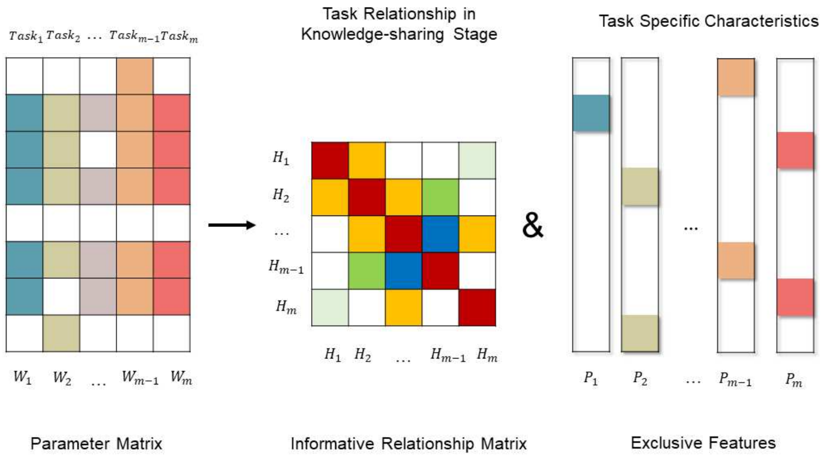
Graphical Abstract

Informative Relationship Multi-task Learning: Exploring Pairwise Contribution across Tasks' Sharing Knowledge

Xiangchao Chang, Menghui Zhou, Xulong Wang, YunYang, Po Yang


$Task_1$ $Task_2$ … $Task_{m-1}$ $Task_m$
W_1 W_2 … W_{m-1} W_m
Parameter Matrix
Task Relationship in Knowledge-sharing Stage
H_1
H_2
…
H_{m-1}
H_m
H_1 H_2 … H_{m-1} H_m
Informative Relationship Matrix
&
Task Specific Characteristics
P_1 P_2 … P_{m-1} P_m
Exclusive Features

Highlights

Informative Relationship Multi-task Learning: Exploring Pairwise Contribution across Tasks' Sharing Knowledge

Xiangchao Chang, Menghui Zhou, Xulong Wang, YunYang, Po Yang

- We propose an informative relationship multi-task learning framework designed to quantify the mutual knowledge contributions in joint learning, while simultaneously capturing the unique characteristics of each task to ensure diversity.
- We propose a sparse informative relationship learning model to enhance efficacy, to handle the sparsity over the shared knowledge, particularly in scenarios characterized by a large number of tasks.
- We propose an optimization algorithm leveraging accelerated gradient descent methodologies to effectively address point-wise regularization.

Informative Relationship Multi-task Learning: Exploring Pairwise Contribution across Tasks' Sharing Knowledge

Xiangchao Chang[a], Menghui Zhou[b], Xulong Wang[b], YunYang[a], Po Yang[b]

[a] *School of Software, Yunnan University, Kunming, 650500, Yunnan, China*
[b] *Department of Computer Science, University of Sheffield, Sheffield, S10 2TN, South Yorkshire, England*

Abstract

Multi-task learning is a machine learning paradigm, that aims to leverage useful domain information to help improve the generalization performance of all tasks. Learning the relationships of tasks helps to identify the latent tasks' associations and access a better performance. However, most of the existing methods hardly pay attention to the determination of knowledge interaction among tasks and instead concentrate solely on certain aspects of task affinity. This compulsory similarity among all tasks leads to deficiencies in both task diversity and model robustness. To address this issue, we emphasize the task relationships within mutual information interaction. We propose a regularized framework from an informative perspective to quantify pairwise contributions during the knowledge-sharing stage, meanwhile utilizing an exclusive Lasso to identify the exclusive characteristics of tasks. An efficient optimization algorithm is developed to solve the proposed objective function. Detailed theoretical analyses and extensive experiments on both synthetic and real-world datasets are provided to demonstrate the effectiveness of our proposed method.

Keywords: Multi-task Learning, Relationship Learning, Sparse Learning

1. Introduction

Multi-Task Learning (MTL) [1, 2], a machine learning paradigm inspired by human learning activities [3], learns multiple related tasks jointly to make a better performance by leveraging the contained knowledge. The most critical issue in MTL is to identify the complex task relations and incorporate the intrinsic relationship into the learning strategy and process estimation to improve the performance and interpretation of all tasks [3, 4].

A common way to utilize the underlying association is seen as the prior knowledge in many scenes. Relying on the selection of experts in the relevant fields [5, 6, 7], experimental works show the benefits of such MTL methods relative to individual task learning. [8] forces the model parameters of each task to approach the average model parameters of all the tasks. Furthermore, task similarity is proposed in [9] to guide joint learning, which means that the more similar two tasks are, the closer the corresponding model parameters are expected to be. The similarity of tasks is also used in modelling the longitudinal progression to approximate the chronological association in [10, 11]. Temporal smoothness methods [12, 13] employ relatively small variations in progression modelling [14]. Spatial association, as discussed by [15], is coupled with temporal relations in spatio-temporal models [16, 17], employing 2-D dependencies to jointly learn relevant tasks. Leveraging task similarity across domains enables effective joint learning and reveals dependencies in selected inspections. As another line of effort, the feature-based MTL methods focus on learning the common feature subset shared by tasks in some structures. The low-dimensional representation in [18] and multi-task feature selection method in [19] learn the features shared across a set of multiple related tasks based on group sparsity regularization. In [20], a shared predictive feature structure learning method is proposed. With the same assumption that overlap features exist, a dirty model in [21] is proposed. The low-rank constraints have been proposed to learn the latent structure of subspace in [22, 23]. It should be noticed that not all tasks share a common feature set or very few features participate in all tasks. Robust methods are proposed in [24, 25] to capture the common feature sets meanwhile modelling the outlier features. Another viewpoint on this problem is that not all learning tasks are related to each other and benefit from joint learning in MTL. Clustered-MTL methods proposed in [26, 27, 28] aim to determine the task groups, each of which consists of similar tasks and learn the latent common structure simultaneously.

However, in real-world applications, it is hard to identify the task association in some respects, and even harder to obtain a clear feature structure shared by tasks. So there are some MTL learning methods were proposed to learn the task relations automatically by the similarity of tasks' parameters to explicit relations with interpretability. A tree structure in [29] is employed to capture the hierarchical relations. Bonilla et al. [30] propose a multi-task Gaussian process model that defines a prior on functional values and uses a covariance-based method to model task relations via a covariance matrix Ω computed from a kernel function. To avoid overfitting from strict Gaussian process assumptions, Zhang and Yeung [31] propose a generalized t process method with a weight-space view, utilizing an inverse-Wishart prior distribution on Σ to generate Ω, where the degree of freedom is determined by the maximum mean discrepancy. Subsequent works [32, 33, 34, 35, 36] introduce regularized frameworks with a matrix-variate normal prior over the parameter matrix, leveraging flexible norm forms to estimate the relation matrix. Based on the likelihood function, sparse prior over the covariance matrix Ω, methods proposed in [35, 37] learned the sparse relations among tasks when the number of tasks is large. And asymmetric relations learning in [33, 38], incorporated with directed acyclic graphs model thus can model complex structure of tasks parameters.

While, most of the existing methods in task relation learning are based on the similarity in some fields, such as prior knowledge, common features, distance of parameters and so on. To quantify the pairwise correlation in an appropriate way is critical, but the matching degree of joint learning has not been taken into consideration. It may cause several following problems. First, the attention to the affinity of tasks brings compulsory similarity constraints over all tasks. It is obvious in the utilization of prior knowledge, and the variations of tasks in the selected inspects forced to be relatively small. Learning relations with automatic methods like graphical models [30, 35] seem no similarity constraints in any way, but the previous means are based on the regularization of model parameter distance, thus providing a constraint of all task diversity. Second, the similarity constraints utilize similar tasks and force the tasks' parameters to lie in a short range, and the similarity may go against the representation of the specific characteristics of tasks. It is indeed to capture the exclusive features meanwhile leveraging the domain knowledge. Third, most of the proposed methods only model the task relation in some ways, ignoring the shared knowledge that may be in some dimensions and varying degrees thus the detailed characterization must bring a high cost in both expert knowledge and calculation.

To address the issues outlined above, an intuitive approach is to discern both the commonalities shared across all tasks and the unique characteristics specific to each task. A comprehensive method for learning relationships should not merely emphasize certain aspects across all tasks, but also differentiate between knowledge that is universally applicable and that which is task-specific. Drawing from the 'informative learning' perspective as outlined in [39] to discern 'what to know' within the established MTL framework, we introduce the concept of the 'informative relationship'. This concept signifies the relationships established during the knowledge interaction phase of MTL and aids in quantifying the mutual information of each subtask. The intuitive nature of this definition facilitates the exploration of methods for relation learning across tasks without imposing compulsory similarity constraints on the unified parameter representation. Essentially, decomposing the learning stage aids in exploring the intrinsic relationships where tasks are genuinely associated.

In this paper, we propose a Multi-Task Informative Relationship Learning (MIRTL) method for exploring the knowledge interaction of tasks and the exclusive characteristics captured. With the simple and intuitive decomposition, we resolve the conflict described above, highlighting the following contributions:

- We propose an informative relationship multi-task learning framework aimed at emphasizing mutual information interactions among tasks in MTL relation learning, while simultaneously capturing their exclusive characteristics to ensure task diversity.
- A sparse informative relationship multi-task learning model was proposed to enhance the appropriateness of task relations. Furthermore, an optimization algorithm was introduced, leveraging accelerated gradient descent to address point-wise regularization.
- Following a comprehensive theoretical analysis and experimental investigation on both synthetic and real-world datasets, the exceptional performance of our framework in target prediction and relation estimation was confirmed. Furthermore, several potential extensions and limitations of our proposed framework were discussed.

Notations: Denote $\mathbb{N}_m = \{1, ..., m\}$. For any $A = [a_1, ..., a_m] \in \mathbb{R}^{d \times m}$, let $a_i \in \mathbb{R}^d$ be the $i-$th column of A; denote by $||a_i||_1$ the l_1-norm of a_i; $||A||_{p,q} = (\Sigma_{j=1}^{n}(\Sigma_{i=1}^{m}|X_{ij}|^{p})^{q/p})^{1/q}$; denote $tr(B) = \Sigma_{i=1}^{n} b_{ii} = b_{11} + b_{22} + ... + b_{nn}$ is the trace of square matrix B; denote the $a_{jk}^{(i)}$ and $a_j^{(i)}$ are the $(j, k)-$th entry and the $j-$th column of matrix A_i; $N(\mu, \sigma^2)$ denotes a normal distribution with mean μ and standard deviation σ.

2. Related work

2.1. Robust multi-task learning

Robust multi-task learning [24] devotes to identifying the irrelevant (outlier) tasks. In robust MTL, the multiple tasks are divided into two groups, the related tasks group and the irrelevant (outlier) tasks. A traditional definition of outlier tasks is 'lies an abnormal distance from the other values in a random from a population'. Other inspects of robustness method include robust feature MTL in [25], robust temporal smoothness in [13] and so on. The robustness based MTL methods take the knowledge shared mechanism into joint learning, and discriminate the tasks' parameters according to the distance to identify the similarity of tasks. While it may be too strict to judge the similarity with only parameter distance, the most glaring flaw is the distance could not reflect the linear correlation. To this end, we reformulate the decomposition method and make an intuitive judgement by multi-task learning mechanism. The biggest difference between us is that we quantify the pairwise relationship in knowledge sharing stage rather than a single distance criterion.

2.2. Multi-task relationship learning

In MTL, tasks are related and the relatedness can be quantitated via similarity, correlation, covariance and so on [3]. As mentioned above, the multi-task relationship learning methods focus on quantifying the similarity of tasks' learned parameters. To learn the task relationship automatically from data, [30] propose a covariance based method built on Gaussian progress. In [35], a sparse prior on the relation matrix is employed to learning the sparse correlation when the number of tasks is pretty large. An exclusive relation learning method in [40] is proposed in a longitudinal research of disease, and applied to finding out the exclusive characteristics of each task. The above MTL relationship learning methods provide an appropriate approach to capturing the pairwise relations of tasks by a graphical model. While the former methods focus on learning the relations in a unitary parameter space, the coherence similarity exists, resulting in a shift to the graphical centre inevitably. The motivation of our work is reasonable in that the intrinsic relationship should only exist in the knowledge sharing stage, and the informative relationship is conducive to finding out the pairwise contribution and comprehension of multi-task learning mechanism.

3. Methods

3.1. Proposed Methods

Assume that there are m learning tasks, associated with the sample data $\{(X_1, y_1), ..., (X_m, y_m)\}$ are given, 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; $y_i \in \mathbb{R}^{n_i}$ is the response target of the $i-$th task(continuous value of y_i for regression tasks and discrete values for classification); the data dimensionality is denoted by d; the number of samples for $i-$th task is denoted by n_i. Denote $W = [W_1, ..., W_m] \in \mathbb{R}^{d \times m}$ is the weight matrix to be estimated, the empirical risk is defined by $\mathcal{L}(W) = \frac{1}{m}\Sigma_{i=1}^{m}\frac{1}{n}\Sigma_{j=1}^{n_i} l((x_j^{(i)^T} w_i, (y_i)_j))$, where the loss function $l(\cdot, \cdot)$ can be selected as squared loss for regression tasks and logistic loss for binary classification tasks. For each task and decomposing the weight matrix W into two components H and P ($W = H + P$) to capture the task relation in learned parameter by domain sharing knowledge to determine the informative relationship, meanwhile capturing the task specific characteristics by P to find out the exclusive feature in different subtasks.

In the following, we proposed a regularized framework for learning multiple task informative relations and identifying tasks' specific characteristics simultaneously:

$$\min_{H,P} \mathcal{L}(W) + \lambda_1 tr(H\Omega^{-1}H^T) + \lambda_2 ||P||_{1,2} \quad s.t. \quad W = H + P \tag{1}$$

where $\Omega \succ \mathbf{0}$ denotes a square positive definite(or positive semidefinite) covariance matrix to capture the relations among tasks; and its inverse(or pseudoinverse) by Ω^{-1}; λ_1 and λ_2 are regularization parameters.

The multi-task informative relations learning (MTIRL) framework proposed in (1) contains three terms. The first terms measures the empirical loss by selected loss function $l(\cdot, \cdot)$ based on the training data. The second term is a regularizer on H. And an intuitive explanation can be formed as $tr(H\Omega^{-1}H^T) = tr(H\Upsilon\Upsilon^T H^T) = tr((H\Upsilon)^T(H\Upsilon)) = ||H\Upsilon||_F^2$, where $\Upsilon\Upsilon^T = \Omega^{-1}$. Similar but different in [11, 13], prior knowledge is employed to model task relations using the Laplacian matrix. To automatically estimate informative relationships among tasks, akin to [35, 40], we interpret the positive semidefinite square matrix Ω as a covariance matrix to capture pairwise task relations during the parameter sharing stage. This matrix can be viewed as analogous to Gaussian process modelling for covariance matrix estimation and reflects the mutual information of tasks by a square logarithmic mapping.

A Gaussian prior is selected to learn the informative relationships because it aligns with the central limit theorem of the model parameters. The informative relationship matrix, denoted as Ω, distinguishes between multi-task learning and single-task learning scenarios. When Ω becomes diagonal, it indicates minimal knowledge interaction across all tasks, resembling single-task learning.

The third component $l_{1,2}-$norm for modeling task specific characteristics with the exclusive lasso term in [40, 41]. Noticed that it is different from the $l_{2,1}-$norm in [11, 42] which selects the common features shared across all tasks. The exclusive lasso constraint provides a competitive relation among the features of every subtask. We argue the exclusive relation learning in [40], for its confusion in intrinsic relationship learning which misjudges the information interaction mechanism in multi-task learning. The proposed MTIRL framework not only captures the intrinsic relation at the interaction stages but also encourages the exclusive feature representation of tasks. As a result, the two above components achieve mining the dependency in multi-task learning to find out the latent associations and help to reveal the exclusive characteristic conveyed in parameter representation space at a particular subtask.

Under the covariance-based relation learning framework in (1), we propose a method to learn tasks' informative relations with a focus on sparsity, acknowledging that not all knowledge contributes equally to joint learning. Similarly, in scenarios with a large number of tasks, the benefits of one task may not extend to all others, potentially increasing model complexity and the risk of overfitting. Effectively capturing sparse task relations thus warrants imposing sparsity on the covariance matrix Ω.

Here we propose the multi-task sparse informative relations learning (MTSIRL) method to capture the informative task relations under the high dimensional tasks with a tight estimation induced by sparse prior. Specially, the objective function of the MTSIRL method is formulated as

$$\underset{H,P}{min}\,\mathcal{L}(W)+\lambda_1 tr(H\Omega^{-1}H^T)+\lambda_2||P||_{1,2}+\lambda_3||\Omega||_1 \quad s.t. \quad W=H+P \tag{2}$$

where the λ_3 is the parameter of the sparse inducing norm l_1-norm on Ω. An intuitive explanation is $||\Omega||_1 = tr(\Omega)+\sum_{i=1}^{m}\sum_{j=1,j\neq i}^{m}|\omega_{ij}|$, so the method (2) could learn the sparse relations by penalization of the off-diagonal entries in covariance matrix Ω. The placement of l_1 regularization on Ω, we restrict the complexity of Ω, and the learned Ω with zero entries Ω_{ij} implies the corresponding tasks with index i and j are no related.

3.2. Optimization Algorithm

In this section, we show the optimization solution of the MTIRL framework. To solve the problem (1) and (2), the alternating descent method in [43] is utilized to access the optimal convergence rate. By alternating optimize H and P with fixed Ω as proposed above, and optimize the Ω by fixed H and P. First, we introduce the optimization algorithm based on accelerated proximal gradient(AGM) for the update step with fixed Ω. Denote

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

$$\mathcal{G}(H,P) = \lambda_1 tr(H\Omega^{-1}H^T) + \lambda_2 ||P||_{1,2} \tag{4}$$

Noted that the objective function in problem (1) is a composite function of a differential term $\mathcal{L}(W)$ and a non-differential term $\mathcal{G}(H,P)$. Denote

$$\begin{aligned} &\mathcal{T}_{R,S,\eta}(R,S) = \mathcal{L}(R,S) + \langle \frac{\partial L(R,S)}{\partial R}, H-R\rangle + \\ &\frac{\eta}{2}||H-R||_F^2 + \langle \frac{\partial L(R,S)}{\partial S}, P-S\rangle + \frac{\eta}{2}||P-S||_F^2 \\ &(H^k,P^k) = \underset{H,P}{argmin}\, \mathcal{T}_{R,S,\eta}\mathcal{L}(W) + \mathcal{G}(H,P) \end{aligned} \tag{5}$$

where $R^1 = H^0, S^1 = P^0$ and $R^k = H^k + \alpha_k(H^k - H^{k-1})$, $S^k = P^k + \alpha_k(P^k - P^{k-1})$ for $k \geq 1$, the value of η_k and α_k apply the strategy in [44], we have the optimal convergence rate among the first-order methods with $O(\frac{1}{k^2})$.

When the H and P are fixed, the update of Ω in problem (1) can be solved with an analytical solution, and the problem with respect to Ω in problem (2) is formulated as

$$\underset{\Omega \succ \mathbf{0}}{min}\, \lambda_1 tr(H\Omega^{-1}H^T) + \lambda_3||\Omega||_1 \tag{6}$$

To solve the problem (6), we utilize the similar method of AGM in [44], and for the optimal Ω, we have the following analysis based on primal dual construction in [45].

Theorem 1. *The optimal Ω in problem (6) satisfies*

$$\Omega \succeq \frac{eig_{min}(H)}{\sqrt{m(\lambda_3/\lambda_1)}} I \tag{7}$$

where $eig_{min}(H)$ denotes the minimum eigenvalue of H; I is identity matrix. The lower bound of Ω depends on the smallest eigenvalue of H. So the optimal Ω is positive definite when the H is of rank m and otherwise positive semidefinite. Thus the covariance matrix Ω can be viewed as calculated by the shared parameter among tasks.

3.3. Implementation Details

The calculation of the proximal operator with non-smooth terms in $\mathcal{G}(H, P)$ and problem (6) are pivotal building blocks of APM. By utilizing the alternating descent algorithm for the decomposed terms H and P. We show the three separate proximal operator problems to update H, P and Ω, the detailed update strategies for calculation are as follow

$$\begin{aligned} H &= \underset{H}{argmin}\,\frac{1}{2}||H-U||_F^2 + \frac{\lambda_1}{\eta_k}\mathcal{G}(H) \\ U &= R^k - \frac{1}{\eta_k}\frac{\partial \mathcal{L}(R,S)}{\partial R} \end{aligned} \tag{8}$$

$$\begin{aligned} P &= \underset{P}{argmin}\,\frac{1}{2}||H-V||_F^2 + \frac{\lambda_2}{\eta_k}\mathcal{G}(P) \\ V &= S^k - \frac{1}{\eta_k}\frac{\partial \mathcal{L}(R,S)}{\partial S} \end{aligned} \tag{9}$$

$$\begin{aligned} \Omega &= \underset{\Omega}{argmin}\,\frac{1}{2}||\Omega-\hat{\Omega}||_F^2 + \frac{\lambda_3}{\eta_k}||\Omega||_1 \\ \hat{\Omega} &= \Omega^k - \frac{\lambda_1}{\eta_k}\frac{\partial tr(H\Omega^{-1}H^T)}{\partial \Omega} \end{aligned} \tag{10}$$

The update step of H can be solved analytically with a space complexity of $O(m^3(1+d^3))$. Similarly, the update step of P involves a typical $l_{1,2}$-norm proximal operator calculation as described in [41], with a space complexity of $O(m^2 d)$. Additionally, the update step of Ω with a Lasso constraint can be solved using the soft-thresholding operator $k(a,b) = \mathrm{sign}(a)\cdot\max(0,|a|-b)$, with a space complexity of $O(m^3)$. Here, sign denotes the sign function, and $|\cdot|$ represents the absolute value. It's worth noting that all the mentioned space complexities are based on the analytical solution of each updating step and can be further reduced through decomposition based on the data scale. Specially, The partial derivative step in (10) can be calculated as

$$\frac{\partial tr(H\Omega^{-1}H^T)}{\partial \Omega} = -\Omega^{-1}H^T H\Omega^{-1} \tag{11}$$

Meanwhile, we provide the analytical solution for updating Ω in the problem (1) with fixed H and P as above.

4. Properties

Since MTIRL does not include the sparsity constraints over the covariance matrix Ω, it can be seen as a special form when $\lambda_3 = 0$, so we provide the theoretical analysis of MTSIRL.

4.1. Basic Assumption

In the following theoretical analyses, we make basic assumptions on the data and model to ensure generality and adaptability. Firstly, assume the normalized data satisfies the orthogonality thus the $(j,k)-$th entry of X_i denoted as $x_{jk}^{(i)}$ satisfies $\sum_{k=1}^{n_i}(x_{jk}^{(i)})^2 = 1, \forall j \in \mathbb{N}_d$. And assume that a linear model with Gaussian noise gives the responses that satisfy

$$y_{ji} = f_i^*(x_j^{(i)}) + \delta_{ji} = (x_j^{(i)})^T w_i^* + \delta_{ji} \tag{12}$$

where $i \in \mathbb{N}_T, j \in \mathbb{N}_n$, the true weight matrix W^* decomposed as the sum of two underlying true components H and P, $W^* = [w_1^*, ..., w_m^*] = H^* + P^* \in \mathbb{R}^{d\times m}$;$X_i = [x_1^{(i)}, ..., x_n^{(i)}] \in \mathbb{R}^{d\times n}, y_i = [y_1^{(i)}, ..., y_n^{(i)}] \in \mathbb{R}^n$ are respectively the training data and targets of the $i-$th task; the noise $\delta_i = [\delta_{1i}, ..., \delta_{ni}]^T \in \mathbb{R}^n, \delta_{ji} \sim N(0, \sigma^2)$; The true evaluation under i.i.d. normal noise is

$$f_i^* = X_i^T w_i^* = [f_i^*(x_1^{(i)})^T, ..., f_i^*(x_n^{(i)})^T] \in \mathbb{R}^n \tag{13}$$

Thus, we have $y_i = f_i^* + \delta_i, i \in \mathbb{N}_m$; and define the index set for sparsity pattern as

$$\mathcal{Q}(A) = \{(i,j)|a_{ij} \neq 0\}, \mathcal{Q}_{\mathcal{C}}(A) = \{(i,j)|a_{i,j} = 0\}, \tag{14}$$

The assumption of a Gaussian distribution aligns with the parameter distribution under the central limit theorem, ensuring generalizability when the sample size is sufficient, meanwhile corresponding to the least biased distribution. The assumption made in this context is that training sample sizes for all tasks are considered equal for simplicity. It is noteworthy that the derivation presented below can be readily extended to scenarios where the training sample size varies across tasks. Based on the above assumptions, we further restrict the eigenvalue [46], similar to some previous studies on MTL[24, 25, 13].

Assumption 1. *For a matrix pair $\Gamma_{sh} \in \mathbb{R}^{d*m}$ and $\Gamma_{sp} \in \mathbb{R}^{d*m}$, let r and c $(1 \leq r \leq d, 1 \leq c \leq m)$ be the upper bounds of $|\mathcal{Q}(H)|$ and $|\mathcal{Q}(P)|$, respectively. Let β_1, β_2 be positive scalars. Given XX^T is positive definite. There exist positive scalars $k_1(r)$ and $k_2(c)$ such that*

$$k_1(r) \triangleq \min_{\Gamma_H, \Gamma_P \in R(r,c)} \frac{||X^T vec(\Gamma_H + \Gamma_P)||}{\sqrt{mn}||\mathcal{Q}_1(\Gamma_H)||_F} \tag{15}$$

$$k_2(c) \triangleq \min_{\Gamma_H, \Gamma_P \in R(r,c)} \frac{||X^T vec(\Gamma_H + \Gamma_P)||}{\sqrt{mn}||\mathcal{Q}_2(\Gamma_P)||_F} \tag{16}$$

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

$$\begin{aligned} R(r,c) = \{\Gamma_H, \Gamma_P \in \mathbb{R}^{d*m} | \Gamma_H \neq 0, \Gamma_P \neq 0, |\mathcal{Q}_1(H)| \leq r, |\mathcal{Q}_2(P)| \leq c, \\ ||\mathcal{Q}_{C1}(\Gamma_H)||_2 \leq \beta_1 ||\mathcal{Q}_1(\Gamma_H)||_2, ||\mathcal{Q}_{C2}(\Gamma_P)||_{1,2} \leq \beta_1 ||\mathcal{Q}_2(\Gamma_P)||_{1,2}\} \end{aligned} \tag{17}$$

4.2. Theoretical Analysis

Our main theoretical result is summarised in the following theorem for performance bounds. Based on the ALM method in the optimization of Ω and H and P, we give the theoretical analysis as follows.

Theorem 2. *Let$(\hat{H}, \hat{P})$ be an optimal solution of (2) under the optimal Ω in Theorem 1 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*

$$m\sqrt{m\lambda_1\lambda_3}, \lambda_2 \geq \alpha, \alpha = \frac{2\sigma}{mn}\sqrt{dm + t} \tag{18}$$

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

$$\begin{aligned} \frac{1}{mn}\sum_{i=1}^{m} ||X_i^T(\hat{H}_i + \hat{P}_i) - f_i^*||^2 \leq \frac{1}{mn}\sum_{i=1}^{m} ||X_i^T(H_i + P_i) - f_i^*||^2 \\ + 2m\sqrt{m\lambda_1\lambda_3}||\mathcal{Q}(\hat{H} - H)||_2 + 2\lambda_2||\mathcal{Q}_2(\hat{P} - P)||_{1,2} \end{aligned} \tag{19}$$

Then (19) can be written as

$$\begin{aligned} &\frac{1}{mn}\sum_{i=1}^{m} ||X_i^T vec(\hat{H} + \hat{P}) - vec(F^*)||^2 \\ &\leq \frac{1}{mn}\sum_{i=1}^{m} ||X_i^T vec(H_i + P_i) - vec(F^*)||^2 \\ &+ 2m\sqrt{m\lambda_1\lambda_3}||\mathcal{Q}_1(\hat{H} - H)||_2 + 2\lambda_2||\mathcal{Q}_2(\hat{P} - P)||_{1,2} \end{aligned} \tag{20}$$

where $F^* = [f_1^*, ..., f_m^*] \in \mathbb{R}^{n*m}$.

Based on Theorem 2 and Assumption 1, we derive the following theorem that lies in the crucial theoretical assurances it offers for MTSIRL. These bounds serve a purpose by assessing the accuracy of our proposed model in approximating the true evaluation values in approximating the true weight matrices ($H^*, P^*, W = H^* + P^*$) and the probability of sparse patterns recovery.

Theorem 3. *Let* $(\hat{H}, \hat{P})$ *be an optimal solution of problem (6) for* $m \geq 2$ *and* $n, d \geq 1$ *and take the regularization parameters* λ_1 *and* λ_2 *as in (18). Then under Assumption 1, the following results hold with the probability of as least* $1 - exp(-\frac{1}{2}(t - dmlog(1 + \frac{t}{dm})))(t > 0)$

$$\frac{1}{mn}\sum_{i=1}^{m} ||X_i^T vec(\hat{H} + \hat{P}) - vec(F^*)||^2 \leq (\frac{2m\sqrt{m\lambda_1\lambda_3}}{k_1(r)} + \frac{2\lambda_2\sqrt{c}}{k_2(c)})^2 \tag{21}$$

$$||\hat{H} - H^*||_2 \leq \frac{(\beta_1 + 1)}{k_1(r)}(\frac{2m\sqrt{m\lambda_1\lambda_3}}{k_1(r)} + \frac{2\lambda_2\sqrt{c}}{k_2(c)}) \tag{22}$$

$$||\hat{P} - P^*||_{1,2} \leq \frac{(\beta_2 + 1)\sqrt{c}}{k_2(c)}(\frac{2m\sqrt{m\lambda_1\lambda_3}}{k_1(r)} + \frac{2\lambda_2\sqrt{c}}{k_2(c)}) \tag{23}$$

Then with the same probability, the following two sets

$$\hat{\mathcal{Q}}_1 = \{q_1 | \|\hat{H}\| > \frac{(\beta_1 + 1)}{k_1(r)}(\frac{2m\sqrt{m\lambda_1\lambda_3}}{k_1(r)} + \frac{2\lambda_2\sqrt{c}}{k_2(c)})\} \tag{24}$$

$$\hat{\mathcal{Q}}_2 = \{q_2 | \|\hat{P}\| > \frac{(\beta_2 + 1)\sqrt{c}}{k_2(c)}(\frac{2m\sqrt{m\lambda_1\lambda_3}}{k_1(r)} + \frac{2\lambda_2\sqrt{c}}{k_2(c)})\} \tag{25}$$

estimate the true sparsity pattern $\mathcal{Q}_1(P^*)$ *and* $\mathcal{Q}_2(H^*)$*, respectively, that is* $\hat{\mathcal{Q}}_1 = \mathcal{Q}_1(P^*)$ *and* $\hat{\mathcal{Q}}_2 = \mathcal{Q}_2(H^*)$*.*

5. Experiments and Analysis

In this section, we conduct experiments to demonstrate the effectiveness of the proposed informative relationship learning methods on both synthetic data and four real-world datasets. The proposed methods are compared with the other six comparative MTL methods and the details of experimental setup and results are presented.

5.1. Datasets

A synthetic dataset and four real-world datasets are conducted in our experiments and the generation of synthetic dataset and the detailed description of real-world datasets are presented as follows.

- **Synthetic Data**: A synthetic dataset is created to simulate multi-task learning progression, incorporating sparse and noisy correlations based on a multivariate Gaussian distribution as a generalization. This dataset is set with the number of tasks $m = 25$ and for each task with samples $n_i = 100$ and 100 dimensions. The $m * m$ matrix U_Ω is generated from random sampling from $N(0, 1)$, then each row in U_Ω has 40% probability to be selected and 80% of entries in selected rows will set to be zero with 50% probability. Then the sparse relation matrix Ω^* is generated by $\Omega^* = U_\Omega U_\Omega^T$, $W \in R^{d*m}$ is generated from the multivariate normal distribution $N(0, \Omega^*)$ for each row. Then prediction target is obtained by the linear model with Gaussian white noise as $y_i = X_i^T W_i + 0.3 * N(0, 1)$.

- **School Dataset**: The dataset is from the Inner London Education Authority (ILEA), including examination records of 15362 students from 139 secondary schools in the years 1985, 1986, 1987. 27 binary attributes consisting of year, gender, etc., and 1 bias attribute represented for each sample. The target is the examination score. So there are 139 tasks with each task corresponding to one school.

- **Parkinson's Disease Dataset**: This dataset is composed of a range of biomedical voice measurements from 42 patients to predict the disease symptom score of Parkinson's for patients at different times using 19 bio-medical features [47]. The dataset has 5,875 data points for all patients, and in our experiments, each patient's disease symptom is treated as a task.

- **SmartFert Dataset**: The dataset is intended for evaluating soil health globally. It comprises 354 geographically dispersed sites across 42 countries, documenting agriculture-related variables such as climate, soil type, yield, and fertilization [48]. Following data preprocessing, 12 features are obtainable for four farming operations executed to a consistent standard. The corresponding target entails measuring the amount of fertilizer, including nitrogen, phosphorus, potassium, and sulfur, applied during each month. The sampling time points of prediction tasks in uniform, from January to December of a year, and SmartFert datasets can be divided into four independent task for predicting different elements: SmartFert.N, SmartFert.P, SmartFert.K, and SmartFert.S.

- **Alzheimer's Disease Dataset**: The dataset is from the Alzheimer's Disease Neuroimaging Initiative (ADNI) funded by the National Institute of Health (NIH) in 2003 [49]. There are five subsets ADAS-cog, MMSE and RAVLT.TOTAL, RAVLT.TOT5 and RAVLT.TOT30. 1092 patients (samples) with 314 MRI features with 5 categories: surface area, volume of white matter parcellation, volume of cortical parcellation, standard deviation and average of cortical thickness. The target is cognitive scores from 12 time points: M00, M06, M12, M24, M36, M48, M60, M72, M84, M96, M108 and M120. We remove the samples with fail the MRI quality controls and with missing entries. And the sample sizes corresponding 12 tasks are 1074, 1064, 1014, 867, 565, 483, 299, 327, 259, 200, 118 and 69, respectively.

5.2. Comparative methods

To evaluate the efficacy and demonstrate the competitiveness of the proposed informative relationship learning models, we compare them with the following methods that are closely related to the task relationship learning domain to verify the prediction performance and ablation.

- **Multi-task learning with exclusive Lasso** (eLasso) in [50]:

$$\underset{W}{min}\, \mathcal{L}(W) + \lambda_1 ||W||_{1,2} \tag{26}$$

  A group regularization models the scenario when variables in the same group compete with each other which has been widely used to estimate the predictive tasks meanwhile modelling the exclusive features at different subtasks separately.

- **Robust multi-task learning** (RMTL) in [24]:

$$\underset{H,P}{min}\, \mathcal{L}(W) + \lambda_1 ||H||_* + \lambda_2 ||P||_{1,2} \quad s.t. \quad W = H + P \tag{27}$$

  RMTL integrates the low-rank regularization under the assumption that all tasks share a common feature set characterized by the trace norm. The proposed robustness method is achieved by similar decomposition in our method, but the implicit theory is to model the tasks' parameters with abnormal values by exclusive Lasso as 'outlier tasks', which we regard as a composition of sharing feature structure and tasks' specific characteristics.

- **Robust multi-task feature learning** (RMTFL) in [25]:

$$\underset{H,P}{min}\,\mathcal{L}(W) + \lambda_1||H||_{2,1} + \lambda_2||P||_{1,2} \quad s.t. \quad W = H + P \tag{28}$$

RMTFL employs the $l_{2,1}-$norm to capture the feature selected across the whole disease progression and $l_{1,2}-$norm to capture the exclusive feature of particular time points. The robustness method focuses the knowledge interaction on the feature level and decomposes the feature into common and specific subsets.

- **Sparse Multi-task relationship learning** (SPATS) in [35]:

$$\underset{W}{min}\,\mathcal{L}(W) + \lambda_1 tr(W\Omega^{-1}W^T) + \lambda_2||\Omega||_1 \tag{29}$$

SPATS learns the pairwise relationships of tasks under the Gaussian graphical prior, and the same sparsity-inducing component in correlation learning. The SPATS utilizes the covariance based method to judge the similarity of two tasks, meanwhile handling a situation in which the number of tasks is pretty large.

- **Multi-task exclusive relationship learning** (MTERL) in [40]:

$$\underset{W,\Omega}{min}\,\mathcal{L}(W) + \lambda_1 tr(W\Omega^{-1}W^T) + \lambda_2||W||_{1,2} \tag{30}$$

The MTERL method learns the pairwise 'exclusive relationship' of the disease progression, employing the exclusive Lasso to capture the prominent characteristics in different disease stages. The main difference between us is the task relationship definition, MTERL explores the relationship in the whole parameter space, while we focus on the knowledge sharing stage in which tasks' interaction happens.

- **Multi-task robust temporal smoothness learning** (RoTS) in [13]:

$$\underset{H,P}{min}\,\mathcal{L}(W) + \lambda_1||RH^T||_{1,1} + \lambda_2||P||_{2,1} \quad s.t. \quad W = H + P \tag{31}$$

The RoTS method utilizes the widely used temporal smoothness (TS) assumption in temporal real-world progression modelling, in which the variance between adjacent time points is relatively small and modelled as total variation reconstruction $||RH^T||_{1,1}$. As a prior knowledge complementary to ambiguous chronological research, TS has been applied in [10, 11, 42] and so on. A similar robustness method is in [24, 25], and it is worth emphasising that we provide a more detailed and intuitive definition of the informative relationship, rather than utilizing the similarity of tasks only. In our experiments, we choose the RoTS method as a competition of tasks' similarity and informative relationship.

5.3. Experimental settings

For the quantitative performance evaluation, we employ the metrics of the root mean squared error (rMSE) for aggregated performance over all tasks and the single subtask evaluation.

$$rMSE(Y, \hat{Y}) = \sqrt{\frac{||Y - \hat{Y}||_2^2}{n}} \tag{32}$$

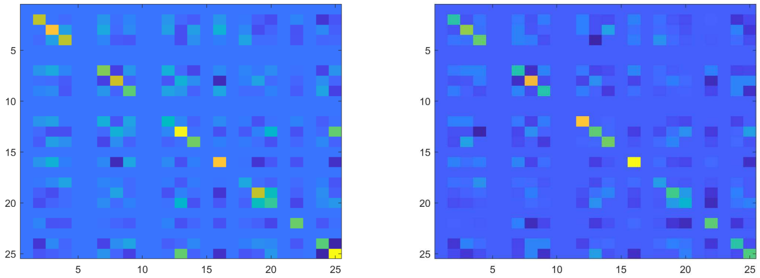
where Y and $\hat{Y}$ are the ground truth prediction target and predicted target value, and n is the number of samples. We partitioned the data into training sets and testing sets using various training ratios. We conducted 10 trials with 5-fold cross-validation to select the optimal hyperparameters $(\lambda_1, \lambda_2, \lambda_3)$, and then utilized these selected hyperparameters to optimize individual prediction performance on the test sets. The regularization parameters were selected from a logarithmic scale spanning from 10^{-3} to 10^3. This range was chosen to accommodate the diverse variances observed in real-world datasets, thereby allowing for an effective structural risk penalization across various datasets and its intrinsic distribution. We randomly select training samples with different training ratios for each dataset and use the rest to test the generalization performance. For synthetic data, we consider 80% as the training ratio. For real-world datasets, we take 60%, 70%, 80%, and 90% as the training ratios. According to the rMSE, we choose the best parameters on the training set and report the performance on the report the performance on the test test for all methods.

5.4. Experimental results on synthetic data

To evaluate the accuracy of the proposed approaches in terms of relationship learning, we first evaluate the performance of the synthetic dataset with

Table 1: Prediction performance comparison of models in terms of rMSE (lower is better). 80% of data is used as training data, and shown data in this table is mean value ± standard derivation.

Error	SPATS	MTERL	MTIRL	MTSIRL
Prediction Error on Y	0.6879±0.1397	0.7311±0.2206	0.6673±0.1126	**0.6651±0.1034**
Estimation Error on Ω^*	0.9366±0.0035	0.9478±0.0017	0.9627±0.0021	**0.9215±0.0014**



(a) Ω^*

(b) $\hat{\Omega}$

Figure 1: Comparison between the true task covariance Ω^* and the estimation by our MTSIRL method $\hat{\Omega}$ on the synthetic dataset.

the competitive MTL relationship learning methods, SPATS and MTERL under the noisy simulated environment. The performance of learning the relationship matrix is shown in Table 1, and we show the rMSE on prediction error in target Y and the error between the ground truth Ω^* and estimation relation matrix $\hat{\Omega}$. As the illustration in Table 1, the proposed sparse informative relationship MTL method shows a better performance in both target prediction and relationship matrix estimation. Noticed that the MTERL method shows a weak performance on both targets, while it is important to emphasise the difference between 'exclusive relationship' and 'informative relationship', which former focuses on the tasks' outward manifestation and we model the process tasks learned from each other, thus better robustness under the noisy condition. A visual demonstration is shown in Figure 1, compared with the true covariance matrix generated from simulation and estimated by our proposed method. The lighter colour indicates the bigger value in the matrix, though the estimated matrix provides a fair performance under a noisy condition, while there is still a small estimation bias in our model. The main reason is the disturbance influences all the tasks with common addictive noise which leads to an error in informative relationship

Table 2: Prediction performance comparison of models in terms of rMSE (lower is better) in the School dataset. The ratio shows the percent used as training data, and the data shown in this table is the mean value ± standard derivation.

Training Ratio	0.6	0.7	0.8	0.9
eLasso	11.4049±0.1154	11.1458±0.1064	10.9882±0.2367	10.9340±0.2317
RMTL	11.0546±0.1311	10.8888±0.1409	10.7291±0.1374	10.6768±0.2312
RMTFL	11.4165±0.1187	11.1829±0.2049	11.0227±0.1923	10.9809±0.2564
SPATS	11.6569±0.1407	11.4516±0.0873	11.2184±0.1618	11.2143±0.2790
MTERL	11.3873±0.1231	11.1368±0.1344	10.9668±0.1929	10.9648±0.2566
RoTS	10.6004±0.0631	10.4908±0.0938	10.4131±0.1511	10.3807±0.1351
MTIRL	10.6145±0.0879	10.5852±0.1488	10.5257±0.1230	10.3204±0.1775
MTSIRL	**10.3051±0.0562**	**10.0596±0.0868**	**10.3608±0.1456**	**10.0046±0.1642**

Table 3: Prediction performance comparison of models in terms of rMSE (lower is better) in the Parkinson's dataset. The ratio shows the percent used as training data, and the data shown in this table is the mean value ± standard derivation.

Methods	0.6	0.7	0.8	0.9
eLasso	0.1549±0.0582	0.1655±0.0609	0.1433±0.1067	0.1556±0.1082
RMTL	0.1415±0.0411	0.1753±0.0508	0.1681±0.0641	0.1481±0.1071
RMTFL	0.1933±0.0809	0.1660±0.0961	0.1408±0.1064	0.2115±0.1605
SPATS	0.1705±0.0659	0.1684±0.0190	0.1055±0.0592	0.1030±0.0957
MTERL	0.1599±0.0522	0.1440±0.0829	0.1408±0.0705	0.1061±0.0716
RoTS	0.1503±0.0504	0.1167±0.0536	**0.0906±0.0354**	**0.0859±0.0487**
MTIRL	0.1439±0.0590	0.1608±0.1012	0.1119±0.0723	0.1160±0.1234
MTSIRL	**0.1361±0.0663**	**0.1090±0.0515**	0.1105±0.0779	0.0897±0.1144

estimation.

5.5. Experimental results on real-world data

The prediction performance conducted on the School dataset is shown in Table 2. With the rise of the training ratio, most methods improve the predictive performance because of the larger number of training instances. Our proposed methods perform well in situations with different training ratios, while RoTS show an approximate predictive performance surprisingly. Though the predictive tasks have no apparent temporal relation among the schools, they still can illustrate the improvement of performance by the utilization of tasks's similarity. Besides, the MTSIRL show better performance in comparison with the SPATS and MTERL, and it illustrates the efficiency of the proposed decomposition method in the task relationship learning approaches.

We demonstrate the prediction performance on the Parkinson's dataset in Table 3. Our proposed methods exhibit superior performance with lower

Table 4: Prediction performance comparison of models in terms of rMSE (lower is better) in SmartFert. The ratio shows the percent used as training data, and the data shown in this table is the mean value ± standard derivation.

Dataset	Methods	0.6	0.7	0.8	0.9
N	eLasso	55.2339±9.8120	50.0739±11.4340	51.0054±14.4822	52.1875±14.3144
	RMTL	60.0938±12.8970	59.3696±7.7862	46.4736±15.4610	40.4928±12.4899
	RMTFL	66.8432±22.0246	59.3716±14.2369	52.2995±16.8092	47.7635±18.2358
	SPATS	58.9708±19.6251	57.2366±16.8723	54.0776±13.5166	43.0291±8.8552
	MTERL	64.9306±15.6415	52.1219±9.6701	51.1431±9.8340	40.6653±8.4338
	RoTS	52.7433±12.5757	48.2904±7.1747	43.2738±8.3501	33.1913±8.5101
	MTIRL	55.0871±19.2175	46.4938±7.6625	**43.1323±5.6423**	32.0887±12.3462
	MTSIRL	**51.0373±13.9577**	**43.0506±6.9202**	45.0990±10.7166	**30.8191±6.9996**
P	eLasso	15.1175±5.7498	16.1312±4.8091	16.5736±5.7106	15.2477±9.9639
	RMTL	15.0136±4.9621	12.8546±2.4684	13.2584±6.1606	14.0132±3.3399
	RMTFL	14.8872±4.6189	16.9388±6.2309	16.2925±2.7575	14.0715±6.4828
	SPATS	14.3851±4.4055	14.1183±3.2163	14.3186±8.1646	15.9009±6.8176
	MTERL	16.1812±4.9246	14.2648±6.1563	14.1543±6.4957	13.6839±3.3818
	RoTS	14.5543±4.2736	13.4671±3.1534	14.6449±4.9837	12.1325±3.7169
	MTIRL	**13.6522±3.6585**	14.2895±4.8874	13.4039±4.4114	12.1931±4.9383
	MTSIRL	14.8991±6.5781	**13.4004±4.0087**	**12.9484±3.6630**	**11.2235±4.0178**
K	eLasso	39.6869±23.0632	34.6040±9.0395	38.6637±8.64882	36.2285±16.2771
	RMTL	40.1564±24.2115	36.7602±8.1751	38.5117±7.4937	36.3774±9.3944
	RMTFL	38.5617±24.5291	33.5959±5.8909	37.1658±3.3030	32.1130±9.3993
	SPATS	33.0144±7.3822	32.1374±6.9506	33.9420±7.5602	30.0107±12.9808
	MTERL	38.1141±17.7324	34.1170±10.9482	32.2685±6.6798	32.9576±8.7970
	RoTS	37.7715±13.5931	32.4376±5.0282	33.9695±4.5422	32.0179±16.0658
	MTIRL	34.2518±21.1542	**30.9235±4.6454**	30.9142±8.0719	**29.0153±12.4076**
	MTSIRL	**32.3273±20.4688**	32.0831±6.6913	**29.2353±6.4537**	31.8171±15.5351
S	eLasso	23.3940±7.8800	22.3194±8.4704	23.5501±8.4900	20.8678±10.8733
	RMTL	22.8227±6.2186	25.3811±7.2993	23.4062±9.2389	17.5113±10.6652
	RMTFL	21.6784±5.9079	22.5237±4.4219	21.5818±4.4859	18.5400±10.7995
	SPATS	20.1999±3.4101	22.8633±10.2144	20.6831±3.6549	16.7218±4.1018
	MTERL	24.3386±9.4860	21.5714±4.9416	23.2879±11.3832	18.2521±5.3309
	RoTS	22.8689±5.1246	21.6916±4.3283	17.8385±4.6142	16.9938±7.8317
	MTIRL	22.5401±9.5597	21.1902±7.4591	**17.3517±3.3483**	16.0500±5.4733
	MTSIRL	**19.7331±4.2782**	**17.2486±4.2821**	17.7444±3.8730	**15.3200±6.0198**

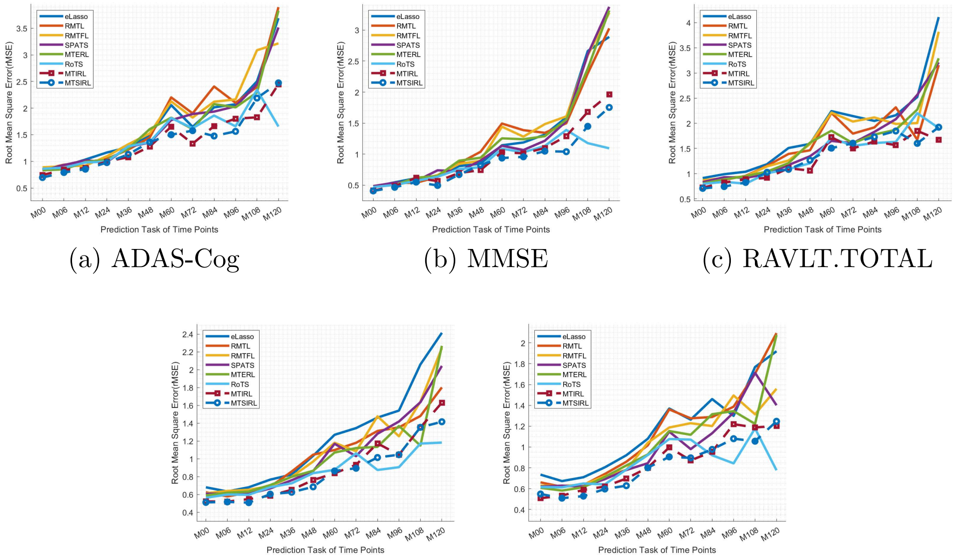
training ratios but perform worse with higher training ratios. This aberration can be attributed to the neglect of the relationship between biomedical voice measurements from subjects in our framework. In this dataset, training samples consist of diverse biomedical voice measurements, leading to greater disparities between the training and test sets as the training ratio increases. In our experiments, with a small number of training samples, implying more biomedical voice measurements are allocated to the test set, the variance of features between the training and test data is relatively small. Conversely, with a higher training ratio, the predictive target exhibits a larger gap due to differences in the biomedical voice measurements.

The prediction performance comparison of models in terms of rMSE conducted on the SmartFert dataset is shown in Table 4. The predictive target in this dataset is to predict the nitrogen, phosphorus, potassium and sulfur content in the soil, while the content of the above elements changes over the seasons and vary from different farms. The proposed informative relationship learning methods outperform others in most single prediction tasks. Noticing that the feature based learning methods with similar decomposition (RMTL and RMTFL) show a relatively larger prediction error, it is reasonable to infer that it is caused by the changes of characteristics in different time points. Specifically, the overall prediction performance of task SmartFert.P stands out due to the relatively small variation in its predictive targets compared to the other tasks. In contrast, all four predictive tasks in this dataset exhibit significant variations, leading to poorer prediction performance. These fluctuations among prediction tasks create challenges in joint learning. A direct consequence is that modeling the progression relationship using decomposition methods (RoTS, MTIRL, and MTSIRL) shows improved prediction performance across tasks. Although a difference in subspace learning, the decomposition methods perform well in the condition that the predictive tasks are various. Another piece of evidence is the performance of RMTL and RMTFL, the assumption that tasks share a subset of features is suitable to some extent in SmartFert.P and SmartFert.S, while inappropriate in the circumstance that the tasks' similarity is lower, such as SmartFert.N and SmartFert.K.

The average and standard deviation of performance measures in terms of the rMSE conducted on the Alzheimer's disease datasets is shown in Table 5 and the performance of each subtask on different time points is shown in Figure 2. Our proposed methods show a better performance in both overall predictive tasks and each time points. In the ADNI cohorts, the data is collected from the baseline to M120 with long-term research up to ten years after the first time checking, and the modelling of progression is challenging. The MTERL learns the exclusive relationship of disease and the RoTS method simulates the gradual changes in progression. The common tendency of all plots in Figure 2 is as time goes by, the worse prediction performance of each task. The main reason why this phenomenon occurs is many patients dropping out from the ADNI study thus the number of instances decreases with the passage of time. Besides, an abnormal fluctuation of the performance curve happens in the district from M48 to M96, especially in ADAS-Cog, RAVLT.TOT6 and RAVLT.TOT30. It seems that the perfor-

Table 5: Prediction performance comparison of models in terms of rMSE (lower is better) in ADNI. The ratio shows the percent used as training data, and the data shown in this table is the mean value ± standard derivation.

Dataset	Methods	0.6	0.7	0.8	0.9
ADAS-Cog	eLasso	15.7015±0.3779	15.0552±0.3445	14.4239±0.2312	14.0114±0.4793
	RMTL	15.3409±0.3742	14.9444±0.1723	14.4464±0.2039	13.9948±0.5116
	RMTFL	15.8653±0.1677	15.1849±0.3523	14.5580±0.2676	14.3239±.05271
	SPATS	12.9562±0.2402	12.5772±0.1557	12.0569±0.2909	12.1118±0.4039
	MTERL	14.8580±0.3246	14.7334±0.4160	14.1492±0.2842	13.8066±0.3446
	RoTS	14.6217±0.3565	14.1171±0.3596	13.7733±0.3288	13.2014±0.2698
	MTIRL	12.3176±0.2070	**12.1113±0.1425**	**11.6162±0.2099**	**11.4096±0.4266**
	MTSIRL	**12.3014±0.2438**	12.1599±0.2006	11.7633±0.2554	11.4111±0.5497
MMSE	eLasso	11.4141±0.5497	8.8135±0.2054	8.5671±0.2199	8.8074±0.5369
	RMTL	8.2303±0.3421	8.0837±0.2119	7.6425±0.2389	7.1122±0.2645
	RMTFL	9.7131±0.4078	8.9356±0.2046	8.7854±0.3006	8.4279±0.4041
	SPATS	9.0279±0.1907	8.3300±0.2324	7.8631±0.3675	7.5838±0.1194
	MTERL	9.1281±0.1998	8.6031±0.2298	8.3006±0.4531	8.0607±0.5817
	RoTS	6.3727±0.2925	6.5939±0.1747	6.3763±0.2058	6.3658±0.2495
	MTIRL	6.4126±0.2605	**6.1824±0.1216**	6.4403±0.1444	6.2120±0.2650
	MTSIRL	**6.2746±0.3405**	6.5803±0.1522	**6.3429±0.2143**	**6.0882±0.2948**
RAVLT.TOTAL	eLasso	5.3755±0.0977	5.0678±0.0998	4.8359±0.1637	4.7767±0.1250
	RMTL	4.8676±0.0766	4.6387±0.0520	4.4673±0.1479	4.1540±0.1088
	RMTFL	5.4789±0.1167	5.0990±0.1045	4.9598±0.1287	4.7922±0.1193
	SPATS	4.8855±0.0745	4.6336±0.0883	4.4870±0.1449	4.4340±0.1866
	MTERL	5.5214±0.0629	4.9767±0.0399	4.8003±0.1121	4.6730±0.1359
	RoTS	4.4387±0.0774	4.2558±0.0974	4.1062±0.0483	4.0188±0.0780
	MTIRL	4.4354±0.0882	4.2642±0.1186	4.1092±0.1049	3.9025±0.1106
	MTSIRL	**4.4152±0.0985**	**4.1763±0.0658**	**4.0783±0.0728**	**3.7872±0.1149**
RAVLT.TOT6	eLasso	6.0839±0.1246	5.7028±0.1175	5.5799±0.0720	5.3902±0.1335
	RMTL	5.7678±0.0746	5.4690±0.1087	5.3087±0.1008	5.1461±0.1611
	RMTFL	6.1640±0.1380	5.7388±0.0865	5.5873±0.1126	5.4081±0.0562
	SPATS	5.2230±0.0741	5.0903±0.1009	5.0707±0.0601	4.9324±0.1000
	MTERL	5.8138±0.1123	5.6149±0.1032	5.3992±0.1264	5.3348±0.1904
	RoTS	5.4959±0.0971	5.2071±0.0717	5.0477±0.0484	4.9265±0.1429
	MTIRL	5.0586±0.0646	4.8730±0.0701	**4.7965±0.0534**	**4.6073±0.1377**
	MTSIRL	**5.0382±0.0890**	**4.8664±0.0788**	4.7995±0.0465	4.6478±0.1297
RAVLT.TOT30	eLasso	6.3493±0.1375	5.9100±0.1299	5.7456±0.0992	5.5945±0.2289
	RMTL	6.1246±0.1277	5.7824±0.1237	5.6081±0.0831	5.5057±0.1852
	RMTFL	6.5142±0.1183	6.0505±0.1729	5.8153±0.1213	5.5249±0.1909
	SPATS	5.3403±0.0609	5.2262±0.0416	5.0116±0.0958	4.9463±0.1135
	MTERL	6.0730±0.1268	5.8038±0.0861	5.6127±0.0908	5.4627±0.2004
	RoTS	5.8117±0.1129	5.5374±0.1021	5.3659±0.0755	5.2364±0.1422
	MTIRL	**5.2528±0.0702**	**5.0927±0.0862**	5.0214±0.0765	4.9016±0.1094
	MTSIRL	5.2714±0.0783	5.1402±0.0807	**4.9709±0.0817**	**4.8509±0.1432**



(a) ADAS-Cog

(b) MMSE

(c) RAVLT.TOTAL

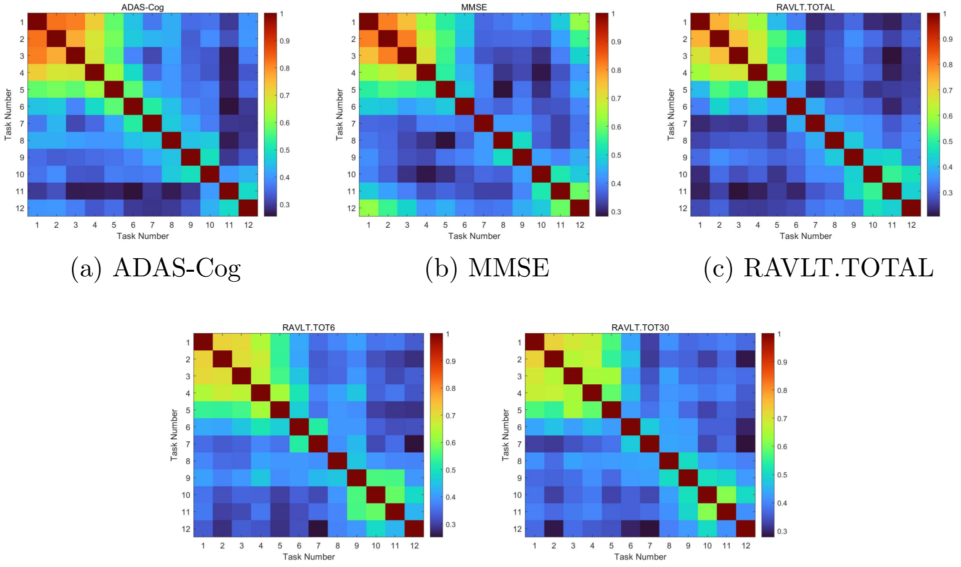
(d) RAVLT.TOT6

(e) RAVLT.TOT30

Figure 2: Prediction performance comparison of each time point in terms of rMSE(lower is better).

mance curves do not follow the number of samples at these time points. The consistency of the progression model changes in this section, no matter the approaches to modelling the disease. It means that both the shared feature sets (eLasso, RMTL and RMTFL) and the chronological prior knowledge (RoTs) do not explicitly capture the progression to some extent. While the MTSIRL method with sparse prior over relationships does not exhibit superior performance compared to MTIRL, this can be attributed to the sparsity of samples for progression estimation, especially for tasks in later sequences.

For a further discussion in AD progression, We show the learned informative relationship matrices from ADNI datasets on Figure 3. The learned informative relationship matrices are accessed from the cross-validation and for a better visual illustration, the learned covariance matrices are transformed into relationship matrices with regularization. In figure 3, first, all of the relationship matrices show a high correlation in the initial stages of the disease, including the baseline, M06, M12, M24 and M48. It reveals a relatively close relationship in the beginning stages, thus there are similarities measured by the proposed frameworks in those cognitive prediction tasks.

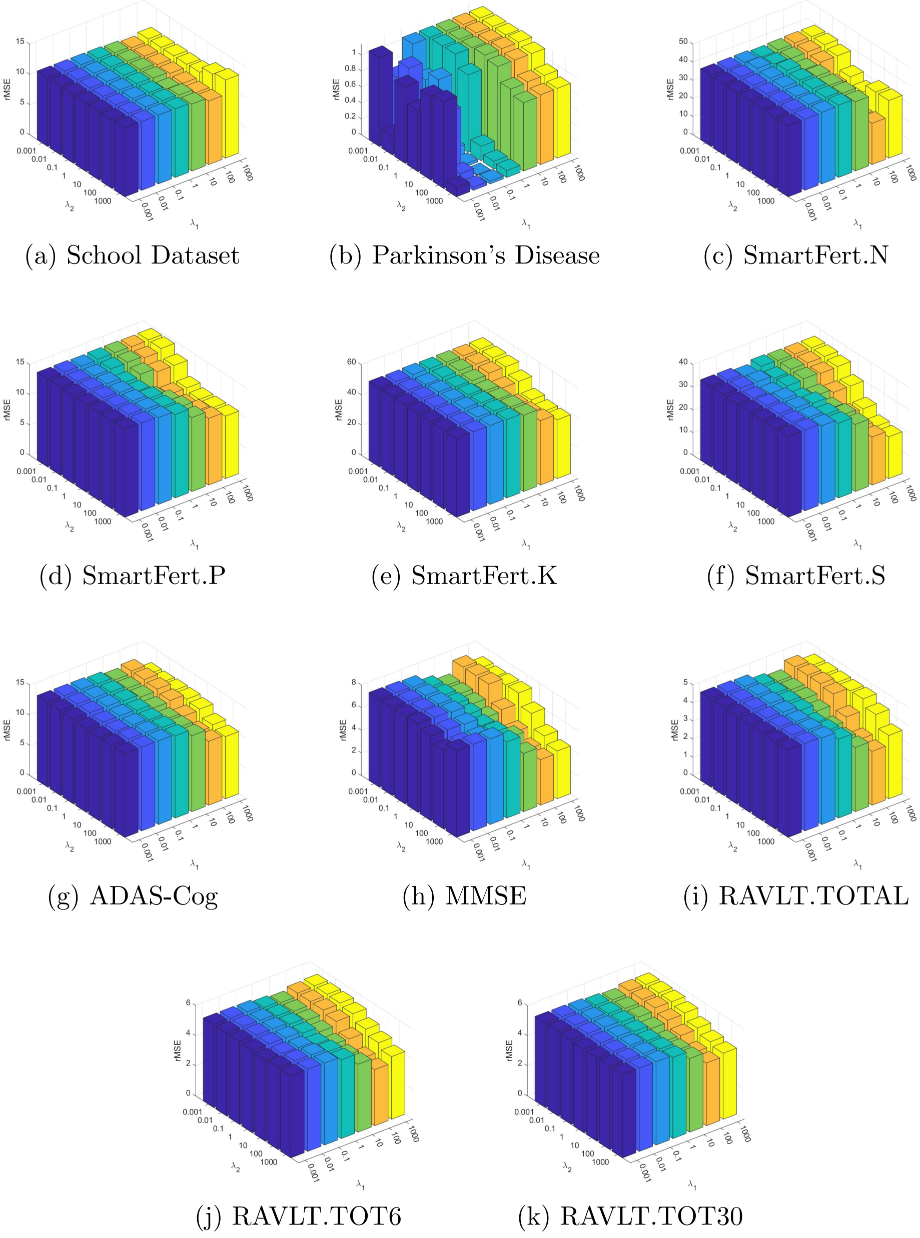


(a) ADAS-Cog (b) MMSE (c) RAVLT.TOTAL

(d) RAVLT.TOT6 (e) RAVLT.TOT30

Figure 3: The informative relationship matrices learned by MTSIRL from the ADNI datasets.

Secondly, the learned informative relationship is quite similar in the latter stages, M96, M108 and M120 in the prediction target of RAVLT, M108 and M120 in the prediction target of ADAS-Cog and MMSE. It means that the changes in the brain gradually come to stability in critical patients. Thirdly, through the learned informative relationship learning, it is shown that coherent tasks in the middle of disease progression, from M60 to M96, appear less similar to others, even compared with adjacent time points. Contacting the abnormal fluctuation in the prediction performance curves in Figure 2, the complex changes in the middle of disease progression bring a challenge in both modelling and determination, because of the limited knowledge sharing under our proposed framework. In conclusion, the proposed informative relationship learning methods provide a deep insight into exploring the disease progression, the experiment results show the limitation of the temporal smoothness prior assumption, and more sophisticated models need to capture the variance of the different stages of Alzheimer's disease.



(a) School Dataset

(b) Parkinson's Disease

(c) SmartFert.N

(d) SmartFert.P

(e) SmartFert.K

(f) SmartFert.S

(g) ADAS-Cog

(h) MMSE

(i) RAVLT.TOTAL

(j) RAVLT.TOT6

(k) RAVLT.TOT30

Figure 4: The prediction performance with variety of configuration for parameters.

Table 6: Average time cost of proposed methods in real-world datasets.

Methods	Cost	School	Parkinson's disease	SmartFert	ADNI
MTIRL	Iterations	93.8000±0.4216	927.9000±65.7139	14.2000±3.7347	327.6000±4.6714
	Time	0.6000±0.1464	0.0500±0.0293	0.0063±0.0132	6.7390±0.6372
MTSIRL	Iterations	93.7000±0.4830	943.4000±46.4954	13.5000±6.6207	331.7000±4.3982
	Time	0.6438±0.1627	0.0484±0.0226	0.0031±0.0099	6.8172±0.3655

5.6. Parameter Sensitivity

For all real-world datasets evaluated in our experiments, we conduct parameter sensitivity analysis based on λ_1 and λ_2 in the MTIRL method. The training ratio is set as 0.9 and the parameters are chosen from the same scale in prediction performance experiments. The result of parameter sensitivity is presented in Figure 4. The performance curves help us to learn the relative importance of shared knowledge and task-specific characteristics under the predictive tasks, as the optimal hyperparameters indicate the variance of data modelling constraints. In the School dataset, the optimal performance at $\lambda_1 = 100$ and $\lambda_2 = 1000$ highlights that task-specific characteristics play a significant role compared to shared knowledge. Conversely, in the Parkinson's disease dataset, the parameter selection indicates less task similarity, treating independent patients with biomedical voice measurements as separate tasks and disregarding their relatedness. On the SmartFert dataset, task similarity is notably important, particularly evident with the turning point at $\lambda_1 = 100$ in the SmartFert.N dataset. In SmartFert.P, SmartFert.K, and SmartFert.S datasets, predictive performance improves as λ_1 and λ_2 increase, which appears contradictory in relative importance analysis. However, it's crucial to note that the significant changes in predictive targets within SmartFert datasets suggest that similarity stems more from prior knowledge than from model parameters. The informative relationship method offers flexibility in learning interactions among tasks. However, in the ADNI dataset, higher values of λ_1 do not consistently lead to improved performance, especially in MMSE and RAVLT.TOTAL datasets. This inconsistency underscores the necessity for accurately simulating disease progression at each time point rather than relying solely on task cohesion to enhance overall performance.

5.7. Time Complexity

Besides the convergency rate and calculation complexity analyses in Section 3, the time complexity of the proposed informative relationship learning methods in real-world datasets is presented in Table 6. Our experiments are conducted on the AMD RYZEN 5800H CPU with 3.20GHz, and the threshold is settled as 10^{-6} when the function value between adjacent iterations is smaller than it, then the iterative progression stops.

Notably, the alternating method employed in optimizing Ω within MTSIRL exhibits comparable convergence rates to directly solving MTIRL. Moreover, despite the additional sparse constraint in informative relationship learning, MTSIRL demonstrates strong scalability in handling a large number of tasks, exhibiting superior performance in both prediction and relation estimation across most experimental scenarios.

However, a higher time cost for each iteration of the objective function optimization is observed, especially notable in the School and ADNI datasets, albeit for distinct reasons. In the School dataset, which features a large number of tasks, the predominant time consumption arises from learning pairwise informative relationship matrices. This process involves extensive computation to establish relationships between multiple tasks. Conversely, in the ADNI dataset, the primary computational burden is attributed to the high-dimensional neuroimaging features. The complexity of processing these features significantly impacts calculation costs, necessitating substantial computational resources. In this paper, the analytic solution is directly adopted in the optimization step to optimize the time complexity. Therefore, for broader practical applications, it is crucial to carefully consider and manage the dimensionality of the data. Efficient handling of data dimensions can mitigate computational overheads and enhance the feasibility of applying the proposed method across different datasets and scenarios.

6. Discussion

We propose a novel multi-task informative relationship learning framework to explore the latent tasks' shared knowledge dependency of multi-task learning. To overcome the compulsory similarity constraints in past relationship learning studies, we employ a decomposition method of learned model parameters, for both informative relationships exploration under the knowledge sharing across the process and tasks' specific features capture. We argue the prior knowledge similarity of tasks assumption which restrains the representation of task-specific characteristics with unclear knowledge sharing mechanism. By applying it to the real-world cohorts, we point out that task dependency changes in different stages, and the existing prior knowledge can not model well the latent association of tasks.

The primary contribution is we propose the concept named 'informative relationship', which means the task relationships in the knowledge-sharing stage of MTL. It helps us to explore the underlying interactive information across all tasks. We utilize graphical regularization to uncover the pairwise informative contributions of tasks using the covariance matrix. This approach offers advantages in terms of generalization performance in data modeling, statistical properties under the central limit theorem, and cost-effectiveness in calculating mutual information. Additionally, we apply an exclusive Lasso constraint to capture unique features within different subtasks. This combined approach allows for a comprehensive exploration of task relationships while efficiently capturing task-specific characteristics.

There are still several limitations to be considered in this study. First, the current research is from the viewpoint of knowledge sharing, while the pairwise knowledge contribution is based on the overall features. It will be helpful to learn the informative relationship with the categories of features to obtain better interpretability, despite the rising expert prior knowledge and calculation cost. Besides, the proposed methods are based on the regularized framework, and fit to extend to the neural networks with the same network structure for each subtask, while lacking generalization in learning with various tasks with diverse network structures due to the unmatched dimensional model parameters. An important work worth studying is about informative relationship learning in multi-modal models to determine the latent knowledge dependency of each modal.

7. Conclusion

In this paper, we propose a novel multi-task informative relationship learning model aimed at uncovering the dependencies of shared knowledge across all tasks. Specifically, our approach utilizes a decomposition method to determine the pairwise knowledge contribution matrix and capture the specific features of each task. This model addresses weaknesses in MTL relationship learning by providing insights into the underlying dependencies through mutual information analysis. Theoretical analysis and experiment results demonstrate the effectiveness of our proposed framework. Furthermore, we discuss some limitations and extensions of informative relationship learning for better model interpretability and data mining.

8. Acknowledgement

This research was supported by the National Natural Science Foundation of China (No. 62061050) and the 14th Postgraduate Research Innovation Project of Yunnan University (KC-22221162).

References

[1] R. Caruana, Multitask learning, Machine learning 28 (1997) 41–75.

[2] J. Baxter, A bayesian/information theoretic model of learning to learn via multiple task sampling, Machine learning 28 (1997) 7–39.

[3] Y. Zhang, Q. Yang, A survey on multi-task learning, IEEE Transactions on Knowledge and Data Engineering 34 (12) (2021) 5586–5609.

[4] X. Sun, R. Panda, R. Feris, K. Saenko, Adashare: Learning what to share for efficient deep multi-task learning, Advances in Neural Information Processing Systems 33 (2020) 8728–8740.

[5] T. Heskes, Empirical bayes for learning to learn (2000).

[6] B. Bakker, T. Heskes, Task clustering and gating for bayesian multitask learning (2003).

[7] S. Thrun, L. Pratt, Learning To Learn, Kluwer Academic Publishers, 1997.

[8] T. Evgeniou, M. Pontil, Regularized multi–task learning, in: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, 2004, pp. 109–117.

[9] T. Evgeniou, C. A. Micchelli, M. Pontil, J. Shawe-Taylor, Learning multiple tasks with kernel methods., Journal of machine learning research 6 (4) (2005).

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

[11] J. Zhou, J. Liu, V. A. Narayan, J. Ye, A. D. N. Initiative, et al., Modeling disease progression via multi-task learning, NeuroImage 78 (2013) 233–248.

[12] M. Zhou, P. Yang, Automatic temporal relation in multi-task learning, in: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2023, pp. 3570–3580.

[13] M. Zhou, Y. Zhang, Y. Yang, T. Liu, P. Yang, Robust temporal smoothness in multi-task learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 37, 2023, pp. 11426–11434.

[14] L. Yuan, Q. Zhu, Y. Zheng, W. Dong, Y. Ke, Z. Li, Temporal smoothness framework: analyzing and exploring evolutionary transition behavior in dynamic networks, in: 2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI), IEEE, 2021, pp. 1206–1210.

[15] E. Meyerson, R. Miikkulainen, The traveling observer model: Multi-task learning through spatial variable embeddings, arXiv preprint arXiv:2010.02354 (2020).

[16] L. Romeo, G. Armentano, A. Nicolucci, M. Vespasiani, G. Vespasiani, E. Frontoni, A novel spatio-temporal multi-task approach for the prediction of diabetes-related complication: a cardiopathy case of study., in: IJCAI, 2020, pp. 4299–4305.

[17] Y. Zhang, M. Zhou, T. Liu, V. Lanfranchi, P. Yang, Spatio-temporal tensor multi-task learning for predicting alzheimer's disease in a longitudinal study, in: 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), IEEE, 2022, pp. 979–985.

[18] A. Argyriou, T. Evgeniou, M. Pontil, Multi-task feature learning, in 'advances in neural information processing systems 19' (2007).

[19] G. Obozinski, B. Taskar, M. Jordan, Multi-task feature selection, Statistics Department, UC Berkeley, Tech. Rep 2 (2.2) (2006) 2.

[20] J. Chen, L. Tang, J. Liu, Jieping ye. a convex formulation for learning a shared predictive structure from multiple tasks. pattern analysis and machine intelligence, IEEE Transactions on 35 (5) (2013) 1025–1038.

[21] A. Jalali, S. Sanghavi, C. Ruan, P. Ravikumar, A dirty model for multi-task learning, Advances in neural information processing systems 23 (2010).

[22] F. Nie, Z. Hu, X. Li, Calibrated multi-task learning, in: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018, pp. 2012–2021.

[23] W. Chang, F. Nie, R. Wang, X. Li, Calibrated multi-task subspace learning via binary group structure constraint, Information Sciences 631 (2023) 271–287.

[24] J. Chen, J. Zhou, J. Ye, 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, 2011, pp. 42–50.

[25] P. Gong, J. Ye, C. Zhang, Robust multi-task feature learning, in: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, 2012, pp. 895–903.

[26] L. Jacob, J.-p. Vert, F. Bach, Clustered multi-task learning: A convex formulation, Advances in neural information processing systems 21 (2008).

[27] J. Zhou, J. Chen, J. Ye, Clustered multi-task learning via alternating structure optimization, Advances in neural information processing systems 24 (2011).

[28] S. Liu, Y. Liang, A. Gitter, Loss-balanced task weighting to reduce negative transfer in multi-task learning, in: Proceedings of the AAAI conference on artificial intelligence, Vol. 33, 2019, pp. 9977–9978.

[29] L. Han, Y. Zhang, Learning tree structure in multi-task learning, in: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2015, pp. 397–406.

[30] E. V. Bonilla, K. Chai, C. Williams, Multi-task gaussian process prediction, Advances in neural information processing systems 20 (2007).

[31] Y. Zhang, D.-Y. Yeung, Multi-task learning using generalized t process, in: Proceedings of the thirteenth international conference on artificial intelligence and statistics, JMLR Workshop and Conference Proceedings, 2010, pp. 964–971.

[32] Y. Zhang, D.-Y. Yeung, A regularization approach to learning task relationships in multitask learning, ACM Transactions on Knowledge Discovery from Data (TKDD) 8 (3) (2014) 1–31.

[33] G. Lee, E. Yang, S. Hwang, Asymmetric multi-task learning based on task relatedness and loss, in: International conference on machine learning, PMLR, 2016, pp. 230–238.

[34] A. R. Gonçalves, F. J. Von Zuben, A. Banerjee, Multi-task sparse structure learning with gaussian copula models, The Journal of Machine Learning Research 17 (1) (2016) 1205–1234.

[35] Y. Zhang, Q. Yang, Learning sparse task relations in multi-task learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 31, 2017.

[36] T. Lan, Z. Li, Z. Li, L. Bai, M. Li, F. Tsung, W. Ketter, R. Zhao, C. Zhang, Mm-dag: Multi-task dag learning for multi-modal data–with application for traffic congestion analysis, arXiv preprint arXiv:2306.02831 (2023).

[37] Y. Zhang, Y. Wei, Q. Yang, Learning to multitask, Advances in Neural Information Processing Systems 31 (2018).

[38] X. Chen, H. Sun, C. Ellington, E. Xing, L. Song, Multi-task learning of order-consistent causal graphs, Advances in Neural Information Processing Systems 34 (2021) 11083–11095.

[39] R. Kegan, What "form" transforms?: A constructive-developmental approach to transformative learning, in: Contemporary theories of learning, Routledge, 2018, pp. 29–45.

[40] M. Wang, D. Zhang, D. Shen, M. Liu, Multi-task exclusive relationship learning for alzheimer's disease progression prediction with longitudinal data, Medical image analysis 53 (2019) 111–122.

[41] F. Campbell, G. I. Allen, Within group variable selection through the exclusive lasso (2017).

[42] X. Liu, P. Cao, A. R. Gonçalves, D. Zhao, A. Banerjee, Modeling alzheimer's disease progression with fused laplacian sparse group lasso, ACM Transactions on Knowledge Discovery from Data (TKDD) 12 (6) (2018) 1–35.

[43] S. P. Boyd, L. Vandenberghe, Convex optimization, Cambridge university press, 2004.

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

[45] P. Bühlmann, S. Van De Geer, Statistics for high-dimensional data: methods, theory and applications, Springer Science & Business Media, 2011.

[46] P. J. Bickel, Y. Ritov, A. B. Tsybakov, Simultaneous analysis of lasso and dantzig selector (2009).

[47] M. Little, Parkinsons, UCI Machine Learning Repository, DOI: https://doi.org/10.24432/C59C74 (2008).

[48] Y. Zhang, X. Wang, T. Liu, R. Wang, Y. Li, Q. Xue, P. Yang, Sustainable fertilisation management via tensor multi-task learning using multi-dimensional agricultural data, Journal of Industrial Information Integration 34 (2023) 100461.

[49] C. R. Jack Jr, M. A. Bernstein, N. C. Fox, P. Thompson, G. Alexander, D. Harvey, B. Borowski, P. J. Britson, J. L. Whitwell, C. Ward, et al., 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) (2008) 685–691.

[50] Y. Zhou, R. Jin, S. C.-H. Hoi, Exclusive lasso for multi-task feature selection, in: Proceedings of the thirteenth international conference on artificial intelligence and statistics, JMLR Workshop and Conference Proceedings, 2010, pp. 988–995.