Significant Anatomy Detection Through Sparse Classification: A Comparative Study

Li Zhang, Dana Cobzas, Alan H. Wilman, and Linglong Kong

***Abstract*— We present a comparative study for discriminative anatomy detection in high dimensional neuroimaging data. While most studies solve this problem using mass univariate approaches, recent works show better accuracy and variable selection using a sparse classification model. Two types of image-based regularization methods have been proposed in the literature based on either a Graph Net (GN) model or a total variation (TV) model. These studies showed increased classification accuracy and interpretability of results when using image-based regularization, but did not look at the accuracy and quality of the recovered significant regions. In this paper, we theoretically prove bounds on the recovered sparse coefficients and the corresponding selected image regions in four models (two based on GN penalty and two based on TV penalty). Practically, we confirm the theoretical findings by measuring the accuracy of selected regions compared with ground truth on simulated data. We also evaluate the stability of recovered regions over cross-validation folds using real MRI data. Our findings show that the TV penalty is superior to the GN model. In addition, we showed that adding an l_2 penalty improves the accuracy of estimated coefficients and selected significant regions for the both types of models.**

***Index Terms*— Sparse classification, logistic regression, voxel based analysis, localized statistics, MRI, l_1 optimization.**

I. Introduction

With the growth of available medical imaging data, the need for powerful tools to perform large neuroimaging studies has increased. One important problem that has been looked at is the detection of discriminant anatomy between two populations, typically consisting of normal and diseased subjects. The majority of studies use voxel-based analysis (VBA) to identify regions where two groups differ [1]. VBA generates statistical maps consisting of p-values characterizing significant differences at the voxel level. These methods have limited ability to identify complex population differences and pathologies that span multiple anatomical regions because they do not take into account correlations between voxels and regions in the brain. In addition a large number of multiple comparisons are needed due to high dimensionality of the data.

Dimensionality reduction methods have been proposed in the literature to overcome these limitations of VBA [2]. High-dimensional pattern classification methods perform feature extraction and selection to achieve good classification accuracy [3]. In this setting, the discriminant anatomy is detected by the feature selection mechanism. One challenge of these methods is the lack of sufficient training samples relative to the high dimensionality of the data. Recently, it has been shown that imposing sparsity in feature extraction [4], [5] leads to a better feature representation. Feature selection and classification or regression can be simultaneously addressed using sparse regularized methods, that were successfully applied to medical data [6]–[8]. However, imposing sparsity can often lead to less stable feature maps that cannot be interpreted from an anatomical viewpoint. For images with high spatial correlation, adjacent voxels contain similar information and only one would be selected for a good prediction or classification. To counter this behavior, several estimators incorporate the notion of spatial smoothness on the coefficient maps through additional penalizers. Two main types of image-based penalizers have been used in the literature. GN formulations use an l_2 penalty on the gradients to force adjacent voxels to have similar weights [9]–[12]. Alternatively regularization could be enforced by imposing sparsity on the spatial gradients through a TV penalty [13]–[18].

Nevertheless, none of those methods looked at the quality and accuracy of the recovered sparse significant regions in the context of classification. Two previous studies report experimental accuracy of recovered coefficients for TV regularized sparse prediction [16], [17]. One study [19] compared the stability of feature selection for a set of sparse classification methods. Note that, as this study is done only on real data, no accuracy for variable selection could be reported. In addition, theoretical results on the coefficient bounds have not yet been shown. This is practically very important, as results on detected regions are often interpreted from a medical viewpoint where the location of the detected image regions is crucial. As most studies are done on real data, the methods are often evaluated only by the fit of the loss function (mean square error for regression and accuracy for classification).

Manuscript received May 26, 2017; revised July 20, 2017; accepted July 26, 2017. Date of publication August 2, 2017; date of current version December 29, 2017. This work was supported in part by the Canadian Institutes of Health Research, in part by the Natural Sciences and Engineering Research Council, in part by the Multiple Sclerosis Society of Canada and in part by The Canadian Statistical Sciences Institute. *(Corresponding author: Dana Cobzas.)*

L. Zhang and L. Kong are with the Mathematical and Statistical Sciences Department, University of Alberta, Edmonton, AB T6G 2G1, Canada.

D. Cobzas and A. H. Wilman are with the Biomedical Engineering Department, University of Alberta, Edmonton, AB T6G 2G1, Canada (e-mail: cobzas@ualberta.ca)

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMI.2017.2735239

With the exception of few mentioned studies, the evaluation of the sparse coefficients is typically based on heuristics - a good result would be considered the one with larger, more compact regions [10], [13], [14].

We fill this gap with the current study on sparse regularized classification, where we theoretically and experimentally look at the accuracy and stability of detected sparse regions in four common sparse logistic regression formulations (two based on TV norm and two on GN). We also compare results with a VBA method. We restrict this study to only penalized logistic regression formulations to be able to prove both theoretical bounds and verify our findings through synthetic and real data experiments. Three main contributions are brought:

- we derive theoretical bounds of the recovered sparse coefficients and the corresponding selected image regions;
- using synthetic data we measure accuracy of detected sparse regions based on the overlap with ground truth;
- using structural MRI data from a multiple sclerosis study, we measure the stability of detected sparse coefficient maps over cross-validation folds.

The rest of this manuscript is organized as follows. In Section II, after we introduce the theoretical aspects related to the image-based penalized sparse classification formulation, we prove bounds for the regression coefficients and the selected regions. We also summarize the numerical solution for the resulting penalized regression problem. Details on the experiments with synthetic and MRI data are presented in Section III. Finally, we report results in Section IV and some discussions in Section V.

II. Methods

A. Sparse Classification

Let X be a $n \times m$ data matrix of n vectorized images $\mathbf{x}_i$ as rows, each with m voxels. Let $\Omega \subseteq \mathbb{R}^3$ be the image domain of $\mathbf{x}_i$. In the context of binary classification, we are given a corresponding set of labels $\mathbf{y}$ as a $n \times 1$ vector where each y_i takes discrete values $\{-1, +1\}$. The goal is to build a classifier that predicts the binary labels given the data. The most common classification method is logistic regression (LR), where we assume that the logarithm of the odds of y_i being $+1$ (as opposed to -1) is a linear function $\mathbf{x}_i\boldsymbol{\beta} + b$ with image coefficients $\boldsymbol{\beta} \in \mathbb{R}^m$ and scalar b. This implies that y_i follows a logistic distribution with location $\mathbf{x}_i\boldsymbol{\beta} + b$ and scale 1:

$$p(y_i|\mathbf{x}_i, \boldsymbol{\beta}, b) = \frac{1}{1 + \exp(-y_i(\mathbf{x}_i\boldsymbol{\beta} + b))} \quad (1)$$

Maximum likelihood estimation in this model is usually carried out by minimizing the negative log-likelihood:

$$\min_{\boldsymbol{\beta},b} \sum_{i=1}^{n} \log\left(1 + \exp(-y_i(\mathbf{x}_i\boldsymbol{\beta} + b))\right) \quad (2)$$

Unlike a regular linear regression formulated in a least square sense, there will in general be no closed-form solution of the logistic regression model. However, accurate numerical solutions could be obtained by minimizing the negative log-likelihood. Nevertheless, there are two main problems with this solution for our imaging data study: (1) the maximum likelihood solution tends to have all coefficients in $\boldsymbol{\beta}$ nonzero, and (2) this solution might over fit, especially when $m \gg n$.

Sparse constraints on the solution address the first issue. However, selecting the best subset of coefficients (l_0 norm) is an NP-hard problem [20], so an l_1 (LASSO) approximation of the l_0 penalizer is used [21]. Under suitable conditions l_1-regularization will choose the correct subset of nonzero variables [22]. The most common method to address the second issue is l_2-regularization (known as Tikhonov regularization or ridge penalty). The effect of this prior is to decrease the variance of the estimator, by shrinking the coefficients towards zero. An approach that simultaneously achieves subset selection and regularization combined the two l_1 and l_2 penalties [8]:

$$\min_{\boldsymbol{\beta},b} \sum_{i=1}^{n} \log\left(1+\exp(-y_i(\mathbf{x}_i\boldsymbol{\beta}+b))\right)+\lambda_1\|\boldsymbol{\beta}\|_1+\lambda_2\|\boldsymbol{\beta}\|_2^2 \quad (3)$$

where λ_1 and λ_2 are model parameters that control the strength of the regularization.

B. Image-Based Penalty

Sparsity is an effective way of regularizing the classification problem, but may select isolated voxels in the brain rather than compact and anatomically meaningful regions. One way to approach this problem is to cluster neighboring voxels [3]. Alternatively, image-based penalties provide a principled way of imposing anatomical continuity of selected regions. Adding such penalties, say, $\mathcal{P}(\boldsymbol{\beta})$, to Eq. 3 leads to models of the following form:

$$\min_{\boldsymbol{\beta},b} \sum_{i=1}^{n} \log\left(1 + \exp(-y_i(\mathbf{x}_i\boldsymbol{\beta} + b))\right) + \lambda_1\|\boldsymbol{\beta}\|_1 + \lambda_2\|\boldsymbol{\beta}\|_2^2 + \lambda_3\mathcal{P}(\boldsymbol{\beta}) \quad (4)$$

Two types of image-based penalizers have been explored in the context of sparse classification or regression. The GN penalty [9]–[11] uses an l_2 norm on the gradients to force adjacent voxels to have similar coefficients: $\mathcal{P}_{GN}(\beta) = \|\nabla\boldsymbol{\beta}\|_2^2$, where $\boldsymbol{\beta}$ are now the coefficients reshaped as an image. As the derivative of the GN penalty is the image Laplacian $\Delta\boldsymbol{\beta}$, it's effect of the coefficients is linked to uniform diffusion.

A second type of image penalizer, the TV penalty, uses an l_1 norm on the image gradients. One interpretation of this term is by linking it to nonlinear diffusion, thus encouraging smoothing while preserving discontinuities in the coefficient map. Alternatively, the TV-norm could also be seen as imposing sparsity on image gradients. Several works explore the use of TV penalty in penalized sparse methods applied to neuorimaging data [13]–[18]. We use an anisotropic formulation of the TV-norm, also referred as the *fused lasso* [12], [23]: $\mathcal{P}_{TV}(\boldsymbol{\beta}) = \|\nabla\boldsymbol{\beta}\|_1 = \|\nabla_i\boldsymbol{\beta}\|_1 + \|\nabla_j\boldsymbol{\beta}\|_1 + \|\nabla_k\boldsymbol{\beta}\|_1$, where (i, j, k) denotes the 3 orthogonal dimensions of the image data.

Throughout the paper we compare four LR regularized models: (M1) $l_1 + GN$; (M2) $l_1 + l_2 + GN$; (M3) $l_1 + TV$; (M4) $l_1 + l_2 + TV$.

C. Theoretical Bounds on the Estimated Coefficients

The literature on various theoretical aspects of l_1 empirical risk minimization has grown over the last decades. Many theoretical properties of l_1 and $l_1 + l_2$ penalized estimates are also well established [24], [25] and several studies looked at regularization by structured sparsity-inducing norm, as extension of the usual l_1 norm [26], [27]. Nevertheless, to the best of our knowledge, there are yet no results for the theoretical properties of the GN and TV sparse regularized penalties. The goal of this section is to study the theoretical properties of the four penalization schemes (M1 to M4) to provide a comprehensive understanding of their merits and possible limitations by seeking answers to the following central question: given the number of variables m and sample size n and assuming certain regularization conditions, how close are the estimated coefficients to the underlying true coefficients with appropriately tuned penalty parameters?

In the following, we present the error bounds on the estimated coefficients in the four models, M1-M4. Let $\tilde{\boldsymbol{\beta}}$ denote the estimated coefficients for all penalties and $\boldsymbol{\beta}^*$ denote the true coefficients, respectively. In the same way, let $\tilde{I}$ and I^* denote the index set of nonzero coefficients in $\tilde{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}^*$, respectively, that is $\tilde{I} = \{j : \tilde{\beta}_j \neq 0\}$ with $\tilde{\beta}_j$ standing for the j-th component of $\tilde{\boldsymbol{\beta}}$ and I^* is defined similarly. Let k^* be the cardinality of the set I^*. For simplicity, we assume that the observations in the data matrix $\mathbf{X}$ are centred and normalized.

We study the estimates of $\boldsymbol{\beta}^*$ corresponding to the two types of image penalties $\mathcal{P}(\boldsymbol{\beta}) = \mathcal{P}_{GN}(\boldsymbol{\beta})$ and $\mathcal{P}(\boldsymbol{\beta}) = \mathcal{P}_{TV}(\boldsymbol{\beta})$ in (4) under the following two assumptions, which are needed to facilitate the technical details, although they may not be the weakest conditions.

Assumption A: There exist a constant $D > 0$ such that $||\boldsymbol{\beta}^*||_1 \leq D$ and another constant $L > 0$ such that $|x_{ij}| < L$ for any element $|x_{ij}|$ in the data matrix $\mathbf{X}$.

Assumption B: Given $\alpha > 0$ and $\epsilon > 0$, there exists a constant $0 < \kappa \leq 1$ such that

$$\mathbb{P}\left(\mathbf{v}^T \mathrm{S}\mathbf{v} \geq \kappa \sum_{j \in I^*} v_j^2 - \epsilon\right) = 1 \tag{5}$$

for any

$$\mathbf{v} \in V_{\alpha,\epsilon} = \left\{\mathbf{v} \in \mathbb{R}^m : \sum_{j \notin I^*} |v_j| \leq \alpha \sum_{j \in I^*} |v_j| + \epsilon\right\} \tag{6}$$

where S is the sample covariance matrix of the data matrix $\mathbf{X}$.

Assumption C: The true coefficients are upper bounded, that is, there exists a constant B_u such that $\max_{j \in I^*} |\beta_j^*| \leq B_u$.

Assumption A implies that $|\mathbf{x}_i \boldsymbol{\beta}| < LD \leq \infty$ for each row vector $\mathbf{x}_i$ in the data matrix $\mathbf{X}$. As a consequence, the probability $p(y|\mathbf{x}_i, \boldsymbol{\beta}, b)$ is bounded away from zero and one for all $i = 1, \ldots n$, by the equation (1).

Assumption B is essentially the *Condition Stabil.* used in [24]. To better understand it, let $\mathrm{H} = (\delta_{ij}^*)_{i,j=1,\ldots,m}$ where $\delta_{ij}^* = 1$ if $i = j \in I^*$ and 0 otherwise. If $\epsilon = 0$, Assumption B is immediately implied by $\mathbb{P}(\mathrm{S} - \kappa \mathrm{H} \geq 0) = 1$, which guarantees the semi-positiveness of the sample covariance matrix S if the diagonal elements corresponding to the true nonzero coefficients slightly decrease with all the others unchanged. In a sense, Assumption B is a stability requirement on the correlation structure as the modification affects only k^* of the m^2 entries in S. Note that Assumption B is weaker than $\mathbb{P}(\mathrm{S} - \kappa \mathrm{H} \geq 0) = 1$. In addition, the term ϵ in $V_{\alpha,\epsilon}$ is not essential and is needed purely for technical reasons.

Assumption C is a typical requirement to quantify the error bound of the estimated coefficients; see [6], [24], and references therein.

In the following theorem, we show that the difference between the estimated coefficients and the true ones can actually be bounded with high probability with appropriately tuned penalty parameters for both GN and TV penalties.

Theorem 1: Under Assumptions A, B and C with α and ϵ specified in the appendix, if the tuning parameter λ_1 satisfies

$$\lambda_1 \geq (16L + 1)\sqrt{2 \log 2(m \vee n)/n} \tag{7}$$

then

$$\mathbb{P}\left(||\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*||_1 \leq C\lambda_1 + (1 + 2/\lambda_1)\epsilon\right) \geq 1 - \frac{1}{n}, \tag{8}$$

where C is a constant depending on the chosen penalty function. In particular, if denote $\rho_{21} = \lambda_2/\lambda_1$ and $\rho_{31} = \lambda_3/\lambda_1$ in (4)*, then for the GN penalty*

$$C := C_{GN} = \frac{2\left(1 + \rho_{21}B_u + 6\rho_{31}(C_1 D + 2B_u)\right)^2 k^*}{s\kappa + \lambda_2} \tag{9}$$

and, for the TV penalty in (4)

$$C := C_{TV} = \frac{2\left(1 + \rho_{21}B_u + 6\rho_{31}\right)^2 k^*}{s\kappa + \lambda_2}, \tag{10}$$

where $C_1 = 5 + 2\rho_{21}B_u + 12\rho_{31}B_u$, s is a constant depending on both L and D while decreases in terms of D.

The proof of Theorem 1 is deferred to the appendix. There are at least three important conclusions we can draw from Theorem 1: (i) the error of the estimated coefficients is bounded with high probability and the error bound tends to zero as the sample size tends to infinity; (ii) the l_2 penalty in (4) plays a significant role in stabilizing the estimated coefficients by tightening the upper bound of the error; and (iii) the TV penalty is superior to the GN penalty in a sense that it provides a tighter upper error bound. We illustrates these in more details in the following three remarks.

Remark 2: In the upper bound, the term $(1 + 2/\lambda_1)\epsilon$ is negligible when $n \to \infty$ as it is of the order $n/2^{(m \vee n)}$. Therefore, $||\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*||_1 = O\left(\sqrt{\log n/n}\right)$ with probability greater than $1 - 1/n$. This implies that the estimator $\tilde{\boldsymbol{\beta}}$ converges in probability towards the true coefficient $\boldsymbol{\beta}^*$ under the l_1 norm.

Remark 3: To see the role that the l_2 penalty plays in the upper bound, we note that the constant C can be very large if $s\kappa$ is small when l_2 penalty is removed ($\lambda_2 = 0$). Therefore, by adding the l_2 penalty in (4)*, the estimation error can be effectively bounded and stable estimators will be obtained. However, care should be taken in choosing the parameter λ_2, as too large λ_2 may destroy the desired sparsity of the estimators; see [8].*

Remark 4: Although the GN penalty and TV penalty have the same convergence rates, say, $O\left(\sqrt{\log n/n}\right)$, the upper

error bound of the TV penalty can be smaller than that of the GN penalty, that is, the TV penalty is superior to the GN penalty. This can be easily verified by noticing that $C_1 D + 2B_u > 1$ *when the upper bound of the* l_1 *norm of the true coefficients is chosen to be greater than* 0.2.

D. Accuracy of Nonzero Coefficient Selection

In the last section, we showed that the error of the coefficient estimators can be upper bounded with high probability. In this session, we will study if the estimators have the desired sparsity thanks to the l_1 penalty in (4), that is, if the estimated nonzero coefficients are indeed the true nonzero coefficients. In particular, to guarantee the accuracy of nonzero coefficient selection we will investigate the probability of $I^* \subset \tilde{I}$ under the following condition on the smallest nonzero coefficient; see [6], [8], [24], and many others. In the literature, it is also called the weakest signal condition; see [24] and [26].

Assumption D: The true nonzero coefficients are lower bounded, that is, there exists a constant B_l such that $\min_{j\in I^*} |\beta_j^*| \geq B_l$.

In what follows, it is showed that $I^* \subset \tilde{I}$ can be achieved with high probability if the upper error bound in Theorem 1 is smaller than the smallest nonzero coefficient B_l.

Theorem 5: Under Assumptions A, B, C and D with α *and* ϵ *specified in the appendix and* λ_1 *and* C *specified in Theorem 1, if*

$$B_l \geq C\lambda_1 + (1 + 2/\lambda_1)\epsilon, \tag{11}$$

then

$$\mathbb{P}\left(I^* \subset \tilde{I}\right) \geq 1 - \frac{1}{n}. \tag{12}$$

The proof of Theorem 5 is straightforward, hence is omitted here, by applying the fact that

$$\mathbb{P}\left(I^* \not\subset \tilde{I}\right) \leq \mathbb{P}\left(||\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*||_1 \geq B_l\right), \tag{13}$$

which can be easily derived by following [24]. Theorem 5 essentially states that to guarantee desired sparsity the smallest true nonzero coefficient can not be too small. Otherwise, it can not be distinguished from noises and the accuracy of nonzero coefficient selection would not be granted. In addition, we make the following two remarks implied by Theorem 5.

Remark 6: The upper error bound in Theorems 1 and 5 is dominated by $C\lambda_1$ *as discussed in remark 2. For stable data matrix* $\mathbf{X}$*, the constant* C *can be very close to* k^* *and thus* $C\lambda_1 + (1 + 2/\lambda_1)\epsilon \approx C\lambda_1 \approx k^*\lambda_1$*. As a consequence, even weak signals can be detected when the true coefficients are sparse enough to make* $k^*\lambda_1 < B_l$ *hold.*

Remark 7: The TV penalty is capable of detecting weaker signals than the GN penalty and thus provides more robust estimations. This can be seen by recalling that the constant C_{TV} *is always less than* C_{GN} *as discussed in remark 4. Therefore, the TV penalty can pick up small coefficients between* $C_{TV}\lambda_1 + (1 + 2/\lambda_1)\,\epsilon$ *and* $C_{GN}\lambda_1 + (1 + 2/\lambda_1)\epsilon$ *with high probability while the GN penalty can not.*

E. Optimization

Solving the optimization problem (4) is complicated due to the non-differentiability of the l_1 and TV penalties. Various numerical methods have been proposed and evaluation studies of l_1-TV regularized regression and logistic regression are implemented in [15]. In this manuscript, we opt to choose the alternating direction method of multipliers (ADMM) [28] which is efficient to solve the convex optimization in (4). The efficiency is coming from the fact that optimizing (4) can be split into two sub-convex optimization problems: one has an explicit closed-form solution and the other one is a smooth function optimization, which can be efficiently solved by many gradient methods.

The convex optimization problem (4) can be rewritten in the following form:

$$\begin{aligned} &\min\ L(\boldsymbol{\beta},\ b) + ||\Lambda^T\boldsymbol{\alpha}||_1 \\ &s.t\ \boldsymbol{\alpha} = \boldsymbol{A}\boldsymbol{\beta} \end{aligned} \tag{14}$$

where the $L(\boldsymbol{\beta}, b)$ is a smooth function of $\boldsymbol{\beta}$, and the matrix $\boldsymbol{A}$ and the vector Λ depend on the penalty types. In particular, for the GN penalty,

$$\begin{aligned} L(\boldsymbol{\beta},\ b) &= \sum_{i=1}^{n} \log\left(1 + \exp(-y_i(\mathbf{x}_i\boldsymbol{\beta} + b))\right) \\ &\quad + \lambda_2\|\boldsymbol{\beta}\|_2^2 + \lambda_3\|\nabla\boldsymbol{\beta}\|_2^2 \\ \boldsymbol{A} &= \mathbb{I}_{m\times m} \\ \Lambda &= \lambda_1\mathbf{1}_m^T, \end{aligned} \tag{15}$$

and for the TV penalty,

$$\begin{aligned} L(\boldsymbol{\beta},\ b) &= \sum_{i=1}^{n} \log\left(1 + \exp(-y_i(\mathbf{x}_i\boldsymbol{\beta} + b))\right) + \lambda_2\|\boldsymbol{\beta}\|_2^2 \\ A &= [\mathbb{I}_{m\times m},\ \boldsymbol{D}^T]^T \\ \Lambda &= [\lambda_1\mathbf{1}_m,\ \lambda_3\mathbf{1}_{3m}]^T, \end{aligned} \tag{16}$$

where $\boldsymbol{D}$ is the 3D different operator matrix; see [29].

The ADMM method alternatively estimates the model coefficients $\boldsymbol{\beta}, b$ and the auxiliary variables $\boldsymbol{\alpha}, \boldsymbol{\eta}$ through the following iterations

$$(\boldsymbol{\beta}, b)^{(t+1)} = \arg\min_{\boldsymbol{\beta},b} L_\rho(\boldsymbol{\beta}, b, \boldsymbol{\alpha}^{(t)}, \boldsymbol{\eta}^{(t)}) \tag{17}$$

$$\boldsymbol{\alpha}^{(t+1)} = \arg\min_{\boldsymbol{\alpha}} L_\rho(\boldsymbol{\beta}^{(t+1)}, b^{(t+1)}, \boldsymbol{\alpha}, \boldsymbol{\eta}^{(t)}) \tag{18}$$

$$\boldsymbol{\eta}^{(t+1)} = \boldsymbol{\eta}^{(t)} + \rho(\boldsymbol{\alpha}^{(t+1)} - \boldsymbol{A}\boldsymbol{\beta}^{(t+1)}). \tag{19}$$

For both the GN and TV penalties, the subproblem (17) involves minimizing a smooth function and hence can be efficiently solved by many gradient methods, such as the limited-memory Broyden-Fletcher-Goldfarb-Shanno algorithm (L-BFGS) [30]. The subproblem (18) involves the minimization of the sum of an l_2 function and an l_1 type penalty and can be solved by the soft thresholding operator, giving the following closed form,

$$\alpha_i^{(t+1)} = S_{\Lambda_i}\left((\boldsymbol{A}\boldsymbol{\beta}^{(t+1)})_i - \eta_i^{(t)}/\rho\right) \tag{20}$$

with $S_\lambda(x) := \mathrm{sgn}(\mathrm{x})(|\mathrm{x}| - \lambda)_+$ a soft thresholding operator.

We use the termination criterion suggested by [28], which is based on primal residuals r_{pri} and dual residuals r_{dual}. At the tth iteration, they are defined as $r_{pri}^{(t)} = \boldsymbol{\alpha}^{(t)} - \boldsymbol{A}\boldsymbol{\beta}^{(t)}$ and $r_{dual}^{(t)} = \rho \boldsymbol{A}^T(\alpha^{(t)} - \alpha^{(t-1)})$. The termination criterion is

$$||r_{pri}^{(t)}||_2 \leq \epsilon^{pri} \quad and \quad ||r_{dual}^{(t)}||_2 \leq \epsilon^{dual} \tag{21}$$

where $\epsilon^{pri} > 0$ and $\epsilon^{dual} > 0$ in the Algorithm 1 are feasibility tolerances for the primal and dual feasibility conditions.

Algorithm 1 ADMM Algorithm

Input: Training sample $(X_1, Y_1), \ldots, (X_n, Y_n)$, tuning parameters $\lambda_1, \lambda_2, \lambda_3$, stopping parameters ϵ^{pri} and ϵ^{dual}

initialization set: $\boldsymbol{\alpha}^{(0)}$ and $\boldsymbol{\eta}^0$ are zero vector; initialize primal and dual residuals $r_{pri}^{(0)}$ and $r_{dual}^{(0)}$;

while $||r_{pri}^{(t)}||_2 > \epsilon^{pri}$ *or* $||r_{dual}^{(t)}||_2 > \epsilon^{dual}$ **do**

Update $\boldsymbol{\beta}^{(t+1)}, b^{(t+1)}$ using (17) by the L-BFGS

Update $\boldsymbol{\alpha}^{(t+1)}$ using (18) by the soft thresholding

Update $\boldsymbol{\eta}^{(t+1)}$ using (19) Update $r_{pri}^{(t+1)}$ and $r_{dual}^{(t+1)}$

end

Output: $\boldsymbol{\beta}$, b, and $\boldsymbol{\alpha}$

III. Experiments

A. Synthetic Data

Medical imaging data has no available ground truth on the significant anatomical regions discriminating two populations. We therefore generated synthetic data $\mathbf{x}_i$ of size $32 \times 32 \times 8$ containing four $8 \times 8 \times 4$ foreground blocks with high spatial coherence (see Figure 2). Background values are generated from a normal distribution $N(0, 1)$, while the correlated values inside the four blocks are drawn from a multinormal distribution $N(0, \Sigma_\rho)$. We tested three levels of coherence $\rho \in \{0.1, 0.25, 0.5\}$. To generate class labels y_i for the synthetic data, we followed the logistic regression distribution from Equation (1). The coefficient vector $\boldsymbol{\beta}$ has fixed values of 0 outside the four blocks and $0.1, 0.2, 0.3, 0.4$ inside, giving different strength for the data signal in each of the four regions. b is set to 0.1. Binary labels are then assigned based on the logistic probability $p(y_i|\mathbf{x}_i, \boldsymbol{\beta}, b)$ following a Bernoulli distribution. Each dataset contains n subjects (n is between 100-300 subjects), making the data matrix X of size $n \times 8192$. For each coherence values ρ we repeated the test 96 times.

B. Neuroimaging Data

Our neuroimaging data belongs to an in-house multiple sclerosis (MS) study [31]. Following recent research that suggests a possible pivotal role for iron in MS [32], we are investigating if iron in deep gray matter is a potential biomarker of disability in MS. High field (4.7T) quantitative traverse relaxation rate (R2*) images are used as they are shown to be highly influenced by non-heme iron [33]. Sample R2* slices can be viewed in Figure 4 left. The focus is subcortical deep gray matter (DGM) structures: caudate, putamen, thalamus and global pallidus, that have a high iron content.

C. Subjects

Institutional ethical approval and informed consent were obtained from each subject prior to the study. Forty subjects with relapsing remitting MS (RRMS) and 40 age- and gender-matched controls were recruited. The mean ages for patients and controls are 38.96 ± 10.01 and 38.92 ± 9.73 years old, while the male/female distribution is 7/33 and 10/30, respectively. Gender matching was not considered and our focus was centered on age-matching because aging is the strongest predictor of DGM MRI changes [34]. Baseline characteristics were compared between groups using one-way analysis of variance (ANOVA) for age. There were no significant age ($p = 0.99$) differences between patients and controls.

D. MRI Methods

Imaging was performed at 4.7T (Varian Inova, Palo Alto, CA). Two imaging sequences were acquired in the same session: 3D T1w volumetric imaging using magnetization-prepared rapid gradient-echo (MPRAGE) (108 flip, TE/TR 4.5/8.5 msec, inversion time to start of readout 300 msec, sequential phase encoding, 84 slices, 2 mm thick, in-plane 0.9×0.9 mm^2, $284 \times 222 \times 84$ matrix, acquisition time 4.8 min) and R2* using 3D multiecho gradient echo (108 flip, TE1/TR 2.9/44 msec, 10 echoes, echo spacing 4.1 msec, 80 slices, 2 mm thick, in-plane 1×1 mm^2, $256 \times 160 \times 80$ matrix, acquisition time 9.4 min).

E. Pre-Processing

Prior to analysis, the MRI data is pre-processed and aligned with an in-house unbiased template using ANTs [35]. The multimodal template is built from 10 healthy controls using both T1w and R2* methods [36]. Pre-processing involves intra-subject alignment of R2* with T1w and bias field intensity normalization for T1w [37]. Nonlinear registration in the template space is done using SyN [38] on the multimodal MRI data. Aligned R2* values are used as iron-related measurements. The measurement row vectors $\mathbf{x}_i$ of size 158865 are formed by selecting only voxels inside a deep gray matter mask that is manually traced on the atlas (Figure 4).

F. Evaluation Methodology

We compare the performance of the four penalized logistic regression models described in section II-B (M1) $LR + l_1 + GN$; (M2) $LR + l_1 + l_2 + GN$; (M3) $LR + l_1 + TV$; (M4) $LR + l_1 + l_2 + TV$ and an additional classic VBA model (M5). For the VBA model, variable selection is done using a two sided t-test with a $p < 0.05$ threshold. Selected variables are next used in a LR model with no penalties to get corresponding coefficients $\boldsymbol{\beta}$. The five methods are uniformly compared as follows. Training and test data is selected for each of the 96 synthetic datasets (100 training and 100 test) and for the real data (5 folds cross-validation). Optimal parameters for the regularized sparse models are tuned using the Bayesian information criterion (BIC) on all data. Results are reported on the test data using the $\boldsymbol{\beta}$ coefficients computed on the training data. The sparse regions are selected from all nonzero

TABLE I

COMPARATIVE RESULTS FOR SYNTHETIC DATA WITH DIFFERENT VALUE OF COHERENCE OF THE MULTINORMAL DISTRIBUTION ρ. SAMPLE SIZE IS 100. MEAN(STD) ARE EVALUATED OVER 96 TRIALS. EVALUATION MEASURES ON COLUMNS: CLASS. ACC.= CLASSIFICATION ACCURACY, SENS. = SENSITIVITY, SPEC.= SPECIFICITY, AUC, DICE = DICE SCORE BETWEEN DETECTED SIGNIFICANT REGIONS AND GROUND TRUTH (MEASURES ACCURACY IN REGION DETECTION); $\|\tilde{\beta} - \beta^*\|_1$ SUM OF ABSOLUTE ERRORS (SAE) OF ESTIMATED AND TRUE COEFFICIENTS. BOLD HIGHLIGHTS BEST RESULTS AMONG METHODS FOR EACH EXPERIMENT

Method	Class. acc.	Sens.	Spec.	AUC	Dice	$\|\tilde{\beta} - \beta^*\|_1$
$\rho = 0.1$						
$l_1 + l_2 + TV$	**0.874**(0.033)	**0.873**(0.034)	**0.870**(0.046)	**0.877**(0.073)	**0.769**(0.053)	**154.34**(12.67)
$l_1 + TV$	0.872(0.035)	0.871(0.037)	0.866(0.049)	**0.877**(0.076)	0.761(0.055)	158.77(15.624)
$l_1 + l_2 + GN$	0.833(0.041)	0.833(0.041)	0.817(0.077)	0.848(0.088)	0.640(0.042)	213.66(4.64)
$l_1 + GN$	0.834(0.039)	0.833(0.039)	0.819(0.081)	0.847(0.089)	0.639(0.049)	211.45(5.03)
VBA	0.806(0.026)	0.806(0.026)	0.800(0.056)	0.811(0.051)	0.462(0.048)	272.92(5.66)
$\rho = 0.25$						
$l_1 + l_2 + TV$	**0.885**(0.027)	**0.884**(0.029)	**0.876**(0.052)	**0.892**(0.061)	**0.778**(0.046)	**151.41**(13.125)
$l_1 + TV$	0.882(0.029)	0.881(0.031)	0.875(0.053)	0.887(0.065)	0.768(0.050)	154.59(14.95)
$l_1 + l_2 + GN$	0.856(0.031)	0.856(0.032)	0.851(0.062)	0.860(0.073)	0.669(0.040)	209.11(5.49)
$l1 + GN$	0.855(0.032)	0.854(0.033)	0.850(0.062)	0.858(0.074)	0.671(0.040)	207.00(5.20)
VBA	0.832(0.021)	0.831(0.021)	0.823(0.054)	0.839(0.045)	0.524(0.053)	268.96(6.00)
$\rho = 0.5$						
$l_1 + l_2 + TV$	**0.904**(0.041)	**0.904**(0.042)	**0.899**(0.052)	**0.910**(0.065)	**0.784**(0.062)	**155.08**(12.38)
$l_1 + TV$	0.902(0.042)	0.902(0.043)	0.898(0.055)	0.906(0.063)	0.773(0.065)	159.91(14.36)
$l_1 + l_2 + GN$	0.878(0.045)	0.878(0.045)	0.876(0.080)	0.880(0.089)	0.711(0.039)	213.51(4.360)
$l_1 + GN$	0.878(0.042)	0.878(0.042)	0.877(0.076)	0.879(0.085)	0.708(0.038)	211.09(4.46)
VBA	0.875(0.016)	0.874(0.016)	0.870(0.041)	0.878(0.039)	0.612(0.060)	265.07(6.73)

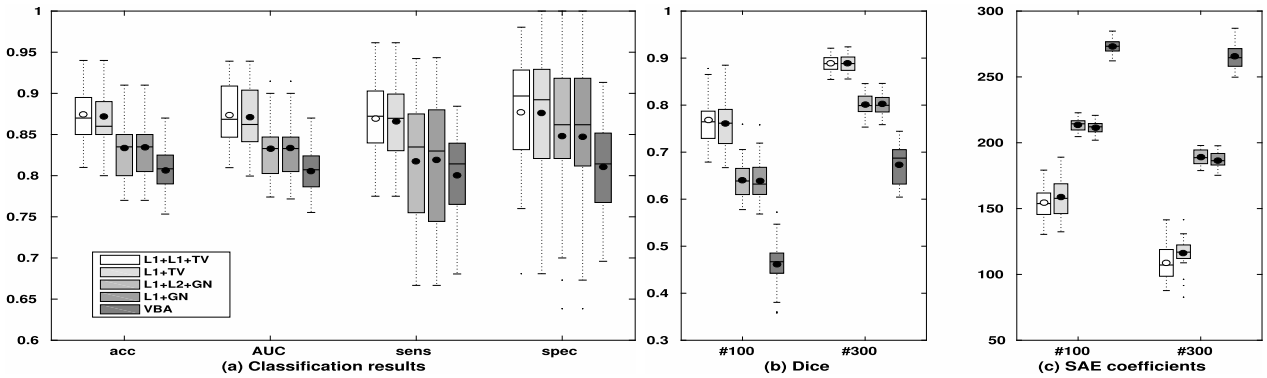

Fig. 1. Results for synthetic experiments. (a) Classification scores for sample size = 100; acc = classification accuracy, AUC, sens = sensitivity, spec = specificity (b) Dice scores between ground truth and estimated nonzero coefficients (c) Sum of Absolute Error (SAE) between ground truth and estimated coefficients for sample size 100 and 300. Results are computed on the test data with optimal parameters and averaged over 96 trials.

coefficients on synthetic data and using a threshold of 5% of the median of non-zero coefficients for the real data.

Classification results are evaluated using accuracy (proportion of correctly classified samples), sensitivity (true positive rate), specificity (true negative rate) and the area-under-the-curve (AUC) for the receiver operating characteristic (ROC) curve. The ROC [39] curve evaluates the performance of a binary classifier as its discrimination threshold is varied.

The accuracy of the detected significant regions when compared to ground truth on synthetic data is evaluated using a dice score between estimated regions and ground truth. We also compute the mean absolute error of recovered vs ground truth coefficients. For real data, as ground truth for variable selection is not available, we measured the stability of the detected regions using a dice score between pairs of estimated nonzero coefficients for each of the 5 folds (10 pairs).

IV. RESULTS

A. Results on Simulated Data

Comparative results on synthetic data with three levels of coherence $\rho \in \{0.1, 0.25, 0.5\}$ for the multinormal distribution are reported in Table I with corresponding bar graphs in Figure 1 (a) and (b). Mean and standard deviation are computed from 96 trials. We used $n = 100$ subjects for training and 100 for test. The average speed for solving Eq 4 on an Intel i7 3.60 GHz machine with 64 GB RAM with the synthetic data, $\rho = 0.25$ and sample size 100, is about 20 sec for the TV case and about 7 sec for the GN case.

When evaluating classification accuracy (accuracy, AUC, sensitivity and specificity), the l_1+TV and l_1+l_2+TV models outperform the other ones, with the l_2-regularized model being slightly superior. The VBA method has the worse performance indicating the need to account for correlations among voxel. All methods perform better with higher coherence.

The last two columns in Table I show results on the accuracy of the detected significant regions using dice score with ground truth as well as the accuracy of the recovered sparse coefficients (last column). The mean dice with ground truth is about 77% for the methods that use TV image penalty compared to 67% for the methods that use GN penalty, and drops to 53% for VBA. The results confirm our theoretical findings. We point out three main links between theory and results. Firstly, as stated in Remark 4 and 7, the TV image-based

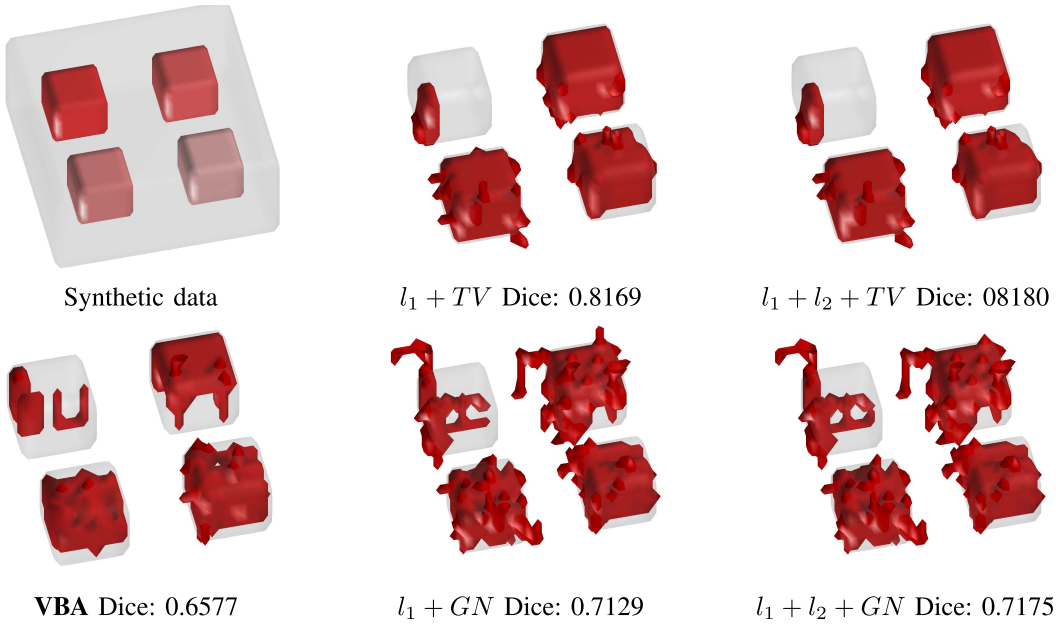

Fig. 2. (top-left) The ground truth coefficients for simulated data. The side of the regression coefficients is 32 × 32 × 8. The nonzero values in the four 8 × 8 × 8 blocks are 0.1 in top-left block, 0.2 in top-right block, 0.3 in bottom-left block and 0.4 in bottom-right block. They determine the strength of the signal. The following figures show significant regions on synthetic data detected by the 5 methods on one of the 96 trials. Shaded gray regions correspond to the true nonzero coefficients and the red regions are the estimated nonzero coefficients. Corresponding dice scores are also shown below each figure.

penalty gives both more accurate coefficient estimation and more accurate variable selection results. Secondly, we notice better results when adding the l_2 penalty to both TV and GN models, in accordance to Remark 3. Finally, to verify the asymptotic behavior of the bounds when $n \to \infty$, we increased the number of subjects in the training set to 300. The size of the test set remains 100 subjects. As shown in the bar plot (c) of Figure 1, the error on the coefficients is smaller when more samples are used to train the model in accordance to Remark 2.

To visualize the synthetic data results we depicted in Figure 2 the recovered nonzero coefficients for one of the 96 trials. The strength of the signal in the four blocks is determined by the regression coefficients $\boldsymbol{\beta}$ that have values of 0.1 in top-left block, 0.2 in top-right block, 0.3 in bottom-left block and 0.4 in bottom-right block. All methods recover the bottom blocks where the coefficients are largest (0.3 and 0.4), the GN-regularized methods recovered part of the 3rd block (coefficients 0.2) and the TV-regularized methods recover most of the 3rd block. The 4th block has the weakest signal and only part of it was only recovered by the TV-regularized methods. It is clear from the figure that the TV regularized methods are superior to the GN ones. While the dice scores show slight improvement when adding the l_2 penalty, there is little visual difference.

B. Results on MRI Data

Comparative classification results on real neuroimaging MRI data are reported in Table II. The receiver operating characteristics (ROC) curves are shown in Figure 3. Average results over 5 crossvalidation folds are reported. Similar to the synthetic data results, the TV-regularized methods achieve slightly better classification accuracy than the GN-regularized methods, and VBA gets lowest scores with an accuracy of only about 0.50. The relatively poor classification accuracy

TABLE II

COMPARATIVE RESULTS FOR REAL MRI DATA. MEANS ON THE 5 FOLDS ARE REPORTED. EVALUATION MEASURES ON COLUMNS: SPARS. = EFFECTIVE SPARSENESS COMPUTED WITH RESPECT TO THE WHOLE VOLUME; SCORES: CLASS. ACC. = CLASSIFICATION ACCURACY, SENS. = SENSITIVITY, SPEC. = SPECIFICITY, AUC; DICE FOLDS = DICE SCORE BETWEEN DETECTED SIGNIFICANT REGIONS DETECTED IN DIFFERENT FOLDS (MEASURES STABILITY IN REGION DETECTION). BOLD HIGHLIGHTS BEST RESULTS AMONG METHODS FOR EACH EXPERIMENT

Method	Class. acc.	Sens.	Spec.	AUC	Dice Folds
$l_1 + l_2 + TV$	0.613	0.613	0.675	0.550	0.586
$l_1 + TV$	0.613	0.613	0.675	0.550	0.556
$l_1 + l_2 + GN$	0.600	0.600	0.700	0.500	0.582
$l_1 + GN$	0.600	0.600	0.700	0.500	0.578
VBA	0.500	0.500	0.600	0.400	0.311

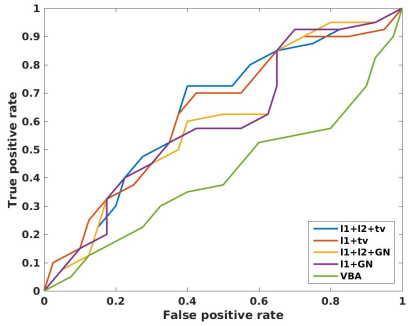

Fig. 3. The receiver operating characteristics (ROC) curves for the 5 compared methods on real MRI data. The average over 5 folds is reported. Area under the curve (AUC) values are listed in Table II on the left side.

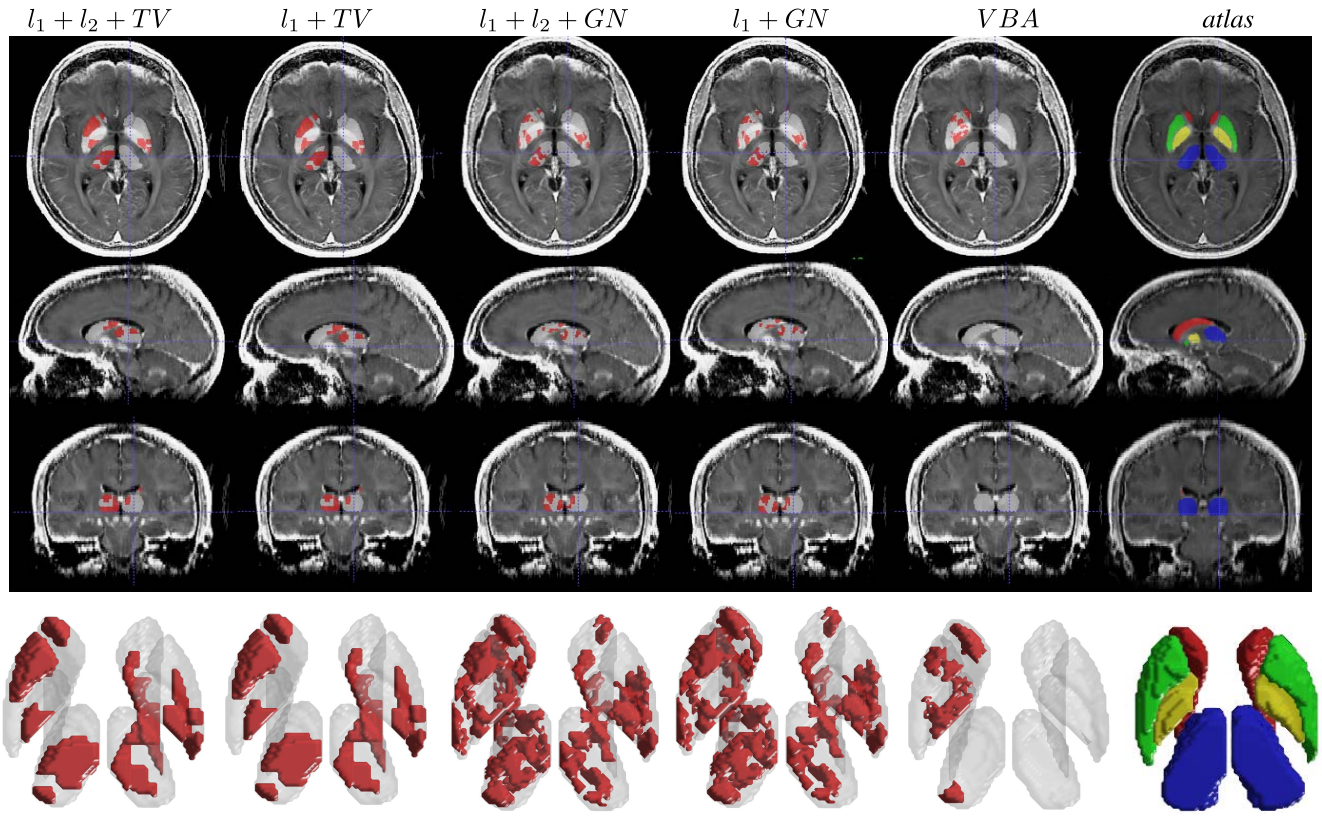

Fig. 4. Illustration of the significant anatomy detected by the 5 methods using R2* MRI data in the first 5 columns. The 3 rows represent orthogonal 2D slices with the R2* data as background and the last row is a 3D view of the result. The deep gray matter mask used for selecting the voxels included in the observation vectors $\mathbf{x}_i$ are displayed in transparent gray, while the thresholded significant regions in red. Last column shows the deep gray matter atlas with the four structures of interest as (red) caudate, (green) putamen, (yellow) global pallidus, (blue) thalamus.

on real data is due to the small data sample. Nevertheless, the results on the stability of selected sparse regions enforce our theoretical results. As ground truth is not available for real data, we measured the quality of the detected sparse regions using between-folds dice scores (DiceFolds). We report the average over the 10 distinct folds combinations. Again, results indicate that the TV-regularized methods achieve best accuracy and region stability, followed by the GN-regularized methods, while VBA has the worst performance and a DiceFolds of 31%. The effect of adding the l_2 penalty is less visible for the accuracy results, but the dice between folds is better for the l_2-regularized methods showing improved stability.

For a qualitative comparison, to visualize the results, Figure 4 displays sample orthogonal slices and a 3D view of the regions recovered by the five methods. The regions were calculated from all data with optimal parameters for each method. We notice that the regions recovered by the TV-regularized methods are more compact than the ones recovered by the GN-regularized methods. VBA failed to recovered anything in the right side of the brain. Iron-sensitive regions that differentiate patients and controls are detected in all four DGM structures similar but more conservative than in previous VBA studies [40]. This is probably due to the fact that the current work uses cross-validation for getting the results, while, in most traditional studies, the variable selection and testing is done on all data.

V. Discussion

Detecting brain regions that discriminate two populations using high dimensional imaging data is an important aspect of many neuroimaging studies. While traditionally this problem was solved using VBA, recent studies showed better formulations using sparse regularized classification [6]–[8]. These methods simultaneously address the problem of classification and feature selection to detect optimal discriminative brain regions. To improve interpretability and stability of results, additional image-based penalizers have been used to encourage spatial smoothness on the coefficient maps. Two main types of penalizers have been proposed, a first group uses an l_2 penalty on the gradients (Graph Net) [9]–[12], while a second group uses a l_1 penalty on the gradients (TV norm) [13]–[18]. Even though these methods showed good classification accuracy and produce visually "good" regions, up to our knowledge, there is no study that looked at the accuracy and quality of the recovered sparse regions from both a theoretical and experimental viewpoint. This is a very important aspect of such studies, as in most cases, results are interpreted from a medical viewpoint depending on the location of the detected image regions. In this current study, we theoretically and experimentally evaluate the accuracy and stability of detected sparse regions in four common sparse classification formulations (two based on TV penalty and two based on GN penalty). We also compare results with a VBA method. Theoretically we derive bounds of the recovered sparse coefficients and the corresponding

selected image regions. Practically, through simulated data we measure accuracy of detected regions based on the overlap with ground truth data. In addition, we used MRI data from an in-house multiple sclerosis study, to measure the stability of detected sparse coefficient maps over cross-validation folds.

Four main conclusions on region selection accuracy of sparse logistic regression methods are drawn from our study:

- We theoretically showed that using a TV penalty results in better bounds on the coefficients and better accuracy of selected regions. Synthetic experiments support this finding confirming that the TV regularized methods achieve higher Dice score between detected discriminative regions and ground truth and also tighter coefficient bounds. For the real MRI data, in the absence of ground truth on variable selection and similar to other studies [19], we used region stability over cross-validation folds (Dice-Folds) as an indication for accuracy. Results confirm better region selection stability when using a TV penalty. Also, in accordance to finding from other studies [14], we visually notice that the recovered regions are more compact and larger when using a TV penalty compared to a GN penalty.
- We showed that adding a l_2 penalty improves coefficient bounds for both types of methods (GN and TV). Quantitative results using synthetic and real data confirm this finding, even though the difference is small and almost unnoticeable in the visualization from Figures 2 and 4.
- We confirmed through theory and synthetic experiments that increasing the number of subjects in the study will result in better bounds on the coefficients and improved accuracy of detected regions
- As noted by many other studies, we showed that all sparse penalized methods are superior to traditional VBA methods. This is clearly confirmed by both numeric and visual results on the synthetic and real data.

We believe that our study filled an important gap in the literature of image-based penalized sparse classification methods. Our results could benefit many practical studies, bringing guarantee on the quality and bounds of the recovered discriminative regions and suggesting an optimal model that uses a combination of l_1, l_2 and TV penalties.

Appendix//Short Proof of Theroem II.1

In this section, we prove a more general case than **Theroem II.1**. We first show that, with high probability, $\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \in V_{\alpha,\epsilon}$ for large enough constant α and $\epsilon = \frac{\log 2}{\lambda_1 2^{(m\vee n)}}$. Then we show $||\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*||_1$ bounded by a small number with high probability. To be simple, we do not consider the intercept b.

The GN penalty has $\mathcal{P}_{GN}(\beta) = \|\nabla\boldsymbol{\beta}\|_2^2$. In this case, the empirical risk can be written as

$$\mathbb{P}_n(l(\boldsymbol{\beta})) = \frac{1}{n}\sum_{i=1}^{n}\left[-Y_i\sum_{j=1}^{m}\beta_j X_{ij} + \log(1+\exp\sum_{j=1}^{m}\beta_j X_{ij})\right].$$

Let $\lambda_1 = \lambda_{n,M}(\delta)$ and the estimator shall satisfy

$$\mathbb{P}_n(l(\tilde{\boldsymbol{\beta}})) + \lambda_1||\tilde{\boldsymbol{\beta}}||_1 + \lambda_2||\tilde{\boldsymbol{\beta}}||_2^2 + \lambda_3||\bigtriangledown\tilde{\boldsymbol{\beta}}||_2^2 \le \mathbb{P}_n(l(\boldsymbol{\beta}^*)) + \lambda_1||\boldsymbol{\beta}^*||_1 + \lambda_2||\boldsymbol{\beta}^*||_2^2 + \lambda_3||\bigtriangledown\boldsymbol{\beta}^*||_2^2$$

where δ in the λ_1 is a small number, for example $\delta = \frac{1}{n}$. Let $I_1 = \lambda_2||\tilde{\boldsymbol{\beta}}||_2^2 - \lambda_2||\boldsymbol{\beta}^*||_2^2$ and $I_2 = \lambda_3||\bigtriangledown\tilde{\boldsymbol{\beta}}||_2^2 - \lambda_3||\bigtriangledown\boldsymbol{\beta}^*||_2^2$, we obtain

$$\begin{aligned}\frac{\lambda_1}{2}||\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*||_1 + \mathbb{P}(l(\tilde{\boldsymbol{\beta}}) - l(\boldsymbol{\beta}^*) + I_1 + I_2 \\ \le (\mathbb{P}_n - \mathbb{P})(l(\boldsymbol{\beta}^*) - l(\tilde{\boldsymbol{\beta}})) + \frac{\lambda_1}{2}||\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*||_1 \\ + \lambda_1(||\boldsymbol{\beta}^*||_1 - ||\tilde{\boldsymbol{\beta}}||_1).\end{aligned} \tag{22}$$

Denote

$$L_n = \sup_{\boldsymbol{\beta}\in\mathbb{R}^m}\frac{(\mathbb{P}_n - \mathbb{P})(l(\boldsymbol{\beta}^*) - l(\boldsymbol{\beta}))}{||\boldsymbol{\beta} - \boldsymbol{\beta}^*||_1 + \epsilon}, \quad \text{and } E = \{L_n \le \frac{\lambda_1}{2}\}.$$

Following proof of [24, Th. 2.4], we have that if

$$\lambda_1 \ge 12\ L\sqrt{\frac{2\log 2(M\vee n)}{n}} + \frac{1}{2(M\vee n)} + 4L\sqrt{\frac{2\log\ (\frac{1}{\delta})}{n}}$$

then $P(E) \ge 1-\delta$. By showing $\sum_j |\tilde{\beta}_j - \beta_j^*| < C_1\ D$ on the set E, we have

$$\mathbb{P}l(\tilde{\boldsymbol{\beta}}) - \mathbb{P}l(\boldsymbol{\beta}^*) \ge ||g_{\tilde{\boldsymbol{\beta}}} - g_{\boldsymbol{\beta}^*}||_2^2 \ge s||f_{\tilde{\boldsymbol{\beta}}} - f_{\boldsymbol{\beta}^*}||_2^2$$

where $g_{\boldsymbol{\beta}}(x) = \frac{\exp(\boldsymbol{\beta}'\boldsymbol{x})}{1+\exp(\boldsymbol{\beta}'\boldsymbol{x})}$, $f_{\boldsymbol{\beta}}(\boldsymbol{x}) = \boldsymbol{\beta}'\boldsymbol{x}$, $A = (C_1+1)LD$, and $s = (1+\exp(A))^{-4}$. On the event E, equation (22) further yields

$$\begin{aligned}\frac{1}{2}\lambda_1||\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*||_1 + s||f_{\tilde{\boldsymbol{\beta}}} - f_{\boldsymbol{\beta}^*}||_2^2 - \frac{\lambda_1}{2}\epsilon \\ \le 2\lambda_1\sum_{j\in I^*}(|\beta_j^* - \tilde{\beta}_j|) + \lambda_2(C_1D + 2B_u)\sum_{j\in I^*}(|\beta_j^* - \tilde{\beta}_j|) \\ + 6\lambda_3(C_1D + 2B_u)\sum_{j\in I^*}(|\beta_j^* - \tilde{\beta}_j|).\end{aligned} \tag{23}$$

Due to the fact

$$\begin{aligned}\sum_{j\notin I^*}|\tilde{\beta}_j - \beta_j^*| \le [3 + 2(\rho_{21} \\ + 6\rho_{31}(C_1D + 2B_u)]\sum_{j\in I^*}(|\beta_j^* - \tilde{\beta}_j|) + \epsilon\end{aligned}$$

we have $\tilde{\beta}_j - \beta_j^* \in V_{\alpha,\epsilon}$, which is defined in Assumption B, with $\alpha = 3 + 2(\rho_{21} + 6\rho_{31})(C_1D + 2B_u)$. Let $\gamma_{kj} = \mathbb{E}X_kX_j$, for $k, j \in \{1,\ 2,\ \ldots,\ m\}$ and Γ be the $m \times m$ matrix with entries γ_{kj}. We can obtain

$$\begin{aligned}\frac{1}{2}\lambda_1\sum_j|\tilde{\beta}_j - \beta_j^*| + \kappa s\sum_{j\in I^*}(\tilde{\beta}_j - \beta_j^*)^2 - (\frac{\lambda_1}{2} + 1)\epsilon \\ \le aC_{GN}(\kappa s + \lambda_2)\lambda_1^2/2 + (\frac{1}{a} - \lambda_2)\sum_{j\in I^*}(\tilde{\beta}_j - \beta_j^*)^2.\end{aligned}$$

Therefore, on region E, we have

$$\sum_j|\tilde{\beta}_j - \beta_j^*| \le C_{GN}\lambda_1 + (1 + \frac{2}{\lambda_1})\epsilon.$$

with probability at least $P(E) > 1-\delta$ by setting $a = \frac{1}{\kappa s+\lambda_2}$ This completes the proof of the GN penalty case.

For the TV penalty, the difference just lies in I_2. In fact, we have the inequality $|I_2| \leq 6\lambda_3||\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*||_1$ and $||\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*||_1 \leq (5+2\rho_{21}B_u+12\rho_{31})D$. Similar as (23), we have

$$\frac{1}{2}\lambda_1||\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_j^*||_1 + s||f_{\tilde{\boldsymbol{\beta}}} - f_{\boldsymbol{\beta}^*}||_2^2 - \frac{\lambda_1}{2}\epsilon \leq (2\lambda_1 + \lambda_2(C_1 D + 2B_u) + 6\lambda_3)\sum_{j\in I^*}(|\beta_j^* - \tilde{\beta}_j|) \quad (24)$$

on the event E. Under Assumption B, we can get

$$\frac{1}{2}\lambda_1\sum_j|\tilde{\beta}_j - \beta_j^*| + \kappa s\sum_{j\in I^*}(\tilde{\beta}_j - \beta_j^*)^2 - (\frac{\lambda_1}{2}+1)\epsilon \leq ak^*(1+\rho_{21}B_u+6\rho_{31})^2\lambda_1^2 + (\frac{1}{a}-\lambda_2)\sum_{j\in I^*}(\tilde{\beta}_j - \beta_j^*)^2.$$

Therefore, on region E, we have

$$\sum_j|\tilde{\beta}_j - \beta_j^*| \leq C_{TV}\lambda_1 + (1+\frac{2}{\lambda_1})\epsilon.$$

with probability at least $P(E) > 1-\delta$ by setting $a = \frac{1}{sb+\lambda_2}$.

REFERENCES

[1] J. Ashburner and K. Friston, "Voxel-based morphometry—The methods," *NeuroImage*, vol. 11, no. 6, pp. 805–821, Jun. 2000.

[2] C. Davatzikos, "Why voxel-based morphometric analysis should be used with great caution when characterizing group differences," *NeuroImage*, vol. 23, no. 1, pp. 17–20, Sep. 2004.

[3] Y. Fan, D. Shen, R. C. Gur, R. E. Gur, and C. Davatzikos, "COMPARE: Classification of morphological patterns using adaptive regional elements," *IEEE Trans. Med. Imag.*, vol. 26, no. 1, pp. 93–105, Jan. 2007.

[4] N. K. Batmanghelich, B. Taskar, and C. Davatzikos, "Generative-discriminative basis learning for medical imaging," *IEEE Trans. Med. Imag.*, vol. 31, no. 1, pp. 51–69, Jan. 2011.

[5] M. R. Sabuncu and K. Van Leemput, "The relevance voxel machine (RVoxM): A self-tuning Bayesian model for informative image-based prediction," *IEEE Trans. Med. Imag.*, vol. 31, no. 12, pp. 2290–2306, Dec. 2012.

[6] B. Krishnapuram, L. Carin, M. A. T. Figueiredo, and A. J. Hartemink, "Sparse multinomial logistic regression: Fast algorithms and generalization bounds," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 6, pp. 957–968, Jun. 2005.

[7] S. Ryali, K. Supekar, D. A. Abrams, and V. Menon, "Sparse logistic regression for whole-brain classification of fMRI data," *NeuroImage*, vol. 51, no. 2, pp. 752–764, Jun. 2010.

[8] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. Roy. Statist. Soc., B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[9] L. Grosenick, B. Klingenberg, K. Katovich, B. Knutson, and J. E. Taylor, "Interpretable whole-brain prediction analysis with GraphNet," *NeuroImage*, vol. 72, pp. 304–321, May 2013.

[10] B. M. Kandel, D. A. Wolk, J. C. Gee, and B. Avants, "Predicting cognitive data from medical images using sparse linear regression," in *Proc. IPMI*, 2013, pp. 86–97.

[11] B. Ng, A. Vahdat, G. Hamarneh, and R. Abugharbieh, "Generalized sparse classifiers for decoding cognitive states in fMRI," in *Proc. MLMI*, Beijing, China, Sep. 2010, pp. 108–115.

[12] T. Watanabe, D. Kessler, C. Scott, M. Angstadt, and C. Sripada, "Disease prediction based on functional connectomes using a scalable and spatially-informed support vector machine," *NeuroImage*, vol. 96, pp. 183–202, Aug. 2014.

[13] M. Dubois *et al.*, "Predictive support recovery with TV-Elastic Net penalty and logistic regression: An application to structural MRI," in *Proc. PRNI*, Jun. 2014, pp. 1–4.

[14] M. Eickenberg, E. Dohmatob, B. Thirion, and G. Varoquaux, "Grouping total variation and sparsity: Statistical learning with segmenting penalties," in *Proc. MICCAI*, 2015, pp. 685–693.

[15] E. D. Dohmatob, A. Gramfort, B. Thirion, and G. Varoquaux, "Benchmarking solvers for TV-ℓ_1 least-squares and logistic regression in brain imaging," in *Proc. Int. Workshop Pattern Recognit. Neuroimag.*, Jun. 2014, pp. 1–4.

[16] V. Michel, A. Gramfort, G. Varoquaux, E. Eger, and B. Thirion, "Total variation regularization for fMRI-based prediction of behavior," *IEEE Trans. Med. Imag.*, vol. 30, no. 7, pp. 1328–1340, Jul. 2011.

[17] A. Gramfort, B. Thirion, and G. Varoquaux, "Identifying predictive regions from fMRI with TV-ℓ_1 prior," in *Proc. Int. Workshop Pattern Recognit. Neuroimag.*, Jun. 2013, pp. 17–20.

[18] L. Baldassarre, J. Mourao-Miranda, and M. Pontil, "Structured sparsity models for brain decoding from fMRI data," in *Proc. Int. Workshop Pattern Recognit. NeuroImag.*, Jul. 2012, pp. 5–8.

[19] J. Tohka, E. Moradi, and H. Huttunen, "Comparison of feature selection techniques in machine learning for anatomical brain MRI in dementia," *Neuroinformatics*, vol. 14, no. 3, pp. 279–296, Jul. 2016.

[20] X. Huo and X. Ni, "When do stepwise algorithms meet subset selection criteria?" *Ann. Statist.*, vol. 35, no. 2, pp. 870–887, Apr. 2007.

[21] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Statist. Soc., B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[22] P. Zhao and B. Yu, "On model selection consistency of Lasso," *J. Mach. Learn. Res.*, vol. 7, pp. 2541–2563, Dec. 2006.

[23] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight, "Sparsity and smoothness via the fused lasso," *J. Roy. Statist. Soc. B (Statist. Methodol.)*, vol. 67, no. 1, pp. 91–108, 2005.

[24] F. Bunea, "Honest variable selection in linear and logistic regression models via ℓ_1 and $\ell_1+\ell_2$ penalization," *Electron. J. Statist.*, vol. 2, pp. 1153–1194, 2008.

[25] S. A. van de Geer, "High-dimensional generalized linear models and the lasso," *Ann. Statist.*, vol. 36, no. 2, pp. 614–645, Apr. 2008.

[26] R. Jenatton, J.-Y. Audibert, and F. Bach, "Structured variable selection with sparsity-inducing norms," *J. Mach. Learn. Res.*, vol. 12, pp. 2777–2824, Feb. 2011.

[27] S. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu, "A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers," *Statist. Sci.*, vol. 27, no. 4, pp. 538–557, 2012.

[28] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, Jan. 2011.

[29] K. Rohr, "On 3D differential operators for detecting point landmarks," *Image Vis. Comput.*, vol. 15, no. 3, pp. 219–233, Mar. 1997.

[30] M. Schmidt. (2005). *minFunc: Differentiable Multivariate Optimization in MATLAB*. [Online]. Available: http://www.cs.ubc.ca/~schmidtm/software/minfunc.html

[31] A. Walsh *et al.*, "Longitudinal MR imaging of iron in multiple sclerosis: An imaging marker of disease," *Radiology*, vol. 1, no. 270, pp. 186–196, 2014.

[32] E. Stephenson, N. Nathoo, Y. Mahjoub, J. F. Dunn, and V. W. Yong, "Iron in multiple sclerosis: Roles in neurodegeneration and repair," *Nature Rev. Neurol.*, vol. 10, no. 8, pp. 459–468, 2014.

[33] J. F. Schenck and E. A. Zimmerman, "High-field magnetic resonance imaging of brain iron: Birth of a biomarker?" *NMR Biomed.*, vol. 17, no. 7, pp. 433–445, Nov. 2004.

[34] S. Ropele *et al.*, "Determinants of iron accumulation in deep grey matter of multiple sclerosis patients," *Multiple Sclerosis J.*, vol. 20, no. 13, pp. 1692–1698, 2014.

[35] (2011). *ANTS*. [Online]. Available: http://www.picsl.upenn.edu/ants/

[36] B. Avants *et al.*, "The optimal template effect in hippocampus studies of diseased populations," *NeuroImage*, vol. 49, no. 3, pp. 2457–2466, Feb. 2010.

[37] N. J. Tustison *et al.*, "N4ITK: Improved N3 bias correction," *IEEE Trans. Med. Imag.*, vol. 29, no. 6, pp. 1310–1320, Jun. 2010.

[38] B. B. Avants, C. L. Epstein, M. Grossman, and J. C. Gee, "Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain," *Med. Image. Anal.*, vol. 12, no. 1, pp. 26–41, Feb. 2008.

[39] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, Jun. 2006.

[40] D. Cobzas *et al.*, "Subcortical gray matter segmentation and voxel-based analysis using transverse relaxation and quantitative susceptibility mapping with application to multiple sclerosis," *J. Magn. Reson. Imag.*, vol. 42, no. 6, pp. 1601–1610, Dec. 2015.