Robust Temporal Smoothness in Multi-Task Learning

Menghui Zhou[1], **Yu Zhang**[2], **Yun Yang**[1], **Tong Liu**[2], **Po Yang**[2]

[1] Deparment of Software, Yunnan University, Kunming, China
[2] Department of Computer Science, Sheffield University, Sheffield, UK
mhzcn@mail.ynu.edu.cn, yzhang489@sheffield.ac.uk, yangyun@ynu.edu.cn, {t.liu, po.yang}@sheffield.ac.uk

Abstract

Multi-task learning models based on temporal smoothness assumption, in which each time point of a sequence of time points concerns a task of prediction, assume the adjacent tasks are similar to each other. However, the effect of outliers is not taken into account. In this paper, we show that even only one outlier task will destroy the performance of the entire model. To solve this problem, we propose two ***Robust Temporal Smoothness*** (RoTS) frameworks. Compared with the existing models based on temporal relation, our methods not only chase the temporal smoothness information, but identify outlier tasks, however, without increasing the computational complexity. Detailed theoretical analyses are presented to evaluate the performance of our methods. Experimental results on synthetic and real-life datasets demonstrate the effectiveness of our frameworks. We also discuss several potential specific applications and extensions of our RoTS frameworks.

Introduction

In recent years, the temporal smoothness assumption (Wei 2006) has been used in a wide range of machine learning applications (Wang, Shi, and Reddy 2020; Zhou et al. 2022; Xu et al. 2021; Romeo et al. 2020; Emrani, McGuirk, and Xiao 2017; Saha et al. 2018). They model the interactions between a time point and its adjacent time points and thus capture the temporal relationship to some extent. Owing to intrinsic correlation among multiple time points, a joint analysis of multiple time points is supposed to be more effective than analysing each time point independently. Therefore, the idea of multi-task learning (MTL) (Shen et al. 2021; Fifty et al. 2021; Zhang and Yang 2021) is applied to analyse multiple time points simultaneously. Specifically, existing methods (Romeo et al. 2020; Wang, Shi, and Reddy 2020; Emrani, McGuirk, and Xiao 2017; Zhao et al. 2015; Zheng and Ni 2013) formulate the prediction of a target at a sequence of time points as a multi-task learning problem, and each task concerns the prediction at a time point. As shown in Figure 1, the t-th time point is treated as the t-th task $\boldsymbol{w_t}$.

The crucial challenge of MTL is to know how the tasks are related and how to capture such complex task relation (Zhang and Yang 2021). One common way is employing the **T**emporal **S**moothness assumption (TS). It assumes the difference between two successive tasks is relatively small and thus chases the temporal correlation among multiple tasks. With TS, advanced MTL benefits many applications, like disease progression prediction, survival analysis, and keypoint tracking.

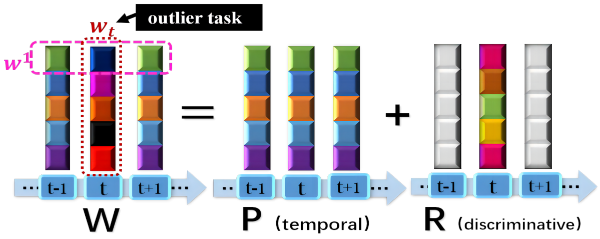

Figure 1: We decompose every $\boldsymbol{w_i} = \boldsymbol{p_i} + \boldsymbol{r_i}$. P satisfies the temporal smoothness, R identifies the outlier tasks.

In (Nie et al. 2016; Zhou et al. 2011), the authors use MTL with TS to predict the progression of Alzheimer's disease. They assume the cognitive score of one patient will not fluctuate dramatically over time, and the difference of cognitive scores between two successive time points is relatively small. So they penalize $\|\boldsymbol{w_i} - \boldsymbol{w_{i+1}}\|_2^2$, known as Laplacian based Temporal Smoothness assumption (LTS). In (Zhou et al. 2022, 2012), the authors argue that LTS only focuses on the smoothing of tasks across different time points. It is a better way to enforce that the nearby time points have similar feature weight, so they penalize $|w_{ij} - w_{i,j+1}|$ using the famous fused Lasso (Tibshirani et al. 2005), regarded as the Fused Lasso based Temporal Smoothness assumption (FTS). Clearly, if FTS is satisfied, so is LTS. Similarly, in (Emrani, McGuirk, and Xiao 2017), the authors use MTL with TS to conduct prognosis and diagnosis of the progression of Parkinson's disease. In (Romeo et al. 2020), the authors propose a novel spatio-temporal MTL model based on TS to predict the progression of diabetes and its complications. Besides in the field of disease diagnosis and treatment, (Wang, Shi, and Reddy 2020) applies TS to propose a tensor based temporal MTL survival model.

Introducing TS into the MTL model has been shown to improve performance and robustness, however, the significant problem is that TS does not consider the difference between tasks and the impacts of potential outlier tasks. Actually, the asymptotic property of fused Lasso proved in (Tibshirani et al. 2005) demonstrates that TS just tends to average all tasks. Due to the usual existence of outlier tasks, TS is too restrictive for real-world applications. Here we first define the outlier task: A task should be considered as outlier if it is vastly different from most tasks. In this study, we identify outlier tasks by comparing the magnitude of the L2-norm of the task coefficients. As shown in Figure 2 (same experimental setup as in Experiment part), LTS and FTS average all tasks, and seem to have a trend, but extremely limited, to capture the outlier task (4-th task). It means even only one outlier task will destroy the entire performance of the MTL models based on TS. Hence, how to detect outlier tasks while chasing the temporal smoothness assumption is a particularly important and challenging problem in models based on TS.

Our motivation comes from an intuitive idea that outlier tasks arise since there is not only the information of temporal smoothness among tasks, but also other information that depends on specific domain knowledge. The outlier task is determined by the domain information, rather than by the noise in data. These outlier tasks also contain a lot of valuable information, that can not be ignored. To implement this idea, we propose two ***R****obust* ***T****emporal* ***S****moothness* (RoTS) frameworks. Mathematically, we write each task model $\boldsymbol{w_i}(i \in \mathbb{N}_m)$ as $\boldsymbol{w_i} = \boldsymbol{p_i} + \boldsymbol{r_i}$ (hence the model coefficient matrix $W = P + R$). *The temporal part* $\boldsymbol{p_i}$ satisfies the temporal smoothness $\boldsymbol{p_i} \approx \boldsymbol{p_{i+1}}$. *The discriminative part* $\boldsymbol{r_i}$ represents the difference beyond the temporal relation among tasks. If $\boldsymbol{r_i}$ is "large", simple temporal smoothness is not suitable, since $\boldsymbol{r_i} \not\approx \boldsymbol{r_{i+1}} \Rightarrow \boldsymbol{w_i} \not\approx \boldsymbol{w_{i+1}}$, i.e., the difference beyond temporal relation among tasks can not be ignored. And the i-th task is regarded as an outlier task.

It is worth noting that it is difficult to give an explicit definition of outlier task, since it depends on the specific case and is governed by the combined effect of the temporal part P and discriminative part R. Traditionally, an outlier is an observation that "lies an abnormal distance from other values in a random sample from a population", where it has significant differences with errors. However, in many practical applications, the outlier may occur randomly and regularly, associating with the definition of tasks. Therefore, through defining different tasks, the threshold in temporal smoothness might be able to classify some errors into outliers, and vice versa. For instance, taking an example of predicting the monthly amount of suitable fertilizers with AI models over historical data, the outliers will differ if we set a 6-month or 12-month task of fertilization over a year. In practice, both of these circumstances are possible and varied with farms. Here, we need to consider outliers associated with tasks, which is an important and common phenomenon facing practical long-term prediction cases.

Specifically, we propose the first RoTS framework, ***Laplacian based RoTS*** (LRoTS), which utilizes LTS to chase the temporal smoothness among $\boldsymbol{p_i}$ and $L2$-norm to measure the "quantity" of $\boldsymbol{r_i}$. The number of outlier tasks is assumed to be small, so we employ the group Lasso (Meier, Van De Geer, and Bühlmann 2008) on column groups of discriminative matrix R to detect outlier tasks. Whereas, LTS only focuses on the smoothness of the prediction models across different time points. Inspired by (Zhou et al. 2022, 2012), we would like to incorporate feature smoothness rather than only task smoothness, so we use FTS to replace LTS to propose the second framework, ***Fused Lasso based RoTS*** (FRoTS), which captures the temporal smoothness not only on task level but also on feature level. In addition, this kind of temporal smoothness based on the extension of fused Lasso has another attractive property, i.e., sparsity continuity (Tibshirani et al. 2005), which is important for us to derive detailed theoretical analyses.

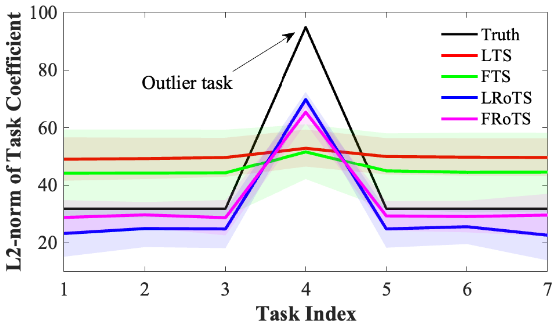

Figure 2: The comparison on **S2** dataset. Both LTS and FTS only have a limited trend to capture the outlier (4th) task.

The main contributions of this work include:

- Our work highlights the importance of outlier tasks in MTL methods and discovers its relationship with temporal smoothness in many real-world applications. We are the first to point out that all MTL models based on TS could not effectively deal with outlier tasks.
- We propose a RoTS assumption to fully utilize both the temporal information between tasks and the specific domain information in outlier tasks. We accomplish this by decomposing the task coefficient and then present two frameworks based on RoTS. Comparing to the model based on TS, our robust frameworks have no additional computational complexity
- Through detailed theoretical analysis and experimental results, we verify the superiority and effectiveness of two RoTS frameworks compared to the TS methods. We discuss several possible specific applications and extensions of our frameworks in broader fields.

Notations: Denote $\mathbb{N}_m = \{1, \cdots, m\}$. $\boldsymbol{x_i}$ a=nd x_{ij} denote the i-th entry of a vector $\boldsymbol{x}$ and the (i, j)-th entry of a matrix X. $\boldsymbol{x^i}$ $(\boldsymbol{x_i})$ denotes the i-th row (column) of a matrix X. $\|X\|_{p,q} = (\sum_{j=1}^{n}(\sum_{i=1}^{m}|x_{ij}|^p)^{q/p})^{1/q}$. $N(\mu, \sigma^2)$ denotes a normal distribution with mean μ and standard deviation σ. $x_{jk}^{(i)}$ and $\boldsymbol{x}_j^{(i)}$ denote the (j, k)-th entry and the j-th column of a matrix X_i. For the implementation code and Appendix, please refer to https://github.com/menghui-zhou/RoTS.

The Proposed Frameworks

Assume that we are given a sequence of time points, the number of which is m. Each time point concerns a task. The training data is $\{(X_1, \boldsymbol{y_1}), \cdots, (X_m, \boldsymbol{y_m}))\}$, where $X_i \in \mathbb{R}^{d \times n_i}$ is the data matrix of the i-th task with each column as a sample; $\boldsymbol{y_i} \in \mathbb{R}^{n_i}$ is the response of the i-th task ($\boldsymbol{y_i}$ has continuous values for regression and discrete values for classification); d is the data dimension; n_i is the number of samples for the i-th task. Denoting $W = [\boldsymbol{w_1}, \cdots, \boldsymbol{w_m}] \in \mathbb{R}^{d \times m}$ as the weight matrix to be estimated, the empirical risk is given by $\mathcal{L}(W) = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{n_i}\sum_{j=1}^{n_i} l((\mathbf{x}_j^{(i)})^T\boldsymbol{w_i}, (y_i)_j)$, where the loss function $l(\cdot,\cdot)$ is squared loss for regression problem and logistic loss for binary classification problem. To learn the m tasks simultaneously, we minimize $\mathcal{L}(W) + \Omega(W)$, where Ω is the regularization term that encodes the prior knowledge.

Laplacian Based Robust Temporal Smoothness

For our RoTS frameworks setting, we decompose the weight matrix $W = P + R$, i.e, $\boldsymbol{w_i} = \boldsymbol{p_i} + \boldsymbol{r_i}$. *The temporal part* $\boldsymbol{p_i}$ satisfies the temporal smoothness $\boldsymbol{p_i} \approx \boldsymbol{p_{i+1}}$. *The discriminative part* $\boldsymbol{r_i}$ represents the difference beyond the temporal relation among tasks. To capture the temporal smoothness, we introduce a regularization term that penalizes large deviations of predictions at neighboring time points; to identify the outlier tasks, we use the group Lasso $l_{2,1}$-norm regularization term. Formally, our first framework is formulated as

$$\min_{P,R} \mathcal{L}(P + R) + \lambda_1 \sum_{i=1}^{m-1} \|\boldsymbol{p_i} - \boldsymbol{p_{i+1}}\|_2^2 + \lambda_2 \|R\|_{2,1},$$

where λ_1 and λ_2 are regularization parameters. For simple notation, we use the following formulation:

$$\min_{P,R} \mathcal{L}(P + R) + \lambda_1 \|PH\|_F^2 + \lambda_2 \|R\|_{2,1}, \tag{1}$$

where $H \in \mathbb{R}^{m \times (m-1)}$ is defined as: $h_{ij} = 1$ if $i = j$, $h_{ij} = -1$ if $i = j + 1$, and $h_{ij} = 0$ otherwise. The regularization term $\|PH\|_F^2$ is also called Laplacian term (Zhou et al. 2011), so we call (1) ***L**aplacian based **Ro**bust **T**emporal **S**moothness framework* (LRoTS).

Fused Lasso Based Robust Temporal Smoothness

Since Laplacian term is differentiable, LRoTS avoids the computational difficulty. However, LRoTS only encourages the smoothness between adjacent tasks. We emphasize that decoupling P into row vectors is usually meaningful. For example, in modeling disease progression scenarios (Zhou et al. 2012; Emrani, McGuirk, and Xiao 2017; Zhou et al. 2022), it is more natural that a feature has similar weights at adjacent time points. So we propose the second framework termed ***F**used Lasso based **Ro**bust **T**emporal **S**moothness* (FRoTS) associated with the following formulation:

$$\min_{P,R} \mathcal{L}(P + R) + \lambda_1 \|FP^T\|_{1,1} + \lambda_2 \|R\|_{2,1}, \tag{2}$$

where $\|FP^T\|_{1,1} = \sum_{i=1}^{d}\sum_{j=1}^{m-1} |p_{i,j} - p_{i,j+1}|$ and $F = H^T$. The term $\|FP^T\|_{1,1}$ is an extension of fused Lasso (Tibshirani et al. 2005) in multi-task setting, that is where the name FRoTS comes from. Compared with LRoTS, FRoTS has another advantage: FPoTS encourages each row of P to get a sparse solution, where sparsity refers to the first difference $|p_{i,j} - p_{i,j+1}|$. It is attractive property for interpretation while LRoTS fails to have. Note that this sparsity property is necessary for us to derive theoretical analyses.

Optimization Algorithm

In this section, we show how to solve the two RoTS frameworks efficiently using the accelerated proximal gradient method (APM) (Li, Fang, and Lin 2020). Denote

$$\mathcal{L}(P, R) = \sum_{i=1}^{m} \frac{1}{n_i} \sum_{j=1}^{n_i} l((X_j^{(i)})^T(\boldsymbol{p_i} + \boldsymbol{r_i}), (y_i)_j), \tag{3}$$

$$\Omega(P, R) = \lambda_1 \Omega(P) + \lambda_2 \Omega(R), \tag{4}$$

where $\Omega(P) = \|PH\|_F^2$ in (1) and $\|FP^T\|_{1,1}$ in (2), $\Omega(R) = \|R\|_{2,1}$. The objective function of two RoTS frameworks is a composite function of a differential term $\mathcal{L}(P, R)$ and a non-differential term $\Omega(P, R)$. Denote

$$T_{Q,S,\eta}(P, R) = \mathcal{L}(Q, S) + \left\langle \frac{\partial \mathcal{L}(Q, S)}{\partial Q}, P - Q \right\rangle +$$

$$\frac{\eta}{2}\|P - Q\|_F^2 + \left\langle \frac{\partial \mathcal{L}(Q, S)}{\partial S}, R - S \right\rangle + \frac{\eta}{2}\|R - S\|_F^2,$$

$$(P^k, R^k) = \arg\min_{P,R} T_{Q^k,S^k,\eta_k}\mathcal{L}(P, R) + \Omega(P, R), \tag{5}$$

where $Q^1 = P^0, S^1 = R^0$ and $Q^k = P^k + \alpha_k(P^k - P^{k-1}), S^k = R^k + \alpha_k(R^k - R^{k-1})$ for $(k \geq 1)$; the value of η_k and α_k applies the strategy in (Beck and Teboulle 2009). According to the theoretical analysis in (Beck and Teboulle 2009; Chen, Zhou, and Ye 2011), we present the following convergence result for our two RoTS frameworks:

Theorem 1 *Let* (P^k, R^k) *be generated by* (5) *where* η_k *satisfies the strategy in (Beck and Teboulle 2009). Then for any* $k \geq 1$, $f(\cdot,\cdot)$ *and* $(P^\star, R^\star)$ *are respectively the objective functions and the optimal solutions of two RoTS formulations* (1) (2)*, we have the optimal convergence rate among the first-order methods:*

$$f(P^k, R^k) - f(P^\star, R^\star) = \mathcal{O}\left(\frac{1}{k^2}\right).$$

Computing the Proximal Operator

A key building block of APM is computing the proximal operator of non-smooth term $\Omega(P, R)$ efficiently. Due to the decomposable property of (5), we cast (5) into the following two separate proximal operator problems:

$$P = \arg\min_{P} \frac{1}{2}\|P - U\|_F^2 + \frac{\lambda_1}{\eta_k}\Omega(P),$$

$$U = Q^k - \frac{1}{\eta_k}\frac{\partial \mathcal{L}(Q^k, S^k)}{\partial Q}), \tag{6}$$

$$R = \arg\min_{R} \frac{1}{2}\|R - V\|_F^2 + \frac{\lambda_2}{\eta_k}\Omega(R),$$

$$V = S^k - \frac{1}{\eta_k}\frac{\partial \mathcal{L}(Q^k, S^k)}{\partial S}. \tag{7}$$

If $\Omega(P) = \|PH\|_F^2$, (6) admits an analytical solution using matrix inverse, but with expensive complexity of $\mathcal{O}(\max(m^3, dm^2))$. We emphasize the matrix $(I + \frac{2\lambda_1}{\eta_k} HH^T) \in \mathbb{R}^{m\times m}$ is tridiagonal and non-singular, this special structure makes us to use the chasing method (Golub and Van Loan 2013) to reduce the complexity to $\mathcal{O}(dm)$. When $\Omega(P) = \|FP^T\|_{1,1}$, (6) no longer admits an analytical solution, however, it can be solved efficiently using FLSA (Fused Lasso Signal Approximation) proposed in (Liu, Yuan, and Ye 2010). It is shown to be scalable to the large-size problem. For updating R, (7) admits closed form solution with the complexity of $\mathcal{O}(dm)$ (Liu, Ji, and Ye 2012). It is concluded that both two frameworks are scalable to large scale datasets using our proposed optimization algorithm.

Theoretical Analysis

Since LTS does not induce the sparsity pattern of the first difference $|p_{i,j} - p_{i,j+1}|$, we do not discuss LRoTS. Here we provide the theoretical analysis of FRoTS.

Basic Assumption

We begin by outlining some fundamental assumptions for the subsequent theoretical analyses. Assume features are normalized, all diagonal elements of the matrix $X_i X_i^T$ equal 1, i.e., $\sum_{k=1}^{n_i}((x_{jk}^{(i)})^2 = 1, \forall j \in \mathbb{N}_d$. Assume that the linear predictive function associated with the i-th task satisfies

$$y_{ji} = f_i^\star(\boldsymbol{x_j^{(i)}}) + \delta_{ji} = (\boldsymbol{x_j^{(i)}})^T \boldsymbol{w_i^\star} + \delta_{ji},$$

where $i \in \mathbb{N}_m, j \in \mathbb{N}_n$, the noise $\boldsymbol{\delta_i} = [\delta_{1i}, \cdots, \delta_{ni}]^T \in \mathbb{R}^n, \delta_{ji} \sim N(0, \sigma^2)$; $X_i = [\boldsymbol{x_1^{(i)}}, \cdots, \boldsymbol{x_n^{(i)}}]^T \in \mathbb{R}^{d\times n}, \boldsymbol{y_i} = [y_{1i}, \cdots, y_{ni}]^T \in \mathbb{R}^n$ are the training data and responses of the i-th task; $W^\star$ is the true weight matrix, decomposed as the sum of two underlying true components $P^\star$ and $R^\star$, i.e., $W^\star = [\boldsymbol{w_1^\star}, \cdots, \boldsymbol{w_m^\star}] = P^\star + R^\star \in \mathbb{R}^{d\times m}$. The true evaluation is

$$\boldsymbol{f_i^\star} = X_i^T \boldsymbol{w_i^\star} = [f_i^\star(\boldsymbol{x_1^{(i)}}), \cdots, f_i^\star(\boldsymbol{x_n^{(i)}})]^T \in \mathbb{R}^n. \quad (8)$$

Thus, we have $\boldsymbol{y_i} = \boldsymbol{f_i^\star} + \boldsymbol{\sigma_i}, i \in \mathbb{N}_m$. We also define the index set $\mathcal{Q}$ and $\mathcal{J}$ for sparsity pattern as

$$\mathcal{Q}(A) = \{(i,j)|a_{ij} \neq 0\}, \mathcal{Q}_\perp(A) = \{(i,j)|a_{ij} = 0\}, \quad (9)$$

$$\mathcal{J}(A) = \{i|\boldsymbol{a_i} \neq 0\}, \mathcal{J}_\perp(A) = \{i|\boldsymbol{a_i} = 0\}. \quad (10)$$

For the sake of simplicity, we assume that the training sample sizes are the same for all tasks; however, the analysis that follows can be easily modified to account for the situation where the training sample sizes differ for various tasks. For notation simplicity, let $X \in \mathbb{R}^{dm\times nm}$ be a block diagonal matrix with $X_i \in \mathbb{R}^{d\times n}(i \in \mathbb{N}_m)$ as the i-th block and $\text{vec}(A) \triangleq [\boldsymbol{a_1}^T, \cdots, \boldsymbol{a_m}^T]^T, A \in \mathbb{R}^{d\times m}$.

Theoretical Analysis for FRoTS

Theorem 2 *Let $(\hat{P}, \hat{R})$ be an optimal solution of* (2) *for $m \geq 2$ and $n, d \geq 1$. Let X_i and y_i satisfy the above assumptions. Take the regularization parameters λ_1 and λ_2 as*

$$\sqrt{2}\lambda_1(m-1), \lambda_2 \geq \alpha, \alpha = \frac{2\sigma}{mn}\sqrt{dm+t}, \quad (11)$$

where $t > 0$ is a universal constant. Then with probability of at least $1 - \exp(-\frac{1}{2}(t - dm\log(1 + \frac{1}{dm})))$, for any $P, R \in \mathbb{R}^{d\times m}$, we have

$$\begin{aligned}\frac{1}{mn}\sum_{i=1}^m \|X_i^T(\boldsymbol{\hat{p}_i} + \boldsymbol{\hat{r}_i}) - \boldsymbol{f_i^\star}\|^2 \\ \leq & \frac{1}{mn}\sum_{i=1}^m \|X_i^T(\boldsymbol{p_i} + \boldsymbol{r_i}) - \boldsymbol{f_i^\star}\|^2 \\ & + 2\sqrt{2}\lambda_1(m-1)\|(P - \hat{P})^T\|_{2,1} \\ & + 2\lambda_2\|(\hat{R} - R)^{\mathcal{J}(R)}\|_{2,1}. \quad (12)\end{aligned}$$

Then (12) can be written as

$$\begin{aligned}\frac{1}{mn}\|X^T\text{vec}(\hat{P} + \hat{R}) - \text{vec}(F^\star)\|^2 \\ \leq & \frac{1}{mn}\|X^T\text{vec}(P + R) - \text{vec}(F^\star)\|^2 \\ & + 2\sqrt{2}\lambda_1(m-1)\|(P - \hat{P})^T\|_{2,1} \\ & + 2\lambda_2\|(\hat{R} - R)^{\mathcal{J}(R)}\|_{2,1} \quad (13)\end{aligned}$$

where $F^\star = [\boldsymbol{f_1^\star}, \cdots, \boldsymbol{f_m^\star}] \in \mathbb{R}^{n\times m}$. We make the following assumption about training data and the weight matrix.

Assumption 1 *For a matrix pair $\Gamma_P \in \mathbb{R}^{d\times m}$ and $\Gamma_R \in \mathbb{R}^{d\times m}$, let r and c $(1 \leq r \leq d(m-1), 1 \leq c \leq m)$ be the upper bounds of $|\mathcal{Q}(FP^{\star T})|$ and $|\mathcal{J}(R^\star)|$, respectively. Let β be positive scalars. Given XX^T is positive definite. There exist positive scalars $k_1(r)$ and $k_2(c)$ such that*

$$k_1(r) \triangleq \min_{\Gamma_P,\Gamma_R \in R(r,c)} \frac{\|X^T vec(\Gamma_P + \Gamma_R)\|}{\sqrt{mn}\|F\|_F\|\Gamma_P\|_F}, \quad (14)$$

$$k_2(c) \triangleq \min_{\Gamma_P,\Gamma_R \in R(r,c)} \frac{\|X^T vec(\Gamma_P + \Gamma_R)\|}{\sqrt{mn}\|(\Gamma_R)^{\mathcal{J}(R)}\|_F}, \quad (15)$$

where the set $R(r,c)$ is defined as

$$\begin{aligned}R(r,c) = & \{\Gamma_P, \Gamma_R \in \mathbb{R}^{d\times m}|\Gamma_P \neq 0, \Gamma_R \neq 0, \\ & |\mathcal{Q}(FP^T)| \leq r, |\mathcal{J}(R)| \leq c, \\ & \|(\Gamma_R)^{\mathcal{J}_\perp(R)}\|_{2,1} \leq \beta\|(\Gamma_R)^{\mathcal{J}(R)}\|_{2,1}\}, \quad (16)\end{aligned}$$

the notations $|\mathcal{J}|$ and $|\mathcal{Q}|$ denote the number of elements in the sets $\mathcal{J}$ and $\mathcal{Q}$ respectively.

Note that Assumption 1 is connected to the restricted eigenvalue assumption, which is essential to (Bickel, Ritov, and Tsybakov 2009). Similar assumptions have also been used in some earlier studies on multi-task learning (Gong, Ye, and Zhang 2012; Chen, Zhou, and Ye 2011; Lounici et al. 2009). The following theorem for performance bounds is a concise statement of our main theoretical finding.

Theorem 3 *Let $(\hat{P}, \hat{R})$ be an optimal solution of* (2) *for $m \geq 2$ and $n, d \geq 1$. Take the regularization parameters λ_1 and λ_2 as in* (11)*. Then under Assumption 1, the following result hold with probability of at least* $1 - \exp(-\frac{1}{2}(t -$

$dm\log(1+\frac{1}{dm}))), t>0$:

$$\frac{1}{mn}\|X^T vec(\hat{P}+\hat{R}) - vec(F^\star)\|^2 \leq \left(\frac{2\lambda_1\sqrt{(m-1)}}{k_1(r)} + \frac{2\lambda_2\sqrt{c}}{k_2(c)}\right)^2, \tag{17}$$

$$\|\hat{R}-R^\star\|_{2,1} \leq \frac{\sqrt{c}(\beta+1)}{k_2(c)}\left(\frac{2\lambda_1\sqrt{(m-1)}}{k_1(r)} + \frac{2\lambda_2\sqrt{c}}{k_2(c)}\right). \tag{18}$$

Theorem 4 *Based on Theorem 3, let*

$$b = \frac{\sqrt{c}(\beta+1)}{k_2(c)}\left(\frac{2\lambda_1\sqrt{(m-1)}}{k_1(r)} + \frac{2\lambda_2\sqrt{c}}{k_2(c)}\right),$$

if the following condition are true:

$$\min_{j\in\mathcal{J}(R^\star)} \|\boldsymbol{r_j^\star}\| > 2b. \tag{19}$$

Define

$$\hat{\mathcal{J}} = \{j \mid \|\boldsymbol{\hat{r}_j}\| > b\}. \tag{20}$$

Then with the same probability, $\hat{\mathcal{J}}$ estimate the true sparsity pattern $\mathcal{J}(R^\star)$. That is $\hat{\mathcal{J}} = \mathcal{J}(R^\star)$.

Theorem 3 gives an essential theoretical guarantee for FRoTS. To be specific, these bounds assess how well FRoTS can approximate the real evaluation values $F^\star$ as well as the real outlier tasks $\boldsymbol{r_i^\star}, i\in\mathbb{N}_m$. Furthermore, we can estimate the true sparsity patterns $\mathcal{J}(R^\star)$ with high probability, i.e., at least $(1-\exp(-\frac{1}{2}(t-dm\log(1+\frac{1}{dm}))))$, if the underlying true weights are above the noise level, i.e, $\min_{j\in\mathcal{J}(R^\star)}\|\boldsymbol{r_j^\star}\| > \frac{2\sqrt{c}(\beta+1)}{k_2(c)}\left(\frac{2\lambda_1\sqrt{(m-1)}}{k_1(r)}+\frac{2\lambda_2\sqrt{c}}{k_2(c)}\right)$.

Experiments

To demonstrate the competitiveness of the proposed approaches, we compare them with Laplacian based temporal similarity (LTS) and fused Lasso based temporal similarity (FTS). The implementation code of all these competitive methods is in the supplementary material. For all the methods, the hyperparameters are selected by grid search and 3-fold cross validation. For each dataset, the experiments on different methods are repeated 5 times by splitting data set randomly, and the mean and standard deviation of the results are reported. Note that for numerical accuracy consideration, we solve the involved formulations with their objective function multiplied by $\sum_{i=1}^m n_i$. The search range of the regularization parameters is $[0.1, 1, 10, 50, 100, 200, 500, 1000, 2500, 5000]$. The root mean square error (rMSE) is used to evaluate the performance of involved methods as used in multi-task learning literature (Yao, Cao, and Chen 2019). We stop the iterative procedure of the algorithms if the change of the objective values in two consecutive iterations is smaller than 10^{-4}. The training ratio is 0.5, defined as the ratio of the training set over the data set.

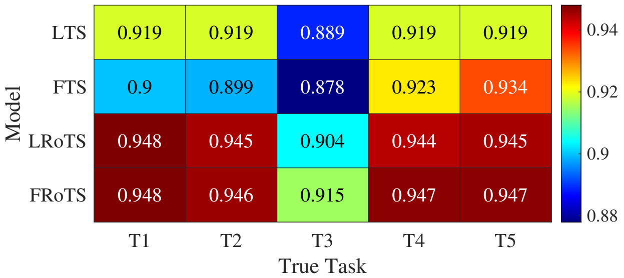

Figure 3: Correlation coefficient with true task on **S1**.

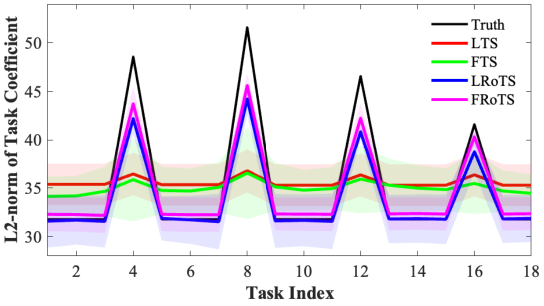

Figure 4: The $L2$-norm of task coefficient on **S3**.

Synthetic Data Sets and Experimental Results

To validate the effectiveness of the proposed approaches in terms of robustness against outlier tasks, we first evaluate our approach on the following three synthetic data sets:

S1: We have 5 tasks ($m=5$), set $\boldsymbol{w_1}=\boldsymbol{w_2}=\frac{2}{3}\boldsymbol{w_3}=\boldsymbol{w_4}=\boldsymbol{w_5}\sim N(0,1)$, hence the 3th task is set as an outlier task. The input data are generated from $X_i\sim N(0,1)$ with feature dimensionality $d=100$, $n_i=100 (i\in\mathbb{N}_5)$, and the output of the i-th task is obtained by $\boldsymbol{y_i}=X_i^T\boldsymbol{w_i}+N(0,1)$.

S2: Denote $\mathbf{1}$ as a vector whose elements are all one. We set 7 tasks ($m=7$), $n_i=20 (i\in\mathbb{N}_7)$, dimensionality $d=20$, $W_1=[\mathbf{1},\cdots,\mathbf{1}]\in\mathbb{R}^{10\times m}$, $W_2=5\cdot[\mathbf{1},\mathbf{1},\mathbf{1},3\cdot\mathbf{1},\mathbf{1},\mathbf{1},\mathbf{1}]\in\mathcal{R}^{10\times m}$, $W=[W_1;W_2]$. Actually, 4-th task $\boldsymbol{w_4}$ is regarded as an outlier task.

S3: This dataset is similar to **S2**, but with 18 tasks ($m=18$). $W_2=5\cdot[\mathbf{1},\mathbf{1},\mathbf{1},\alpha_1\mathbf{1},\mathbf{1},\mathbf{1},\mathbf{1},\alpha_2\mathbf{1},\mathbf{1},\mathbf{1},\mathbf{1},\alpha_3\mathbf{1},\mathbf{1},\mathbf{1},\mathbf{1},\alpha_4\mathbf{1},\mathbf{1},\mathbf{1}]$. $\alpha_1,\alpha_2,\alpha_3,\alpha_4$ are generated from a uniform distribution with the range of $[0.3,0.8]$. It means the 4th, 8th, 12th, and 16th tasks are outliers.

We verify the performance of different methods on **S1** dataset, and we calculate the correlation coefficients between the model parameters learned by different methods and the real model. As shown in Figure 3, the correlation coefficient associated with the 3th task is generally lower than the others, which indicates that the influence of the outlier task is obvious. Note that the correlation coefficients corresponding to LRoTS, and FRoTS are significantly better than those of LTS and FTS, which shows the effectiveness of our proposed methods. However, there is no clear illustration to show how well our methods capture the outlier tasks. To more intuitively analyze the differences between the various methods, we design the two datasets **S2** and **S3**. Note that if we only look at W_1, there is no outlier task. This setting is designed to analyze the difference between the two ways, LTS and FTS, of chasing the temporal information. Since LTS focuses on the task level, and FTS focuses on the feature level (every entry of task coefficient). As shown in Figure 2 and 4, two RoTS frameworks are significantly better than LTS and FTS in detecting outlier tasks, and also better than LTS and FTS in terms of fitting non-outlier tasks. This also indicates that we can not simply average all tasks to chase the temporal information, and both LTS and FTS are too strict.

Real Datasets and Experimental Results

Here we introduce the two used datasets in this work.

SmartFert Dataset The dataset is designed for global soil health assessment. The data are collected from 354 geographic sites from 42 countries. It includes many factors describing agriculture, such as climate, soil type, yield, and fertilization. After data preprocessing, the SmartFert dataset has available data of four farms with same standard and 12 features. The corresponding label is the amount of fertilizer applied for the months of the year, including nitrogen, phosphorus, and potash content. We emphasize that in the SmartFert dataset, heavy fertilization is only applied in the 6th, 7th, 8th, and 9th months. Some farms apply additional fertilization in 11th months. From the perspective of our proposed methods, the 6th, 7th, 8th, and 9th month can be regarded as outlier time points, since the amount of fertilization in these months is extremely different from other months.

Alzheimer's Disease (AD) Dataset This dataset (Jack Jr et al. 2008) consists of three subsets, including RAVLT, MMSE, and ADAS-Cog, ADAS. National Institute of Health (NIH) in 2003 funded the Alzheimer's Disease Neuroimaging Initiative (ADNI) to facilitate the scientific evaluation of neuroimaging data including magnetic resonance imaging (MRI), and clinical and neuropsychological assessments for predicting the onset and progression of mild cognitive impairment (MRI) and AD. The three data sets RAVLT, MMSE, and ADAS are all from ADNI (Weiner et al. 2017). Every dataset has 313 MRI features and corresponding six time points.

Evaluation of Performance

We verify our methods on the SmartFert dataset, the results are shown in Table 1. Note that the variances of the four methods are all large. The possible reason is the sample number of the SmartFert dataset is small and we can not train the model adequately. However, in this scenario with limited data, both two RoTS frameworks achieve significant improvements compared to LTS and FTS, with FRoTS performing the best. Compared to LTS, FRoTS reduces rMSE from 44.26 to 29.22, almost 34% lower. This shows that our methods have greater potential to achieve good performance with limited data. To visualize the detection ability and practical significance of outlier tasks, we compute the l_2-norm of each column of the discriminant matrix R. As shown in Figure 5, the l_2-norm of the 6th, 7th, 8th, 9th, and 11th task is significantly higher than the others and thus can be considered as outlier tasks. It is consistent with the reality, since in the SmartFert dataset, heavily fertilization is only applied in the 6th, 7th, 8th, 9th months. In the 11th month, there is occasional extra fertilizer.

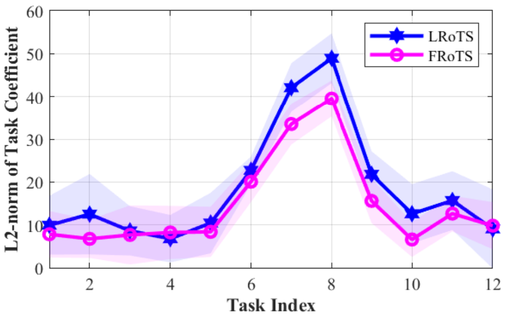

Figure 5: The L2-norm of each column of the matrix R, generated by two RoTS frameworks on the SmartFert dataset.

To analyze the performance of our methods comprehensively, we conduct experiments on AD datasets with training ratio as 0.2. As shown in Table 1, both RoTS methods outperform LTS and FTS clearly. Note that the RoTS frameworks do not improve the baseline on three AD datasets as much as on SmartFert dataset. Possibly because the cognitive scores of AD patients is a somewhat smooth process (Zhou et al. 2022). It tells us the limitation of our methods: The stronger the temporal information is, the more limited improvement will be achieved by our RoTS frameworks. We find that in most cases, FRoTS performs better than LRoTS. This seems to suggest that FRoTS is the better choice of the two metrics. However, we would like to emphasize that, although the optimization algorithm we designed has high efficiency and can be extended to a large-scale dataset, computing the proximal operator of the Fused Lasso penalty is required in FRoTS, which makes FRoTS more complicated than LRoTS. We conclude that if more efficiency is needed, LRoTS is a better option; If better performance is necessary, FRoTS is the better choice.

We also make visual analysis of the detection of outlier tasks on the three AD sub datasets. Refer to Figure 6, the detection result of outlier task on three AD datasets is not as clear as that on SmartFert dataset. That means the temporal relation on the AD dataset is stronger than on SmartFert dataset. And we also notice that there are big differences between the experimental results conducted on the three datasets. For example, on ADAS dataset , the second and third tasks are clearly identified as outlier tasks (right subfigure of Figure 6), but on MMSE dataset, only 2nd task is an obvious outlier task (middle subfigure of Figure 6); on the RAVLT dataset, 1st and 3rd are clear outlier tasks. The reason for this phenomenon may be the differences of the three datasets themselves. For example, the ADAS dataset focuses on the analysis of the patient's language and cognitive ability, the MMSE dataset focuses on the analysis of the patient's arithmetic, memory and direction recognition ability, and the RAVLT dataset focuses on the assessment of

Data set	LTS	FTS	**LRoTS**	**FRoTS**
SmartFert	44.26 ± 10.92	40.95 ± 11.72	35.19 ± 6.27	**29.22** ± **5.39**
RAVLT	4.74 ± 0.22	4.72± 0.25	4.80± 0.51	**4.67** ± **0.37**
MMSE	5.21 ± 0.41	5.13 ± 0.26	**5.01** ± **0.21**	5.07± 0.32
ADAS	9.38 ± 0.03	9.39 ± 0.03	9.35± 0.03	**9.33**± **0.04**

Table 1: The comparison of performance in terms of rMSE ('mean ± std').

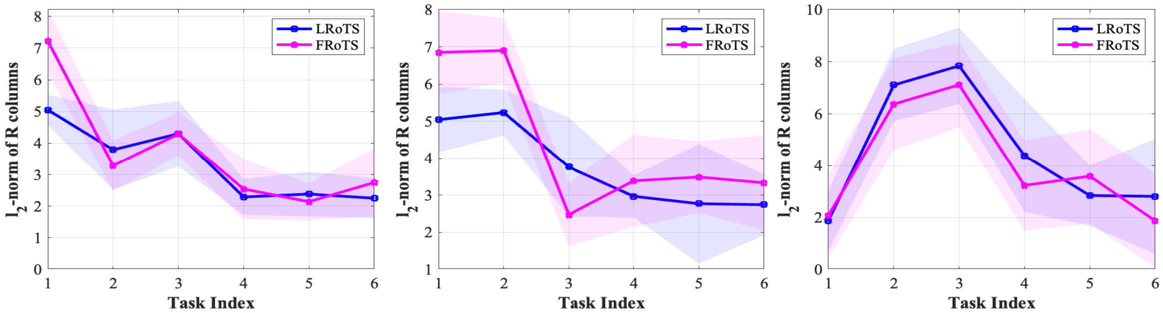

Figure 6: The results of detecting the outlier tasks of LRoTS and FRoTS on RAVLT (left subfigure), MMSE (middle subfigure), and ADAS (right subfigure) datasets.

the patient's learning ability. It is worth emphasizing that, the outlier task mainly appears in the early stages. The possible reason is in initial stages of the disease, the patient's condition is relatively good, but in later stages will deteriorate rapidly because of a rapid loss of many cognitive functions. It makes a huge difference between the state of the patient at the beginning and the state of the patient at other time points.

Possible Specific Application and Extension

We point out that if temporal smoothness assumption (TS) is useful in some scenarios, the two RoTS frameworks are a better option. For instance, (Zhou et al. 2012) proposed cFSGL based on TS for modeling disease progression. We can easily extend cFSGL to: $\mathcal{L}(P + R) + \lambda_1\|P^T\|_{1,1} + \lambda_2\|P^T\|_{2,1} + \lambda_3\|FP^T\|_{1,1} + \lambda_4\|R\|_{2,1}$. It employs the sparse group Lasso ($\lambda_1\|P^T\|_{1,1} + \lambda_2\|P^T\|_{2,1}$) (Simon et al. 2013) to conduct simultaneous joint feature selection for all tasks and selection of a specific set of features for each task. And the FRoTS term ($\lambda_3\|FP^T\|_{1,1} + \lambda_4\|R\|_{2,1}$) is used to capture the robust temporal smoothness. The decomposition property of ($\lambda_1\|P^T\|_{1,1} + \lambda_2\|P^T\|_{2,1} + \lambda_3\|FP^T\|_{1,1}$), proved in (Zhou et al. 2012), enables to compute the proximal operator efficiently and be scalable to the large size problem. Similarly, two RoTS frameworks also have a potential extension on temporal survival model (Wang, Shi, and Reddy 2020).

Our RoTS assumption can be possibly extended to tackle other kinds of sequence data. Gene expression sequence data usually shows some order patterns (Robinson, McCarthy, and Smyth 2010). Tibshirani et al. (Tibshirani et al. 2005) proposed the famous fused Lasso to encourage the orderly successive features to be similar. However, they did not consider the outlier features. We may propose the robust Fused Lasso formulation for tackling it: $\mathcal{L}(\boldsymbol{p} + \boldsymbol{r}) + \lambda_1\|\boldsymbol{p}\|_1 + \lambda_2\|F\boldsymbol{p}\|_1 + \lambda_3\|\boldsymbol{r}\|_1$. Another example is spatio sequence data. Some works (Xu et al. 2016; Gao et al. 2019) utilize the spatio smoothness assumption, which means the closer two objects are, the more similar they are. Similar to RoTS assumption, the robust spatio smoothness assumption is possibly proposed, which simultaneously captures the spatio smoothness and detects outliers.

Conclusion

Temporal smoothness assumption is widely used in multi-task learning setting to simultaneously analyze multiple time points. However, it treats all tasks equally, without considering the difference between them, which means ignoring the negative effect of the outlier tasks. In this paper, we assumed every task consists of one temporal part and one discriminative part. Based on it, we proposed two Robust Temporal Smoothness (RoTS) frameworks that simultaneously chase the temporal smoothness among tasks and capture the outlier tasks, but with no additional computational complexity. The effectiveness of our approach is demonstrated by experimental results and theoretical analyses. Finally, we presented some possible applications in modeling disease progression, tensor multi-task model, and survival model. We also discussed the potential extension of our idea of RoTS frameworks to deal with other kinds of sequence data, like gene expression data and spatio data. Our future work focuses on using these frameworks in a broader area.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (No. 62061050).

References

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

Bickel, P. J.; Ritov, Y.; and Tsybakov, A. B. 2009. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of statistics*, 37(4): 1705–1732.

Chen, J.; Zhou, J.; and Ye, J. 2011. Integrating low-rank and group-sparse structures for robust multi-task learning. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 42–50.

Emrani, S.; McGuirk, A.; and Xiao, W. 2017. 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*, 1457–1466.

Fifty, C.; Amid, E.; Zhao, Z.; Yu, T.; Anil, R.; and Finn, C. 2021. Efficiently identifying task groupings for multi-task learning. *Advances in Neural Information Processing Systems*, 34.

Gao, Y.; Zhao, L.; Wu, L.; Ye, Y.; Xiong, H.; and Yang, C. 2019. Incomplete label multi-task deep learning for spatio-temporal event subtype forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 3638–3646.

Golub, G. H.; and Van Loan, C. F. 2013. *Matrix computations*. JHU press.

Gong, P.; Ye, J.; and Zhang, C. 2012. Robust multi-task feature learning. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, 895–903.

Jack Jr, C. R.; Bernstein, M. A.; Fox, N. C.; Thompson, P.; Alexander, G.; Harvey, D.; Borowski, B.; Britson, P. J.; L. Whitwell, J.; Ward, C.; et al. 2008. The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 27(4): 685–691.

Li, H.; Fang, C.; and Lin, Z. 2020. Accelerated first-order optimization algorithms for machine learning. *Proceedings of the IEEE*, 108(11): 2067–2082.

Liu, J.; Ji, S.; and Ye, J. 2012. Multi-task feature learning via efficient l2, 1-norm minimization. *arXiv preprint arXiv:1205.2631*.

Liu, J.; Yuan, L.; and Ye, J. 2010. An efficient algorithm for a class of fused lasso problems. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, 323–332.

Lounici, K.; Pontil, M.; Tsybakov, A. B.; and Van De Geer, S. 2009. Taking advantage of sparsity in multi-task learning. *arXiv preprint arXiv:0903.1468*.

Meier, L.; Van De Geer, S.; and Bühlmann, P. 2008. The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1): 53–71.

Nie, L.; Zhang, L.; Meng, L.; Song, X.; Chang, X.; and Li, X. 2016. Modeling disease progression via multisource multitask learners: A case study with Alzheimer's disease. *IEEE transactions on neural networks and learning systems*, 28(7): 1508–1519.

Robinson, M. D.; McCarthy, D. J.; and Smyth, G. K. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1): 139–140.

Romeo, L.; Armentano, G.; Nicolucci, A.; Vespasiani, M.; Vespasiani, G.; and Frontoni, E. 2020. A Novel Spatio-Temporal Multi-Task Approach for the Prediction of Diabetes-Related Complication: a Cardiopathy Case of Study. In *IJCAI*, 4299–4305.

Saha, T. K.; Williams, T.; Hasan, M. A.; Joty, S.; and Varberg, N. K. 2018. Models for capturing temporal smoothness in evolving networks for learning latent representation of nodes. *arXiv preprint arXiv:1804.05816*.

Shen, J.; Zhen, X.; Worring, M.; and Shao, L. 2021. Variational Multi-Task Learning with Gumbel-Softmax Priors. *Advances in Neural Information Processing Systems*, 34.

Simon, N.; Friedman, J.; Hastie, T.; and Tibshirani, R. 2013. A sparse-group lasso. *Journal of computational and graphical statistics*, 22(2): 231–245.

Tibshirani, R.; Saunders, M.; Rosset, S.; Zhu, J.; and Knight, K. 2005. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1): 91–108.

Wang, P.; Shi, T.; and Reddy, C. K. 2020. Tensor-based Temporal Multi-Task Survival Analysis. *IEEE Transactions on Knowledge and Data Engineering*.

Wei, W. W. 2006. Time series analysis. In *The Oxford Handbook of Quantitative Methods in Psychology: Vol. 2*.

Weiner, M. W.; Veitch, D. P.; Aisen, P. S.; Beckett, L. A.; Cairns, N. J.; Green, R. C.; Harvey, D.; Jack Jr, C. R.; Jagust, W.; Morris, J. C.; et al. 2017. Recent publications from the Alzheimer's Disease Neuroimaging Initiative: Reviewing progress toward improved AD clinical trials. *Alzheimer's & Dementia*, 13(4): e1–e85.

Xu, J.; Tan, P.-N.; Luo, L.; and Zhou, J. 2016. Gspartan: a geospatio-temporal multi-task learning framework for multi-location prediction. In *Proceedings of the 2016 SIAM International Conference on Data Mining*, 657–665. SIAM.

Xu, Y.; Sun, S.; Zhang, H.; Yi, C.; Miao, Y.; Yang, D.; Meng, X.; Hu, Y.; Wang, K.; Min, H.; et al. 2021. Time-aware graph embedding: A temporal smoothness and task-oriented approach. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 16(3): 1–23.

Yao, Y.; Cao, J.; and Chen, H. 2019. 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*, 1408–1417.

Zhang, Y.; and Yang, Q. 2021. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*.

Zhao, L.; Li, X.; Xiao, J.; Wu, F.; and Zhuang, Y. 2015. Metric learning driven multi-task structured output optimization for robust keypoint tracking. In *Twenty-ninth AAAI conference on artificial intelligence*.

Zheng, J.; and Ni, L. M. 2013. Time-dependent trajectory regression on road networks via multi-task learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 27, 1048–1055.

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

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

Zhou, M.; Zhang, Y.; Liu, T.; Yang, Y.; and Yang, P. 2022. 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*, 2743–2752.