GSplit LBI: Taming the Procedural Bias in Neuroimaging for Disease Prediction

Xinwei Sun[1], Lingjing Hu[2(✉)], Yuan Yao[3,4(✉)], and Yizhou Wang[5]

[1] School of Mathematical Science, Peking University, Beijing 100871, China
[2] Yanjing Medical College, Capital Medical University, Beijing 101300, China
hulj@ccmu.edu.cn
[3] Hong Kong University of Science and Technology, Hong Kong, Hong Kong
[4] Peking University, Beijing, China
yuany@ust.hk
[5] National Engineering Laboratory for Video Technology, Key Laboratory of Machine Perception, School of EECS, Peking University, Beijing 100871, China

Abstract. In voxel-based neuroimage analysis, lesion features have been the main focus in disease prediction due to their interpretability with respect to the related diseases. However, we observe that there exist another type of features introduced during the preprocessing steps and we call them "**Procedural Bias**". Besides, such bias can be leveraged to improve classification accuracy. Nevertheless, most existing models suffer from either under-fit without considering procedural bias or poor interpretability without differentiating such bias from lesion ones. In this paper, a novel dual-task algorithm namely *GSplit LBI* is proposed to resolve this problem. By introducing an augmented variable enforced to be structural sparsity with a variable splitting term, the estimators for prediction and selecting lesion features can be optimized separately and mutually monitored by each other following an iterative scheme. Empirical experiments have been evaluated on the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. The advantage of proposed model is verified by improved stability of selected lesion features and better classification results.

Keywords: Voxel-based structural magnetic resonance imaging · Procedural bias · Split Linearized Bregman Iteration · Feature selection

1 Introduction

Usually, the first step of voxel-based neuroimage analysis requires preprocessing the T_1-weighted image, such as segmentation and registration of grey matter (GM), white matter (WM) and cerebral spinal fluid (CSF). However, some

Electronic supplementary material The online version of this chapter (doi:10.1007/978-3-319-66179-7_13) contains supplementary material, which is available to authorized users.

© Springer International Publishing AG 2017
M. Descoteaux et al. (Eds.): MICCAI 2017, Part III, LNCS 10435, pp. 107–115, 2017.
DOI: 10.1007/978-3-319-66179-7_13

systematic biases due to scanner difference and different population etc., can be introduced in this pipeline [2]. Part of them can be helpful to the discrimination of subjects from normal controls (NC), but may not be directly related to the disease. For example in structural Magnetic Resonance Imaging (sMRI) images of subjects with Alzheimer's Disease (AD), after spatial normalization during simultaneous registration of GM, WM and CSF, the GM voxels surrounding lateral ventricle and subarachnoid space etc. may be mistakenly enlarged caused by the enlargement of CSF space in those locations [2] compared to normal template, as shown in Fig. 1. Although these voxels/features are highly correlated with disease, they can't be regarded as lesion features in an interpretable model. In this paper we refer to them as "**Procedural Bias**", which should be identified but is neglected in the literature. We observe that it can be harnessed in our voxel-based image analysis to improve the prediction of disease.

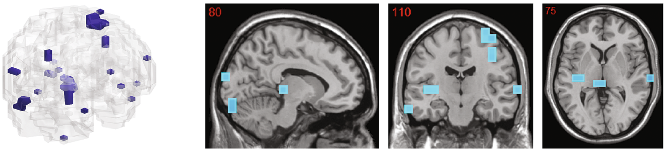


Fig. 1. The overlapped voxels among top 150 negative value voxels in each fold of β_{pre} at the time corresponding to the best average prediction result in the path of GSplit LBI using 10-fold cross-validation. For subjects with AD, they represent enlarged GM voxels surrounding lateral ventricle, subarachnoid space, edge of gyrus, etc.

Together with procedural bias, the lesion features are vital for prediction and lesion regions analysis tasks, which are commonly solved by two types of regularization models. Specifically, one kind of models such as general losses with l_2 penalty, elastic net [13] and graphnet [5] select strongly correlated features to minimize classification error. However, such models don't differentiate features either introduced by disease or procedural bias and may also introduce redundant features. Hence, the interpretability of such models are poor and the models are prone to over-fit. The other kind of models with sparsity enforcement such as TV-L_1 (Combination of Total Variation [9] and L_1) and particularly n^2 GFL [12] enforce strong prior of disease on the parameters of the models introduced in order to capture the lesion features. Although such features are disease-relevant and the selection is stable, the models ignore the inevitable procedural bias, hence, they are losing some prediction power.

To incorporate both tasks of prediction and selection of lesion features, we propose an iterative dual-task algorithm namely *Generalized Split LBI* (GSplit LBI) which can have better model selection consistency than generalized lasso [11]. Specifically, by the introduction of variable splitting term inspired by Split LBI [6], two estimators are introduced and split apart. One estimator is for prediction and the other is for selecting lesion features, both of which can be pursued separately with a gap control. Following an iterative scheme, they will be mutually monitored by each other: the estimator for selecting lesion features is gradually monitored to pursue stable lesion features; on the other hand, the estimator for prediction is also monitored to exploit both the procedural bias and lesion features to improve prediction. To show the validity of the proposed method, we successfully apply our model to voxel-based sMRI analysis for AD, which is challenging and attracts increasing attention.

2 Method

2.1 GSplit LBI Algorithm

Our dataset consists of N samples $\{x_i, y_i\}_1^N$ where $x_i \in \mathbb{R}^p$ collects the i^{th} neuroimaging data with p voxels and $y_i = \{\pm 1\}$ indicates the disease status (-1 for Alzheimer's disease in this paper). $X \in \mathbb{R}^{N \times p}$ and $y \in \mathbb{R}^p$ are concatenations of $\{x_i\}_i$ and $\{y_i\}_i$. Consider a general linear model to predict the disease status (with the intercept parameter $\beta_0 \in \mathbb{R}$),

$$\log P(y_i = 1|x_i) - \log P(y_i = -1|x_i) = x_i^T \beta_{pre} + \beta_0. \tag{2.1}$$

A desired estimator $\beta_{pre} \in \mathbb{R}^p$ should not only fit the data by maximizing the log-likelihood in logistic regression, but also satisfy the following types of structural sparsity: (1) the number of voxels involved in the disease prediction is small, so β_{pre} is sparse; (2) the voxel activities should be geometrically clustered or 3D-smooth, suggesting a TV-type sparsity on $D_G\beta_{pre}$ where D_G is a graph difference operator[1]; (3) the degenerate GM voxels in AD are captured by nonnegative component in β_{pre}. However, the existing procedural bias may violate these *a priori* sparsity properties, *esp.* the third one, yet increase the prediction power.

To overcome this issue, we adopt a variable splitting idea in [6] by introducing an auxiliary variable $\gamma \in \mathbb{R}^{|V|+|E|}$ to achieve these sparsity requirements separately, while controlling the gap from $D\beta_{pre}$ with penalty $S_\rho(\beta_{pre}, \gamma) := \|D\beta_{pre} - \gamma\|_2^2 := \|\beta_{pre} - \gamma_V\|_2^2 + \|\rho D_G \beta_{pre} - \gamma_G\|_2^2$ with $\gamma = \left[\gamma_V^T \; \gamma_G^T\right]^T$ and $D = \left[I \; \rho D_G^T\right]^T$. Here ρ controls the trade-off between different types of sparsity. Our purpose is thus of two-folds: (1) use β_{pre} for prediction; (2) enforce sparsity on γ. Such a dual-task scheme can be illustrated by Fig. 2.

To implement it, we generalize the Split Linearized Bregman Iteration (Split LBI) algorithm in [6] to our setting with generalized linear models (GLM) and the three types of structural sparsity above, hence called Generalized Split LBI (or GSplit LBI). Algorithm 1 describes the procedure with a new loss:

$$\ell(\beta_0, \beta_{pre}, \gamma; \{x_i, y_i\}_1^N, \nu) := \ell(\beta_0, \beta_{pre}; \{x_i, y_i\}_1^N) + \frac{1}{2\nu} S_\rho(\beta_{pre}, \gamma), \tag{2.2}$$

where $\ell(\beta_{pre}; \{x_i, y_i\}_1^N)$ is the negative log-likelihood function for GLM and $\nu > 0$ tunes the strength of gap control. The algorithm returns a sequence of estimates

[1] Here $D_G : \mathbb{R}^V \to \mathbb{R}^E$ denotes a graph difference operator on $G = (V, E)$, where V is the node set of voxels, E is the edge set of voxel pairs in neighbour (e.g. 3-by-3-by-3), such that $D_G(\beta)(i, j) := \beta(i) - \beta(j)$.

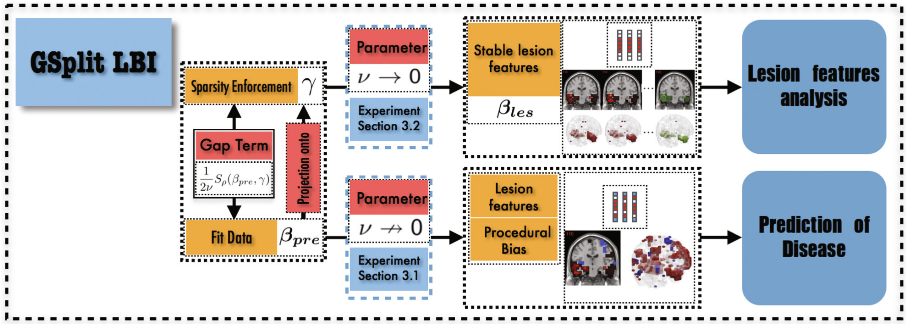


Fig. 2. Illustration of *GSplit LBI*. The gap between β_{pre} for fitting data and γ for sparsity is controlled by $S_\rho(\beta_{pre}, \gamma)$. The estimate β_{les}, as a projection of β_{pre} on support set of γ, can be used for stable lesion features analysis when $\nu \to 0$ (Sect. 3.2). When $\nu \nrightarrow 0$ (Sect. 3.1) with appropriately large value, β_{pre} can be used for prediction by capturing both lesion features and procedural bias.

Algorithm 1. GSplit LBI

Input: Loss function $\ell(\beta_0, \beta_{pre}, \gamma; \{x_i, y_i\}_{i=1}^N, \nu)$, parameters ν, ρ, κ, $\alpha > 0$.

Initialize: $k = 0$, $t^k = 0$, $\beta_0^k = 0$, $\beta_{les}^k = 0$, $\beta_{pre}^k = 0$, $\gamma_V^k = 0_p$, $\gamma_G^k = 0_m$, $z_V^k = 0_p$, $z_G^k = 0_m$ and $S_k := \mathrm{supp}(\gamma^k) = \emptyset$.

Iteration

$\beta_0^{k+1} = \beta_0^k - \kappa\alpha\nabla_{\beta_0}\ell(\beta_0^k, \beta_{pre}^k, \gamma^k; \{x_i, y_i\}_1^N, \nu)$

$\beta_{pre}^{k+1} = \beta_{pre}^k - \kappa\alpha\nabla_{\beta_{pre}}\ell(\beta_0^k, \beta_{pre}^k, \gamma^k; \{x_i, y_i\}_1^N, \nu)$

$z^{k+1} = z^k - \alpha\nabla_\gamma \ell(\beta_0^k, \beta_{pre}^k, \gamma^k; \{x_i, y_i\}_1^N, \nu)$

$\gamma_V^{k+1} = \kappa \cdot \mathcal{S}^+(z_V^{k+1}, 1)$, where $\mathcal{S}^+(x, 1) = \max(x - 1, 0)$

$\gamma_G^{k+1} = \kappa \cdot \mathcal{S}(z_G^{k+1}, 1)$, where $\mathcal{S}(x, 1) = \mathrm{sign}(x) \cdot \max(|x| - 1, 0)$

$\beta_{les}^{k+1} = P_{S_{k+1}}\beta_{pre}^{k+1}$, where $P_S = P_{ker(D_{S^c})} = I - D_{S^c}^\dagger D_{S^c}$

$t^{k+1} = (k+1)\alpha$

Output: $\{\beta_0^k, \beta_{pre}^k, \beta_{les}^k, \gamma^k\}$, where $\gamma^{k+1} = \begin{bmatrix} \gamma_V^{k+1} \\ \gamma_G^{k+1} \end{bmatrix}$ and $z^{k+1} = \begin{bmatrix} z_V^{k+1} \\ z_G^{k+1} \end{bmatrix}$.

---

as a regularization path, $\{\beta_0^k, \beta_{pre}^k, \gamma^k, \beta_{les}^k\}_{k\geq 0}$. In particular, γ^k shows a variety of sparsity levels and β_{pre}^k is generically dense with different prediction powers. The projection of β_{pre}^k onto the subspace with the same support of γ^k gives estimate β_{les}^k, satisfying those *a priori* sparsity properties (sparse, 3D-smooth, nonnegative) and hence being regarded as the interpretable lesion features for AD. The remainder of this projection is heavily influenced by procedural bias; in this paper the non-zero elements in β_{pre}^k which are negative (-1 denotes disease label) with comparably large magnitude are identified as procedural bias, while others with tiny values can be treated as nuisance or weak features. In summary, β_{les} only selects lesion features; while β_{pre} also captures additional procedural bias. Hence, such two kinds of features can be differentiated, as illustrated in Fig. 2.

2.2 Setting the Parameters

A stopping time at t^k (line 10) is the regularization parameter, which can be determined via cross-validation to minimize the prediction error [7]. Parameter ρ is a tradeoff between geometric clustering and voxel sparsity. Parameter κ, α is damping factor and step size, which should satisfy $\kappa\alpha \leq \nu/\kappa(1 + \nu\Lambda_H + \Lambda_D^2)$ to ensure the stability of iterations. Here $\Lambda_{(\cdot)}$ denotes the largest singular value of a matrix and H denotes the Hessian matrix of $\ell(\beta_0, \beta_{pre}; \{x_i, y_i\}_1^N)$.

Parameter ν balances the prediction task and sparsity enforcement in feature selection. In this paper, it is task-dependent, as shown in Fig. 2. For prediction of disease, β_{pre} with appropriately larger value of ν may increase the prediction power by harnessing both lesion features and procedural bias. For lesion features analysis, β_{les} with a small value of ν is helpful to enhance stability of feature selection. For details please refer to supplementary information.

3 Experimental Results

We apply our model to AD/NC classification (namely ADNC) and MCI (Mild Cognitive Impairment)/NC (namely MCINC) classification, which are two fundamental challenges in diagnosis of AD. The data are obtained from ADNI[2] database, which is split into 1.5 T and 3.0T (namely 15 and 30) MRI scan magnetic field strength datasets. The 15 dataset contains 64 AD, 208 MCI and 90 NC; while the 30 dataset contains 66 AD and 110 NC. DARTEL VBM pipeline [1] is then implemented to preprocess the data. Finally, the input features consist of 2,527 $8 \times 8 \times 8\,\text{mm}^3$ size voxels with average values in GM population template greater than 0.1. Experiments are designed on 15ADNC, 30ADNC and 15MCINC tasks.

3.1 Prediction and Path Analysis

10-fold cross-validation is adopted for classification evaluation. Under exactly the same experimental setup, comparison is made between GSplit LBI and other classifiers: SVM, MLDA (univariate model via t-test + LDA) [3], Graphnet [5], Lasso [10], Elastic Net, TV+L_1 and n^2GFL. For each model, optimal parameters are determined by grid-search. For GSplit LBI, ρ is chosen from $\{1, 2, ..., 10\}$, κ is set to 10; $\alpha = \nu/\kappa(1 + \nu\Lambda_X^2 + \Lambda_D^2)$[3]; specifically, ν is set to 0.2 (corresponding to $\nu \nrightarrow 0$ in Fig. 2)[4]. The regularization coefficient λ is ranged in $\{0, 0.05, 0.1, ..., 0.95, 1, 10, 10^2\}$ for lasso[5] and $2^{\{-20,-19,...,0,...,20\}}$ for SVM. For other models, parameters are optimized from $\lambda : \{0.05, 0.1, ..., 0.95, 1, 10, 10^2\}$ and $\rho : \{0.5, 1, .., 10\}$ (in addition, the mixture parameter α: $\{0, 0.05, ..., 0.95\}$ for Elastic Net).

[2] http://adni.loni.ucla.edu.

[3] For logit model, $\alpha < \nu/\kappa(1 + \nu\Lambda_H^2 + \nu\Lambda_X^2)$ since $\Lambda_X > \Lambda_H$.

[4] In this experiment, comparable prediction result will be given for $\nu \in (0.1, 10)$.

[5] 0 corresponds to logistic regression model.

Table 1. Comparison of GSplit LBI with other models

	MLDA	SVM	Lasso	Graphnet	Elastic net	TV + l_1	n^2GFL	GSplit LBI (β_{pre})
15ADNC	85.06%	83.12%	87.01%	86.36%	88.31%	83.77%	86.36%	**88.96%**
30ADNC	86.93%	87.50%	87.50%	88.64%	89.20%	87.50%	87.50%	**90.91%**
15MCINC	61.41%	70.13%	69.80%	72.15%	70.13%	73.83%	69.80%	**75.17%**

The best accuracy in the path of GSplit LBI and counterpart are reported. Table 1 shows that β_{pre} of our model outperforms that of others in all cases. Note that although our accuracies may not be superior to models with multi-modality data [8], they are the state-of-the-art results for only sMRI modality.

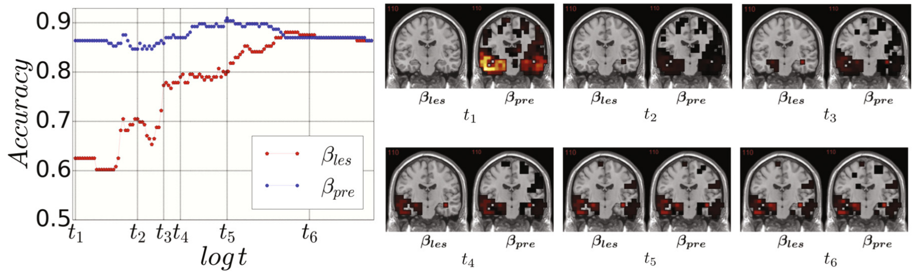


Fig. 3. Left image: Accuracy of $(\beta_{pre}, \beta_{les})$ vs $log\, t$ (t: regularization parameter). Right image: Six 2-d brain slice images of selected degenerative voxels of β_{les} and β_{pre} are sorted orderly at $\{t_1, ...t_6\}$. As t grows, β_{pre} and β_{les} identify similar lesion features.

The process of feature selection combined with prediction accuracy can be analyzed together along the path. The result of 30ADNC is used as an illustration in Fig. 3. We can see that β_{pre} (blue curve) outperforms β_{les} (red curve) in the whole path for additional procedural bias captured by β_{pre}. Specifically, at β_{pre}'s highest accuracy (t_5), there is a more than 8% increase in prediction accuracy by β_{pre}. Early stopping regularization at t_5 is desired, as β_{pre} converges to β_{les} in prediction accuracy with overfitting when t grows. Recall that positive (negative) features represent degenerate (enlarged) voxels. In each fold of β_{pre} at t_5, the commonly selected voxels among top 150 negative (enlargement) voxels are identified as procedural bias shown in Fig. 1, where most of these GM voxels are enlarged and located near lateral ventricle or subarachnoid space etc., possibly due to enlargement of CSF space in those locations that are different from the lesion features.

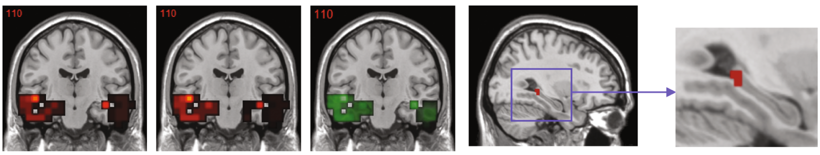
3.2 Lesion Features Analysis

To quantitatively evaluate the stability of selected lesion features, multi-set Dice Coefficient (mDC)[6] [4,12] is applied as a measurement. The 30ADNC task is again applied as an example, the mDC is computed for β_{les} which achieves highest accuracy by 10-fold cross-validation. As shown from Table 2, when $\nu = 0.0002$ (corresponding to $\nu \to 0$ in Fig. 2), the β_{les} of our model can obtain more stable lesion feature selection results than other models with comparable prediction power. Besides, the average number of selected features (line 3 in Table 2) are also recorded. Note that although elastic net is of slightly higher accuracy than β_{les}, it selects much more features than necessary.

Table 2. mDC comparison between GSplit LBI and other models

	Lasso	Elastic Net	Graphnet	TV + l_1	n^2 GFL	GSplit LBI (β_{les})
Accuracy	87.50%	89.20%	88.64%	87.50%	87.50%	88.64%
mDC	0.1992	0.5631	0.6005	0.5824	0.5362	**0.7805**
$\sum_{k=1}^{10} \|S(k)\|/10$	50.2	777.8	832.6	712.6	443.9	129.4

For the meaningfulness of selected lesion features, they are shown in Fig. 4(a)–(c), located in hippocampus, parahippocampal gyrus and medial temporal lobe etc., which are believed to be early damaged regions for AD patients.



(a) fold 2 (b) fold 10 (c) overlap (d) coarse-to-fine

Fig. 4. (a)–(c): Stability of selected lesion features of β_{les} shown in 2-d 110 slice brain images when $\nu = 0.0002$. (a)–(b): Results of fold 2 and fold 10. (c): The overlapped features in 10 folds. (d): The 2-d slice brain image of selected voxels with $2 \times 2 \times 2\,\text{mm}^3$ using coarse-to-fine approach.

To further investigate the locus of lesion features, we conduct a coarse-to-fine experiment. Specifically, we project the selected overlapped voxels of $8 \times 8 \times 8\,\text{mm}^3$ size (shown in Fig. 4(c)) onto MRI image with more finer scale voxels, i.e. in size of $2 \times 2 \times 2\,\text{mm}^3$. Totally 4,895 voxels are served as input

[6] In [12], $mDC := \frac{10|\cap_{k=1}^{10} S(k)|}{\sum_{k=1}^{10} |S(k)|}$ where $S(k)$ denotes the support set of β_{les} in k-th fold.

features after projection. Again, the GSplit LBI is implemented using 10-fold cross-validation. The prediction accuracy of β_{pre} is 90.34% and on average 446.6 voxels are selected by β_{les}. As desired, these voxels belong to parts of lesion regions, such as those located in hippocampal tail, as shown in Fig. 4(d).

4 Conclusions

In this paper, a novel iterative dual task algorithm is proposed to incorporate both disease prediction and lesion feature selection in neuroimage analysis. With variable splitting term, the estimators for prediction and selecting lesion features can be separately pursued and mutually monitored under a gap control. The gap here is dominated by procedural bias, some specific features crucial for prediction yet ignored in *a priori* disease knowledge. With experimental studies conducted on 15ADNC, 30ADNC and 15MCINC tasks, we have shown that the leverage of procedural bias can lead to significant improvements in both prediction and model interpretability. In future works, we shall extend our model to other neuroimaging applications including multi-modality data.

Acknowledgements. This work was supported in part by 973-2015CB351800, 2015CB85600, 2012CB825501, NSFC-61625201, 61370004, 11421110001 and Scientific Research Common Program of Beijing Municipal Commission of Education (No. KM201610025013).

References

1. Ashburner, J.: A fast diffeomorphic image registration algorithm. Neuroimage **38**(1), 95–113 (2007)
2. Ashburner, J., Friston, K.J.: Why voxel-based morphometry should be used. Neuroimage **14**(6), 1238–1243 (2001)
3. Dai, Z., Yan, C., Wang, Z., Wang, J., Xia, M., Li, K., He, Y.: Discriminative analysis of early alzheimer's disease using multi-modal imaging and multi-level characterization with multi-classifier. Neuroimage **59**(3), 2187–2195 (2012)
4. Dice, L.R.: Measures of the amount of ecologic association between species. Ecology **26**(3), 297–302 (1945)
5. Grosenick, L., Klingenberg, B., Katovich, K., Knutson, B., Taylor, J.E.: Interpretable whole-brain prediction analysis with graphnet. Neuroimage **72**, 304–321 (2013)
6. Huang, C., Sun, X., Xiong, J., Yao, Y.: Split lbi: An iterative regularization path with structural sparsity. In: Advances In Neural Information Processing Systems, pp. 3369–3377 (2016)
7. Osher, S., Ruan, F., Xiong, J., Yao, Y., Yin, W.: Sparse recovery via differential inclusions. Appl. Comput. Harmonic Anal. **41**(2), 436–469 (2016)
8. Peng, J., An, L., Zhu, X., Jin, Y., Shen, D.: Structured sparse kernel learning for imaging genetics based alzheimer's disease diagnosis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 70–78 (2016)

9. Rudin, L.I., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. Physica D **60**(1–4), 259–268 (1992)
10. Tibshirani, R.: Regression shrinkage and selection via the lasso. J. Roy. Stat. Soc.: Ser. B (Methodol.) **58**, 267–288 (1996)
11. Tibshirani, R.J., Taylor, J.E., Candes, E.J., Hastie, T.: The solution path of the generalized lasso. Ann. Stat. **39**(3), 1335–1371 (2011)
12. Xin, B., Hu, L., Wang, Y., Gao, W.: Stable feature selection from brain smri. In: AAAI, pp. 1910–1916 (2014)
13. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. J. Roy. Stat. Soc. Ser. B (Statistical Methodology) **67**(2), 301–320 (2005)