An Unbiased Penalty for Sparse Classification with Application to Neuroimaging Data

Li Zhang, Dana Cobzas(✉), Alan Wilman, and Linglong Kong

University of Alberta, Edmonton, Canada
cobzas@ualberta.ca

Abstract. We present a novel formulation 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. Such methods typically use an l_1 penalty for imposing sparseness and a graph net (GN) or a total variation (TV) penalty for ensuring spatial continuity and interpretability of the results. However it is known that the l_1 and TV penalties have inherent bias that leads to less stable region detection and less accurate prediction. To overcome these limitations, we propose a novel variable selection method in the context of classification, based on the Smoothly Clipped Absolute Deviation (SCAD) penalty. We experimentally show superiority of three models based on the SCAD and SCADTV penalties when compared to the classical l_1 and TV penalties in both simulated and real MRI data from a multiple sclerosis study.

Keywords: Sparse classification · Variable selection · Localized statistics · l_1 optimization · SCAD penalty

1 Introduction

With the growth of available medical imaging data, the need for good methods to perform large neuroimaging studies has increased. The majority of studies use voxel-based analysis (VBA) to identify regions where two groups differ [2]. 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 the high dimensionality of the data.

To overcome these limitations of VBA [6], alternative methods reformulate the region selection problem as a simultaneous feature selection and classification (or regression). Such methods typically use a sparse l_1 (LASSO) penalty and have been successfully applied to medical imaging data [12]. However, imposing sparsity can often lead to less stable feature maps that cannot be interpreted from an anatomical viewpoint. To counter this behavior, several estimators incorporate the notion of spatial smoothness on the coefficient maps through additional

M. Descoteaux et al. (Eds.): MICCAI 2017, Part III, LNCS 10435, pp. 55–63, 2017.
DOI: 10.1007/978-3-319-66179-7_7

penalizers. Two main types of image-based penalizers have been used in the literature. Graph net (GN) formulations use an l_2 penalty on the gradients to force adjacent voxels to have similar weights [10,11]. Alternatively regularization could be enforced by imposing sparsity on the spatial gradients through a total variation (TV) penalty [7,9]. These two types of penalties correspond to the linear and nonlinear diffusion and have been used in many image analysis applications like denoising, segmentation and registration.

However it is known that l_1 and TV penalties have inherent bias and often lead to less stable predictions [16]. To address these limitations of LASSO penalty, the Smoothly Clipped Absolute Deviation (SCAD) penalty [8] have been proposed in the context of high dimensional regression with variable selection. SCAD has become quite popular in the statistical community and proved to have some desired properties such as continuity, asymptotic unbiasedness and sparsity [8]. However these statistical works are limited to the one-dimensional case and there exist very few application of SCAD in image analysis [5,13].

In the current study we propose a novel regularized variable selection method, in the context of classification, based on the SCAD penalty. We use SCAD for enforcing sparsity of solution and SCAD of TV as the image regularization penalty for enforcing spatial continuity. Using synthetic and real MRI data from a multiple sclerosis study, we show superiority for variable selection for models based on the SCAD penalty when compared to the classical l_1 or TV penalties.

2 Methods

2.1 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) that can be formulated as minimizing the negative log-likelihhod of a logistic regression distribution:

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

One main problem with the solution of this problem is that all coefficients in $\boldsymbol{\beta}$ are usually nonzero. Sparse constraints on the solution address this issue. However, selecting the best subset of coefficients (l_0 norm) is an NP-hard problem, so an l_1 approximation of the l_0 penalizer is usually used [16]:

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

2.2 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. Image-based penalties provide a principled way of imposing anatomical continuity of selected regions. Two types of image-based penalizers have been explored in the context of sparse classification or regression: the GN penalty [10,11] and the TV penalty [7,9]. We limit our discussion to the TV penalty as it shown superior at variable selection compared to GN [7] and provides the base for the proposed extension to SCADTV. The TV penalty uses an l_1 norm on the image gradients. We use an anisotropic formulation of the TV-norm: $\|\nabla\boldsymbol{\beta}\|_1 = \|\nabla_i\boldsymbol{\beta}\|_1 + \|\nabla_j\boldsymbol{\beta}\|_1 + \|\nabla_k\boldsymbol{\beta}\|_1$, where (i, j, k) denotes the 3 orthogonal dimensions of the image data. Denoting by λ, γ two tuning parameters, the resulting penalized classification can then be written as:

$$\min_{\boldsymbol{\beta},b} \sum_{i=1}^{n} \log\left(1 + \exp(-y_i(\mathbf{x}_i\boldsymbol{\beta} + b))\right) + \mathcal{P}_{l_1+TV}(\boldsymbol{\beta}) \tag{3}$$

$$\mathcal{P}_{l_1+TV} = \lambda\left(\gamma\|\boldsymbol{\beta}\|_1 + (1-\gamma)\|\nabla\boldsymbol{\beta}\|_1\right) \tag{4}$$

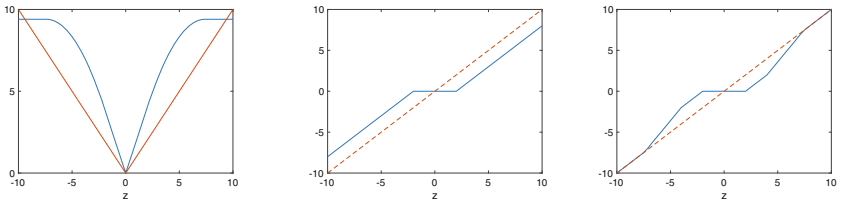
2.3 SCAD and SCADTV Penalties

The SCAD penalty $\rho_\lambda(.)$ is more conveniently defined by its derivative

$$\rho'_\lambda(t) = \lambda\left\{I(t \le \lambda) + \frac{(a\lambda - t)_+}{(a-1)\lambda} I(t > \lambda)\right\}, t > 0 \tag{5}$$

with $\rho_\lambda(0) = 0$, $(z)_+ = \max(z, 0)$, I the indicator function, λ and a model parameters. As usual, $a = 3.7$ is used [8]. Figure 1(a) shows the SCAD penalty (blue) and l_1 penalty (red).

To better understand the behavior of SCAD penalty, consider the penalized least square problem $\min_\beta(z-\beta)^2 + \mathcal{P}(\beta)$, where $\mathcal{P}(\beta)$ is chosen as the LASSO or the SCAD penalty. The solution is unique $\hat{\beta} = S_\lambda(z)$ where S_λ is a thresholding function. Figure 1 displays the thresholding function for LASSO (b) and SCAD (c) with $\lambda = 2$. We notice that the SCAD penalty shrinks small coefficients to zero, while keeping large coefficients intact, while the l_1 penalty tends to shrink



(a) SCAD and l_1 penalties **(b) l_1 thresholding** **(c) SCAD thresholding**

Fig. 1. Illustration of SCAD penalty: (a) SCAD (blue) and l_1 (red) penalty functions; thresholding function with l_1 (b) and SCAD (c) penalty and $\lambda = 2$.

all coefficients. This unbiased property of SCAD penalty comes from the fact that $\rho_\lambda(t) = 0$, when t is large enough.

Extending the SCAD definition for vector data and discrete gradients of the coefficients we define the combined SCAD and SCADTV penalties as:

$$\mathcal{P}_{SCAD} = \sum_{l=1}^{m} \rho_\lambda(|\beta_l|) \tag{6}$$

$$\mathcal{P}_{SCADTV} = \sum_{l=1}^{m} \rho_\lambda(|\nabla_i \beta_l|) + \rho_\lambda(|\nabla_j \beta_l|) + \rho_\lambda(|\nabla_k \beta_l|) \tag{7}$$

where (i, j, k) denotes the 3 orthogonal dimensions of the image data as in the definition of TV norm. Similar to SCAD, SCADTV shrinks small gradients encouraging neighboring coefficients to have the same values, but leaves large gradients unchanged. We propose three types of penalty functions that are compared with the classic $\mathcal{P}_{l_1+TV}$ model in the context of logistic regression classification: $\mathcal{P}_{SCAD+SCADTV}$, $\mathcal{P}_{l_1+SCADTV}$ and $\mathcal{P}_{SCAD+TV}$.

2.4 Optimization and Parameter Tuning

Note that the SCAD penalty, unlike l_1 and TV, is not convex. We solve this problem using ADMM [4] that was successfully applied to convex problems. Recently it was shown [18] that several ADMM algorithms including SCAD are guaranteed to converge. The tuning parameters λ, γ are chosen by generalized information criterion (GIC).

3 Experimental Results

3.1 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 Fig. 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 $N(0, \Sigma_r)$, with $r \in \{0.25, 0.5\}$. Binary labels y_i are then assigned based on the logistic probability following a Bernoulli distribution. The coefficient vector $\boldsymbol{\beta}$ has fixed values of 0 outside the four blocks and piecewise smooth values inside, with increasing strength for the data signal in the following order: top-left, top-right, bottom-right, bottom-left. Figure 2 top-left presents a 2D slice of the synthetic data and bottom-left presents the 3D view of the nonzero coefficients. Binary labels are assigned based on the logistic probability following a Bernoulli distribution. Each dataset contains $n = 300$ subjects, making the data matrix X of size $n \times 8192$. For each coherence value r we repeated the test 96 times.

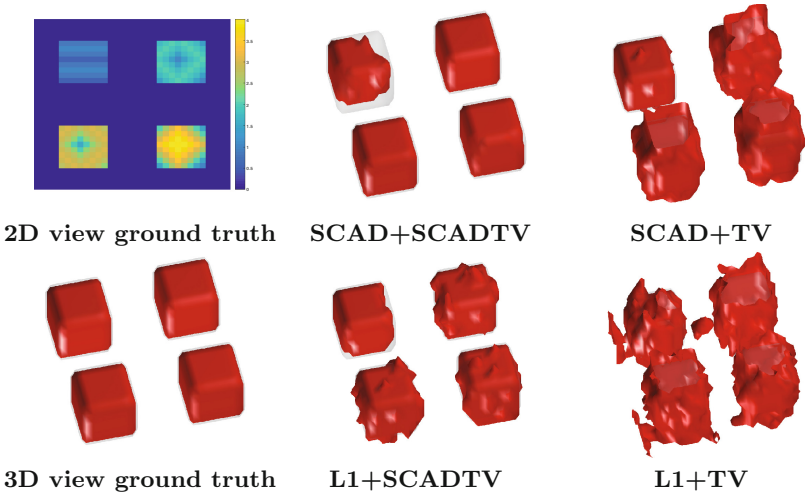


Fig. 2. (top-left) shows a 2D slice of the ground truth coefficients for simulated data. (bottom-left) shows a 3D view of the ground truth nonzero coefficients. The following figures show significant regions on synthetic data detected by the 4 methods. Shaded gray regions correspond to the true nonzero coefficients and the red regions are calculated from the estimated nonzero coefficients averaged over 96 trials.

3.2 Neuroimaging Data

Our neuroimaging data belongs to an in-house multiple sclerosis (MS) study. Following recent research that suggests a possible pivotal role for iron in MS [15], 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 [14]. Sample R2* slices can be viewed in Fig. 4 (top). The focus is subcortical deep gray matter structures: caudate, putamen, thalamus and global pallidus. Forty subjects with relapsing remitting MS (RRMS) and 40 age- and gender-matched controls were recruited. Ethical approval and informed consent were obtained.

Prior to analysis, the MRI data is pre-processed and aligned with an in-house unbiased template using ANTs [1]. The multimodal template is built from 10 healthy controls using both T1w and R2*. Pre-processing involves intra-subject alignment of R2* with T1w and bias field intensity normalization for T1w [17]. Nonlinear registration in the template space is done using SyN [3]. 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 manually traced on the atlas.

3.3 Evaluation Methodology

We compare the performance of the four penalized logistic regression models described in Sect. 2: $SCAD + SCADTV$, $SCAD + TV$, $l_1 + SCADTV$ and

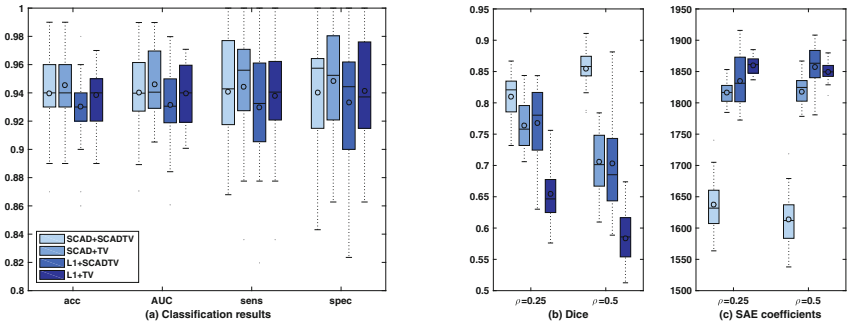


Fig. 3. Results for synthetic experiments (a) Classification scores for noise level $r = 0.25$. (b) Dice scores between ground truth and estimated nonzero coeff. (c) Sum of Absolute Error (SAE) between ground truth and estimated coeff. for $r = 0.25, 0.5$.

$l_1 + TV$. Training and test data is selected for each of the 96 synthetic datasets (200 training and 100 test) and for the real data (5 folds cross-validation). 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 coefficients.

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. Variable selection accuracy compared to ground truth for synthetic data is evaluated using a dice score. We also compute the mean absolute error of recovered vs ground truth coefficients. For real data, we measured the stability of the detected regions using a dice score between the estimated regions in each of the 5 folds (dice folds).

3.4 Results

Comparative results on **synthetic data** with two levels of coherence $r \in \{0.25, 0.5\}$ for the multinormal distribution are reported in the bar graphs in Fig. 3(a), (b) and (c). When evaluating the classification accuracy in plot (a) results are comparable for all four methods with a mean of about 94% for $SCAD + SCADTV$, $SCAD + TV$ and $L1 + TV$ and a bit lower for $L1 + SCADTV$. But, when looking at the accuracy of variable selection using dice score (b) as well as the accuracy of the recovered sparse coefficients (c), we see that the $SCAD + SCADTV$ penalizer is superior compared to the others. It achieves the highest dice score and the lowest SAD of recovered coefficients. To visualize the results of the 96 trials, we average the estimated nonzero coefficients, as binary masks, and threshold at 0.2. Results as illustrated in Fig. 2 confirm the numerical evaluation showing that the $SCAD + SCADTV$ penalty gives the cleanest and closest to ground truth variable selection results while the $l_1 + TV$ penalty archives the worse performance.

Table 1. Results for real MRI data. Means on the 5 folds are reported. Class. rate = classification rate, Sens. = sensitivity, Spec. = specificity, AUC; Dice Folds = Dice score between detected sparse regions. Bold highlights best results among methods.

Method	Class. rate	Sens.	Spec.	AUC	Dice Folds
$SCAD + SCADTV$	0.75	0.75	0.75	**0.75**	**0.79**
$SCAD + TV$	0.71	0.71	0.75	0.67	0.71
$L1 + SCADTV$	**0.76**	**0.76**	**0.77**	**0.75**	0.68
$L1 + TV$	0.74	0.74	0.75	0.72	0.61

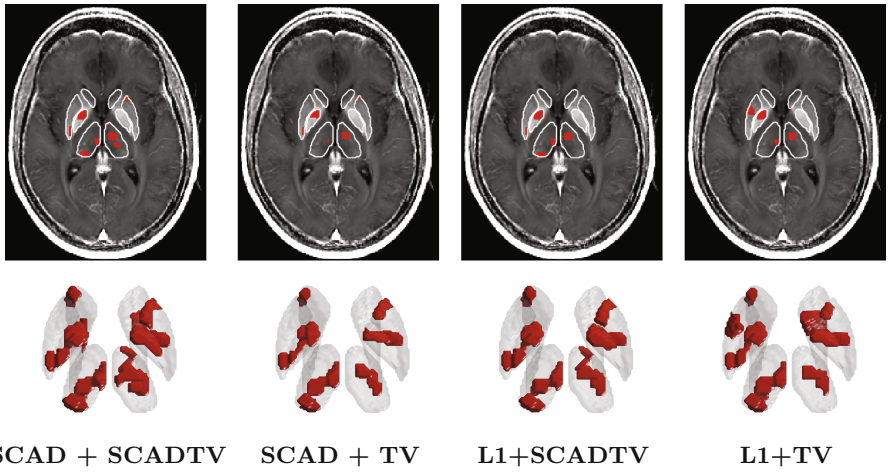


Fig. 4. Illustration of the significant anatomy detected by the 4 methods using MRI data. Top: 2D axial slices with the R2* data as background; Bottom: a 3D view of the result. The deep gray matter mask used for selecting the voxels included in the observation vectors $\mathbf{x}_i$ is contoured in white and the selected significant regions in red.

Comparative classification results on **real neuroimaging MRI data** are reported in Table 1. As ground truth on selected sparse regions is not available for real data, we estimated the quality of the detected sparse regions using a stability over folds measured using between-folds dice scores (DiceFolds). We report the average over the 10 distinct folds combinations. While classification results are comparable among proposed penalizers, results on stability of detected regions clearly show that the new penalties SCAD and SCADTV achieve superior results. To visualize the results, Fig. 4 displays sample axial slices and a 3D view of the regions recovered by the four methods. The regions were calculated from all data with optimal parameters for each method. Most methods recover compact regions in very similar brain locations.

4 Discussion

We introduced a new penalty based on SCAD for variable selection in the context of sparse classification in high dimensional neuroimaging data. While SCAD penalty was proposed in statistical literature to overcome the inherent bias of l_1 and TV penalties, it was not yet used in medical imaging population studies. We experimentally shown on simulated and real MRI data that the proposed models based on SCAD are better at selecting the true nonzero coefficients and achieve higher accuracy. Part of our future work, we are looking at deriving theoretical results on coefficients bounds and accuracy of variable selection for the SCAD based models. Extending our work, similar penalizers could be used for regression or data representation (ex. PCA, CCA).

References

1. ANTS (2011). http://www.picsl.upenn.edu/ants/
2. Ashburner, J., Friston, K.: Voxel-based morphometry - the methods. NeuroImage **11**(6), 805–821 (2000)
3. Avants, B., Epstein, C., Grossman, M., Gee, J.: Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. Med. Image Anal. **12**(1), 26–41 (2008)
4. Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J.: Distributed optimization and statistical learning via the alternating direction method of multipliers. Found. Trends Mach. Learn. **3**(1), 1–122 (2011)
5. Chopra, A., Lian, H.: Total variation, adaptive total variation and nonconvex smoothly clipped absolute deviation penalty for denoising blocky images. Pattern Recogn. **43**(8), 2609–2619 (2010)
6. Davatzikos, C.: Why voxel-based morphometric analysis should be used with great caution when characterizing group differences. Neuroimage **23**(1), 17–20 (2004)
7. Eickenberg, M., Dohmatob, E., Thirion, B., Varoquaux, G.: Grouping total variation and sparsity: statistical learning with segmenting penalties. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9349, pp. 685–693. Springer, Cham (2015). doi:10.1007/978-3-319-24553-9_84
8. Fan, J., Li, R.: Variable selection via nonconcave penalized likelihood and its Oracle properties. J. Am. Stat. Assoc. **96**(456), 1348–1360 (2001)
9. Gramfort, A., Thirion, B., Varoquaux, G.: Identifying predictive regions from fMRI with TV-L1 prior. In: International Workshop on PRNI, pp. 17–20 (2013)
10. Grosenick, L., Klingenberg, B., Katovich, K., Knutson, B., Taylor, J.: Interpretable whole-brain prediction analysis with graphnet. Neuroimage **72**, 304–21 (2013)
11. Kandel, B., Avants, B., Gee, J., Wolk, D.: Predicting cognitive data from medical images using sparse linear regression. In: IPMI, pp. 86–97 (2013)
12. Krishnapuram, B., Carin, L., Figueiredo, M., Hartemink, A.: Sparse multinomial logistic regression: fast algorithms and generalization bounds. IEEE Trans. Pattern Anal. Mach. Intell. **27**(6), 957–68 (2005)
13. Mehranian, A., Rad, H.S., Rahmim, A., Ay, M.R., Zaidi, H.: Smoothly clipped absolute deviation (SCAD) regularization for compressed sensing MRI using an augmented lagrangian scheme. Magn. Reson. Imaging **31**(8), 1399–1411 (2013)
14. Schenck, J., Zimmerman, E.: High-field MRI of brain iron: birth of a biomarker? NMR Biomed. **17**, 433–45 (2004)

15. Stephenson, E., Nathoo, N., Mahjoub, Y., et al.: Iron in multiple sclerosis: roles in neurodegeneration and repair. Nat. Rev. Neurol. **10**(8), 459–68 (2014)
16. Tibshirani, R.: Regression shrinkage and selection via the Lasso. J. Royal Stat. Soc. **58**(1), 267–288 (1996)
17. Tustison, N., Avants, B., Cook, P., et al.: N4ITK: improved N3 bias correction. IEEE Trans. Med. Imaging **29**(6), 1310–20 (2010)
18. Wang, Y., Yin, W., Zeng, J.: Global convergence for ADMM in nonconvex nonsmooth optimization, arXiv:1551.06324