

Komodakis, N., Tziritas, G., Paragios, N., 2008. Performance vs computational efficiency for optimizing single and dynamic MRFs: setting the state of the art with primal-dual strategies. Comput. Vis. Image Underst. 112, 14–29.

Lao, Z., Shen, D., Liu, D., Jawad, A.F., Melhem, E.R., Launer, L.J., Bryan, R.N., Davatzikos, C., 2008. Computer-assisted segmentation of white matter lesions in 3D MR images using support vector machine. Acad. Radiol. 15, 300–313.

Lladó, X., Oliver, A., Cabezas, M., Freixenet, J., Vilanova, J.C., Quiles, A., Valls, L., Rami-Torrent, L., Rovira, A., 2012. Segmentation of multiple sclerosis lesions in brain MRI: a review of automated approaches. Inf. Sci. 186, 164–185.

Manjón, J.V., Coupé, P., Martí-Bonmatí, L., Collins, D.L., Robles, M., 2010. Adaptive non-local means denoising of MR images with spatially varying noise levels. J. Magn. Reson. Imaging 31, 192–203.

McLachlan, G., Peel, D., 2005. Finite mixture models. Wiley Series in Probability and Statistics. Wiley, Hoboken, NJ, USA.

Neykov, N., Filzmoser, P., Dimova, R., Neytchev, P., 2007. Robust fitting of mixtures using the trimmed likelihood estimator. Comput. Stat. Data Anal. 52, 299–308.

Prastawa, M., Gerig, G., 2008. Brain lesion segmentation through physical model estimation. Proc. 4th Int. Symposium on Advances in Visual Computing (ISVC 2008), pp. 562–571.

Prins, N.D., Scheltens, P., 2015. White matter hyperintensities, cognitive impairment and dementia: an update. Nat. Rev. Neurol. 11, 157–165.

Rocca, M.A., Anzalone, N., Falini, A., Filippi, M., 2013. Contribution of magnetic resonance imaging to the diagnosis and monitoring of multiple sclerosis. Radiol. Med. 118, 251–264.

Rouaïnia, M., Medjram, M.S., Doghmane, N., 2006. Brain MRI segmentation and lesions detection by EM algorithm. Proc. of World Academy of Science, Engineering and Technology, pp. 301–304.

Rovira, A., Wattjes, M.P., Tintoré, M., Tur, C., Yousry, T.A., Sormani, M.P., De Stefano, N., Filippi, M., Auger, C., Rocca, M.A., Barkhof, F., Fazekas, F., Kappos, L., Polman, C., Miller, D., Montalban, X., on behalf of the MAGNIMS study Group, 2015. Evidence-based guidelines: MAGNIMS consensus guidelines on the use of MRI in multiple sclerosis-clinical implementation in the diagnostic process. Nat. Rev. Neurol. 11, 471–482.

Roy, S., He, Q., Carass, A., Jog, A., Cuzzocreo, J.L., Reich, D.S., Prince, J., Pham, D., 2014. Example based lesion segmentation. Proc. SPIE Medical Imaging (pp. 90341Y–90341Y).

Scherrer, B., Forbes, F., Garbay, C., Dojat, M., 2009. Distributed local MRF models for tissue and structure brain segmentation. IEEE Trans. Med. Imaging 28, 1278–1295.

Schmidt, P., Gaser, C., Arsic, M., Buck, D., Frschler, A., Berthele, A., Hoshi, M., Ilg, R., Schmid, V.J., Zimmer, C., Hemmer, B., Mhlau, M., 2012. An automated tool for detection of FLAIR-hyperintense white-matter lesions in Multiple Sclerosis. NeuroImage 59, 3774–3783.

Shah, M., Xiao, Y., Subbanna, N., Francis, S., Arnold, D.L., Collins, D.L., Arbel, T., 2011. Evaluating intensity normalization on MRIs of human brain with multiple sclerosis. Med. Image Anal. 15, 267–282.

Shattuck, D.W., Sandor-Leahy, S.R., Schaper, K.A., Rottenberg, D.A., Leahy, R.M., 2001. Magnetic resonance image tissue classification using a partial volume model. NeuroImage 13, 856–876.

Shiee, N., Bazin, P.L., Ozturk, A., Reich, D.S., Calabresi, P.A., Pham, D.L., 2010. A topology-preserving approach to the segmentation of brain images with multiple sclerosis lesions. NeuroImage 49, 1524–1535.

Souplet, J.C., Lebrun, C., Ayache, N., Malandain, G., 2008. An automatic segmentation of T2-FLAIR multiple sclerosis lesions. Proc. MICCAI Grand Challenge Workshop: Multiple Sclerosis Lesion Segmentation Challenge, pp. 1–11 (URL: http://hdl.handle.net/10380/1451).

Steenwijk, M.D., Pouwels, P.J.W., Daams, M., van Dalen, J.W., Caan, M.W.A., Richard, E., Barkhof, F., Vrenken, H., 2013. Accurate white matter lesion segmentation by k nearest neighbor classification with tissue type priors (kNN-TTPs). NeuroImage Clin. 3, 462–469.

Sudre, C.H., Cardoso, M.J., Bouvy, W., Biessels, G.J., Barnes, J., Ourselin, S., 2014. Bayesian model selection for pathological data. Proc. 17th Int. Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI 2014), pp. 323–330.

Sweeney, E.M., Shinohara, R.T., Shiee, N., Mateen, F.J., Chudgar, A.A., Cuzzocreo, J.L., Calabresi, P.A., Pham, D.L., Reich, D.S., Crainiceanu, C.M., 2013. OASIS is automated statistical inference for segmentation, with applications to multiple sclerosis lesion segmentation in MRI. NeuroImage Clin. 2, 402–413.

Thompson, S.K., 2012. Stratified sampling. Sampling. Wiley, pp. 141–156.

Tohka, J., Dinov, I.D., Shattuck, D.W., Toga, A.W., 2010. Brain MRI tissue classification based on local markov random fields. Magn. Reson. Imaging 28, 557–573.

Tomas-Fernandez, X., Warfield, S., 2015. A model of population and subject (MOPS) intensities with application to multiple sclerosis lesion segmentation. IEEE Trans. Med. Imaging 34, 1349–1361.

Tustison, N., Avants, B., Cook, P., Zheng, Y., Egan, A., Yushkevich, P., Gee, J., 2010. N4ITK: improved N3 bias correction. IEEE Trans. Med. Imaging 29, 1310–1320.

Van Leemput, K., Maes, F., Vandermeulen, D., Colchester, A., Suetens, P., 2001. Automated segmentation of multiple sclerosis lesions by model outlier detection. IEEE Trans. Med. Imaging 20, 677–688.

Vovk, U., Pernus, F., Likar, B., 2007. A review of methods for correction of intensity inhomogeneity in MRI. IEEE Trans. Med. Imaging 26, 405–421.

Vrenken, H., Jenkinson, M., Horsfield, M.A., Battaglini, M., Schijndel, R.A.V., Rostrup, E., Geurts, J.J.G., Fisher, E., Zijdenbos, A., Ashburner, J., Miller, D.H., Filippi, M., Fazekas, F., Rovaris, M., Rovira, A., Barkhof, F., Stefano, N.D., Group, M.S., 2013. Recommendations to improve imaging and analysis of brain lesion load and atrophy in longitudinal studies of multiple sclerosis. J. Neurol. 260, 2458–2471.

Wakimoto, K., 1971. Stratified random sampling (II) estimation of the population covariance. Ann. Inst. Stat. Math. 23, 327–337.
Warfield, S.K., Kaus, M., Jolesz, F.A., Kikinis, R., 2000. Adaptive, template moderated, spatially varying statistical classification. Med. Image Anal. 4, 43–55.
Weiss, N., Rueckert, D., Rao, A., 2013. Multiple sclerosis lesion segmentation using dictionary learning and sparse coding. Proc. 16th Int. Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI 2013), pp. 735–742.
Wu, X., Bricq, S., Collet, C., 2011. Brain MRI segmentation and lesion detection using generalized Gaussian and Rician modeling. Proc. SPIE Medical Imaging (pp. 90341Y–90341Y).
Xiao, Y., Shah, M., Francis, S., Arnold, D., Arbel, T., Collins, D., 2010. Optimal Gaussian mixture models of tissue intensities in brain MRI of patients with multiple-sclerosis. Proc. MICCAI Workshop on Machine Learning in Medical Imaging, pp. 165–173.