Proceedings Paper:
Zhou, M., Liu, T., Wang, X. et al. (3 more authors) (2024) Integrating automatic temporal relation graph into multi-task learning for Alzheimer's disease progression prediction. In: 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 05-08 Dec 2023, Istanbul, Turkey. Institute of Electrical and Electronics Engineers (IEEE) , pp. 3265-3272. ISBN 979-8-3503-3749-5

https://doi.org/10.1109/BIBM58861.2023.10385873

Integrating Automatic Temporal Relation Graph into Multi-Task Learning for Alzheimer's Disease Progression Prediction

Menghui Zhou
Department of Computer Science
The University of Sheffield
Sheffield, United Kingdom
mzhou47@sheffield.ac.uk

Xulong Wang
Department of Computer Science
The University of Sheffield
Sheffield, United Kingdom
xl.wang@sheffield.ac.uk

Yu Zhang
Department of Computer Science
The University of Sheffield
Sheffield, United Kingdom
yzhang489@sheffield.ac.uk

Tong Liu
Department of Chemical Engineering
Imperial College London
London, United Kingdom
tliu.soton@gmail.com

Kang Liu
Department of Computer Science
The University of Sheffield
Sheffield, United Kingdom
kang.liu@sheffield.ac.uk

Po Yang
Department of Computer Science
The University of Sheffield
Sheffield, United Kingdom
po.yang@sheffield.ac.uk

***Abstract*—Alzheimer's disease (AD), the most prevalent dementia, gradually reduces the cognitive abilities of patients while also posing a significant financial burden on the healthcare system. A variety of multi-task learning methods have recently been proposed to identify potential MRI-related biomarkers and accurately predict the progression of AD. These methods, however, all use a predefined task relation structure that is rigid and insufficient to adequately capture the intricate temporal relations among tasks. Instead, we propose a novel mechanism for directly and automatically learning the temporal relation and constructing it as an Automatic Temporal relation Graph (AutoTG). We use the sparse group Lasso to select a universal MRI feature set for all tasks and particular sets for various tasks in order to find biomarkers that are useful for predicting the progression of AD. To solve the biconvex and nonsmooth objective function, we adopt the alternating optimization and show that the two related suboptimization problems are amenable to closed-form solution of the proximal operator. To solve the two problems efficiently, the accelerated proximal gradient method is used, which has the fastest convergence rate of first-order method. We have preprocessed two latest AD datasets, and the experimental results verify our proposed novel multi-task approach outperforms several baseline methods. To demonstrate the high interpretability of our approach, we visualize the automatically learned temporal relation graph and investigate the temporal patterns of the important MRI features. The implementation source is at https://github.com/menghui-zhou/MAGPP.**

***Index Terms*—Alzheimer's disease, multi-task learning, automatic temporal relation graph, disease progression**

I. Introduction

Alzheimer's disease (AD), the most prevalent neurodegenerative disorder, is marked by the deterioration of cognitive abilities over time [1]. Since only a challenging brain biopsy or autopsy can provide a conclusive diagnosis of AD, it is of great importance to accurately predict AD progression over time. There are currently no treatments that can halt or reverse AD progression, it is hence crucial to identify the biomarkers that are significant to the emergence of AD [1].

Previous studies have demonstrated that a variety of cognitive scores, such as ADAS-Cog (the Alzheimer's Disease Assessment Scale Cognitive Sub-scale) and MMSE (the Mini-Mental State Examination), are capable of assessing the state of AD patients [2]. Noninvasive structural magnetic resonance imaging (MRI) can identify atrophic changes in the brain [3]. Due to an intrinsic relation between a series of time points, it is anticipated that a joint examination of several time points will enhance model performance. To achieve this goal, in recent years, several multi-task learning (MTL) strategies have been put forth to forecast how AD will develop [4], [5], [6]. They consider predicting a target at a series of time points to be a MTL problem, with each task focusing on the prediction at a specific time point. As illustrated in Fig. 1, the k-th time point is regarded as the k-th task $\mathbf{w_k}$. The goal of MTL is to improve generalization ability and model performance by utilizing the inherent relations between various related tasks [7]. Despite recent great advancements made in investigating AD through MTL, a significant challenge is determining how to fully capture and hence exploit the complex temporal relation between multiple tasks.

A typical approach is employing the temporal smoothness relation, which assumes there is a limited difference between two adjacent tasks. Zhou et al. [4] propose a MTL method with the temporal group Lasso (TGL) and assume that the cognitive score of patients will not change significantly over time, i.e., there will not be much of a difference in cognitive scores between two successive time points. TGL penalizes the difference between adjacent tasks $\|\mathbf{w_k} - \mathbf{w_{k+1}}\|_2^2$ to achieve temporal smoothness at task level. Similar to TGL, a MTL formulation with convex sparse group Lasso (cFSGL) is

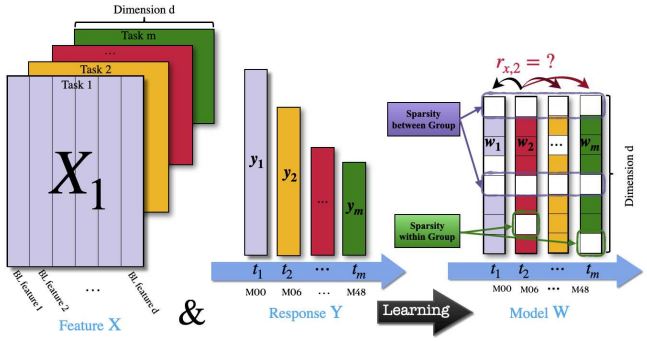

Fig. 1. Illustration of MTL prediction model. We use baseline MRI features to predict the progression of AD patients, whose states are measured by cognitive scores. The notation BL and M00 both mean the baseline time point. Mx means x months after baseline time point and $x \in \{0, 6, 12, 24, 36, 48\}$.

proposed in [5] which assumes that nearby time points have similar features, so they penalize $\sum_k |w_{i,k} - w_{i,k+1}|$ to pursue temporal smoothness at feature level. Clearly, these two kinds of methods seek the same outcome, i.e., $\mathbf{w_k} \approx \mathbf{w_{k+1}}$.

However, the main limitation is that both two kinds of temporal smoothness relation is a type of *local* and *predefined* structure. It only takes into account how the task relates with its neighbors, potentially ignoring other important task relations. In essence, if each task is viewed as a node in a graph, with edges determining task relation, TGL and cFSGL both utilize a graph with only edges between successive tasks, but on other edges. Different from TGL and cFSGL, Liu et al. [2] propose a multi-task formulation with fused Laplacian sparse group Lasso (FLSGL), which enables a fully connected graph with decreasing task weights. This type of relation is also based on a *predefined* Gaussian kernel. Recently Zhou et al. [6] propose an adaptive global temporal relation structure LSA. As this structure is built on a *predefined* and *specific* iterative convex combination, it has limited capability to handle complicated temporal relations among tasks.

Different from all mentioned existing methods, the motivation of this work comes from a common but extremely complicated situation, i.e., *the time points are not evenly distributed and the corresponding notation is usually inaccurate when collecting the data.* Specifically, as shown in Fig. 1, the notation $M00$ is the baseline time point and Mx represents x months after $M00$. Clearly, the time points are not evenly distributed since the intervals between two successive time points are not the same, i.e., 6 months or a year. Furthermore, even when the time points are evenly distributed, the given time notation is frequently inaccurate. The data at M24 may come from M23, M25, or M26 in practice [8].

To handle this common but extremely complicated problem, it should be far preferable to learn the complex temporal relation between tasks directly and automatically from the given data, rather than relying on a predefined temporal relation structure. So we present a novel mechanism, termed ***Auto**matic **T**emporal relation **G**raph* (AutoTG), to automatically capture the complex temporal relation between tasks and construct it as a relation graph.

Note that MTL based on temporal relation is found in a vast variety of applications. Except for the study of AD, Emrani et al. [9] use multi-task learning with temporal smoothness relation to diagonalize the progression of Parkinson's disease. Romeo et al. [10] suggest a novel spatio-temporal MTL with the temporal smoothness to predict the development of diabetes. Wang et al. [11] propose a temporal MTL model for survival analysis. Though this paper focuses on AD, we believe that AutoTG has great potential to be a building block for other MTL models based on temporal relation.

In the area of AD research, finding the biomarkers associated with the progression is crucial as well. We apply the sparse group Lasso [12] to introduce the sparsity between groups and within each group, as shown in Fig. 1. It means that we select a universal MRI feature set for all time points and particular sets for specific time points. Combing sparse group Lasso with AutoTG, we propose a novel **M**ulti-task learning approach with **A**utomatic temporal relation **G**raph for **P**redicting Alzheimer's disease **P**rogression (MAGPP).

We summarize main contributions of this work as follows:

- We present a novel multi-task approach MAGPP. It automatically captures the complex temporal relation between tasks and constructs it as a relation graph, while also selecting a universal MRI feature set for all time points and particular sets for specific time points. Experimental findings on two latest AD datasets show that MAGPP outperforms several baseline methods in terms of overall performance and nearly every task-specific performance.
- To explore the complex temporal relation among tasks, we visualize the automatically learned relation graph. It reveals that the temporal relation among tasks is not strictly symmetric. Not only that, tasks that are too far apart may even, although very slightly, repel rather than approximate each other which has never been considered in all previous works [2], [4], [5], [6].
- To show the high interpretability of MAGPP, we utilize the method of stability selection [5] to identify stable biomarkers from the MRI feature set and investigate their temporal patterns in the progression of AD. The features selected are consistent with previous work in bioinformatics, possibly facilitating the understanding of AD progression.

Notation: $\mathbb{N}_m = \{1, \cdots, m\}$. x_i and $x_{i,j}$ denote the i-th element of a vector $\boldsymbol{x}$ and the (i,j)-th element of a matrix X. $\boldsymbol{x_i}$ ($\boldsymbol{x^i}$) denotes the i-th column (row) of a matrix X. Euclidean and Frobenius norms are denoted by $\|\cdot\|_2$ and $\|\cdot\|_F$, $\langle A, B\rangle$ is the inner product, $A \odot B$ is component-wise multiplication of A and B. $\|X\|_{p,q} = (\sum_j (\sum_i x_{i,j}^p)^{q/p})^{1/q}$.

II. Related Work

A. Classification Methods for AD

Models based on classification attempt to group the condition of patients into various recognized disease stages, which are typically divided into Alzheimer's disease (AD), Mild Cognitive Impairment (MCI), and Cognitively Normal (CN).

For the purpose of using structural MRI to diagnose AD and localize joint atrophy, [3] suggest a hierarchical fully convolutional network. In [13], the multi-view input is regarded as the first layer in a multi-layer multi-view classification strategy, and a latent representation is constructed to examine the relation between class labels and features. In [14], the authors present an iterative sparse and deep learning model for diagnosing AD. In [15], the authors applied the generative adversarial network for the assessment of AD. Although the aforementioned models worked well for classification problems, they failed to predict AD progression.

B. Progression Models for AD

Different from the above classification methods, in [16], the authors use a graph convolutional network to assess skeleton-based human behavior and subsequently track the progression of AD. But this kind of monitoring can only judge the state of patients from the perspective of action, there are still shortcomings, since the symptoms of AD patients occur before the behavioral abnormality [17]. In [18], the authors propose a generalized training rule for long short-term memory (LSTM). This method only focuses on predicting progression of biomarkers of AD patients. However, only one biomarker cannot accurately measure the state of AD patients. In contrast, in our approach MAGPP, we use two kinds of cognitive scores to measure the state of AD and simultaneously explore the correlation between the cognitive scores and the MRI features. The limited interpretability of LSTM is another possible drawback. The high interpretability of MAGPP not only reveals the complex temporal relation between tasks, but also enables us to investigate the temporal pattern of selected important features, which has the potential to improve understanding of AD.

As discussed, although much effort has been dedicated to AD study, the noted methods suffer from the several limitations. ① The MTL models based on fixed temporal relation structure [4], [5], [2], [6] are rigid and insufficient to capture the complicated temporal relation among tasks. ② Classification methods [3], [13], [15] make progress in diagnosing AD but fail to predict AD development. ③ The deep methods based on graph convolutional network [16] or LSTM [18] have limited ability to directly measure the state of AD patients and are not capable to explore the relation between different kinds of cognitive scores and MRI features.

III. Methods

A. Multi-task Learning

Given m tasks, each task $i \in \mathbb{N}_m$ has a set of samples $(X_i, \boldsymbol{y_i})$, where $X_i \in \mathbb{R}^{n_i \times d}, \boldsymbol{y_i} \in \mathbb{R}^{n_i}$. $X = [X_1, \cdots, X_m], Y = [\boldsymbol{y_1}, \cdots, \boldsymbol{y_m}]$, $W = [\boldsymbol{w_1}, \cdots, \boldsymbol{w_m}] \in \mathbb{R}^{d \times m}$ is model coefficient matrix. We minimize the empirical risk to learn the m tasks concurrently: $\min_W L(W) + \Omega(W)$, where $\Omega(W)$ is the penalty, $L(W)$ is the empirical loss. We use the square loss to fit the relation between X and Y.

Fig. 1 is the illustration of model. Each time point concerns a prediction of a single task. For the i-th task, each row in X_i represents all features of one patient. One MRI feature is represented by each column of X_i at the baseline time point. The cognitive score at each time point is represented by a column of $Y = [\boldsymbol{y_1}, \cdots, \boldsymbol{y_t}]$. We have total 6 time points, every time point corresponds to a task for predicting disease progression. The notation "Mx" denotes x months after the baseline time point (BL, M00). When modeling disease progression using a MTL approach, the following two major challenges need to be solved: ① How are the tasks related to one another? ② Which concrete method should be used to capture such task relation?

To address the challenging problems, we propose the following novel mechanism, termed ***Auto**matic **T**emporal relation **G**raph* (AutoTG), to automatically capture the complex temporal relation among tasks rather than using several predefined temporal relation structures [4], [5], [2], [6].

B. Automatic Temporal Relation Graph

We start with the widely used temporal smoothness assumption [4], [5], [19], [10], which assumes every time point is similar to its adjacent time points. If every task concerns a prediction of a time point, every task has a trend to be similar to its neighboring tasks, i.e., $\mathbf{w_k} \approx \mathbf{w_{k+1}}$. To achieve this goal, the models based on temporal smoothness usually penalize the difference between two successive tasks $\|\mathbf{w_k} - \mathbf{w_{k+1}}\|_2^2$ [20], [10] or $\sum_k |w_{i,k} - w_{i,k+1}|$ [5], [6]. Despite that many experiments have proved that the introduction of temporal smoothness can effectively enhance the model performance, it is actually only a *local* and *predefined* temporal relation.

To make our statement clear, we explain this temporal relation from the perspective of graph theory. In [4], [5], [21], they consider total six time points and each time point corresponds to a task. If we view each task as a node, the temporal relation between a pair of nodes is an edge, so all tasks and their temporal relation form a graph. However, the adjacency matrix R of temporal smoothness relation graph is *a fixed and symmetric tridiagonal matrix* as

$$r_{i,j} \overset{\text{predefine}}{=\!=\!=} \begin{cases} 1, \forall i = j+1 \ \& \ \forall i = j-1 \\ 0, \text{otherwise} \end{cases}$$

This structure at has least three limitations. ① Every task is only related to its adjacent tasks, potentially missing helpful and informative relation with other tasks. ② The weights of temporal relations are fixed, which is not sufficient and flexible to capture the complex temporal relation between tasks. ③ The weights of temporal relations are also identical, which is not appropriate in terms of the asymmetry in time.

Motivated by the discussion above, first of all, we allow that each task can be connected to every other task, and the weight of temporal relation can be learned directly and automatically from every given dataset, rather than predefined. So we write this type of temporal relation mathematically as

$$\begin{aligned} \mathbf{w_k} \approx \ & r_{1,k}\mathbf{w_1} + \cdots + r_{k-1,k}\mathbf{w_{k-1}} \\ & + r_{k+1,k}\mathbf{w_{k+1}} + \cdots + r_{m,k}\mathbf{w_m}. \end{aligned}$$

Clearly, as shown in Fig. 1, $\mathbf{w_k}$ is related to all other tasks $\mathbf{w_i}, \forall i \neq k$. The weight of temporal relation $r_{x,k}$ (the relation from task $\mathbf{w_k}$ to $\mathbf{w_x}$) is not fixed yet and needs to be learned from data. Another important point is that in this structure, the temporal relation is not symmetric as predefined [4], [5], [2], since we do not constrain $r_{x,k} = r_{k,x}$. In fact, this asymmetry corresponds to the real-life temporal relation. For instance, $r_{k-1,k}$ represents analyzing the past state of one patient in the current k-th time point, whereas $r_{k,k-1}$ represents predicting future state from $(k-1)$-th time point. They have completely different meanings in practice and should be allowed to have different values, rather than being predefined as the same value which is too strict in real-life applications.

Not only that, we do not assume that tasks are necessarily similar to others, i.e., we do not constrain $r_{x,k} \geqslant 0$. In fact, as the results show in Section V, we found that sometimes if two tasks that are too far apart can have a slightly negative relation with $r_{x,k} < 0$, i.e., they slightly repel, rather than approximate each other. This phenomenon has never been considered in all existing works such as [2], [4], [5], [6].

We Integrate temporal relation between all tasks to have

$$W \approx W \begin{bmatrix} 0 & r_{1,2} & \cdots & r_{1,m} \\ r_{2,1} & 0 & \cdots & r_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ r_{m-1,1} & r_{m-1,2} & \cdots & r_{m-1,m} \\ r_{m,1} & r_{m,2} & \cdots & 0 \end{bmatrix} = WR, \quad (1)$$

where R is the adjacency matrix of the temporal relation graph between tasks.

Based on above description, we propose a novel mechanism, termed ***Auto**matic **T**emporal relation **G**raph* (AutoTG), to automatically capture the complex temporal relation among tasks, and construct it as a temporal graph adjacency matrix:

$$\min_{W,R} \frac{1}{2} \sum_{i=1}^{m} \|X_i \mathbf{w_i} - \mathbf{y_i}\|_2^2 + \lambda_1 \|W - WR\|_F^2 + \lambda_2 \|R\|_{1,1}, \quad \text{s.t.} \quad r_{i,i} = 0, i \in \mathbb{N}_m. \quad (2)$$

The first penalty $\|W - WR\|_F^2$ is applied to chase the complex temporal relation among all tasks. We use the second penalty $\|R\|_{1,1}$ to encourage only the tasks that are most pertinent to share common temporal information.

In order to constrain $r_{i,i} = 0$, we need to penalize the main diagonal elements of R much more heavily than other entries. So we introduce the auxiliary matrix S which is formulated as $S = (s-1) \cdot I_{m \times m} + \mathbf{1}_{m \times m}$. The optimization problem (2) becomes

$$\min_{W,R} \frac{1}{2} \sum_{i=1}^{m} \|X_i \mathbf{w_i} - \mathbf{y_i}\|_2^2 + \lambda_1 \|W - WR\|_F^2 + \lambda_2 \|R \odot S\|_{1,1}. \quad (3)$$

We just need to give s an enough large number to constrain $r_{i,i} = 0$ for $i \in \mathbb{N}_m$. In our experimental setting, we let $s = 10^9$ to achieve the constraint of $r_{i,i} = 0$. Please refer to Section V for more detailed information. We conclude that introducing S will not increase the computational complexity of the associated optimization problem.

C. A Novel Multi-task Learning Formulation

In the area of AD research, finding the biomarkers associated with AD progression is crucial, so we utilize the group Lasso to choose a universal set of biomarkers for all tasks. The group Lasso constraint, however, fails to select particular feature sets for each task. Then, we use the Lasso to add sparsity to the matrix of model coefficients. The sparse group Lasso $\beta\|W^T\|_{2,1} + \alpha\|W^T\|_{1,1}$, the mixture of L_1-norm and $L_{2,1}$-norm, introduces sparsity into both group and within-group levels, as illustrated in 1. In the context of AD study, it promotes choosing a particular MRI feature set for each task as well as selecting a universal MRI feature set for all tasks [19], [6]. Then the proposed novel mechanism AutoTG is applied to capture the temporal task relation automatically.

After integrating AutoTG with sparse group Lasso, we present a novel approach, termed **M**ulti-task learning with **A**utomatic temporal relation **G**raph for **P**redicting Alzheimer's disease **P**rogression (MAGPP). The mathematical formulation of MAGPP is defined as

$$\min_{W,R} \frac{1}{2} \sum_{i=1}^{m} \|X_i \mathbf{w_i} - \mathbf{y_i}\|_2^2 + \lambda_1 \|W - WR\|_F^2 + \lambda_2 \|R \odot S\|_{1,1} + \lambda_3 \|W^T\|_{2,1} + \lambda_4 \|W^T\|_{1,1}. \quad (4)$$

$\lambda_1, \lambda_2, \lambda_3, \lambda_4$ are all fine-tuned hyperparameters. The AutoTG part of two penalties $\lambda_1\|W - WR\|_F^2 + \lambda_2\|R \odot S\|_{1,1}$ is applied to automatically capture the complex temporal relation among tasks. The sparse group Lasso part of two penalties $\lambda_3\|W^T\|_{2,1} + \lambda_4\|W^T\|_{1,1}$ is employed to conduct feature selection at both group and within-group levels.

IV. Optimization Algorithm

Note that the objective function (4) is not easy to solve, since it is nonsmooth and biconvex. In this section, we first introduce the whole alternating optimization for solving (4). Then we show how to customize the accelerated proximal gradient method (APM) [22] to solve the associated two subproblems about W and R with high efficiency.

The alternating optimization is widely used for solving the biconvex objective function [23]. We conclude the overall alternating optimization algorithm for solving our proposed MAGPP in Algorithm 1. The procedure is stopped when the relative changes in W and R between two successive iterations ΔW and ΔR are both not bigger than the threshold τ.

A. Accelerated Proximal Gradient Method

To update W and R efficiently, we use the accelerated proximal gradient method (APM). Because of the fastest convergence rate for the class of first-order methods, APM has been widely used to address issues with MTL [24]. It has the form

$$\min_{W} F(W) = f(W) + g(W), \quad (5)$$

Algorithm 1 Alternating Optimization for MAGPP.

Input: $X, Y, \lambda_1, \lambda_2, \lambda_3, \lambda_4, s, \epsilon$.
Output: W, R

```
Initialize: W = 0, R = 0.
for k = 1 to ··· do
    Fix R, update W.
    Fix W, update R.
    if then ΔW ⩽ τ and ΔR ⩽ τ
        break
    end if
end for
```

where $f(W)$ is smooth and convex, and $g(W)$ is nonsmooth and convex. APM is built on two sequences, the search point $\{S^k\}$ and the approximation point $\{W^k\}$. S^k is a linear combination of W^{k-1} and W^k. $S^{k+1} = W^k + \alpha_k(W^k - W^{k-1})$, where α_k is the combination coefficient. The approximation point W^k is computed as

$$W^k = \pi(S^k - \eta_k \nabla f(S^k)), \tag{6}$$

where η_k is step size, $\pi(V)$ is the proximal operator of V. We follow the line search schemes [25] to estimate η_k.

Emphasize that the computation of the proximal operator (6) is the crucial step in using APM. The complexity for solving (6) dominates the whole complexity of APM-based algorithms. As usual, the proximal operator of the nonsmooth part is not easy to solve, e.g., [5], [6]. However, in our proposed novel MAGPP (4), we will show no matter updating W or R, the proximal operators admit a closed-form solution, which enables to design an efficient algorithm.

B. Fix R, Update W

For updating W, we fix R. In order to find the proximal operator of $\lambda_3\|W^T\|_{2,1} + \lambda_4\|W^T\|_{1,1}$, we need to solve

$$\arg\min_V \frac{1}{2}\|V - W\|_F^2 + \lambda_3\|V^T\|_{2,1} + \lambda_4\|V^T\|_{1,1}. \tag{7}$$

According to [26], the complexity to get the closed-form solution of (7) is only $\mathcal{O}(md)$, so we can update W efficiently.

C. Fix W, Update R

For updating R, the sub-optimization problem is

$$\min_R \lambda_1\|W - WR\|_F^2 + \lambda_2\|R \odot S\|_{1,1}. \tag{8}$$

To obtain the proximal operator of $\lambda_2\|R \odot S\|_{1,1}$, we solve

$$\pi(R) = \arg\min_Q \frac{1}{2}\|Q - R\|_F^2 + \lambda_2\|R \odot S\|_{1,1}. \tag{9}$$

Clearly, (9) is an extension of Lasso problem, we also apply soft-thresholding method to arrive the closed-form solution. It means We only need the complexity of $\mathcal{O}(m^2)$ to solve (8).

D. Complexity Analysis

For simplicity, we make an assumption that each task has identical n training samples.

1) The Complexity of Updating W: When optimizing W, each iteration needs to compute the gradient of the smooth part which is $\frac{1}{2}\sum_{i=1}^m \|X_i\mathbf{w_i} - \mathbf{y_i}\|_2^2 + \lambda_1\|W - WR\|_F^2$ and the proximal operator of $\lambda_3\|W^T\|_{2,1} + \lambda_4\|W^T\|_{1,1}$. The complexity for computing the gradient is $\mathcal{O}(nmd + m^2(m+d))$. Here we emphasize that in our implementation MATLAB code, we compute the loss part $\mathcal{L}(W)$ parallelly with the complexity of $\mathcal{O}(nd)$, so the complexity of every iteration reduces to $\mathcal{O}(nd + m^2(m+d))$. The cost for computing the proximal operator of $\lambda_3\|W^T\|_{2,1} + \lambda_4\|W^T\|_{1,1}$ is $\mathcal{O}(md)$. The convergence rate of APM is proved to be $\mathcal{O}(1/\sqrt{\epsilon})$ iterations for a desired accuracy ϵ [27], so the overall complexity for updating W is $\mathcal{O}\left((nd + m^3 + m^2d)/\sqrt{\epsilon}\right)$.

2) The Complexity of Updating R: When optimizing R, each iteration needs to compute the gradient of smooth part $\lambda_1\|W - WR\|_F^2$ and the proximal gradient of nonsmooth part $\lambda_2\|R \odot S\|_{1,1}$. The complexity for computing the gradient is $\mathcal{O}(m^2d)$. The cost for computing the proximal operator of $\lambda_2\|R \odot S\|_{1,1}$ is $\mathcal{O}(m^2)$. So for updating R, each iteration has the complexity of $\mathcal{O}(m^2d)$. So the overall complexity for updating R is $\mathcal{O}\left(m^2d/\sqrt{\epsilon}\right)$.

3) The Overall Complexity: In Algorithm 1, W and R will be updated once each, which counts as a full iteration. Therefore, a full iteration has the complexity of

$$\mathcal{O}\left(\frac{nd + m^2(m+d)}{\sqrt{\epsilon}}\right).$$

V. Experimental Result

The experiment hardware is an Apple M1 Max chip with 32 GB memory. The implementation code runs on MATLAB and can be found at https://github.com/menghui-zhou/MAGPP.

A. The Latest Dataset from ADNI

The Alzheimer's Disease Neuroimaging Initiative (ADNI) database [8] supports research efforts [4], [5], [2] and is the data source for this study. In ADNI, the baseline (BL) marks the initial patient screening, serving as a time point reference."M12," for example, signifies a year after baseline. Some patients have follow-up data up to 120 months, but many leave the study. Due to limited data at later points, we follow prior methods [4], [5], [2] by using only the first six time points. This paper employs MMSE and ADAS-Cog to measure AD cognitive state. Data preprocessing aligns with previous work [4], [5], [2], [6]. Ultimately, we extract 314 features, with further details in Table I.

TABLE I
THE SPECIFIC DETAILS OF THE SAMPLE NUMBER AT EACH TIME POINT IN THE SEQUENCE ON MMSE AND ADAS-COG DATASETS.

Time point	M00	M06	M12	M24	M36	M48
MMSE	1092	1078	1027	883	579	494
ADAS-Cog	1074	1064	1014	867	556	483

B. Empirical Evaluation

In this section, we thoroughly assess the efficacy of our proposed MAGPP in comparison to several baseline methods. We randomly select β of the dataset as the training set, where the training ratio $\beta \in \{0.4, 0.6, 0.8\}$ and the rest is divided randomly and equally into validation set and test set. We repeat 5 trials. In each trial, we train the model on the training set and use the validation set to select the best hyperparameters $\lambda_1, \lambda_1, \lambda_3, \lambda_4 \in \{10^0, 10^1, 10^2, 10^3\}$, the pseudo hyperparameter s is set as 10^9. The feature matrix X is normalized.

1) Evaluation Metrics: We use the Root Mean Squared Error (rMSE) for task-specific regression performance. Additionally, we measure overall performance across all tasks using the weighted R-value (wR) and the normalized mean squared error (nMSE), both of which are frequently used in the MTL literature [5], [6]. Higher performance is indicated by lower nMSE and rMSE or higher wR.

2) Comparative Models and Ablation Experiments: We thoroughly contrast our MAGPP with a number of MTL baseline techniques. All comparative models include TGL [4], cFSGL [5], VSTG [28], FLSGL [2], and LSA [6]. In addition, in order to further demonstrate the superiority of our algorithm, we also compare the performance of the neural network based method, LSTM (Long Short Term Memory). The number of training iterations was 1000 epochs. In multiple training iterations, we train the model using the Adam optimizer, set rMSE as the loss function, and the batch size is 2. The learning rate starts at 0.0001. Since LSTM does not allow the patient to have missing cognitive score at specific time points, after we keep all the patient data with cognitive scores at six time points, MMSE and ADAS-Cog datasets have 331 and 365 samples, respectively. Given that MAGPP consists of two components—AutoTG and sparse group Lasso (SGLasso), we validate the efficiency of both AutoTG and SGLasso on two AD datasets, reaffirming the effectiveness of MAGPP.

Table II presents the results, revealing inferior performance of LSTM across all cases possibly attributed to inadequate training data (around 300 samples). SGLasso exhibits suboptimal performance due to its disregard of inter-task relations. This underscores the significance of incorporating task relations in the study of AD progression. VSTG does not perform well on two datasets. The possible reason is that VSTG is capable of feature selection, but the low-rank task relation based on the k-support norm is not a great choice for the case of the progression of AD. The poor performance of FLSGL suggests that using a specific exponential format to capture the temporal relation between tasks is insufficient. TGL also performs poorly, owing to the fact that it does not introduce sparsity within and between groups as cFSGL does. It constrains all tasks to share a single feature set, which is overly restrictive in practice. AutoTG surpasses SGLasso, yet falls short of the performance of MAGPP. This discrepancy highlights the value of introducing sparsity within and between groups for effective feature selection. Overall, MAGPP

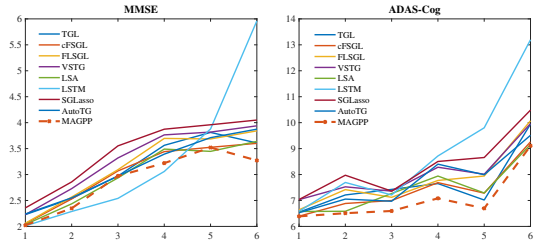

Fig. 2. The comparison of single task performance, between our MAGPP and several baseline methods.

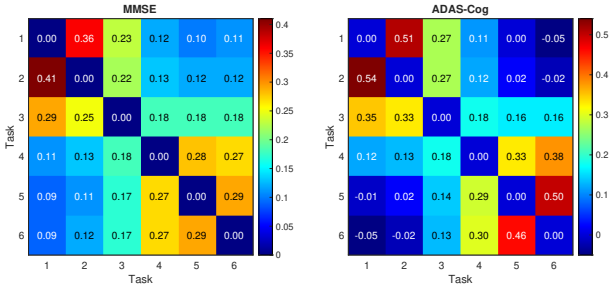

Fig. 3. The adjacency matrix of temporal relation graph between tasks, which MAGPP automatically learns from the MMSE and ADAS-Cog datasets.

demonstrates superior efficacy, except in instances such as the ADAS-Cog dataset with $\beta = 0.8$, where cFSGL marginally outperforms it by a small metric margin ($wR = 0.756$ for cFSGL versus $wR = 0.755$ for MAGPP).

Apart from assessing overall model performance, we also evaluate efficacy of MAGPP at individual time points. Due to space constraints, we only present results for a training ratio of $\beta = 0.8$ in Fig. 2. Notably, other β cases yield similar outcomes. Across diverse datasets, LSTM consistently demonstrates the poorest performance. Conversely, SGLasso consistently ranks second worst at each time point, underscoring the imperative of integrating task relationships in AD progression. In contrast, AutoTG delivers intermediate performance compared to alternative methods, highlighting the necessity of feature selection in AD progression. The performance of MAGPP emerges as consistently superior at individual time points, regardless of the dataset.

C. Visualization of Temporal Relation

For a comprehensive analysis of inter-task temporal relations, we visualize the adjacency matrix R captured by MAGPP, focusing on a training ratio of $\beta = 0.8$. From the findings in Fig. 3, MAGPP unveils both shared patterns and distinctions in its learned temporal relations across the two datasets. Notably, the adjacency matrices are not strictly symmetrical, mirroring real-world temporal dynamics. For example, $r_{k-1,k}$ signifies an evaluation of previous state of patient at the current time point k, while $r_{k,k-1}$ entails predicting the future state from the $(k-1)$-th time point. These divergent practical implications warrant distinct values rather than a predetermined uniformity.

In both datasets, most tasks are closely connected to their adjacent tasks. The strong connectivity within the ADAS-Cog

TABLE II

THREE DIFFERENT TYPES OF COGNITIVE SCORES ARE USED. THE AVERAGE NMSE AND WR OVER 5 REPETITIONS ARE DISPLAYED IN THE RESULTS. THE BOLD FONT HIGHLIGHTS THE STATISTICALLY SUPERIOR MODELS. SGLASSO AND AUTOTG ARE THE TWO PARTS OF MAGPP.

Ratio β	Metric	TGL	cFSGL	FL-SGL	VSTG	LSA	LSTM	SGLasso	AutoTG	MAGPP
Dataset: MMSE										
0.4	nMSE	**0.620**	0.631	0.651	0.649	0.630	0.850	0.671	0.642	0.624
	wR	0.616	0.610	0.594	0.588	0.610	0.438	0.590	0.601	**0.619**
0.6	nMSE	0.621	0.597	0.639	0.659	0.601	0.860	0.666	0.638	**0.585**
	wR	0.618	0.632	0.607	0.593	0.619	0.444	0.599	0.607	**0.636**
0.8	nMSE	0.602	0.583	0.626	0.641	0.579	0.756	0.653	0.617	**0.567**
	wR	0.631	0.650	0.619	0.608	0.650	0.562	0.612	0.629	**0.663**
Dataset: ADAS-Cog										
0.4	nMSE	0.494	0.490	0.511	0.527	0.493	0.870	0.526	0.498	**0.485**
	wR	0.717	0.730	0.700	0.691	0.734	0.493	0.673	0.695	**0.740**
0.6	nMSE	0.482	0.470	0.491	0.500	0.463	0.796	0.519	0.487	**0.460**
	wR	0.729	0.747	0.713	0.698	**0.749**	0.513	0.685	0.709	0.748
0.8	nMSE	0.471	0.459	0.474	0.483	0.463	0.703	0.505	0.476	**0.453**
	wR	0.734	**0.756**	0.729	0.711	0.753	0.591	0.698	0.717	0.755

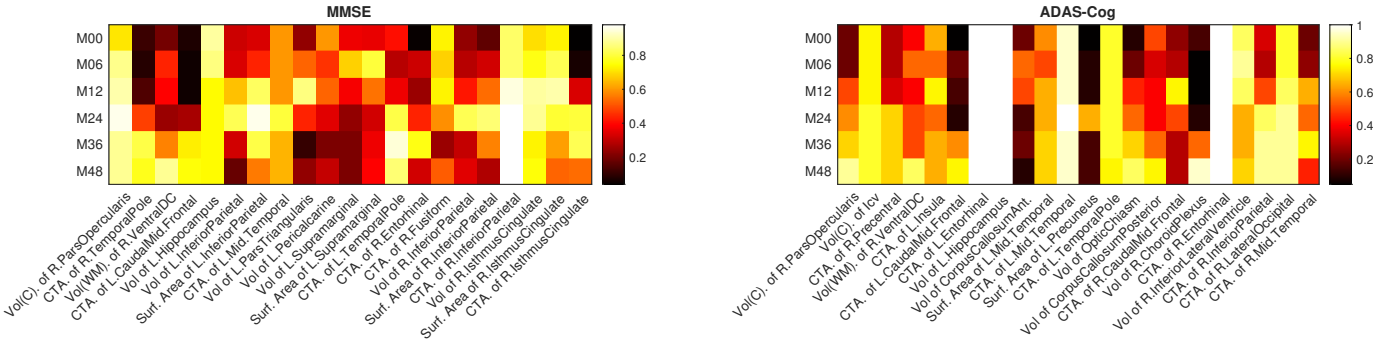

Fig. 4. The stability vector of table MRI features using MAGPP on two datasets. We choose top 6 stable features on each time point, finally we get 21 stable features on MMSE dataset and 22 stable features on ADAS-Cog dataset.

dataset results from its extensive 11-item cognitive tests, tailored to gauge changes in AD severity. This strong connection aligns with prior research [4], [5] advocating local temporal relations in AD studies, considering its status as a prevalent chronic condition. However, within the MMSE dataset, inter-neighboring task relationships are notably weaker. Specifically, the relation weight between the 3rd and 4th tasks is merely 0.18. A plausible explanation lies in the primary role of MMSE as a widely-used scale for gauging AD-related cognitive decline. However, compared to ADAS-Cog, MMSE lacks detailed evaluations of memory, executive function, and language ability. Notably, in ADAS-Cog dataset, certain temporal relation weights are negative. For instance, task 1 and tasks 5 and 6 have relation weights of -0.01 and -0.05 respectively. This indicates that when time intervals are considerable, corresponding tasks might become less similar and exhibit some repulsion. Importantly, this phenomenon has been overlooked in prior studies [2], [4], [5], [6].

It is concluded that the temporal relation learned by MAGPP shows that ① in AD progression, the temporal relation between tasks is asymmetric and global, which proves the deficiency of using local temporal relation in previous works [4], [5]. ② The temporal relation between tasks is extremely complex, which indicates that the previous works use a predefined Gaussian kernel method [2] or a iterative convex structure [6] can not fully capture the complex task relation. ③ All existing works [4], [5], [2], [6] do not consider the negative temporal relation.

D. Temporal Pattern of Stable Biomarkers

MAGPP offers the advantage of analysing temporal MRI feature patterns for better understanding of AD progression. To explore discovered MRI biomarkers, we employ the longitudinal stability selection method [19], used in prior studies [2], [6]. The term "stability vector" denotes the calculated frequency vector.

To begin, we notice that the volume of the left hippocampus (Vol. of L.Hippocampus) is considered a stable biomarker in all datasets, particularly in ADAS-Cog dataset, where the volume of the left hippocampus is selected to be stable the biomarkers with the probability close to 1. In MMSE dataset, the volume of the left hippocampus is selected to stable the biomarkers with a probability greater than 0.8. This is in line with other AD studies [6] because it has long been known that the hippocampus plays a key role in the development of AD. There are many different discoveries between the two datasets.

In ADAS-Cog dataset, we also discover that the cortical thickness average of the left entorhinal (CTA. of L. Entorhinal)

and the cortical thickness average of the right entorhinal (CTA. of R. Entorhinal) are both chosen as stable biomarkers with a probability close to 1. The most stable biomarker in the MMSE dataset is the volume of right IsthmusCingulate (Vol of R.IsthmusCingulate). However, in ADAS-Cog dataset, the volume of right IsthmusCingulate only shows stability in the last few moments from M24 to M48. The cortical thickness average of middle temporal (CTA. of Mid. Temporal) always has a high selection frequency of about 0.7 in the MMSE dataset, in ADAS-Cog dataset, it is selected as a stable biomarker with a higher frequency, close to 0.9.

The distinct temporal patterns of the stable biomarkers of two cognitive scores also suggest that it may be less effective to confine the model to a shared set of features [5], [2], [6].

VI. Conclusion

In this study, we introduce AutoTG, a novel method to capture intricate task temporal relations via a graph adjacency matrix. Our approach MAGPP combines sparse group Lasso and AutoTG, outperforming baselines for overall and individual task performance. The nonsmooth, biconvex objective function is tackled through customized alternating optimization and an accelerated proximal gradient method. The graph adjacency matrix of MAGPP visualizes complex temporal relations. The results reveal asymmetry, indicating distinct tasks even with negative weights. This insight furthers our understanding of AD progression. We employ stability selection to identify stable MRI features, enhancing interpretability. This approach aids in discovering potential new biomarkers.

As MAGPP is a general method for modelling disease progression, in the future, we hope to investigate the efficacy in a border area like Parkinson's disease [9] and diabetes [10].

References

[1] John Wiley. Alzheimer's disease facts and figures. *Alzheimers Dement*, 17:327–406, 2021.

[2] Xiaoli Liu, Peng Cao, André R Gonçalves, Dazhe Zhao, and Arindam Banerjee. Modeling alzheimer's disease progression with fused laplacian sparse group lasso. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 12(6):1–35, 2018.

[3] Chunfeng Lian, Mingxia Liu, Jun Zhang, and Dinggang Shen. Hierarchical fully convolutional network for joint atrophy localization and alzheimer's disease diagnosis using structural mri. *IEEE transactions on pattern analysis and machine intelligence*, 42(4):880–893, 2018.

[4] Jiayu Zhou, Lei Yuan, Jun Liu, and Jieping 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*, pages 814–822, 2011.

[5] Jiayu Zhou, Jun Liu, Vaibhav A Narayan, and Jieping Ye. Modeling disease progression via fused sparse group lasso. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1095–1103, 2012.

[6] Menghui Zhou, Yu Zhang, Tong Liu, Yun Yang, and Po Yang. Multi-task learning with adaptive global temporal structure for predicting alzheimer's disease progression. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 2743–2752, 2022.

[7] Yu Zhang and Qiang Yang. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, 2021.

[8] Michael W Weiner, Paul S Aisen, Clifford R Jack Jr, William J Jagust, John Q Trojanowski, Leslie Shaw, Andrew J Saykin, John C Morris, Nigel Cairns, Laurel A Beckett, et al. The alzheimer's disease neuroimaging initiative: progress report and future plans. *Alzheimer's & Dementia*, 6(3):202–211, 2010.

[9] Saba Emrani, Anya McGuirk, and Wei Xiao. Prognosis and diagnosis of parkinson's disease using multi-task learning. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1457–1466, 2017.

[10] Luca Romeo, Giuseppe Armentano, Antonio Nicolucci, Marco Vespasiani, Giacomo Vespasiani, and Emanuele Frontoni. A novel spatio-temporal multi-task approach for the prediction of diabetes-related complication: a cardiopathy case of study. In *IJCAI*, pages 4299–4305, 2020.

[11] Ping Wang, Tian Shi, and Chandan K Reddy. Tensor-based temporal multi-task survival analysis. *IEEE Transactions on Knowledge and Data Engineering*, 2020.

[12] Noah Simon, Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A sparse-group lasso. *Journal of computational and graphical statistics*, 22(2):231–245, 2013.

[13] Changqing Zhang, Ehsan Adeli, Tao Zhou, Xiaobo Chen, and Dinggang Shen. Multi-layer multi-view classification for alzheimer's disease diagnosis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[14] Yuanyuan Chen and Yong Xia. Iterative sparse and deep learning for accurate diagnosis of alzheimer's disease. *Pattern Recognition*, 116:107944, 2021.

[15] Wen Yu, Baiying Lei, Shuqiang Wang, Yong Liu, Zhiguang Feng, Yong Hu, Yanyan Shen, and Michael K Ng. Morphological feature visualization of alzheimer's disease via multidirectional perception gan. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[16] XB Bruce, Yan Liu, Keith CC Chan, Qintai Yang, and Xiaoying Wang. Skeleton-based human action evaluation using graph convolutional network for monitoring alzheimer's progression. *Pattern Recognition*, 119:108095, 2021.

[17] Fatih Altay, Guillermo Ramón Sánchez, Yanli James, Stephen V Faraone, Senem Velipasalar, and Asif Salekin. Preclinical stage alzheimer's disease detection using magnetic resonance image scans. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 15088–15097, 2021.

[18] Mostafa Mehdipour Ghazi, Mads Nielsen, Akshay Pai, M Jorge Cardoso, Marc Modat, Sébastien Ourselin, Lauge Sørensen, Alzheimer's Disease Neuroimaging Initiative, et al. Training recurrent neural networks robust to incomplete data: application to alzheimer's disease progression modeling. *Medical image analysis*, 53:39–46, 2019.

[19] Jiayu Zhou, Jun Liu, Vaibhav A Narayan, Jieping Ye, Alzheimer's Disease Neuroimaging Initiative, et al. Modeling disease progression via multi-task learning. *NeuroImage*, 78:233–248, 2013.

[20] Jiayu Zhou, Jianhui Chen, and Jieping Ye. Malsar: Multi-task learning via structural regularization. *Arizona State University*, 2011.

[21] Qiang Zhou, Gang Wang, Kui Jia, and Qi Zhao. Learning to share latent tasks for action recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2264–2271, 2013.

[22] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2003.

[23] Yaqiang Yao, Jie Cao, and Huanhuan Chen. Robust task grouping with representative tasks for clustered multi-task learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1408–1417, 2019.

[24] Menghui Zhou, Yu Zhang, Yun Yang, Tong Liu, and Po Yang. Robust temporal smoothness in multi-task learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 11426–11434, 2023.

[25] Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[26] Jun Liu, Shuiwang Ji, and Jieping Ye. Multi-task feature learning via efficient l2, 1-norm minimization. *arXiv preprint arXiv:1205.2631*, 2012.

[27] Huan Li, Cong Fang, and Zhouchen Lin. Accelerated first-order optimization algorithms for machine learning. *Proceedings of the IEEE*, 108(11):2067–2082, 2020.

[28] Jun-Yong Jeong and Chi-Hyuck Jun. Variable selection and task grouping for multi-task learning. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1589–1598, 2018.