**Proceedings Paper:**
Wang, X., Zhou, M., Zhang, Y. et al. (2024) Empirical analysis of regularized multi-task learning for modelling Alzheimer's disease progression. In: 2023 IEEE International Conference on Bioinformatics & Biomedicine. 2023 IEEE International Conference on Bioinformatics & Biomedicine (BIBM), 05-08 Dec 2023, Istanbul and Turkey. Institute of Electrical and Electronics Engineers (IEEE), pp. 4444-4451. ISBN: 979-8-3503-3748-8. ISSN: 2156-1125. EISSN: 2156-1133.

https://doi.org/10.1109/BIBM58861.2023.10385476

# Empirical Analysis of Regularized Multi-task Learning for Modelling Alzheimer's Disease Progression


Xulong Wang
*Department of Computer Science*
*University of Sheffield*
Sheffield, UK
xl.wang@sheffield.ac.uk

Menghui Zhou
*Department of Computer Science*
*University of Sheffield*
Sheffield, UK
mzhou47@sheffield.ac.uk

Yu Zhang
*Department of Computer Science*
*University of Sheffield*
Sheffield, UK
yzhang489@sheffield.ac.uk

Kang Liu
*Department of Computer Science*
*University of Sheffield*
Sheffield, UK
kang.liu@sheffield.ac.uk

Jun Qi
*Department of Computing*
*Xi'an JiaoTong-Liverpool University*
Suzhou, China
Jun.Qi@xjtlu.edu.cn

Po Yang
*Department of Computer Science*
*University of Sheffield*
Sheffield, UK
po.yang@sheffield.ac.uk



***Abstract*—Recently, there have been a wide spectrum of machine learning models developed to model Alzheimer's disease (AD) progression. Multi-Task Learning (MTL) approaches has been commonly used by these studies to address challenges of missing and insufficient AD data. Typical MTL studies in AD focuses on obtaining high quality of baselines (MRI features and cognitive scores) from AD raw data and exploring advanced regression models for exploring their relationship and correlations. These studies follow a unified regularized MTL framework to process AD datasets with simple evaluation matrix. But another easy-ignorable issue here is whether experimental evaluation strategies are objective and reliable to access MTL performance. There is little attention on studying how to design feasible experimental protocols and evaluation matrix for assessment of regularized MTL models. In this paper, we describe an empirical study and analysis that investigate above question. Four typical structural regularization approaches in MTL study are examined, including (Ridge, Lass, TGL and cFSGL) [1], [2]. Four issues affecting evaluation process of regularised MTL models are evaluated by experiments: 1) evaluation indicators, 2) repeated experimental times; 3) size and portion of training data; 4) number of tasks in MTL. The results demonstrate that regularized MTL models like cFSGL are capable of predicting AD progression with high accuracy, in many challenging cases of data missing, insufficiency or even single MRI data input. One important finding is the performance gain of cFSGL may not only from its ability on dealing with sparsity of AD feature data labels. It is more likely due to existence of a low rank space inside original AD data features. We also discover and proof some limitations of regularized MTL in AD study: the assumption of temporal smoothness in regularized MTL models for AD study limits their performance improvement of the initial task. It is a special relationship that fails to accurately capture certain tasks. Some MTL models like cFSGL have great potential of improvement at late stage prediction of AD progression.**



***Index Terms*—Multi-task learning, Regression model, Alzheimer's disease**


## I. Introduction

Alzheimer's disease (AD), as one of the most common forms of dementia, is a neurodegenerative disease that causes problems with progressive cognitive decline and memory loss. With rates projected to increase by 75% in the next quarter of a century [3]–[5], AD is a leading contributor to disability amongst older people and causes significant morbidity as well as personal family burden. So far, there is no effective cure for AD where science has not yet identified any treatments that can slow or halt the progression of this disease. Yet, timely diagnosis and early intervention in AD can be still useful and cost-effective. It poses an important research area that understands how the AD progresses and identify their related pathological biomarkers for the progression. In order to accelerate AD's research, the Alzheimer's Disease Neuroimaging Initiative (ADNI) funded by NHI provided a large boundary of publicly available neuroimaging data including MRI, PET, other biomarkers and cognitive measures for scientific study. A variety of data driven based machine learning techniques [6]–[8] like deep learning models [6], [9], multi-task modeling [2], [6], [10],and survival model [7], have been investigated to deal with these data for better prediction of AD progression.

In traditional machine learning paradigm, an accurate learner is usually treated as one single learning task (e.g., classification, regression) and learnt by a large number of training samples. For instance, deep learning model can train an accurate AD prediction model of neural network with hundreds of layers contacting a great amount of parameters via massive labelled biomarkers at baseline from ADNI. But one key challenge here is that sufficient and well-labelled longitudinal AD data at multiple time points are hardly collected from AD patients. The problem of missing, sparse and insufficient

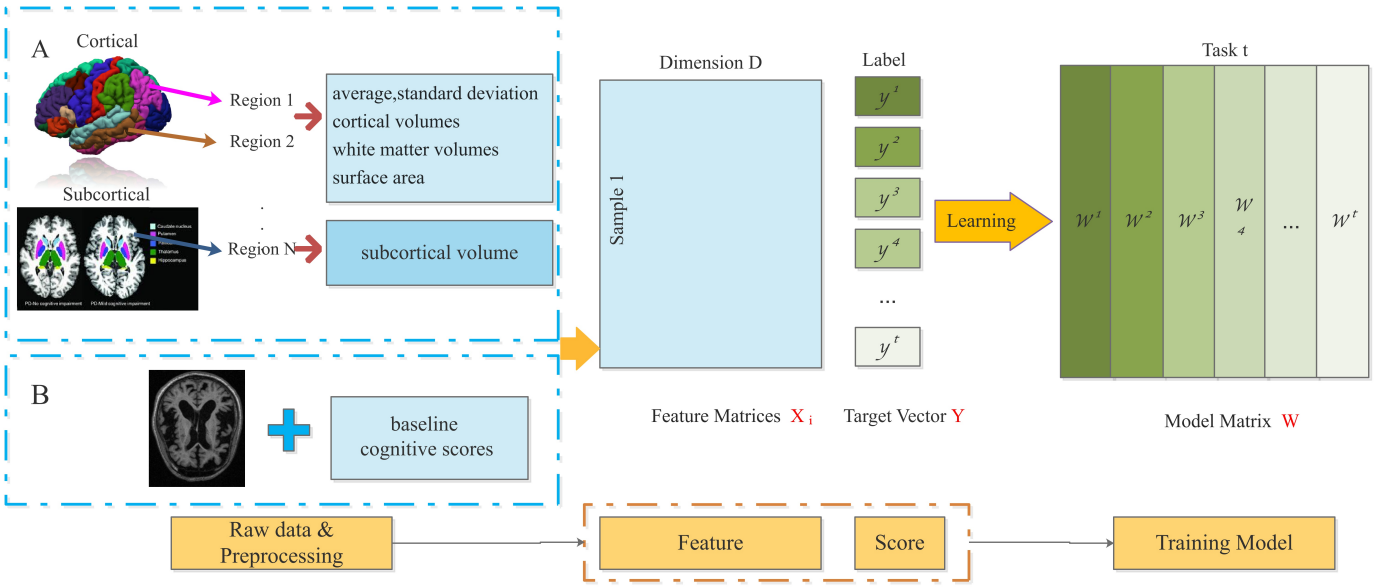



Fig. 1. Schematic illustration of AD progression model based on regularized multi-task learning

data strongly impacts on learning a fine model. Differing with traditional ML approaches, Multi-Task Learning (MTL) [11] considers the prediction of AD progression as multiple learning tasks each of which can be a general prediction task at certain time point. Among these prediction tasks, all of them are assumed to be related to each other in time domain with relevant temporal features (e.g., biomarkers in MRI). As shown in Fig. 1, we demonstrate a typical pipeline of leveraging MTL algorithms for predicting cognitive functionality of AD patients from their brain imaging scans [12], where the predictive information is shared and transferred among related models to reinforce their generalization performance. The data sources employed are A (Extracted features from MRI like Volume of Hippocampus) and B (AD cognitive scores like MMSE or ADAS-cog [4], [13]) from selected AD patients repeatedly by multiple time points. By considering the prediction of cognitive scores at a single time point (like 6, 12 or 18 months) as a regression task. The prediction of clinical scores at multiple future time points as a multi-task regression problem. MTL model matrix is trained and optimized through processing pre-extracted features from MRI and baseline cognitive scores.

Two important issues affect the progress of applying MTL in AD modelling problems. First, it is important to obtain good quality of baselines from AD raw data, where Magnetic resonance imaging (MRI) reflects changes in brain structure, such as the cerebral cortex and ventricle; cognitive score directly shows cognitive functions of AD patients. Sparse representation [14] is a popular method in MTL for capturing key biomarkers in AD, which uses sparseness as a regularization condition, image blocks with key characteristics. Cognitive measure can be achieved by using worldwide standard AD cognitive assessment, such as Mini Mental State Exam score (MMSE), Alzheimer's Disease Assessment Scale cognitive total score (ADAS-cog) and Rey Auditory Verbal Learning Test (RAVLT) [5], [15]. As the second issue, utilizing and improving advanced regression models [16]–[19], in MTL are highly critical, where they could better explore the relationship and correlations between MRI features and cognitive measures. Here, structural regularization [2] is a common approach in MTL for minimize the penalized empirical loss and bundling the correlations between tasks in the assumption. In the field of MTL in AD, there are many prior work that model relationships among tasks using novel regularizations [6], [20], [21]. The addition of kernel method problems allows the algorithm to fit non-linear relationships [6], [22]. The benchmark of this paradigm is derived from Zhou et al. [10] and subsequent achievements are mostly aimed at theoretical structure, relevance, and fusing the multi-modality data applications. So far to our best knowledge, above regularized MTL approaches deliver promising performance in many AD prediction applications.

Notably, it is worth mentioning that most existing MTL studies in AD progression only focus on above two research issues, where they follow a unified regularized MTL framework to process ANDI datasets with simple evaluation matrix. Another important aspect here is if their experimental evaluation strategies are objective and reliable to access their performance. In different settings of dataset and parameters, MTL algorithms would perform differently in the tasks of prediction. Current empirical understanding of evaluation and judgement process in MTL for AD study is very limited. There is little attention on studying how to design suitable experimental protocols and evaluation matrix for assessment of regularized MTL algorithms.

In this paper, we describe an empirical study and analysis

that investigate above question. We examined typical MTL models via structural regularization approaches in AD study, and choose two typical single task models (Ridge regression and Lasso regression) and two state-of-the-art MTL models (TGL and cFSGL) [8], [10] as targeted methods. Considering that MTL features shared parameters and representations, we conduct four important points potentially affecting evaluation process of regularised MTL models in AD study: 1) evaluation indicators (e.g. use correlation coefficients and mean square errors to obtain two different sets of models hyperparameters.) 2) repeated experimental times (e.g., results of 10 repeated experiments and 100 repeated experiments are different results; 3) size and portion of training data; 4) number of tasks in MTL (e.g., time points in AD progression). For each point, we design and set up experimental protocols for comparison and exploration, highlighting following multi-fold contributions:

- We demonstrate that regularized MTL models like cFSGL are capable of predicting AD progression with high accuracy, in many cases of data missing, insufficiency or single MRI data input. This confirms a fact that cFSGL is under the fused Lasso penalty where selected features of AD across different time points are similar to each other, satisfying the temporal smoothness property in AD progression study.
- We provide a solid evidence verification point on whether regularized MTL model perform well in complex practical experimental settings. One important finding is that the performance gain of cFSGL may not only from its ability on dealing with sparsity of AD feature data labels. It is more likely due to existence of a low rank space inside original AD data features. Collinearity of low rank sub-spaces implies that the model actually needs fewer features than the input features at present. This provides a checkpoint for whether the model works well in more complex practical applications.
- By leveraging verification in experimental progress, we also discover and proof that some limitations of regularized MTL models in AD study: 1) mMSE is the best indicator to evaluate these models due to relatively stable performance, but other evaluation indicators are not so reliable to objectively access model performance. 2) cFSGL has a great potential for further improvement at late stage prediction of AD progression. 3) The assumption of temporal smoothness in regularized MTL models for AD study limits the performance improvement of the initial task. This hypothesis is a special relationship that fails to accurately capture certain tasks.

## II. Methodology

### A. Problem formulation of AD Progression

In the longitudinal AD study, the cognitive scores and MRI of AD patients are repeatedly measured at multiple time points. MTL approaches usually consider the prediction of cognitive scores at each single time point as a regression task, and then formulate the prediction of clinical scores at multiple future time points as a multi- task regression problem. Also, the prior knowledge of intrinsic temporal smoothness information among different tasks can be incorporated into the MTL model.

Consider a multi-task learning of k tasks with n training samples of d features. Let $x_1, x_2, ..., x_n$be the input data for the samples, and $y_1, y_2, ..., y_n$be the predicted value for each sample, where each $x_i \in \mathbb{R}^d$ represents the feature data of an AD patient, and $y_i \in \mathbb{R}$ is the predicted value of cognitive score of different types of scales.

Then, let$X = [x_1, ..., x_n]^T \in \mathbb{R}^{n\times d}$ be the data matrix, $Y = [y_1, ..., y_n]^T \in \mathbb{R}^{n\times k}$ be the predicted matrix, and $W = [w_1, ..., w_k]^T \in \mathbb{R}^{d\times k}$ be the weight matrix. The process of establishing a MTL model is to estimate the value of W, which is the parameter to be estimated from the training samples.

In order to solve above problem, many prior works in MTL that model relationships among tasks using regularization methods. Normally, they assume the empirical loss to be square loss and common regularization terms are $\mathcal{L}_1$ and $\mathcal{L}_2$ norms, separately named as Lasso regression and ridge regression models as shown in Eq. 1 and 2. Ridge regression constrains variables to a smaller range for reducing some factors with little impacts on model's prediction. Unfortunately, this reduction means that these variables are still considered. To solve this problem, Lasso was proposed as a new sparse representation linear algorithm, which simultaneously performs feature selection and regression. Some variables are set to zero directly to achieve sparsity and dimensionality reduction.

$$\min_w L(Y, X, W) + \lambda||W||_1 \tag{1}$$

$$\min_w L(Y, X, W) + \lambda||W||_2 \tag{2}$$

In AD study, the task of predicting AD patient's cognitive score at certain time point is strongly associated with other tasks at adjacent time points. Thus, many recent studies have focused on designing novel structural regularization methods to improve their performance in AD study.

### B. Structural regularization methods

Structural regularization methods in MTL constrains optimization by using regularization terms and shares information between tasks. In this article, we mainly considering two state-of-the-art models proposed by Zhou [29]: Temporal Group Lasso (TGL) and Convex Fused Sparse Group Lasso (cFSGL).

Specifically, TGL contains a time smoothing term and a group Lasso term as constraints, which ensures that all regression models at different time points share a common set of features. The TGL formulation solves the following convex optimization problem:

$$\min_w ||XW - Y||_F^2 + \theta_1||W||_F^2 + \theta_2||WH||_F^2 + \delta||W||_{2,1} \tag{3}$$

where the first term measures the empirical error on the training data, $||W||_F$ is the Frobenius norm, $||WH||_F^2$ is the temporal smoothness term, which ensures a small deviation between two regression models at successive time points, and

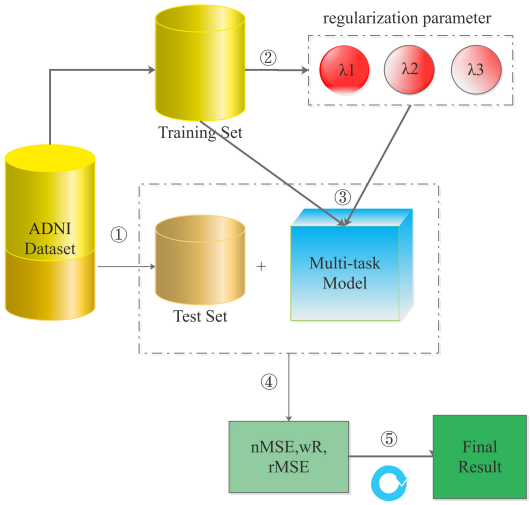



Fig. 2. Pipeline of empirical protocol design

$||W||_{2,1}$ is the group lasso penalty, which ensures that a small subset of features will be selected for the regression models at all-time points.

cFSGL involves sparsity between tasks, where it considers both common features at different points in time and unique features to each task. This feature is helpful to improve the overall performance of the model. cFSGL formulation solves the following convex optimization problem:

$$\min_{w} ||XW - Y||_F^2 + \theta_1||W||_1 + \theta_2||RW^T||_1 + \delta||W||_{2,1} \quad (4)$$

where the first term measures the empirical error on the training data, $||W||_1$ is the lasso penalty, $||RW^T||_1$ is the fused lasso penalty, and $||W||_{2,1}$ is the group lasso penalty.

### C. Empirical protocol design

Our empirical protocol design is based on a pipeline shown in 2 The complete experimental process mainly includes 5 steps: 1) split the data set; 2) select the hyperparameters; 3) train the model; 4) evaluate the model using the test set; 5) iterate the above operations. Different colors donate the source or generation of different data, arrows indicate the flow of data, and serial numbers re the steps of the experiment.

Our first goal is to perform quantitative reproducibility analysis of typical 4 regularized MTL methods (Ridge, Lasso, TGL and cFSGL) in comparing to Zhou's [29] experiment results. Plus, we consider one practical application case with only MRI data as input data to predict cognitive scores at baseline and future time points. In many real-world AD application scenario, it is hard to acquire both precise MRI and cognitive measures. Then we would set up individual experimental protocol for exploring four important points of evaluating MTL models in AD study: 1) evaluation indicators, 2) repeated experimental times; 3) size and portion of training data; 4) number of tasks in MTL.

The evaluation metric of cross-validation is employed to evaluate the performance of AD progression model. When a metric is set in the cross-validation experiment process, a set of hyperparameters can be obtained. By comparing the pros and cons of the results, the suitable metric for the model is finally determined. The regression performance metric often employed in MTL is normalized mean square error (nMSE) and root mean square error (rMSE) is employed to measure the performance of each specific regression task. In particular, nMSE has been normalized to each task before evaluation, so it is widely used in multi-task learning methods based on regression tasks. Also, weighted correlation coefficient (wR) as employed in the medical literature addressing AD progression problems[10, 19, 30] nMSE, rMSE and wR are defined as follows:

$$\mathrm{nMSE}(Y, \hat{Y}) = \frac{\sum_{i=1}^{t} \left\| Y_i -, \hat{Y}_i \right\|_2^2 / \sigma\left(Y_i\right)}{\sum_{i=1}^{t} n_i} \quad (5)$$

$$\mathrm{rMSE}(y, \hat{y}) = \sqrt{\frac{\|y - \hat{y}\|_2^2}{n}} \quad (6)$$

$$\mathrm{wR}(\mathrm{Y}, \hat{Y}) = \frac{\sum_{i=1}^{t} \mathrm{Corr}\left(\mathrm{Y}_i, \hat{Y}_i\right) n_i}{\sum_{i=1}^{t} n_i} \quad (7)$$

As for repeated experimental times, one evaluation consensus in MTL models for AD study is that one experiment result is usually accidental and unreliable. To reduce experiment accidental errors, repeated experiments are required. So we evaluate the performance of four selected regularized MTL models under different repeated experimental times. Lastly, we will evaluate typical factors like data size and number of tasks affecting MTL models.

## III. Data and Implementation

To verify the effectiveness of disease progression models, data in Alzheimer's Disease Neuroimaging Initiative (ADNI) are used as research. The ADNI is a longitudinal multicenter study designed to develop clinical, imaging, genetic, and biochemical biomarkers for the early detection and tracking of AD. High quality standardized MRI data and cognitive function measures data available can be acquired on the website. In this study, image preprocessing was performed to obtain relevant statistical indicators of relevant regions-of-interest (ROI) of the subject's brain in baseline period, such as average cortical thickness, standard deviation in cortical thickness, the volumes of cortical parcellations, the volumes of specific white matter parcellations, and the total surface area of the cortex. This step was employed the Freesufer [23] image analysis suite for processing and analyzing brain MRI images to complete the cortical reconstruction and cortical segmentation. Furthermore, we performed data cleaning to remove individuals that failed image processing. The columns with a small number of missing values were filled with the average, and the columns with a large number of missing or completely missing were directly deleted. Moreover, for the purpose of establishing a longitudinal study, the experiment will set a baseline for subsequent follow-up observations, and we define the follow-up observation after the 6th month of

TABLE I
THE VALIDITY OF AD DISEASE PROGRESSION MODEL

| | Ridge | Lasso | TGL [12] | cFSGL [8] |
|---|---|---|---|---|
| *Target:MMSE* | | | | |
| nMSE | 1.185±0.286 | 0.641±0.156 | 0.562±0.106 | **0.459±0.095** |
| wR | 0.545±0.057 | 0.694±0.034 | 0.734±0.057 | **0.777±0.034** |
| rMSE M06 | 2.770±0.360 | 2.044±0.472 | 1.853±0.225 | **1.845±0.259** |
| rMSE M12 | 3.029±0.293 | 2.226±0.466 | 1.972±0.244 | **1.873±0.266** |
| rMSE M24 | 3.375±0.470 | 2.690±0.664 | 2.544±0.535 | **2.374±0.479** |
| rMSE M36 | 4.533±0.513 | 3.287±0.584 | 3.060±0.437 | **2.932±0.594** |
| *Target:ADAS-cog* | | | | |
| nMSE | 0.693±0.116 | 0.417±0.052 | 0.408±0.073 | **0.358±0.057** |
| wR | 0.660±0.052 | 0.777±0.034 | 0.789±0.042 | **0.809±0.034** |
| rMSE M06 | 4.517±0.412 | 3.387±0.496 | 3.500±0.561 | **3.319±0.401** |
| rMSE M12 | 3.387±0.393 | 3.644±0.462 | **3.467±0.437** | 3.485±0.473 |
| rMSE M24 | 5.519±0.713 | 4.248±0.828 | 4.260±0.913 | **3.553±0.453** |
| rMSE M36 | 7.655±1.200 | 6.088±1.077 | **5.707±0.824** | 5.739±1.037 |

TABLE II
THE RESULT OF ONLY USE MRI DATA AS INPUT DATA TO PREDICT COGNITIVE SCORES AT BASELINE AND FUTURE TIME POINT

| | Ridge | Lasso | TGL [12] | cFSGL [8] |
|---|---|---|---|---|
| *Target: ADAS-cog* | | | | |
| nMSE | 1.180±0.140 | 0.727±0.058 | 0.537±0.066 | **0.521±0.081** |
| wR | 0.410±0.072 | 0.541±0.049 | 0.703±0.040 | **0.712±0.055** |
| BL rMSE | 5.439±0.543 | 4.138±0.501 | **3.794±0.389** | 3.880±0.356 |
| M06 rMSE | 6.128±0.906 | 4.599±0.906 | 4.125±0.475 | **3.776±0.427** |
| M12 rMSE | 6.225±0.835 | 4.879±0.835 | 4.091±0.468 | **3.965±0.809** |
| M24 rMSE | 7.216±1.112 | 5.857±1.007 | 4.514±0.790 | **4.742±0.646** |
| M36 rMSE | 9.914±1.242 | 7.501±1.309 | **7.091±0.971** | 7.156±1.054 |

the baseline period as M06. In terms of cognitive function scores as dependent (target) of each task, we selected subjects from baseline to several future time points, such as M06, M12, M24, and M36, and none of them had missing records. After the preprocessing procedure, there are a total of 429 subjects and 327 MRI features, which together with the baseline scores will be used as input data.

## IV. EXPERIMENTS

### A. Reproducible analysis

Our first goal is to perform quantitative reproducibility analysis of typical 4 regularized MTL methods (Ridge, Lasso, TGL and cFSGL) in comparing to Zhou's [29] experiment results. Specifically, dataset was randomly split into training and testing sets using a ratio 9:1, i.e., models were built on 90% of the data and evaluated on the remaining 10% of the data. Models parameters were selected by 5-fold cross validation.

The only difference is that our tasks and samples are slightly less due to available extracted AD data. There are two different settings from [29], namely samples and tasks, making sure that the labels of data are not missing, our data samples are less than the Zhou et al. [29] which increases the risk of underfitting to some extent, especially for single task models. The learning process of multi-task learning models will also be affected. The number of tasks affecting the performance of MTL models will be discussed later.

In Table I, it shows that our experimental results under similar settings are quite close to Zhou's [12] outcomes. It implies that four selected structural regularization methods are all robust. Also, MTL models (TGL and cFSGL) outperforms single-task learning model (Ridge and Lasso), in terms of prediction accuracy. This accords with our previous survey of features of MTL in dealing with data insufficiency cases. Notably, cFSGL performs the best in all 4 methods. It is probably because in AD study, the model built by cFSGL has two levels of sparsity: 1) a small set of features shared across all tasks, 2) task-specific features for each time point. One key advantage of fused Lasso in cFSGL is that under the fused Lasso penalty the selected features across different time points are similar to each other, satisfying the temporal smoothness property, while the Laplacian-based penalty focuses on the smoothing of the prediction models across different time points.

### B. Application with only MRI data input

In many real-world AD application scenario, clinicians expect the prediction model to be simple and with less input data required for giving timely early screening. In this case, it is hard to acquire both precise MRI and cognitive measures. Normally, doctors need to spend few hours to measure AD patients' cognitive scores though some tests. Thus, in this case, we consider one application with only MRI data as input data to predict cognitive scores at baseline and future time points. It is necessary for doctors to perform a cognitive scale assessment, but time-consuming to complete a set of cognitive measures. Using baseline cognitive measures as a predictive target have far-reaching significance. The arrangement of experimental and results are shown in Table II.

The results in Table II show that cFSGL still performs the best of prediction in all 4 methods, especially at the later point time. That proves one assumption in Zhou's work [29] of a linear relationship between MRI features and cognitive scores. We also find out that joint analysis of multiple time points is capable of improving the predict performance of MTL approaches.

However, one importantly arguable points in cFSGL we find out is that many work believes the performance improvement of cFSGL may gain from its ability on dealing with sparsity of AD feature data labels. (For various reasons, many values of AD data are missing at an individual later point time). Our experimental results show that it may not be the key factor in improving the performance of MTL model, because the labels of our dataset are not missing.

Therefore, this may be due to existence of a low rank space inside original data features. Collinearity of low rank subspaces implies that the model actually needs fewer features than the input features at present. Specifically, since the features of input data are various statistical indicators of the subject's brain area, we can simply draw a conclusion that the shrinkage of one area (reduction of indicators) may cause the shrinkage of another area (reduction of indicators synchronization).

TABLE III
THE RESULT BASED ON DIFFERENT METRIC

| | Lasso | TGL | cFSGL |
|---|---|---|---|
| *cv: nMSE* | | | |
| nMSE | 0.779±0.077 | 0.718±0.137 | **0.629±0.077** |
| wR | 0.516±0.043 | 0.630±0.049 | **0.677±0.049** |
| BL rMSE | 1.805±0.232 | **1.803±0.251** | 1.816±0.286 |
| M06 rMSE | 2.345±0.337 | 2.132±0.293 | **1.962±0.182** |
| M12 rMSE | 2.393±0.537 | 2.393±0.385 | **1.966±0.312** |
| M24 rMSE | 3.087±0.633 | 3.087±0.572 | **2.345±0.400** |
| M36 rMSE | 3.924±0.751 | 3.924±0.683 | **3.232±0.550** |
| *cv: wR* | | | |
| nMSE | 0.783±0.072 | **0.712±0.192** | 0.750±0.269 |
| wR | 0.514±0.050 | 0.667±0.043 | **0.710±0.041** |
| BL rMSE | 1.702±0.225 | **1.813±0.291** | 2.112±0.329 |
| M06 rMSE | 2.293±0.218 | 2.109±0.312 | **2.059±0.309** |
| M12 rMSE | 2.385±0.425 | **2.040±0.296** | 2.092±0.330 |
| M24 rMSE | 2.975±0.648 | **2.570±0.470** | 2.579±0.809 |
| M36 rMSE | 3.635±0.577 | 3.741±1.118 | **3.528±0.888** |
| *cv: rMSE* | | | |
| nMSE | 0.788±0.091 | 0.684±0.194 | **0.630±0.007** |
| wR | 0.522±0.044 | 0.648±0.062 | **0.691±0.042** |
| BL rMSE | 1.776±0.229 | **1.823±0.293** | 1.879±0.277 |
| M06 rMSE | 2.275±0.348 | 1.996±0.262 | **1.943±0.208** |
| M12 rMSE | 2.523±0.543 | 2.133±0.272 | **1.907±0.243** |
| M24 rMSE | 3.180±0.411 | **2.424±0.544** | 2.563±0.515 |
| M36 rMSE | 3.788±0.556 | 3.345±0.596 | **3.149±0.584** |
| *cv: MSE* | | | |
| nMSE | 0.765±0.057 | 0.650±0.087 | **0.613±0.132** |
| wR | 0.527±0.032 | 0.658±0.039 | **0.684±0.039** |
| BL rMSE | 1.806±0.218 | 1.748±0.148 | **1.738±0.252** |
| M06 rMSE | 2.304±0.354 | **1.952±0.234** | 2.059±0.267 |
| M12 rMSE | 2.338±0.486 | 2.083±0.261 | **1.992±0.236** |
| M24 rMSE | 3.138±0.759 | 2.689±0.541 | **2.472±0.576** |
| M36 rMSE | 3.876±0.597 | 3.391±0.645 | **3.228±0.579** |

### C. Evaluation indicators

In MTL for AD study, cross-validation with evaluation metric is widely utilized to select suitable model hyper-parameters. Good hyperparameters can make MTL models have better generalization performance. When an evaluation metric is set in cross-validation experiment process, a set of hyperparameters can be obtained. By comparing the pros and cons of the results, the suitable metric for the model is finally determined. However, different metrics have different preferences and emphasis on the model. It has become a consensus to employ metrics to evaluate the pros and cons of models.

Three models (Lasso, TGL and cFSGL) are selected for evaluation. Dataset is randomly split into training and testing sets using a ratio 9:1. Models parameters were selected by 5-fold cross validation. The mean and standard deviation based on 20 iterations of experiments. The experimental results in Table III confirm our concern that selection of evaluation metrics significantly affect performance assessment of regularized MTL models.

As shown in the Table III, we could find out that the 1) results obtained by metrics such as square error (MSE, rMSE, nMSE) are basically the same; 2) mMSE is the best indicator to evaluate these models due to relatively stable performance. The reason is that data distribution of each task is not the same, sharing with each other will have the effect of noise. Therefore, using the variance of tasks in nMSE will reduce the impact of task differences, and the results can better take into account each other's tasks.

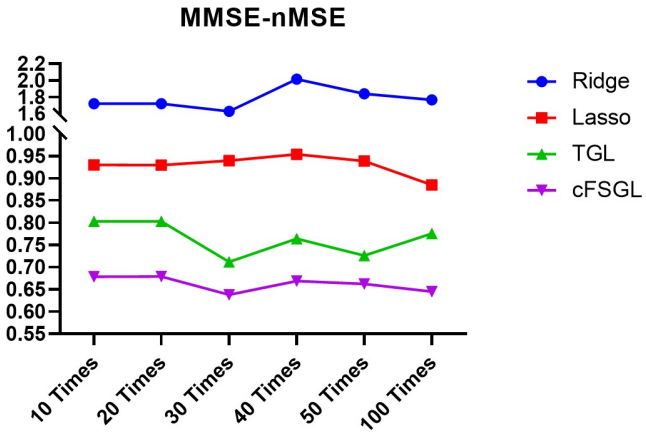



Fig. 3. Evaluation results of repeated experiments times

### D. Repeated experimental times

In MTL for AD study, one typical evaluation consensus is that one experiment result is usually accidental and unreliable. To reduce experiment accidental errors, repeated experiments are required. So we evaluate the performance of four regularized MTL models under different repeated experimental times. We conducted six sets of experiments, and the number of iterations in each set was 10, 20, 30, 40, 50, 100. Also, in each set of experiments, other conditions remained the same, namely: dataset was randomly split into training and testing sets using a ratio 9:1, data includes tasks at 5 time points, the final result is shown in 3. The horizontal axis represents iteration, the vertical axis represents the nMSE value of each algorithm, and different colors represent algorithm.

In 3, it appears that the effect of different experiments on four algorithm are visually observed. All 4 MTL models maintains good performance in each set of experiments. From the fluctuation range of the model mean: Ridge not only performs poorly overall, but also has a large range of fluctuations, which may be the reason for the underfitting. The average volatility of Lasso, TGL, cFSGL is 0.06, 0.11, 0.05. As the number of iterations increased, four algorithms are fluctuating to varying degrees. The reason may be caused by abnormal information, for example, the existence of abnormal points during training. By taking more repeated experimental times, the probability of anomalous information being hit is also higher, which more conforms to the realistic scenario. Lasso and cFSGL are relatively less affected, which implies that sparsity plays a key role in real-world scenarios.

### E. Size and portion of training data

One important advantage of MTL is to deal with the problem of data missing and insufficiency. In other words,

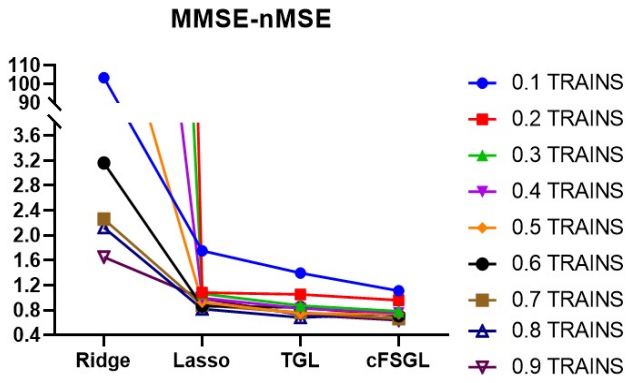



Fig. 4. nMSE values for predicting MMSE cognitive scores under different data sizes

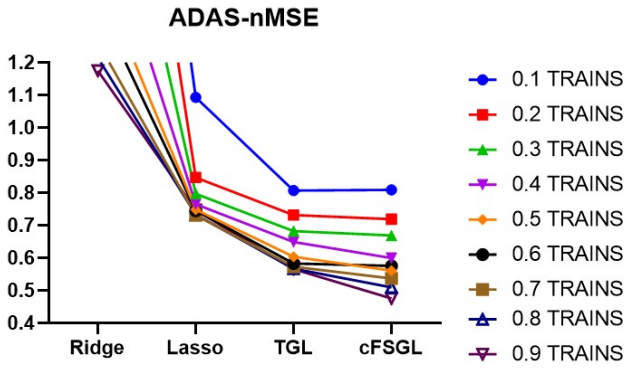



Fig. 5. nMSE values for predicting ADAS-cog cognitive scores under different data sizes

regularized MTL models reduce the risk of underfitting while improving overall performance. To prove this assumption and further examine degree of resistance to underfitting risk in the prediction MTL model of AD disease progression, we would like to evaluate different portion of training AD data over these regularized MTL models.

We train four MTL models with datasets of different data sizes. Nine groups of experiments were performed. When splitting the training and test sets, we followed 1: 9, 2: 8, 3: 7, 4: 6, 5: 5, 6: 4, 7: 3, 8: 2, 9: 1 operation. In order to compare the experimental results, the other condition settings of each group of experiments are kept consistent: two datasets with MMSE and ADAS-cog scores as learning labels are conducted, with 429 and 425 samples respectively. The same data set was used to predict the trend of cognitive scores of the MMSE and ADAS-cog scales at baseline and in the next three years. The result based on 20 iterations of experiments on different splits of data using 5-fold cross validation. Each group of experiments uses four algorithms (Ridge, Lasso, TGL and cFSGL) for comparison. The results are shown in the Fig **??** . The finding shows that:

- Ridge and Lasso are underfitting definitely, and multi-task learning methods represented by TGL and cFSGL show advantages.
- cFSGL does not play the advantage of spare even 9:1, which implies that this cFSGL has greater potential to improve performance.
- From the comparison of the two MTL methods, as the amount of training data increases, the performance of cFSGL is gradually better than TGL.

### *F. Number of tasks in MTL*

Another key issue to regularized MTL models is resource exchange and sharing between multiple tasks. The common method is to propose an assumption, which can be transformed into a constraint and put into an optimization function. But whether this assumption relationship is worth scrutinizing needs to be paid more attention. Therefore, several sets of experiments were designed to test the validity of this relationship.

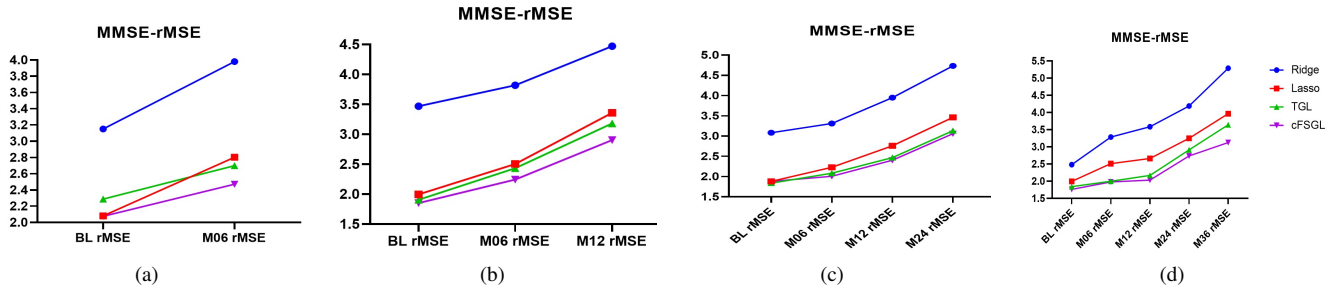
We carried out four sets of experiments using 2-5 tasks together to build a MTL model. The purpose of the experiment is to find whether the performance of the model can be improved under this task relationship. The other condition settings of each group of experiments are kept consistent: the same data set is exploited to predict the trend of cognitive scores of the MMSE. The results are based on 20 iterations of experiments on different splits of data with 9:1 using 5-fold cross validation. Four algorithms (Ridge, Lasso, TGL and cFSGL) are conducted in each group for comparison. The results are shown in the Fig 6 (a)-(d). The finding shows that:

- On the assumption of temporal smoothness of AD study, the results of MTL models (TGL and cFSGL) are much better than single-task models (Ridge and Lasso).
- As the number of tasks in MTL increases, the accuracy gain of MTL models in AD progression prediction become more obvious.
- At the beginning, the errors of the two MTL models are small.
- As time goes by, the error of the task increase, this may be due to a non-linear relationship of MRI features and cognitive scores in the late stage of AD progression.

Combining (3) and (4), we can observe one key limitation of the temporal smoothness assumption in regularized MTL models for AD study, where the performance improvement of the initial task is limited. The problem may be because the relationship between MRI and cognitive scores is non-linear, or the temporal smoothness hypothesis is a special relationship that fails to accurately capture certain tasks.

## V. Conclusion

In this paper, we describe an empirical study and analysis of evaluation and judgement process in MTL for AD study. We examined four typical MTL models via structural regularization approaches in AD study and conduct four important points




(a) (b) (c) (d)

Fig. 6. Schematic diagram of nMSE values under different experiment

potentially affecting evaluation process of regularised MTL models in AD study. Our finding shows that regularized MTL models are capable of predicting AD progression with high accuracy, in many cases of data missing, insufficiency or single MRI data input. But they also suffer from some limitations: first concern is that the performance gain of cFSGL may not only from its ability on dealing with sparsity of AD feature data labels. It is more likely due to existence of a low rank space inside original AD data features. Collinearity of low rank subspaces implies that the model actually needs fewer features than the input features at present. Secondly, the assumption of temporal smoothness in regularized MTL models for AD study limits the performance improvement of the initial task. MTL like cFSGL has a great potential for further improvement at late stage prediction of AD progression. Our work could guide how to design suitable experimental protocols and evaluation matrix for assessment of regularized MTL algorithms. Also it highlights the future possible directions of utilising and improving regularized MTL models in AD progression study.

## REFERENCES


[1] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 58, no. 1, pp. 267–288, 1996.

[2] J. Zhou, J. Chen, and J. Ye, "Malsar: Multi-task learning via structural regularization," *Arizona State University*, vol. 21, pp. 1–50, 2011.

[3] L. Jönsson, P. Lindgren, A. Wimo, B. Jönsson, and B. Winblad, "Costs of mini mental state examination-related cognitive impairment," *Pharmacoeconomics*, vol. 16, pp. 409–416, 1999.

[4] P. Doraiswamy, F. Bieber, L. Kaiser, K. Krishnan, J. Reuning-Scherer, and B. Gulanski, "The alzheimer's disease assessment scale: patterns and predictors of baseline cognitive performance in multicenter alzheimer's disease trials," *Neurology*, vol. 48, no. 6, pp. 1511–1517, 1997.

[5] A. Association, "2019 alzheimer's disease facts and figures," *Alzheimer's & dementia*, vol. 15, no. 3, pp. 321–387, 2019.

[6] J. Peng, X. Zhu, Y. Wang, L. An, and D. Shen, "Structured sparsity regularized multiple kernel learning for alzheimer's disease diagnosis," *Pattern recognition*, vol. 88, pp. 370–382, 2019.

[7] M. Sun, I. M. Baytas, L. Zhan, Z. Wang, and J. Zhou, "Subspace network: Deep multi-task censored regression for modeling neurodegenerative diseases," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, pp. 2259–2268.

[8] J. Zhou, J. Liu, V. A. Narayan, and J. Ye, "Modeling disease progression via fused sparse group lasso," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2012, pp. 1095–1103.

[9] M. Nguyen, N. Sun, D. C. Alexander, J. Feng, and B. T. Yeo, "Modeling alzheimer's disease progression using deep recurrent neural networks," in *2018 International Workshop on Pattern Recognition in Neuroimaging (PRNI)*. IEEE, 2018, pp. 1–4.

[10] J. Zhou, J. Liu, V. A. Narayan, J. Ye, A. D. N. Initiative *et al.*, "Modeling disease progression via multi-task learning," *NeuroImage*, vol. 78, pp. 233–248, 2013.

[11] K.-H. Thung and C.-Y. Wee, "A brief review on multi-task learning," *Multimedia Tools and Applications*, vol. 77, pp. 29 705–29 725, 2018.

[12] J. Zhou, L. Yuan, J. Liu, and J. Ye, "A multi-task learning formulation for predicting disease progression," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2011, pp. 814–822.

[13] L. Chu, K. Chiu, S. Hui, G. Yu, W. Tsui, and P. Lee, "The reliability and validity of the alzheimer's disease assessment scale cognitive subscale (adas-cog) among the elderly chinese in hong kong." *Annals of the Academy of Medicine, Singapore*, vol. 29, no. 4, pp. 474–485, 2000.

[14] Z. Zhang, Y. Xu, J. Yang, X. Li, and D. Zhang, "A survey of sparse representation: algorithms and applications," *IEEE access*, vol. 3, pp. 490–530, 2015.

[15] M. Schmidt *et al.*, *Rey auditory verbal learning test: A handbook*. Western Psychological Services Los Angeles, CA, 1996, vol. 17.

[16] 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 *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 940–947.

[17] M. Wang, D. Zhang, D. Shen, and M. Liu, "Multi-task exclusive relationship learning for alzheimer's disease progression prediction with longitudinal data," *Medical image analysis*, vol. 53, pp. 111–122, 2019.

[18] C. M. Stonnington, C. Chu, S. Klöppel, C. R. Jack Jr, J. Ashburner, R. S. Frackowiak, A. D. N. Initiative *et al.*, "Predicting clinical scores from magnetic resonance scans in alzheimer's disease," *Neuroimage*, vol. 51, no. 4, pp. 1405–1413, 2010.

[19] G. B. Frisoni, N. C. Fox, C. R. Jack Jr, P. Scheltens, and P. M. Thompson, "The clinical use of structural mri in alzheimer disease," *Nature Reviews Neurology*, vol. 6, no. 2, pp. 67–77, 2010.

[20] P. Cao, X. Shan, D. Zhao, M. Huang, and O. Zaiane, "Sparse shared structure based multi-task learning for mri based cognitive performance prediction of alzheimer's disease," *Pattern Recognition*, vol. 72, pp. 219–235, 2017.

[21] X. Liu, A. R. Goncalves, P. Cao, D. Zhao, A. Banerjee, A. D. N. Initiative *et al.*, "Modeling alzheimer's disease cognitive scores using multi-task sparse group lasso," *Computerized Medical Imaging and Graphics*, vol. 66, pp. 100–114, 2018.

[22] P. Cao, X. Liu, J. Yang, D. Zhao, M. Huang, and O. Zaiane, "2, 1- 1 regularized nonlinear multi-task representation learning based cognitive performance prediction of alzheimer's disease," *Pattern Recognition*, vol. 79, pp. 195–215, 2018.

[23] B. Fischl, "Freesurfer," *Neuroimage*, vol. 62, no. 2, pp. 774–781, 2012.