

Spatial Modeling of Multiple Sclerosis for Disease Subtype Prediction

Bernd Taschler¹, Tian Ge², Kerstin Bendfeldt³, Nicole Müller-Lenke³,
Timothy D. Johnson⁴, and Thomas E. Nichols²

¹ Centre for Complexity Science, University of Warwick, Coventry, United Kingdom

² Department of Statistics, University of Warwick, Coventry, United Kingdom

³ Medical Image Analysis Center, University Hospital Basel, Basel, Switzerland

⁴ Department of Biostatistics, University of Michigan, Ann Arbor, MI, USA

Abstract. Magnetic resonance imaging (MRI) has become an essential tool in the diagnosis and managing of Multiple Sclerosis (MS). Currently, the assessment of MS is based on a combination of clinical scores and subjective rating of lesion images by clinicians. In this work we present an objective 5-way classification of MS disease subtype as well as a comparison between three different approaches. First we propose two spatially informed models, a Bayesian Spatial Generalized Linear Mixed Model (BSGLMM) and a Log Gaussian Cox Process (LGCP). The BSGLMM accounts for the binary nature of lesion maps and the spatial dependence between neighboring voxels, and the LGCP accounts for the random spatial variation in lesion location. Both improve upon mass univariate analyses that ignore spatial dependence and rely on some level of arbitrarily defined smoothing of the data. As a comparison, we consider a machine learning approach based on multi-class support vector machine (SVM). For the SVM classification scheme, unlike previous work, we use a large number of quantitative features derived from three MRI sequences in addition to traditional demographic and clinical measures. We show that the spatial models outperform standard approaches with average prediction accuracies of up to 85%.

1 Introduction

Multiple Sclerosis (MS) is a chronic inflammatory and demyelinating disease of the central nervous system that leaves behind scarred tissue in various regions of the brain and the spinal cord. MS patients can be grouped into five distinct clinical categories according to disease progression, which differs significantly between these subtypes [3]. Currently, in the assessment of MS, MRI data is to a large extent only used in a qualitative way to assess the dissemination of lesions in space and time; the most common quantitative measure is ‘lesion load’, the total lesion volume. Previous work on using lesion load for classification have shown mixed results, see e.g. [12] and [9]. Other studies have shown that conventional MRI measures have rather low predictive value and are therefore poor indicators for determining the clinical outcomes in MS [8].

Relapsing-remitting (RLRM) is the most common subtype of MS (70-80% of all MS patients) and is characterized by episodes of acute relapses and periods of remission. Secondary progressive (SCP), primary progressive (PRP) and progressive relapsing (PRL) forms of MS are comparatively rare and differ in severity and progression of the disease. Clinically isolated syndrome (CIS) describes a singular instance of inflammatory demyelination. Patients with CIS may or may not subsequently develop one of the other courses of MS.

We propose an objective classification of MS subtype using two spatially informed approaches, a Bayesian spatial generalized linear mixed model (BS-GLMM) on voxel-wise lesion maps and a Log Gaussian Cox process (LGCP) model on lesion location data, and contrast our results with a classification model based on support vector machine (SVM). In addition to traditional demographic and clinical measures, the features for SVM include aspects of lesion geometry, measured by Minkowski functionals, and statistics of the image intensities within individual lesions. Each of these three approaches has relative merits: The BSGLMM has the interpretability of a traditional regression model but can only account for local spatial dependence; the LGCP doesn't enjoy a regression framework but explicitly accounts for (larger scale) spatial variation in lesion location; and while the previous two only operate on a single type of MRI image, the SVM uses a rich constellation of features from all available MRI sequences. While our BSGLMM method is in press [4], the other two approaches and such a detailed comparison of predictive accuracy are to our knowledge completely novel.

2 Methods

2.1 Bayesian Spatial Generalized Linear Mixed Model

As an extension of generalized linear mixed models, a spatial generalized linear mixed model (BSGLMM) has covariates $\mathbf{x}(s)$ and coefficients $\boldsymbol{\beta}(s)$ that comprise the systematic component, $\eta(s) = \mathbf{x}^T(s)\boldsymbol{\beta}(s)$; here $s \in \mathbb{R}^3$. The form of the systematic component includes any combination of spatially constant or varying covariates and coefficients. For computational reasons [4], we use the probit link function, Φ^{-1} . Let $Y_i(s_j) \in \{0, 1\}$ denote a Bernoulli random variable for subject i , indicating the presence or absence of a lesion at voxel s_j , and let $\Pr[Y_i(s_j)=1] \equiv p_i(s_j)$. Thus, the random, link and systematic components in our BSGLMM are specified as

$$[Y_i(s_j)|p_i(s_j)] \sim \text{Bernoulli}[p_i(s_j)], \quad (1)$$

$$\Phi^{-1}\{E[Y_i(s_j)|p_i(s_j)]\} = \eta_i(s_j), \quad (2)$$

$$\eta_i(s_j) = \mathbf{x}_i^T[\boldsymbol{\alpha} + \boldsymbol{\beta}(s_j)] + w(s_j)\gamma. \quad (3)$$

The parameters $\boldsymbol{\alpha}$ and γ represent fixed effects, whereas the elements of $\boldsymbol{\beta}(s_j)$ are spatially varying random effects and $w(s_j)$ is an optional spatially varying covariate. The parameters $\boldsymbol{\alpha}$ and γ have flat, improper, uninformative priors and $\boldsymbol{\beta}(s_j)$ have zero-centered multivariate conditional autoregressive model (MCAR)

priors. The full conditional posterior distribution of $\beta(s)$ is multivariate normal with an improper, uninformative Wishart prior on the hyperparameter Σ . By Brook’s lemma, the joint posterior is therefore proportional to

$$\pi[\beta|\Sigma] \propto \exp\{-1/2 \sum_{s_i \sim s_j} [\beta(s_i) - \beta(s_j)]^T \Sigma^{-1} [\beta(s_i) - \beta(s_j)]\}, \quad (4)$$

where $s_i \sim s_j$ indicates two neighboring voxels that share a common face. Note that the degree of spatial regularization induced by the MCAR prior, as parameterized by Σ , is *not* fixed, but rather is estimated from the data. Our model comprises 12 subject-specific covariates: clinical subtype (coded as five dummy variables), sex, age, disease duration, seven EDSS subscores and PASAT score. Additionally, the white matter probability map represents one spatially varying covariate that is shared by all subjects and accounts for gross differences in lesion incidence over the brain. Leave-one-out-cross-validation (LOOCV) and importance sampling is employed, as described in Section 2.2. See [4] for details.

Naïve Bayesian Classifier. For comparison we also perform LOOCV on a naïve Bayesian classifier. The full T₂ binary lesion mask is used as feature vector and the classifier assumes mutual independence between each element (i.e. voxel). The naïve Bayesian classifier consists of a binomial model at each voxel for each subtype, where we assign Jeffrey’s prior, i.e. Beta(0.5,0.5), to the conditional probability of having a lesion at voxel j given a certain subtype.

2.2 Log Gaussian Cox Process

To better account for random spatial variation in the lesion data, we consider summarizing each lesion with their center of mass. FSL’s cluster command was used to find the center of mass of each lesion (order 6 connectivity) in MNI atlas coordinates, on T₁ black-hole and T₂ lesion masks. Sample size was slightly reduced as some subjects had no lesions (T₁: 232 and T₂: 248 data points).

Our model builds on the principles of a spatial Poisson process. A point process X is a Poisson point process with (non-random) intensity $\lambda : \mathbb{R}^3 \rightarrow \mathbb{R}^+$ if a) $N(B) \sim \text{Poi}(\Lambda(B))$, $B \subseteq \mathbb{R}^3$, as long as $\Lambda(B) = \int_B \lambda(y) dy < \infty$; b) conditional on $N(B)$, the points in X are i.i.d. with density $\lambda(y)/\Lambda(B)$. A generalization of a Poisson process is a Cox process, where the Poisson intensity function is a realization of a random field. Precisely, if λ is a non-negative random field, then X is a Cox process driven by λ if $X|\lambda$ is an (inhomogeneous) Poisson process with intensity λ [10]. The special case of a Log Gaussian Cox process (LGCP) is obtained by letting $Y = \ln \lambda$ be a Gaussian random field, i.e. a random process whose finite dimensional distributions are Gaussian, $Y \sim \text{GP}\{m(\cdot), c(\cdot, \cdot)\}$ [13]. The distribution of (X, Y) is completely determined by the mean and covariance function: $m(r) = E[Y(r)]$ and $c(r, s) = \text{Cov}[Y(r), Y(s)]$. The intensity function of the LGCP is given by

$$\lambda(r) = \exp [m(r) + c(r, r)/2], \quad (5)$$

where we assume that $c(r, s)$ is isotropic and translation invariant with the form $c(r, s) = \sigma^2 \exp[-\|r - s\|^\delta / \alpha]$. The exponential term represents the power correlation function and $\delta \in [0, 2]$, which for $\delta=1$ gives an exponential and for $\delta=2$ a Gaussian correlation; σ^2 denotes the variance and $\alpha > 0$ is a correlation parameter.

Estimation of the scale α of the process is challenging. Thus, before fitting the model, we use minimum contrast estimation to estimate (and fix) the scale α and the exponent δ via a fit of $c(r, s)$ to the pair-correlation-function of all data [11]. Benefits of the LGCP include its nonparametric nature and thus flexibility in modeling spatial data, and its computational tractability resulting from its relation to multivariate Gaussian distributions.

The posterior distribution is estimated via Markov chain Monte Carlo. We use Bayes' theorem and an importance sampling approach [5] to obtain leave-one-out-cross-validation without running our sampler N times. An estimate of the predictive probability that subject i is categorized as subtype g_i is obtained from a sample from the posterior, weighted to discount the (independent) contribution of subject i to the likelihood.

2.3 Geometric Measures and SVM

The two previous spatial models require computationally intensive and carefully tuned MCMC sampling procedures. To see if a simpler, machine learning approach could provide similar prediction accuracies we consider a SVM model based on detailed geometric and texture features of three types of MRI lesion data. We use Minkowski functionals to describe the geometry of each lesion. In \mathbb{R}^3 , Minkowski functionals are directly related to the geometric quantities volume, surface area, mean breadth and Euler-Poincaré characteristic [2,7]. We employ a radial basis function kernel and adopt an one-vs-one approach [1] for multi-class classification problems based on pairwise classifiers, combined with a majority voting scheme to make predictions.

For model evaluation we carry out stratified k -fold cross-validation, where k is determined by the number of elements in the smallest class. In our case, one group (PRL) consists only of ten subjects, thus $k=10$. Within each cross-validation fold, a parameter optimization via grid search is performed and different parameter sets are evaluated by means of (nested) cross-validation. Along with demographic and clinical covariates, the full feature set includes the Euler-Poincaré characteristic, and, mean, median, maximum, minimum and standard deviation across the whole brain for lesion volume, surface area and mean breadth. In addition, we include the fraction of gray matter (GM) volume to whole brain volume as well as total, mean, median and standard deviation of intra-lesion intensities. In order to encode spatial information about the location of individual lesions, we also look at splitting these whole brain measures according to 13 regions of interest (ROI's) based on white matter (WM) track segmentations. All features are centered and normalized to zero mean and unit variance.

Model evaluation. The predictive performance of all models is estimated by computing the confusion matrix. The average accuracy is given by the mean of the diagonal elements. Due to the large discrepancy in the number of RLRM subjects (173 out of 250) with respect to other subtypes, the average accuracy is more representative of total performance than the overall accuracy. (Note that chance level of average accuracy is 20%.)

Data. The full data set used for our analysis consists of 250 subjects scanned on a 1.5T Siemens Avanto scanner, collecting T_1 -weighted (T_1), T_2 -weighted (T_2) and T_1 -weighted Gd-enhanced (T_1 -Gd) images; native resolution is $0.9766 \times 0.9766 \times 3.0\text{mm}^3$. White matter lesion masks for T_1 black-hole, T_2 and T_1 -Gd images were created in native space by a semi-automatic procedure [6]. Each scan was then affine registered to MNI space at $1 \times 1 \times 1\text{mm}^3$ resolution using trilinear interpolation, and thresholded at 0.5 to retain binary values. All patients have been categorized into one of the five MS subgroups following the current diagnostic standard (cf. McDonald criteria). Patient specific covariates include sex, age, disease duration, the seven subscores of the Expanded Disability Status Scale (EDSS) and the Paced Auditory Serial Addition Test (PASAT) score.

3 Results

Figure 1 shows the empirical lesion probabilities (left) and the estimated mean posterior probabilities (right) from our BSGLMM. Lesion incidence for CIS (top panel) differs significantly from the other subtypes. This is likely due to the fact that CIS patients have the lowest lesion load. A sagittal slice of standardized parameter estimates of PASAT and total EDSS score is given in Figure 2 (top left and right). PASAT scores are negatively and EDSS scores are positively correlated with lesion occurrence throughout regions of high lesion counts, reflecting higher levels of disability (lower PASAT and higher EDSS scores correspond to more severe MS). Note the strong positive correlation of EDSS scores with lesion occurrence in the minor and major forceps (indicated by arrows). In comparison, Figure 3 shows an axial slice of the mean posterior for the LGCP.

With respect to SVM, an exhaustive combinatorial search across all features is computationally infeasible. Instead we consider a subset of possible feature combinations which are guided by the magnitude of weights of the support vectors. The best performance was obtained by using GM volume, T_2 median volume split into WM ROI's, whole brain summaries for standard deviation of T_1 mean breadth, the median of T_2 mean breadth, and T_1 and T_1 -Gd total intra-lesion intensities, alongside demographic and clinical covariates; resulting in an average accuracy of 47.8% (Table 2). A feature set containing GM volume by lobar ROI's, T_1 and T_2 lesion count and lesion volume by WM ROI's, but excluding any demographic or clinical covariates yields a value of 39.4% for average accuracy. This is well above chance level and indicates that not the covariates but instead the information contained in MRI data is predominantly driving the predictions. Furthermore, the relevance of different geometry and intensity

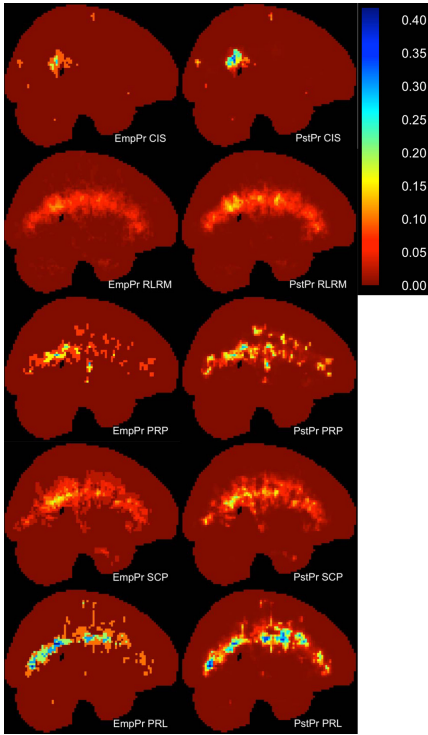


Fig. 1. BSGLMM: Comparison of the empirical probabilities (*left*) and the estimated mean posterior probabilities (*right*) for all subtypes [4].

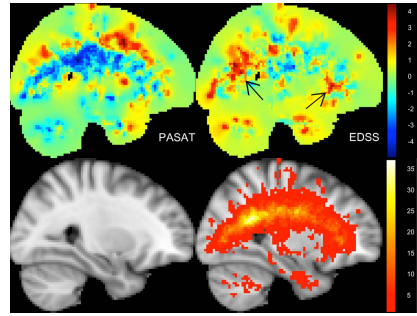


Fig. 2. Statistical significance maps for PASAT and EDSS (*top*). Arrows indicate minor and major forceps; T₁ template with empirical counts overlaid (*bottom*).

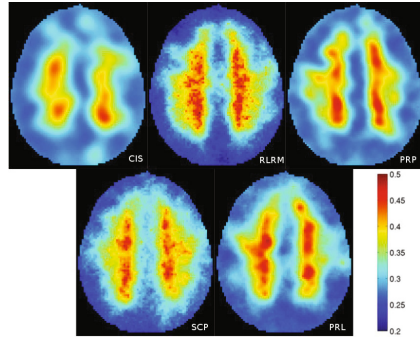


Fig. 3. LGCP: Axial slice of mean posterior probabilities for all subtypes

based features varies depending on which groups are involved in the classification. For instance, maximum T₁ lesion area is very significant in RLRM vs. PRP, and T₁ maximum lesion area and mean breadth are prominent in three out of four classifiers concerning RLRM. In general, a comparison across all SVM pairwise classifiers indicates that median and maximum measures of lesion volume, area and mean breadth for a single lesion are often more meaningful than other features, and that the Euler-Poincaré characteristic is more significant than a simple lesion count.

Tables 1–6 show confusion matrices for all models. A comparison of classification performance reveals the superiority of the spatial approaches. The naïve Bayesian approach is only slightly above chance level and although SVM is doing better, it does not reach satisfactory levels. The BSGLMM shows strong prediction results with an average accuracy of 78.3% for T₁ and 82.3% for T₂ data. As can be seen from Tables 3 and 4, misclassification predominantly occurs in the CIS subtype. These patients tend to have fewer and smaller lesions than those correctly classified.

Table 1. Naïve Bayes (T_2): overall & average accuracy: 0.552 & **0.245**.

	CIS	RLRM	PRP	SCP	PRL
CIS	0.000	1.000	0.000	0.000	0.000
RLRM	0.046	0.757	0.017	0.093	0.087
PRP	0.077	0.769	0.000	0.077	0.077
SCP	0.023	0.744	0.023	0.070	0.140
PRL	0.000	0.600	0.000	0.000	0.400

Table 2. best SVM feature set: overall & average accuracy: 0.560 & **0.478**.

	CIS	RLRM	PRP	SCP	PRL
CIS	0.818	0.182	0.000	0.000	0.000
RLRM	0.162	0.584	0.058	0.081	0.116
PRP	0.000	0.231	0.308	0.231	0.231
SCP	0.023	0.093	0.116	0.581	0.186
PRL	0.000	0.400	0.200	0.300	0.100

Table 3. BSGLMM (T_1): overall & average accuracy: 0.654 & **0.783**.

	CIS	RLRM	PRP	SCP	PRL
CIS	1.000	0.000	0.000	0.000	0.000
RLRM	0.348	0.598	0.030	0.024	0.000
PRP	0.083	0.000	0.917	0.000	0.000
SCP	0.216	0.054	0.027	0.703	0.000
PRL	0.100	0.100	0.100	0.000	0.700

Table 4. BSGLMM (T_2): overall & average accuracy: 0.748 & **0.823**.

	CIS	RLRM	PRP	SCP	PRL
CIS	1.000	0.000	0.000	0.000	0.000
RLRM	0.238	0.713	0.006	0.043	0.000
PRP	0.083	0.000	0.917	0.000	0.000
SCP	0.162	0.000	0.054	0.784	0.000
PRL	0.200	0.000	0.000	0.100	0.700

Table 5. LGCP (T_1): overall & average accuracy: 0.753 & **0.510**.

	CIS	RLRM	PRP	SCP	PRL
CIS	0.250	0.375	0.125	0.125	0.125
RLRM	0.056	0.850	0.069	0.019	0.006
PRP	0.167	0.333	0.333	0.083	0.083
SCP	0.071	0.119	0.119	0.667	0.024
PRL	0.111	0.222	0.111	0.111	0.445

Table 6. LGCP (T_2): overall & average accuracy: 0.895 & **0.851**.

	CIS	RLRM	PRP	SCP	PRL
CIS	0.800	0.000	0.100	0.100	0.000
RLRM	0.040	0.913	0.029	0.006	0.012
PRP	0.000	0.154	0.769	0.077	0.000
SCP	0.023	0.047	0.000	0.884	0.047
PRL	0.000	0.000	0.111	0.000	0.889

The LGCP does not consider any covariates which, per se, should make it harder to do predictions. Regarding the T_1 data, it performs well on the largest subtype (RLRM) but has difficulty with the much smaller groups (CIS, PRP) which in part can be attributed to the small number of data points available for these subtypes, i.e. there are only eight CIS patients with T_1 lesions. However, in case of T_2 data, predictions are much more accurate, resulting in overall and average accuracies of 89.5% and 85.1% respectively. Among the three models considered here, the LGCP is also the closest to a generative model for lesion data, i.e. when simulating new data, it would give much more realistic predictions than the BSGLMM for instance, which assumes independent lesion data conditional on (spatially regularized) coefficients.

4 Discussion and Conclusion

In contrast to standard methods, the Bayesian spatial models presented here exploit the spatial structure of MS lesion maps and take into account the binary nature of lesion data without an arbitrary smoothing parameter. The BSGLMM explicitly includes covariates and spatially varying coefficients and is able to provide spatial information, e.g. estimates for the spatially varying effects of age, sex, disease duration, EDSS and PASAT, which current empirical methods cannot. We have demonstrated that MRI scans of MS lesions contain more information about the specific subtype of the disease than is currently utilized in clinical assessments. To our knowledge this is the first attempt to build spatially informed models. A future extension of the LGCP is a marked point process which allows for the inclusion of lesion specific features as additional *marks*.

References

1. Allwein, E., Schapire, R., Singer, Y.: Reducing multiclass to binary: a unifying approach for margin classifiers. *J. Mach. Learn. Res.* 1, 113–141 (2001)
2. Arns, C., Knackstedt, M., Pinczewski, W., Mecke, K.: Euler-Poincaré characteristics of classes of disordered media. *Phys. Rev. E* 63, 31112 (2001)
3. Cohen, J., Rae-Grant, A.: *Handbook of multiple sclerosis*. Springer Healthcare LCC, London (2010)
4. Ge, T., Müller-Lenke, N., Bendfeldt, K., Nichols, T., Johnson, T.: Analysis of multiple sclerosis lesions via spatially varying coefficients. *Ann. Appl. Stat.* (in press, 2014)
5. Gelfand, A., Dey, D., Chang, H.: Model determination using predictive distributions with implementation via sampling-based methods. *Bayes. Stat.* 4, 147–167 (1992)
6. Kappos, L., Antel, J., Comi, G., Montalban, X., O'Connor, P., Polman, C., Haas, T., Korn, A., Karlsson, G., Radü, E.: Oral fingolimod (FTY720) for relapsing multiple sclerosis. *N. Engl. J. Med.* 355, 1124–1140 (2006)
7. Lang, C., Ohser, J., Hilfer, R.: On the analysis of spatial binary images. *J. Microsc.* 203, 303–313 (2001)
8. Lövblad, K., Anzalone, N., Dörfler, A., Essig, M., Hurwitz, B., Kappos, L., Lee, S.-K., Filippi, M.: MR imaging in multiple sclerosis: Review and recommendations for current practice. *AJNR* 31, 983–989 (2010)
9. MacKay-Altman, R., Petkau, J., Vrecko, D., Smith, A.: A longitudinal model for magnetic resonance imaging lesion count data in multiple sclerosis patients. *Stat. Med.* 31, 449–469 (2011)
10. Møller, J., Syversveen, A., Waagepetersen, R.: Log Gaussian Cox processes. *Scand. J. Stat.* 25, 451–482 (1998)
11. Møller, J., Waagepetersen, R.: *Statistical inference and simulation for spatial point processes*. Chapman & Hall/CRC (2004)
12. Morgan, C., Aban, I., Katholi, C., Cutter, G.: Modeling lesion counts in multiple sclerosis when patients have been selected for baseline activity. *Mult. Sci.* 16, 926–934 (2010)
13. Rasmussen, C.E., Williams, C.K.I.: *Gaussian Processes for Machine Learning*. MIT Press (2006)