Temporally-Constrained Group Sparse Learning for Longitudinal Data Analysis

Daoqiang Zhang[1,2], Jun Liu[3], and Dinggang Shen[1]

[1] Dept. of Radiology and BRIC, University of North Carolina at Chapel Hill, NC 27599
[2] Dept. of Computer Science and Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China
[3] Imaging and Computer Vision Dept., Siemens Corporate Research, Princeton, NJ 08540
dqzhang@nuaa.edu.cn, junliu.nt@gmail.com, dgshen@med.unc.edu

Abstract. Sparse learning has recently received increasing attentions in neuroimaging research such as brain disease diagnosis and progression. Most existing studies focus on cross-sectional analysis, i.e., learning a sparse model based on single time-point of data. However, in some brain imaging applications, multiple time-points of data are often available, thus longitudinal analysis can be performed to better uncover the underlying disease progression patterns. In this paper, we propose a novel temporally-constrained group sparse learning method aiming for longitudinal analysis with multiple time-points of data. Specifically, for each time-point, we train a sparse linear regression model by using the imaging data and the corresponding responses, and further use the *group regularization* to group the weights corresponding to the same brain region across different time-points together. Moreover, to reflect the smooth changes between adjacent time-points of data, we also include two *smoothness regularization* terms into the objective function, i.e., one *fused smoothness* term which requires the differences between two successive weight vectors from adjacent time-points should be small, and another *output smoothness* term which requires the differences between outputs of two successive models from adjacent time-points should also be small. We develop an efficient algorithm to solve the new objective function with both group-sparsity and smoothness regularizations. We validate our method through estimation of clinical cognitive scores using imaging data at multiple time-points which are available in the Alzheimer's Disease Neuroimaging Initiative (ADNI) database.

1 Introduction

Neuroimaging plays an important role in characterizing the neurodegenerative process of many brain diseases such as Alzheimer's disease (AD). At present, a lot of pattern classification and regression methods have been developed for brain disease diagnosis and progression. Recently, sparse learning techniques have attracted more and more attentions due to their excellent performances in a series of neuroimaging applications on different modalities. For example, in a recent study [1], a voxel-based sparse classifier using L_1-norm regularized linear regression model, also known as the least absolute shrinkage and selection operator (LASSO) [2], was applied for classification

N. Ayache et al. (Eds.): MICCAI 2012, Part III, LNCS 7512, pp. 264–271, 2012.

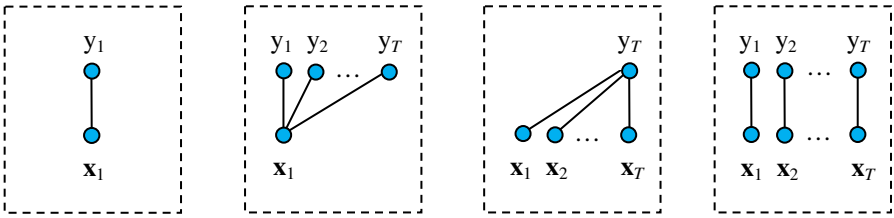
of AD and mild cognitive impairment (MCI) using magnetic resonance imaging (MRI) data, showing better performance than support vector machine (SVM) which is one of the state-of-the-art methods in brain imaging classification.

Following LASSO, several other advanced sparse learning models (i.e., LASSO variants) have also been recently used for solving problems in neuroimaging applications. For example, in [3], the elastic net which extends LASSO by imposing extra L_2-norm based regularizer to encourage a grouping effect, was recently used for identifying both neuroimaging and proteomic biomarkers for AD and MCI using MRI and proteomic data. In [4], a generalized sparse regularization with domain-specific knowledge was proposed for functional MRI (fMRI) based brain decoding. More recently, group LASSO [5], based on $L_{2,1}$-norm regularization, was used for jointly learning multiple tasks including both classification tasks (e.g., AD/MCI vs. healthy controls) and regression tasks (e.g., estimation of clinical cognitive scores) using MRI data in [6] and multimodal data including MRI, fluorodeoxyglucose positron emission tomography (FDG-PET) and cerebrospinal fluid (CSF) in [7], respectively. Here, the assumption of both methods is that multiple regression/classification variables are inherently related and essentially determined by the same underlying pathology, i.e., the diseased brain regions, and thus they can be solved together.

One commonplace of all above mentioned methods (i.e., LASSO and its variants) is that they aimed for cross-sectional analysis. In other words, only single-time-point imaging data (input) and single-time-point responses (output) are used for learning models in those methods. However, in some practical brain imaging applications, multiple-time-point data and/or multi-time-point responses are often available, thus longitudinal analysis can be performed to better uncover the underlying disease progression patterns [8]. According to the number of time-points in input and output of learning models, we can categorize them into the following four different learning problems: 1) Single-time-point Input and Single-time-point Output (SISO), 2) Single-time-point Input and Multi-time-points Output (SIMO), 3) Multi-time-points Input and Single-time-point Output (MISO), and 4) Multi-time-points Input and Multi-time-points Output (MIMO). Fig. 1 gives an illustration for these four different learning problems, with more detailed explanations given later in Section 2. To the best of our knowledge, most existing sparse models are aimed for the SISO problem (Fig. 1(a)), and it remains unknown in the literature on how to effectively use the longitudinal information in sparse learning to solve the other three problems (Fig. 1(b)-(d)).

In this paper, we address the above problems, i.e., SIMO, MISO and MIMO, which involves longitudinal information in either output or input or both. For this purpose, we develop a novel temporally-constrained group LASSO method, named as tgLASSO, which simultaneously includes the *group regularization* and the temporally *smoothness regularization* into its objective function. On one hand, as in group LASSO (gLASSO), for each time-point we train a sparse linear regression model by using the corresponding imaging data and responses at that time-point, and further use the *group regularization* to group the weights corresponding to the same brain region across different time points together. On the other hand, to reflect the smooth changes between adjacent time-points of data, we also introduce two smoothness regularization terms: 1) *fused smoothness* term which originates from fused LASSO [9] , for constraining the differences between two successive weight vectors from adjacent time-points to be small; 2) *output smoothness* term, for

constraining the differences between outputs of two successive models from adjacent time-points to be small. To the best of our knowledge, no previous sparse models ever use both the group-sparsity and the (fused plus output) smoothness regularizations into the objective function, for which we further develop a new efficient algorithm. We will use our proposed method for estimating clinical cognitive scores, e.g., Mini Mental State Examination (MMSE) and Alzheimer's Disease Assessment Scale - Cognitive Subscale (ADAS-Cog), by using MRI data from different time-points.



(a) SISO (b) SIMO (c) MISO (d) MIMO

Fig. 1. Illustration on four different learning problems. Here, each edge represents a model, and the nodes $\mathbf{x}_j$ and y_j denote the imaging data (input) and clinical scores (output) at j-th time-point, respectively.

2 Method

In this section, we will introduce our temporally-constrained group LASSO (tgLASSO) method for longitudinal data analysis. We will first give our motivation and problem formulation in Section 2.1, followed by providing the objective function in Section 2.2 and the algorithmic solution in Section 2.3.

2.1 Motivation and Problem Formulation

Because of the neurodegenerative property of many brain diseases, e.g., AD and MCI, patients usually undergo a series of temporal changes reflected in MRI data and clinical scores (e.g., MMSE and ADAS-Cog for AD). Here, we want to estimate the clinical scores using MRI data. There are four different learning problems according to different number of time-points in both MRI data (input) and clinical scores (output), as shown in Fig. 1.

In the first learning problem, i.e., SISO as shown in Fig. 1(a), we want to estimate the clinical scores at a certain time-point, e.g., time-point 1 (baseline), by using imaging data from single time-point (e.g., baseline). Because both input and output are from single time-point, no longitudinal information is involved in this problem, and it can be easily solved by the existing sparse linear models, e.g., LASSO.

In the second learning problem, i.e., SIMO as shown in Fig. 1(b), we want to estimate the clinical scores at each time-point (ranging from 1 to T), by using imaging data from single time-point 1 (baseline). Similarly, in the third learning problem, i.e., MISO as shown in Fig. 1(c), we want to estimate the clinical scores at time-point T,

by using imaging data from all time-points (from 1 to T). Finally, in the fourth learning problem, i.e., MIMO as shown in Fig. 1(d), we want to estimate the clinical scores at each time-point j, by using imaging data from its corresponding time-point j, for j =1, …, T.

Unlike the first learning problem (SISO), the last three learning problems all involve longitudinal information, and thus cannot be directly solved using the existing sparse models. Also, it is worth noting that SIMO can be gotten from MIMO if setting $\mathbf{x}_j = \mathbf{x}_1$ (for j =1, …, T), and similarly MISO can be gotten from MIMO if setting $y_j = y_T$ (for j =1, …, T). For this reason, in this section we focus on MIMO and will further develop a new efficient algorithm to solve this new problem as below.

2.2 Objective Function

Assume that we have N training subjects, and each subject has T imaging data at T different time-points, represented as $\{\mathbf{x}_{i1}, \ldots, \mathbf{x}_{ij}, \ldots, \mathbf{x}_{iT}\}$, where $\mathbf{x}_{ij} \in \Re^{1\times D}$ is a D-dimensional row vector. Denote $\mathbf{X}_j = [\mathbf{x}_{1j}; \ldots; \mathbf{x}_{ij}; \ldots; \mathbf{x}_{Nj}]$ ($\in \Re^{N\times D}$) and $\mathbf{y}_j$ ($\in \Re^{N\times 1}$) as the training data matrix (input) and the corresponding clinical scores at the j-th time-point, respectively. We use the linear model to estimate the clinical score from the imaging data $\mathbf{x}$ at the j-th time-point as $h_j(\mathbf{x}) = \mathbf{x}\mathbf{w}_j$, where the feature weight vector $\mathbf{w}_j \in \Re^D$. Let $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_j, \ldots, \mathbf{w}_T]$ ($\in \Re^{D\times T}$), then the objective function of our temporally-constrained group LASSO (tgLASSO) can be defined as follows

$$\min_{\mathbf{W}} \quad J(\mathbf{W}) = \frac{1}{2}\sum_{j=1}^{T}\left\|\mathbf{y}_j - \mathbf{X}_j\mathbf{w}_j\right\|_2^2 + R_g(\mathbf{W}) + R_s(\mathbf{W}) \tag{1}$$

Where $R_g(\mathbf{W})$ and $R_s(\mathbf{W})$ are the *group regularization* term and the *smoothness regularization* term, respectively, which are defined as below

$$R_g(\mathbf{W}) = \lambda_1\|\mathbf{W}\|_{2,1} = \lambda_1\sum_{d=1}^{D}\|\mathbf{w}^d\|_2 \tag{2}$$

and

$$R_s(\mathbf{W}) = \lambda_2\sum_{j=1}^{T-1}\left\|\mathbf{w}_j - \mathbf{w}_{j+1}\right\|_1 + \lambda_3\sum_{j=1}^{T-1}\left\|\mathbf{X}_j\mathbf{w}_j - \mathbf{X}_{j+1}\mathbf{w}_{j+1}\right\|_2^2 \tag{3}$$

In Eq. 2, $\mathbf{w}^d$ is the d-th row vector of $\mathbf{W}$. It is worth noting that the use of L_2-norm on row vectors forces the weights corresponding to the d-th feature across multiple time-points to be grouped together and the further use of L_1-norm tends to select features based on the strength of T time-points jointly. The regularization parameter λ_1 controls the group sparsity of the linear models.

On the other hand, as shown in Eq. 3, the smoothness regularization consists of two parts. The first one as defined in the first term in Eq. 3 is called as the *fused smoothness* term which originates from fused LASSO [9], and its function is to constrain the differences between two successive weight vectors from adjacent time-points to be small. Also, it is worth noting that, due to the use of L_1-norm in the fused

smoothness term which encourages the sparsity on differences of weight vectors, there will be a lot of zeros in the components of the weigh difference vectors. In other words, a lot of components from adjacent weight vectors will be identical because of using the fused smoothness regularization. The second term in Eq. 3 is called as the *output smoothness* term which constrains the differences between outputs of two successive models from adjacent time-points to be small as well. The regularization parameters λ_2 and λ_3 balance the relative contributions of the two terms and also control the smoothness of the linear models. It is easy to know that when both λ_2 and λ_3 are zero, our method will reduce to group LASSO.

To the best of our knowledge, the objective function in Eq. 1 is the first time to simultaneously include both the group and the fused regularizations, which cannot be solved by the existing sparse models. Also, no previous studies consider using the output smoothness as extra regularizer. In the next section, we will develop a new efficient algorithm to solve the objective function in Eq. 1.

2.3 Efficient Iterative Solution

To minimize Eq. 1, we propose to use the iterative projected gradient descent approach [10]. Specifically, we separate the objective function in Eq. 1 to the smooth term

$$s(\mathbf{W}) = \frac{1}{2}\sum_{j=1}^{T}\left\|\mathbf{y}_j - \mathbf{X}_j\mathbf{w}_j\right\|_2^2 + \lambda_3\sum_{j=1}^{T-1}\left\|\mathbf{X}_j\mathbf{w}_j - \mathbf{X}_{j+1}\mathbf{w}_{j+1}\right\|_2^2 \tag{4}$$

and the non-smooth term

$$n(\mathbf{W}) = \lambda_1\|\mathbf{W}\|_{2,1} + \lambda_2\sum_{j=1}^{T-1}\left\|\mathbf{w}_j - \mathbf{w}_{j+1}\right\|_1 \tag{5}$$

In each iteration k, the projected gradient descent contains two steps. Firstly, from $\mathbf{W}^{(k)}$, we compute

$$\mathbf{V}^{(k)} = \mathbf{W}^{(k)} - \gamma_k \mathrm{s}'(\mathbf{W}^{(k)}) \tag{6}$$

where $\mathrm{s}'(\mathbf{W}^{(k)})$ denotes the gradient of $s(\mathbf{W})$ at $\mathbf{W}^{(k)}$, and γ_k is the step size that can be determined by line search. Secondly, we set

$$\mathbf{W}^{(k+1)} = arg\min\frac{1}{2}\left|\left|\mathbf{W} - \mathbf{V}^{(k)}\right|\right|_2^2 + n(\mathbf{W}) \tag{7}$$

The problem in Eq. 7 is the proximal operator associated with the non-smooth term $n(\mathbf{W})$, and it can be computed by sequentially solving the proximal operator associated with the group Lasso penalty [5] and the proximal operator associated with the fuse Lasso penalty [9].

By utilizing the technique discussed in [10], the above projected gradient descent can be further accelerated to yield the accelerated gradient descent approach. Specifically, instead of performing gradient descent based on $\mathbf{W}^{(k)}$, we compute the search point

$$\mathbf{S}^{(k)} = \mathbf{W}^{(k)} + \alpha_k\left(\mathbf{W}^{(k)} - \mathbf{W}^{(k-1)}\right) \tag{8}$$

where α_k is a pre-defined variable [10], Then, we set

$$\mathbf{V}^{(k)} = \mathbf{S}^{(k)} - \gamma_i \mathrm{s}'(\mathbf{S}^{(k)}) \tag{9}$$

Finally, we compute the new approximate solution as in Eq. 7. It can be shown that such a scheme can achieve a convergence rate of $O(1/l^2)$ for l iterations. For more details, please refer to [10].

3 Results

In this section, we validate our proposed tgLASSO method, with comparison to the existing LASSO and gLASSO methods, using 445 subjects (including 91 AD, 202 MCI, and 152 healthy controls) from the ADNI database. For each subject, there are MRI data as well as clinical scores including MMSE and ADAS-Cog, for the four different time-points, i.e., baseline, 6 months, 12 months, and 24 months which are denoted as T1, T2, T3 and T4, respectively. Our goal is to estimate the MMSE and ADAS-Cog scores at each of the four time-points using MRI data from corresponding time-point, which is a MIMO problem as shown in Fig. 1. It is worth noting that both SIMO and MISO problems can also be solved by our method as mentioned before. However, due to space limit, we do not report those results in this paper.

Standard image pre-processing is performed for all MRI images, including anterior commissure (AC) - posterior commissure (PC) correction, skull-stripping, removal of cerebellum, and segmentation of structural MR images into three different tissues: grey matter (GM), white matter (WM), and cerebrospinal fluid (CSF). Then, an atlas warping method [11] is used to register all different time-point images of each subject to a template with 93 manually labeled regions of interests (ROIs). For each of the 93 ROIs, we compute the GM tissue volume from the subject's MRI image as features.

In our experiments, 10-fold cross-validation is adopted to evaluate the performances of LASSO, gLASSO, and tgLASSO, by measuring the correlation coefficient between the actual clinical score and the estimated one. For all methods, the values of the parameters are determined by performing another cross-validation on the training data.

Fig. 2 shows the feature weight maps gotten from three different methods. Here, gLASSO and tgLASSO jointly learn the weight vectors for the four time-points, while LASSO learns each weight vector independently for each time-point. As can be seen from Fig. 2, due to the use of group regularization, gLASSO and tgLASSO obtain more grouped weights across different time-points than LASSO. Furthermore, due to the use of smoothness regularization, tgLASSO achieves more smooth weights across different time-points than other two methods. These properties are helpful to discover those intrinsic biomarkers relevant to brain diseases. For example, as shown in Fig. 2, among other disease related brain regions, both left and right hippocampal regions which are well-known AD-relevant biomarkers, are detected by tgLASSO, while only the left one can be detected by the other two methods.

On the other hand, Fig. 3 gives the comparisons of regression performances of the three methods in estimating MMSE and ADAS-Cog scores at four different time-points. As can be seen from Fig. 3, tgLASSO consistently outperforms the other two methods in estimating clinical scores for multiple time-points. In average, tgLASSO

achieves correlation coefficients of 0.613 and 0.639 for estimating MMSE and ADAS-Cog scores across all four time-points, respectively, while LASSO and gLASSO respectively achieve correlation coefficients of 0.569 and 0.587 for MMSE and 0.591 and 0.605 for ADAS-Cog. Fig. 3 also indicates that estimating later time-point scores often achieves better performance than estimating previous time-point scores. This may be because the relationship between imaging features and clinical scores becomes much stronger with progress of disease or brain aging, i.e., atrophy in the brain is more obvious in advanced disease and thus the related features are more distinctive and correlated to the clinical scores.

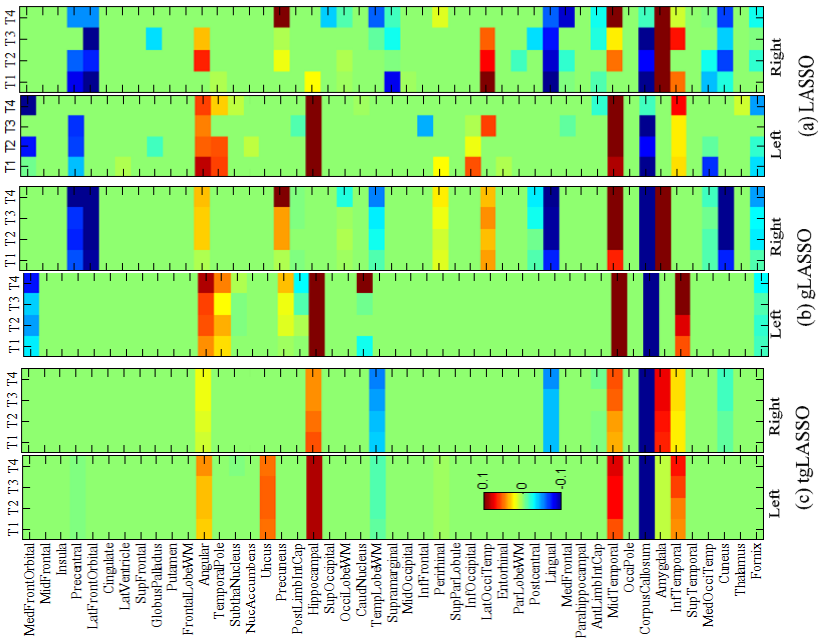


Fig. 2. Comparison of the feature weight maps of three different methods: (a) LASSO, (b) gLASSO, and (c) tgLASSO

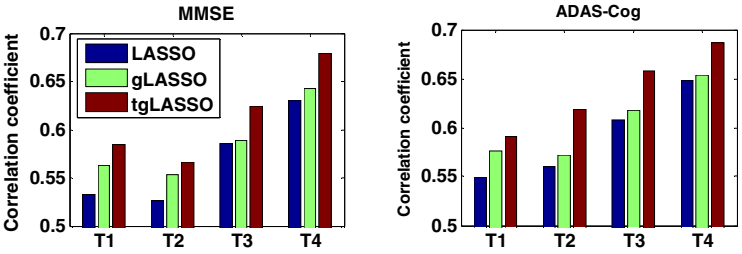


Fig. 3. Comparisons of regression performances of three different methods in estimating MMSE (left) and ADAS-Cog (right) scores

4 Conclusions

We have presented a new sparse learning method called tgLASSO for longitudinal data analysis with multiple time-points of data, which is different from most existing sparse learning methods focusing on cross-sectional analysis with single time-point of data. Our methodological contributions include: 1) proposing to simultaneously use group and (fused plus output) smoothness regularizations in sparse learning; 2) developing an efficient iterative algorithm for solving the new objective function. Experimental results on estimating clinical scores from imaging data at multiple time-points show the advantages of our method over the existing sparse methods on both regression performance and ability in discovering disease related imaging biomarkers.

Acknowledgments. This work was supported in part by NIH grants EB006733, EB008374, EB009634, MH088520 and AG041721 also by NSFC grants (No. 60875030 and 60905035).

References

1. Liu, M., Zhang, D., Shen, D.: Ensemble sparse classification of Alzheimer's disease. NeuroImage 60, 1106–1116 (2012)
2. Tibshirani, R.: Regression shrinkage and selection via the Lasso. J. Roy. Stat. Soc. B Met. 58, 267–288 (1996)
3. Shen, L., Kim, S., Qi, Y., Inlow, M., Swaminathan, S., Nho, K., Wan, J., Risacher, S.L., Shaw, L.M., Trojanowski, J.Q., Weiner, M.W., Saykin, A.J.: Identifying Neuroimaging and Proteomic Biomarkers for MCI and AD via the Elastic Net. In: Liu, T., Shen, D., Ibanez, L., Tao, X. (eds.) MBIA 2011. LNCS, vol. 7012, pp. 27–34. Springer, Heidelberg (2011)
4. Ng, B., Abugharbieh, R.: Generalized sparse regularization with application to fMRI brain decoding. Inf. Process Med. Imaging 22, 612–623 (2011)
5. Yuan, M., Lin, Y.: Model selection and estimation in regression with grouped variables. J. Roy Stat. Soc. B 68, 49–67 (2006)
6. Wang, H., Nie, F., Huang, H., Risacher, S., Saykin, A.J., Shen, L.: Identifying AD-Sensitive and Cognition-Relevant Imaging Biomarkers via Joint Classification and Regression. In: Fichtinger, G., Martel, A., Peters, T. (eds.) MICCAI 2011, Part III. LNCS, vol. 6893, pp. 115–123. Springer, Heidelberg (2011)
7. Zhang, D., Shen, D.: Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease. NeuroImage 59, 895–907 (2012)
8. Xu, S., Styner, M., Gilmore, J., Piven, J., Gerig, G.: Multivariate nonlinear mixed model to analyze longitudinal image data: MRI study of early brain development. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pp. 1–8 (2008)
9. Liu, J., Yuan, L., Ye, J.: An efficient algorithm for a class of fused lasso problems. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 323–332. ACM, Washington, DC (2010)
10. Beck, A., Teboulle, M.: A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems. SIAM J. Img. Sci. 2, 183–202 (2009)
11. Shen, D., Resnick, S.M., Davatzikos, C.: 4D HAMMER Image Registration Method for Longitudinal Study of Brain Changes. In: Proceedings of the Human Brain Mapping, New York City, USA (2003)