Article:
Zhou, M., Wang, X., Liu, T. et al. (2 more authors) (2024) Integrating visualised automatic temporal relation graph into multi-task learning for Alzheimer's disease progression prediction. IEEE Transactions on Knowledge and Data Engineering, 36 (10). pp. 5206-5220. ISSN 1041-4347

https://doi.org/10.1109/TKDE.2024.3385712

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

Menghui Zhou, Xulong Wang, Tong Liu, Yun Yang, and Po Yang*, *Senior Member, IEEE*

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 in order 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 non-smooth objective function, we adopt the alternating optimization and show that the two related sub-optimization 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 any first-order method. We have preprocessed three 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 visualise the automatically learned temporal relation graph and investigate the temporal patterns of the important MRI features. The implementation source can be found at https://github.com/menghui-zhou/MAGPP.

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

1 Introduction

Alzheimer's disease (AD), the most prevalent neurodegenerative disorder, is marked by the deterioration of cognitive abilities over time. AD accounts for 60% to 80% of dementia cases and eventually leads to irreversible neuronal loss and death [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 the progression of AD, it is hence crucial to identify the biomarkers that are significant to the emergence of this illness [1].

Previous studies have demonstrated that a variety of cognitive scores, such as ADAS-Cog (the Alzheimer's Disease Assessment Scale Cognitive Sub-scale), RAVLT (the Rey Auditory Verbal Earning Test), and MMSE (the Mini-Mental State Examination), are capable to assess the state of AD patients [2], [3], [4]. Non-invasive structural magnetic resonance imaging (MRI) can identify atrophic changes in the brain [5]. Machine learning techniques have been used to investigate the relationship between various cognitive scores and MRI features. Owing to an inherent relationship among multiple time points, it is expected that analysing multiple time points simultaneously will improve model performance. To achieve this goal, in recent years, several multi-task learning strategies have been put forth to forecast how AD will develop [6], [7], [8], [9]. They consider predicting a target at a series of time points to be a multi-task learning 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 multi-task learning is to improve generalization ability and model performance by utilizing the inherent relations between various related tasks [10]. Despite the recent great advancements made in investigating AD through multi-task learning, 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. [6] propose a multi-task learning method with the temporal group Lasso (TGL) and assume that the cognitive score of patients will not change significantly over time, i.e., there won't be much of a difference in cognitive scores between two successive time points. TGL penalizes the difference between adjacent tasks $\|\mathbf{w_k} - \mathbf{w_{k+1}}\|_2^2$ in order to achieve temporal smoothness at task level. Similar to TGL, a multi-task learning formulation with convex sparse group Lasso (cFSGL) is proposed in [7] which assumes that nearby time points have similar

- *M. Zhou is with the Department of Computer Science, University of Sheffield, Sheffield S10 2TT, UK (e-mail: menghuizhoucn@gmail.com).*
- *X. Wang is with the Department of Computer Science, University of Sheffield, Sheffield S10 2TT, UK (e-mail: xl.wang@sheffield.ac.uk).*
- *T. Liu is with Department of Chemical Engineering, Imperial College London, London SW7 2AZ, UK (e-mail: tliu.soton@gmail.com).*
- *Y. Yang is with the Department of Software, Yunnan University, Kunming 674199, China (e-mail: yangyun@ynu.edu.cn).*
- *P. Yang is with the Department of Computer Science, University of Sheffield, Sheffield S10 2TT, UK (e-mail: po.yang@sheffield.ac.uk).*

Corresponding author: Po Yang (e-mail: po.yang@sheffield.ac.uk).

Manuscript received April 20, 2023; Revised XXX; Accepted XXXX.

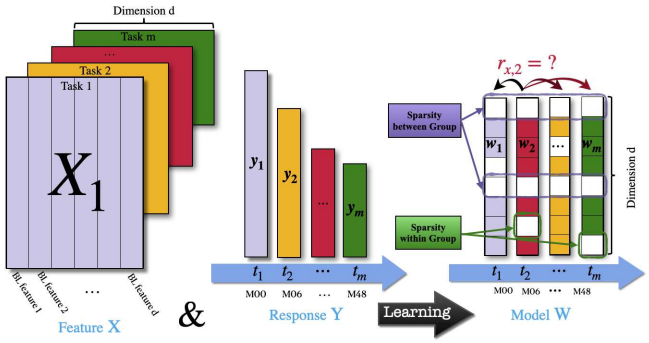

Fig. 1. Illustration of MTL prediction model. We use baseline MRI features to predict the progression of AD patients, whose states are measured by cognitive scores. The notation BL and M00 both mean the baseline time point. Mx means x months after baseline time point and $x \in \{0, 6, 12, 24, 36, 48\}$. The time points are not evenly distributed. Moreover, in practice, the notation Mx is usually inaccurate.

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 are a type of *local* and *predefined* structure. It only takes into account how the task relates to its neighbours, 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 utilise a graph with only edges between successive tasks, but on other edges. Different from TGL and cFSGL, Liu et al. [8] 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* kernel function. Recently Zhou et al. [9] 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 [11].

To handle this challenging 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 multi-task learning based on temporal relation is found in a vast variety of applications. Except for the study of AD, in [12], the authors use multi-task learning with temporal smoothness relation to diagonalize the progression of Parkinson's disease. Romeo et al. [13] suggest a novel spatio-temporal multi-task learning with the temporal smoothness relation to predict the development of diabetes and its complications. Wang et al. [14] propose a temporal multi-task learning model for survival analysis. Though this paper focuses on the study of AD, AutoTG has great potential to be a building block for other multi-task learning 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 [15] 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. Combining 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 the main contributions as follows:

- We point out that existing multi-task works for AD progression prediction all rely on strong assumptions and employ specific predefined structures to mine temporal relationships among tasks. They disregard the possibility of a complex asymmetric relationship and negative correlation between tasks, thereby limiting model performance.
- 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 three latest AD datasets show that MAGPP outperforms several baseline methods in terms of overall performance and nearly every task-specific performance.
- To solve the non-smooth and biconvex objective function, we utilize the widely used alternating optimization [16]. To improve the efficiency even further, we design a warm start strategy based on a variant of the Gaussian kernel. Experiments show that it can reduce iterations by up to 87% and computation time by 85%.
- To explore the complex temporal relation among tasks, we visualise 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, though not frequently, repel rather than approximate each other which has never been considered in all previous works [6], [7], [8], [9].
- To show the high interpretability of MAGPP, we utilise the method of stability selection [7] 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. $\|X\|_{p,q} = (\sum_j (\sum_i x_{i,j}^p)^{q/p})^{1/q}$. The component-wise operator $\mathrm{sgn}(\cdot)$ satisfies: $t < 0, \mathrm{sgn}(t) = -1; t = 0, \mathrm{sgn}(t) = 0$, and $t > 0, \mathrm{sgn}(t) = 1$.

Organization: The remainder of this work is structured as follows. The related work is in Section 2. In Section 3, we present the formulation of MAGPP. We go into great detail about the related optimization algorithm in Section 4. Section 5 presents the experimental findings. Sections 6 and 7 serve as the discussion and conclusion of this paper, respectively.

2 Related Work

This section includes a brief discussion of some frequently associated works, roughly divided into single task learning, multi-task learning and deep learning-based methods.

2.1 Single task learning for AD prediction

Existing single task learning problems basically include classification, survival analysis, and regression models. Single task classification model [17] attempts to group the condition of patients into various recognised disease stages, which are typically divided into AD, Mild Cognitive Impairment (MCI), and Cognitively Normal (CN). For the purpose of using structural MRI to diagnose AD and localize joint atrophy, Lian et al. [5] suggest a hierarchical fully convolutional network. Zhang et al. [18] propose a multi-layer multi-view classification strategy, with the input serving as the first layer and a latent representation built to investigate the relationship between class labels and features. Yu et al. [19] propose a tensorizing GAN with high-order pooling to make full use of the second-order statistics of the holistic MRI images and thus enhance the assessment of AD patients. Different from single task classification model, survival analysis models [20] try to answer how long the patient can live rather than the state of AD patient. Several single task regression models tend to predict the cognitive score at a single time point, e.g., baseline [21] or one year [22]. However, these single task learning methods only focus on the prediction of AD patients at a single time point. Given the intrinsic relationship between a sequence of time points, it is expected that a joint examination of all time points will improve model performance, especially when the amount of data is small and the feature dimension is high [7].

2.2 Multi-task learning for AD progression prediction

To achieve this, recently several traditional multi-task learning strategies have been put forth to forecast the progression of AD [6], [7], [8], [9]. Specifically, in order to fully capture and hence exploit the complex temporal relation between multiple tasks, Zhou et al. [6], [7] assume every task is similar to its neighbouring tasks and propose a local temporal relation structure, namely temporal smoothness relation. Then Liu et al. [8] try to extend this local structure to be adaptively global, so they propose a novel kernel function based global temporal relation structure and the underlying assumption is tasks that are further apart are considered to be less connected. Recently Zhou et al. [9] claim that when predicting AD progression, every task should be related to all previous tasks and then propose a convex combination-based global temporal relation. However, considering all aforementioned temporal relations are based on predefined structures, they have limited capability to handle complicated temporal relations among tasks in the case of AD progression.

2.3 Deep learning-based methods for AD progression prediction

In the past decade or so, deep neural networks have developed rapidly, and several works have attempted to use deep learning-based models to predict AD progression. Bruce et al. [23] use a graph convolutional network to assess skeleton-based human behaviour and subsequently track the progression of AD. However, this kind of monitoring can only analyse the motor ability of AD patients, not their cognitive states, especially considering that cognitive deterioration of AD patients occurs before the behavioural abnormality [24]. Ghazi et al. [25] propose a generalised training rule for long short-term memory (LSTM) to model the progression of AD using six volumetric MRI features.

The specialised training rule enhances to effective management of both absent predictor values and target values, consequently leading to notable improvements in overall model performance. In a similar vein, Nguyen et al. [26] introduce a minimal recurrent neural network (minimalRNN), which harnesses three distinct strategies designed to proficiently address missing data. Liang et al. [27] recently bring attention to the fact that minimalRNN frequently produces many inaccurate values at unobserved time points, potentially degrading performance. Then, based on LSTM, they propose a multi-task learning framework that can predict future AD patient progression and adaptively impute the missing data.

Although these deep learning-based approaches significantly advance AD research, a technical hurdle that prevents their widespread adoption by practitioners is the difficulty of meaningful interpretation of deep neural networks [28]. This is largely because deep network architectures and their parts are frequently created after trial and error, and then deployed as a "black box." Many of the widely used heuristic and empirical methods are developed for designing and training deep networks [29]. For any new data and tasks, practitioners are constantly faced with a number of challenges, such as which network architecture or specific components they should use. How wide or deep should the network be? etc. Not only that, but in research on healthcare, we frequently care more about why the model produces the result rather than just concentrating on the result [24], [28].

As a result, modelling and forecasting the progression of AD still benefit significantly from traditional multi-task learning with full interpretability. We understand the meaning of each model weight, allowing us to directly incorporate a fully interpretable correlation of features into multi-task learning models [30]. MRI features can also be further grouped according to human brain regions of

interest (ROIs), and the interpretable correlation of ROIS can also be introduced into the multi-task learning model [31]. Similarly, we can study the selection of MRI features with different penalties based on the high interpretability of model [3]. We can also statistically perform longitudinal stability analysis of biological features, which entails examining how the importance of each feature changes over time, similar to what has been done in earlier studies [6], [7], [8], [9].

As discussed, although much effort has been dedicated to the study of AD, the noted algorithms suffer from the following limitations and challenges:

- Numerous existing single task learning models for AD prediction [17], [5], [19], [20], [21], [22] only focus on a single time point, which hinders model performance since it ignores the intrinsic and valuable temporal relationship among a sequence of time points.
- Existing multi-task learning models with predefined temporal relation structures [6], [7], [8], [9] are promising in the field of AD progression prediction since they are capable of jointly analysing all time points simultaneously. However, these predefined structures are rigid and insufficient to adequately capture and thus utilise the intricate temporal relation among tasks.
- Many deep learning-based models for progression prediction methods [25], [26], [27] have achieved great progress in AD research field. The possible barrier is their limited interpretability, which raises a number of questions for practitioners when designing models, such as what special architectures are required for neural networks, how deep the network should be, how wide each layer should be, and so on. The limited interpretability also restricts our understanding of the model results as well as the exploration of pathological causes of AD.

3 METHODS

3.1 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 the model coefficient matrix. We minimize the following empirical risks so that we can 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:

$$L(Y, X, W) = \frac{1}{2} \sum_{i=1}^{m} \|X_i \boldsymbol{w_i} - \boldsymbol{y_i}\|_2^2.$$

Fig. 1 is the illustration of the model. Each time point concerns a prediction of a single task. For the i-th task, each row in X_i represents all the features of one patient. An 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 modelling disease progression using a multi-task learning approach, the following two major challenges need to be solved:

- How are the tasks related to one another?
- Which concrete method should be used to capture such task relation?

To address the challenging problems mentioned above, 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 graph.

3.2 Automatic Temporal Relation Graph

We start with the widely used temporal smoothness assumption [6], [7], [32], [13], 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 neighbouring 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 $\|\mathbf{w_k} - \mathbf{w_{k+1}}\|_2^2$ [33], [13] or $\sum_k |w_{i,k} - w_{i,k+1}|$ [7], [9]. 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 [6], [7], [34], they consider a total of 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 of the temporal smoothness relation graph is *fixed and symmetric tridiagonal*. This structure at has least three drawbacks as follows: ① Every task is only related to its adjacent tasks, potentially missing the 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 heterogeneity of time.

Different from the temporal smoothness relation, Liu et al. [8] propose a multi-task formulation with a type of global temporal relation which enables a fully connected graph with decreasing task weights. The underlying assumption of this approach is tasks that are further apart are considered to be less connected. They use a local approximation method based on a kernel to compute the temporal weight of the inter-task relationship in advance as

$$\mathbf{w}_t \approx \sum_{\substack{\ell=1 \\ \ell \neq t}}^{m} h_{\ell,t} \mathbf{w}_\ell, \quad h_{\ell,t} = \frac{\exp\left(-\frac{(\ell-t)^2}{\sigma^2}\right)}{\sum_{\ell' \neq t}^{m} \exp\left(-\frac{(\ell'-t)^2}{\sigma^2}\right)}. \quad (1)$$

The kernel-based temporal relation is a global structure because each task can have a relationship with any other task. This approach, however, is still based on a predefined kernel function, and the adjacency matrix is non-negative.

Compared to the above two relation structures, our previous work [9] proposes a multi-task learning with a convex combination-based global temporal relation. The intuitive idea is that when diagnosing a patient in practice, the expert should consider not only the current state of AD patient, but also all previous states. So the basic assumption is every task is related to all previous tasks and the corresponding mathematical form is

$$\mathbf{W}H(\alpha) = \mathbf{W}\Theta A_1(\alpha)A_2(\alpha)\cdots A_{t-2}(\alpha), \tag{2}$$

where the matrix $\Theta \in R^{m\times(m-1)}$ satisfies $\Theta_{ij} = 1$ if $i = j, \Theta_{ij} = -1$ if $i = j+1$, and $\Theta_{ij} = 0$ otherwise. $A_i(\alpha) \in R^{(m-1)\times(m-1)}$ is an identity matrix except that $A_{i_{m,n}}(\alpha) = \alpha$ if $m = i, n = i+1$, $A_{i_{m,n}}(\alpha) = 1-\alpha$ if $m = n = i+1$. Even though this method has achieved good performance, it is still a predefined structure which is rigid and inflexible. Moreover, this structure makes a trade-off in the temporal relationships between all tasks, which can easily lead to $\alpha = 0$ [9]. That means this temporal relation structure could easily degenerate into the temporal smoothness relation [6], [7], [12], [13].

Motivated by the discussion above, first of all, a better approach is to make no assumptions about the temporal relationships between tasks. The weight of temporal relation can be learned directly and automatically from every given dataset, rather than being 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 by temporal smoothness [6], [7] or the relation based on Gaussian kernel [8], 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 5, we found that sometimes two tasks that are too far apart will 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 [6], [7], [8], [9].

After integrating the temporal relation between all tasks, we have

$$W \approx W \begin{bmatrix} 0 & r_{1,2} & \cdots & r_{1,m} \\ r_{2,1} & 0 & \cdots & r_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ r_{m-1,1} & r_{m-1,2} & \cdots & r_{m-1,m} \\ r_{m,1} & r_{m,2} & \cdots & 0 \end{bmatrix} = WR, \tag{3}$$

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

Based on the above description, we propose the following 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:

$$\begin{aligned}\min_{W,R} \frac{1}{2}\sum_{i=1}^{m} \|X_i\mathbf{w_i} - \mathbf{y_i}\|_2^2 + \lambda_1\|W - WR\|_F^2 + \lambda_2\|R\|_{1,1}, \\ \text{s.t.} \quad r_{i,i} = 0, i \in \mathbb{N}_m.\end{aligned} \tag{4}$$

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.

We emphasize that the penalty $\|W - WR\|_{1,1}$ is an alternate option to chase the temporal relation, however, with extremely expensive computational cost. Please refer to Section 4 for the detailed discussion about the reason for using $\|W - WR\|_F^2$, rather than $\|W - WR\|_{1,1}$.

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 (4) becomes

$$\begin{aligned}\min_{W,R} \frac{1}{2}\sum_{i=1}^{m} \|X_i\mathbf{w_i} - \mathbf{y_i}\|_2^2 \\ + \lambda_1\|W - WR\|_F^2 + \lambda_2\|R \odot S\|_{1,1}.\end{aligned} \tag{5}$$

We want to emphasize that s is only a "pseudo" hyperparameter, not a hyperparameter like λ_1 and λ_2. 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 5 for more detailed information. We conclude that introducing the auxiliary matrix S will not increase the computational complexity of the associated optimization problem.

Note that the optimization problem (4) is biconvex. We can employ the alternating optimization algorithm to update both variables W and R. Based on AutoTG, we have the capability of automatically and directly learning the complex temporal relation among tasks from every specific dataset.

3.3 A Novel Multi-task Learning Formulation

In the area of AD research, finding the biomarkers associated with the progression is crucial, so we utilise the group Lasso [35] 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 $\beta\|W^T\|_{2,1}+\alpha\|W^T\|_{1,1}$ [15], 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 [32], [9]. 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

$$\begin{aligned}\min_{W,R}\ &\frac{1}{2}\sum_{i=1}^{m}\|X_i\mathbf{w_i}-\mathbf{y_i}\|_2^2+\lambda_1\|W-WR\|_F^2\\&+\lambda_2\|R\odot S\|_{1,1}+\lambda_3\|W^T\|_{2,1}+\lambda_4\|W^T\|_{1,1}.\end{aligned}\tag{6}$$

λ_1, λ_2, λ_3, λ_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.

4 Optimization Algorithm

Note that the objective function (6) is not easy to solve, since it is non-smooth and biconvex. In this section, we first introduce the whole alternating optimization for solving (6). Then we show how to customize the accelerated proximal gradient method (APM) [36] 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 [37]. We conclude the overall alternating optimization algorithm for solving our proposed MAGPP in Alg. 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.

Algorithm 1 Alternating Optimization for MAGPP.

Input:
$X=[X_1,\cdots,X_m]$: feature dataset for m tasks.
$Y=[\mathbf{y_1},\cdots,\mathbf{y_m}]$: response for m tasks.
$\lambda_1,\lambda_2,\lambda_3,\lambda_4$: hyperparameter.
s: the pseudo hyperparameter to constrain $r_{i,i}=0$.
ϵ: the threshold for terminating the procedure.
Output:
W: the model coefficient matrix.
R: the temporal relation between tasks.

```
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
```

4.1 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 multi-task learning [38], [39]. It has the following form:

$$\min_W F(W)=f(W)+g(W),\tag{7}$$

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. According to [40], let $\alpha_k=\frac{(t_{k-1}-1)}{t_k}$, $t_0=1$ and $t_k=\frac{1}{2}(1+\sqrt{4t_{k-1}^2+1})$ for $k\geqslant 1$.

The approximation point W^k is computed as

$$W^k=\pi(S^k-\eta_k\nabla f(S^k)),\tag{8}$$

where η_k is the chosen step size, $\pi(V)$ is the proximal operator of V.

The global convergence of APM is dependent on an appropriate step size of η_k. Many sophisticated line search schemes [41] can estimate the step size η_k. Updates are made to the value of η_k up until the following condition is met:

$$\begin{aligned}f(W^k)\leqslant& f_\eta(W^k,S^k)\\=&f(S^k)+\langle\nabla f(S^k),W^k-S^k\rangle\\&+\frac{1}{2\eta_k}\|W^k-S^k\|_F^2.\end{aligned}\tag{9}$$

We summarize the procedure of APM in Alg. 2.

Algorithm 2 The Accelerated Proximal Gradient Algorithm.

Input: $X=[X_1,\cdots,X_m]$, $Y=[\mathbf{y_1},\cdots,\mathbf{y_m}]$
Output: W: the model coefficient matrix.

```
Initialize: η_0 = 1, t_0 = 0, t_1 = 1, W^1 = W^0.
for k = 1 to ··· do
    α_k = (t_{k-1} − 1)/t_k, S^k = W^k + α_k(W^k − W^{k−1}). ▷ search point
    for m = 0 to ··· do
        η_k = 2^m η_{k−1}
        Solving (8) for W^{k+1}.
        if (9) is satisfied then                ▷ line search
            break
        end if
    end for
    t_k = (1/2)(1 + sqrt(1 + 4t_{k−1}^2))
    if convergence crition is satisfied then
        Output W^k, break
    end if
end for
```

Emphasize that the computation of the proximal operator (8) is the crucial step in using APM. The complexity for solving (8) dominates the whole complexity of APM-based algorithms. As usual, the proximal operator of the

non-smooth part is not easy to solve, e.g., [7], [9]. However, in our proposed novel MAGPP (6), we will show no matter updating W or R, the proximal operators admit a closed-form solution, which enables to design an efficient algorithm.

4.2 Fix R, Update W

For updating W, we fix the matrix R, the sub-optimization problem is

$$\min_W \frac{1}{2}\sum_{i=1}^{m}\|X_i\mathbf{w_i}-\mathbf{y_i}\|_2^2+\lambda_1\|W-WR\|_F^2 + \lambda_3\|W^T\|_{2,1}+\lambda_4\|W^T\|_{1,1}. \quad (10)$$

The last two terms $\lambda_3\|W^T\|_{2,1}+\lambda_4\|W^T\|_{1,1}$ are non-smooth. In order to find the proximal operator, we need to solve

$$\pi(W)=\arg\min_V \frac{1}{2}\|V-W\|_F^2 + \lambda_3\|V^T\|_{2,1}+\lambda_4\|V^T\|_{1,1}. \quad (11)$$

Each row of V and W is decoupled in (11). To get the i-th row $\mathbf{v^i}$, we need to solve

$$\pi(\mathbf{w^i})=\arg\min_{\mathbf{v^i}} \frac{1}{2}\|\mathbf{v^i}-\mathbf{w^i}\|_2^2+\lambda_3\|\mathbf{v^i}\|_2+\lambda_4\|\mathbf{v^i}\|_1. \quad (12)$$

We introduce the following Lemma 1 to get the closed-form solution.

Lemma 1. [42] For any λ_1,λ_2,

$$\pi_{Lasso}(\boldsymbol{w})=\arg\min_{\boldsymbol{v}} \frac{1}{2}\|\boldsymbol{v}-\boldsymbol{w}\|_2^2+\lambda_1\|\boldsymbol{v}\|_1. \quad (13)$$

$$\pi_{GLasso}(\boldsymbol{w})=\arg\min_{\boldsymbol{v}} \frac{1}{2}\|\boldsymbol{v}-\boldsymbol{w}\|_2^2+\lambda_2\|\boldsymbol{v}\|_2. \quad (14)$$

$$\pi(\boldsymbol{w})=\arg\min_{\boldsymbol{v}} \frac{1}{2}\|\boldsymbol{v}-\boldsymbol{w}\|_2^2+\lambda_1\|\boldsymbol{v}\|_1+\lambda_2\|\boldsymbol{v}\|_2.$$

Then the following holds:

$$\pi(\boldsymbol{w})=\pi_{GLasso}(\pi_{Lasso}(\boldsymbol{w})).$$

We use the soft-thresholding method to get the closed-form solution for (13). Each element of $\pi_{Lasso}(\mathbf{v})$ satisfies

$$\pi_{Lasso}(\mathbf{w})_i=\max(|v_i|-\lambda,0)\cdot sgn(v_i). \quad (15)$$

For the closed-form solution for (14), we use Lemma 2.

Lemma 2. [43] For $\lambda\geqslant 0, \boldsymbol{w}\neq\mathbf{0}$,

$$\pi(\boldsymbol{w})=\arg\min_{\boldsymbol{v}} \frac{1}{2}\|\boldsymbol{v}-\boldsymbol{w}\|_2^2+\lambda\|\boldsymbol{v}\|_2 = \max\{\|\boldsymbol{w}\|_2-\lambda,0\}\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_2}.$$

We conclude that the complexity for solving (11) is only $\mathcal{O}(md)$, so we can update W efficiently.

4.3 Fix W, Update R

For updating R, the sub-optimization problem is

$$\min_R \lambda_1\|W-WR\|_F^2+\lambda_2\|R\odot S\|_{1,1}. \quad (16)$$

To obtain the proximal operator of $\lambda_2\|R\odot S\|_{1,1}$, we must resolve the problems below.

$$\pi(R)=\arg\min_Q \frac{1}{2}\|Q-R\|_F^2+\lambda_2\|R\odot S\|_{1,1}. \quad (17)$$

Clearly, (17) is an extension of Lasso problem, so we also apply the soft-thresholding method to arrive the closed-form solution:

$$\pi(R)=\max(|R|-\lambda_2 S,0)\odot sgn(R). \quad (18)$$

We only need the complexity of $\mathcal{O}(m^2)$ to solve (16).

4.4 The Reason for Using $\|W-WR\|_F^2$

Based on the above discussion, here we explain the reason why we choose $\|W-WR\|_F^2$, rather than $\|W-WR\|_{1,1}$, to capture the complex temporal relation between tasks.

If we apply $\|W-WR\|_{1,1}$, the associated optimization problem for updating W becomes from (10) to

$$\min_W \frac{1}{2}\sum_{i=1}^{m}\|X_{\mathbf{w_i}}-\mathbf{y_i}\|_2^2+\lambda_1\|W-WR\|_{1,1} + \lambda_3\|W^T\|_{2,1}+\lambda_4\|W^T\|_{1,1}. \quad (19)$$

The proximal operator problem (11) becomes

$$\pi(W)=\arg\min_V \frac{1}{2}\|V-W\|_F^2+\lambda_1\|V-VR\|_{1,1} + \lambda_3\|V^T\|_{2,1}+\lambda_4\|V^T\|_{1,1}. \quad (20)$$

This problem (20) no longer admits a closed-form solution. In fact, we can solve (20) using the alternating direction method of multipliers (ADMM) [44]. Despite ADMM being widely used [16], [9], [9], for a desired accuracy ϵ, the worst-case convergence rate of ADMM is only $\mathcal{O}(1/\epsilon^2)$. It is quite slow, and the actual speed of implementation of ADMM may be affected by the penalty parameter ρ chosen [42]. It is concluded that applying $\|W-WR\|_{1,1}$ will result in expensive computational costs for updating W.

Similarly, if $\|W-WR\|_{1,1}$ is applied, the associated optimization problem for updating R becomes from (16) to

$$\min_R \lambda_1\|W-WR\|_{1,1}+\lambda_2\|R\odot S\|_{1,1}. \quad (21)$$

Due to the all non-smooth terms, (21) is challenging to solve. The subgradient method [41] is a viable option. However, the low convergence rate of subgradient method, say $\mathcal{O}(1/\epsilon^2)$ for a desired accuracy ϵ, will also lead to extremely expensive computational cost for updating R.

We conclude that the utilization of $\|W-WR\|_F^2$ is for reducing the computational cost. $\|W-WR\|_{1,1}$ is an alternative option, however, only from the perspective of theory. In practice, we can hardly accept such expensive computational costs leading by the use of $\|W-WR\|_{1,1}$.

4.5 Complexity Analysis

For simplicity, we make an assumption that each task has identical n training samples. The computational cost of our proposed optimization algorithm is composed of two parts, the complexity of updating W and R.

4.5.1 The Complexity of Updating W

When optimizing W, each iteration needs to compute the gradient of the smooth part $\frac{1}{2}\sum_{i=1}^{m}\|X_i\mathbf{w_i}-\mathbf{y_i}\|_2^2+\lambda_1\|W-WR\|_F^2$ and the proximal operator of the non-smooth part $\lambda_3\|W^T\|_{2,1}+\lambda_4\|W^T\|_{1,1}$. 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 $\mathcal{O}(md)$. So in the procedure of updating W, each iteration has the complexity of $\mathcal{O}(nd+m^2(m+d))$. The convergence rate of APM is proved to be $\mathcal{O}(1/\sqrt{\epsilon})$ iterations for a desired accuracy ϵ [45], so the overall complexity for updating W is $\mathcal{O}\left((nd+m^3+m^2d)/\sqrt{\epsilon}\right)$.

4.5.2 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 $\mathcal{O}\left(m^2d/\sqrt{\epsilon}\right)$.

4.5.3 The Overall Complexity of Algorithm 1

In Alg. 1, W and R will be updated once each, which counts as a full iteration. Therefore, a full iteration has the following complexity:

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

4.6 A Warm Start Strategy

Note that, there is currently no theory work that can guarantee the convergence rate of the alternating optimization [46]. In order to further improve the efficiency, we propose a warm start technique to initialize R. Specifically, this strategy starts from an intuitive idea that the larger the interval between two time points, the less similar they are. We use a variant of the Gaussian kernel to measure the similarity between two time points i and j. The corresponding weight of the temporal relation is initialised as

$$r_{i,j}\overset{\text{Initialize}}{=\!=}\begin{cases}\dfrac{e^{-|i-j|}}{\sum_{i=1,i\neq j}^{m}e^{-|i-j|}},\forall i\neq j\\ 0, \qquad i=j.\end{cases}$$

According to experimental results, it can, at most, reduce the number of iterations by 87% and the computational time by 85%, compared to initialization using the zero matrix in our experiments. Please refer to Section 5 for details. It is worth noting that we can use $e^{-|i-j|^\alpha}$ to propose different initialization strategies. The parameter α adjusts the decay of the temporal relation. In fact, we have tried $\alpha\in\{0.5,1,2,e,5,10\}$ and it works best when $\alpha=1$. As a result, in this work, we uniformly set $\alpha=1$.

5 Experimental Result

In this section, the three AD datasets used in this study are first described. Then we show the effectiveness of the warm start strategy and compare the performance of MAGPP with several baseline methods. To investigate the complex temporal relations between tasks, we visualise the adjacency matrix of the temporal relation graph which MAGPP automatically learns from the datasets. To show the high interpretability of our method, and possibly facilitate the understanding of AD progression, we perform stability selection to find stable biomarkers from the MRI feature set and examine their temporal trends in the development of AD. The hardware condition is an Apple M1 Max chip with 32 GB memory. The implementation source runs on MATLAB and can be found at https://github.com/menghuizhou/MAGPP.

5.1 The Latest Dataset from ADNI

The Alzheimer's Disease Neuroimaging Initiative (ADNI) database [11] helps many research works like [47], [48] and also provides the data for this study. The primary goal of ADNI has been to ascertain whether magnetic resonance imaging (MRI), positron emission tomography (PET), and neuropsychological tests can be used in conjunction to track the development of early AD. Finding sensitive and precise biomarkers of very early AD progression will help clinical trials move more quickly and affordably. This will assist medical professionals in developing new treatments and evaluating their effectiveness. The initial hospital screening of patients is known as the baseline (BL). The moment the baseline began serves as a representation of the follow-up time point. For instance, the notation "M12" indicates a time point that is a year after the initial visit (baseline time point). For certain patients, the most recent ADNI provides follow-up data from up to 120 months. But many participants leave the study for a variety of reasons. Due to the small amount of data at the last time points, according to the method of previous works [6], [7], [8], only the data from the first six time points are used. The three measurements for AD cognitive state used in this paper are MMSE, ADAS-Cog, and RAVLT.

Here is a list of the data preprocessing steps we take:

- Patients without baseline MRI records are excluded.
- Delete any participants whose MRI picture quality control failed.
- Use the average value to fill in missing entries of features.

Finally, we get 314 features. The specifics of datasets are shown in Table 1.

5.2 Effectiveness of Warm Start Strategy

Here we compare the efficiency of two different initialization strategies for R on three datasets. We refer to MAGPP initialised with zero as MAGPP-0, and initialised with our suggested warm start strategy as MAGPP-w.

In order to comprehensively compare the efficiency of the two initialization strategies, we randomly select 5 times of hyperparameters and run them on three

TABLE 1
The specific details of the sample number at each time point in the sequence. The number of patients who have baseline MRI features at subsequent time points is the sample size.

Time point	MMSE	ADAS-Cog	RAVLT
M00	1092	1074	1091
M06	1078	1064	1074
M12	1027	1014	1021
M24	883	867	877
M36	579	556	576
M48	494	483	468

datasets, respectively. The hyperparameters $\lambda_1, \lambda_2, \lambda_3, \lambda_4 \in \{10^0, 10^1, 10^2, 10^3, 10^4\}$, the pseudo hyperparameter s is set as 10^9. The feature matrix X is normalised. When the relative changes of objective function value in two successive iterations are not greater than the stopping criterion $\tau \in \{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\}$, the optimization algorithm is terminated. The maximum iteration is 200. We record the average number of iterations on three datasets, respectively. We put the results about the iteration number in Table 2 and computational cost in Table 3.

TABLE 2
Comparing the number of iterations of MAGPP-0 and MAGPP-w on three datasets.

Dataset	Method	Stopping Criterion ($\leqslant \tau$)				
		10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}
MMSE	MAGPP-0	43.5	74.0	75.4	78.6	86.5
	MAGPP-w	8.1	9.9	12.9	17.1	67.6
	Rate (%)	81↓	**87**⇓	83↓	78↓	22↓
ADAS-Cog	MAGPP-0	35.4	41.6	45.0	55.4	63.1
	MAGPP-w	7.4	9.0	16.4	19.9	26.5
	Rate (%)	79↓	78↓	64↓	64↓	58↓
RAVLT	MAGPP-0	57.5	64.5	65.1	91.9	111.9
	MAGPP-w	10.9	13.8	15.3	19.0	21.0
	Rate (%)	81↓	79↓	77↓	79↓	81↓

TABLE 3
Comparing the computation time (second) of MAGPP-0 and MAGPP-w on three datasets.

Dataset	Method	Stopping Criterion ($\leqslant \tau$)				
		10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}
MMSE	MAGPP-0	5.5	8.9	9.3	10.2	11.3
	MAGPP-w	1.0	1.3	1.7	2.4	8.2
	Rate (%)	83↓	**85**⇓	83↓	77↓	27↓
ADAS-Cog	MAGPP-0	3.5	4.4	4.9	6.2	7
	MAGPP-w	0.8	1	1.7	1.9	3
	Rate (%)	78↓	78↓	65↓	69↓	57↓
RAVLT	MAGPP-0	5.8	6.5	6.7	9.6	13
	MAGPP-w	1	1.2	1.5	1.8	2.1
	Rate (%)	83↓	82↓	78↓	81↓	84↓

As shown in Table 2, regardless of the stopping criterion on either dataset, the number of iterations needed by MAGPP-w is rather less than the number needed by MAGPP-0. When $\tau = 10^{-2}$, on the MMSE dataset, our proposed warm start strategy can reduce the number of iterations by up to 87%. Similarly, as shown in Table 3, it can also effectively reduce the computation time of the algorithm. When $\tau = 10^{-2}$, on the MMSE dataset, the computation time of the algorithm is reduced by at most 85%, which means that the efficiency has increased by 6.84 times. Overall, our proposed simple warm start strategy based on a variant of the Gaussian kernel can effectively reduce the number of iterations and computation time required by the algorithm on all datasets and under different stopping criteria.

5.3 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 are divided randomly and equally into the 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 $\lambda_1, \lambda_1, \lambda_3, \lambda_4$ where $\lambda_1 \in \{10^2, 10^3, 10^4\}, \lambda_2 \in \{10^0, 10^1\}, \lambda_3, \lambda_4 \in \{10^0, 10^1, 10^2, 10^3\}$, the pseudo hyperparameter s is set as 10^9. The feature matrix X is normalised.

5.3.1 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 multi-task learning literature [7], [9]. Higher performance is indicated by lower nMSE and rMSE or higher wR. The nMSE, wR, rMSE are defined as follows:

$$\text{nMSE}(Y, \hat{Y}) = \frac{\sum_{i=1}^{t} \|Y_i - \hat{Y}_i\|_2^2 / \sigma^2(Y_i)}{\sum_{i=1}^{n} n_i},$$

$$\text{wR}(Y, \hat{Y}) = \frac{\sum_{i=1}^{t} \text{Corr}(Y_i, \hat{Y}_i) n_i}{\sum_{i=1}^{t} n_i},$$

$$\text{rMSE}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \sqrt{\frac{\|\boldsymbol{y} - \hat{\boldsymbol{y}}\|_2^2}{n}},$$

where Y and $\hat{Y}$ are the ground truth cognitive scores and the predicted cognitive scores, respectively. $\sigma^2(\cdot)$ is variance.

5.3.2 Comparative Models and Ablation Experiments

We thoroughly contrast our MAGPP with a number of multi-task learning baseline techniques. All comparative models include TGL [6], cFSGL [7], VSTG [16], NCCMTL [49], FLSGL [8], LSA [9] and GAMTL [50]. The tuning range of all hyperparameters is $\in \{10^1, 10^2, 10^3, 10^4\}$. For VSTG, the hyperparameter k of the k-support norm is $\in \{1, 3, 5\}$. For FLSGL, the bandwidth hyperparameter $\sigma \in \{1, 5, 10\}$. For LSA, the hyperparameter is $\alpha \in \{0, 0.05, 0.1, 0.2\}$. We emphasize that although VSTG is not specifically proposed for predicting AD progression like other baseline methods, because it can use task relation to select the important MRI

TABLE 4
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	FLSGL	VSTG	NCCMTL	LSA	GAMTL	LSTM	SGLasso	AutoTG	MAGPP
Dataset: MMSE												
0.4	nMSE	**0.620**	0.631	0.651	0.649	0.626	0.630	0.635	0.850	0.671	0.642	0.624
	wR	0.616	0.610	0.594	0.588	0.611	0.610	0.608	0.438	0.590	0.601	**0.619**
0.6	nMSE	0.621	0.597	0.639	0.659	0.632	0.601	0.627	0.860	0.666	0.638	**0.585**
	wR	0.618	0.632	0.607	0.593	0.609	0.619	0.611	0.444	0.599	0.607	**0.636**
0.8	nMSE	0.602	0.583	0.626	0.641	0.627	0.579	0.591	0.756	0.653	0.617	**0.567**
	wR	0.631	0.650	0.619	0.608	0.600	0.650	0.637	0.562	0.612	0.629	**0.663**
Dataset: ADAS-Cog												
0.4	nMSE	0.494	0.490	0.511	0.527	0.519	0.493	0.511	0.870	0.526	0.498	**0.485**
	wR	0.717	0.730	0.700	0.691	0.690	0.734	0.718	0.493	0.673	0.695	**0.740**
0.6	nMSE	0.482	0.470	0.491	0.500	0.498	0.463	0.493	0.796	0.519	0.487	**0.460**
	wR	0.729	0.747	0.713	0.698	0.705	**0.749**	0.736	0.513	0.685	0.709	0.748
0.8	nMSE	0.471	0.459	0.474	0.483	0.480	0.463	0.475	0.703	0.505	0.476	**0.453**
	wR	0.734	**0.756**	0.729	0.711	0.717	0.753	0.724	0.591	0.698	0.717	0.755
Dataset: RAVLT												
0.4	nMSE	0.618	0.620	0.634	0.629	0.627	0.599	0.625	0.882	0.639	0.614	**0.587**
	wR	0.632	0.633	0.609	0.605	0.607	0.639	0.621	0.445	0.576	0.618	**0.646**
0.6	nMSE	0.600	0.605	0.617	0.615	0.620	0.587	0.609	0.829	0.635	0.609	**0.575**
	wR	0.640	0.651	0.616	0.607	0.609	0.656	0.622	0.514	0.584	0.624	**0.665**
0.8	nMSE	0.593	0.574	0.589	0.601	0.596	0.571	0.587	0.781	0.629	0.601	**0.549**
	wR	0.648	0.659	0.632	0.618	0.625	0.660	0.652	0.556	0.597	0.635	**0.673**

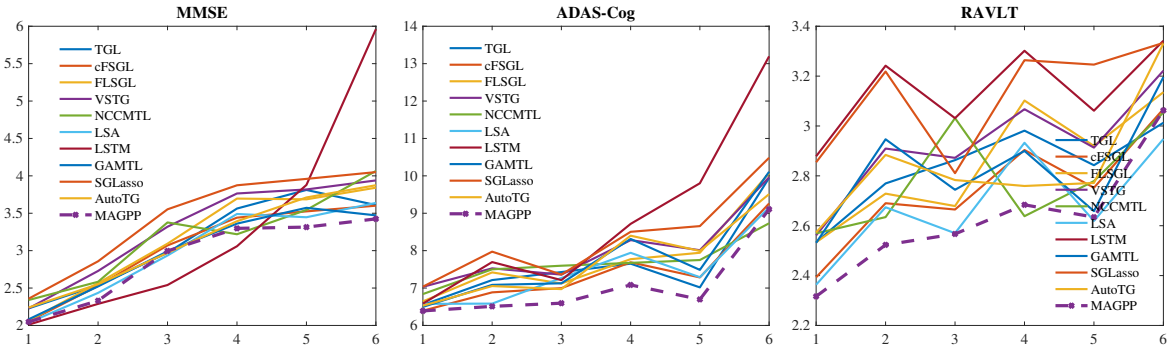

Fig. 2. The comparison of the single task performance, between our MAGPP and several baseline methods. The results show the average rMSE with 5 repetitions. The index of x-axis represents the time points M00, M06, M12, M24, M36, and M48, respectively. The training ratio $\beta = 0.8$.

features, we regard it to be a baseline method. For GAMTL, we use the same MRI feature grouping strategy as [50], i.e., we group all features according to brain regions of interest.

It is worth noting that the number of MRI features is relatively high in comparison to the number of samples, as shown in Table 1. Several baseline works TGL, cFSGL, VSTG, FLSGL, LSA, GAMTL, and our proposed MAGPP all have used a penalty term associated with Lasso to select the most important feature subset. Although NCCMTL does not involve the part of feature selection, it considers the case that different tasks have varying noise levels, so we still use it as a kind of baseline in order to provide deeper Insights into AD progression.

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 is 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 scores at specific time points, after we keep all the patient data with cognitive scores at six time points, MMSE, ADAS-Cog, RAVLT datasets have 331, 365, and 350 samples, respectively.

Considering that MAGPP is composed of two parts, the first is AutoTG, which automatically captures the temporal relation between tasks, and the second is sparse group Lasso (SGLasso). We test the effectiveness of AutoTG and SGLasso on three AD datasets, respectively, to further confirm the efficacy of our MAGPP.

As shown in Table 4, we first notice that LSTM has the worst performance in all cases, mainly because about 300 samples could not train a good enough LSTM

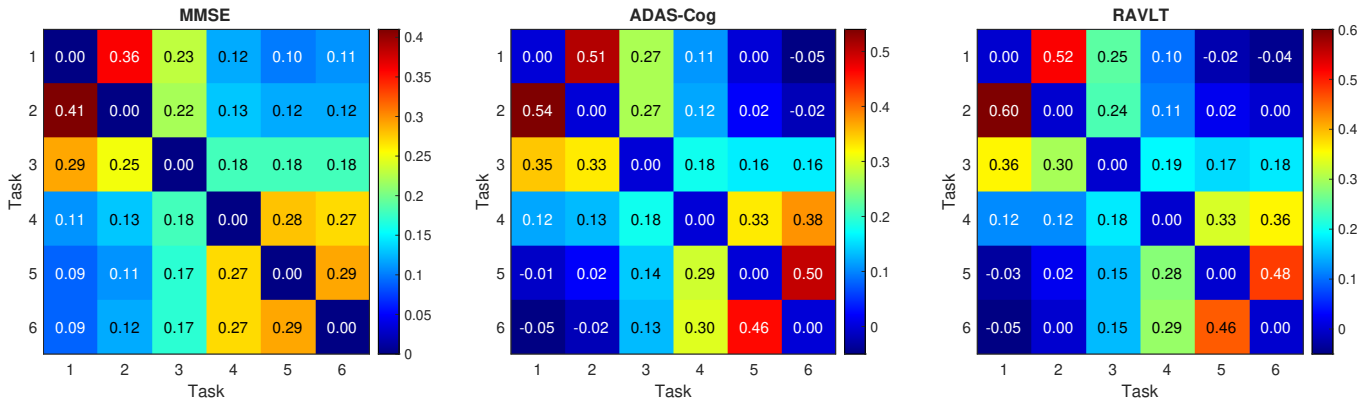

Fig. 3. The adjacency matrix of the temporal relation graph between tasks, which MAGPP automatically learns from the MMSE, ADAS-Cog, and RAVLT datasets, respectively.

model. SGLasso does not perform well. The performance of SGLasso is essentially the second worst when compared to all other methods, with a noticeable performance gap in MMSE and RAVLT datasets. The reason is that SGLasso does not consider the relation between tasks, which also indicates that it is necessary to take advantage of the relation between tasks in the study of AD progression. Compared to SGLasso, AutoTG performs significantly better, but not as well as MAGPP. It suggests that it is useful to introduce sparsity within and between groups for feature selection. The preceding discussion fully demonstrates the effectiveness of the components of MAGPP.

In most cases, MAGPP achieves the best performance, except in several cases like on the ADAS-Cog dataset with $\beta = 0.8$, cFSGL achieves the best result with $wR = 0.756$ as the metric, but only slightly better than MAGPP with $wR = 0.755$. We also notice that on the RAVLT dataset with $\beta = 0.8$, compared with the best baseline performance $nMSE = 0.571$ of LSA, MAGPP significantly reduces the nMSE to 0.549. Other than MAGPP, LSA gets the best performance. It indicates the adaptive global temporal structure used in LSA is effective in handling the temporal relation between tasks in the AD progression. VSTG does not perform well in all of these 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. NCCMTL achieves good performance. The possible reason is that it adopts square root loss function to deal with different noise levels of different tasks. This also shows that the effect of noise cannot be easily ignored in the three used datasets. We also note that GAMTL performs moderately, which again demonstrates the effectiveness of feature selection in AD progression prediction, considering that GAMTL does not do feature selection.

In addition to analyzing the model's overall performance, we also examine how well MAGPP performs at each time point. Considering the limited paper space, we only show the results of training ratio $\beta = 0.8$ in Fig. 2. We emphasize that other cases with different β have similar results. First of all, in most cases, LSTM has the worst performance of every task at different datasets. We discover that SGLasso basically gets the second worst performance at each time point on all three datasets. It demonstrates once again that the relation between tasks must be fully considered and utilised in the study of AD progression. In contrast, the other component of the MAGPP, AutoTG, gets a middle performance, compared to other methods, explaining the need for feature selection in the AD progression. It is evident that, regardless of the outcomes on the MMSE, ADAS-Cog, or RAVLT datasets, the prediction performance of MAGPP is generally the best at single time points.

5.4 Visualisation of Temporal Relation

To fully analyse the temporal relation between multiple tasks, we visually analyze the temporal relation captured automatically by MAGPP, where all training ratio $\beta = 0.8$. We specifically visualise the adjacency matrix of the learned task relation graph. According to the results in Fig. 3, the temporal relation automatically learned by MAGPP on the three datasets has both similarities and differences. To begin with, all adjacency matrices are not strictly symmetric and this asymmetry corresponds to the real-life temporal relation. For instance, $r_{k-1,k}$ represents analyzing the past state of one patient at 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.

In each of the three datasets, almost every task is primarily related to its neighbours. The difference is that the relation between adjacent tasks is basically strongest in the RAVLT dataset. For example, the average weight of the 1-st task and the 2-nd task is 0.56. This could be because the RAVLT is designed to assess episodic memory, but AD patients do not lose episodic memory very quickly. The ADAS-Cog dataset also has a strong neighbouring relation which is due to that the ADAS-Cog is a detailed and comprehensive measurement to evaluate the cognitive state of AD patients, and it includes 11-item cognitive tests designed to detect

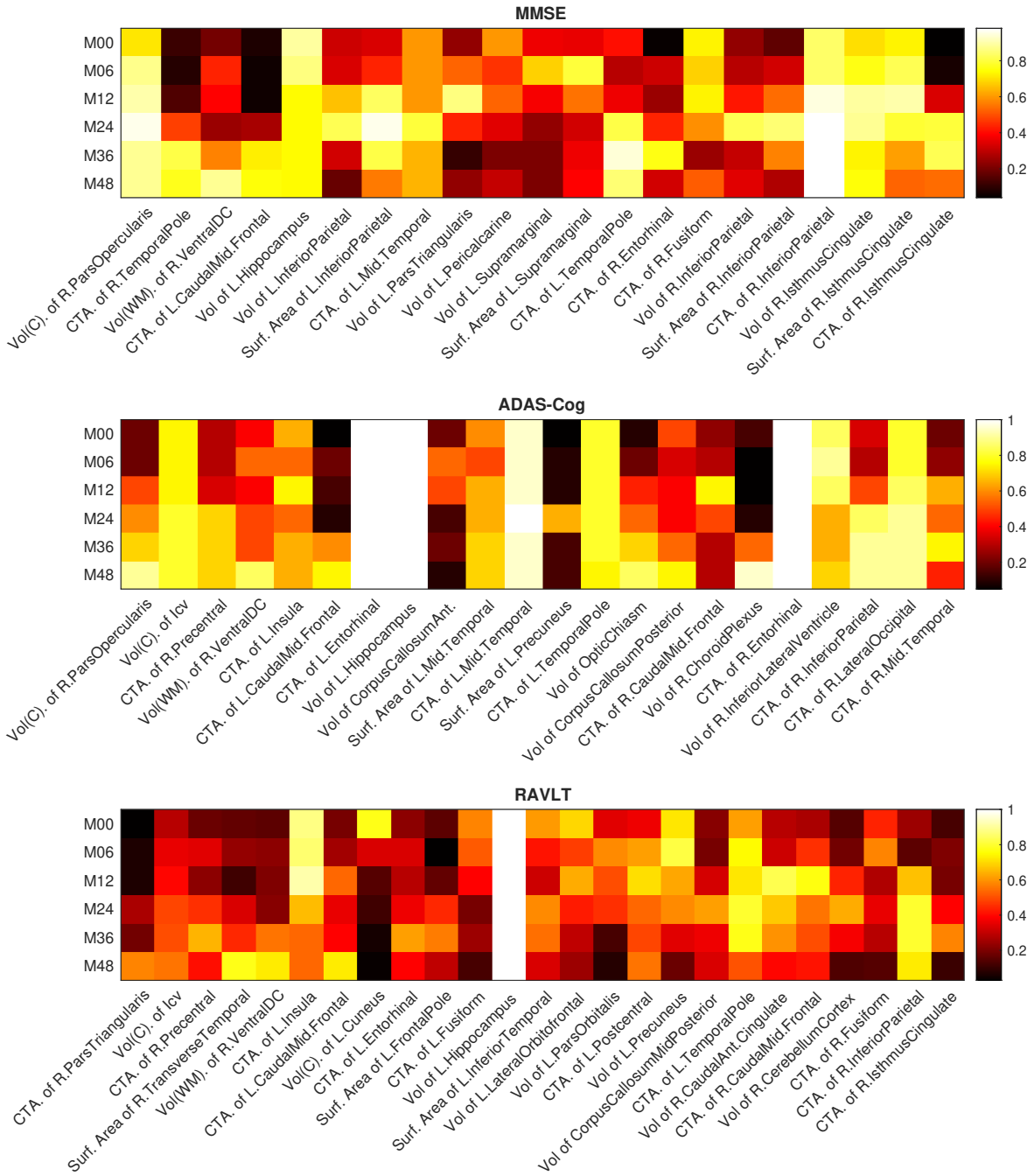

Fig. 4. The stability vector of table MRI features using MAGPP on three datasets. We choose top 6 stable features on each time point, finally we get 21 stable features on MMSE dataset, 22 stable features on ADAS-Cog dataset, and 25 stable features on RAVLT dataset. The feature is more stable the higher the value.

changes in AD severity. Given that AD is the most common chronic disease, the relations between adjacent time points should be strong, as evidenced by previous literature that studies AD with a local temporal relation [6], [7].

However, in the MMSE dataset, the relation between neighbouring tasks is the weakest, especially the weight of the relation between the 3-th and 4-th tasks is only 0.18. A possible reason is that, while MMSE is the most common scale for measuring the degree of mental impairment of AD in clinical practice, compared with ADAS-Cog, MMSE cannot further perform detailed neuropsychological tests, including memory, executive function, and language ability as ADAS-Cog does. MMSE only has brief content and measurements. As a result, when compared to ADAS-Cog, the measurement of MMSE may be rougher and unable to accurately assess patients' cognitive state.

It is very worth noting that on the two datasets ADAS-Cog and RAVLT, some temporal relation weights are negative. For example, on the ADAS-Cog dataset, the relation weights of the 1-st task and the 5-th and 6-th tasks are -0.01 and -0.05, respectively. On the RAVLT dataset, they are -0.03 and -0.05. This shows that when the time points are too far apart, the corresponding tasks are no longer similar to each other, but a little repulsive. This phenomenon has

never been considered in all previous works [6], [7], [8], [9].

We conclude that the temporal relation learned by MAGPP shows that:

- In the study of AD progression, the temporal relation between tasks is asymmetric and global, which proves the deficiency of using local temporal relation in previous works [6], [7].
- The temporal relation between tasks is extremely complex, which indicates that the previous works use a predefined Gaussian kernel method [8] or a predefined iterative convex structure [9] can not fully capture the complex task relation.
- All existing works [6], [7], [8], [9] do not consider the possible negative temporal relation.

5.5 Temporal Pattern of Stable Biomarkers

One of the advantages of MAGPP is its capacity to examine the temporal patterns of MRI features, which makes it easier to comprehend how AD progresses. To further explore the MRI biomarkers discovered by our formulations, we use the longitudinal stability selection method [32], which has been employed in numerous prior studies [8], [9]. The details are in [51], [32]. In this context, the term "stability vector" refers to 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 the ADAS-Cog and RAVLT datasets, where the volume of the left hippocampus is selected to be stable biomarker with the probability close to 1. In the 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 [52] 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 three datasets. In the 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 the ADAS-Cog dataset, the volume of right IsthmusCingulate only shows stability in the last few moments from M24 to M48, and in the RAVLT dataset, the volume of the right IsthmusCingulate only shows stability in the last few moments, from M12 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 the ADAS-Cog dataset, it is selected as a stable biomarker with a higher frequency, close to 0.9. However, in the RAVLT dataset, the cortical thickness average of middle temporal is not selected as stable biometrics all the time.

The distinct temporal patterns of the stable biomarkers of these three cognitive scores also suggest that it may be less effective to confine the model to a shared set of features, as do previous methods [7], [8], [9].

6 Discussion

Although our proposed method achieves good performance on datasets corresponding to multiple cognitive scores, including visualised temporal relationships, longitudinal stability selection analysis of MRI features, etc., its performance and generalization capacity could be further improved in the future by carefully addressing the limitations or challenges listed below.

- Our model analyses different cognitive scores separately, which means that the input data only contains one type of cognitive score. Given that different cognitive scores measuring the state of the same patient are intrinsically related, it is possible to improve the model performance and generalisation ability by jointly analysing multiple cognitive scores. In fact, we already have some preliminary ideas. For example, Romeo et al. [13] integrated the input data of the five diabetes complications into the same multi-ask learning model without losing interpretability, and we can try to borrow the spirit to extend MAGPP to deal with multiple cognitive scores at the same time.
- Recently, in the field of deep learning, especially in transformer related research, the positional encoding technique has been recognised as an effective way to exploit temporal relationships. The main idea is to treat the time point information as a feature and part of the model input. Temporal information can be treated as an input to our MAGPP, or more specifically, as a special type of feature. In our future work, we plan to investigate the feasibility of incorporating the positional encoding technique into our MAGPP.

7 Conclusion

In this paper, we investigated AD progression using the baseline MRI feature set and cognitive scores at future time points. We conclude this paper in the following three parts.

We propose AutoTG, a novel mechanism for automatically capturing the complex temporal relation between tasks, and build it as a graph adjacency matrix. Then, to predict the progression of Alzheimer's disease, we combine the sparse group Lasso and AutoTG to propose MAGPP, a novel multi-task formulation that outperforms several baseline methods on three AD datasets. To solve the non-smooth and biconvex objective function, we customize the alternating optimization and utilize the accelerated proximal gradient method to handle the two associated sub-optimization problems efficiently. Since there is currently no theory work to prove the convergence rate of alternating optimization, to improve the efficiency of our algorithm even further, we design a warm start strategy based on a local temporal relation between multiple tasks. When compared to the method without the warm start strategy, the outcomes of the experiment indicate that the warm start strategy can, at most, cut the number of iterations by 87% and the computation time by 85%.

To demonstrate the ability of MAGPP for capturing the complex temporal relation, and also explore the temporal relation between tasks in the study of AD progression, we visualise the automatically learned graph adjacency matrix.

The results on three datasets demonstrate that every task is temporally related to all other tasks, with different relation weights. The asymmetry of the learned adjacency matrices on three AD datasets reveals that the temporal heterogeneity can not be ignored. It means that analyzing the relation between time points m and n at time point m is not equivalent to analyzing the relation between time points m and n at time point n. Not only that, we also show two tasks that are far apart in time can even have a negative weight, meaning that they are mutually exclusive rather than similar. Furthermore, to show the high interpretability of our method, and also aid the understanding of AD progression, we use stability selection to identify stable MRI features and investigate their temporal patterns. Some of the selected features are consistent with previous works, while others have the potential to facilitate the discovery of new biomarkers.

Since AutoTG is a general method for capturing the complex temporal relation between multiple tasks, in the future, we hope to investigate the efficacy in a border area, such as the development prediction of Parkinson's disease [12] and diabetes [13].

8 Acknowledgment

This work was supported by the National Natural Science Foundation of China (No. 62061050). We very appreciate the valuable comments from anonymous reviewers.

References

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

[2] Jing Wan, Zhilin Zhang, Bhaskar D Rao, Shiaofen Fang, Jingwen Yan, Andrew J Saykin, and Li Shen. Identifying the neuroanatomical basis of cognitive impairment in alzheimer's disease by correlation-and nonlinearity-aware sparse bayesian learning. *IEEE transactions on medical imaging*, 33(7):1475–1487, 2014.

[3] Peng Cao, Xuanfeng Shan, Dazhe Zhao, Min Huang, and Osmar Zaiane. Sparse shared structure based multi-task learning for mri based cognitive performance prediction of alzheimer's disease. *Pattern Recognition*, 72:219–235, 2017.

[4] Jialin Peng, Xiaofeng Zhu, Ye Wang, Le An, and Dinggang Shen. Structured sparsity regularized multiple kernel learning for alzheimer's disease diagnosis. *Pattern recognition*, 88:370–382, 2019.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

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

[11] 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.

[12] 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.

[13] 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.

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

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

[16] 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.

[17] Feng Li, Loc Tran, Kim-Han Thung, Shuiwang Ji, Dinggang Shen, and Jiang Li. A robust deep model for improved classification of ad/mci patients. *IEEE journal of biomedical and health informatics*, 19(5):1610–1616, 2015.

[18] 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.

[19] Wen Yu, Baiying Lei, Michael K Ng, Albert C Cheung, Yanyan Shen, and Shuqiang Wang. Tensorizing gan with high-order pooling for alzheimer's disease assessment. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[20] P Vemuri, HJ Wiste, SD Weigand, LM Shaw, JQ Trojanowski, MW Weiner, David S Knopman, Ronald Carl Petersen, CR Jack, et al. Mri and csf biomarkers in normal, mci, and ad subjects: predicting future clinical change. *Neurology*, 73(4):294–301, 2009.

[21] Cynthia M Stonnington, Carlton Chu, Stefan Klöppel, Clifford R Jack Jr, John Ashburner, Richard SJ Frackowiak, Alzheimer Disease Neuroimaging Initiative, et al. Predicting clinical scores from magnetic resonance scans in alzheimer's disease. *Neuroimage*, 51(4):1405–1413, 2010.

[22] Simon Duchesne, Anna Caroli, Cristina Geroldi, D Louis Collins, and Giovanni B Frisoni. Relating one-year cognitive change in mild cognitive impairment to baseline mri features. *Neuroimage*, 47(4):1363–1370, 2009.

[23] 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.

[24] Fatih Altay, Guillermo Ramón Sánchez, Yanli James, Stephen V Faraone, Şenem 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.

[25] 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.

[26] Minh Nguyen, Tong He, Lijun An, Daniel C Alexander, Jiashi Feng, BT Thomas Yeo, Alzheimer's Disease Neuroimaging Initiative, et al. Predicting alzheimer's disease progression using deep recurrent neural networks. *NeuroImage*, 222:117203, 2020.

[27] Wei Liang, Kai Zhang, Peng Cao, Xiaoli Liu, Jinzhu Yang, and Osmar Zaiane. Rethinking modeling alzheimer's disease progression from a multi-task learning perspective with deep recurrent neural network. *Computers in Biology and Medicine*, 138:104935, 2021.

[28] Aya Abdelsalam Ismail, Mohamed Gunady, Hector Corrada Bravo, and Soheil Feizi. Benchmarking deep learning interpretability in time series predictions. *Advances in neural information processing systems*, 33:6441–6452, 2020.

[29] Kwan Ho Ryan Chan, Yaodong Yu, Chong You, Haozhi Qi, John Wright, and Yi Ma. Redunet: A white-box deep network from the principle of maximizing rate reduction. *J Mach Learn Res*, 23(114):1–103, 2022.

[30] Wei Liang, Kai Zhang, Peng Cao, Xiaoli Liu, Jinzhu Yang, and Osmar R Zaiane. Exploiting task relationships for alzheimer's disease cognitive score prediction via multi-task learning. *Computers in Biology and Medicine*, 152:106367, 2023.

[31] Shanshan Tang, Peng Cao, Min Huang, Xiaoli Liu, and Osmar Zaiane. Dual feature correlation guided multi-task learning for alzheimer's disease prediction. *Computers in Biology and Medicine*, 140:105090, 2022.

[32] 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.

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

[34] 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.

[35] Ming Yuan and Yi 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.

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

[37] 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.

[38] Jiayu Zhou, Jianhui Chen, and Jieping Ye. Clustered multi-task learning via alternating structure optimization. *Advances in neural information processing systems*, 2011:702, 2011.

[39] Pinghua Gong, Jieping Ye, and Changshui Zhang. Robust multi-task feature learning. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 895–903, 2012.

[40] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

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

[42] Lei Yuan, Jun Liu, and Jieping Ye. Efficient methods for overlapping group lasso. *IEEE transactions on pattern analysis and machine intelligence*, 35(9):2104–2116, 2013.

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

[44] Stephen Boyd, Neal Parikh, and Eric Chu. *Distributed optimization and statistical learning via the alternating direction method of multipliers*. Now Publishers Inc, 2011.

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

[46] John Wright and Yi Ma. *High-dimensional data analysis with low-dimensional models: Principles, computation, and applications*. Cambridge University Press, 2022.

[47] Shuai Liu, Shichen Huang, Shuai Wang, Khan Muhammad, Paolo Bellavista, and Javier Del Ser. Visual tracking in complex scenes: A location fusion mechanism based on the combination of multiple visual cognition flows. *Information Fusion*, 2023.

[48] Nasir Rahim, Shaker El-Sappagh, Sajid Ali, Khan Muhammad, Javier Del Ser, and Tamer Abuhmed. Prediction of alzheimer's progression based on multimodal deep-learning-based fusion and visual explainability of time-series data. *Information Fusion*, 92:363–388, 2023.

[49] Feiping Nie, Zhanxuan Hu, and Xuelong Li. Calibrated multi-task learning. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2012–2021, 2018.

[50] Saullo HG Oliveira, Andre R Goncalves, and Fernando J Von Zuben. Asymmetric multi-task learning with local transference. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 16(5):1–30, 2022.

[51] Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010.

[52] Alzheimer's Association. 2019 alzheimer's disease facts and figures. *Alzheimer's & dementia*, 15(3):321–387, 2019.