Group guided sparse group lasso multi-task learning for cognitive performance prediction of Alzheimer's disease

Xiaoli Liu[1]*, Peng Cao**[1], Jinzhu yang[1], Dazhe Zhao[2], and Osmar Zaiane[3]

[1] College of Computer Science and Engineering, Northeastern University, China
[2] Key Laboratory of Medical Image Computing of Ministry of Education, Northeastern University, China
[3] Department of Computing Science, University of Alberta, Canada

Abstract. Alzheimers disease (AD), the most common form of dementia, causes progressive impairment of cognitive functions of patients. There is thus an urgent need to (1) accurately predict the cognitive performance of the disease, and (2) identify potential MRI (Magnetic Resonance Imaging)-related biomarkers most predictive of the estimation of cognitive outcomes. The main objective of this work is to build a multi-task learning based on MRI in the presence of structure in the features. In this paper, we simultaneously exploit the interrelated structures within the MRI features and among the tasks and present a novel Group guided Sparse group lasso (GSGL) regularized multi-task learning approach, to effectively incorporate both the relatedness among multiple cognitive score prediction tasks and useful inherent group structure in features. An Alternating Direction Method of Multipliers (ADMM) based optimization is developed to efficiently solve the non-smooth formulation. We demonstrate the performance of the proposed method using the Alzheimer's Disease Neuroimaging Initiative (ADNI) datasets and show that our proposed methods achieve not only clearly improved prediction performance for cognitive measurements, but also finds a compact set of highly suggestive biomarkers relevant to AD.

1 Introduction

Alzheimer's disease (AD) is a gradually progressive syndrome that mainly affects memory function, ultimately culminating in a dementia state. It has been proved that brain atrophy detected by MRI is correlated with neuropsychological deficits. Many clinical/cognitive measures have been designed to evaluate the cognitive status of the patients and used as important criteria for clinical diagnosis of probable AD. Many cognitive measures including Mini Mental S-tate Examination (MMSE) and Alzheimers Disease Assessment Scale cognitive subscale (ADAS-Cog) have been designed to evaluate the cognitive status of the

* Supported by the National Natural Science Foundation of China (No.61502091), the Fundamental Research Funds for the Central Universities (No.N161604001 and N150408001).

** Corresponding author.

patients and used as important criteria for clinical diagnosis of probable AD. It is known that there exist inherent correlations among multiple clinical variables of a subject, and a joint analysis of data from multiple cognitive tasks is expected to improve the performance[11, 8, 5]. The assumption of the commonly used Multi-task learning (MTL) is that all tasks share the same data representation with $\ell_{2,1}$ regularization, since a given imaging marker can affect multiple cognitive scores and only a subset of the imaging features (brain region) are relevant. This assumption of $\ell_{2,1}$ regularization is restrictive since it encourages all the tasks to share the same data representation. Sparse group Lasso (SGL) [6] allows the simultaneous selection of a common set of biomarkers for all the tasks and the selection of a specific set of biomarkers for different tasks.

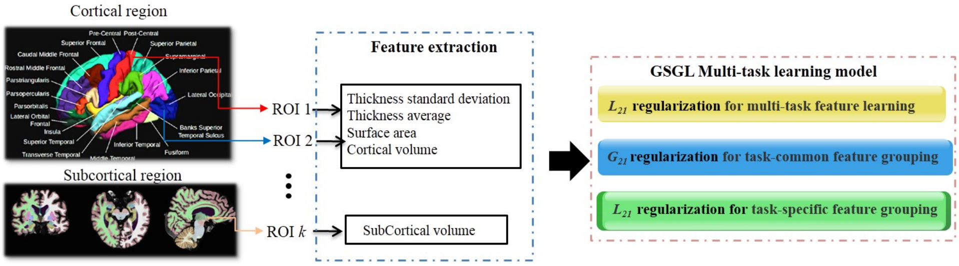

Fig. 1: Flow chart of the proposed GSGL-MTL method.

Many previous works extract only the volume or thickness measures of cortical regions of interest (ROIs) as the features[14, 5]. To avoid manual measure bias caused by the single feature in this study, multiple features are extracted to measure the atrophy information of each ROI involving cortical thickness, surface area and volume from gray matter and white matter. The multiple shape measures from the same region provide a comprehensively quantitative evaluation of cortical atrophy, and tend to be selected together as joint predictors. It is hypothesized that not only a subset of MRI features, but also a subset of ROIs are relevant to each assessment. Therefore, we use this prior knowledge of interrelated structure to group relevant shape features together in the same region to guide the learning process. Based on this intuitive motivation, we simultaneously exploit the interrelated structures within features as well as among the tasks, and present a novel multi-task learning method to effectively incorporate both the relatedness among multiple cognitive score prediction tasks and useful inherent group structure in features. Inspired by the recent success of the group lasso regularization [10] as well as the term bi-level analysis [7], we propose a unified bi-level learning framework to jointly perform both individual feature-level and ROI-level analysis by group lasso regularization with the grouping effect such that it helps reduce the variances in the estimation of coefficient and improves the stability of biomarkers selection. Specially, we develop a novel multi-task learning formulation based on a group guided SGL.

The regularizer consists of three components including an $\ell_{2,1}$ penalty, which ensures that a small subset of features will be selected for the regression models, and a $G_{2,1}$ penalty, which encourages the task-common ROI across multi-task. To relax the restrictive assumption of shared ROI imposed in the correlation among the cognitive tasks, a task-specific ROI based $\ell_{2,1}$-norm for each task is incorporated. The proposed formulation is challenging to solve due to the use of multiple non-smooth penalties. We present an Alternating Direction Method of Multipliers (ADMM)-type algorithm for solving the proposed non-smooth optimization problems efficiently. We conducted extensive experiments using data from the ADNI dataset to demonstrate our methods along various dimensions including prediction performance and biomarkers identification.

2 Proposed Method

2.1 Group guided sparse group lasso multi-task learning

The high feature-dimension problem is one of the major challenges in the study of computer aided Alzheimer's Disease (AD) diagnosis. Variable selection is of great importance to improve the prediction performance and model interpretation for high-dimensional data. Lasso is a widely used technique for high-dimensional association mapping problems, which can yield a sparse and easily interpretable solution via an ℓ_1 regularization. However, despite the success of Lasso, it is limited to considering each task separately and ignores the inherent structure of features. However, Lasso fails to capture the correlation information among the pairwise of group features. The pairwise correlations among group of features are very high, Lasso tends to select only one of the pairwise correlated features, resulting in ignoring the group effect.

Group regularizers like group lasso [10] via an $\ell_{2,1}$ regularization assumes covarying variables in groups, and have been extensively studied in the multi-task feature learning. The difference of Lasso and group lasso is illustrated in Figure 2. The key assumption behind the group lasso regularizer is that if a few features in a group are important, then most of the features in the same group should also be important.

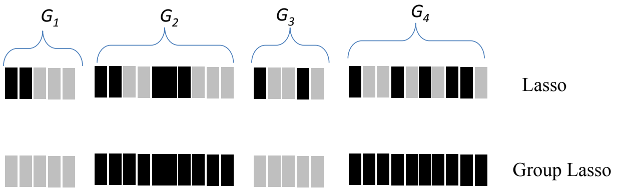

Fig. 2: The difference between Lasso and Group lasso

Multi-Task Learning (MTL) is a statistical learning framework which seeks at learning several models in a joint manner. It has been commonly used to obtain better generalization performance than learning each task individually [1, 4]. The critical issues in MTL is to identify how the tasks are related and build learning models to capture such task relatedness. Consider a multi-task learning (MTL) setting with k tasks. Let $X \in \mathbb{R}^{n \times p}$ denote the matrix of covariates, $Y \in \mathbb{R}^{n \times k}$ be the matrix of responses with each row corresponding to a sample, and $\Theta \in \mathbb{R}^{p \times k}$ denote the parameter matrix, with column $\theta_{.h} \in \mathbb{R}^p$ corresponding to task h, $h = 1, \dots, k$, and row $\theta_{i.} \in \mathbb{R}^k$ corresponding to feature i, $i = 1, \dots, p$. The MRI measure features in the same brain region belong to a group. We assume the p features to be divided into q disjoint groups $\mathcal{G}_l, l = 1, \dots, q$, with each group having m_l features respectively. The MTL problem can be set-up as one of estimating the parameters based on suitable regularized loss function:

$$\min_{\Theta} \quad L(Y, X, \Theta) + \lambda R(\Theta) \ , \tag{1}$$

where $L(\cdot)$ denotes the loss function and $R(\cdot)$ is the regularizer. In the current context, we assume the loss to be square loss, i.e.,

$$L(Y, X, \Theta) = \|Y - X\Theta\|_F^2 = \sum_{i=1}^{n} \|\mathbf{y}_i - \mathbf{x}_i\Theta\|_2^2 \ , \tag{2}$$

where $\mathbf{y}_i \in \mathbb{R}^{1 \times k}, \mathbf{x}_i \in \mathbb{R}^{1 \times p}$ are the i-th rows of Y, X, respectively corresponding to the multi-task response and covariates for the i-th sample. We note that the MTL framework can be easily extended to other loss functions. Base on some prior knowledge, we then add penalty $R(\Theta)$ to encode the relatedness among tasks.

Group Lasso regularized multi-task learning (GL-MTL) aims to obtain better generalization performance by exploiting the shared features among different tasks [4]. In our case, given that one imaging marker can affect multiple cognitive scores, the coefficients of the coefficient matrix of the same row is largely correlated. It has been successfully applied to capture biomarkers having affects across most or all responses in the application of AD prediction [3, 13]. The GL-MTL model via the $\ell_{2,1}$-norm regularization considers

$$R(\Theta) = \|\Theta\|_{2,1} = \sum_{i=1}^{p} \|\theta_{i.}\|_2 \ , \tag{3}$$

and is suitable for simultaneously enforcing sparsity over features for all tasks.

The key point of Eq. (3) is the use of ℓ_2-norm for $\theta_{i.}$, which forces the weights corresponding to the i-th feature across multiple tasks to be grouped together and tends to select features based on the strength of k tasks jointly. There is a correlation in multiple cognitive measures, and the associated imaging predictors usually have more or less effect on all of these scores, which leads to a correlation between regression coefficients. By employing GL-MTL, the correlation information among different tasks can be incorporated into the model to build a more appropriate predictive model and identify a subset of the features.

One appealing property of the group lasso regularization in GL-MTL is that it encourages multiple predictors from related tasks to share a subset of features. However, the $\ell_{2,1}$-norm regularization only consider the shared representation from the features, neglecting the potentially grouping information among multiple neuroimaging measures. In order to address it, we consider prior information group information in features and multi-task learning simultaneously in one single framework. Specifically, We propose a Group guided Sparse Group Lasso regularized multi-task learning (GSGL-MTL) algorithm exploiting both the group structure of features and the multi-task correlation, to unify feature-level and ROI-level analysis in an unified multi-task learning framework. The GSGL-MTL formulation focuses on the following regularized loss function:

$$\min_{\Theta \in \mathbb{R}^{p \times k}} \frac{1}{2}\|Y - X\Theta\|_F^2 + \lambda_1\|\Theta\|_{2,1} + \lambda_2\|\Theta\|_{G_{2,1}} + \lambda_3\|\mathrm{vec}(\Theta)\|_{2,1} \ . \tag{4}$$

where $\|\Theta\|_{G_{2,1}} = \sum_{l=1}^{q} w_l \sqrt{\sum_{j \in \mathcal{G}_l} \|\theta_{j.}\|_2}$, $\|\mathrm{vec}(\Theta)\|_{2,1} = \sum_{h=1}^{k} \sum_{l=1}^{q} w_l \|\theta_{\mathcal{G}_l h}\|_2$, and $w_l = \sqrt{m_l}$ is the weight for each group. The second and third norms are called Group guided Sparse Group Lasso norm (GSGL), where $\|\Theta\|_{G_{2,1}}$ encourages the task-common ROIs to induce the same group sparsity patterns across different tasks (coupling all tasks) and $\|\mathrm{vec}(\Theta)\|_{2,1}$ encourages the task-specific ROIs to induce the different group sparsity patterns across different tasks (decoupled for each task), as illustrated in Figure 3.

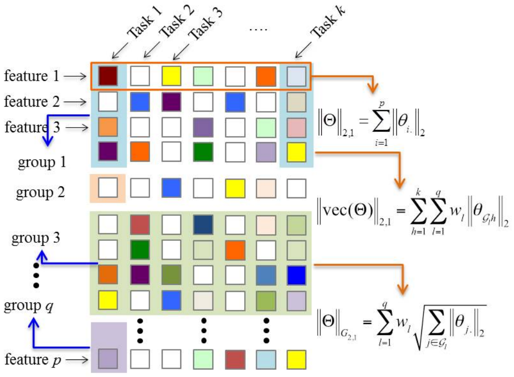
Task 1
Task 2
Task 3
....
Task k
feature 1
feature 2
feature 3
group 1
group 2
group 3
group q
feature p
$\|\Theta\|_{2,1} = \sum_{i=1}^{p}\|\theta_{i.}\|_2$
$\|\mathrm{vec}(\Theta)\|_{2,1} = \sum_{h=1}^{k}\sum_{l=1}^{q} w_l\|\theta_{\mathcal{G}_l h}\|_2$
$\|\Theta\|_{G_{2,1}} = \sum_{l=1}^{q} w_l \sqrt{\sum_{j\in\mathcal{G}_l}\|\theta_{j.}\|_2}$

Fig. 3: The illustration of the GSGL-MTL method

2.2 Optimization

In this section, we present a novel solver for the problem in Equation (1) based on the ADMM. The proposed formulation is, however, challenging to solve due to the use of three non-smooth penalties. It is easy to show that the objective function of the GSGL-MTL method is convex. To efficiently handle the two non-smooth constraints, we propose an optimization method which employs ADMM algorithm [2] to solve the proposed multi-task learning problem by decomposing a large global problem into a series of smaller local subproblems and coordinates the local solutions to identify the globally optimal solution[2].

Assume $R^{\lambda_1}_{\lambda_2,\lambda_3}(\Theta) = \lambda_1\|\Theta\|_{2,1} + \lambda_2\|\Theta\|_{G_{2,1}} + \lambda_3\|\mathrm{vec}(\Theta)\|_{2,1}$, then Eq. (4) is equivalent to the following constrained optimization problem:

$$\min_{\Theta\in\mathbb{R}^{p\times k}} \frac{1}{2}\|Y - X\Theta\|_F^2 + R^{\lambda_1}_{\lambda_2,\lambda_3}(Q) \quad \text{subject to } \Theta - Q = 0\ . \tag{5}$$

where Q is slack variables. Then Eq. (5) can be solved by ADMM. The augmented Lagrangian is $L_\rho(\Theta, Q, U) = \frac{1}{2}\|Y - X\Theta\|_F^2 + R^{\lambda_1}_{\lambda_2,\lambda_3}(Q) + \mathrm{Tr}(U^T(\Theta - Q)) + \frac{\rho}{2}\|\Theta - Q\|^2$, where U is augmented Lagrangian multiplier.

Update Θ^{t+1}: In the $(t+1)$-th iteration, Θ^{t+1} can be updated by minimizing L_ρ with Q ,U fixed: $\Theta^{t+1} = \underset{\Theta}{\mathrm{argmin}}\frac{1}{2}\|Y - X\Theta\|_F^2 + \mathrm{Tr}((U^t)^T(\Theta - Q^t)) + \frac{\rho}{2}\|\Theta - Q^t\|^2$. The optimization problem is quadratic. The optimal solution is given by $\Theta^{t+1} = F^{-1}B^t$, where $F = X^TX + \rho I$ and $B^t = X^TY - U^t + \rho Q^t$.

Update Q: The update for Q effectively needs to solve the following problem: $Q^{t+1} = \underset{Q}{\mathrm{argmin}}\ \frac{\rho}{2}\|Q - \Theta^{t+1}\|^2 + R^{\lambda_1}_{\lambda_2,\lambda_3}(Q) - \mathrm{Tr}((U^t)^TQ$, which is equivalent to computing the proximal operator for $R^{\lambda_1}_{\lambda_2,\lambda_3}(\cdot)$. In particular, we need to solve

$$\Psi^{\lambda_1/\rho}_{\lambda_2/\rho,\lambda_3/\rho}(O^{t+1}) = \underset{Q}{\mathrm{argmin}}\ \left\{R^{\lambda_1/\rho}_{\lambda_2/\rho,\lambda_3/\rho}(Q) + \frac{1}{2}\|Q - O^{t+1}\|^2\right\}, \tag{6}$$

where $O^{t+1} = \Theta^{t+1} + \frac{1}{\rho}U^t$.

The goal is to be able to compute $Q^{t+1} = \Psi^{\lambda_1/\rho}_{\lambda_2/\rho,\lambda_3/\rho}(O^{t+1})$ efficiently. It can be shown [9] that the proximal operator for the composite regularizer can be computed efficiently in three steps, and all of these steps can be executed efficiently using suitable extensions of soft-thresholding.

$$\Pi^{t+1} = \Psi^{\lambda_1/\rho}_{0,0}(O^{t+1}) = \underset{\Pi}{\mathrm{argmin}}\left\{\frac{\lambda_1}{\rho}\|\Pi\|_{2,1} + \frac{1}{2}\|\Pi - O^{t+1}\|\right\} \tag{7a}$$

$$\begin{aligned}\Gamma^{t+1} &= \Psi^{0}_{\lambda_2/\rho,0}(\Pi^{t+1}) = \Psi^{\lambda_1/\rho}_{\lambda_2/\rho,0}(O^{t+1}) \\ &= \underset{\Gamma}{\mathrm{argmin}}\left\{\frac{\lambda_2}{\rho}\|\Gamma\|_{G_{2,1}} + \frac{1}{2}\|\Gamma - \Pi^{t+1}\|\right\}\end{aligned} \tag{7b}$$

$$\begin{aligned}Q^{t+1} &= \Psi^{0}_{0,\lambda_3/\rho}(\Gamma^{t+1}) = \Psi^{0}_{\lambda_2/\rho,\lambda_3/\rho}(\Pi^{t+1}) = \Psi^{\lambda_1}_{\lambda_2/\rho,\lambda_3/\rho}(O^{t+1}) \\ &= \underset{Q}{\mathrm{argmin}}\left\{\frac{\lambda_3}{\rho}\|\mathrm{vec}(Q)\|_{2,1} + \frac{1}{2}\|Q - \Gamma^{t+1}\|\right\}\end{aligned} \tag{7c}$$

The row-wise updates of (7a) -(7c) can be done by soft-thresholding as:

$$\pi_{i.} = \frac{\max\left\{\|o_{i.}\|_2 - \frac{\lambda_1}{\rho}, 0\right\}}{\|o_{i.}\|_2} o_{i.} \ , \tag{8a}$$

$$\gamma_{j.} = \frac{\max\left\{\sqrt{\sum_{j\in\mathcal{G}_l} \|\pi_{j.}\|_2} - \frac{\lambda_2 w_l}{\rho}, 0\right\}}{\sqrt{\sum_{j\in\mathcal{G}_l} \|\pi_{j.}\|_2}} \pi_{j.} \ , \tag{8b}$$

$$q_{\mathcal{G}_l h} = \frac{\max\left\{\|\gamma_{\mathcal{G}_l h}\|_2 - \frac{\lambda_3 w_l}{\rho}, 0\right\}}{\|\gamma_{\mathcal{G}_l h}\|_2} \gamma_{\mathcal{G}_l h} \ , \tag{8c}$$

where $\pi_{i.}$, $o_{i.}$, $\gamma_{j.}$ are the i-th row of Π^{t+1}, O^{t+1}, Γ^{t+1}, $q_{\mathcal{G}_l h}$, $\gamma_{\mathcal{G}_l h}$ are rows in group $\mathcal{G}_l$ for task h of Q^{t+1} and Γ^{t+1}, respectively.

Dual Update for U: Following standard ADMM dual update, the update for the dual variable for our setting is as follows: $U^{t+1} = U^t + \rho(\Theta^{t+1} - Q^{t+1})$.

3 Experimental Results

3.1 Data and Experimental Setting

In this work, only ADNI subjects with no missing features or cognitive scores are included. This yields a total of $n = 816$ subjects, who are categorized into 3 baseline diagnostic groups: Cognitively Normal (CN, $n_1 = 228$), Mild Cognitive Impairment (MCI, $n_2 = 399$), and Alzheimer's Disease (AD, $n_3 = 189$). The dataset has been processed by a team from UCSF (University of California at San Francisco), who performed cortical reconstruction and volumetric segmentations with the FreeSurfer image analysis suite. There were $p = 319$ MRI features in total, including the cortical thickness average (TA), standard deviation of thickness (TS), surface area (SA), cortical volume (CV) and subcortical volume (SV) for a variety of ROIs. In order to sufficiently investigate the comparison, we further evaluate the performance on all the cognitive assessments (e.g. ADAS, MMSE and RAVLT, totally $k = 20$ tasks). To our best knowledge, no previous work uses all the cognitive scores for training and evaluation.

We use 10-fold cross valuation to evaluate our model and conduct the comparison. In each of twenty trials, a 5-fold nested cross validation procedure for all the comparable methods in our experiments is employed to tune the regularization parameters. Data was z-scored before applying regression methods. To have a fair comparison, we validate the regularization parameters of all the methods in the same search space (from 10^{-1} to 10^3) on a subset of the training set, and use the optimal parameters to train the final models. We evaluate all the algorithms in terms of both root mean squared error (rMSE), normalized mean squared error (nMSE) and the weighted R-value (wR) which are commonly used in multi-task learning problem.

3.2 The results of comparing with the comparable methods

In this section, we conduct empirical evaluation for the proposed methods by comparing with three single task learning methods: Ridge and Group Lasso, both of which are applied independently on each task. To verify the effect of individual components in our framework and show the contribution of individual components, we evaluate the three components of our approach: GL-MTL ($\lambda_2 = \lambda_3 = 0$), GSGL-MTL-s ($\lambda_2 = 0$) with promoting task-specific ROI and GSGL-MTL-c ($\lambda_3 = 0$) with promoting task-common ROI. Moreover, to illustrate how well our GSGL-MTL works, we comprehensively compare our proposed methods with several popular state-of-the-art MTL methods: SGL-MTL and Sparse regularized multi-task learning formulation (SRMTL)[12]. The experimental results are shown in Table 1.

Table 1: Performance comparison of various methods on twenty cognitive prediction tasks. The best results are bolded, and superscript symbol $*$ indicate that GSGL-MTL significantly outperformed that method on that score (Student's t-test at a level of 0.05 was used).

	Ridge	Group Lasso	GL-MTL	GSGL-MT-s	GSGL-MT-c	SGL-MTL	RMTL	SRMTL	GSGL-MTL
ADAS	7.445±0.369	6.769±0.395	6.662±0.411	6.650±0.429	**6.632±0.455**	6.653±0.427	7.338±0.548	6.925±0.463	6.647±0.462
MMSE	2.567±0.146	2.212±0.074	2.190±0.106	2.186±0.098	**2.173±0.086**	2.191±0.097	2.811±0.131	2.404±0.316	2.175±0.086
TOTAL	11.16±0.734	9.966±0.877	9.656±0.695	9.644±0.753	**9.595±0.765**	9.646±0.754	10.82±0.814	10.39±0.813	9.606±0.778
TOT6	3.909±0.366	3.361±0.283	3.324±0.259	3.314±0.270	**3.308±0.262**	3.315±0.271	3.586±0.332	4.066±0.878	3.309±0.259
TOTB	1.981±0.124	1.664±0.156	1.670±0.148	1.664±0.150	1.657±0.150	1.663±0.149	1.729±0.138	3.027±1.785	**1.654±0.152**
T30	4.061±0.287	3.468±0.236	3.440±0.232	3.430±0.247	**3.424±0.259**	3.431±0.245	3.742±0.238	4.232±0.940	3.428±0.264
RECOG	4.310±0.429	3.980±0.210	3.626±0.272	3.622±0.247	3.614±0.227	3.626±0.246	3.887±0.363	4.115±0.736	**3.611±0.218**
ANIM	6.307±0.551	5.514±0.698	5.266±0.448	5.261±0.497	5.243±0.491	5.259±0.499	5.762±0.491	5.432±0.494	**5.236±0.494**
VEG	4.276±0.385	3.711±0.178	3.676±0.180	3.672±0.207	**3.661±0.201**	3.676±0.207	3.948±0.306	4.242±0.868	3.666±0.199
A	26.18±3.764	23.19±4.199	23.01±3.492	22.99±3.565	22.88±3.659	22.99±3.568	26.51±3.501	24.06±3.793	**22.87±3.668**
B	80.01±8.102	71.15±6.039	69.88±5.280	69.82±5.177	69.17±4.702	69.82±5.183	85.29±7.558	75.16±8.208	**69.13±4.658**
IMM	4.695±0.365	4.202±0.300	4.144±0.302	4.140±0.326	**4.123±0.324**	4.142±0.327	4.436±0.373	4.895±1.047	4.126±0.323
DEL	5.277±0.508	4.636±0.469	4.589±0.435	4.584±0.456	**4.562±0.464**	4.585±0.455	4.909±0.430	5.197±0.629	4.566±0.461
DRAW	1.152±0.108	0.978±0.107	0.967±0.119	0.960±0.114	0.961±0.116	0.960±0.116	1.001±0.125	2.697±2.062	**0.958±0.116**
COPY	0.774±0.060	0.671±0.078	0.647±0.103	0.642±0.092	0.644±0.092	0.644±0.092	0.698±0.097	3.356±3.077	**0.641±0.091**
BOSNAM	4.563±0.539	4.026±0.385	**3.951±0.477**	3.964±0.423	3.961±0.440	3.965±0.424	4.488±0.354	4.040±0.519	3.953±0.441
ANART	11.23±0.756	9.760±0.895	9.611±0.722	9.585±0.742	9.533±0.716	9.584±0.742	10.73±0.650	10.07±0.852	**9.524±0.714**
FOR	2.570±0.262	1.999±0.154	2.009±0.122	2.001±0.133	2.000±0.127	1.997±0.133	2.145±0.142	3.634±2.038	**1.995±0.132**
BAC	2.559±0.193	2.144±0.195	2.160±0.177	2.141±0.186	2.145±0.180	2.143±0.186	2.227±0.182	3.130±1.287	**2.134±0.185**
DIGIT	12.76±1.273	11.73±1.337	11.21±1.224	11.23±1.261	**11.16±1.230**	11.23±1.260	12.55±1.275	12.14±1.570	11.18±1.209
nMSE	10.35±1.088*	8.079±0.682*	7.762±0.633*	7.740±0.615*	7.641±0.555	7.742±0.617*	10.44±1.069*	11.68±4.227*	**7.636±0.543**
wR	0.292±0.046*	0.392±0.049*	0.404±0.055*	0.409±0.053*	0.414±0.050*	0.409±0.054*	0.327±0.058*	0.395±0.046*	**0.416±0.048**

As can be seen from the Table 1, GSGL-MTL significantly outperformed the single task learning methods (Ridge and Group Lasso), and the recent state-of-the-art algorithms proposed in terms of nMSE and wR, which indicates that the interrelated structures within features and the correlation among the tasks are effectively captured by the GSGL norm.

3.3 Identification of MRI biomarkers

Finally, we examined the biomarkers identified by different methods. The proposed GSGL-MTL is a group guided model which is able to identify a compact set of relevant neuroimaging biomarkers from the region level due to the group lasso on the features, which would provide us with better interpretability of the brain region.

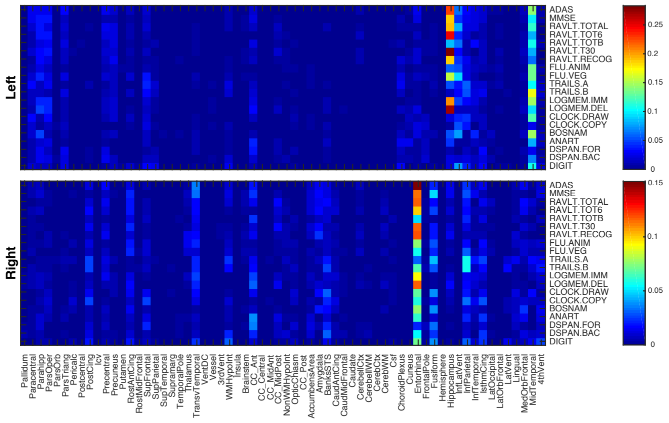

Fig. 4: Baseline matrix sparsity features.

Fig. 4 is the heat maps of the regression weights of all ROIs in each hemisphere for each cognitive score at the baseline time calculated by GSGL-MTL through 10-fold cross validation experiments. Each item (i, j) indicates the weight of the i-th ROI for the j-th task, and is calculated by $w_i\sqrt{\sum_{q\in\mathcal{G}_i}\|\theta_{qi}\|_2}$, where q is the q-th MRI feature in the i-th ROI. The larger the absolute value of a coefficient, the more important its corresponding brain region is in predicting the corresponding cognitive score. The figure illustrates that the proposed GSGL-MTL clearly presented a sparsity across all the cortical measures from the level of ROI, which indicates a small portion of the brain regions is relevant to the cognitive outcome. We found that the imaging biomarkers identified by GSGL-MTL yielded promising patterns that are expected from prior knowledge on neuroimaging and cognition. Some important brain regions are selected, such as R.Middle Temporal, L.Hippocampus and R.Entorhinal, which are highly relevant to the cognitive impairment.

4 Conclusions

In this paper, we propose a Group guided Sparse group lasso (GSGL) regularized multi-task learning to learn the relationship between images and corresponding clinical scores from feature level and ROI level with taking the inherent group structure of the features into account. The experiments on the ADNI dataset have verified the effectiveness of GSGL-MTL, which offers consistently better performance than the baseline single task learning and several state-of-the-art multi-task learning algorithms. These promising results justify that by inducing both sparsity of feature and ROI level, GSGL-MTL captures useful information about AD. In the current work, only apriori group information is incorporated into multi-task predictive model, we are interested in the investigation of other structures in features, such as graph structure, which can help gain additional insights to understand and interpret data in future work.

References

1. T. E. A. Argyriou and M. Pontil. Convex multi-task feature learning. In *Machine Learning*, volume 73, pages 243–272, 2008.
2. S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. *Foundation and Trends in Machine Learning*, pages 1–122, 2011.
3. R. Guerrero, C. Ledig, A. Schmidt-Richberg, D. Rueckert, Alzheimer's Disease Neuroimaging Initiative, et al. Group-constrained manifold learning: Application to AD risk assessment. *Pattern Recognition*, 63:570–582, 2017.
4. J. Liu, S. Ji, and J. Ye. Multi-task feature learning via efficient $\ell_{2,1}$-norm minimization. In *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*, pages 339–348. AUAI Press, 2009.
5. J. Wan, Z. Zhang, J. Yan, T. Li, B. D. Rao, S. Fang, S. Kim, S. L. Risacher, A. J. Saykin, and L. Shen. Sparse bayesian multi-task learning for predicting cognitive outcomes from neuroimaging measures in alzheimer's disease. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 940–947, 2012.
6. J. Wang and J. Ye. Two-layer feature reduction for sparse-group lasso via decomposition of convex sets. In *Advances in Neural Information Processing Systems*, pages 2132–2140, 2014.
7. S. Xiang, L. Yuan, W. Fan, Y. Wang, P. M. Thompson, J. Ye, A. D. N. Initiative, et al. Bi-level multi-source learning for heterogeneous block-wise missing data. *NeuroImage*, 102:192–206, 2014.
8. J. Yan, H. Huang, S. L. Risacher, S. Kim, M. Inlow, J. H. Moore, A. J. Saykin, and L. Shen. Network-guided sparse learning for predicting cognitive outcomes from mri measures. In *International Workshop on Multimodal Brain Image Analysis*, pages 202–210. Springer, 2013.
9. L. Yuan, J. Liu, and J. Ye. Efficient Methods for Overlapping Group Lasso. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(9):21042116, 2013.
10. M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.
11. D. Zhang, D. Shen, A. D. N. Initiative, et al. Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in alzheimer's disease. *Neuroimage*, 59(2):895–907, 2012.
12. J. Zhou. Multi-task learning in crisis event classification. Technical report, Tech. Rep., http://www. public. asu. edu/jzhou29.
13. X. Zhu, H.-I. Suk, S.-W. Lee, and D. Shen. Subspace regularized sparse multitask learning for multiclass neurodegenerative disease identification. *IEEE Transactions on Biomedical Engineering*, 63(3):607–618, 2016.
14. X. Zhu, H.-I. Suk, and D. Shen. Sparse discriminative feature selection for multiclass alzheimers disease classification. In *International Workshop on Machine Learning in Medical Imaging*, pages 157–164. Springer, 2014.